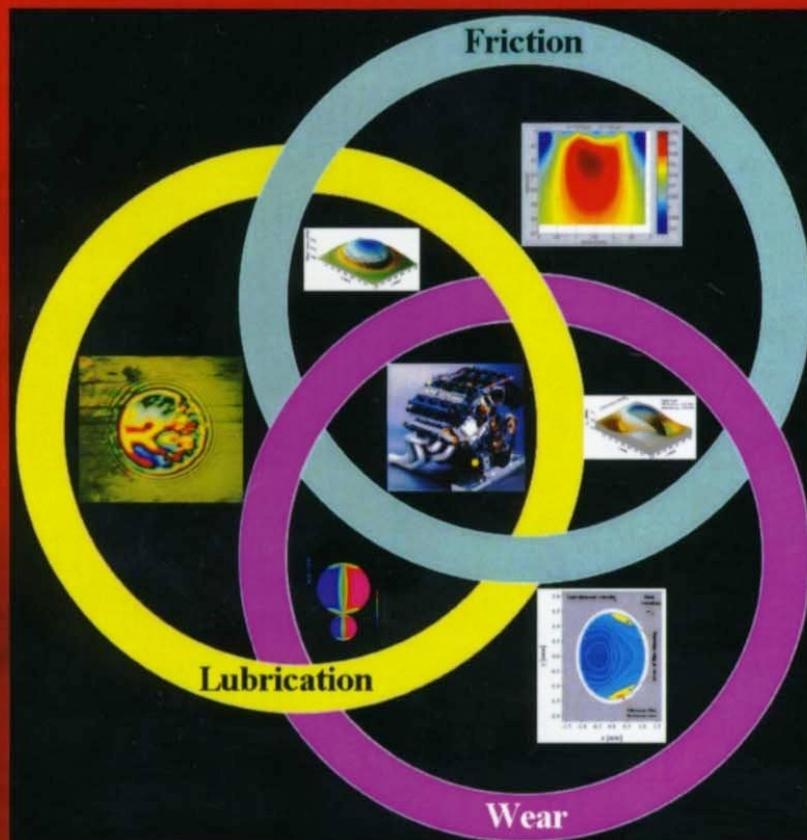


RECENT DEVELOPMENTS IN WEAR PREVENTION, FRICTION AND LUBRICATION

EDITOR
GEORGE K. NIKAS



RESEARCH SIGNPOST

Recent Developments in Wear Prevention, Friction and Lubrication

2010

Editor

George K. Nikas

Research Associate, Tribology Group, Department of Mechanical Engineering
Imperial College London, England



Research Signpost, T.C. 37/661 (2), Fort P.O., Trivandrum-695 023
Kerala, India

Published by Research Signpost

2010; Rights Reserved.

Research Signpost

T.C. 37/661(2), Fort P.O.,

Trivandrum-695 023, Kerala, India

Editor

George K. Nikas

Managing Editor

S. G. Pandalai

Publication Manager

A. Gayathri

Research Signpost and the Editor assume no responsibility
for the opinions and statements advanced by contributors

ISBN: 978-81-308-0377-7

Preface

The knowledge gained from studies on *friction*, *lubrication* and *wear* of contacting surfaces (three terms collectively known as *tribology*) is met in every branch of engineering. Tribology is a multidisciplinary science involving mechanical engineering, materials and lubrication science, physics, and chemistry. Tribological principles are normally used in characterising the mechanical behaviour of surfaces in relative motion. This involves a very large number of applications ranging in size from the nano-scale (molecular and nano-tribology) to the macro-scale (bearings, tyres, rock drills, etc.).

Friction, lubrication and wear are inherent characteristics of the physical world regardless of scale. From the intermolecular forces holding a gecko's feet on a vertical wall to the tractive forces at the contact patches of aircraft tyres; from the lubricating film at the rolling contact of a ball on a rolling-bearing raceway to the soft magma supporting tectonic plates of the earth; from the erosion of human teeth from toothpaste particles to the particle erosion of turbine blades; such diverse phenomena can be described and analysed in tribological terms.

Every type of machinery includes parts affected by friction, lubrication, and wear. Therefore, the role of tribology in machine operation and reliability is a major one. As a result, the effect of tribology to modern world economy, even though unknown to the majority of the population, is crucial. Although the effect on the economy has been quantified and found to represent a significant proportion of the gross domestic product of economically advanced countries, the true effect pertaining to the development of technologically advanced products is far greater and not immediately obvious.

There are several good books on tribology covering the fundamental theory and various applications. However, technology is progressing at an immense pace and engineering norms are quickly outdated. This book is based on the collective experience of some of the world's top experts on friction, lubrication and wear. It presents recent developments in the field of tribology and its contributors are equally divided among theoreticians and experimentalists. Furthermore, some chapters contain new theory and results, which, in the editor's experience, is at the frontier of research. The topics covered in the eight chapters of this book include thin-film lubrication from a theoretical perspective, a critical review of rolling-bearing life-prediction models, a review of Laser Surface Texturing, a proposed unification of friction and wear via thermodynamic principles, the role of tribofilms in concentrated contacts from a materials-science perspective, transient phenomena in elasto-hydrodynamic lubrication from an experimental point of view, a theoretical assessment of the Stribeck curve in lubricated contacts, and, finally, a thorough presentation of modelling and adhesion problems of microelectromechanical systems (MEMS) and miniature devices.

Chapter 1 by Professor Szeri is an in-depth discussion of the “thin-film” lubrication theory referring to hydrodynamic and elastohydrodynamic lubrication. Professor Szeri (University of Delaware, USA), a well-known author with many decades of research experience, attempts to define the limits of the classical Reynolds’ theory of lubrication in both laminar and turbulent flows of liquids and gases. He discusses the continuity limitation of the classical theory in terms of film thickness, wall slip, and the Reynolds number. He reaches to quantitative conclusions on the validity of the film continuity assumption and the necessity to use the full Navier-Stokes equations or “first principles” in some applications for which the Reynolds’ theory is inaccurate.

Chapter 2 by Dr Zaretsky is a critical evaluation of the main models proposed for the prediction of the fatigue lives of rolling-element bearings, comprising the past and present ISO standards. Those models have been adopted by major bearing manufacturers such as SKF and include the classical work of Palmgren, Weibull, Lundberg, Ioannides, Harris, and Zaretsky. Dr Zaretsky, a Chief Engineer of NASA and Adjunct Professor at Case Western Reserve University (USA), is an authority on bearing life prediction. In his comprehensive tutorial, which is suitable both to experts and non-experts on this topic, he explains the advantages and disadvantages of the various models from a theoretical point of view, and includes a selection of endurance test results to compare the relative accuracy of the theoretical models. Among Dr Zaretsky’s conclusions is that the load-life exponent (10/3) used in the ANSI/ABMA and ISO standards to predict rolling bearing life is not reflective of modern rolling bearings and actually underestimates bearing lives.

Chapter 3 by Professor Etsion is a review of Laser Surface Texturing (LST). Surface texturing of tribological, mechanical components, has emerged in the last decade as a viable option of surface engineering resulting in significant improvements in load capacity, wear resistance, reduction of friction etc. Professor Etsion (Chair in Fluid Mechanics and Heat Transfer, Technion – Israel Institute of Technology) is one of the most recognised experts on the topic of surface texturing. In his review of the LST, he outlines the technique and discusses the potential of this technology in various engineering applications, including automotive (piston rings), bearings, seals, magnetic storage devices (hard disk sliders and magnetic tapes), and others.

Chapter 4 by Professor Bryant is rather unique in the literature. Professor Bryant, a professor in the University of Texas at Austin (USA) and the Editor of the Journal of Tribology of the American Society of Mechanical Engineers, attempts to unify friction and wear via thermodynamic principles, focusing on the dissipative processes found at sliding interfaces. He presents a review of friction and wear mechanisms, and relates them in terms of their associated thermodynamics, energy losses and entropy produced by common dissipative processes. He shows that the expressions for entropy generation in dissipative processes operative at a sliding interface, which are common to both friction and wear, could result in unified friction and wear models.

Chapter 5 by Professors Jacobson and Hogmark provides an overview of the role of tribofilms on surfaces in sliding contacts. The authors, both professors of materials science and tribology at Uppsala University in Sweden, have several decades of research experience on such topics. Through illustrative examples based on real applications, they show that tribologically induced surface modifications result in the creation of surface layers or tribofilms ranging in thickness from a few nanometres to tens of micrometres. The surface modifications include topographical changes, formation of micro-cracks, material phase transformations, formation of oxides, formation of solid films by reaction with lubricant additives, material transfer from the counter surface, etc. It is such tribofilms that actually dominate the contact properties between sliding solids in terms of friction and wear and not the original materials.

Chapter 6 by Dr Glovnea, Reader in the School of Engineering and Design at Sussex University in England, presents a review of recent experimental research on non-steady-state elastohydrodynamic lubrication, including transient loading, sudden variation of entrainment speed and variation of micro-geometry. His study is relevant to all industrial machinery involving lubricated contacts and susceptible to vibrations or normal operation involving load and speed variations. Dr Glovnea, who has spent several years studying such phenomena in the Tribology Group of Imperial College London, shows how transient conditions and various parameters affect (normally sub-micrometre) elastohydrodynamic films, their thickness distribution and overall cohesion. It is thus possible to assess the risk of contact damage from film thinning or collapse.

Chapter 7 by Professor Khonsari (Louisiana State University, USA) and Dr Booser, both well-respected researchers in the tribology community and authors of best-selling tribology books, deals with the different lubrication regimes met in contacts at relatively low sliding speeds. These are related via the so-called Stribeck curve, which demonstrates the transition from boundary lubrication at the start-up of motion to mixed lubrication as the speed is increased, and on to full-film lubrication at higher speeds. The review covers issues pertaining to roughness asperity interactions (particularly in the mixed lubrication regime), wear, stick-slip phenomena, and the effects of lubricant additives.

Chapter 8 by Dr Xue (Analog Devices, Inc., USA) and Professor Polycarpou (University of Illinois at Urbana-Champaign, USA) covers microelectromechanical systems (MEMS) from an experimental and theoretical point of view. The authors have extensive experience in this research field and are focused on the problem of adhesion or stiction of MEMS and miniature devices, which reduces their reliability and hinders their advancement and wider commercialisation. They develop an experimentally validated, adhesive-contact model, which is valid for a wide range of adhesion parameter values, covering the practical range of application of MEMS and other miniature devices.

The editor embarked on this project in June 2007 and invited eminent tribologists to make contributions on specific topics based on the editor's plans for the book. The project was completed in about 18 months before the manuscripts were sent to the publisher. The editor is grateful to the authors/contributors of this book for their cooperation and patience throughout the lengthy process of communication, chapter editing, and book compilation. The editor is also grateful to the expert reviewers (see the list of reviewers) who kindly reviewed the chapters, one chapter each reviewer. Finally, the editor would like to thank the publisher, Research Signpost, for the original invitation and kind assistance.

George K. Nikas

E-mail: gnikas@teemail.gr

London, England

March 2009

Contributors

Editor

Dr George K. Nikas

Research Associate

Tribology Group, Department of Mechanical Engineering
Imperial College London, England

Authors

Dr E. R. Booser

Engineering Consultant, USA

Professor Michael D. Bryant

Accenture Endowed Professor of Manufacturing Systems Engineering
Mechanical Engineering Department, University of Texas at Austin, USA

Professor Izhak Etsion

Yeshayahu Winograd Chair in Fluid Mechanics and Heat Transfer
Mechanical Engineering Department, Technion – Israel Institute of Technology,
Israel

Dr Romeo Glovnea

Reader in Mechanical Engineering

School of Engineering and Design, Sussex University, England

Professor Sture Hogmark

Professor Emeritus in Materials Science and Tribology
Ångström Tribomaterials Group, Department of Engineering Sciences
Uppsala University, Sweden

Professor Staffan Jacobson

Professor in Materials Science

Ångström Tribomaterials Group, Department of Engineering Sciences
Uppsala University, Sweden

Professor Michael M. Khonsari
Dow Chemical Endowed Chair in Rotating Machinery
Department of Mechanical Engineering, Louisiana State University, USA

Professor Andreas A. Polycarpou
Kritzer Faculty Scholar; Professor
Department of Mechanical Science and Engineering
University of Illinois at Urbana-Champaign, USA

Professor Andras Z. Szeri
Robert Lyle Spencer Professor of Mechanical Engineering
Department of Mechanical Engineering, University of Delaware, USA

Dr Xiaojie Xue
Analog Devices Inc., USA

Dr Erwin Zaretsky
Chief Engineer (Structures and Materials), NASA Glenn Research Center, USA
Adjunct Professor, Case Western Reserve University, USA

Chapter reviewers

Professor George G. Adams
Professor of Mechanical and Industrial Engineering
Department of Mechanical and Industrial Engineering, Northeastern University,
USA

Professor Liming Chang
Professor of Mechanical Engineering
Department of Mechanical and Nuclear Engineering
The Pennsylvania State University, USA

Professor Rob S. Dwyer-Joyce
Professor of Lubrication Engineering; Head of the Mechanical Engineering
Department
Tribology Group, Department of Mechanical Engineering
The University of Sheffield, England

Professor Ian M. Hutchings

GKN Professor of Manufacturing Engineering, Institute for Manufacturing
University of Cambridge, England

Professor Emeritus Bo O. Jacobson

Professor Emeritus
Machine Elements Division, Mechanical Engineering Department
Lund University, Sweden

Dr George K. Nikas

Research Associate
Tribology Group, Department of Mechanical Engineering
Imperial College London, England

Professor Homer Rahnejat

Professor of Dynamics
Dynamics Research Group, Department of Mechanical, Aeronautical and
Manufacturing Engineering, Loughborough University, England

Professor Richard F. Salant

Georgia Power Distinguished Professor in Mechanical Engineering
The George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology, USA

Professor Ray W. Snidle

Professor
Tribology and Contact Mechanics Research Group, School of Engineering
Cardiff University, UK

Contents

Chapter 1

The thin film approximation in hydrodynamic, including
elastohydrodynamic, lubrication 1
Andras Z. Szeri

Chapter 2

Rolling bearing life prediction, theory, and application 45
Erwin V. Zaretsky

Chapter 3

Laser surface texturing and applications 137
Izhak Etsion

Chapter 4

Unification of friction and wear 159
Michael D. Bryant

Chapter 5

Tribofilms – On the crucial importance of tribologically
induced surface modifications 197
Staffan Jacobson and Sture Hogmark

Chapter 6

Transient phenomena in elastohydrodynamic lubrication 227
Romeo P. Glovnea

Chapter 7

On the Stribeck curve 263
Michael M. Khonsari and E. Richard Booser

Chapter 8

Surface characterization, adhesion measurements and modeling of
microelectromechanical systems 279
Xiaojie Xue and Andreas A. Polycarpou



Research Signpost
37/661 (2), Fort P.O.
Trivandrum-695 023
Kerala, India

Recent Developments in Wear Prevention, Friction and Lubrication, 2010: 1-43
ISBN: 978-81-308-0377-7 Editor: George K. Nikas

1. The thin film approximation in hydrodynamic, including elastohydrodynamic, lubrication

Andras Z. Szeri

Department of Mechanical Engineering, University of Delaware, Newark, DE 19716, USA

Abstract. The equations that describe the motion of viscous fluids are difficult to solve even with present day computing facilities and we constantly search for ways to simplify them. A notable simplification that arises in lubrication is the so-called “thin film” or “lubrication” approximation. The resulting Reynolds theory of lubrication is a constant viscosity, quasi two-dimensional theory, valid when the ratio of the characteristic lengths is vanishingly small. It breaks down where there is a sudden change in film thickness, or if the Reynolds number is increased even with the flow remaining laminar. Among the additional circumstances that negate validity of the classical Reynolds theory is the viscosity being strongly dependent on the pressure or on another component of stress. The Reynolds theory also breaks down if the film becomes too thin for the continuum model to remain applicable.

Introduction

Osborne Reynolds developed the thin film approximation in his efforts to explain the experimental results of Beauchamp Tower. While studying Tower’s report on railroad bearings, Reynolds identified “crucial proof ... that the surfaces were completely and continuously separated by a film of oil; this film being maintained by the motion of the journal, although the pressure in the oil at the crown of the bearing was shown by actual measurement to be as much as 625 lbs. per sq. inch above the pressure in the oil bath” [1]. It further occurred to Reynolds as possible that “the film of oil might be sufficiently thick for the unknown boundary actions to disappear, in which case the results would be deducible from the equations of hydrodynamics”. During the course of his research, Reynolds linearized the Navier-Stokes equations for flow between slightly inclined, rigid surfaces. The resulting Reynolds theory of lubrication is a constant viscosity, quasi two-dimensional theory, valid when the ratio of the characteristic lengths is

vanishingly small. It breaks down where there is a sudden change in film thickness, or if the Reynolds number is increased even with the flow remaining laminar. Among the additional circumstances that negate validity of the classical Reynolds theory is the viscosity being strongly dependent on the pressure or on another component of stress. The Reynolds theory also breaks down if the film becomes too thin for the continuum model to remain applicable.

1. The Reynolds equation

Reynolds based his theory of lubrication on the following assumptions [1].

1. The continuum description is valid.
2. The Navier-Stokes equations hold (viscosity depends at most on temperature).
3. The lubricant is incompressible.
4. The film is thin, therefore
 - (a) lubricant flow is laminar;
 - (b) lubricant inertia is negligible;
 - (c) lubricant film curvature is negligible;
 - (d) lubricant body force is negligible.

According to Reynolds, as consequence of assumption 4(c) it is permissible to describe fluid film lubrication relative to orthogonal Cartesian coordinates. Conventionally, the y -axis of the Cartesian system is in the direction of the minimum film dimension while the ‘plane’ of the lubricant film coincides with the (x,z) plane.

Applications of the above assumptions to the Navier-Stokes equations led Reynolds to the reduced equations of motion

$$\frac{\partial p}{\partial x} = \mu \frac{\partial^2 u}{\partial y^2}, \quad \frac{\partial p}{\partial y} = 0, \quad \frac{\partial p}{\partial z} = \mu \frac{\partial^2 w}{\partial y^2} \quad (1)$$

The equation of continuity, in contrast, retained its original form

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0. \quad (2)$$

The boundary condition assigned to these equations by Reynolds were no-slip on the solid surfaces

$$\begin{aligned} u &= U_1, \quad w = 0 \quad \text{at} \quad y = 0 \\ u &= U_2, \quad w = 0 \quad \text{at} \quad y = h \end{aligned} \quad (3)$$

To the present order of approximation, the pressure does not vary across the film (1) and the equations of motion could be integrated with respect to y

$$\begin{aligned} u &= \frac{1}{2\mu} \frac{\partial p}{\partial x} (y^2 - yh) + \left(1 - \frac{y}{h}\right) U_1 + \frac{y}{h} U_2, \\ w &= \frac{1}{2\mu} \frac{\partial p}{\partial z} (y^2 - yh) \end{aligned} \quad (4)$$

The pressure is, yet, unknown. However, since it is an induced pressure (Reynolds postulated ambient pressure at film's edges) with the sole function of enforcing conservation of mass, it can be evaluated from the equation of continuity. This seems like a reasonable scheme, but has one flaw. If u and w are substituted into equation (2) the resulting single equation will contain two unknowns, v and p , and unless v is given, we have insufficient information to determine p . Reynolds overcome the problem by integrating the equation of continuity across the film; the integrated equation of continuity contained the velocity component v only in the values it assumed at the boundaries $y = 0$ and $y = h(x, t)$. As the approach velocity of the surfaces was presumed known during this analysis, integration across the film eliminated one of the two unknowns.

Substituting for u and w into the equation of continuity and integrating across the film resulted in what we now call the *Reynolds equation of lubrication*

$$\frac{\partial}{\partial x} \left(\frac{h^3}{\mu} \frac{\partial p}{\partial x} \right) + \frac{\partial}{\partial z} \left(\frac{h^3}{\mu} \frac{\partial p}{\partial z} \right) = 6(U_1 - U_2) \frac{\partial h}{\partial x} + 6h \frac{\partial (U_1 + U_2)}{\partial x} + 12(V_2 - V_1). \quad (5)$$

Here $V_1 - V_2$ is the velocity of approach of the surfaces,

It is emphasized that $V_1 = (U_1, V_1)$ and $V_2 = (U_2, V_2)$ are the velocities of "corresponding" points, each fixed to one of the bearing surfaces.¹ The velocities V_1 and V_2 result from rigid body motion that may include both rotation and translation of the bearing surfaces. To take cognizance of this we recast the equation in a form that contains relative, rather than absolute, velocities (see, for e.g. [2])

¹ We call two points, one fixed to the bearing surface and the other to the runner surface, corresponding points at the instant when they are located on the same normal to the reference surface. For journal bearings, the pad surface is the reference surface. For the plane slider, on the other hand, it is expedient to designate the runner surface as the reference surface [2].

$$\frac{\partial}{\partial x} \left(\frac{h^3}{\mu} \frac{\partial p}{\partial x} \right) + \frac{\partial}{\partial z} \left(\frac{h^3}{\mu} \frac{\partial p}{\partial z} \right) = 6U_0 \frac{\partial h}{\partial x} + 6h \frac{\partial U_0}{\partial x} + 12V_0. \quad (6)$$

The interpretation put on the velocity U_0 is distinctly different in the two cases. For thrust bearings $U_0 = U_1 - U_2$, in contrast $U_0 = U_1 + U_2$ for journal bearings. In both cases, however, $V_0 = V_1 - V_2$.

Table 1 demonstrates convergence of the Sommerfeld number $S = \mu N(R/C)^2/P$ with the clearance ratio (C/R) for a journal bearing under Gumbel's boundary conditions [3].

In this Chapter, we will examine some of the consequences and restrictions that follow from the assumptions of Reynolds. In Section 2, we will examine the restrictions brought about by the thin-film assumption. In Section 3, we discuss behavior of piezoviscous fluids and Section 4 will highlight the lower limit in film thickness for molecularly thin films.

Table 1. Convergence of Sommerfeld number with the clearance ratio.

<i>Model</i>	<i>C/R</i>	<i>S</i>
Navier-Stokes	0.002	0.33692
	0.001	0.33698
	0.0005	0.33701
Reynolds lubrication	0.0	0.33704

2. The thin-film assumption for Navier-Stokes fluids

In this section, we investigate the thin film assumption for a Navier-Stokes fluid. The equations of motion and continuity governing the flow of constant property Newtonian fluid are

$$\rho \frac{dv_i}{dt} = \frac{\partial}{\partial x_j} \left(-p \delta_{ij} + 2\mu D_{ij} \right) \quad (7)$$

$$\frac{\partial v_i}{\partial x_i} = 0$$

We normalize these equations with characteristic length scales L_{xz} , and L_y and characteristic velocity scales U_* and $V_* = (L_y/L_{xz})U_*$, along and across the film, respectively

$$\begin{aligned} (x_1, x_2, x_3) &= L_{xz} (X, \varepsilon Y, Z) \\ (v_1, v_2, v_3) &= U_* (U, \varepsilon V, W) \end{aligned} \quad (8)$$

Here $\varepsilon = L_y/L_{xz}$ is the ratio of characteristic lengths, in conventional lubricant films $\varepsilon = O(10^{-3})$. We normalize time and pressure according to

$$t = \left(\frac{L_{xz}}{U_*}\right) \tau, \quad p = \frac{\rho U_*^2}{Re^*} P \quad (9)$$

where

$$Re = \frac{U_* L_y}{\nu}, \quad \text{and} \quad Re^* = \varepsilon Re,$$

are the Reynolds number and the reduced Reynolds number, respectively. This choice for normalization leaves the continuity equation formally invariant and retains the pressure term in the limit $Re^* \rightarrow 0$.

2.1. Laminar flow

Recasting the Navier-Stokes equations in terms of normalized variables, one obtains [4]

$$-\varepsilon^2 \left(\frac{\partial^2 U}{\partial X^2} + \frac{\partial^2 U}{\partial Z^2} \right) + Re^* \frac{dU}{d\tau} = -\frac{\partial P}{\partial X} + \frac{\partial^2 U}{\partial Y^2} \quad (10a)$$

$$\varepsilon^2 \left\{ -\varepsilon^2 \left(\frac{\partial^2 V}{\partial X^2} + \frac{\partial^2 V}{\partial Z^2} \right) + Re^* \frac{dV}{d\tau} - \frac{\partial^2 V}{\partial Y^2} \right\} = -\frac{\partial P}{\partial Y} \quad (10b)$$

$$-\varepsilon^2 \left(\frac{\partial^2 W}{\partial X^2} + \frac{\partial^2 W}{\partial Z^2} \right) + Re^* \frac{dW}{d\tau} = -\frac{\partial P}{\partial Z} + \frac{\partial^2 W}{\partial Y^2} \quad (10c)$$

The system represented by equations (7₂) and (10) contains two parameters, the aspect ratio ε and the reduced Reynolds number Re^* . We will now investigate the significance of two asymptotic cases provided by limiting values of these parameters.

Case (A): $Re^* > \varepsilon^2 \rightarrow 0$

Neglecting terms multiplied by ε^2 equation (10) yields

$$Re^* \frac{dU}{d\tau} = -\frac{\partial P}{\partial X} + \frac{\partial^2 U}{\partial Y^2} \quad (11a)$$

$$P = P(X, Z, \tau) \quad (11b)$$

$$Re^* \frac{dW}{d\tau} = -\frac{\partial P}{\partial Z} + \frac{\partial^2 W}{\partial Y^2} \quad (11c)$$

The second of these equations states that the flow is a quasi two-dimensional Navier-Stokes flow: the pressure is invariant along L_y . Clearly, it is not possible to characterize this flow by a single equation in pressure. It is only the second approximation $Re^* \rightarrow 0$, that makes possible the derivation of a single pressure equation.

Case (B): $\varepsilon^2 > Re^* \rightarrow 0$

Neglecting terms multiplied by Re^* in equation (10) leads to a three-dimensional Stokes flow

$$\nabla^2 \mathbf{V} = -\nabla P \quad (12)$$

It is, again, not possible to characterize this flow by a single equation in pressure and the full three-dimensional problem must be solved [5, 6]. However, in lubrication this limit is of interest only in the most unusual of circumstances; the Reynolds number must approach zero faster than ε locally. For machined surfaces L_y might be equated to the asperity height, δ , and the lateral length scale L_{xz} to the distance between asperities, l ; for ground surfaces the local value of the aspect

ratio is then $\sim 1.25\mu/12.5\mu$, or, $\varepsilon^2 \sim 0.01 > Re^*$, forcing the Reynolds number to be so small as to be outside interest in most lubrication application.

2.2. Turbulent flow

For sufficiently large Reynolds number assumption (4a) no longer holds as the flow might become turbulent even in thin lubricant films. In this case the equations of motion for the mean flow are obtained by substituting the assumptions

$$p = \bar{P} + p' \quad v_i = \bar{V}_i + v'_i$$

into equation (7₁) and averaging.

$$\rho \frac{d\bar{V}_i}{dt} = \frac{\partial}{\partial x_j} \left(-\bar{P} \delta_{ij} + 2\mu \bar{D}_{ij} - \overline{\rho v'_i v'_j} \right) \quad (13)$$

The over-score bar signifies the average value of the quantity and the prime its instantaneous departure from the average; the material derivative d/dt and the stretching tensor \bar{D}_{ij} both refer to mean quantities

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \bar{V}_k \frac{\partial}{\partial x_k}, \quad \bar{D}_{ij} = \frac{1}{2} (\bar{V}_{ij} + \bar{V}_{ji})$$

The coordinates are normalized as for laminar flow (8₁) while the velocities and the pressure according to

$$\left\{ \begin{array}{l} (\bar{V}_1, \bar{V}_2, \bar{V}_3) = \frac{1}{U_*} (U, \varepsilon V, W), \quad \bar{P} = \frac{\rho U_*^2}{Re^*} P \\ (v'_1, v'_2, v'_3) = u_*(u, v, w), \quad \kappa = (u_*/U_*)^2 \end{array} \right. \quad (14)$$

Here we assume that the fluctuating components of the velocity are all of the same order of magnitude, $v'_1 \sim v'_2 \sim v'_3$ [7]. We have not yet specified the value of κ , the square of the ratio of characteristic velocities of the fluctuations and of the mean flow. In fact, we shall investigate the consequences of selecting for κ one value or another.

Substitution into equation (13) leads to the normalized form of the equations for turbulent flow

$$Re^* \left[\frac{dU}{d\tau} + \kappa \left(\frac{\partial \bar{u}u}{\partial X} + \frac{\partial \bar{u}w}{\partial Z} \right) \right] + \kappa Re \frac{\partial \bar{u}v}{\partial Y} = -\frac{\partial P}{\partial X} + \frac{\partial^2 U}{\partial Y^2} + \varepsilon^2 \left(\frac{\partial^2 U}{\partial X^2} + \frac{\partial^2 U}{\partial Z^2} \right) \quad (15a)$$

$$\varepsilon^2 \left[Re^* \frac{dV}{d\tau} - \frac{\partial^2 V}{\partial Y^2} - \varepsilon^2 \left(\frac{\partial^2 V}{\partial X^2} + \frac{\partial^2 V}{\partial Z^2} \right) \right] + \kappa \varepsilon Re^* \left(\frac{\partial \bar{v}u}{\partial X} + \frac{\partial \bar{v}w}{\partial Z} \right) = -\frac{\partial P}{\partial Y} - \kappa Re^* \left(\frac{\partial \bar{v}v}{\partial Y} \right) \quad (15b)$$

$$Re^* \left[\frac{dW}{d\tau} + \kappa \left(\frac{\partial \bar{u}w}{\partial X} + \frac{\partial \bar{w}w}{\partial Z} \right) \right] + \kappa Re \frac{\partial \bar{v}w}{\partial Y} = -\frac{\partial P}{\partial Z} + \frac{\partial^2 W}{\partial Y^2} + \varepsilon^2 \left(\frac{\partial^2 W}{\partial X^2} + \frac{\partial^2 W}{\partial Z^2} \right) \quad (15c)$$

These equations contain three independent parameters ε , Re^* or Re , and κ . Based on the previous discussion, we focus attention on the limit $Re^* > \varepsilon^2 \rightarrow 0$ while investigating the role of κ . Setting the condition $\varepsilon^2 = 0$ we obtain

$$Re^* \left[\frac{dU}{d\tau} + \kappa \left(\frac{\partial \bar{u}u}{\partial X} + \frac{\partial \bar{u}w}{\partial Z} \right) \right] + \kappa Re \frac{\partial \bar{u}v}{\partial Y} = -\frac{\partial P}{\partial X} + \frac{\partial^2 U}{\partial Y^2} \quad (16a)$$

$$\kappa Re^* \left[\varepsilon \left(\frac{\partial \bar{v}u}{\partial X} + \frac{\partial \bar{v}w}{\partial Z} \right) + \frac{\partial \bar{v}v}{\partial Y} \right] = -\frac{\partial P}{\partial Y} \quad (16b)$$

$$Re^* \left[\frac{dW}{d\tau} + \kappa \left(\frac{\partial \bar{u}w}{\partial X} + \frac{\partial \bar{w}w}{\partial Z} \right) \right] + \kappa Re \frac{\partial \bar{v}w}{\partial Y} = -\frac{\partial P}{\partial Z} + \frac{\partial^2 W}{\partial Y^2} \quad (16c)$$

Most current turbulent lubrication theories that are based on a single equation for pressure [8-10], take their departure from equation (16) on the assumption that

$$\kappa = (u_*/U_*)^2 = O(1) \quad (17)$$

Hinze [7] shows, however, that this assumption is valid only in free turbulent flows such as jets and wakes, i.e., in flows that differ qualitatively from flows in confined, narrow spaces. Nevertheless, when inserting this value of κ into equation (16) one obtains

The thin film approximation in hydrodynamic, including elasto-hydrodynamic, lubrication

$$Re^* \left[\frac{dU}{d\tau} + \frac{\partial \bar{u}}{\partial X} + \frac{\partial \bar{u}}{\partial Z} \right] + Re \frac{\partial \bar{u}}{\partial Y} = -\frac{\partial P}{\partial X} + \frac{\partial^2 U}{\partial Y^2} \quad (18a)$$

$$Re^* \frac{\partial \bar{v}}{\partial Y} = -\frac{\partial P}{\partial Y} \quad (18b)$$

$$Re^* \left[\frac{dW}{d\tau} + \frac{\partial \bar{u}}{\partial X} + \frac{\partial \bar{w}}{\partial Z} \right] + Re \frac{\partial \bar{w}}{\partial Y} = -\frac{\partial P}{\partial Z} + \frac{\partial^2 W}{\partial Y^2} \quad (18c)$$

Only under the highly non-physical assumption

$$Re^* \rightarrow 0, Re = O(1)$$

is it possible to combine equation (18) with the equation of continuity to yield a single equation in pressure

$$\frac{\partial}{\partial x} \left(\frac{h^3}{\mu k_x(Re_h)} \frac{\partial \bar{P}}{\partial x} \right) + \frac{\partial}{\partial z} \left(\frac{h^3}{\mu k_z(Re_h)} \frac{\partial \bar{P}}{\partial z} \right) = \frac{U_0}{2} \frac{\partial h}{\partial x} \quad (19)$$

Here $k_x = k_x(Re_h)$, $k_z = k_z(Re_h)$, and $Re_h(x) = [h(x)/L_y] Re$ stands for the local Reynolds number [2].

Constantinescu [11] attempted to demonstrate that the conditions leading up to equation (19) hold in typical lubrication situations. He did not address, however, the validity of the 'free turbulence' assumption (17).

Instead of using equation (17), one might recognize that due to the thinness of the lubricant film the bounding walls have controlling effect on the developing super laminar flow and, following Hinze [7], employ

$$\kappa = (u_*/U_*)^2 = O(\varepsilon) \quad (20)$$

Applying equation (20), equation (16) yields

$$Re^* \left(\frac{dU}{d\tau} + \frac{\partial \bar{u}}{\partial Y} \right) = -\frac{\partial P}{\partial X} + \frac{\partial^2 U}{\partial Y^2} \quad (21a)$$

$$P = P(X, Z, \tau) \quad (21b)$$

$$Re^* \left(\frac{dW}{d\tau} + \frac{\partial \overline{vw}}{\partial Y} \right) = -\frac{\partial P}{\partial Z} + \frac{\partial^2 W}{\partial Y^2} \quad (21c)$$

Although a Reynolds type pressure equation is not available for turbulent flow under assumption (20) either; however, in first approximation the pressure no longer varies across the film. Furthermore, for $Re^* \geq 1$ inertia terms are of the same order of magnitude as the turbulence terms: if there is no inertia, there can be no turbulence.

We may thus conclude that when classical lubrication assumptions no longer yield acceptable results in 'thin' flows, extension to quasi two-dimensional Navier-Stokes, rather than three-dimensional Stokes, equations should be made. The equations governing this flow are

$$\begin{aligned} Re^* \left(\frac{dU}{d\tau} + \alpha \frac{\partial \overline{uv}}{\partial Y} \right) &= -\frac{\partial P(X, Z, \tau)}{\partial X} + \frac{\partial^2 U}{\partial Y^2} \\ Re^* \left(\frac{dW}{d\tau} + \alpha \frac{\partial \overline{vw}}{\partial Y} \right) &= -\frac{\partial P(X, Z, \tau)}{\partial Z} + \frac{\partial^2 W}{\partial Y^2} \\ \frac{\partial U}{\partial X} + \frac{\partial V}{\partial Y} + \frac{\partial W}{\partial Z} &= 0 \end{aligned} \quad (22)$$

where $\alpha = 0$ for laminar flow and $\alpha = 1$ for turbulent flow.

2.3. The plane slider

We aim now to provide some evidence for the validity of the system in (22), at least in laminar flow. For infinite extent of the flow domain in the z-direction the steady laminar flow equation becomes quasi one-dimensional

$$\begin{aligned} Re^* \left(U \frac{\partial U}{\partial X} + V \frac{\partial U}{\partial Y} \right) &= -\frac{\partial P}{\partial X} + \frac{\partial^2 U}{\partial Y^2} \\ \frac{\partial U}{\partial X} + \frac{\partial V}{\partial Y} &= 0 \end{aligned} \quad (23)$$

The results we shall quote here relate to flow between inclined, nominally flat planes, i.e., the ‘plane slider’. We put $h_1 = h(x_1)$ and $h_2 = h(x_2)$ for the film thickness at outlet and inlet, respectively, and specify the length scales by $L_y = (h_1 + h_2)/2$ and $L_{xz} = B = x_2 - x_1$, where x_1 and x_2 defines the position of the outlet and the inlet, respectively.

Introduction of the stream function $\Psi(X, Y)$ into (23) leads to [4]

$$Re^* \left(H \frac{\partial \Psi}{\partial \eta} \frac{\partial^3 \Psi}{\partial \eta^2 \partial \xi} - 2 \frac{dH}{dX} \frac{\partial \Psi}{\partial \eta} \frac{\partial^2 \Psi}{\partial \eta^2} - H \frac{\partial \Psi}{\partial \xi} \frac{\partial^3 \Psi}{\partial \eta^3} \right) - \frac{\partial^4 \Psi}{\partial \eta^4} = 0, \quad 0 \leq \xi, \eta \leq 1 \quad (24)$$

where we also introduced a change of variables

$$\begin{aligned} \xi &= X - X_1 \\ \eta &= Y / H(X), \quad H = h(x) / L_y \end{aligned}$$

The no-slip boundary conditions on the solid boundaries are

$$\begin{aligned} \Psi &= 0, \quad \frac{\partial \Psi}{\partial \eta} = -H, \quad \text{at } \eta = 0 \\ \Psi &= Q^*, \quad \frac{\partial \Psi}{\partial \eta} = 0, \quad \text{at } \eta = 1 \end{aligned} \quad (25)$$

where Q^* is the dimensionless flow rate, yet unknown. To insure that the problem remains mathematically well posed we increase the number of independent equations by constraining the average pressure at outlet to equal its value at inlet:

$$\int_0^1 \int_0^1 H(X) \frac{\partial P}{\partial \xi} d\xi d\eta = 0 \quad (26)$$

We approximate $\Psi(\xi, \eta)$ by piecewise polynomial functions [12] and apply Galerkin's method to evaluate the coefficients in the approximation. The resulting system of nonlinear algebraic equations can be written in the form

$$G(\omega) = 0, \quad \omega = (u, \sigma), \quad (27)$$

where u is the vector of state variables and σ is the vector of parameters. The computational scheme for solving equation (27), i.e., parametric continuation followed by the Gauss – Newton method, can be found in [4].

The principal conclusion from equation (22) is the invariance of the pressure across the film. To investigate the upper bound of ε for this conclusion to hold, we look at flow between inclined planes of various aspect ratios. As long as equations (22) hold, the pressure on the upper plate, $P(h)$, and the pressure on the lower plate, $P(0)$, are approximately equal, becoming identical at the limit $\varepsilon \rightarrow 0$. This may be investigated quantitatively by computing a pressure difference coefficient, d_p

$$d_p = |P(h) - P(0)|_{max} / P(h)_{max} \quad (28)$$

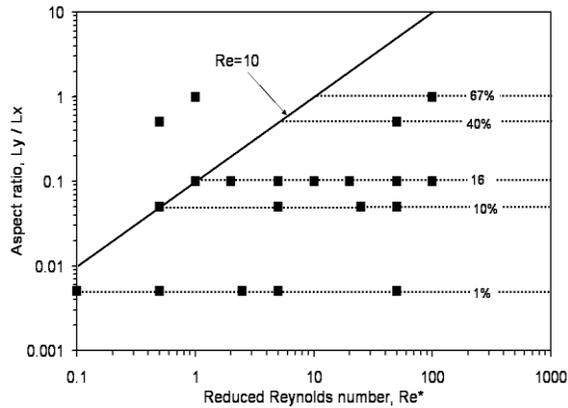


Figure 1. Pressure difference coefficient d_p for various values of the aspect ratio and reduced Reynolds number.

In figure 1, we indicate the value of d_p , calculated from the three-dimensional Navier-Stokes problem employing FIDAP, as a function of the parameters ε and Re^* . We restrict attention here arbitrarily to $Re > 10$, accepting this as a lower bound on the Reynolds number for applications. For 'small' values of the aspect ratio, figure 1 appears to support the assertion of equation (22): for $\varepsilon \leq 0.05$, $d_p \leq 0.01$, and even for the wider range $\varepsilon \leq 0.1$, $d_p < 0.16$, though the increase in d_p for $\varepsilon > 0.1$ is quite rapid. Thus, for $\varepsilon \leq 0.1$, we have the approximate relationship $d_p \approx (Re^*)$. This conclusion seems to hold well for $Re^* \leq 100$.

Figure 2 plots the ratio of actual pressure over its zero Reynolds number value against Re^* , as calculated by FIDAP from the three-dimensional Navier-Stokes problem at various values of $\varepsilon \leq 0.1$. Data for different ε values collapse onto a

single curve, confirming, again, that under the stated conditions the aspect ratio is not a strong parameter of the flow, that is $P_{max}/P_{0,max} \approx \Phi(Re^*)$.

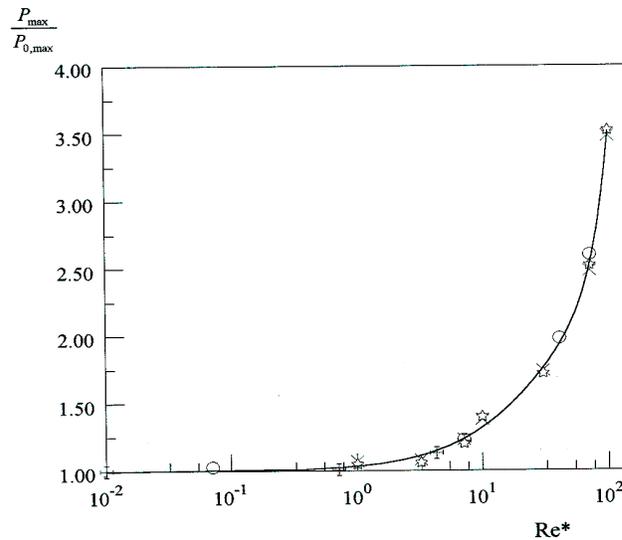


Figure 2. Variation of $P_{max}/P_{0,max}$ with Re^* , FIDAP (+, $\varepsilon = 0.005$; o, $\varepsilon = 0.05$; *, $\varepsilon = 0.08$; x, $\varepsilon = 0.1$).

Figure 3 compares lubricant force from two sources, FIDAP solution of the full Navier-Stokes problem and the stream function-Galerkin formulation of equations (24). In this plot the force is normalized with its zero Reynolds number value, and two channel geometries, $h_2/h_1 = 2$ and $h_2/h_1 = 3/2$ are depicted.

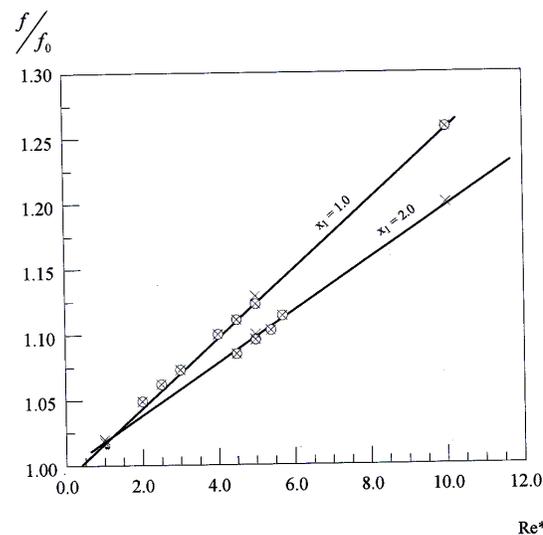


Figure 3. Normalized force, $\varepsilon \leq 0.1$ (o, approximation; x, Navier-Stokes).

The approximation (22) will now be used to investigate the effect of waviness of the runner. To this end, we perturb the film thickness to $H(\xi) = X_1 + \delta \cos(n\pi\xi)$. Figure 4 displays results for $n=5$, $\delta = \pm 0.1$ and $Re^* = 4.0$. When $\delta > 0$, the film shape is convergent in the direction of the flow at inlet and divergent at outlet, we characterize this as a c/d film shape. When $\delta < 0$, the film shape is d/c. It may be seen from figure 4 that for film shape c/d the effect of convective fluid inertia is to lower the pressure within the channel, while for film shape d/c the pressure is raised relative to inertialess flow. By changing to $n = 5.5$, figure 5, $\delta > 0$ yields a d/d film shape, resulting in the channel walls being pulled together. $\delta < 0$, on the other hand, establishes a c/c film and the channel walls are forced apart.

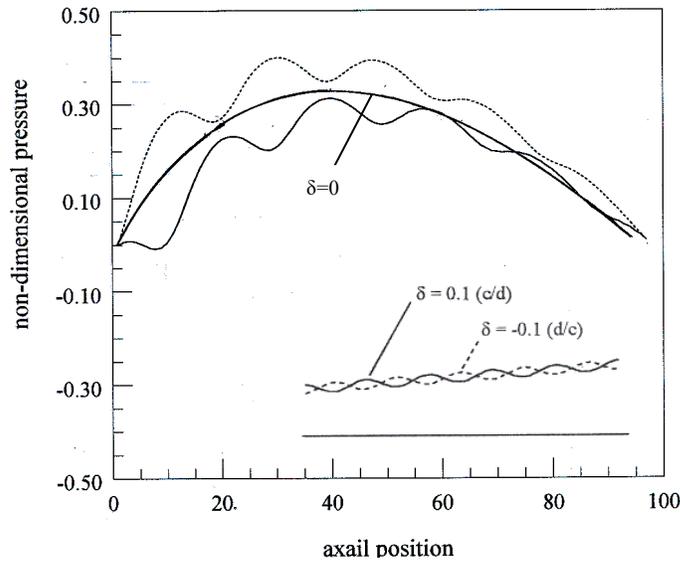


Figure 4. Perturbed slider ($n=5$, $X_1 = 2.0$, $\delta = \pm 0.1$, $Re^* = 4.0$) (— convergent at inlet, divergent at outlet; ---- divergent at inlet, convergent at outlet).

We suggest here that the system of equations (22) may constitute a logical first extension of the Reynolds thin film model. Provided that $\varepsilon \leq 0.1$, the normalized maximum pressure varies with the Reynolds number but is independent of the aspect ratio. Furthermore, the excess force due to convective inertia is additive for sinusoidal films convergent in the flow direction at exit [4]

$$\text{sgn}(f - f_0) = -\text{sgn} \left. \frac{dh}{dx} \right|_{x_{out}} \quad (29)$$

Here x is now increasing in the flow direction.

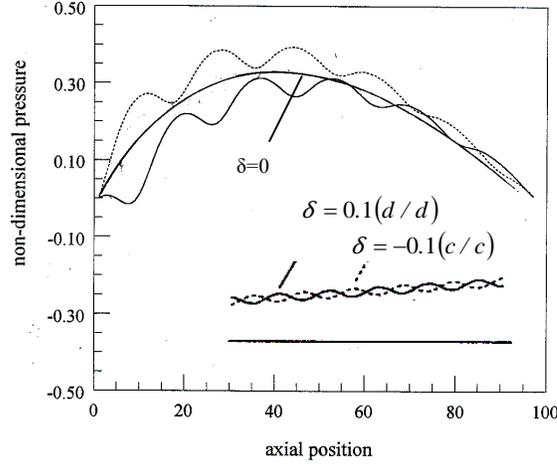


Figure 5. Perturbed slider ($n=5.5$ $X_1=2.0$, $\delta = \pm 0.1$, $Re^*=4.0$) (— convergent at inlet, divergent at outlet; ---- divergent at inlet, convergent at outlet).

2.4. Film curvature

To study the effect of curvature of the film (journal bearings) we employ a bipolar coordinate system $\{\hat{\alpha}, \hat{\beta}\}$ that is related to our Cartesian coordinate system $\{X, Y\}$ through

$$\hat{\alpha} + i\hat{\beta} = -2 \coth^{-1} \left(\frac{X + iY}{a} \right) \quad (30)$$

where a is the separation between the pole and the origin of the $\{X, Y\}$ system. In the bipolar coordinate system the cylinders of radii r_1 and r_2 , $r_1 < r_2$, have the simple representation $\hat{\alpha} = \hat{\alpha}_1$ and $\hat{\alpha} = \hat{\alpha}_2$, $\hat{\alpha}_1 < \hat{\alpha}_2 < 0$. The scale factor of the bipolar coordinate system [13] is $H = a / (\cosh \hat{\alpha} - \cos \hat{\beta})$.

The equations of motion and continuity defining the two-dimensional flow field are first written relative to the bipolar coordinate system [13] and then nondimensionalized [2]. In terms of bipolar coordinates the eccentricity ratio is given by

$$\varepsilon = \frac{\sinh(\hat{\alpha}_1 - \hat{\alpha}_2)}{\sinh \hat{\alpha}_1 - \sinh \hat{\alpha}_2} \quad (31)$$

The objective here is to find approximate solutions for thin films without neglecting film curvature and compare with Reynolds equation results. To discover

the correct approximation to use, we adopt the length scales $L_\beta = r_1$ along the principal dimension of the film and $L_\alpha = r_1/\Delta$, $\Delta = 1/(\hat{\alpha}_2 - \hat{\alpha}_1)$, across it. The characteristic velocities are $U_* = r_1\omega$ and $V_* = U_*/2\pi\Delta$. The dynamic condition in the flow are represented by the Reynolds number

$$r_e = \frac{U_* L_\alpha}{\nu}$$

Neglecting terms of the order $(\hat{\alpha}_2 - \hat{\alpha}_1)^2$ or smaller, the normalized equations of motion thus take the form

$$\begin{aligned} \frac{\partial P}{\partial \alpha} &= 0 \\ -\frac{1}{(2\pi)^2 H} \frac{\partial P}{\partial \beta} &= \frac{1}{H^2 \sinh \hat{\alpha}_1} \frac{\partial^2 v}{\partial \alpha^2} + r_e^* \left[\frac{1}{H} \left(u \frac{\partial v}{\partial \alpha} + v \frac{\partial v}{\partial \beta} \right) \right] \end{aligned} \quad (32)$$

Here u, v , are normalized velocities, $\alpha = (\hat{\alpha} - \hat{\alpha}_1)\Delta$ and $\beta = \hat{\beta}/2\pi$, and $r_e^* = r_e/2\pi\Delta$ is the reduced Reynolds number of the problem analogous to Re^* . The condition $Re^* \rightarrow 0$ of classical lubrication theory is equivalent, thus, to $r_e^* \rightarrow 0$ and upon applying this limit the equations reduces to

$$-\sinh \hat{\alpha}_1 \frac{H}{(2\pi)^2} \frac{\partial P}{\partial \beta} = \frac{\partial^2 v}{\partial \alpha^2}, \quad \frac{\partial P}{\partial \alpha} = 0 \quad (33)$$

To solve equation (33) we integrate twice with respect to α . Substitution into the integrated (across the film) continuity equation yields [3]

$$\begin{aligned} \frac{\partial}{\partial \beta} \left\{ \frac{\partial P}{\partial \beta} \int_0^1 H(\alpha, \beta) \left[\int_0^\alpha I(\sigma, \beta) d\sigma - \alpha \int_0^1 I(\sigma, \beta) d\sigma - C(1-\alpha) \right] d\alpha \right\} &= 0 \\ I(\alpha, \beta) &= \int_0^\alpha H(\varphi, \beta) d\varphi \end{aligned} \quad (34)$$

The innermost integral of equation (34) was obtained analytically while the other integrals necessary to solve for P were performed via Gaussian quadrature [3]. The error committed by neglecting curvature of the film maybe gauged from table 2, which displays values of the nondimensional group $P/\mu N$ under various conditions.

Table 2. Effect of film curvature on $(P/\mu N) \times 10^{-5}$.

Boundary condition	(C/R)	Navier-Stokes	Reynolds lub.	Bipolar lub.
Sommerfeld	0.002	7.4202	7.4175	7.4052
	0.001	29.6754	29.6701	29.6454
	0.0005	118.6908	118.6802	118.6732
Gümbel	0.002	14.8104	14.8052	14.7824
	0.001	59.2311	59.2207	59.1486
	0.0005	236.8966	236.8826	236.8125

The exact, zero Reynolds number solution in [14] valid for arbitrary clearance ratio, can be employed to study film curvature effects. When this solution is expanded in powers of the clearance ratio, the first two terms correspond to the Myllerup and Hamrock [15] solution, which employs regular perturbation:

$$\begin{aligned}
 P = & \frac{12\pi\varepsilon \sin \theta (2 + \varepsilon \cos \theta)}{(2 + \varepsilon^2)(1 + \varepsilon \cos \theta)^2} \\
 & + \left(\frac{C}{R}\right) \frac{4\pi\varepsilon \sin \theta (1 + 5\varepsilon^2 + 2\varepsilon(2 + \varepsilon^2)\cos \theta)}{(2 + \varepsilon^2)(1 + \varepsilon \cos \theta)^3} + O\left(\frac{C}{R}\right)^2
 \end{aligned} \tag{35}$$

The first term is identical to the solution of the Reynolds equation under full film boundary conditions, while the second term is the first order curvature correction [16].

3. Departure from Newtonian fluid behavior

The lubrication approximation has been widely used far outside the confines of lubrication, becoming one of the cornerstones of fluid mechanics [17-20]. It is important, therefore, to be cognizant of the assumptions under which it is derived, so that it is not used outside its range of applicability.

3.1. The Navier-Stokes model

The compressible Navier-Stokes equation assumes that the Cauchy stress \mathbf{T} depends only on the density and the velocity gradient.

$$\mathbf{T} = \mathbf{T}(\rho, \mathbf{L}), \quad \mathbf{L} = \text{grad} \mathbf{v} \quad (36)$$

The requirement of frame-indifference then implies that the velocity gradient occur only through its symmetric part. Restrictions due to isotropy and the assumption that the stress be linear in the symmetric part \mathbf{D} of the velocity gradient lead to the compressible Navier-Stokes model

$$\mathbf{T} = \alpha_0 (\rho, I_D, II_D, III_D) \mathbf{1} + \alpha_1 (\rho, I_D, II_D, III_D) \mathbf{D} + \alpha_2 (\rho, I_D, II_D, III_D) \mathbf{D}^2 \quad (37)$$

where

$$\mathbf{D} = \frac{1}{2} (\mathbf{L} + \mathbf{L}^T),$$

$$I_D = \text{tr} \mathbf{D}, \quad II_D = \frac{1}{2} [(\text{tr} \mathbf{D})^2 - \text{tr} \mathbf{D}^2], \quad III_D = \det \mathbf{D}.$$

If the fluid is assumed to be incompressible, then it will follow (using the assumption that the constraint response does no work) that

$$\mathbf{T} = -p \mathbf{1} + \hat{\alpha}_1 (II_D, III_D) \mathbf{D} + \hat{\alpha}_2 (II_D, III_D) \mathbf{D}^2 \quad (38)$$

where p is the Lagrange multiplier due to the incompressibility constraint.

Requiring that the stress depend linearly on \mathbf{D} leads to

$$\mathbf{T} = -p \mathbf{1} + 2\hat{\mu} \mathbf{D}, \quad (39)$$

where $\hat{\mu}$ is a constant.

During his derivation of the model, Stokes [21] already recognized that the viscosity of a fluid could depend on the pressure, and he was concerned with the question as to when it would be reasonable to assume that the viscosity is a constant. “Let us now consider in what cases it is allowable to suppose μ to be independent of pressure. Du Buat has concluded it from his experiments on the motion of water in pipes and canals, that the total retardation of the velocity due to friction is not increased by increasing the pressure... I shall therefore suppose that for water, and by analogy for other incompressible fluids, μ is independent of the pressure” [21].

That the viscosity for liquids could depend on the pressure and could change significantly with sufficiently large variations of the pressure has been well

recognized [22]. Experiments that are more recent also confirm the fact that the viscosity of certain lubricants can change dramatically with the pressure (see [23-27]). Incorporating this notion into the development of constitutive theories leads to an interesting departure from classical theories.

Assuming that the constraint forces, in our case the pressure, can influence the work done, we have the possibility that the stress \mathbf{T} is of the form

$$\mathbf{T} = -p\mathbf{1} + \tilde{\alpha}_1(p, II_D, III_D)\mathbf{D} + \tilde{\alpha}_2(p, II_D, III_D)\mathbf{D}^2. \quad (40)$$

Further assuming linearity in \mathbf{D} , we arrive at

$$\mathbf{T} = -p\mathbf{1} + 2\mu(p)\mathbf{D}. \quad (41)$$

There is a very fundamental difference between the two models (39) and (41). While the model (39) provides an explicit relation between \mathbf{T} and \mathbf{D} , model (41) is implicit with the more general the form

$$\mathbf{f}(\mathbf{T}, \mathbf{D}) = \mathbf{0}. \quad (42)$$

A generalization of the model (42) in which the viscosity depends on both the pressure and the symmetric part of the velocity gradient permits one to describe shear thinning and shear-thickening, observed in some fluids. In such fluids the stress is of the form

$$\mathbf{T} = -p\mathbf{1} + \beta(p, \mathbf{D})\mathbf{D}. \quad (43)$$

While many models for viscoelastic fluids are implicit models, the popular Maxwell model being one such, they are not of the type (42); they usually involve higher derivatives of the stress and the symmetric part of the velocity gradient. Rigorous global existence of solutions for a sub-class of fluids of the form (43) in which the viscosity satisfies certain conditions, met by the models used in elastohydrodynamics, can be found in Malek *et al.* [28, 29]. Of course, for fluids of the implicit type the starting point of analysis cannot be the Navier-Stokes equations but Cauchy's equation of balance of linear momentum

$$\rho \frac{dv}{dt} = \operatorname{div} \mathbf{T} + \mathbf{b} \quad (44)$$

where \mathbf{b} is the body force.

Models of the type (41) have a much richer class of solutions than the Navier-Stokes model (39). Even in the case of flow between infinite parallel plates, non-unique solutions arise that have no counterparts in the Navier-Stokes theory. In addition, the structure of the solutions to (41) can qualitatively differ from those for the Navier-Stokes fluid (see [30]). In the case of simple Poiseuille flow, the solution may vary from that of plug flow to a V shaped profile, while it is parabolic for the Navier-Stokes fluid.

As the model (41) leads to solutions that qualitatively differ from the solutions to the Navier-Stokes equations, it is also possible that approximations derived from it could lead to equations that predict response that is not only quantitatively but also qualitatively different from the classical approximation. The question here is the magnitude of the error made by not acknowledging the dependence of the viscosity on the pressure consistently while deriving the equations governing the problem of elastohydrodynamics.

3.2. The piezoviscous fluid

Ever since Stokes assumed that the viscosity is a constant, it has been treated as such in most, though not all, subsequent studies. Of course, there are many fluids that shear-thin or shear-thicken; for such fluids, the viscosity is considered a function of the symmetric part of the velocity gradient. However, in one area of research that presupposes the fluid to be a Navier-Stokes fluid, the viscosity is not treated as a constant but is allowed to depend on the pressure, namely in elastohydrodynamic lubrication. Here, we come across a rather intriguing inconsistency. The Reynolds approximation is based on the assumption of constant viscosity. However, in developing the elastohydrodynamic approximation, pressure dependence of the viscosity is acknowledged only after the equation has been obtained under the assumption of constant viscosity [31]. It is astonishing that this obvious inconsistency in the derivation of the equation for elastohydrodynamic lubrication has gone unnoticed until quite recently.

Current research on piezoviscous fluids has raised questions concerning (1) the appropriateness of the Reynolds equation in elastohydrodynamic lubrication (EHL), because of possible change of type of the equations of motion at high pressures, and (2) the errors inherent in the lubrication approximation due to potential existence of cross-film pressure gradient.

There have been some rigorous studies concerning the existence of solutions to the equations governing the flows of fluids with pressure dependent viscosity. Renardy [32] recognized that the equations could change type if the class of viscosity functions that he picked did not satisfy a certain condition. However, his choice of the viscosity functions is unrealistic, as he demands that it satisfy

$\mu(p)/p \rightarrow 0$, as $p \rightarrow \infty$. Numerous experimental results clearly contradict this assumption [33-36]. Gazzola [37] and Gazzola and Secchi [38] have proved local-in-time existence of solutions to the flow of fluids with pressure dependent viscosities. More recently, Malek *et al.* [28, 29] have established existence of solutions that are global-in-time when the viscosity depends on the pressure and the shear rate. However, these results were proven only for spatially periodic flows. Existence of solutions that are global-in-time for the standard Dirichlet problem when $\mu(p)/p \rightarrow \infty$, as $p \rightarrow \infty$ is an open problem.

Bair *et al.* [39] picked up on the criterion that Renardy [32] required for the equations to remain elliptic. Adopting the Barus equation, $\mu = \mu_0 \exp(\alpha p)$, $\mu' = \alpha \mu$, to represent the pressure dependence of viscosity, they re-cast Renardy's criterion. According to Bair *et al.*, in two-dimensional flow change of type from elliptic to hyperbolic occurs when $\tau_I = \alpha^{-1}$. Here $\tau_I = 2\mu d_I$ is the principal shear stress. For mineral oils $\alpha^{-1} \approx 50$ MPa, and the criterion sets a limiting value for the principal shear stress. The practical value of this finding to EHL remains questionable, however, as the Barus formula is unrealistic for glass forming liquids. Moreover, in view of Renardy's analysis requiring unrealistic physical conditions, the modifications proposed are not relevant to the flows of lubricants.

Bair *et al.* [39] appear to be the first to argue, "the Reynolds equation adequately captures the mechanics of the piezoviscous liquid only when the shear stress is much less than the reciprocal of the pressure viscosity coefficient." Schafer, *et al.* [40] continued along this line of investigation and, starting from the Navier-Stokes equations with pressure dependent viscosity, derived a corrected Reynolds equation. They concluded, "application of Reynolds equation is permissible for the case of pure rolling in the contact, but not when considering partial or pure sliding." Greenwood [41], in a discussion to Schafer's paper, offered a simpler derivation of the same equation. Referring to Schafer's paper, Greenwood remarks "the author's astonishing claim that the whole EHL theory is based on an incorrect equation, seems to this discussor to be entirely correct." Full Navier-Stokes solutions of the EHL problem applying the Roelands viscosity-pressure relationship have been produced by Almqvist and Larsson [42].

We follow the procedure that is usually employed to derive equation (6), but use equation (44) as our starting point instead of equation (7). On substituting (41) into the balance of linear momentum (44) we obtain [43]

$$-\text{grad } p + \mu(p)\Delta \mathbf{v} + 2\mathbf{D}[\text{grad } \mu(p)] + \rho \mathbf{b} = \rho \frac{d\mathbf{v}}{dt} \quad (45)$$

We drop the body force \mathbf{b} and restrict our attention to steady two-dimensional plane flows

$$-\frac{\partial p}{\partial x} + \mu(p)\Delta u + 2\mu'(p)\frac{\partial u}{\partial x}\frac{\partial p}{\partial x} + \mu'(p)\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)\frac{\partial p}{\partial y} = \rho\left[u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y}\right], \quad (46a)$$

$$-\frac{\partial p}{\partial y} + \mu(p)\Delta v + \mu'(p)\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)\frac{\partial p}{\partial x} + 2\mu'(p)\frac{\partial v}{\partial y}\frac{\partial p}{\partial y} = \rho\left[u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y}\right]. \quad (46b)$$

Equations (46a, b), will now be recast in terms of non-dimensional variables, defined by

$$(X, Y) = \frac{1}{L_{xz}}\left(x, \frac{1}{\varepsilon}y\right), \quad (U, V) = \frac{1}{U_*}\left(u, \frac{1}{\varepsilon}v\right), \quad P = \frac{p}{P_*}, \quad \bar{\mu} = \frac{\mu}{\mu_*} \quad (47)$$

$$Re^* = \frac{\rho U_* L_y}{\mu_*} \varepsilon, \quad \varepsilon = \frac{L_y}{L_{xz}}$$

μ_* and P_* represent characteristic viscosity and pressure, respectively. For expediency, we chose, $\mu_* = \mu_0$, $P_* = \mu_* U_x L_{xz} / L_y^2$ and employ the Barus formula

$$\bar{\mu} = \exp(\bar{\alpha}P), \quad \bar{\alpha} = \alpha P_* \quad (48)$$

with constant coefficient α , to characterize the pressure dependence of the viscosity.

It is not suggested here that the Barus formula (48) has much validity in the context of EHL calculations and is employed here only for illustrative purposes. Better fit to experiments, at least far from glass transition, is provided by Roelands [33], whose formula

$$\bar{\mu} = \exp\{(\ln \mu_0 + 9.67)[-1 + (1 + 5.1 \times 10^{-9} p)^{const}]\}$$

has been employed in EHL calculations by numerous investigators (see e.g., [44]). Recent viscosity measurements indicate, however, that for glass forming liquids the increase of viscosity with pressure is far more severe when nearing glass transition. Paluch *et al.* [34] find that in certain low-molecular-weight liquids viscosity variation with pressure can be described by

$$\bar{\mu} = \exp\left[\left(\frac{\text{const}}{P_0 - p}\right)p\right]$$

in effect replacing the constant exponent, α , of Barus by $[\text{const}/(P_0 - p)]$ where P_0 is the pressure of the ideal glass transition. Based on their experiments, Irving and Barlow [35] recommended a double exponential in the form

$$\bar{\mu} = \exp(Ae^{Bp} - Ce^{-Dp})$$

where A , B , C , and D , are constants at a given temperature.

Using (48), equations (46) are transformed into (here we drop the overhead bar on μ and α)

$$-\frac{\partial P}{\partial X} + \mu\left(\varepsilon^2 \frac{\partial^2 U}{\partial X^2} + \frac{\partial^2 U}{\partial Y^2}\right) + \alpha\mu\left[\frac{\partial U}{\partial Y} \frac{\partial P}{\partial Y} + \varepsilon^2\left(\frac{\partial V}{\partial X} \frac{\partial P}{\partial Y} + 2\frac{\partial U}{\partial X} \frac{\partial P}{\partial X}\right)\right] = Re^*\left(U \frac{\partial U}{\partial X} + V \frac{\partial U}{\partial Y}\right) \quad (49a)$$

$$-\frac{1}{\varepsilon^2} \frac{\partial P}{\partial Y} + \mu\left(\varepsilon^2 \frac{\partial^2 V}{\partial X^2} + \frac{\partial^2 V}{\partial Y^2}\right) + \alpha\mu\left[\frac{\partial U}{\partial Y} \frac{\partial P}{\partial X} + \varepsilon^2 \frac{\partial V}{\partial X} \frac{\partial P}{\partial X} + 2\frac{\partial V}{\partial Y} \frac{\partial P}{\partial Y}\right] = Re^*\left(U \frac{\partial V}{\partial X} + V \frac{\partial V}{\partial Y}\right) \quad (49b)$$

Let us now examine the consequence of taking $\varepsilon^2 \rightarrow 0$. On close examination of (49), we find that while the non-dimensional velocities and their derivatives are $O(1)$, the same does not hold true for the derivatives of the non-dimensional pressure. We must keep this in mind and neglect only $O(1)$ terms among those that are multiplied by ε^2 . Consistent with classical lubrication theory, we also assume that $Re^* \rightarrow 0$ and obtain

$$-\frac{\partial P}{\partial X} + \mu\left\{\frac{\partial^2 U}{\partial Y^2} + \alpha\left[\frac{\partial U}{\partial Y} \frac{\partial P}{\partial Y} + \varepsilon^2\left(\frac{\partial V}{\partial X} \frac{\partial P}{\partial Y} + 2\frac{\partial U}{\partial X} \frac{\partial P}{\partial X}\right)\right]\right\} = 0, \quad (50a)$$

$$-\varepsilon^{-2} \frac{\partial P}{\partial Y} + \mu\left\{\frac{\partial^2 V}{\partial Y^2} + \alpha\left[\frac{\partial U}{\partial Y} \frac{\partial P}{\partial X} + \varepsilon^2 \frac{\partial V}{\partial X} \frac{\partial P}{\partial X} + 2\frac{\partial V}{\partial Y} \frac{\partial P}{\partial Y}\right]\right\} = 0. \quad (50b)$$

We now appeal to equation (50) to estimate the order of magnitude of the pressure derivatives. To do this, we require two assumptions, the first of which has already been employed in arriving at equation (50).

Assumptions

(i) All velocities and their derivatives are $O(1)$, i.e., in (50) we have

$$\frac{\partial U}{\partial Y} \frac{\partial P}{\partial X} \gg \varepsilon^2 \frac{\partial V}{\partial X} \frac{\partial P}{\partial X}.$$

(ii) The pressure derivatives are not of the same order, in fact

$$\begin{aligned} \frac{\partial P}{\partial X} &\gg \frac{\partial P}{\partial Y}, \\ \alpha \frac{\partial U}{\partial Y} \frac{\partial P}{\partial X} &\gg \frac{\partial^2 V}{\partial Y^2}. \end{aligned}$$

On employing (i) and (ii) above, in conjunction with equation (50), we obtain the relative order of magnitude of the pressure derivatives

$$\frac{\partial P}{\partial Y} = \alpha \mu \varepsilon^2 \frac{\partial P}{\partial X} \frac{\partial U}{\partial Y}.$$

It follows immediately from assumptions (i) and (ii) that (now in primitive variables)

$$\begin{aligned} \frac{\partial p}{\partial x} &= \mu \frac{\partial^2 u}{\partial y^2} + \frac{d\mu}{dp} \left[\frac{\partial u}{\partial y} \frac{\partial p}{\partial y} + 2 \frac{\partial u}{\partial x} \frac{\partial p}{\partial x} \right], \\ \frac{\partial p}{\partial y} &= \frac{d\mu}{dp} \frac{\partial p}{\partial x} \frac{\partial u}{\partial y}, \\ \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} &= 0. \end{aligned} \tag{51}$$

To first approximation, flow of a lubricant with pressure dependent viscosity is governed by the system of equations (51). This system contains three equations in three unknowns and, presumably, can be solved to provide a more accurate estimate of the consequences of ignoring the pressure dependence of the viscosity. Nevertheless, here we make the additional simplifying assumption of neglecting the second of equations (51), merely to render the system more amenable to computations, as we demonstrate the effect of the additional terms. Under the approximations made here, the system of equations (51) reduces to

$$\frac{dp}{dx} = \mu \frac{\partial^2 u}{\partial y^2} + 2 \frac{d\mu}{dp} \frac{\partial u}{\partial x} \frac{dp}{dx} , \quad (52)$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 . \quad (53)$$

Following the analysis of Reynolds, we formally integrate equation (52) subject to the usual boundary conditions and substitute the result into equation (53). Upon integration across the film and observing that $v(h) = v(0) = 0$, we obtain

$$\frac{d}{dx} \left[\left(\frac{h^3}{\mu} - 12\alpha \int_0^h y(h-y) \frac{\partial u}{\partial x} dy \right) \frac{dp}{dx} \right] = 6\tilde{U} \frac{dh}{dx} . \quad (54)$$

To illustrate the kind of difference that might arise on using the modified equation (54), compared to conventional analysis, we examine the elastic cylinder rolling on a plane under the following conditions

$$\frac{w}{E'R} = 1.05 \times 10^{-5} , \quad \frac{\mu_0 \tilde{U}}{E'R} = 1.0 \times 10^{-11} , \quad \alpha E' = 3.0 \times 10^3$$

$$X_{in} = -3.0 , \quad X_{out} = 1.5 ,$$

Figure 6 displays the pressure distribution and figure 7 the viscosity distribution for the classical Reynolds equation and the modified Reynolds equation, within the vicinity of the pressure peak. As may be concluded, these figures show slightly increased peak pressure and significantly higher peak viscosity for the modified equation, relative to the classical solution. This result is in line with our earlier findings for rigid cylinders.

Numerical solutions of the modified pressure equation show that proper accounting for pressure dependence of the lubricants viscosity yields slightly higher pressures, but at much increased viscosity, relative to classical analysis. This conclusion holds for both rigid and elastic cylinders in a cylinder-rolling-on-a-plane problem.

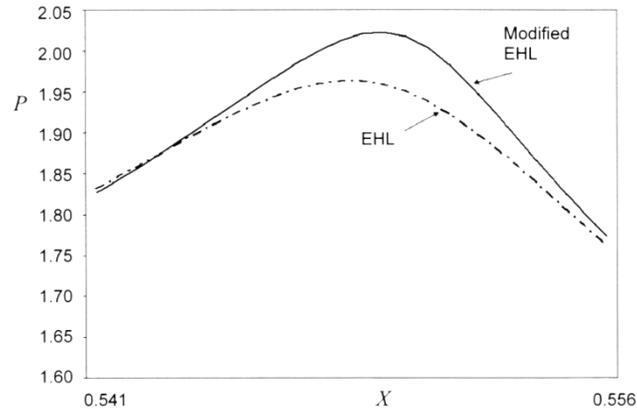


Figure 6. Non-dimensional pressure.

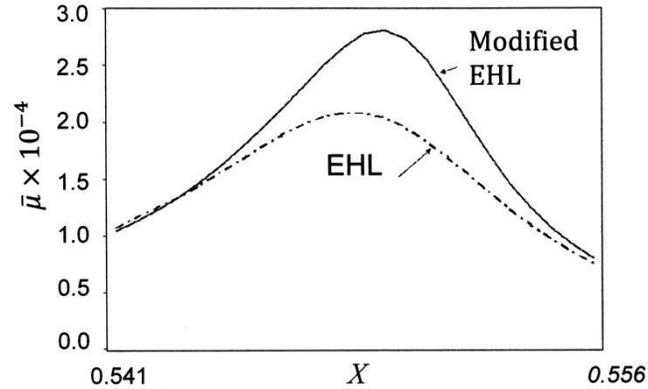


Figure 7. Non-dimensional viscosity.

4. Molecularly thin films

Although the lubrication approximation is derived for thin films, there is, nevertheless, a thin film limit to the approximation's validity. When the characteristic dimensions of the device containing the fluid approach the mean free path (for gases) or the dimension of the molecules (for liquids) the continuum assumption breaks down. There are two distinct representations at our disposal for fluids, continuum and particle. The former is applicable only with restrictions while the particle representation is valid over the whole range of conditions.

The mathematical models that specify the representation of molecular interactions in fluids are shown in table 3 [45]. Particle based representation, which is at the most fundamental level of this hierarchy, is of two kinds, deterministic, in

which the motion of each particle in an ensemble of molecules is followed, and statistical, in which the evolution of the probability density function for the molecules is investigated [46].

Table 3. Types of analysis.

	Type	Equation (Method)	Fluid
Continuum	Deterministic	Navier-Stokes, no-slip (CFD)	Gas-liquid
		Navier-Stokes, slip (CFD)	Gas-liquid
Particle	Deterministic	Newton (MD)	Liquid
	Statistical	Liouville (DSMC)	Gas
		Boltzmann (CFD)	Gas

The deterministic particle based method, molecular dynamics (MD) simulation [47], although theoretically valid for a whole range of conditions, is employed mainly for liquids, as the long flight paths between collisions for gases makes forward integration of the equations prohibitively expensive. In a liquid, the molecules are densely packed, leading to a more efficient application of MD simulation. Sanbonmatsu and Tung, [48] simulated the dynamics of 2.64×10^6 atoms for a total of 22 ns sampling.

The statistics based particle method, on the other hand, presupposes well developed kinetic theory, which is not available for liquids. In addition, the equations employed here are derived for low-density packing of molecules, making these methods applicable to gases. Both Monte Carlo approaches and the Boltzmann equation are derived from the Liouville equation, a conservation equation of the n-dimensional probability function. The direct simulation Monte Carlo (DSMC) method [49, 50] may be used for dilute fluids when the ratio of average molecular spacing to molecular diameter $\sigma/d \geq 10$. The basic assumption of DSMC is to uncouple molecular motions from intermolecular collisions over small time intervals. Particle motions are modeled deterministically, while collisions are treated statistically. The Boltzmann equation for the one-particle distribution function $f(\mathbf{x}, \mathbf{c}, t)$, where \mathbf{x} is the location of the particle and \mathbf{c} its velocity, is applicable over the whole range of the Knudsen number, $Kn = \lambda/h$; it is usually solved by computational fluid dynamics (CFD) methods.

Breakdown of the continuum model is best illustrated for gas flow, as gases have well-developed kinetic theory.

When the characteristic dimension of the flow device can accommodate a large enough number of gas molecules, the fluid can be considered to have matter continuously distributed throughout the space it occupies. However, as devices are made smaller and smaller, attention must be paid eventually to the fact that the

gases consist of discrete molecules. The generally chaotic motion of these molecules, at speeds comparable to the velocity of sound, is punctuated by frequent collisions of about 10^{10} collisions per second. If the density of the gas is small, i.e., the average spacing of the molecules relative to their dimension is large, the collisions between molecules will be binary collisions.

There are numerous technical applications of gas flow in ultrathin channels. One such application is data recording. The density of information storage on a disk increases dramatically with a decrease of the flying height of the read/write head above the disk. However, the flying height cannot be reduced to zero, as this would cause excessive wear of the surfaces. Currently, computer hard drives are manufactured in which the minimum separation of the read/write head from the disk is of the order of the mean free path.

For air at standard temperature and pressure (STP) the ratio of mean free path λ , average spacing of molecules σ , and molecular diameter d are $\lambda : \sigma : d \sim 170 : 10 : 1$. A cube $1 \mu\text{m}$ at the edges contains $n \sim 2.9 \times 10^7$ molecules, and the mean free path is $\lambda \sim 60 \text{ nm}$. Between collisions, the molecules travel along straight-line trajectories, with no intermolecular forces acting on them [58].

The conditions that apply to a gas in thin films are best described with reference to the Knudsen number, $Kn = \lambda/L_y$ (cf. figure 8). $Kn \rightarrow 0$ is the domain of continuum flow while $Kn \rightarrow \infty$ typifies collisionless molecular flow. The various Knudsen number regimes are: $Kn = 0$ for Euler flow, for $0 < Kn < 0.001$ the flow is governed by the Navier-Stokes equation. For flows with Knudsen number above $Kn = 0.001$, the continuum approach is still usable if we allow slip to occur at the boundaries [45]. This was demonstrated by Shaaf and Sherman [51] among others, who measured the drag on a flat plate in a wind tunnel. The Boltzmann equation holds for the full Kn range, but, as the equation is difficult to solve, the continuum approach is advocated whenever applicable. For $Kn \rightarrow \infty$ the collisionless Boltzmann equation applies.

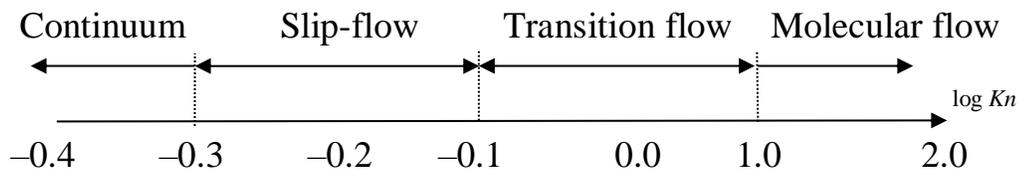


Figure 8. Knudsen number range of various gas flow regimes.

4.1 Velocity slip

It was observed by Kundt and Warburg as early as 1857 that in rarefied gas flow the solid boundary does not support no-slip (velocity) and no-jump (temperature) boundary conditions. A consistent slip condition to make the Navier-Stokes equations valid in the $0.001 < Kn < 0.1$ regime, the so-called slip flow regime, is

$$u_{slip} = L_s \left. \frac{\partial u}{\partial y} \right|_0 \quad (55)$$

where $(\partial u / \partial y)_0$ is the gradient at a point $P \rightarrow wall$. The coefficient L_s is called the slip coefficient; its importance in the slip-flow regime is comparable to the coefficients of viscosity and heat conduction. Albertoni *et al.* [52] tabulated the slip coefficient for various authors, they calculate $L_s = 1.1466\lambda$ from the BGK (Bhatnagar-Gross-Krook) model of the collision term in the Boltzmann equation [53].

To examine the slip-flow boundary conditions we let $u(y)$ represent the velocity parallel to the boundary at y . The shear stress on the small area ds , oriented parallel to the boundary, is given by the momentum transport across it

$$\tau = \nu m (u_+ - u_-), \quad (56)$$

where ν is the frequency of collisions of molecules with ds , m is the molecular mass and $u_+ = u(y^+)$, $u_- = u(y^-)$. The velocities u_+ and u_- are provided by a Taylor expansion of u about y :

$$u_{+,-} = u(y) + \frac{\partial u}{\partial y} (y^{+,-} - y) + \frac{1}{2} \frac{\partial^2 u}{\partial y^2} (y^{+,-} - y)^2. \quad (57)$$

Substituting (57) into (56) we obtain

$$\tau = \nu m \frac{\partial u}{\partial y} (y^+ - y^-) \quad (58)$$

and by comparing (58) with the macroscopic equation $\tau = \mu (\partial u / \partial y)$, the viscosity is found to be

$$\mu = \nu m (y^+ - y^-). \quad (59)$$

Now applying the foregoing to the boundary at y , where y^- is the location of the “displaced” solid wall, so that

$$u_- = u_{plate}, \quad u(y) = u_{gas} \text{ (velocity of gas at wall),}$$

we have

$$u_+ = u_{gas} + \left. \frac{\partial u}{\partial y} \right|_{y=0} (y^+ - y) + \frac{1}{2} \left. \frac{\partial^2 u}{\partial y^2} \right|_{y=0} (y^+ - y)^2 \quad (60)$$

Substituting equation (60) into equation (59), and making use of equation (58), we find

$$\begin{aligned} \tau &= \alpha \nu m \left[\left. \frac{\partial u}{\partial y} (y^+ - y) + \frac{1}{2} \frac{\partial^2 u}{\partial y^2} (y^+ - y)^2 + u_{gas} - u_{plate} \right] \\ &= \nu m (y^+ - y^-) \left. \frac{\partial u}{\partial y} \right|_{y=0}. \end{aligned} \quad (61)$$

The slip velocity at the boundary is calculated as the difference between the apparent velocity at the wall, u_{gas} , and the prescribed plate velocity, u_{plate} , [54]

$$u_{slip} = u_{gas} - u_{plate} = \frac{(y^+ - y^-) - \alpha (y^+ - y)}{\alpha} \left. \frac{\partial u}{\partial y} \right|_{y=0} - \frac{1}{2} \left. \frac{\partial^2 u}{\partial y^2} \right|_{y=0} (y^+ - y)^2 \quad (62)$$

Here we introduced α , $0 < \alpha < 1$, the momentum accommodation coefficient (see e.g., Bird, 1994).

In the 1st order slip [55] and the 2nd order slip [56] models $y^+ - y = \lambda$, while $y^+ - y = (2/3)\lambda$ in the 1.5 order slip model of [57], taking into account that the y -direction component of the averaged distance between collisions is $(2/3)\lambda$ (see Vincenti and Kruger [58] and Bird [49] on this topic).

Using equation (8) and $U_* = \sqrt{2RT_0}$ to define nondimensional coordinates and velocity components, respectively, and putting $(2 - \alpha)/\alpha = a$

The thin film approximation in hydrodynamic, including elastohydrodynamic, lubrication

$$\begin{aligned}
 U|_{Y=1} &= U_0 + a Kn \frac{\partial U}{\partial Y} \Big|_n + \frac{1}{2} (\beta Kn)^2 \frac{\partial^2 U}{\partial Y^2} \Big|_n \\
 U|_{Y=0} &= -a Kn \frac{\partial U}{\partial Y} \Big|_0 - \frac{1}{2} (\beta Kn)^2 \frac{\partial^2 U}{\partial Y^2} \Big|_0
 \end{aligned} \tag{63}$$

In the notation of equation (8), $\beta = 0$ for 1st order slip, $\alpha = \beta = 1$ for 2nd order slip, and $\beta = 2/3$ for 1.5 order slip (however, Mitsuya multiplies the first order term by 3/2). Using the boundary conditions (63), the velocity distribution along x is given as

$$u = \frac{h^2}{2\mu} \frac{\partial p}{\partial x} \left[\left(\frac{y}{h} \right)^2 - \left(\frac{y}{h} \right) - a Kn - (\beta Kn)^2 \right] + U \left[1 - \frac{(y/h) + Kn}{1 + 2 Kn} \right] \tag{64}$$

Using equation (64) and a similar expression for w , a Reynolds equation can be derived by methods similar to those for incompressible fluids (see e.g., [2]). The procedure yields

$$\begin{aligned}
 \frac{\partial}{\partial X} \left\{ PH^3 \frac{\partial P}{\partial X} \left[1 + 6a \left(\frac{Kn}{PH} \right) + 6 \left(\beta \frac{Kn}{PH} \right)^2 \right] \right\} \\
 + \frac{\partial}{\partial Z} \left\{ PH^3 \frac{\partial P}{\partial Z} \left[1 + 6a \left(\frac{Kn}{PH} \right) + 6 \left(\beta \frac{Kn}{PH} \right)^2 \right] \right\} = \Lambda \frac{\partial PH}{\partial X}
 \end{aligned} \tag{65}$$

Here p_a is the ambient pressure, h_0 is the minimum film thickness, $P = p/p_a$, $H = h/h_0$ and the bearing number Λ has the definition

$$\Lambda = \frac{6\mu UL}{p_a h_0^2}$$

The Reynolds equation for continuum gas flow can be obtained from equation (65) by setting $\alpha = \beta = 0$.

Wu *et al.* [59] derived first and second order slip models by summing the contributions from each group of molecules impinging on the surface at an angle; this, in effect, means relaxing the requirement that the length scale in the Taylor

expansion (62) equals the mean free path (20). The coefficients multiplying the Kn/PH and the $(Kn/PH)^2$ terms in their scheme are $4a$ and 3 , respectively.

Comparison with experiments suggest that the continuum model allowing for slip at the boundaries yields good results for $Kn \leq 1$ according to Odaka *et al.* (quoted by Fukui and Kaneko [60]), and even for $Kn \leq 2.5$ according to Hsia and Domoto [56]. Nevertheless, it is difficult to justify its use for $Kn > 1$.

4.2. Molecular gas lubrication

To construct a model of gas flow valid for arbitrary Knudsen number, we must make recourse to the Boltzmann equation, a conservation equation of the one-particle distribution function $f(x, y, z, c_x, c_y, c_z, t)$

$$\frac{\partial f}{\partial t} + \mathbf{c} \cdot \nabla f = \left\{ \frac{\delta f}{\delta t} \right\}_{coll} \quad (66)$$

Here, $f(c_i)dV_x dV_c$ is the expected number of molecules that lie simultaneously in the volume elements $dV_x = dx_1 dx_2 dx_3$ of physical space, located at \mathbf{x} , and $dV_c = dc_1 dc_2 dc_3$ of velocity space, located at \mathbf{c} [58]. The right hand side of equation (66) represents the collision integral that contains the probability distribution $f(c_i)$, making equation (66) a non-linear integro-differential equation.

Because of the nonlinearity of the collision term, the Boltzmann equation is not easy to solve. A particular class of solutions, namely the Maxwellians, describes equilibrium states. The left hand side of the Boltzmann equation (66) is then zero and the equilibrium distribution function, f_0 , is its solution [58].

$$\left\{ \frac{\delta f_0}{\delta t} \right\}_{coll} = 0$$

The Maxwellian distribution function is given by

$$f_0 = n_0 \left(\frac{m}{2\pi kT} \right)^{3/2} \exp \left[\frac{-m(\mathbf{c} - \mathbf{v})^2}{2kT} \right], \quad (67)$$

where k is the Boltzmann constant, and T is the temperature.

Expanding the Boltzmann equation in powers of the Knudsen number [61]

$$f = f_0 + Kn f_1 + Kn^2 f_2 + \dots,$$

the first term is the Maxwellian equilibrium distribution and the corresponding conservation equation simplifies to Euler's equation of inviscid flow. The second equation reduces to the Navier-Stokes equation while the third term produces the Burnett equation.

For small departures from equilibrium, when the macroscopic velocity of the gas is small relative to the thermal velocity of its molecules, the collision integral can be linearized [60, 62]. Note that since the velocity of sound is of the order of the root mean square of the molecular velocity, the condition of linearity may be written in terms of the Mach number as $M \rightarrow 0$.

To derive a Reynolds type equation from the linearized form of equation (66) the flow rate must first be calculated [60]. Due to linearity, the flow rate q is a superposition of the pressure flow q_p , the Couette flow q_C and the thermal creep flow q_T ,

$$q = q_p + q_C + q_T$$

Owing to the symmetry property of the Couette velocity profile, q_C is not dependent on the Knudsen number. As for the other two flow components, q_p has been evaluated by Cercignani *et al.* [63, 64]

$$q_p = -Q_p \frac{h^2}{\sqrt{2RT_0}} \left(\frac{dp}{dx} \right) \quad (68a)$$

and q_T by Loyalka [65]

$$q_T = Q_T \frac{ph^2}{T_0 \sqrt{2RT_0}} \left(\frac{dT_w}{dx} \right). \quad (68b)$$

In equation (68) the Q_p and Q_T are functions of the Knudsen number, they can be found in the above cited papers and also reproduced in Fukui and Kaneko [60].

A generalized Reynolds equation based on the Boltzmann equation can now be expressed as [60]

$$\text{div} \left\{ PH^3 \left[\overline{Q}_p(D_0PH, \alpha) \cdot \text{grad}P - \overline{Q}_T(D_0PH, \alpha) \cdot P \cdot \text{grad}\tau_w \right] \right\} = \Lambda \text{grad}(PH), \quad (69)$$

where $D_0 = p_0 h_0 / \mu \sqrt{2RT}$ is the “characteristic” inverse Knudsen number calculated on the reference state, and is an assigned parameter.

The problem, as presented by equation (69) is laborious to solve. The nonlinear coefficients must be obtained from certain integral equations at each step of the iteration. Fukui and Kaneko presented an alternative method of solution in a later paper [66], in which they solved the finite difference approximation to the integral equations referred to above, ahead of time then used the newly created database to interpolate for flow coefficients in terms of the Knudsen number.

In the next few figures, reproduced from Fukui and Kaneko [60] and Mitsuya [57], we shall compare results from the various models with one another and with experimental data. The variation of the pressure flow rate, Q_p , with the inverse Knudsen number is shown in figure 9. The continuum theory with no-slip yields acceptable results for $Kn < 0.01$, while the 2nd order slip result is good for $Kn < 1$. For $Kn > 1$, the Boltzmann equation results differ drastically from slip flow model predictions.

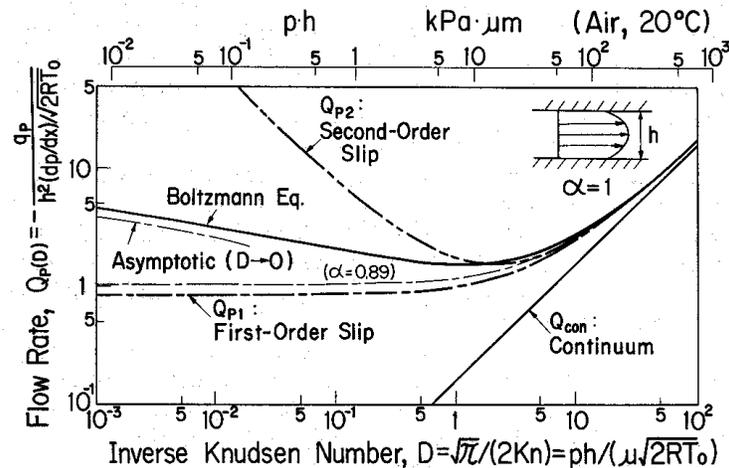


Figure 9. Variation of pressure flow coefficient with the inverse Knudsen number [60].

The load capacity of inclined plates in relative sliding at $A = 10$ and varying D is shown in figure 10. The Boltzmann solution seems to be bracketed by the 1st order slip and 2nd order slip solutions.

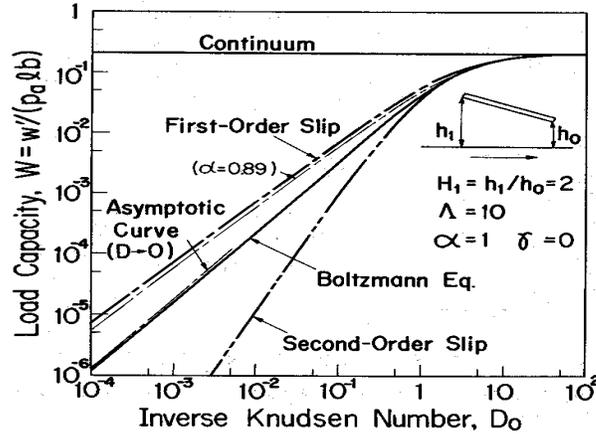


Figure 10. Load capacity of plane slider [60].

As a final comparison between model predictions and experiment, figure 11 plots minimum film thickness between a read/write head and computer disk. The 1.5 order slip model seems to perform best in this case, after the Boltzmann model, of course.

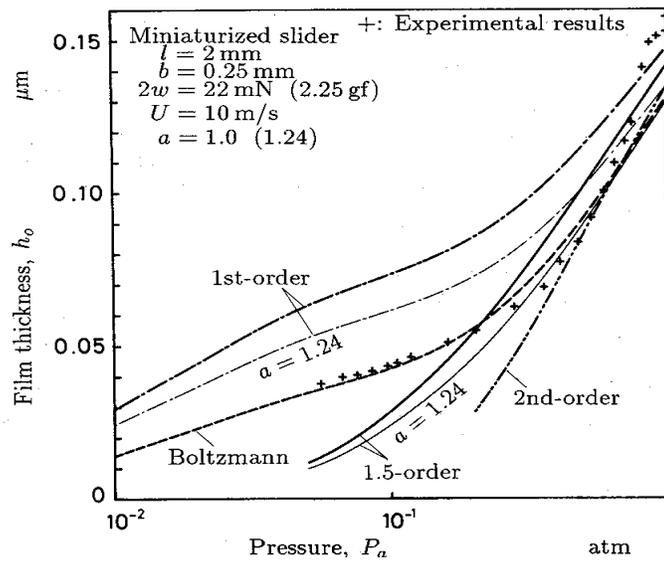


Figure 11. Minimum film thickness for slider [57].

4.3. Liquids

As the temperature of a gas is lowered, the thermal velocity of the molecules decreases and the molecules become more densely packed. A cube of liquid $1 \mu\text{m}$ at the edges now contains $n \sim 3.35 \times 10^{10}$ molecules, and the average molecular spacing decreases to $\sigma = 0.31 \text{ nm}$. We can no longer speak of mean free path, as

the molecules are now too closely packed. Liquid molecules are in a continuous state of collision.

The difference between liquids and gases is that for gases there are no forces acting on a molecule between successive collisions, while the molecules of a liquid are so closely packed that the forces of attraction between molecules can manifest. Therefore, it might be expected that the behavior of liquids and gases close to a solid boundary will also show similarity and under certain surface-liquid conditions the no-slip boundary condition will not be satisfied. The degree of slip will then depend on the strength of the solid-fluid coupling.

Experimental evidence is contradictory in this respect. Derjaguin (quoted in [67]) found that the viscosity of water in quartz capillaries with a diameter less than 100 nm, attains a value 40-50% higher than in bulk. However, this conclusion did not hold for non-polar liquids. Chan and Horn [68] studied the drainage of fluids between two atomically smooth mica surfaces. Their results are in excellent agreement with the Reynolds theory of lubrication for film thickness $h > 50$ nm. For thinner films, they find that drainage is somewhat slower than predicted by continuum theory, as about two molecular layers on each surface undergo no shear. Thus, in thinner films there is an apparent enhancement of viscosity, which can be accounted for by allowing the plane of shear to be displaced into the liquid. Israelachvili [67] reported that in films as thin as 5 nm the “plane of slip” is within a few Angstrom units of the interface and the viscosity is within 10% of its bulk value. Viscosity increase in thin channels was reported by Migun and Prokhorenko [69], while Debye and Cleland [70], and Pfahler *et al.* [71] found the apparent viscosity, μ_a , consistently smaller than μ .

Liquids in large gaps, or liquids above a single wall, are fluid all the way to within one or two molecular layers of the solid surfaces. When the gap is squeezed down to the thickness of a few molecules, the confined liquid is solid-like. Klein and Kumacheva [72] find that as the film is made increasingly thinner, the effective viscosity of the liquid changes by at least seven orders of magnitude over a change in film thickness of a single molecular spacing (figure 12)². When $n > n_c$, where n represents the number of molecules across the film and n_c is the critical number of molecules for liquid-solid transition, there is liquid like behavior, and solid-like behavior when $n < n_c$. In their experiments Klein and Kumacheva [72] find $n_c = 7$.

² Reprinted from Klein, J. and Kumacheva, E., *Physica A*, 249, 206-215 (1998) with permission by Elsevier.

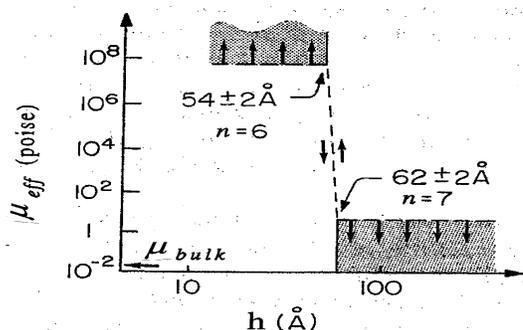


Figure 12. Variation of the effective mean viscosity, μ_{eff} , with film thickness, h . Confined liquid film [72].

This fluid-solid like transition can be attributed to a change in the geometric packing of the molecules, as they undergo layering, extending a few molecules away from the surface. The dramatic increase in viscosity is due to abrupt transition caused by confinement alone. Hu and Granick [73] recognize a special lubrication regime, located between EHL and boundary lubrication. In this thin film regime, “lubricant flow and fluid dynamics are still in action but behave differently from expectations of classical theory.”

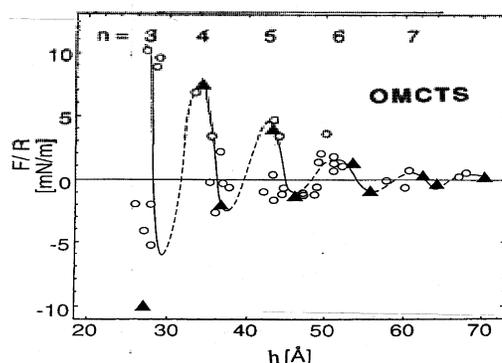


Figure 13. Solvation force, indicating the number n of molecular layers corresponding to each ‘hump’ [72].

When atomically smooth mica sheets are made to approach one another, the normal force acting at large separations is the continuum van der Waals forces. As the separation between the surfaces is decreased, the solvation force becomes an oscillating function of the separation [67, 68], due to the layering of the molecules (deviations from the predictions of continuum theory are often referred to as solvation effects). The period of oscillation is roughly the same as the diameter of the liquid molecules. The number of layers, n , can be counted by dividing the spacing at the maxima by the molecular diameter, as indicated in figure 13².

Jang and Tichy [74] used the exponential-cosine curve fit of Chan and Horn [68]

$$F_{solv} = -RB \exp\left(-\frac{h}{d}\right) \cos\left(\frac{2\pi h}{d}\right), \quad B = 172 \text{ MPa}$$

to represent the solvation pressure in their thin film lubrication (TFL) correction to EHL theory. The film thickness begins to deviate from the conventional EHL theory [75] when the separation is less than 7-8 nm, and changes stepwise. Matsuoka and Kato [76], who calculated the solvation pressure from a force potential [77], in figure 14 compare experimental data with theoretical predictions.

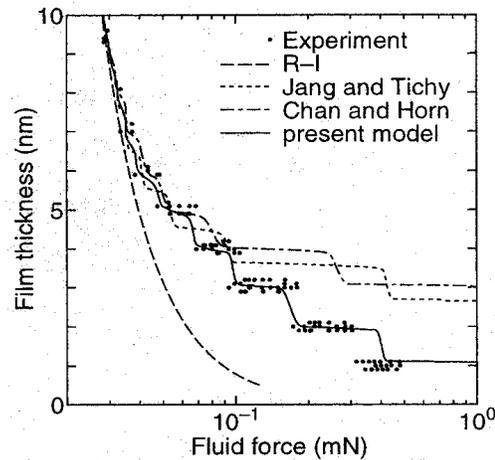


Figure 14. Comparison of three solvation pressure models with experiment for OMCTS [76].

5. Summary

Often, it is not too difficult to state that the just-derived approximation becomes increasingly accurate as a defining parameter approaches a definite value, say zero. It is considerably more challenging but at the same time, more useful, to state with certainty that just how small the parameter must be for the approximation to yield useful results. Nevertheless, we will now attempt to define the range of validity of the major assumptions of Reynolds' lubrication theory.

The continuum assumption with no-slip boundary conditions is valid for films whose thickness is considerably greater than the mean free path for gases or significantly more than 10-12 molecular thickness for liquids. The assumption will

still yield acceptable answers for films whose thickness is comparable to the mean free path, if velocity slip is specified at the boundaries. Liquid films are far less understood than gas films; nevertheless, it is clear that for confined liquids there is complete agreement with the Navier–Stokes theory, including the no-slip boundary condition except under conditions of extraordinarily high rates of shear, when the film is more than 10 molecules thick. However, a liquid-solid transition occurs when the number of molecules across the film falls below a critical value, the transition resulting in an increase of viscosity by several orders of magnitude. This sequence of events points to the existence of a thin-film lubrication regime, located between EHL and boundary lubrication.

Lubricant inertia is negligible and the classical Reynolds equation yields good results when the reduced Reynolds number $Re^* \leq 1$ and the local value of the film slope $\varepsilon \leq 0.1$. If the reduced Reynolds number is larger than this value but the local value of the slope remains small, the flow is essentially two-dimensional. The pressure, however, must be calculated in such cases from equations that retain fluid inertia; the classical Reynolds equation is no longer valid. The same holds true whether the flow is laminar or turbulent.

Though the derivation of the Reynolds equation currently in use for elastohydrodynamic lubrication is less than rigorous, current EHL theory yields acceptable results in most applications [78]. There might be cases, however, especially those associated with significant heat generation in the film, where the governing equations must be derived from first principles.

6. References

1. Reynolds, O. 1986. On the theory of lubrication and its application to Mr. Beuchamp Tower's experiments. *Phil Trans. Roy. Soc.*, **177**, 157-234.
2. Szeri, A. Z. 1998. *Fluid Film Lubrication: Theory and Design*, Cambridge University Press.
3. Dai, R. X., Dong, Q. M. and Szeri, A. Z. 1992. Approximations in Hydrodynamic Lubrication. *J. Tribology*, **114**, 14-25.
4. Szeri, A. Z. and Snyder, V. 2006. Convective inertia effects in wall-bounded thin film flows. *Meccanica*, **41**, 473-482.
5. Sun, D. C. and Chen, K. K. 1977. First effects of Stokes roughness on hydrodynamic Lubrication. *J. Lub. Tech.*, **99**, 2-9.
6. Van Odyck, D. E. A. and Venner, C. H. 2003. Compressible Stokes flow in thin films. *J. Tribology*, **125**, 543-551.
7. Hinze, J. O. 1975. *Turbulence*, 2d ed., McGraw-Hill, New York.
8. Constantinescu, V. N. 1962. Analysis of bearings operating in turbulent regime. *Trans. ASME*, **82**, 139-151.

9. Elrod, H. G. and Ng, C. W. 1967. A theory for turbulent films and its application to bearings. *Trans. ASME*, **89**, 347-362.
10. Ng, C. W. and Pan, C. H. T. 1965. A linearized turbulent Lubrication theory, *Trans. ASME*, **87**, 675-688.
11. Constantinescu, V. N. 1959. On turbulent lubrication. *Proc. Inst. Mech. Eng.*, **173**, 881-900.
12. de Boor, C. 1978. *A Practical Guide to Splines*. Springer-Verlag.
13. Ritchie, G. S. 1968. On the stability of viscous flow between eccentric rotating cylinders. *J. Fluid Mech.*, **32**, 131-144.
14. Ballal, B. and Rivlin, R. S. 1976. Flow of a Newtonian fluid between eccentric rotating cylinders. *Arch. Rat. Mech. Anal.*, **62**, 237-274.
15. Myllerup, C. M. and Hamrock, B. J. 1994. Perturbation approach to hydrodynamic lubrication theory. *J. Tribology*, **116**, 110-118.
16. Nazarov, S. A. and Videman, J. H. 2007. A modified nonlinear Reynolds equation for thin viscous flows in lubrication. *Asymptotic Analysis*, **52**, 1-36.
17. Becker, J. and Grun, G. 2005. The thin film equation: recent advances and some new perspectives. *J. Phys. Cond. Matter*, **17**, 291-307.
18. Schwartz, L., Roy, R. V., Elly R. R. and Princen H. M. 2004. "Surfactant-driven motion and splitting of droplets on a substrate" *J. Eng. Math.*, **50**, 157-175.
19. Alvarez, A. and Soto, R. 2005. Dynamics of a suspension confined in a thin cell. *Phys. Fluids* **17**, 093103.
20. Munch, A. and Wagner, B. 2005. Contact-line instability of dewetting thin films. *Physica D – Nonlin. Phenomena* **209**, 178-190.
21. Stokes, G. G. 1845. On the theories of the internal friction of fluids in motion, and of the equilibrium and motion of elastic solids. *Trans. Cambridge Phil. Soc.* **8**, 287-305.
22. Bridgman, P. W. 1931. *The Physics of High-Pressure*, New York, MacMillan.
23. Cutler, W. G., McMickles, R. J., Webb, W. and Schiessler, R. W. 1958. Study of the compressions of several high molecular weight hydrocarbons. *J. Chem. Phys.* **29**, 727-740.
24. Griest, E. M., Webb, W. and Schiessler, R. W. 1958. Effect of pressure on viscosity of higher hydrocarbons and their mixtures, *J. Chem. Phys.*, **29**, 711-720.
25. Johnson, K. L. and Cameron, R. 1967. Shear behaviour of elastohydrodynamic oil films at high rolling contact pressures. *Proc. Instn. Mech. Engrs.*, **182**, 307-319.
26. Johnson, K. L. and Greenwood, J. A. 1980. Thermal analysis of an Eyring fluid in elasto-hydro-dynamic traction. *Wear* **61**, 355-374.
27. Johnson, K. L. and Tevaarwerk, J. L. 1977. Shear behaviour of elastohydrodynamic oil films. *Proc. R. Soc. Lond.*, **A356**, 215-236.
28. Malek, J., Necas, J. and Rajagopal, K. R. 2002. Global analysis of the flow of fluids with pressure dependent viscosities. *Arch. Ratl. Mech. Anal.*, **165**, 243-269.
29. Malek, J., Necas, J. and Rajagopal, K. R. 2002. Global existence of solutions for flows of fluids with pressure and shear dependent viscosities. *Applied Math. Letters*, **15**, 961-967.
30. Hron, J., Malek, J. and Rajagopal, K.R. 2001. Simple flows of fluids with pressure-dependent viscosities. *Proc. Royal Society Lond.*, **A457**, 1603-1622.
31. Dowson, D. and Higginson, G. R. 1966. *Elastohydrodynamic Lubrication. The Fundamentals of Roller and Gear Lubrication*, Pergamon, Oxford.

32. Renardy, M. 1986. Some remarks on the Navier-Stokes equations with a pressure-dependent viscosity. *Comm. Partial Diff. equation*, **11**, 779-793.
33. Roelands, C. J. A. 1966. Correlation aspects of the viscosity-temperature-pressure relationship of lubricating oils. PhD dissertation, Technische Hogeschool Delft, The Netherlands.
34. Paluch, M., Dzendzik, Z. and Rzożka, S. J. 1999. Scaling of high -pressure viscosity data in low-molecular-weight glass-forming liquids. *Phys. Rev. B*, **60**, 2979-2982.
35. Irving J. B. and Barlow, A. J. 1971. An automatic high pressure viscometer. *J. Phys. E.*, **4**, 233-236.
36. Bendler, J. T., Fontanella, J. J. and Schlesinger, M. F. 2001. A new Vogel-like law: ionic conductivity, dielectric relaxation, and viscosity near the glass transition. *Phys. Rev. Let.*, **87**, 195503-1-4.
37. Gazzola, F. 1997. A note on the solution of Navier-Stokes equations with a pressure dependent viscosity. *Z. Angew. Math. Phys.* **48**, 760-773.
38. Gazzola, F. and Secchi, P. 1998. Some results about stationary Navier-Stokes equations with a pressure dependent viscosity. In *Navier-Stokes Equations: Theory and Numerical Methods* (ed. R. Salvi), 31-37, New York, Longman.
39. Bair, S., Khonsari, M. and Winer, W. O. 1998. High-pressure rheology of lubricants and limitations of the Reynolds equation. *Trib. Int.*, **10**, 573-586.
40. Schafer, C. T., Giese, P., Rowe, W. B. and Woolley, N. H. 1999. Elastohydrodynamically lubricated line contact based on the Navier-Stokes equations. *Proc. Leeds-Lyon Symp. Trib.*, 57-68, Elsevier. *Soc.*, **177**, 157-234.
41. Greenwood, J. A. 2000. In *Thinning Films and Tribological Interfaces. Proc 26th Leeds-Lyon Symp.* (ed. D. Dowson), Tribology series, **28**, 793-794.
42. Almqvist, T. and Larsson, R. 2002. The Navier-Stokes approach for thermal EHL line contact solution. *Trib. Int.*, **35**, 163-170.
43. Rajagopal, K. R. and Szeri, A. Z. 2003. On an inconsistency in the derivation of the equations of elastohydrodynamic lubrication. *Proc. Roy. Soc. Lond.*, **A 459**, 2771-2786.
44. Venner, C. H. and Lubrecht, A. A. 1994. Transient analysis of surface features in an EHL line contact in the case of sliding. *J. Tribology*, **116**, 186-193.
45. Gad-el-Hak, M. 1999. The fluid mechanics of micro devices - The Freeman Scholar Lecture. *J. Fluids Eng.*, **121**, 5-33.
46. Ungerer, P., Nieto-Draghi, C., Rousseau, B., Ahunbay, G. and Lachet, V. 2007. Molecular simulation of the thermophysical properties of fluids: From understanding toward quantitative predictions. *J. Molecular Liquids*, **134**, 71-89.
47. Koplik, J. and Banavar, J. R. 1995. Continuum Deductions from Molecular Hydrodynamics. *Ann. Rev. Fluid Mech.*, **27**, 957-992.
48. Sanbonmatsu, K.Y and Tung, C.S. 2006. High performance computing in biology: Multimillion atom simulations of nanoscale systems. *J. Structural Biology*, **157**, 470-480.
49. Bird, G. A. 1994. *Molecular Gas Dynamics and the Direct Simulation Of Gas Flows*, Clarendon Press, Oxford.
50. Oran, E. S., Oh, C. K. and Cybyk, B. Z. 1998. Direct Simulation Monte Carlo: Recent Advances and Applications. *Annual Rev. Fluid Mech.*, **30**, 403-441.
51. Schaaf, S. A. and Sherman, F. S. 1954. Skin friction in slip flow. *J. Aero. Sci.*, **21**, 85-90.

52. Albertoni, S., Cercignani, C. and Gotusso, L. 1963. Numerical evaluation of the slip coefficient. *Phys. Fluids*, **6**, 993-996.
53. Bhatnagar, P. L., Gross, E. P. and Krook, M. 1954. A model for collision processes in gases. I. Small amplitude processes in charged and neutral one-component systems. *Phys. Review*, **94**, 511-525.
54. Szeri, A. Z. 2001. Flow modeling of thin films from macroscale to nanoscale. *Fundamentals of Tribology and Bridging the Gap between the Macro-Micro/Nanoscales*, 767-798, Kluwer, Netherlands.
55. Burgdorfer, A. 1959. The influence of molecular mean free path on the performance of hydrodynamic gas lubricated bearings. *J. Basic Engr.*, **81**, 94-100.
56. Hsia, Y. T. and Domoto, G. A. 1983. An experimental investigation of molecular rarefaction effects in gas lubricated bearings at ultra-low clearances. *J. Tribology*, **105**, 120-130.
57. Mitsuya, Y. 1993. Modified Reynolds equation for ultra-thin film gas lubrication using 1.5 order slip flow model and considering surface accommodation coefficient. *J. Tribology*, **115**, 289-294.
58. Vincenti, W. G. and Kruger, C. H. 1965. *Introduction To Physical Gas Dynamics*, John Wiley & Sons, New York.
59. Wu, Lin and Bogy, D. B. 2003. New first and second order slip models for the compressible Reynolds equation. *J. Tribology*, **125**, 558-561.
60. Fukui, S. and Kaneko, R. 1988. Analysis of ultra-thin gas film lubrication based on linearized Boltzmann equation. *J. Tribology*, **110**, 253-261.
61. Kogan, M. N. 1969. *Rarefied Gas Dynamics*, Plenum Press, New York.
62. Gross, F. P., Jackson, E. A. and Ziering, S. 1957. Boundary value problems in kinetic theory of gases. *Ann. of Physics*, **1**, 141-167.
63. Cercignani, C. and Daneri, A. 1963. Flow of a rarefied gas between two parallel plates. *J. Appl. Phys.*, **34**, 3509-3513.
64. Cercignani, C. and Pagani, C. D. 1966. *Phys. Fluids*, **6**, 1167-1175.
65. Loyalka, S. K. 1971. Kinetic theory of thermal transpiration and mechanocaloric effect. *I. J. Chem. Phys.*, **55**, 4497-4503.
66. Fukui, S. and Kaneko, R. 1990. A database for interpolation of Poiseuille flow rates for high Knudsen number lubrication problems. *J. Tribology*, **112**, 78-83.
67. Israelachvili, J. N. 1986. Measurement of the viscosity of liquids in very thin films. *J. Coll. Interface Sci.*, **110**, 263-271.
68. Chan, D. Y. C. and Horn, R. G. 1985. The drainage of thin liquid films between solid surfaces. *J. Chem. Phys.*, **83**, 5311-5324.
69. Migun, N. P. and Prokhorenko, P. P. 1987. Measurement of the viscosity of polar liquids in micro-capillaries. *Colloid J. USSR*, 849-897.
70. Debye, P. and Cleland, R. L. 1959. Flow of liquid hydrocarbons in porous vycor. *J. Appl. Phys.*, **30**, 843-849.
71. Pfahler, J., Harley, J., Bau, H. and Zemel, J. N. 1991. Gas and liquid flow in small channels. *Symposium on Micromechanical Sensors, Actuators and Systems*, ed. D. Cho *et al.* ASME DSC, **32**, 49-60.
72. Klein, J. and Kumacheva, E. 1998. Liquid-to-solid transition in thin liquid films induced by confinement. *Physica A*, **249**, 206-215.

73. Hu, YZ. and Granick, S. 1998. Microscopic study of thin film lubrication and its contribution to macroscopic tribology. *Tribology Letters*, **5**, 81-88.
74. Jang, S. and Tichy, J. A. 1995. Rheological models for thin film EHL contacts. *J. Tribology*, **117**, 22-28.
75. Luo, J., Huang, P. and Li, L. K. Y. 1999. Characteristics of liquid lubricant films at the nano-scale. *ASME J. Trib.*, **121**, 872-878.
76. Matsuoka H. and Kato, T. 1997. An ultrathin liquid film lubrication theory – Calculation method of solvation pressure and its application to the EHL problem. *ASME J. Tribology*, **119**, 217-226.
77. Henderson, D. and Lozada-Cassou, M. 1986. A simple theory for the force between spheres immersed in liquid. *J. Colloid and Interface Sci.*, **114**, 180-183.
78. Gohar, R. 2001. *Elastohydrodynamic Lubrication*. 2nd edition, Imperial College Press.

Copyright

- **Figures 1, 2, 3, 4, 5:** reprinted with permission from Szeri, A. Z. and Snyder, V. Convective inertia effects in wall-bounded thin film flows. *Meccanica*, 2006, **41**, 473-482.
- **Figures 6, 7:** reprinted from Rajagopal, K. R., and Szeri, A. Z. On an inconsistency in the derivation of the equations of elastohydrodynamic lubrication. *Proc. Roy. Soc. Lond. A*, 2003, **459**, 2771-2786.
- **Figures 9, 10:** reprinted with permission from Fukui, S., and Kaneko, R. Analysis of ultra-thin gas film lubrication based on linearized Boltzman equation. *ASME J. Tribology*, 1988, **110**, 253-262.
- **Figure 11:** reprinted with permission from Mitsuya, Y. Modified Reynolds equation for ultra-thin film gas lubrication using 1.5 order slip flow model and considering surface accommodation coefficient. *ASME J. Tribology*, 1993, **115**, 289-294.
- **Figures 12, 13:** reprinted from *Physica A* **249**. Klein, J. and Kumacheva, E. Liquid-to-solid transition in thin liquid films induced by confinement. 206-215, copyright (1998), with permission from Elsevier.
- **Figure 14:** reprinted with permission from Matsuoka, H., and Kato, T. An ultrathin liquid film lubrication theory – calculation method of solvation pressure and its application to the EHL problem. *ASME J. Tribology*, 1997, **119**, 217-226.



Research Signpost
37/661 (2), Fort P.O.
Trivandrum-695 023
Kerala, India

Recent Developments in Wear Prevention, Friction and Lubrication, 2010: 45-136
ISBN: 978-81-308-0377-7 Editor: George K. Nikas

2. Rolling bearing life prediction, theory, and application

Erwin V. Zaretsky

Chief Engineer (Structures and Materials), NASA Glenn Research Center, USA

Abstract. A tutorial is presented outlining the evolution, theory, and application of rolling-element bearing life prediction from that of A. Palmgren, 1924, W. Weibull, 1939, G. Lundberg and A. Palmgren, 1947 and 1952, E. Ioannides and T. Harris, 1985, and E. Zaretsky, 1987. Comparisons are made between these life models. The Ioannides-Harris model without a fatigue limit is identical to the Lundberg-Palmgren model. The Weibull model is similar to that of Zaretsky if the exponents are chosen to be identical. Both the load-life and Hertz stress-life relations of Weibull, Lundberg and Palmgren, and Ioannides and Harris reflect a strong dependence on the Weibull slope. The Zaretsky model decouples the dependence of the critical shear stress-life relation from the Weibull slope. This results in a nominal variation of the Hertz stress-life exponent.

For 9th- and 8th-power Hertz stress-life exponents for ball and roller bearings, respectively, the Lundberg-Palmgren model best predicts life. However, for 12th- and 10th-power relations reflected by modern bearing steels, the Zaretsky model based on the Weibull equation is superior. Under the range of stresses examined, the use of a fatigue limit would suggest that (for most operating conditions under which a rolling-element bearing will operate) the bearing will not fail from classical rolling-element fatigue. Realistically, this is not the case. The use of a fatigue limit will significantly overpredict life over a range of normal operating Hertz stresses. (The use of ISO 281:2007 with a fatigue limit in these calculations would result in a bearing life approaching infinity.) Since the predicted lives of rolling-element bearings are high, the problem can become one of undersizing a bearing for a particular application.

Rules had been developed to distinguish and compare predicted lives to those actually obtained. Based upon field and test results of 51 ball and roller bearing sets, 98 percent of these bearing sets had acceptable life results using the Lundberg-Palmgren equations with life adjustment factors to predict bearing life. That is, they had lives equal to or greater than that predicted.

The Lundberg-Palmgren model was used to predict the life of a commercial turboprop gearbox. The life prediction was compared with the field lives of 64 gearboxes. From these results, the roller bearing lives exhibited a load-life exponent of 5.2, which correlated with

the Zaretsky model. The use of the ANSI/ABMA and ISO standards load-life exponent of $10/3$ to predict roller bearing life is not reflective of modern roller bearings and will underpredict bearing lives.

Introduction

By the close of the 19th century, the rolling-element bearing industry began to focus on sizing of ball and roller bearings for specific applications and determining bearing life and reliability. In 1896, R. Stribeck [1] in Germany began fatigue testing full-scale rolling-element bearings. J. Goodman [2] in 1912 in Great Britain published formulae based on fatigue data that would compute safe loads on ball and cylindrical roller bearings. In 1914, the “American Machinists Handbook” [3], devoted 6 pages to rolling-element bearings that discussed bearing sizes and dimensions, recommended (maximum) loading, and specified speeds. However, the publication did not address the issue of bearing life. During this time, it would appear that rolling-element bearing fatigue testing was the only way to determine or predict the minimum or average life of ball and roller bearings.

In 1924, A. Palmgren [4] in Sweden published a paper in German outlining his approach to bearing life prediction and an empirical formula based upon the concept of an L_{10} life, or the time that 90 percent of a bearing population would equal or exceed without rolling-element fatigue failure. During the next 20 years he empirically refined his approach to bearing life prediction and matched his predictions to test data [5]. However, his formula lacked a theoretical basis or an analytical proof.

In 1939, W. Weibull [6,7] in Sweden published his theory of failure. Weibull was a contemporary of Palmgren and shared the results of his work with him. In 1947, Palmgren in concert with G. Lundberg, also of Sweden, incorporated his previous work along with that of Weibull and what appears to be the work of H. Thomas and V. Hoersch [8] into a probabilistic analysis to calculate rolling-element (ball and roller) life. This has become known as the Lundberg-Palmgren theory [9,10]. (In 1930, H. Thomas and V. Hoersch [8] at the University of Illinois, Urbana, developed an analysis for determining subsurface principal stresses under Hertzian contact [11]. Lundberg and Palmgren [9,10] do not reference the work of Thomas and Hoersch [8] in their papers.)

The Lundberg-Palmgren life equations have been incorporated into both the International Organization for Standardization (ISO) and the American National Standards Institute (ANSI)/American Bearing Manufacturers Association (ABMA)¹ standards for the load ratings and life of rolling-element [12 to 14] as well as in current bearing codes to predict life.

¹ABMA changed their name from the Anti-Friction Bearing Manufacturers Association (AFBMA) in 1993.

In the post World War II era the major technology drivers for improving the life, reliability, and performance of rolling-element bearings have been the jet engine and the helicopter. By the late 1950s most of the materials used for bearings in the aerospace industry were introduced into use. By the early 1960s the life of most steels was increased over that experienced in the early 1940s primarily by the introduction of vacuum degassing and vacuum melting processes in the late 1950s [15].

The development of elastohydrodynamic (EHD) lubrication theory in 1939 by A. Ertel [16] and later A. Grubin [17] in 1949 in Russia showed that most rolling bearings and gears have a thin EHD film separating the contacting components. The life of these bearing and gears is a function of the thickness of the EHD film [15].

Computer programs modeling bearing dynamics that incorporate probabilistic life prediction methods and EHD theory enable optimization of rolling-element bearings based on life and reliability. With improved manufacturing and material processing, the potential improvement in bearing life can be as much as 80 times that attainable in the late 1950s or as much as 400 times that attainable in 1940 [15].

While there can be multifailure modes of rolling-element bearings, the failure mode limiting bearing life is contact (rolling-element) surface fatigue of one or more of the running tracks of the bearing components. Rolling-element fatigue is extremely variable but is statistically predictable depending on the material (steel) type, the processing, the manufacturing, and operating conditions [18].

Rolling-element fatigue life analysis is based on the initiation or first evidence of fatigue spalling on a loaded, contacting surface of a bearing. This spalling phenomenon is load cycle dependent. Generally, the spall begins in the region of maximum shear stresses, located below the contact surface, and propagates into a crack network. Failures other than that caused by classical rolling-element fatigue are considered avoidable if the component is designed, handled, and installed properly and is not overloaded [18]. However, under low EHD lubricant film conditions, rolling-element fatigue can be surface or near-surface initiated with the spall propagating into the region of maximum shearing stresses.

The database for ball and roller bearings is extensive. A concern that arises from these data and their analysis is the variation between life calculations and the actual endurance characteristics of these components. Experience has shown that endurance tests of groups of identical bearings under identical conditions can produce a variation in L_{10} life from group to group. If a number of apparently identical bearings are tested to fatigue at a specific load, there is a wide dispersion of life among these bearings. For a group of 30 or more bearings, the ratio of the

longest to the shortest life may be 20 or more [18]. This variation can exceed reasonable engineering expectations.

Bearing life theory

Foundation for bearing life prediction

Hertz contact stress theory

In 1917, Arvid Palmgren began his career at the A.–B. Svenska Kullager-Fabriken (SKF) bearing company in Sweden. In 1924 he published his paper [4] that laid the foundation for what later was to become known as the Lundberg-Palmgren theory [9]. Because the 1924 paper was missing two elements, it did not allow for a comprehensive rolling-element bearing life theory. The first missing element was the ability to calculate the subsurface principal stresses and hence, the shear stresses below the Hertzian contact of either a ball on a nonconforming race or a cylindrical roller on a race. The second missing element was a comprehensive life theory that would fit the observations of Palmgren. Palmgren discounted Hertz contact stress theory [11] and depended on the load-life relation for ball and roller bearings based on testing at SKF Sweden that began in 1910 [19]. Zaretsky discusses the 1924 Palmgren work in [20].

Palmgren did not have confidence in the ability of the Hertzian equations to accurately predict rolling bearing stresses. Palmgren [4] states, “The calculation of deformation and stresses upon contact between the curved surfaces...is based on a number of simplifying stipulations, which will not yield very accurate approximation values, for instance, when calculating the deformations. Moreover, recent investigations (circa 1919–1923) made at A.–B. Svenska Kullager-Fabriken (SKF) have proved through calculation and experiment that the Hertzian formulae will not yield a generally applicable procedure for calculating the material stresses....As a result of the paramount importance of this problem to ball bearing technology, comprehensive in-house studies were performed at SKF in order to find the law that describes the change in service life that is caused by changing load, rpm, bearing dimensions, and the like. There was only one possible approach: tests performed on complete ball bearings. It is not acceptable to perform theoretical calculations only, since the actual stresses that are encountered in a ball bearing cannot be determined by mathematical means.”

Palmgren later recanted his doubts about the validity of Hertz theory and incorporated the Hertz contact stress equation in his 1945 book [5]. In their 1947 paper [9], Lundberg and Palmgren state, “Hertz theory is valid under the

assumptions that the contact area is small compared to the dimensions of the bodies and that the frictional forces in the contact areas can be neglected. For ball bearings, with close conformity between rolling elements and raceways, these conditions are only approximately true. For line contact the limit of validity of the theory is exceeded whenever edge pressure occurs.”

Lundberg and Palmgren exhibited a great deal of insight as to the other variables modifying the resultant shear stresses calculated from Hertz theory. They state [9], “No one yet knows much about how the material reacts to the complicated and varying succession of (shear) stresses which then occur, nor is much known concerning the effect of residual hardening stresses or how the lubricant affects the stress distribution within the pressure area. Hertz theory also does not treat the influence of those static stresses which are set up by the expansion or compression of the rings when they are mounted with tight fits.” These effects are now understood, and life factors are currently being used to account for them so as to more accurately predict bearing life and reliability [18].

Equivalent Load

Palmgren [4] recognized that it was necessary to account for combined and variable loading around the circumference of a ball bearing. He proposed a procedure in 1924 “to establish functions for the service life of bearings under purely radial load and to establish rules for the conversion of axial and simultaneous effective axial and radial loads into purely radial loads.” Palmgren used Stribeck’s equation [1] to calculate what can best be described as a stress on the maximum radially loaded ball-race contact in a ball bearing. The equation attributed to Stribeck by Palmgren is as follows:

$$k = \frac{5Q}{Zd^2} \quad (1)$$

where Q is the total radial load on the bearing, Z is the number of balls in the bearing, d is the ball diameter, and k is Stribeck’s constant.

Palmgren modified Stribeck’s equation to include the effects of speed and load as well as modifying the ball diameter relation. For brevity, this modification is not presented. It is not clear whether Palmgren recognized at that time that Stribeck’s equation was valid only for a diametral clearance greater than zero with fewer than half of the balls being loaded. However, he stated that the corrected constant yielded good agreement with tests performed.

Palmgren [4] states, “It is probably impossible to find an accurate and, at the same time, simple expression for the ball pressure as a function of radial and axial pressure...” According to Palmgren, “Adequately precise results can be obtained by using the following equation:

$$Q = R + yA \quad (2)$$

where Q is the imagined, purely radial load that will yield the same service life as the simultaneously acting radial and axial forces, R is the actual radial load, and A is the actual axial load.” For ball bearings, Palmgren presented values of y as a function of Stribeck’s constant k . Palmgren stated that these values of y were confirmed by test results [4].

By 1945, Palmgren [5] modified Eq. (2) as follows:

$$Q = P_{eq} = XF_r + YF_a \quad (3)$$

where

- P_{eq} the equivalent load
- F_r the radial component of the actual load
- F_a the axial component of the actual load
- X a rotation factor
- Y the thrust factor of the bearing

The rotation factor X is an expression for the effect on the bearing capacity of the conditions of rotation. The thrust factor Y is a conversion value for thrust loads [5].

Fatigue Limit

Palmgren [4] states that bearing “limited service life is primarily a fatigue phenomenon. However, under exceptional high loads there will be additional factors such as permanent deformations, direct fractures, and the like....If we start out from the assumption that the material has a certain fatigue limit, meaning that it can withstand an unlimited number of cyclic loads on or below a certain, low level of load, the service life curve will be asymptotic. Since, moreover, the material has an elastic limit and/or fracture limit, the curve must yield a finite load even when there is only a single load value, meaning that the number of cycles equals zero. If we further assume that the curve has a profile of an exponential function, the

general equation for the relationship existing between load and number of load cycles prior to fatigue would read:

$$k = C(an + e)^{-x} + u \quad (4)$$

where k is the specific load or Stribeck's constant, C is the material constant, a is the number of load cycles during one revolution at the point with the maximum load exposure, n is the number of revolutions in millions, e is the material constant that is dependent on the value of the elasticity or fracture limit, u is the fatigue limit, and x is an exponent."

According to Palmgren, "This exponent x is always located close to 1/3 or 0.3. Its value will approach 1/3 when the fatigue limit is so high that it cannot be disregarded, and 0.3 when it is very low." Palmgren reported test results that support a value of $x = 1/3$. Hence, Eq. (4) can be written as

$$\text{Life (millions of stress cycles)} = \left(\frac{C}{k - u} \right)^3 - e \quad (5)$$

The value e suggests a finite time below which no failure would be expected to occur. By letting $e = 0$ and eliminating the concept of a fatigue limit for bearing steels, Eq. (5) can be rewritten as

$$L(\text{million of race revolutions}) = \left(\frac{CZd^2/5}{Q} \right)^3 \quad (6)$$

In Eq. (6), by letting $f_c = C/5$, and $P_{eq} = Q$, the 1924 version of the dynamic load capacity C_D for a radial ball bearing would be

$$C_D = f_c Z d^2 \quad (7)$$

and Eq. (6) becomes

$$L_{10} = \left(\frac{C_D}{P_{eq}} \right)^3 \quad (8)$$

where L_{10} is the life in millions of inner-race revolutions, at which 10 percent of a bearing population will have failed and 90 percent will have survived. This is also referred to as 10-percent life or L_{10} life.

By 1945, Palmgren [5] empirically modified the dynamic load capacity C_D for ball and roller bearings as follows:

For ball bearings

$$C_D = f_c \frac{id^2 Z^{2/3} \cos \beta}{1 + 0.02d} \quad (9)$$

For roller bearings

$$C_D = f_c id^2 l_t Z^{2/3} \cos \beta \quad (10)$$

where

- f_c material-geometry coefficient²
- i number of rows of rolling elements (balls or rollers)
- d ball or roller diameter
- l_t roller length
- Z number of rolling elements (balls or rollers) in a row i
- β bearing contact angle

From Anderson [21], for a constant bearing load, the normal force between a rolling element and a race will be inversely proportional to the number of rolling elements. Therefore, for a constant number of stress cycles at a point, the capacity is proportional to the number of rolling elements. Alternately, the number of stress cycles per revolution is also proportional to the number of rolling elements, so that for a constant rolling-element load the capacity for point contact is inversely proportional to the cube root of the number of rolling elements. This comes from the inverse cubic relation between load and life for point contact. Then the dynamic load capacity varies with number of balls as

$$C_D \sim \frac{Z}{Z^{1/3}} = Z^{2/3} \quad (11)$$

² Post 1990, the coefficient f_c is designated as f_{cm} in the ANSI/ABMA/ISO standards [12 to 14].

Equation (11) is reflected in the dynamic load capacity of Eqs. (9) and (10).

According to Palmgren [5], the coefficient f_c (in Eqs. (9) and (10)) is dependent, among other things, on the properties of the material, the degree of osculation (bearing race-ball conformity), and the reduction in capacity on account of uneven load distribution within multiple row bearings and bearings with long rollers. The magnitude of this coefficient can be determined only by numerous laboratory tests. It has one definite value for all sizes of a given bearing type.

In all of the above equations, the units of the input variables and the resultant units used by Palmgren have been omitted because they cannot be reasonably used or compared with engineering practice today. As a result, these equations should be considered only for their conceptual content and not for any quantitative calculations.

L_{10} life

The L_{10} life, or the time that 90 percent of a group of bearings will exceed without failing by rolling-element fatigue, is the basis for calculating bearing life and reliability today. Accepting this criterion means that the bearing user is willing in principle to accept that 10 percent of a bearing group will fail before this time. In Eq. (8) the life calculated is the L_{10} life.

The rationale for using the L_{10} life was first laid down by Palmgren in 1924. He states [4], “The (material) constant C (Eq. (4)) has been determined on the basis of a very great number of tests run under different types of loads. However, certain difficulties are involved in the determination of this constant as a result of service life demonstrated by the different configurations of the same bearing type under equal test conditions. Therefore, it is necessary to state whether an expression is desired for the minimum, (for the) maximum, or for an intermediate service life between these two extremes...In order to obtain a good, cost effective result, it is necessary to accept that a certain small number of bearings will have a shorter service life than the calculated lifetime, and therefore the constants must be calculated so that 90 percent of all the bearings have a service life longer than that stated in the formula. The calculation procedure must be considered entirely satisfactory from both an engineering and a business point of view, if we are to keep in mind that the mean service life is much longer than the calculated service life and that those bearings that have a shorter life actually only require repairs by replacement of the part which is damaged first.”

Palmgren is perhaps the first person to advocate a probabilistic approach to engineering design and reliability. Certainly, at that time, engineering practice dictated a deterministic approach to component design. This approach by Palmgren was decades ahead of its time. What he advocated is designing for finite life and

reliability at an acceptable risk. This concept was incorporated in the ANSI/ABMA and ISO standards [12 to 14].

Linear damage rule

Most bearings are operated under combinations of variable loading and speed. Palmgren recognized that the variation in both load and speed must be accounted for in order to predict bearing life. Palmgren reasoned: “In order to obtain a value for a calculation, the assumption might be conceivable that (for) a bearing which has a life of n million revolutions under constant load at a certain rpm (speed), a portion M/n of its durability will have been consumed. If the bearing is exposed to a certain load for a run of M_1 million revolutions where it has a life of n_1 million revolutions, and to a different load for a run of M_2 million revolutions where it will reach a life of n_2 million revolutions, and so on, we will obtain

$$\frac{M_1}{n_1} + \frac{M_2}{n_2} + \frac{M_3}{n_3} + \dots = 1 \quad (12)$$

In the event of a cyclic variable load we obtain a convenient formula by introducing the number of intervals p and designate m as the revolutions in millions that are covered within a single interval. In that case we have

$$p \left(\frac{m_1}{n_1} + \frac{m_2}{n_2} + \frac{m_3}{n_3} + \dots \right) = 1 \quad (13)$$

where n still designates the total life in millions of revolutions under the load and rpm (speed) in question (and M in Eq. (12) equal pm .”

Equations (12) and (13) were independently proposed for conventional fatigue analysis by B. Langer [22] in 1937 and M. Miner [23] in 1945, 13 and 21 years after Palmgren, respectively. The equation has been subsequently referred to as the linear damage rule or the Palmgren-Langer-Miner rule. For convenience, the equation can be written as follows:

$$\frac{1}{L} = \frac{X_1}{L_1} + \frac{X_2}{L_2} + \frac{X_3}{L_3} + \dots + \frac{X_n}{L_n} \quad (14)$$

and

$$X_1 + X_2 + X_3 + \dots X_n = 1 \quad (15)$$

where L is the total life in stress cycles or race revolutions, $L_1 \dots L_n$ is the life at a particular load and speed in stress cycles or race revolutions, and $X_1 \dots X_n$ is the fraction of total running time at load and speed. The values of M_1, M_2 , etc. in Eq. (12) equal $X_1 L, X_2 L$, etc. from Eq. (14). Equation (14) is the basis for most variable-load fatigue analysis and is used extensively in bearing life prediction.

Weibull analysis

Weibull distribution function

In 1939, W. Weibull [6,7] developed a method and an equation for statistically evaluating the fracture strength of materials based upon small population sizes. This method can be and has been applied to analyze, determine, and predict the cumulative statistical distribution of fatigue failure or any other phenomenon or physical characteristic that manifests a statistical distribution. The dispersion in life for a group of homogeneous test specimens can be expressed by

$$\ln \ln \frac{1}{S} = e \ln \left(\frac{L - L_\mu}{L_\beta - L_\mu} \right) \quad \text{where } 0 < L < \infty; 0 < S < 1 \quad (16)$$

where S is the probability of survival as a fraction ($0 \leq S \leq 1$); e is the slope of the Weibull plot; L is the life cycle (stress cycles); L_μ is the location parameter, or the time (cycles) below which no failure occurs; and L_β is the characteristic life (stress cycles). The characteristic life is that time at which 63.2 percent of a population will fail, or 36.8 percent will survive.

The format of Eq. (16) is referred to as a three-parameter Weibull analysis. For most—if not all—failure phenomenon, there is a finite time period under operating conditions when no failure will occur. In other words, there is zero probability of failure, or a 100-percent probability of survival, for a period of time during which the probability density function is nonnegative. This value is represented by the location parameter L_μ . Without a significantly large data base, this value is difficult to determine with reasonable engineering or statistical certainty. As a result, L_μ is usually assumed to be zero and Eq. (16) can be written as

$$\ln \ln \frac{1}{S} = e \ln \left(\frac{L}{L_\beta} \right) \text{ where } 0 < L < \infty; 0 < S < 1 \quad (17)$$

This format is referred to as the two-parameter Weibull distribution function. The estimated values of the Weibull slope e and L_β for the two-parameter Weibull analysis may not be equal to those of the three-parameter analysis. As a result, for a given survivability value S , the corresponding value of life L will be similar but not necessarily the same in each analysis.

By plotting the ordinate scale as $\ln \ln(1/S)$ and the abscissa scale as $\ln L$, a Weibull cumulative distribution will plot as a straight line, which is called a “Weibull plot.” Usually, the ordinate is graduated in statistical percent of specimens failed F where $F = [(1 - S) \times 100]$. Figure 1(a) is a generic Weibull plot with some of the values of interest indicated. Figure 1(b) is a Weibull plot of actual bearing fatigue data. The derivation of the Weibull distribution function can be found in Appendix A.

The Weibull plot can be used to evaluate any phenomenon that results in a statistical distribution. The tangent of the resulting plot, called the “Weibull slope” (also called the “Weibull shape parameter” or “Weibull modulus”) and designated by e , defines the statistical distribution. Weibull slopes of 1, 2, and 3.57 represent exponential, Rayleigh, and Gaussian (normal) distributions, respectively.

The scatter in the data is inversely proportional to the Weibull slope; that is, the lower the value of the Weibull slope, the larger the scatter in the data, and vice versa. The Weibull slope is also liable to statistical variation depending on the sample size (data base) making up the distribution [24]. The smaller the sample size, the greater the statistical variation in the slope.

A true fit of a two-parameter Weibull distribution function (Fig. 1) would imply a zero minimum life of $L_\mu = 0$ in Eq. (16). Tallian [25] analyzed a composite sample of 2500 rolling-element bearings and concluded that a good fit was obtained in the failure probability region between 10 and 60 percent. Outside this region, experimental life is longer than that obtained from the two-parameter Weibull plot prediction. In the early failure region, bearings were found to behave as shown in Fig. 2. From the Tallian data, it was found that the location parameter for the three-parameter Weibull distribution of Eq. (16) is $0.053 L_{10}$, where L_{10} is that value obtained from the two-parameter Weibull plot (Eq. (17) and Fig. 1) [15].

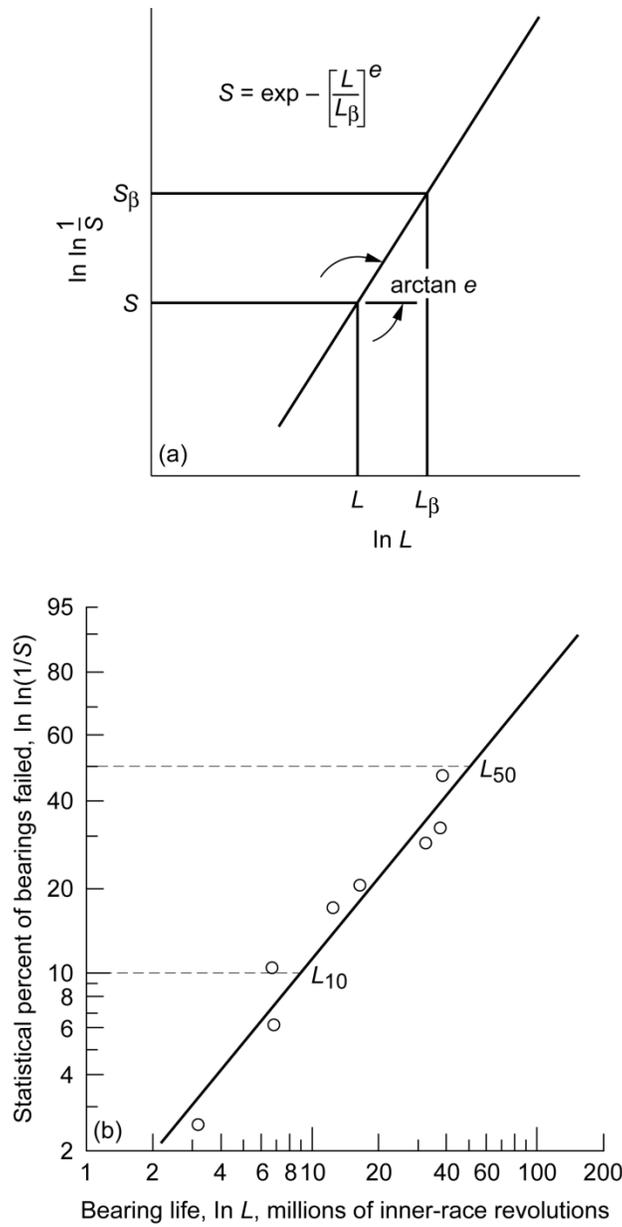


Figure 1. Weibull plot where (Weibull) slope or tangent of line is e ; probability of survival, S_β , is 36.8 percent at which $L = L_\beta$, or $L = L_\beta = 1$. (a) Schematic. (b) Rolling-element bearing fatigue data.

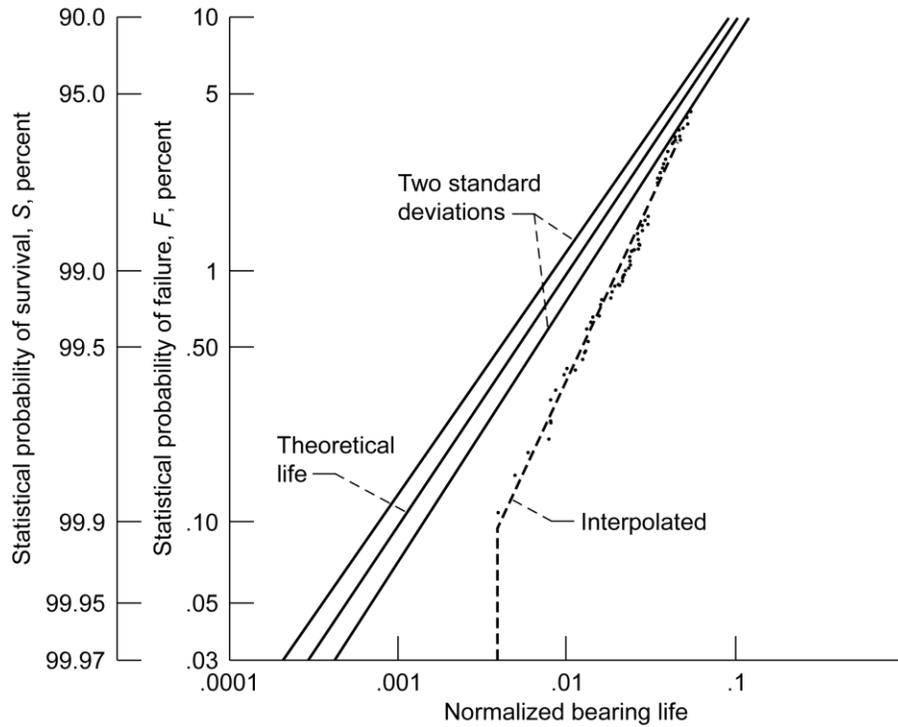


Figure 2. Two-parameter Weibull plot of bearing life distribution in early failure region [25].

Weibull fracture strength model

Weibull [6,7,26,27] related the material strength to the volume of the material subjected to stress. If the solid were to be divided in an arbitrary manner into n volume elements, the probability of survival for the entire solid can be obtained by multiplying the individual survivabilities together as follows

$$S = S_1 \cdot S_2 \cdot S_3 \cdots S_n \quad (18)$$

where the probability of failure F is

$$F = 1 - S \quad (19)$$

Weibull further related the probability of survival S , the material strength σ , and the stressed volume V according to the following relation

$$\ln \frac{1}{S} = \int_V f(X) dV \quad (20)$$

where

$$f(X) = \sigma^e \quad (21)$$

For a given probability of survival S ,

$$\sigma \sim \left[\frac{1}{V} \right]^{1/e} \quad (22)$$

From Eq. (22), for the same probability of survival the components with the larger stressed volume will have lower strength (or shorter life).

Bearing life models

Weibull fatigue life model

In conversations E.V. Zaretsky had with W. Weibull on January 22, 1964, Weibull related that he suggested to his contemporaries A. Palmgren and G. Lundberg in Gothenberg, Sweden (circa 1944), to use his equation (Eq. (20)) to predict bearing (fatigue) life where

$$f(X) = \tau^c \eta^e \quad (23)$$

and where τ is the critical shear stress and η is the number of stress cycles to failure.

In the past E.V. Zaretsky has credited this relation to Weibull. However, there appears to be no documentation of the above nor any publication of the application of Eq. (23) by Weibull in the open literature. However, in [28] Poplawski *et al.* applied Eq. (23) to Eq. (20) where

$$\eta \sim \left[\frac{1}{\tau} \right]^{c/e} \left[\frac{1}{V} \right]^{1/e} \quad (24)$$

The parameter c/e is the stress-life exponent. This implies that the inverse relation of life with stress is a function of the life scatter (Weibull slope) or data dispersion.

Referring to Figs. 3 and 4 for point contact and line contact, respectively, the stressed volume [9] is defined as

Point contact: $V = al_Lz$ (25a)

Line contact: $V = l_t l_L z$ (25b)

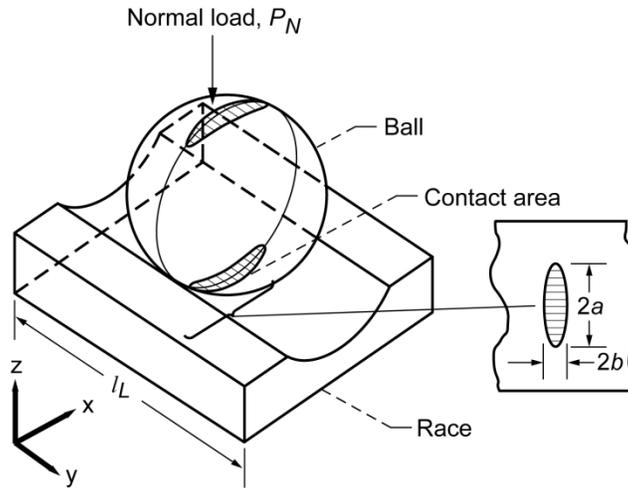


Figure 3. Ball-race model for point contact.

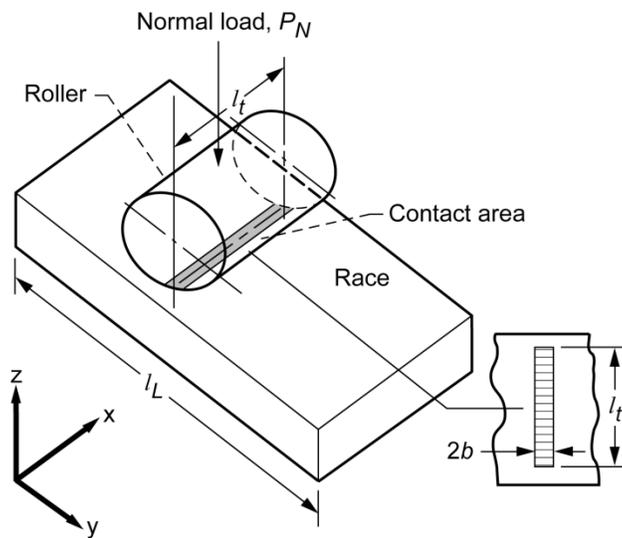


Figure 4. Roller-race model for line contact.

The depth z to the critical shear stress τ below the Hertzian contact in the running track is shown in Fig. 5. The length of the running track is l_L , and l_t is the roller width.

The critical shearing stress can be any one or a combination of the maximum shearing stress, τ_{\max} , the maximum orthogonal shearing stress, τ_o , the octahedral shearing stress τ_{oct} , or the von Mises shearing stress τ_{VM} . The von Mises shearing stress is a variation of the octahedral shearing stress.

From Hertz theory [11,29] for point contact (Fig. 3), V and τ can be expressed as a function of the maximum Hertz (contact) stress, S_{\max} [29], where

$$\tau \sim S_{\max} \quad (26a)$$

$$V \sim S_{\max}^2 \quad (26b)$$

Substituting Eqs. (26a) and (26b) in Eq. (24) and L for η ,

$$L \sim \left(\frac{1}{S_{\max}} \right)^{c/e} \left[\frac{1}{S_{\max}^2} \right]^{1/e} \sim \frac{1}{S_{\max}^n} \quad (27)$$

From [28], solving for the value of the exponent n for point contact (ball in a raceway) from Eq. (27) gives

$$n = \frac{c+2}{e} \quad (28)$$

From Hertz theory for line contact (roller in a raceway, Fig. 4),

$$V \sim S_{\max} \quad (29)$$

Substituting Eqs. (26a) and (29) in Eq. (24) and L for η ,

$$L \sim \left[\frac{1}{S_{\max}} \right]^{c/e} \left[\frac{1}{S_{\max}} \right]^{1/e} \sim \frac{1}{S_{\max}^n} \quad (30)$$

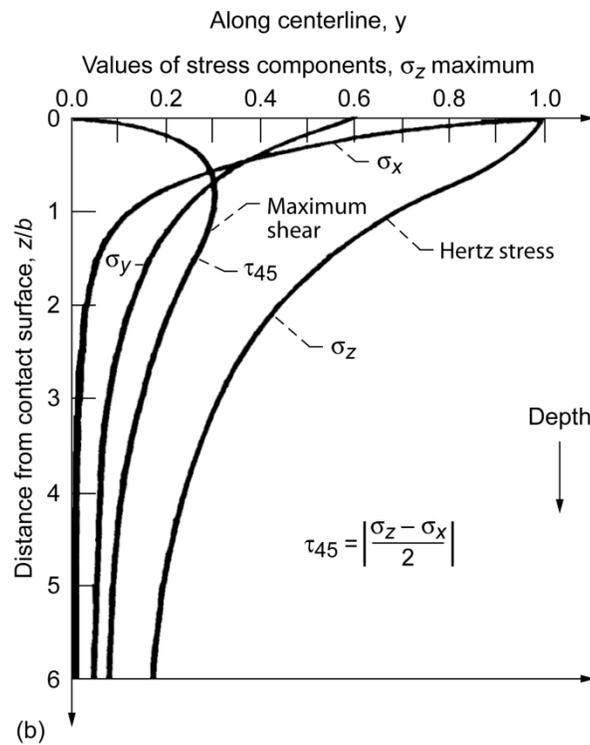
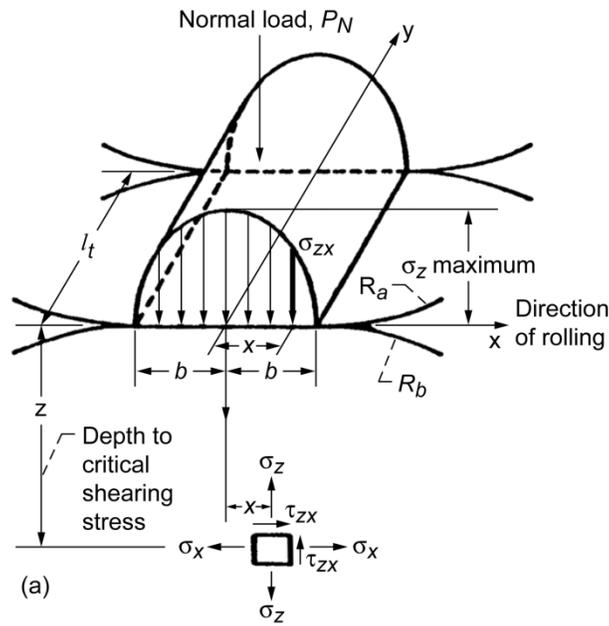


Figure 5. Subsurface stress field under line contact. (a) Hertz stress distribution for roller on raceway showing principal stresses at distance z below surface. (b) Distribution of principal and shearing stress as a function of depth z below surface.

Solving for the value of n for line contact by substituting Eqs. (25a) and (28) into Eq. (26) gives

$$n = \frac{c+1}{e} \quad (31)$$

From Lundberg and Palmgren [9] for point contact, $c = 10.33$ and $e = 1.11$. Then from Eq. (28),

$$n = \frac{c+2}{e} = \frac{10.33+2}{1.11} = 11.12 \quad (32)$$

From Hertz theory [29] for point contact,

$$S_{\max} \sim P^{1/3} \quad (33)$$

From Eq. (27) for point contact,

$$L \sim \frac{1}{S_{\max}^n} \sim \frac{1}{P_N^p} \quad (34a)$$

Combining Eqs. (33) and (34a) for point contact, and solving for p ,

$$p = \frac{n}{3} = \frac{c+2}{3e} \quad (34b)$$

From Eq. (32) where $n = 11.12$,

$$p = \frac{11.12}{3} = 3.7 \quad (34c)$$

For line contact from Eq. (31),

$$n = \frac{c+1}{e} = \frac{10.33+1}{1.11} = 10.21 \quad (35)$$

From Eq. (30) for line contact,

$$L \sim \frac{1}{S_{\max}^n} \sim \frac{1}{P_N^p} \quad (36a)$$

From Hertz theory [29] for line contact,

$$S_{\max} \sim P^{1/2} \quad (36b)$$

Combining Eqs. (36a) and (36b) and solving for p for line contact,

$$p = \frac{n}{2} = \frac{c+1}{2e} = \frac{10.21}{2} = 5.1 \quad (36c)$$

In their 1952 publication [10], Lundberg and Palmgren assume e for line contact equals 1.125, then from Eq. (35) $n = 10.1$, and from Eq. (36c) $p = 5$. From Weibull, the values of the stress-life and the load-life exponents are dependent on the Weibull slope e , which for rolling-element bearings can and usually varies between 1 and 2. As a result, the values can be only valid for a single value of the Weibull slope. As an example, if in Eq. (32) for point contact, a Weibull slope e of 1.02 were selected, $n = 12$ and $p = 4$ from Eq. (34b). These values did not fit the bearing data base that existed in the 1940s.

Lundberg-Palmgren model

In 1947 Lundberg and Palmgren [9] applied the Weibull analysis to the prediction of rolling-element bearing fatigue life. In order to account for the variation between the values of the Hertz stress-life exponent n and the load-life exponent p from those experimentally determined at the time, they introduced another variable, the depth to the critical shearing stress z to the h power where $f(x)$ in Eq. (20) can be expressed as

$$f(x) = \frac{\tau^c \eta^e}{z^h} \quad (37)$$

The rationale for introducing z^h was that it took a finite time period for a crack to initiate at a distance from the depth of the critical shearing to the rolling surface. Lundberg and Palmgren assumed that the time for crack propagation was a function of z^h .

Equation (24) thus becomes

$$\eta \sim \left[\frac{1}{\tau} \right]^{c/e} \left[\frac{1}{V} \right]^{1/e} [z]^{h/e} \quad (38)$$

For their critical shearing stress, Lundberg and Palmgren chose the orthogonal shearing stress. From Hertz theory [29],

$$z \sim S_{\max} \quad (39)$$

For point contact, substituting Eqs. (26a), (26b), and (39) in Eq. (38) and L for η ,

$$L \sim \left[\frac{1}{S_{\max}} \right]^{c/e} \left[\frac{1}{S_{\max}^2} \right]^{1/e} [S_{\max}]^{h/e} \sim \frac{1}{S_{\max}^n} \quad (40)$$

From [28], solving for the value of the exponent n for point contact (ball on a raceway) from Eq. (40) gives

$$n = \frac{c + 2 - h}{e} \quad (41a)$$

From Lundberg and Palmgren [9], using values of 1.11 for e , $c = 10.33$, and $h = 2.33$, from Eq. (41a) for point contact

$$n = \frac{10.33 + 2 - 2.33}{1.11} = 9 \quad (41b)$$

From Eq. (34b) for point contact, where $n = 9$,

$$p = \frac{n}{3} = \frac{9}{3} = 3 \quad (41c)$$

For line contact, substituting Eqs. (26a), (29), and (39) in Eq. (38) and L for η ,

$$L \sim \left[\frac{1}{S_{\max}} \right]^{c/e} \left[\frac{1}{S_{\max}} \right]^{1/e} \left[\frac{1}{S_{\max}} \right]^{h/e} \sim \frac{1}{S_{\max}^n} \quad (42)$$

From Eq. (42) solving for n for line contact,

$$n = \frac{c + 1 - h}{e} \quad (43a)$$

Using previous values of c and h , and $e = 1.125$ for line contact,

$$n = \frac{10.33 + 1 - 2.33}{1.125} = 8 \quad (43b)$$

From Eq. (36b) for line contact,

$$p = \frac{n}{2} = \frac{8}{2} = 4 \quad (43c)$$

These values of n and p for point and line contacts correlated to the then-existing rolling-element bearing database.

In their 1952 paper [10], Lundberg and Palmgren modified their value of the load-life exponent p for roller bearings from 4 to 10/3. The rationale for doing so was that various roller bearing types had one contact that is line contact and other that is point contact. They state “. . . as a rule the contacts between the rollers and the raceways transforms from a point to a line contact for some certain load so that the life exponent varies from 3 to 4 for differing loading intervals within the same bearing.” The ANSI/ABMA and ISO standards [12,14] incorporate $p = 10/3$ for roller bearings. Computer codes for rolling-element bearings incorporate $p = 4$.

Strict series reliability

Figures 6 and 7 show schematics of deep-groove and angular-contact ball bearings. Figure 8 is a schematic of a roller bearing. From Eqs. (20) and (30), the fatigue life L of a bearing inner or outer race determined from the Lundberg-Palmgren theory [9] can be expressed as follows:

$$L = A \left(\frac{1}{\tau} \right)^{c/e} \left(\frac{1}{V} \right)^{1/e} \left(\frac{1}{N} \right) [Z]^{h/e} \quad (44)$$

where N is the number of stress cycles per inner-race revolution and A is a material life factor based upon air-melt, pre-1940 AISI 52100 steel³ and mineral oil lubricant.

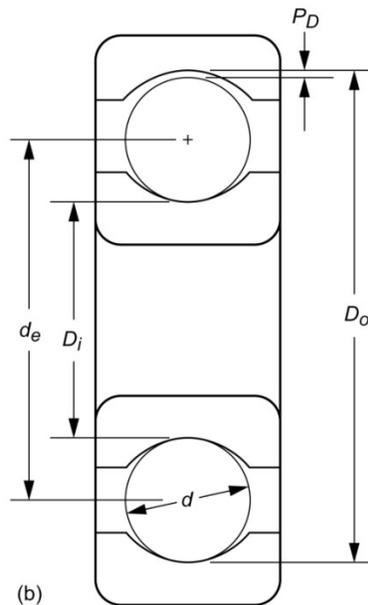
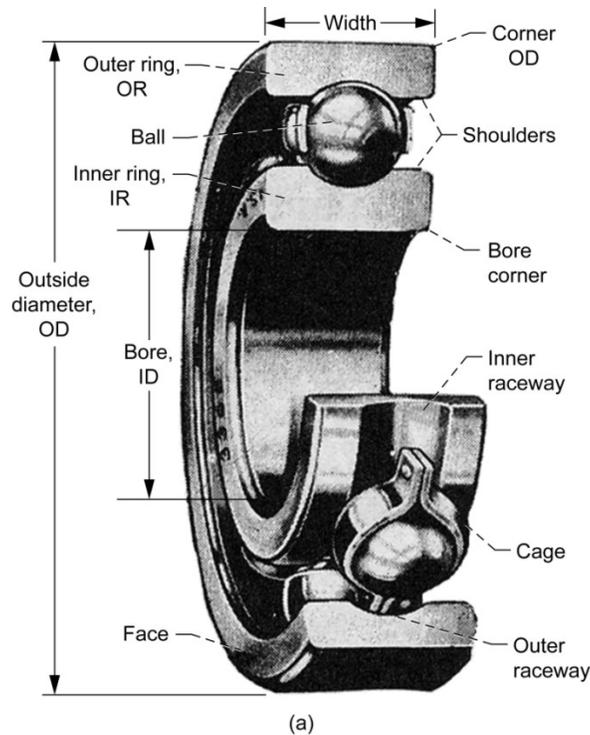


Figure 6. Deep-groove ball bearing. (a) Schematic. (b) Cross section without cage.

³ Numbered AISI steel grades are standardized by the American Iron and Steel Institute (AISI).

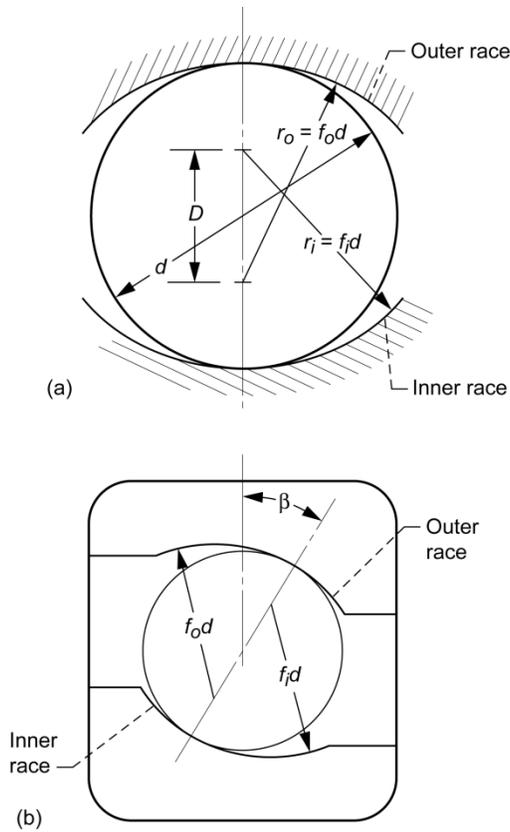


Figure 7. Ball-race conformity. (a) Deep-groove ball bearing. (b) Angular-contact ball bearing.

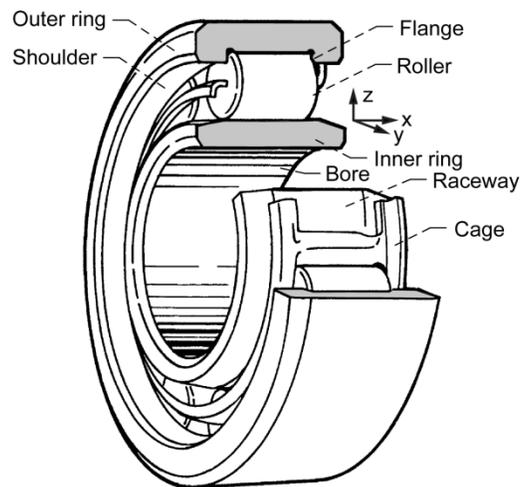


Figure 8. Schematic of cylindrical roller bearing with inner raceway. Bearing accommodates axial movement by not restraining rollers axially on inner raceway. Similar bearing with flanged inner ring allows axial roller movement on outer raceway.

In general, for ball and roller bearings, the running track lengths for Eqs. (25a) and (25b) for the inner and outer raceways are, respectively,

$$l_{L_{ir}} = \pi D_i = \pi(d_e - d \cos \beta) \quad (45a)$$

and

$$l_{L_{or}} = \pi D_o = \pi k(d_e + d \cos \beta) \quad (45b)$$

where d_e is the bearing pitch diameter (see Fig. 6).

In Eq. (45b), k is a correction factor that can account for variation of the stressed volume in the outer raceway. Equations (45a) and (45b) without the correction factor k are used in the Lundberg-Palmgren theory [9] to develop the capacity of a single contact on a raceway, assuming that all the ball-raceway loads are the same. In Eq. (45b), for an angular-contact bearing under thrust load only, $k = 1$.

Under radial load and no misalignment, the stressed volume V of a stationary outer race in a roller bearing or deep-groove ball bearing varies along the outer raceway in a load zone equal to or less than 180° . In the ANSI/ABMA and ISO standards [12,14] for radially loaded, rolling-element bearings, Eqs. (45a) and (45b) are adjusted for inner-race rotation and a fixed outer race with zero internal clearance, using system-life equations for multiple single contacts to calculate the bearing fatigue life. The outer raceway has a maximum load zone of 180° . An equivalent radial load P_{eq} was developed by Lundberg and Palmgren [9] and is used in the standards [12,14]. The equivalent load P_{eq} mimics a 180° ball-race load distribution assumed in the standards when pure axial loads are applied. It is also used throughout the referenced standards when combined axial and radial loads are applied in an angular-contact ball bearing.

Equations (45a) and (45b) are applicable for radially loaded roller bearing and deep-groove ball bearings where the rolling element-raceway contact diameters are at the pitch diameter plus or minus the roller/ball diameter, $\cos \beta = 1$, and $k < 1$. The maximum Hertz stress values are different at each ball or roller-race contact, at the inner and outer races, and vary along the arc in the zone of contact in a predictable manner. The width of the contact $2a$ for a ball bearing (Fig. 3) and the depth z for both ball and roller bearings (Fig. 5) to the critical shearing stress τ are functions of the maximum Hertz stress and are different at the inner and outer race contacts.

From Jones [29], for a ball bearing with a rotating inner race and a stationary outer race, the number of stress cycles N_{ir} and N_{or} for a single inner-race rotation for single points on the inner- and outer-races, respectively, are

$$N_{ir} = \frac{Z}{2} \left(1 + \frac{d}{d_e} \cos \beta \right) \quad (46a)$$

$$N_{or} = \frac{Z}{2} \left(1 - \frac{d}{d_e} \cos \beta \right) \quad (46b)$$

From Eqs. (12) and (17) from Weibull [6,7], Lundberg and Palmgren [9] first derived the relationship between individual component life and system life. A bearing is a system of multiple components, each with a different life. As a result, the life of the system is different from the life of an individual component in the system. The L_{10} bearing system life, where 90 percent of the population survives, can be expressed as

$$\frac{1}{L_{10}^e} = \frac{1}{L_{10ir}^e} + \frac{1}{L_{10or}^e} \quad (47)$$

where the life of the rolling elements, by inference, is incorporated into the life of each raceway tacitly assuming that all components have the same Weibull slope e where the L_{10} life of the bearing will be less than the L_{10} life of the lowest lived component in the bearing, which is usually that of the inner race. This is referred to as a “strict series reliability” equation and is derived in Appendix B. In properly designed and operated rolling-element bearings, fatigue of the cage or separator should not occur and, therefore, is not considered in determining bearing life and reliability. From Eqs. (17) and (44), Lundberg and Palmgren [9] derived the following relation:

$$L_{10} = \left(\frac{C_D}{P_{eq}} \right)^p \quad (48)$$

Equation (48) is identical to Eq. (8) proposed by Palmgren [4] in 1924 if p equal 3. From Lundberg-Palmgren [9], the load-life exponent p equals 3 for ball

bearings and 4 for roller bearings. However, as previously discussed, Lundberg and Palmgren in 1952 [10] proposed $p = 10/3$ for roller bearings.

Dynamic load capacity, C_D

Palmgren [4] proposed the concept of a dynamic load rating or capacity for a rolling-element bearing, defined as the load placed on a bearing that will theoretically result in a L_{10} life of 1 million inner-race revolutions. He first characterized this concept as that shown in Eq. (6) that subsequently evolved as Eqs. (9) and (10).

From Anderson [21], according to the Hertz theory, the dynamic load capacity should be proportional to the square of the rolling-element diameter. From experimental data, Palmgren [30] found that capacity varied as $d^{1.8}$ for balls up to about 25 mm in diameter and $d^{1.4}$ for balls larger than 25 mm in diameter.

From Eq. (11), the dynamic load capacity varies with the number of rolling elements Z to the $2/3$ power ($Z^{2/3}$). However, this would only be correct for an inverse cubic relation between load and life.

From Anderson [21], multiple-row bearings with i rows of balls may be considered as a combination of i single-row bearings [21]. From strict series reliability (Appendix B) the following relation between the life of a multirow bearing and the lives of the i individual rows is obtained assuming that all rows carry equal load:

$$\frac{1}{L^e} = \frac{1}{L_1^e} + \frac{1}{L_2^e} + \cdots + \frac{1}{L_i^e} \quad (49a)$$

Then

$$\frac{1}{L^e} = \frac{i}{L_i^e} \quad (49b)$$

If each row of the bearing is loaded with a load equal to the dynamic load capacity of one row C_i , then $L_i = 1$ (i.e., one million inner-race revolutions) and from Eq. (49b),

$$L^e = \frac{1}{i} \quad (50a)$$

or

$$L = \frac{1}{i^{1/e}} \quad (50b)$$

The load P_{eq} on the entire bearing is iC_i , where P_{eq} is the equivalent bearing load. In this case,

$$P_{eq} = iC_i \quad (51)$$

From Eqs. (50b) and (51),

$$\left(\frac{C_D}{iC_i} \right)^n = \frac{1}{i^{1/e}} \quad (52a)$$

or

$$C_D = C_i i^{1-(1/ep)} \quad (52b)$$

For ball bearings, $p = 3$ and e is approximately 1.1, so that the capacity of multirow bearings varies as $i^{0.7}$. For radial ball bearings, the normal force between a ball and a race varies as $1/\cos \beta$, so that the capacity is proportional to $\cos \beta$, where β is the contact angle (see Fig. 7). The influence of the ball-race conformity, bearing type, and internal dimensions expressed by $f_{cm}/(\cos \beta)^{0.3}$, where f_{cm} is the material and geometry coefficient. Therefore the capacity of a radial ball bearing varies as $(i \cos \beta)^{0.7}$.

For thrust ball bearings, the normal force between a ball and a race varies as $1/\sin \beta$, so that the capacity is proportional to $\sin \beta$ or to $(\cos \beta)(\tan \beta)$. When the influences of the degree of conformity, of bearing type, and of internal dimensions are included, the capacity of a thrust ball bearing varies as $(i \cos \beta)^{0.7}(\tan \beta)$.

For roller bearings with line contact, the load-life exponent in the life equation is 4, so that the capacity varies as $Z^{3/4}$. From Eq. (52b) with $p = 4$, the capacity of a multirow-roller bearing is found to vary as $i^{0.78}$. Theoretically, the capacity of roller bearings should be proportional to $l_i d$. Experimental data [9] indicate that capacity varies as $l_i^{0.78} d^{1.07}$.

Formulas for the dynamic load capacity CD as developed by Palmgren [30] and Lundberg and Palmgren [9, 10] are dependent on

- (1) Size of rolling elements, d (ball or roller diameter) and l_t (ball or roller length)
- (2) Size of rolling elements, d (ball or roller diameter) and l_t (ball or roller length)
- (3) Number of rolling elements per row, Z
- (4) Number of rows of rolling elements, i
- (5) Contact angle, β (see Fig. 7)
- (6) Material and geometry coefficient, f_{cm}

They are incorporated into the ANSI/ABMA and ISO standards [12 to 14], are semiempirical, and are as follows:

For radial ball bearings with $d \leq 25$ mm,

$$C_D = f_{cm} (i \cos \beta)^{0.7} Z^{2/3} d^{1.8} \quad (53a)$$

For radial ball bearings with $d > 25$ mm,

$$C_D = f_{cm} (i \cos \beta)^{0.7} Z^{2/3} d^{1.4} \quad (53b)$$

For radial roller bearings,

$$C_D = f_{cm} (i l_t \cos \beta)^{7/9} Z^{3/4} d^{29/27} \quad (53c)$$

For thrust ball bearings with $\beta \neq 90^\circ$,

$$C_D = f_{cm} (i \cos \beta)^{0.7} (\tan \beta) Z^{2/3} d^{1.8} \quad (53d)$$

For thrust roller bearings with $\beta \neq 90^\circ$,

$$C_D = f_{cm} (i l_t \cos \beta)^{7/9} (\tan \beta) Z^{3/4} d^{29/27} \quad (53e)$$

For thrust ball bearings with $\beta = 90^\circ$,

$$C_D = f_{cm} i^{0.7} Z^{2/3} d^{1.8} \quad (53f)$$

For thrust roller bearings with $\beta = 90^\circ$,

$$C_D = f_{cm} (i l_t)^{7/9} Z^{3/4} d^{29/27} \quad (53g)$$

The material and geometry coefficient f_{cm} (originally designated f_c by Lundberg and Palmgren [9]) in turn depends on the bearing type, material, and processing and the conformity between the rolling elements and the races. Representative values of f_{cm} are given in Table 1 from the ANSI/ABMA standards [13,14]. It should be noted that the coefficient f_{cm} and the various exponents of Eqs. (53a) through (53g) were chosen by Lundberg and Palmgren [9] and Palmgren [30] to match their bearing data base at the time of their writing. However, the values of f_{cm} have been updated periodically in the ANSI/ABMA and ISO standards [18,31].

Table 1. Representative values of rolling-element bearing geometry and material coefficient f_{cm} in ANSI/ABMA Standards 9 and 11 ([13], [14]) for representative rolling-element bearing sizes [18].

Bearing envelope size, $\frac{d \cos \beta}{d_e}$	Bearing geometry and material coefficient, f_{cm} ^a	
	Deep-groove and angular-contact ball bearings ^c	Cylindrical (radial) roller bearing
0.05	6070 (4610)	81.51 (7329)
.10	72.16 (5480)	92.62 (8322)
.16	77.56 (5890)	97.35 (8747)
.22	77.56 (5890)	97.02 (8767)
.28	74.27 (5640)	93.02 (8767)
.34	69.26 (5260)	-----
.40	62.94 (4780)	-----

^aValues of f_{cm} are for use with newtons and millimeters; those in parentheses are for use with pounds and inches.

^bPrior to 1990, f_{cm} was designated as f_c .

^cInner- and outer-race conformities are equal to 0.52.

Substituting the bearing geometry and the Hertzian contact stresses for a given normal load P_N into Eqs. (44) through (47), the dynamic load capacity C_D can be calculated from Eq. (48). Since P_N is the normal load on the maximum-loaded rolling element, it is required that the equivalent load P_{eq} be calculated. Once C_D is determined, f_{cm} can be calculated for the appropriate bearing type from Eq. (53).

The equivalent load P_{eq} can be obtained from Eq. (3) where values of X and Y for different bearing types are given in the ANSI/ABMA standards [13,14]. The dynamic load capacity C_D in the standards should be C_r (Eqs. (53a) to (53c)) for a radial bearing or C_a (Eqs. (53d) to (53g)) for a thrust bearing.

Lives determined using Eq. (53) are based on the “first evidence of fatigue.” This can be a small spall or surface pit that may not significantly impair the function of the bearing. The actual useful bearing life can be much longer. It should be also noted that in these Eqs. (53) where derived exponents differed from those obtained experimentally, those exponents obtained experimentally were substituted by Lundberg and Palmgren [9,10] for those that they analytically derived.

Ioannides-Harris model

Ioannides and Harris [32], using Weibull [6,7] and Lundberg and Palmgren [9,10] introduced a fatigue-limiting shear stress τ_u where from Eq. (37),

$$f(X) = \frac{(\tau - \tau_u)^c \eta^e}{z^h} \quad (54)$$

The equation is identical to that of Lundberg and Palmgren (Eq. 37) except for the introduction of a fatigue-limiting stress where

$$\eta \sim \left[\frac{1}{\tau - \tau_u} \right]^{c/e} \left[\frac{1}{V} \right]^{1/e} [z]^{h/e} \quad (55)$$

Equation (55) can be expressed as a function of S_{\max} where

$$L \sim \left(\frac{1}{\tau - \tau_u} \right)^{c/e} \left[\frac{1}{V} \right]^{1/e} [z]^{h/e} \sim \frac{1}{S_{\max}^n(\tau_u)} \quad (56)$$

Ioannides and Harris [32] use the same values of Lundberg and Palmgren for e , c , and h . If τ_u equals 0, then the values of the Hertz stress-life exponent n are identical to those of Lundberg and Palmgren (Eqs. (41b) and (43b)). However, for values of $\tau_u > 0$, n is also a function of $(\tau - \tau_u)$. For their critical shearing stress, Ioannides and Harris chose the von Mises stress.

From the above, Eq. (48) can be rewritten to include a “fatigue-limiting” load P_u :

$$L_{10} = \left(\frac{C_D}{P_{eq} - P_u} \right)^p \quad (57a)$$

where

$$P_u = f(\tau_u) \quad (57b)$$

When $P_{eq} \leq P_u$, bearing life is infinite and no failure would be expected. When $P_u = 0$, the life is the same as that for Lundberg and Palmgren.

The concept of a fatigue limit for rolling-element bearings was first proposed by Palmgren in 1924 (Eq. (5)) [4]. It was apparently abandoned by him first in 1945 [5] and then again with Lundberg in 1947 [9]. In 1985, Ioannides and Harris [32] applied Palmgren's concept of a fatigue limit to the Lundberg-Palmgren equations in the form shown in Eq. (54). The ostensible reason Ioannides and Harris used the fatigue limit was to replace the material and processing life factors [18] that are used as life modifiers in conjunction with the bearing lives calculated from the Lundberg-Palmgren equations.

There are two problems associated with the use of a fatigue limit for rolling-element bearing. The first problem is that the form of Eq. (55) may not reflect the presence of a fatigue limit but the presence of a compressive residual stress [18]. The second problem is that there are no data in the open literature that would justify the use of a fatigue limit for through-hardened bearing steels such as AISI 52100 and AISI M-50. In fact, a paper presented by Tosha *et al.* [33], reporting the results of rotating beam fatigue experiments for through-hardened AISI 52100 steel at very low stress levels, shows conclusively that a fatigue limit does not exist for this bearing steel.

Recent publications by the ASME [34] and the ISO [35,36] for calculating the life of rolling-element bearings include a fatigue limit and the effects of ball-race conformity on bearing fatigue life. These methods do not, however, include the effect of ball failure on bearing life. The ISO method is based on the work reported by Ioannides, Bergling, and Gabelli [37]. The ASME method as contained in their ASMELIFE software [34] uses the von Mises stress as the critical shearing stress with a fatigue limit value of 684 MPa (99,180 psi). This corresponds to a Hertz surface contact stress of 1140 MPa (165,300 psi). The ISO 281:2007 [36] method uses a fatigue limit stress of 900 MPa (130,500 psi), which corresponds to a Hertz contact stress of 1500 MPa (217,500 psi) [31].

The concepts of a fatigue limit load (bearing load under which the fatigue stress limit is just reached in the most heavily loaded raceway contact) introduced in the

new ISO rating methods [36] is proportional to the fatigue limit load raised to the 3rd power for ball bearings (point contact). These differing values of load would result in a 128-percent higher load below which no fatigue failure would be expected to occur [31] using ISO 281:2007 [36] than ASMELIFE [34].

The effect of using different values of fatigue limit or no fatigue limit on rolling-element fatigue life prediction is shown in Table 2. This table summarizes the qualitative results obtained for maximum Hertz stresses of 1379, 1724, and 2068 MPa (200, 250, and 300 ksi) for point contact using Eq. (38) for Lundberg-Palmgren without a fatigue limit and Eq. (55) for fatigue limits of 684 MPa (99,180 psi) (from ASMELIFE) and 900 MPa (130,500 psi) (from ISO 281:2007). The results are normalized to a maximum Hertz stress of 1379 MPa (200 ksi) with no fatigue limit where the quotient of Eq. (55) divided by Eq. (38) is taken to the c/e power of 9.3 (taken from Lundberg and Palmgren). The effect of stressed volume was also factored into these calculations [31]:

$$L_{IH} \approx L \left[\frac{\tau}{(\tau - \tau_u)} \right]^{c/e} \quad (58)$$

where L_{IH} is the life with the fatigue limit τ_u , L is the life without a fatigue limit τ_u and τ is the critical shearing stress.

Table 2. Effect of fatigue limit τ on rolling-element fatigue life [31].

Fatigue limit, ^a τ_u , MPa (ksi)	Relative life ^{b,c} (Eq. (58))		
	Maximum Hertz stress, MPa (ksi)		
	1379 (200)	1724 (250)	2068 (300)
0 (0), Lundberg-Palmgren [9]	1	0.134	0.026
684 (99.2), ASMELIFE [34]	11.9×10^6	3152	44.6
900 (130.5), ISO 281: 2007 [36]	∞	23.3×10^6	4258

^a The von Mises stress.

^b Includes effect of stressed volume.

^c Normalized to life at maximum Hertz stress of 1379 MPa (200 ksi) with no fatigue limit.

Zaretsky model

Both the Weibull and Lundberg-Palmgren models relate the critical shear stress-life exponent c to the Weibull slope e . The parameter c/e thus becomes, in

essence, the effective critical shear stress-life exponent, implying that the critical shear stress-life exponent depends on bearing life scatter or dispersion of the data. A search of the literature for a wide variety of materials and for nonrolling-element fatigue reveals that most stress-life exponents vary from 6 to 12. The exponent appears to be independent of scatter or dispersion in the data. Hence, Zaretsky [38] has rewritten the Weibull equation to reflect that observation by making the exponent c independent of the Weibull slope e , where

$$f(X) = \tau^{ce} \eta^e \quad (59)$$

From Eqs. (5) and (59),

$$\eta \sim \left[\frac{1}{\tau} \right]^c \left[\frac{1}{V} \right]^{1/e} \quad (60)$$

For critical shearing stress τ , Zaretsky chose the maximum shearing stress, τ_{45} .

Lundberg and Palmgren [9] assumed that once initiated, the time a crack takes to propagate to the surface and form a fatigue spall is a function of the depth to the critical shear stress z . Hence, by implication, bearing fatigue life is crack propagation time dependent. However, rolling-element fatigue life can be categorized as “high-cycle fatigue.” Crack propagation time is an extremely small fraction of the total life or running time of the bearing. The Lundberg-Palmgren relation implies that the opposite is true. To decouple the dependence of bearing life on crack propagation rate, Zaretsky [38,39] dispensed with the Lundberg-Palmgren relation of $L \sim z^{h/e}$ in Eq. (60). (It should be noted that at the time (1947) Lundberg and Palmgren published their theory [9], the concepts of “high-cycle” and “low-cycle” fatigue were only then beginning to be formulated.)

Equation (60) can be written as

$$L \sim \left[\frac{1}{\tau} \right]^c \left[\frac{1}{V} \right]^{1/e} \sim \frac{1}{S_{\max}^n} \quad (61)$$

From [28], solving for the value of the Hertz stress-life exponent n , for point contact from Eq. (61) gives

$$n = c + \frac{2}{e} \quad (62a)$$

and for line contact,

$$n = c + \frac{1}{e} \quad (62b)$$

If it is assumed that $c = 9$ and $e = 1.11$, $n = 10.8$ for point contact and $n = 9.9$ for line contact. If it is further assumed that $c = 10$ and $e = 1.0$, $n = 12$ for point contact and $n = 11$ for line contact.

What differentiates Eq. (61) and those of Weibull (Eq. 24), Lundberg and Palmgren (Eq. (38)) and Ioannides and Harris (Eq. (56)) is that the relation between shearing stress and life is independent of the Weibull slope, e , or the distribution of the failure data. However, in all four models, there is a dependency of the Hertz stress-life exponent, n , on the Weibull slope. The magnitude of the variation is least with the Zaretsky model.

Although Zaretsky [38,39] does not propose a fatigue-limiting stress, he does not exclude that concept either. However, his approach is entirely different from that of Ioannides and Harris [32]. For critical stresses less than the fatigue-limiting stress, the life for the elemental stressed volume is assumed to be infinite. Thus, the stressed volume of the component would be affected where $L \sim 1/V^{1/e}$. As an example, a reduction in stressed volume of 50 percent results in an increase in life by a factor of 1.9.

Ball and roller set life

Lundberg and Palmgren [9] do not directly calculate the life of the rolling-element (ball or roller) set of the bearing. However, through benchmarking of the equations with bearing life data by use of a material-geometry factor f_{cm} , the life of the rolling-element set is implicitly included in the life calculation of Eqs. (53a) to (53g).

The rationale for not including the rolling-element set in Eq. (47) appears in the 1945 edition of A. Palmgren's book [5] wherein he states that, "...the fatigue phenomenon which determines the life (of the bearing) usually develops on the raceway of one ring or the other. Thus, the rolling elements are not the weakest parts of the bearing ...". The database that Palmgren used to benchmark his and later the Lundberg-Palmgren equations were obtained under radially loaded conditions. Under these conditions, the life of the rolling elements as a system (set) will be equal to or greater than that of the outer race. As a result, failure of the rolling elements in determining bearing life was not initially considered by Palmgren. Had it been, Eq. (47) would have been written as follows:

$$\left(\frac{1}{L_{10}}\right)^e = \left(\frac{1}{L_{ir}}\right)^e + \left(\frac{1}{L_{re}}\right)^e + \left(\frac{1}{L_{or}}\right)^e \quad (63)$$

where the Weibull slope e is the same for each of the components as well as for the bearing as a system.

Comparing Eq. (63) with Eq. (47), the value of the L_{10} bearing life will be the same. However, the values of the L_{ir} and L_{or} between the two equations will not be the same, but the ratio of L_{or}/L_{ir} will remain unchanged.

The fraction of failures due to the failure of a bearing component is expressed by Johnson [24] as

$$\text{Fraction of inner-race failures} = \left[\frac{L_{10}}{L_{ir}}\right]^e \quad (64a)$$

$$\text{Fraction of rolling-element failures} = \left[\frac{L_{10}}{L_{re}}\right]^e \quad (64b)$$

$$\text{Fraction of outer-race failures} = \left[\frac{L_{sys}}{L_{or}}\right]^e \quad (64c)$$

From Eqs. (64a) to (64c), if the life of the bearing and the fractions of the total failures represented by the inner race, the outer race and the rolling element set are known, the life of each of these components can be calculated. Hence, by observation, it is possible to determine the life of each of the bearing components with respect to the life of the bearing.

Equations (64a) to (64c) were verified using radially loaded and thrust-loaded 50-mm-bore ball bearings. Three hundred and forty virtual bearing sets totaling 31,400 bearings were randomly assembled and tested by Monte Carlo (random) number generation [40]. From the Monte Carlo simulation, the percentage of each component failed was determined and compared to those predicted from Eqs. (64a) to (64c). These results are shown in Table 3. There is excellent agreement between these techniques [40].

Table 3. Comparison of bearing failure distributions based upon Weibull-based Monte Carlo method and those calculated from equations (64a) to (64c) for 50-mm-bore deep-groove and angular-contact ball bearings [40].

Ball bearing type	Component	Percent failure	
		Weibull-based Monte Carlo results	Results from Eqs. (64a) to (64c)
Deep groove	Inner race	70.1	69.9
	Rolling element	14.8	15.0
	Outer race	15.1	15.0
Angular contact	Inner race	45.4	45.1
	Rolling element	45.2	45.1
	Outer race	9.4	9.7

Figure 9 summarizes rolling-element fatigue life data for ABEC 7 204-size angular-contact ball bearings⁴ made from AISI 52100 steel [41]. The bearings had a free contact angle of 10°. Operating conditions were an inner-ring speed of 10,000 rpm, an outer-ring temperature of 79 °C (175 °F), and a thrust load of 1108 N (249 lb). The thrust load produced maximum Hertz stresses of 3172 MPa (460 ksi) on the inner race and 2613 MPa (379 ksi) on the outer race. From a Weibull analysis of the data, the bearing L_{10} life was 20.5 million inner-race revolutions, or approximately 34.2 hr of operation [41].

Seven of the twelve bearings failed from rolling-element fatigue. Two of the failed bearings had fatigue spalls on a ball and an inner race. Two bearings had inner-race fatigue spalls. Two bearings had fatigue spalls on a ball, and one bearing had an outer-race fatigue spall. Counting each component that failed as an individual failure independent of the bearing, there were four inner-race failures, four ball failures, and one outer-race failure for a total of nine failed components. Inner-race failures were responsible for 44.4 percent of the failures; ball failures, 44.4 percent; and outer-race failures, 11.2 percent. Using each of these percentages in Eqs. (64a) to (64c) together with the experimental L_{10} life, the lives of the inner and outer races and the ball set were calculated. For purposes of the calculation, the Weibull slope e was assumed to be 1.11, the same as Lundberg and Palmgren [9]. The resultant component L_{10} lives were 53 million inner-race revolutions (88.3 hr) for both the inner race and ball set and 183.3 million inner-race revolutions (305.5 hr) for the outer race.

⁴The ABEC scale is a system for rating the manufacturing tolerances of precision bearings developed by the Annular Bearing Engineering Committee (ABEC) of the ABMA.

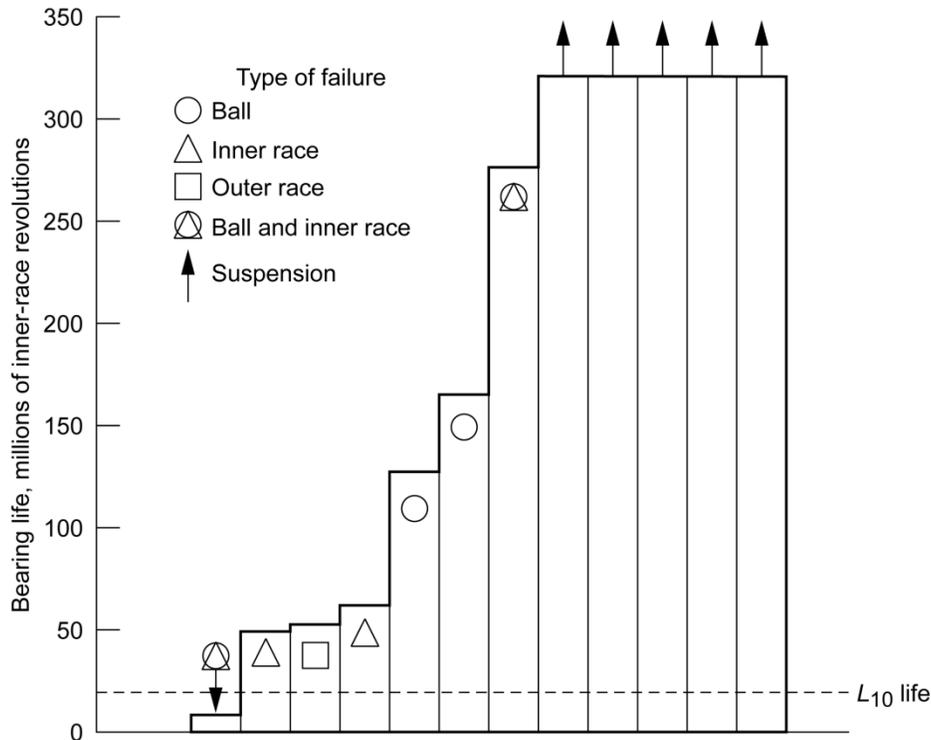


Figure 9. Rolling-element fatigue lives of AISI 52100 204-size angular-contact ball bearings. Contact angle is 10° ; outer-race temperature, 79°C (175°F); thrust load, 1108 N (249 lb); inner-ring speed, 10,000 rpm; lubricant, MIL-L-7808; L_{10} life, 20.5×10^6 inner-ring revolutions (34.2 hr); and failure index, 7 out of 12 [41].

For nearly all rolling-element bearings the number of inner-race failures is greater than those of the outer race. Accordingly, from Eqs. (64a) and (64c), the life of the outer race will be greater than that of the inner race. Zaretsky [18] noted that for radially loaded bearings (ball or roller), the percentage of failures of the rolling-element set was generally equal to and/or less than that of the outer race. For thrust-loaded ball or roller bearings, Zaretsky [18] further noted that the percent for the rolling-element set was equal to or less than that for the inner race but more than for the outer race. In order to account for material and processing variations, Zaretsky developed what is now referred to as Zaretsky's Rule [18]:

For radially loaded ball and roller bearings, the life of the rolling element set is equal to or greater than the life of the outer race. Let the life of the rolling element set (as a system) be equal to that of the outer race.

From Eq. (63)

$$\left(\frac{1}{L_{10}}\right)^e = \left(\frac{1}{L_{ir}}\right)^e + 2\left(\frac{1}{L_{or}}\right)^e \quad (65)$$

where $L_{re} = L_{or}$.

For thrust-loaded ball and roller bearings, the life of the rolling-element set is equal to or greater than the life of the inner race but less than that of the outer race. Let the life of the rolling-element set (as a system) be equal to that of the inner race.

From Eq. (63),

$$\left(\frac{1}{L_{10}}\right)^e = 2\left(\frac{1}{L_{ir}}\right)^e + \left(\frac{1}{L_{or}}\right)^e \quad (66)$$

where $L_{re} = L_{ir}$.

Examples of using Eqs. (65) and (66) are given in [18]. As previously stated, the resulting values for L_{ir} and L_{or} from these equations are not the same as those from Eq. (47). They will be higher.

H. Takata [42], using a modified approach to the Lundberg-Palmgren theory [9], derived the basic dynamic load capacity of the rolling-element bearing set in addition to those for the inner and outer races for radial and thrust-loaded ball and roller bearings. For radially loaded ball bearings, Takata assumes random ball rotation. For thrust-loaded ball bearings, he assumes a single or fixed running track on each ball. According to Takata, the basic dynamic load capacity C_D of a bearing system can be expressed as

$$C_D = \left(C_{ir}^{-w} + C_{re}^{-w} + C_{or}^{-w}\right)^{-1/w} \quad (67)$$

where C_D is calculated from Eqs. (53a) and (53b) (from Lundberg-Palmgren [9]), and the exponent w is equal to 10/3 for ball bearings. Takata [42] provides equations for calculating the dynamic load capacity of the rolling-element (ball) set, C_{re} . The resulting values for C_{ir} and C_{or} will be higher than those from the Lundberg-Palmgren equations [9].

Takata [42] performed a single ball-set life calculation within his paper for a 30-mm-bore deep-groove ball bearing. From this calculation he concluded that for this bearing

$$C_{ir} < C_{re} < C_{or} \quad (68a)$$

This would imply that

$$L_{ir} < L_{re} < L_{or} \quad (68b)$$

However, Takata [42] did not validate his example or his equations to determine ball- or roller-set life with a bearing life data base.

Ball-race conformity effects

ANSI/ABMA and ISO standards based on the Lundberg-Palmgren bearing life model [9] are normalized for ball bearings having inner- and outer-race conformities of 52 percent (0.52) and made from pre-1940 bearing steel. As discussed previously, the Lundberg-Palmgren model incorporates an inverse 9th-power relation between Hertz stress and fatigue life for ball bearings. Except for differences in applied loading, deep-groove and angular-contact ball bearings are treated identically. The effect of race conformity on ball set life independent of race life is not incorporated into the Lundberg-Palmgren model. An analysis by Zaretsky, Poplawski, and Root [31,43] considered the life of the ball set independently from race life, resulting in different life relations for deep-groove and angular-contact ball bearings. Both a 9th- and a 12th-power relation between Hertz stress and life were considered by them.

Rolling-element bearing computer models are capable of handling various race conformities in combination with Lundberg-Palmgren theory, but they universally do not include the influence of ball-set life on overall bearing life. Computer programs acknowledging the influence of ball-set life are typically used for more rigorous analysis of bearing systems but are not commonly used in the general bearing design community.

The conformities at the inner and outer races affect the resultant Hertz stresses and the lives of their respective raceways. The determination of life factors LF_i and LF_o based on the conformities at the inner and outer races, respectively, can be calculated by normalizing the equations for Hertz stress for the inner and outer races to a conformity of 0.52 (the value of 0.52 was chosen as a typical reference value). Stresses are evaluated for the same race diameter as a function of conformity. Based on Eq. (27), the ratio of the stress at a 0.52 conformity to the value at the same normal load P_N at another ball-race conformity, where $n = 9$ or 12 , gives the appropriate life factor

$$LF = \left(\frac{S_{\max 0.52}}{S_{\max}} \right)^n \quad (69a)$$

For the inner race,

$$LF_i = \left[\frac{\left(\frac{2}{d_e - d} + \frac{4}{d} - \frac{1}{0.52d} \right)^{2/3} (\mu\nu)_i}{\left(\frac{2}{d_e - d} + \frac{4}{d} - \frac{1}{f_i d} \right)^{2/3} (\mu\nu)_{0.52}} \right]^n \quad (69b)$$

and for the outer race,

$$LF_o = \left[\frac{\left(-\frac{2}{d_e + d} + \frac{4}{d} - \frac{1}{0.52d} \right)^{2/3} (\mu\nu)_o}{\left(-\frac{2}{d_e + d} + \frac{4}{d} - \frac{1}{f_o d} \right)^{2/3} (\mu\nu)_{0.52}} \right]^n \quad (69c)$$

The values of $(\mu\nu)_{0.52}$ are different for the inner and outer races.

For various ball bearing series (see Fig. 10), values of these life factors for conformities ranging from 0.505 to 0.570, subject to round-off error, are given in Table 4 for inner and outer races, for $n = 9$ and 12.

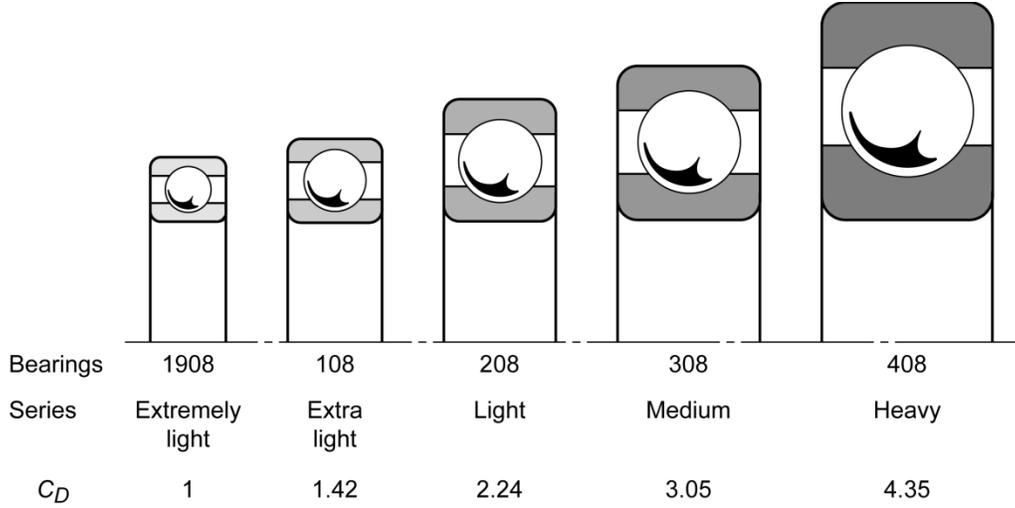


Figure 10. Effect of bearing series on relative sizes and dynamic capacities, C_D , of 40-mm bore deep-groove ball bearings [31].

From [31] and [43], the ratio of the outer- to the inner-race life can be approximated as

$$X = \frac{L_o}{L_i} \approx \left(\frac{S_{\max_i}}{S_{\max_o}} \right)^n \quad (70)$$

Referring to Figs. 7 and 8 and Hertzian contact theory [29] (see Appendix C),

$$X = \frac{L_o}{L_i} \approx \left[\frac{\left(\frac{2 \cos \beta}{d_e - d \cos \beta} + \frac{4}{d} - \frac{1}{f_i d} \right)^{2/3} (\mu\nu)_o}{\left(\frac{2 \cos \beta}{d_e + d \cos \beta} + \frac{4}{d} - \frac{1}{f_o d} \right)^{2/3} (\mu\nu)_i} \right]^n \quad (71)$$

Values of $\mu\nu$ for representative ball bearing series can be obtained from Table 4. Assuming $\cos \beta = 1$ and $f_i = f_o = 0.52$, values of L_o/L_i were calculated from Eq. (71) for $n = 9$ and 12 and are summarized in Table 5. It should be noted that in the development of these relationships it was assumed that the contact angle β does not change with speed and load [31,43].

Table 4. Effect of race conformity and Hertz stress-life exponent n on ball bearing life as function of ball bearing series^a [43].

Conformity, f	Ball bearing series					
	Extremely light, $\frac{d \cos\beta}{d_e} = 0.15$					
	Inner race			Outer race		
	$(\mu\nu)_i$	Life factor, LF_i		$(\mu\nu)_o$	Life factor, LF_o	
$n = 9$		$n = 12$	$n = 9$		$n = 12$	
0.505	2.013	14.7	36.05	1.826	11.84	27.00
0.510	1.776	4.53	7.51	1.673	3.61	5.53
0.515	1.641	2.12	2.73	1.551	1.71	2.05
0.520	1.517	1.00	1.00	1.471	1.00	1.00
0.525	1.503	0.88	0.84	1.415	0.66	0.57
0.530	1.452	0.62	0.53	1.369	0.46	0.36
0.535	1.409	0.45	0.35	1.335	0.35	0.25
0.540	1.376	0.35	0.25	1.304	0.27	0.17
0.545	1.361	0.30	0.20	1.282	0.22	0.13
0.550	1.328	0.23	0.14	1.262	0.18	0.10
0.555	1.306	0.19	0.11	1.244	0.15	0.08
0.560	1.296	0.17	0.10	1.227	0.13	0.06
0.565	1.278	0.15	0.08	1.211	0.11	0.05
0.570	1.262	0.13	0.06	1.196	0.09	0.04
	Extra light, $\frac{d \cos\beta}{d_e} = 0.18$					
0.505	2.048	12.58	29.27	1.887	11.88	27.10
0.510	1.784	3.47	5.25	1.662	3.54	5.40
0.515	1.654	1.68	1.99	1.541	1.68	2.00
0.520	1.570	1.00	1.00	1.465	1.00	1.00
0.525	1.505	0.66	0.57	1.407	0.65	0.57
0.530	1.458	0.47	0.37	1.364	0.47	0.36
0.535	1.441	0.41	0.30	1.345	0.39	0.28
0.540	1.398	0.30	0.20	1.303	0.28	0.18
0.545	1.366	0.23	0.14	1.276	0.22	0.13
0.550	1.336	0.18	0.10	1.254	0.17	0.10
0.555	1.307	0.15	0.08	1.234	0.14	0.08
0.560	1.301	0.13	0.07	1.212	0.12	0.06
0.565	1.283	0.11	0.06	1.204	0.10	0.05
0.570	1.267	0.10	0.05	1.192	0.09	0.04
	Light, $\frac{d \cos\beta}{d_e} = 0.23$					
0.505	2.062	12.32	28.38	1.870	11.83	26.94
0.510	1.843	4.28	6.95	1.652	3.61	5.54
0.515	1.658	1.58	1.84	1.549	1.89	2.34
0.520	1.583	1.00	1.00	1.454	1.00	1.00
0.525	1.520	0.67	0.58	1.397	0.66	0.57
0.530	1.471	0.48	0.37	1.354	0.47	0.36
0.535	1.431	0.36	0.25	1.333	0.38	0.28
0.540	1.401	0.28	0.19	1.292	0.27	0.18
0.545	1.374	0.22	0.14	1.267	0.21	0.13
0.550	1.347	0.19	0.11	1.235	0.16	0.09
0.555	1.324	0.15	0.08	1.227	0.14	0.08
0.560	1.312	0.14	0.07	1.211	0.12	0.06
0.565	1.298	0.12	0.06	1.198	0.10	0.05
0.570	1.280	0.10	0.05	1.180	0.09	0.04

Table 4. Continued

Conformity, f	Ball bearing series					
	Medium, $\frac{d \cos \beta}{d_e} = 0.25$					
	Inner race			Outer race		
	$(\mu\nu)_i$	Life factor, LF_i		$(\mu\nu)_o$	Life factor, LF_o	
$n = 9$		$n = 12$	$n = 9$		$n = 12$	
0.505	2.067	11.78	26.82	1.878	12.34	28.53
0.510	1.893	5.11	8.81	1.623	3.09	4.50
0.515	1.684	1.71	2.04	1.524	1.64	1.93
0.520	1.594	1.00	1.00	1.454	1.00	1.00
0.525	1.548	0.74	0.67	1.393	0.64	0.55
0.530	1.483	0.48	0.38	1.350	0.45	0.35
0.535	1.442	0.36	0.26	1.314	0.33	0.23
0.540	1.409	0.28	0.19	1.285	0.26	0.16
0.545	1.382	0.23	0.14	1.253	0.19	0.11
0.550	1.360	0.19	0.11	1.234	0.16	0.09
0.555	1.338	0.16	0.09	1.214	0.13	0.07
0.560	1.318	0.14	0.07	1.206	0.12	0.06
0.565	1.304	0.12	0.06	1.186	0.09	0.04
0.570	1.289	0.10	0.05	1.190	0.09	0.04
	Heavy, $\frac{d \cos \beta}{d_e} = 0.28$					
0.505	2.056	11.87	27.08	1.874	13.02	30.62
0.510	1.784	3.18	4.17	1.637	3.58	5.47
0.515	1.668	1.62	1.97	1.519	1.70	2.03
0.520	1.583	1.00	1.00	1.443	1.00	1.00
0.525	1.547	0.78	0.72	1.387	0.66	0.57
0.530	1.472	0.48	0.38	1.345	0.47	0.36
0.535	1.435	0.37	0.27	1.310	0.34	0.24
0.540	1.400	0.29	0.19	1.282	0.27	0.17
0.545	1.373	0.23	0.14	1.260	0.22	0.13
0.550	1.346	0.19	0.11	1.231	0.17	0.09
0.555	1.327	0.16	0.09	1.217	0.14	0.07
0.560	1.310	0.14	0.07	1.206	0.12	0.06
0.565	1.292	0.12	0.06	1.191	0.10	0.05
0.570	1.283	0.11	0.05	1.180	0.09	0.04

^aAll values of LF_i and LF_o are normalized to 1.00 for conformity f of 0.520.

Table 5. Representative ratios of outer- to inner-race life, L_o/L_i for representative ball bearing series as a function of Hertz stress-life exponent, n , at inner- and outer-race conformities of 0.52 (from Eq. (70)).

Ball bearing series	$\frac{d \cos \beta}{d_e}$	Outer- to inner-race life ratio, $X = L_o/L_i$	
		Hertz stress-life exponent, n	
		9	12
Extremely light	0.15	4.35	7.11
Extra light	0.18	4.39	7.18
Light	0.23	6.61	12.40
Medium	0.25	8.36	16.96
Heavy	0.28	12.04	27.60

Deep-groove ball bearings

Zaretsky's Rule for radially loaded deep-groove ball bearings (Eq. (65)), the life of the rolling element set is equal to or greater than the life of the outer race. Therefore, $L_{re} = L_o$ with $X = L_o/L_i$, Eq. (65) can be rewritten as

$$L_{10} = \left(\frac{X^e L_i^e}{X^e + 2} \right)^{1/e} \quad (72)$$

Applying life factors based on the effect of conformity to the respective lives of the inner and outer races, Eq. (72) becomes

$$L_{10_m} = \frac{(LF_i)(LF_o)XL_i}{\left[(LF_o)^e X^e + 2(LF_i)^e \right]^{1/e}} \quad (73)$$

Dividing Eq. (73) by (72) provides the bearing life factor LF_c for the radially loaded deep-groove ball bearing based on conformity:

$$LF_c = \left[\frac{(LF_i)^e (LF_o)^e (X + 2)}{(LF_o)^e X^e + 2(LF_i)^e} \right]^{1/e} \quad (74)$$

Angular-contact ball bearings

From Zaretsky's Rule for thrust-loaded ball bearings (Eq. (66)), the life of the rolling element set is equal to or greater than the life of the inner race but less than that of the outer race with $L_{re} = L_i$ and $X = L_o/L_i$. Equation (66) can be written as follows:

$$L_{10} = \left(\frac{X^e L_i^e}{2X^e + 1} \right)^{1/e} \quad (75)$$

Applying life factors based on the effect of conformity on the respective lives of the inner and outer races, Eq. (75) becomes

$$L_{10_m} = \frac{(LF_i)(LF_o)XL_i}{\left[2(LF_o)^e X^e + (LF_i)^e\right]^{1/e}} \quad (76)$$

Dividing Eq. (76) by (75) provides the bearing life factor LF_c for the thrust-loaded, angular-contact ball bearing recognizing conformity:

$$LF_c = \left[\frac{(LF_i)^e (LF_o)^e (2X^e + 1)}{2(LF_o)^e X^e + (LF_i)^e} \right]^{1/e} \quad (77)$$

Representative life factors LF_c from Eqs. (22), (27), and (31) were determined with the Lundberg-Palmgren theory and Zaretsky's Rule based on four combinations of inner- and outer-race conformities and for three representative series. These results are summarized in Table 6, for the two Hertz stress-life exponents $n = 9$ and 12 [43].

From the above, the ANSI/ABMA and ISO life calculations can be modified based upon ball-race conformity as follows:

$$L_{10} = LF_c \left(\frac{C_D}{P_{eq}} \right)^p \quad (78)$$

The bearing fatigue lives in actual application will usually be equal to or greater than those calculated using the ANSI/ABMA and ISO standards that incorporate the Lundberg-Palmgren model. The relative fatigue life of an individual race is more sensitive to changes in race conformity for a Hertz stress-life exponent n of 12 than where $n = 9$. However, when the effects are combined to predict actual bearing life for a specified set of conditions and bearing geometry, the predicted life of the bearing will be greater for a value of $n = 12$ ($p = 4$) than $n = 9$ ($p = 3$) [43].

Table 6. Life factors based on combinations of ball-race conformities for Hertz stress-life exponent $n = 9$ and 12 normalized to inner- and outer-race conformities of 0.52 [43].

Ball bearing series	Ball-race conformity		Bearing life factor for conformity, ^a LF_c				
	Inner race, IR	Outer race, OR	Lundberg-Palmgren ^b	Deep-groove ball bearing		Angular-contact ball bearing	
				From Eq. (72)	Change from Lundberg-Palmgren, %	From Eq. (77)	Change from Lundberg-Palmgren, %
Hertz stress-life exponent $n = 9$ and $L_{10} = \left(\frac{C_D}{P_{eq}}\right)^3$							
Extremely light, $\frac{d \cos \beta}{d_e} = 0.15$	0.505	0.52	4.16	3.15	-24.30	7.40	-77.83
	0.57	0.52	0.15	0.15	-----	0.14	-6.59
	0.52	0.505	1.16	1.19	2.59	1.09	-6.23
	0.52	0.57	0.35	0.22	-37.11	0.45	29.83
Light, $\frac{d \cos \beta}{d_e} = 0.23$	0.505	0.52	5.10	3.05	-40.28	6.97	36.76
	0.57	0.52	0.11	0.10	-9.09	0.11	-----
	0.52	0.505	1.10	1.04	-5.68	7.05	-4.42
	0.52	0.57	0.49	0.27	-39.04	0.59	35.15
Heavy, $\frac{d \cos \beta}{d_e} = 0.28$	0.505	0.52	6.77	4.66	-32.24	8.51	25.73
	0.57	0.52	0.12	0.10	-18.53	0.09	-26.03
	0.52	0.505	1.05	0.89	-15.08	1.03	-2.22
	0.52	0.57	0.59	0.31	-46.87	0.73	24.15
Hertz stress-life exponent $n = 12$ and $L_{10} = \left(\frac{C_D}{P_{eq}}\right)^4$							
Extremely light, $\frac{d \cos \beta}{d_e} = 0.15$	0.505	0.52	$6.83 \left(\frac{C_D}{P_{eq}}\right)$	$3.65 \left(\frac{C_D}{P_{eq}}\right)$	-46.56	$10.79 \left(\frac{C_D}{P_{eq}}\right)$	57.98
	0.57	0.52	$0.07 \left(\frac{C_D}{P_{eq}}\right)$	$0.06 \left(\frac{C_D}{P_{eq}}\right)$	-14.29	$0.06 \left(\frac{C_D}{P_{eq}}\right)$	-14.29
	0.52	0.505	$1.10 \left(\frac{C_D}{P_{eq}}\right)$	$1.02 \left(\frac{C_D}{P_{eq}}\right)$	-7.27	$1.05 \left(\frac{C_D}{P_{eq}}\right)$	-4.55
	0.52	0.57	$0.26 \left(\frac{C_D}{P_{eq}}\right)$	$0.14 \left(\frac{C_D}{P_{eq}}\right)$	-46.15	$0.39 \left(\frac{C_D}{P_{eq}}\right)$	50.00
Light, $\frac{d \cos \beta}{d_e} = 0.23$	0.505	0.52	$9.67 \left(\frac{C_D}{P_{eq}}\right)$	$5.03 \left(\frac{C_D}{P_{eq}}\right)$	-47.98	$14.03 \left(\frac{C_D}{P_{eq}}\right)$	45.09
	0.57	0.52	$0.05 \left(\frac{C_D}{P_{eq}}\right)$	$0.04 \left(\frac{C_D}{P_{eq}}\right)$	-20.00	$0.05 \left(\frac{C_D}{P_{eq}}\right)$	-----
	0.52	0.505	$1.05 \left(\frac{C_D}{P_{eq}}\right)$	$0.89 \left(\frac{C_D}{P_{eq}}\right)$	-15.24	$1.03 \left(\frac{C_D}{P_{eq}}\right)$	-1.90
	0.52	0.57	$0.37 \left(\frac{C_D}{P_{eq}}\right)$	$0.20 \left(\frac{C_D}{P_{eq}}\right)$	-45.95	$0.53 \left(\frac{C_D}{P_{eq}}\right)$	43.24
Heavy, $\frac{d \cos \beta}{d_e} = 0.28$	0.505	0.52	$14.97 \left(\frac{C_D}{P_{eq}}\right)$	$7.81 \left(\frac{C_D}{P_{eq}}\right)$	-47.83	$19.13 \left(\frac{C_D}{P_{eq}}\right)$	27.79
	0.57	0.52	$0.05 \left(\frac{C_D}{P_{eq}}\right)$	$0.04 \left(\frac{C_D}{P_{eq}}\right)$	-20.00	$0.05 \left(\frac{C_D}{P_{eq}}\right)$	-----
	0.52	0.505	$1.02 \left(\frac{C_D}{P_{eq}}\right)$	$0.77 \left(\frac{C_D}{P_{eq}}\right)$	-24.51	$1.01 \left(\frac{C_D}{P_{eq}}\right)$	-0.01
	0.52	0.57	$0.57 \left(\frac{C_D}{P_{eq}}\right)$	$0.30 \left(\frac{C_D}{P_{eq}}\right)$	-47.37	$0.72 \left(\frac{C_D}{P_{eq}}\right)$	26.32

^aAll life factors are benchmarked to $L_{10} = \left(\frac{C_D}{P_{eq}}\right)^3$ and inner- and outer-race conformities of 0.52.

^bFor deep-groove and angular-contact ball bearings.

Stress effects

Hertz Stress-Life Relation

There is an issue of what the value of the Hertz stress-life relation and, hence, the value of the load-life exponent p should be for purposes of analysis. The generally accepted relation between load and life in a rolling-element bearing is that life varies with the inverse cubic power of load for ball bearings (point contact) and with the inverse 4th power of load for roller bearings (line contact). Work reported by Parker and Zaretsky [44] suggests that for air-melted steels the exponent n is approximately 9 for point contact. However, for the cleaner post-1960 vacuum-processed steels, $n = 12$ for point contact. A definitive data base does not exist for line contact (roller bearings).

For ball bearings, the Hertz stress-life relation where life is inversely proportional to the maximum Hertz stress to the 9th power and the cubic root of load, has been generally accepted by ball bearing manufacturers and users. There is at least one exception where a manufacturer had indicated based on its unpublished data base that the life of ball bearings varies inversely with the 4th power of load (or 12th power of Hertz stress). Nevertheless, the inverse cubic load-life relation has been included in the ANSI/ABMA standards for ball bearings [13].

Varying the Hertz stress-life exponent n can significantly affect life predictions for bearings. Using $n = 9$ results in a more conservative estimate of bearing life than using $n = 12$. Also, the ratio of the predicted lives at the two values of the Hertz stress-life exponent n is a function of load and is directly related to C_D/P_{eq} . Hence, in the normal operating load envelope, life may be underpredicted by a factor of 20 when using the ANSI/ABMA standards or a 9th power Hertz stress-life exponent n . A proper stress-life exponent then becomes more than of mere academic interest, since a design engineer requires a reliable analytic tool to predict bearing life and performance.

Fatigue tests of ball bearings by several investigators tended to verify this inverse cubic relation. Styri [45] presented data for two types of ball bearings. For one group of 6207-size, deep-groove ball bearings under various radial loads from 3.5 to 17.3 kN (775 to 3880 lbf), life was inversely proportional to load to the 3.3 power. Another group of 1207-size double-row self-aligning bearings was tested such that the maximum Hertz stress at the outer-race-ball contact varied from 4.0 to 5.6 GPa (580 to 810 ksi). Here it was found that life varied inversely with the 9th power of stress (or the 3rd power of load). Cordiano *et al.* [46] reported load-life data for 217-size, thrust-loaded ball bearings. The resultant Hertz stress-life exponents were 8.1, 9.6, and 12.6 for three lubricants, a water-glycol base, a

phosphate ester base, and a phosphate ester, respectively. McKelvey and Moyer [47] reported that, with four groups of AISI 4620, carburized-steel, crowned rollers (elliptical contact), fatigue life varied inversely with maximum Hertz stress to a range of the 8th to 9th power. Maximum Hertz stress in these tests varied from 1.8 to 3.3 GPa (262 to 478 ksi). Townsend *et al.* [48] surface fatigue tested three groups of case-carburized, consumable-electrode-vacuum-arc-remelted AISI 9310 8.89-cm- (3.5-in.-) pitch-diameter spur gears at maximum Hertz stresses of 1.5, 1.7, and 1.9 GPa (222, 248, and 272 ksi). The gears were run at 10,000 rpm and 77 °C (170 °F). The lubricant was superrefined naphthenic mineral oil with an additive package. The L_{10} life varied inversely with stress to the 8.4 power, but the L_{50} life varied inversely with stress to the 10.2 power. The average Hertz stress-life exponent n was 9.3.

Several other investigators [49 to 57] have reported data with bench-type, rolling-element fatigue testers, rather than with full-scale bearings or gears. Data are summarized in Table 7. The L_{10} lives as a function of maximum Hertz stress for these data are shown in Fig. 11. From the table the maximum Hertz stress for these data ranged from 3.7 to 9.0 GPa (526 to 1300 ksi). With the exception of Greenert's work [54] the stress-life exponents ranged from 8.4 to 12.4. The data are all for AISI 52100 and AISI M-50 steels (except for the data reported by Barwell and Scott [49], who do not state the type of steel). At least two sets of data were for air-melted steel, three were for vacuum-degassed steel, and one was for vacuum-arc-remelted steel. The other references do not state the melting process.

Table 7. Published bench-type rig rolling-element fatigue data related to stress-life effects [44].

Number from Fig. 11	Reference	Material		Lubricant	Load range, kg	Maximum Hertz stress range, GPa, (ksi)	Load-life exponent	Stress-life exponent	Test type
		Type	Melting process						
1	Barwell and Scott [49]	—	(a)	Mineral oil	200 to 600	^b 4.5 to 6.6 (^b 660 to 950)	2.8	8.4	Four ball
2	Butler and Carter [50]	AISI 52100	Air melt	SAE 10 mineral oil	-----	4.1 to 5.2 (600 to 750)	3.5	10.4	Spin rig
3	Butler and Carter [50]	MV-1 (AISI M-50)	Air melt	SAE 10 mineral oil	-----	4.1 to 5.2 (600 to 750)	3.2	9.7	Spin rig
4	Baughman [51]	MV-1 (AISI M-50)	(a)	MIL-L-7808	-----	4.4 to 5.4 (640 to 777)	3.2	9.7	Rolling-contact rig
5	Scott [52]	EN-31 (AISI 52100)	(a)	Diester	400 to 600	^b 5.7 to 6.6 (830 to 950)	3.6	10.8	Four ball
6	Utsumi and Okomoto [53]	AISI 52100	(a)	#60 Spindle oil	-----	3.7 to 5.1 (526 to 730)	^c 2.8	^c 8.5	Crowned disk
7	Greenert [54]	AISI 52100	(a)	Navy 2190 TEP lubricating oil	-----	5.8 to 6.1 (840 to 890)	^c 5 to 6.3	^c 15 to 19	Toroids
8	Valori <i>et al.</i> [55]	AISI 52100	Vacuum-arc remelted	Mineral oil	-----	4.2 to 5.5 (610 to 800)	^d 4.1	^d 12.4	Four ball
9	Schatzberg and Felsen [56]	AISI 52100	Vacuum degassed	Squalene	-----	6.4 to 9.0 (925 to 1300)	3.8	11.5	Four ball
10	Schatzberg and Felsen [56]	AISI 52100	Vacuum degassed	Squalene + 100 ppm H ₂ O	-----	6.4 to 9.0 (925 to 1300)	3.9	11.7	Four ball
11	Parker and Zaretsky [44]	AISI 52100	Vacuum degassed	Paraffinic mineral oil	-----	4.5 to 6.0 (6.50 to 8.57)	4	12	Five ball

^aMelting process not reported.

^bApproximate stress range, not reported by authors of reference.

^cEstimated, not reported by authors of reference.

^dReanalyzed by Zaretsky and Parker [57].

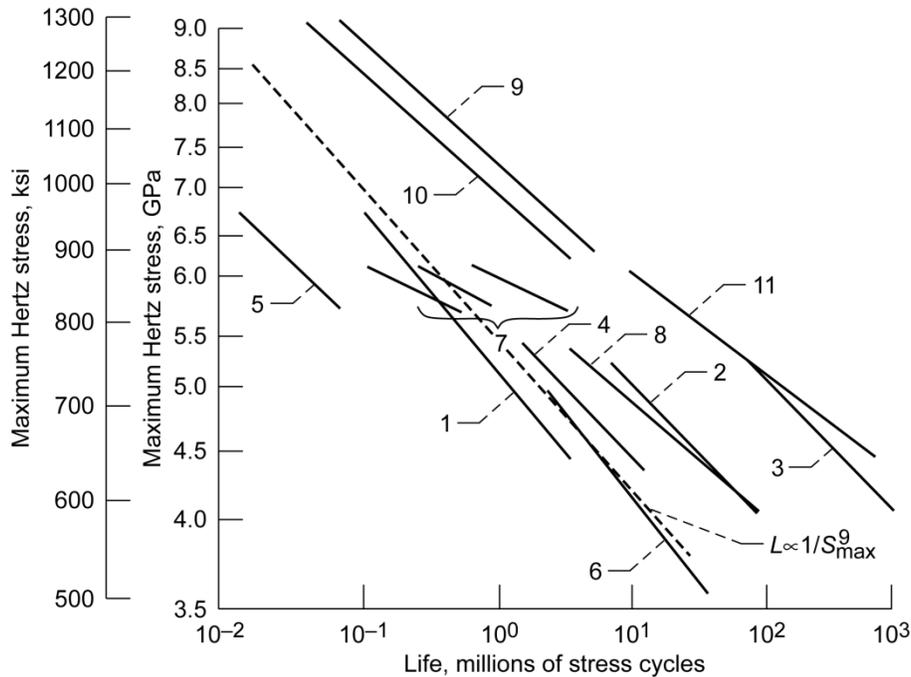


Figure 11. Summary of published stress-life relation data for Hertzian contacts failing from rolling-element fatigue. Refer to Table 7 [44].

The Hertz stress-life exponents from Greenert [54], ranging from 15 to 19, are much higher than those from other published data. This lack of correlation is unexplained. However, there is a probability that compressive residual stresses present in the steel could account for the resultant high values of the Hertz stress-life exponents.

Lorosch [58] fatigue tested three groups of vacuum-degassed 7205B-size AISI 52100 inner races at maximum Hertz stresses of 2.6, 2.8, and 3.5 GPa (370, 406, and 500 ksi), respectively. Each group consisted of 20 races, for a total of 60. (It is assumed that all 60 races came from a single heat of material and were of the same hardness, but Lorosch does not state so in his paper.) After the test runs Lorosch examined cross sections of the races. He observed that only at the lowest stress, 2.6 GPa (370 ksi), did no measurable plastic deformation occur. At the two higher stresses plastic deformations of different magnitudes were measured. Lorosch divided the races at the two higher stresses into groups according to these magnitudes. Group A included the races with the smaller deformations, and group B included the races with the larger deformations. The resultant stress-life exponent of groups A and B was 12. However, when E.V. Zaretsky [28] reconstituted the data for groups A and B into a single group, designated group AB, the stress-life exponent varied from 12 to 27 depending on the stress range over which the exponent was calculated. Lorosch [58] did not calculate the stress

reduction resulting from plastic deformation of the races, nor did he report on component hardness, hardness differential between the balls and races, or residual stresses induced during operation—all of which would also have affected his results. From these tests Lorosch concluded that “under low loads and with elastohydrodynamic lubrication there is no material fatigue, thus indicating that under such conditions bearing life is practically unlimited.”

Zwirlein and Schlicht [59], in a companion paper published concurrently with that of Lorosch [58] and using the same 7205B-size bearing inner race data, state that “contact pressures less than 2.6 GPa (370 ksi) do not lead to the formation of pitting within a foreseeable period. This corresponds to ‘true endurance’.” This observation would support the assumption by Ioannides and Harris [32] of the existence of a “fatigue limit for bearing steels.” However, this observation is not supported by rolling-element fatigue data in the open literature for stress levels under 2.6 GPa (370 ksi) – such as those reported in [60–64], which exhibited classical rolling-element fatigue. In rotating machinery nearly all rolling-element bearings operate at a maximum Hertz stress less than 2.1 GPa (300 ksi). Therefore, if Lorosch [58] and Zwirlein and Schlicht [59] were correct, no bearing in rotating machinery applications would fail by classical rolling-element fatigue.

To the author’s knowledge there are no reported laboratory-generated, full-scale bearing rolling-element fatigue data at stresses significantly below 2.1 GPa (300 ksi) that would either establish or refute the presumption of a “fatigue limit.” However, Townsend *et al.* [65] reported rolling-element and bending fatigue tests for four sets of spur gears made from through-hardened, vacuum-induction-melted, consumable-electrode-vacuum-arc-remelted (VIM–VAR) AISI M–50 steel and case-carburized, vacuum-arc-remelted (VAR) AISI 9310 steel. These gears were tested at a maximum Hertz stress of 1.7 GPa (248 ksi) and a maximum bending stress at the tooth root of 0.27 GPa (39 ksi). This results in a maximum shearing stress at the tooth root of 0.14 GPa (20 ksi). The AISI 9310 gears were “standard” machined as was one set of the AISI M–50 gears. A second set of AISI M–50 gears was “standard” near-net-shape forged using a controlled-energy-flow forming technique (CEFF). The third set of AISI M–50 gears was ausforged in a CEFF machine [65].

The results of these tests are shown in Fig. 12 [65]. The entire set of standard machined, case carburized VAR AISI 9310 gears failed by classical rolling-element fatigue at or near the gear tooth pitch diameter. There was no tooth-bending fatigue failure with the AISI 9310 gears. The entire set of standard machined AISI M–50 gears failed by bending fatigue. The standard forged and ausforged AISI M–50 gears failed by classical rolling-element fatigue and had approximately 5 times the L_{10} life of the AISI 9310 gears. The standard machined

AISI M-50 gears had an L_{10} bending fatigue life of about 40 percent that of the AISI 9310 gear surface fatigue life.

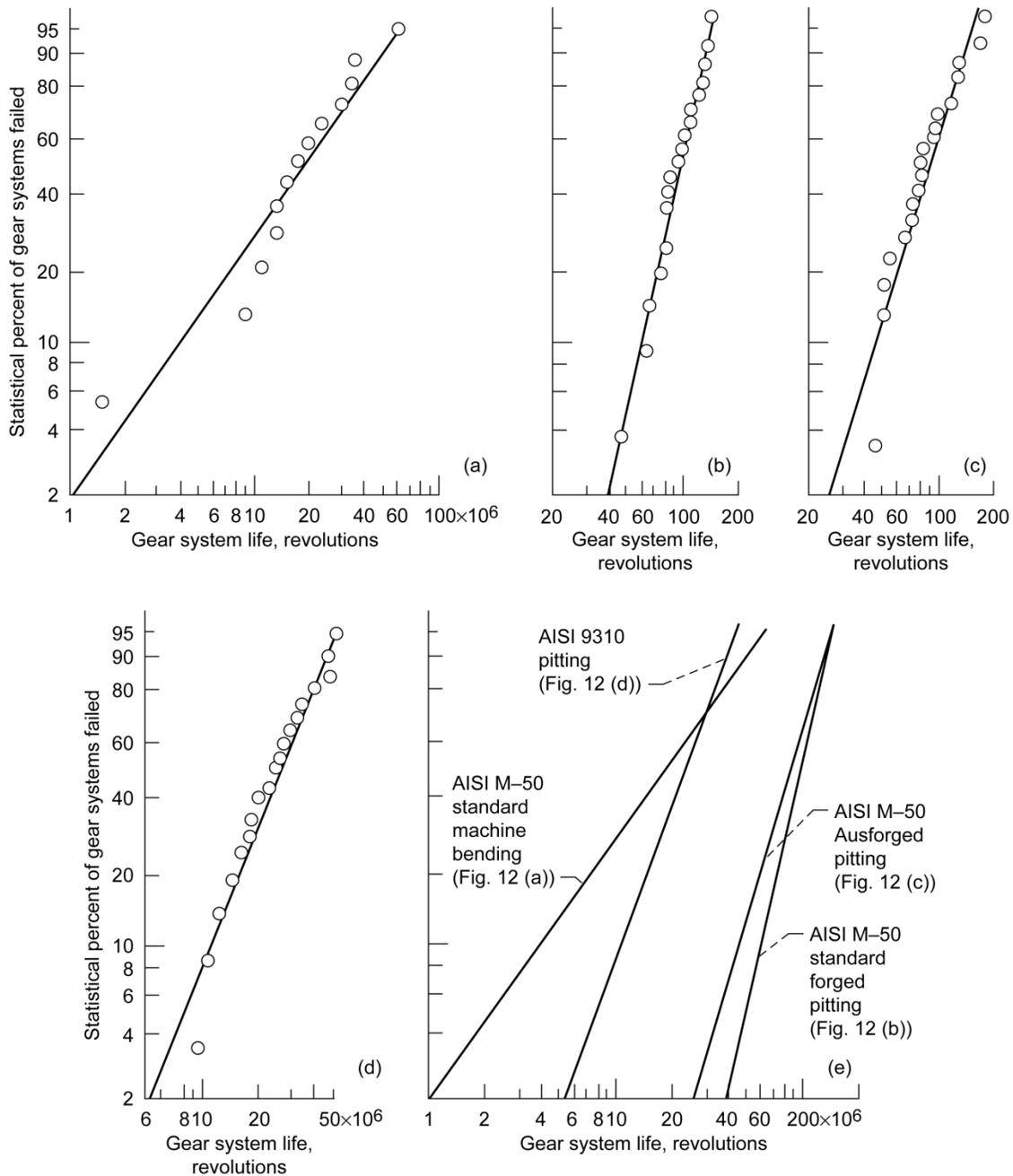


Figure 12. Fatigue lives of spur gear systems made of VIM-VAR AISI M-50 and VAR AISI 9310 [65]. Maximum Hertz stress, 1.71 GPa (248 ksi); maximum bending stress at tooth root, 0.27 GPa (39 ksi); speed, 10,000 rpm; temperature 350 K (170 °F); lubricant, super-refined naphthenic mineral oil. (a) Standard machined AISI M-50 bending fatigue. (b) Standard forged M-50 pitting fatigue. (c) Ausforged M-50 pitting fatigue. (d) AISI 9310 pitting fatigue. (e) Summary of gear life.

What is significant about these tests [65] is that bending fatigue is reported for through hardened bearing steel (AISI M-50) at shearing stresses of 0.14 GPa (20 ksi). These results suggest that if a fatigue limit exists for AISI M-50 it would be less than 0.14 GPa (20 ksi).

While it can be reasonably argued that the resultant stresses may be higher than that reported because of gear tooth dynamic loads, tooth bending fatigue was not experienced with the standard machined case-carburized AISI 9310 gears under the same conditions.

The standard forged and the ausforged AISI M-50 gears did not fail from bending fatigue but did fail from classical rolling-element (surface) fatigue. These results further suggest that if there is a fatigue limit for through-hardened bearing steels it would be less than at a maximum Hertz stress of 1.7 GPa (248 ksi).

As previously discussed, a paper presented by Tosha *et al.* [33] reporting the results of rotating beam fatigue experiments for through-hardened AISI 52100 steel at very low stress levels, shows conclusively that a fatigue limit does not exist for this bearing steel.

The explanation for the trend in the Lorosch [58] and Zwirlein and Schlicht [59] data is the inducement of compressive residual stresses in the AISI 52100 steel caused by the transformation of retained austenite into martensite during rolling-element cycling. These compressive residual stresses were reported by Zwirlein and Schlicht. Compressive residual stresses reduce the effective magnitude of the maximum shear stresses caused by Hertzian loading. This lower stress results in longer bearing life and deviation from the Hertz stress-life exponent n of 9 or 12 to a significantly higher value. Lorosch [58] and Zwirlein and Schlicht [59] extrapolated their data, leading them to conclude the existence of a fatigue-limiting stress rather than concluding that induced compressive residual stresses had increased bearing life.

Lorosch [58] performed another series of rolling-element fatigue experiments with 20 inner races, designated group C, from the same heat as groups A and B. He used a Rockwell hardness tester to make 0.1-mm-diameter indentations at four evenly spaced locations around the center of the race circumference. He divided group C into 10-bearing sets, which were tested at 2.6 and 2.8 GPa (370 and 406 ksi), respectively. The L_{10} lives for group C were significantly reduced from the group AB lives. These tests exhibited a 9.5 stress-life exponent. Because these indentations are analogous to wear debris denting during bearing operation, the results suggest that surface damage is another factor affecting the stress-life exponent.

Residual and hoop stresses

Residual stresses can be induced in a material by heat treating, rolling, shot peening, diamond burnishing, and severe grinding. Each of these methods (except heat treating) is a separate mechanical process that is performed after heat treating [15,18]. Figure 13 shows representative residual stresses as a function of depth below the surface for three heat-treated bearing steels [15,18]. Carburized-grade steels generally have high compressive residual stresses in their cases, as shown for the carburized AISI 9310 and carburized AMS 6278 (VIM-VAR M50 NiL) steels in Fig. 13. Residual stresses are virtually nonexistent in unrun through-hardened steels but are induced at the surface by grinding to a depth of approximately 25 μm (0.001 in.) below the original surface. They can also be induced during operation or stressing of the bearing race surface [15,18].

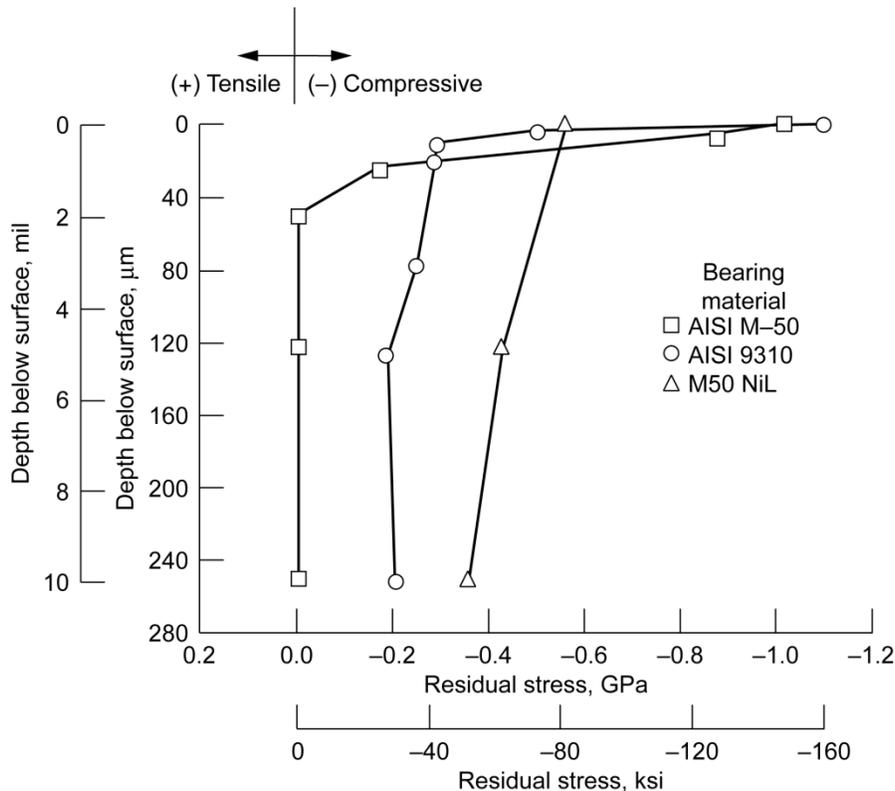


Figure 13. Representative principal residual stress as a function of depth below surface for heat-treated AISI M-50, AISI 9310, and M50 NiL (AMS 6278) [18].

Koistinen [66] reported a method of producing compressive residual stresses in the surface of AISI 52100 steel by austenitizing it in an atmosphere containing ammonia. Stickels and Janotik [67] also induced compressive residual stresses in

the surfaces of AISI 52100 steel rolling-element specimens by austenitizing them in a carbide atmosphere, even though the austenitizing temperature was below that needed to dissolve all primary carbides. The compressive residual stress extended to 300 μm (0.012 in.) below the surface and had a maximum value of 600 MPa (87 ksi). The carburized case (surface layer) contained a larger volume fraction of primary carbides and more retained austenite and was slightly harder than the core [15,18].

Pioneering research on the effect of residual stress on rolling-element fatigue life conducted by the staff of the General Motors Research Laboratories [68 to 71] found that compressive residual stresses induced beneath the surface of ball bearing race grooves prolong rolling-element fatigue life. According to [70], ball bearing lives were doubled when metallurgically induced (“pre-nitrided”) compressive residual stress was present in the inner races. Scott *et al.* [71] found that compressive residual stresses induced by unidentified “mechanical processing” extend the fatigue life of ball bearings.

Naisong *et al.* [72] conducted rolling-element fatigue tests in the rolling-contact fatigue tester with AISI 52100 steel specimens that had compressive residual stress induced by treatment in a carburizing atmosphere. The tests, conducted at a maximum Hertz stress of 5 GPa (729 ksi), resulted in a depth to the maximum shear stress of 0.15 mm (0.006 in.). The treated material with the induced compressive residual stresses had approximately 1.6 times the life of untreated AISI 52100 in spite of having more carbides at or near the surface.

From Naisong *et al.* [72] it is apparent that the distribution of the induced compressive residual stresses is a function of the carbon potential. For ball and roller bearings the zone of maximum resolved shear stresses due to Hertzian loading occurs from 0.10 to 0.25 mm (0.004 to 0.010 in.) below the surface. With a carbon potential of 0.9 wt% carbon in iron, effective compressive residual stresses were available to a depth of 0.30 mm (0.012 in.). This depth was sufficient to have a beneficial effect. For carbon potentials less than 0.7 wt% carbon, high tensile residual stresses were present at depths to 0.20 mm (0.008 in.). Hence, without taking due care it is also possible to reduce the fatigue life by using the carburizing process to induce residual stresses [15,18].

The results of this research indicated that a compressive residual stress at the depth of the zone of maximum resolved shear stresses does prolong rolling-element fatigue life. Jones [73], Carter [74], Akoaka [75], and Zaretsky *et al.* [76] suggested that the maximum shear stress τ_{max} is the most significant stress in the rolling-element fatigue process. Zaretsky *et al.* [77] developed an analysis that superimposes the residual stresses upon the principal subsurface stress, lowering the maximum shear stress. A similar analysis was reported by Foord *et al.* [78] and then by Cioclov [79] in a discussion to Foord *et al.* The theoretical effect of

compressive residual stresses due to heat treatment on rolling-element fatigue is shown in Table 8.

Table 8. Theoretical effect of compressive residual stresses due to heat treatment on rolling-element fatigue life [18].

Material	Compressive residual stress, GPa (ksi)	Maximum Hertz stress, GPa (ksi)			
		1.4 (200)	1.9 (275)	2.4 (350)	4.8 (700)
		Relative life ^a			
AISI M-50	0	1	5.7×10^{-2}	6.5×10^{-3}	1×10^{-5}
AISI 9310	0.2 (29)	12.1	32.4×10^{-2}	24.8×10^{-3}	2×10^{-5}
AMS 6278 (VIM-VAR M50 NiL)	0.4 (58)	381	280×10^{-2}	119.1×10^{-3}	5×10^{-5}

^aRelative life normalized to AISI M-50 at a maximum Hertz stress of 1.4 GPa (200 ksi) and no subsurface residual stress. These relative lives are not life factors. This table shows the potential effect of both applied stress and subsurface residual compressive stress on theoretical life.

The analysis of Zaretsky *et al.* [77], illustrated in Fig. 14, shows the principal stresses in the tangential (rolling) direction S_t and in the normal direction S_n for a Hertzian contact. The maximum shear stress is

$$\tau_{\max} = \frac{1}{2}(S_n - S_t) \tag{79}$$

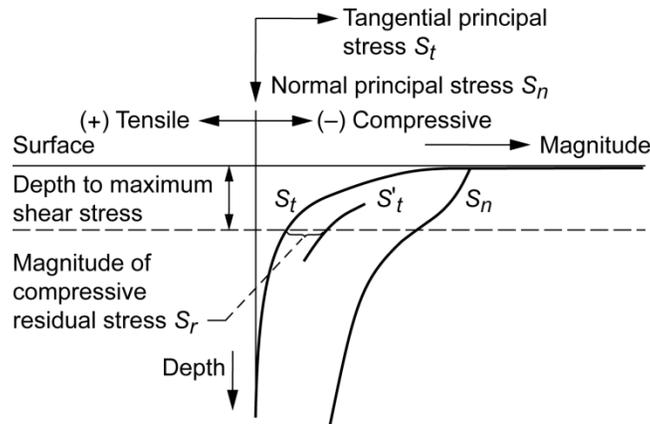


Figure 14. Effect of superimposing compressive residual stress on tangential principal stress in direction of rolling under Hertzian contact [15].

Superimposing the value of the compressive residual stress on the tangential principal stress gives the modified principal tangential stress

$$S'_t = S_t + S_r \quad (80)$$

By combining Eqs. (79) and (80), the maximum shear stress due to the effect of residual stresses $(\tau_{\max})_r$ can be determined:

$$(\tau_{\max})_r = \frac{1}{2}(S_n - S'_t) = \frac{1}{2}[S_n - (S_t + S_r)] = -\tau_{\max} - \frac{1}{2}(-S_r) \quad (81)$$

The negative sign before S_r designates a compressive residual stress. A positive sign before S_r would designate a tensile residual stress [77]. Hence, the residual stress can either increase or decrease the maximum shear stress. Accordingly, a compressive residual stress would reduce the maximum shear stress and increase the fatigue life according to the inverse relation of life and stress to the c/e power,

$$L \sim \left[\frac{1}{(\tau_{\max})_r} \right]^{c/e} \quad (82a)$$

for Lundberg and Palmgren, where $c/e = 9.3$, or

$$L \sim \left[\frac{1}{(\tau_{\max})_r} \right]^c \quad (82b)$$

for Zaretsky, where $9 \leq c \leq 10$.

For rolling-element bearings the beneficial effect of the compressive residual stresses is offset by the presence of tensile (hoop) stresses in the bearing inner race.

These hoop stresses are induced and affected primarily by the press fit of the inner race over the shaft, centrifugal loading, and thermal effects between the race and the shaft. Coe and Zaretsky [80] analyzed the effect of hoop stresses on rolling-element fatigue. Czyzewski [81] showed that, in the absence of compressive residual stress, hoop stresses are generally tensile, are designated by a + sign, and can negatively affect fatigue life. As with the compressive residual stresses, hoop stresses alter the critical subsurface shear stress and affect the life of the bearing inner race. Equation (81) can be rewritten to account for the effect of both residual and tensile (hoop) stresses as

$$(\tau_{\max})_{rh} = -\tau_{\max} - \frac{1}{2}(\pm S_r \pm S_h) \quad (83)$$

Again the positive or negative sign indicates a tensile or compressive stress, respectively. Equation (82) can be written to reflect the effect of both the compressive and tensile stresses on life as

$$L \sim \left[\frac{1}{(\tau_{\max})_{rh}} \right]^{c/e} \quad (84a)$$

for Lundberg and Palmgren, where $c/e = 9.3$, or

$$L \sim \left[\frac{1}{(\tau_{\max})_{rh}} \right]^c \quad (84b)$$

for Zaretsky, where $9 \leq c \leq 10$.

Since life $L \sim \left(\frac{1}{\tau}\right)^{c/e}$ or $\left(\frac{1}{\tau}\right)^c$, Eqs. (48) and (84) have been combined to modify the predicted bearing life [18]:

$$L_{10} = \left[\frac{\tau_{\max}}{(\tau_{\max})_{rh}} \right]^{c/e} \left[\frac{C_D}{P_{eq}} \right]^p \quad (85a)$$

From Lundberg and Palmgren [9], $c/e = 9.3$ and $p = 3$ for ball bearings and $p = 4$ for roller bearings. From Zaretsky,

$$L_{10} = \left[\frac{\tau_{\max}}{(\tau_{\max})_{rh}} \right]^c \left[\frac{C_D}{P_{eq}} \right]^p \quad (85b)$$

where $p = 4$ for ball bearings, $p = 5$ for roller bearings, and $9 \leq c \leq 10$. From Hertz theory [29], $\tau_{\max} \approx 0.32S_{\max}$ for ball bearings and $0.3S_{\max}$ for roller bearings.

Comparison of bearing life models

The Ioannides-Harris model without a fatigue limit is identical to Lundberg-Palmgren model. The Weibull model is similar to that of Zaretsky if the exponents are chosen to be identical. Comparison of the Lundberg-Palmgren model and the Zaretsky model with the ANSI/ABMA and ISO standards is shown in Fig. 15 for

ball and cylindrical roller bearings. The theoretical lives were normalized to a maximum Hertz stress of 4.14 GPa (600 ksi) and subsequently normalized to the calculated ANSI/ABMA and ISO standards at each stress level. For the Ioannides-Harris comparison shown in Fig. 15, a fatigue-limiting stress of 276 MPa (40 ksi) was assumed. For ball bearings, the ANSI/ABMA and ISO standards and the Lundberg-Palmgren model give identical results. For roller bearings, the results are not identical [82].

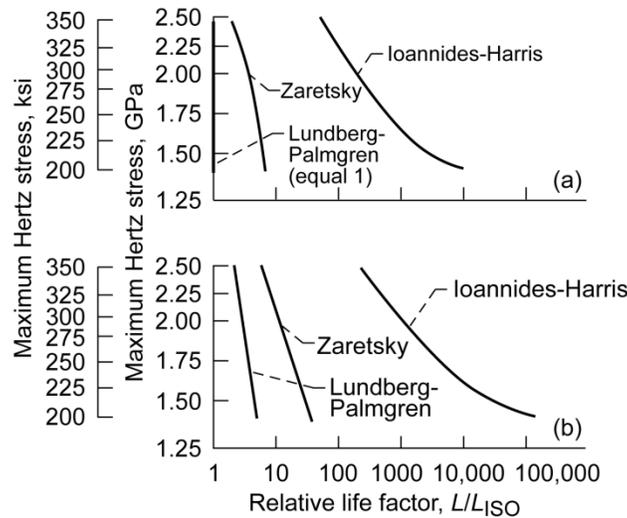


Figure 15. Comparison of life models for rolling-element bearings with ANSI/ABMA and ISO standard without a fatigue limit. Fatigue-limiting shear stress for τ_{45} assumed for Ioannides-Harris model, 276 MPa (40 ksi) [82]. (a) Ball bearings. (b) Cylindrical roller bearings.

The ANSI/ABMA and ISO standards use a load-life exponent p of $10/3$ (3.33) for line contact (roller bearings). This results in a value of n equal to 6.6 and can account in part for lower life predictions than those experienced in the field. From Lundberg and Palmgren [9], the load-life exponent p for line contact should be 4. However, Lundberg and Palmgren’s justification for a p of $10/3$ was that a roller bearing can experience “mixed contact”; that is, one raceway can experience line contact and the other raceway point contact [10]. This may be true in limited number of roller bearing designs, but it is certainly not consistent with the vast majority of cylindrical roller and tapered roller bearings designed and used today.

Both the load-life and stress-life relations of Weibull, Lundberg and Palmgren, and Ioannides and Harris reflect a strong dependence on the Weibull slope. The existing rolling-element fatigue data reported by Parker and Zaretsky [44] reflect slopes in the range of 1 to 2 and some cases higher or lower. If the slope were factored into these equations, then the stress-life (load-life) exponent significantly

decreases with increases in Weibull slope, whereby the relation no longer matches reality.

The Zaretsky model that decouples the dependence of the critical shear stress-life relation and the Weibull slope shows only a slight variation of the maximum Hertz stress-life exponent n with Weibull slope e .

These results would indicate that for 9th- and 8th-power Hertz stress-life exponents for ball and roller bearings, respectively, the Lundberg-Palmgren model best predicts life. However, for 12th- and 10th-power relations reflected by modern bearing steels, the Zaretsky model based on the Weibull equation is superior [82].

Under the range of stresses examined, the use of a fatigue limit would suggest that (for most operating conditions under which a rolling-element bearing will operate) the bearing will not fail from classical rolling-element fatigue. Realistically, this does not occur. The use of a fatigue limit will significantly overpredict life over a range of normal operating Hertz stresses. (The use of ISO 281:2007 [36] in these calculations would result in a bearing life approaching infinity.) Since the predicted lives of rolling-element bearings are high, the problem can become one of undersizing a bearing for a particular application [31].

Comparing life data with predictions

Bearing life variation

Vlcek *et al.* [83] randomly assembled and tested 340 virtual bearing sets totaling 31,400 radially loaded and thrust-loaded rolling-element bearings. It was assumed that each bearing was assembled from three separate bins of components, with one bin containing 1000 inner rings; one with 1000 rolling element sets, and one with 1000 outer rings. The median ranks of the individual components were assigned and then virtual bearing assemblies were created using a Monte Carlo technique. The corresponding lives of the bearing components were determined using Eq. (17). The weakest link theory was applied; that is, it was assumed that the life of the shortest-lived component of the system was the life of the system. A linear curve fit of these system lives results in a Weibull plot. Weibull parameters from the plot and Eq. (17) can be used to determine lives at any percentage of survivability.

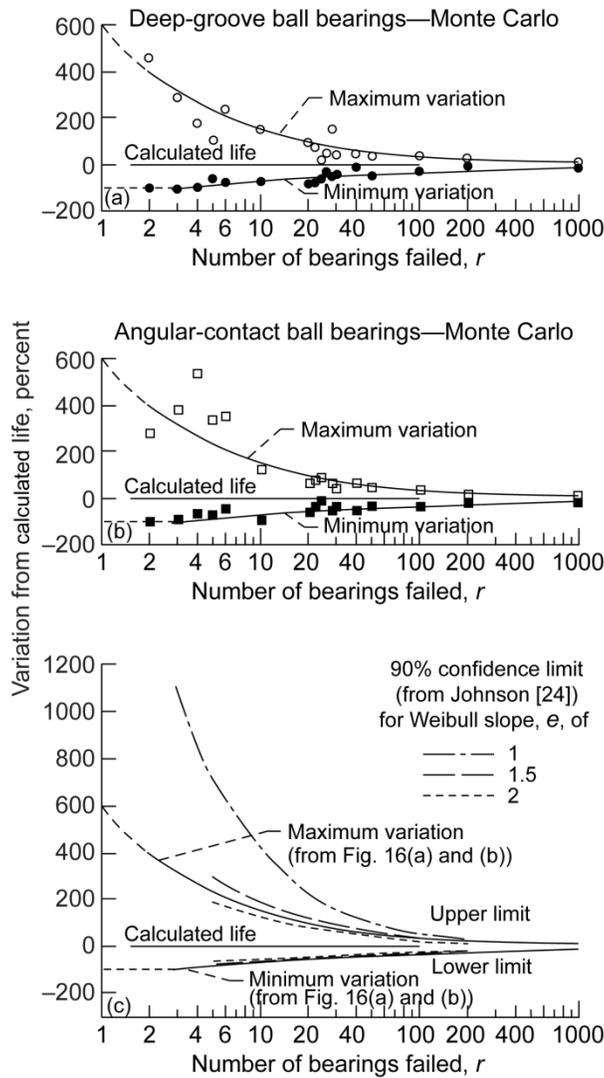


Figure 16. Maximum and minimum variation of L_{10} lives as percent of calculated L_{10} for groups of r bearings and 90% confidence limits based on Weibull slope and number of bearings failed r [83]. (a) 50-mm-bore deep-groove ball bearing. (b) 50-mm-bore angular-contact ball bearing. (c) 90% confidence limits.

Vlcek *et al.* [83] determined the L_{10} maximum limit and L_{10} minimum limit for the number of bearings failed, r , using a Weibull-based Monte Carlo method. By fitting the resultant lives for different size populations of failed bearings (Fig. 16), equations were determined for both of these limits:

$$\text{Maximum variation } L_{10} \text{ life} = \text{calculated } L_{10} \text{ life} (1+6r^{-0.6}) \quad (86a)$$

$$\text{Minimum variation } L_{10} \text{ life} = \text{calculated } L_{10} \text{ life} (1-1.5r^{-0.33}) \text{ where } r > 3 \quad (86b)$$

$$\text{Minimum } L_{10} \text{ life} = 0 \text{ where } r \leq 3 \quad (86c)$$

These curves compared favorably with the 90-percent confidence limits of Johnson [24] at a Weibull slope of 1.5 (Fig. 16) [83].

Rules can be implied from these results to compare and distinguish resultant lives of similar bearings from two or more sources and/or made using different manufacturing methods. The following rules are suggested to determine if the bearings are acceptable for their intended application and if there are significant differences between two groups of bearings.

1. If the L_{10} lives of both bearing sets are between the maximum and minimum L_{10} life variations, there can be no conclusion that there is a significant difference between the two sets of bearings regardless of the ratio of the L_{10} lives. The bearing sets are acceptable for their intended application (Fig. 17(a)).
2. If the L_{10} life of one set of bearings is greater than the maximum variation and the second set is less than the minimum value, there exists a significant difference between the bearing sets. Only one bearing set is acceptable for its intended application (Fig. 17(b)).
3. If the L_{10} lives of both sets of bearings exceed the maximum variation, the bearing life differences may or may not be significant and should be evaluated based upon calculation of confidence numbers according to the method of Johnson [24]. Both sets of bearings are acceptable for their intended application (Fig. 17(c)).
4. If the L_{10} lives of both sets of bearings are less than the minimum variation, the bearing life differences may or may not be significant. However, neither set of bearings is acceptable for its intended application (Fig. 17(d)).
5. If the L_{10} life of one set of bearings exceeds the maximum variation and the other set is between the maximum and minimum variation, the bearing life differences may or may not be significant and should be evaluated based upon calculation of confidence numbers according to the method of Johnson [24]. Both sets of bearings are acceptable for their intended application (Fig. 17(e)).
6. If the L_{10} life of one set of bearings is less than the minimum variation and the other set is between the maximum and minimum variation, there exists a significant difference between the bearing sets. Only one set of bearings is acceptable for its intended application (Fig. 17(f)).

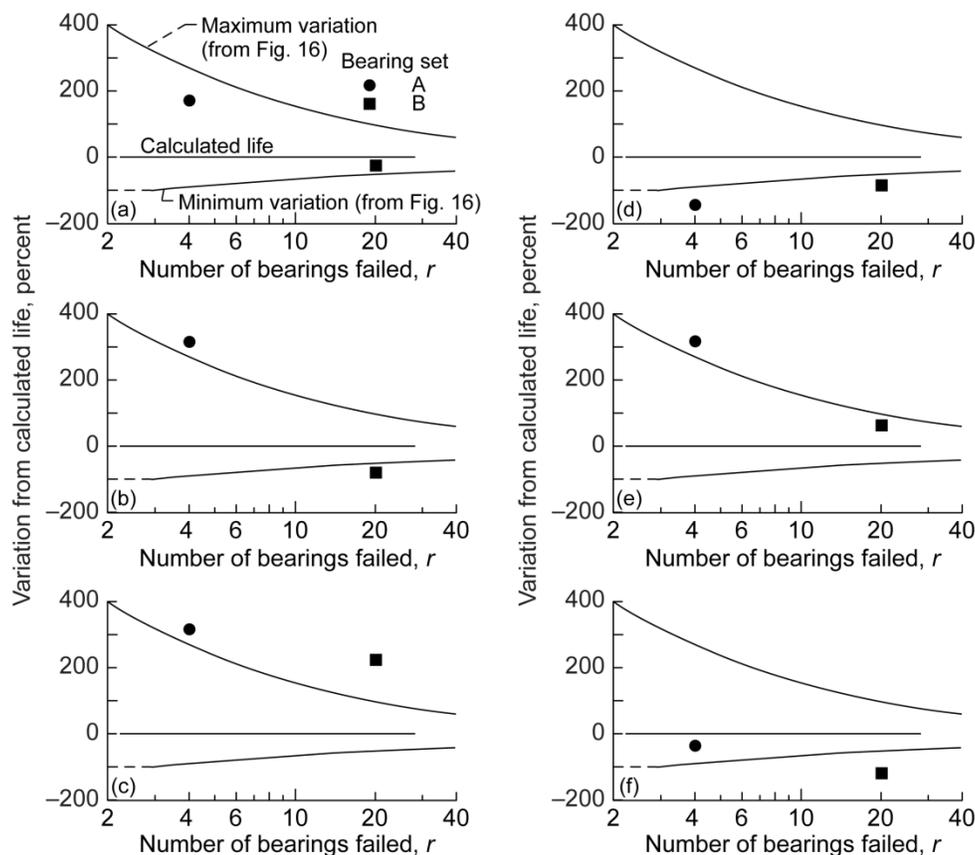


Figure 17. Rules for comparing bearing life results to calculated life [83]. (a) Bearing sets A and B are acceptable. (b) Bearing set A is acceptable. Bearing set B is not acceptable. (c) Bearing sets A and B are acceptable. (d) Bearing sets A and B are not acceptable. (e) Bearing sets A and B are acceptable. (f) Bearing set A is acceptable. Bearing set B is not acceptable.

Harris [84] and Harris and McCool [85] analyzed endurance data for 62 rolling-element bearing sets. These data were obtained from four bearing manufacturers, two helicopter manufacturers, three aircraft engine manufacturers, and U.S. Government agency-sponsored technical reports. The data sets comprised deep-groove radial ball bearings, angular-contact ball bearings, and cylindrical roller bearings for a total of 7935 bearings. Of these, 5321 bearings comprised one sample size for a single cylindrical roller bearing, leaving 2614 bearings distributed among the remaining bearing types and sizes. Among the 62 rolling-element bearing endurance sets, 11 had one or no failure and could not be used for the analysis. These data are summarized in [83] and plotted in Fig. 18. A discussion of the Harris data can be found in [28] and [43].

Of these data, 39 percent fall between the maximum and minimum life variations suggesting that the statistical variations of these lives are within that

predicted. Four bearing sets representing 8 percent of the bearing sets had lives less than what was predicted. Thirty bearing sets, or 59 percent of the bearing sets, exceeded the maximum life variation of the Monte Carlo study of Vlcek *et al.* [83]. Eight of these bearing sets or 16 percent exceeded the 90-percent confidence upper limit of Johnson [24]. However, only one bearing set representing 2 percent of the bearing sets fall below the lower life limit. Therefore, it can be reasonably concluded that 98 percent of the bearing sets have acceptable life results using the Lundberg-Palmgren equations with the life adjustment factors from [18] to predict bearing life.

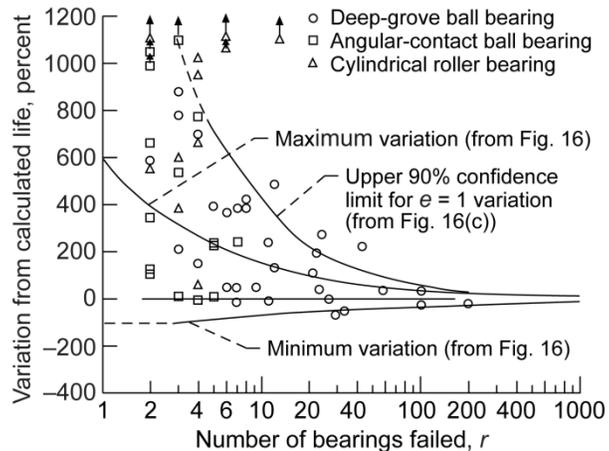


Figure 18. Variation between actual and calculated L_{10} bearing lives for 51 sets of deep-groove and angular-contact ball bearings and cylindrical roller bearings from [84] and [85] compared with Monte Carlo variations and 90% confidence limit [83].

The life calculations for the data of Fig. 18 have material and steel processing life factors from [18] incorporated in them. Table 9 summarizes these life factors for each of the materials. The data of Fig. 18 are broken down and plotted in parts (a) of Figs. 19 through 22 based on the steel type and processing. These data were adjusted for a load-life exponent p of 4 for ball bearings and 5 for roller bearings and are shown in parts (b) of Figs. 19 through 22. The adjusted life results correlated with those of the Monte Carlo tests shown in Fig. 16. Based upon these material and processing life factors and load-life exponents, each bearing data set appears consistent with the other.

Lundberg and Palmgren [9] assumed the value of the Weibull slope e in Eq. (17) to be 1.11 for ball bearings [9] and 1.125 for roller bearings [10]. These values were necessary in their analysis because it approximated those values exhibited by their experimental data, and it made the end result of their life prediction analysis correlate with their bearing life database at that time. Experience has shown that

most rolling bearing life data exhibit Weibull slopes between 1 and 2. For the analysis in [83] a Weibull slope of 1.11 was assumed for all of the components for each bearing. This should theoretically result in a bearing Weibull slope of 1.11 as shown in Fig. 23 for the deep-groove and angular-contact ball bearings.

Table 9. Life factors for bearing steels and processing [18].

Material and process	Life factor		
	Material	Process	Resultant
CVD ^a AISI 52100	3	1.5	4.5
CVD ^a AISI 8620	1.5	1.5	2.25
VAR ^b AISI M-50	2	3	6
VIM-VAR ^c AISI M-50	2	6	12
VIM-VAR ^c M50 NiL	4	6	24

^aCarbon vacuum degassing.

^bVacuum arc remelting.

^cVacuum induction melting and vacuum arc remelting.

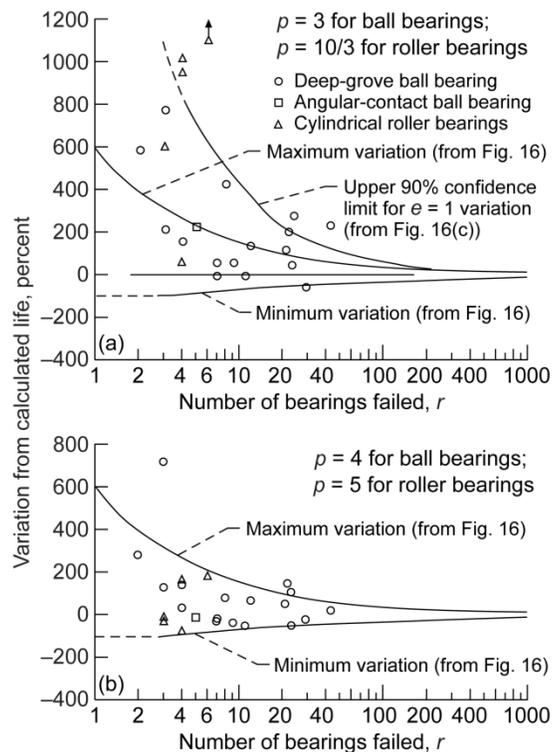


Figure 19. Effect of CVD AISI 52100 steel and load-life exponent on bearing life [83]. (a) Load-life exponent p is 3 for ball bearings and 10/3 for cylindrical roller bearings (from Fig. 18). (b) Load-life exponent p is 4 for ball bearings and 5 for cylindrical roller bearings.

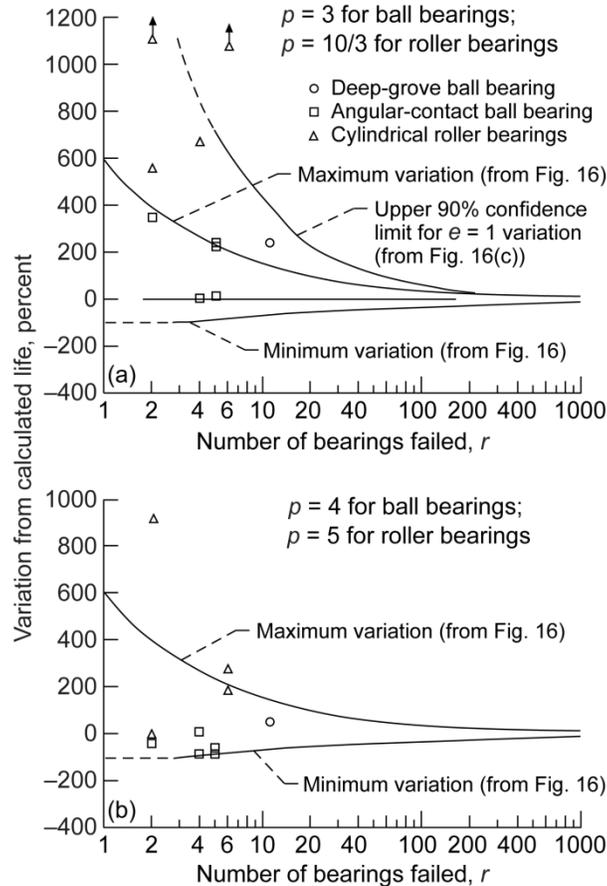


Figure 20. Effect of VAR AISI M-50 steel and load-life exponent on bearing life [83]. (a) Load-life exponent p is 3 for ball bearings and 10/3 for cylindrical roller bearings (from Fig. 18). (b) Load-life exponent p is 4 for ball bearings and 5 for cylindrical roller bearings.

Weibull slope variation

Johnson [24] analyzed the probable variation of the Weibull slope as a function of the number of bearings tested to failure. Based on the Johnson analysis, in 90 percent of all possible cases the resultant Weibull slope will be within the limits shown in Fig. 23 based upon a Weibull slope of 1.11. Based on Johnson, the approximate relation for the number of bearings failed r and the limits of the value of Weibull slope e equal to 1.11 are as follows:

$$\text{Maximum Weibull slope} = 1.11 + 1.31 r^{-0.5} \quad (87a)$$

$$\text{Minimum Weibull slope} = 1.11 - 1.31 r^{-0.5} \quad (87b)$$

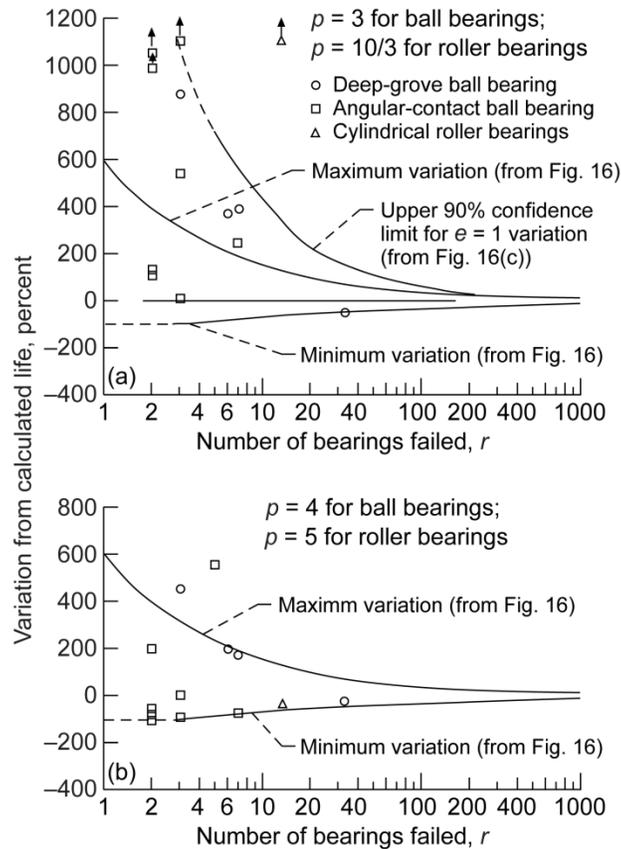


Figure 21. Effect of VIM-VAR AISI M-50 steel and load-life exponent on bearing life [83]. (a) Load-life exponent p is 3 for ball bearings and 10/3 for cylindrical roller bearings (from Fig. 18). (b) Load-life exponent p is 4 for ball bearings and 5 for cylindrical roller bearings.

The results of the extremes in the Weibull slopes for each group of the 10 bearing trials of r bearings are compared with the Johnson analysis in Fig. 23(a). (Note that the Weibull slopes for the data summarized in Fig. 18 for the maximum and minimum bearing lives are not necessarily the same as the maximum and minimum values of the Weibull slopes for each of trials of r bearings.) For the data shown in Fig. 18 the relation between the number of bearings tested and the limits of the Weibull slope are as follows:

$$\text{Maximum Weibull slope} = 1.2 + 5(\ln r)^{-3} \quad (88a)$$

$$\text{Minimum Weibull slope} = 1.11 - 0.95r^{-0.33} \quad (88b)$$

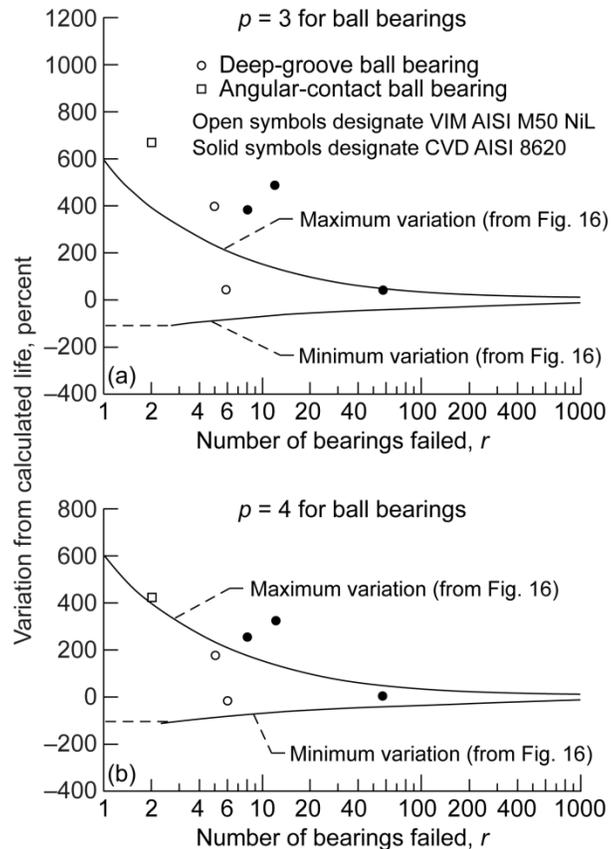


Figure 22. Effect of VIM-VAR AISI M50 NiL and CVD AISI 8620 steels and load-life exponent on bearing life [83]. (a) Load-life exponent p is 3 for ball bearings (from Fig. 18). (b) Load-life exponent p is 4 for ball bearings.

Where the number of bearings failed is 10 or greater, there is a reasonably good correlation between the limits of the slopes generated from the Johnson analysis [24] and those from the Monte Carlo bearing tests shown in Fig. 23(a). Where the number of failed bearings is below 10, there are differences between the extremes in Weibull slope between the Monte Carlo bearing tests and those of Johnson, especially at the upper limits for the Weibull slopes [83].

Case study

Turboprop gearbox

The commercial turboprop gearbox used for this analysis (Fig. 24) consists of 2 stages with a single mesh spur reduction followed by a 5-planet planetary gearbox consisting of 11 rolling-element bearings and 9 spur gears [86]. The first stage consists of the input pinion gear meshing with the main drive gear. The second

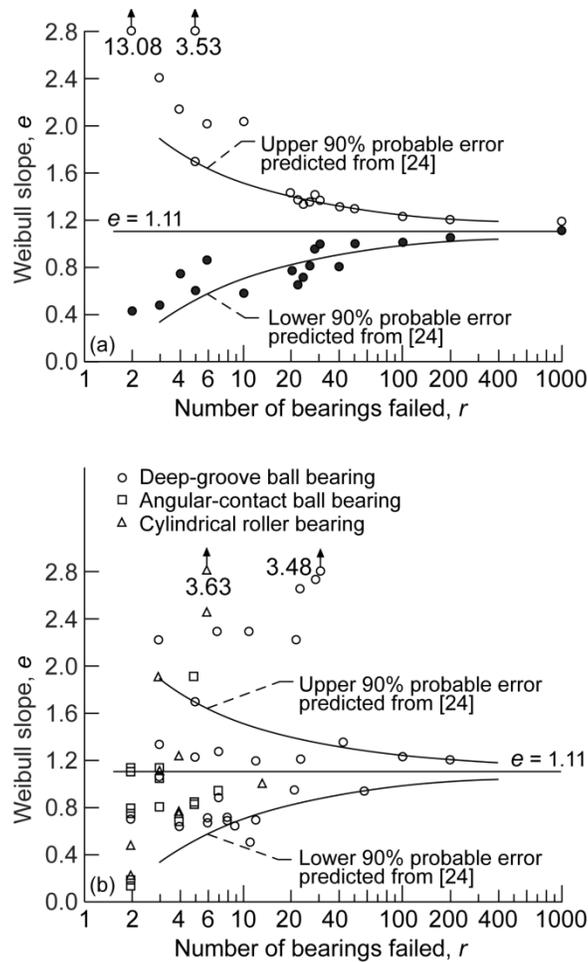


Figure 23. Variation of Weibull slope e compared with predicted 90% probable error. (a) Extremes of Weibull slope from Monte Carlo testing for each group of 10 bearing trials of r bearings (data from [83]). (b) Weibull slopes from 51 sets of ball and roller bearings (data from [84] and [85]).

stage is provided by the fixed-ring planetary gear driven by a floating sun gear as input with a five-planet carrier as output. At cruise conditions, the input pinion speed is constant at 13,820 rpm, producing a carrier output speed of 1021 rpm. A list of the component parts of the gearbox is given in Table 10.

A typical mission profile for this commercial gearbox is given in Table 11, which presents the duration as a percentage and the propeller shaft power for each flight condition. This profile includes loads for (a) takeoff, (b) climb, (c) cruise, and (d) descent. The cruise segment of the profile consumes 68 percent of the flight time with a little less than half of the power required for the takeoff, which lasts for less than 3 percent of the flight time [87].

The cause for removal can be assumed to be one or more bearings or gears that had fatigue or damage resulting in wear and/or vibration detected by magnetic chip

detectors and/or vibration pickups. The gearbox is removed from service before secondary damage occurs and is inspected. After the failed part or parts are replaced, the gearbox is put back into service [87].

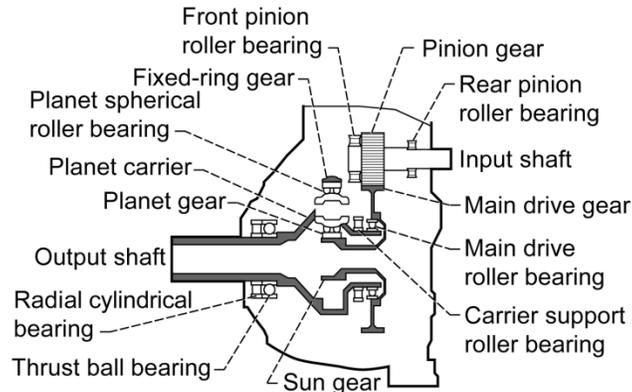


Figure 24. Commercial turboprop gearbox [87].

Table 10. Predicted turboprop gearbox component lives from Lundberg-Palmgren analysis with life factors and strict-series system reliability [87].

Component description (see Fig. 24)	Predicted life, hr		Weibull slope, e
	L_{10}	L_{50}	
Rolling-element bearings			
Cylindrical roller bearings			
Front pinion	20,890	111,476	↓
Rear pinion	21,312	113,728	
Main drive	26,459	141,194	
Carrier support	312,881	1,669,635	
Propeller radial	68,194	363,905	
Propeller thrust ball bearing	33,065	180,484	1.11
Planet double-row spherical roller bearing set, five bearings	844	4,503	1.125
Bearing system life	774	4,132	1.125
Gears			
Pinion	53,477	131,552	↓
Ring	4,540,212	11,168,760	
Sun	19,033	46,821	
Main drive	108,148	266,040	
Planet gear set, five gears	28,092	69,105	
Gear system life	16,680	44,032	2.5
Gearbox life	774	4,132	1.125

Table 11. Mission profile of commercial turboprop gearbox [87].

Mission segment	Percent time of segment	Propeller shaft power, kW (hp)
Takeoff	2.84	3132 (4200)
Climb	17.02	2461 (3300)
Cruise	68.08	1516 (2033)
Descent	12.06	945 (1267)
Equivalent	100.00	1833 (2457)

Individual failure occurrences are not predictable but are probabilistic. No two gearboxes run under the same conditions fail necessarily from the same cause and/or at the same time. At a given probability of survival, the life of the gearbox system will always be less than that of the lowest lived element in it.

Historical field data for 64 gearboxes were collected. The first part of these data covered the time from their field installation and first field operation to their removal for cause (failure) and refurbishment [87].

Analysis

Bearing life analysis

Equations based on the Lundberg-Palmgren model with life modifying factors [18] were used to calculate bearing lives in Table 10 and shown in Fig. 25(a). Equation (48) can be rewritten to include the bearing life modifying factors [18] as follows:

$$L = a_1 a_2 a_3 L_{10} = a_1 a_2 a_3 \left[\frac{C_D}{P_{eq}} \right]^p \quad (89)$$

where a_1 is a reliability factor, a_2 is a materials and processing factor, and a_3 is an operating conditions factor such as lubrication [18]. Table 12 contains a list of representative variables affecting bearing life that contribute to these factors [18]. For the purpose of the bearing life analysis summarized in Table 10, for ball bearings $p \leq 3$, and for roller bearings $p = 4$ [87].

The L_{10} life of a single double-row spherical bearing is 3529 hr. The system L_{10} life for the five-bearing planetary set is 774 hr. For all the bearings in the gearbox, the bearing system L_{10} life is also predicted to be 774 hr [87].

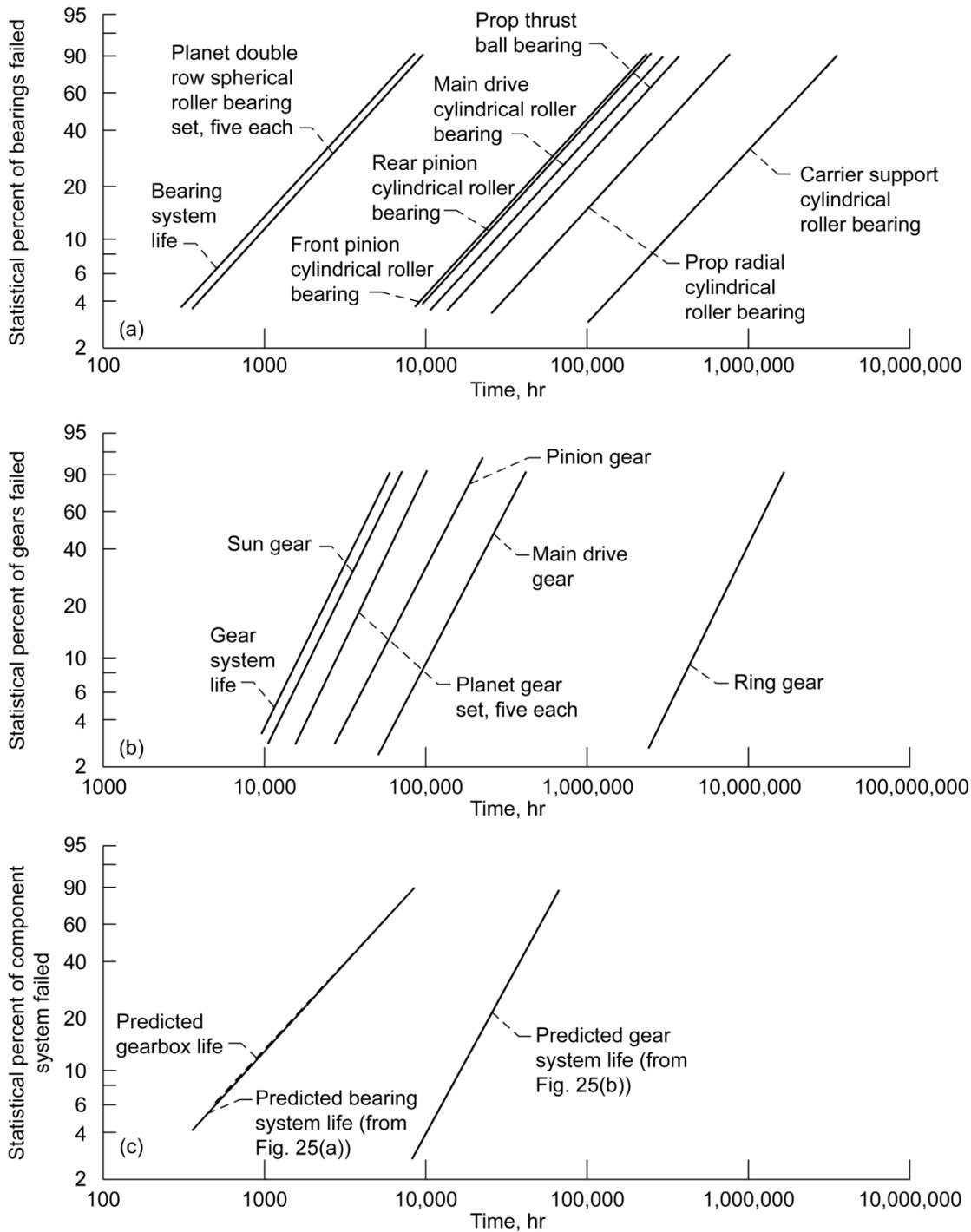


Figure 25. Weibull plots of predicted lives for commercial turboprop gearbox and its respective bearing gear components using Lundberg-Palmgren life theory [87]. (a) Bearing component lives. (b) Gear component lives. (c) Gearbox life and component lives.

Table 12. Representative variables affecting bearing life and reliability [18].

Life adjustment factor	Variable
Reliability, a_1	Probability of failure
Materials and processing, a_2	Bearing steel Material hardness Residual stress Melting process Metal working
Operating conditions, a_3	Load Misalignment Housing clearance Axially loaded cylindrical bearings Rotordynamics Hoop stresses Speed Temperature Steel Lubrication Lubricant film thickness Surface finish Water Oil Filtration

Gear life analysis

Between 1975 and 1981, Coy, Townsend, and Zaretsky [88 to 90] published a series of papers developing a methodology for calculating the life of spur and helical gears based upon the Lundberg-Palmgren theory and methodology for rolling-element bearings. Townsend, Coy, and Zaretsky [48] reported that for AISI 9310 spur gears, the Weibull slope e is 2.5. Based on Eq. (20), for all gears except a planet gear, the gear life can be written as

$$L_{10G} = \frac{N^{-1/e_G} (\eta_{10t})}{k} \quad (90)$$

For a planet gear, the life is

$$L_{10G} = \frac{N^{-1/e_G} (\eta_{10t1}^{-e_G} + \eta_{10t2}^{-e_G})^{-1/e_G}}{k} \quad (91)$$

The L_{10} life in millions of stress cycles of a single gear tooth can be written as

$$\eta_{10_t} = a_2 a_3 \left(\frac{C_t}{P_t} \right)^{p_G} \quad (92)$$

where

$$C_t = Bf^{0.907} \rho^{-1.165} l^{-0.093} \quad (93)$$

and

$$\rho = \left(\frac{1}{r_1} + \frac{1}{r_2} \right) \frac{1}{\sin \phi} \quad (94)$$

The value for η_{10_t} can be determined by using Eq. (92) where for bearings C_t is the basic load capacity of the gear tooth, P_t is the normal tooth load, p_G is the load-life exponent usually taken as 4.3 for gears based on experimental data for AISI 9310 steel, and a_2 and a_3 are life adjustment factors similar to those for rolling-element bearings. The value for C_t can be determined by using Eq. (93), where B is a material constant that is based on experimental data and is approximately equal to 1.39×10^8 when calculating C_t in SI units (Newtons and meters) and is 21,800 in English units (pounds and inches) for AISI 9310 steel spur gears. Also, f is the tooth width, and ρ is the curvature sum at the start of single-tooth contact.

The L_{10_G} life of the gear (all teeth) in millions of input shaft revolutions at which 90 percent will survive can be determined from Eq. (90) or (91) where N is the total number of teeth on the gear, e_G is the Weibull slope for the gear and is assumed to be 2.5 (from [48]), and k is the number of load (stress) cycles on a gear tooth per input shaft revolution.

For all gears except the planet gears, each tooth will see a load on only one side of its face for a given direction of input shaft rotation. However, each tooth on a planet gear will see contact on both sides of its face for a given direction of input shaft rotation. One side of its face will contact a tooth on the sun gear, and the other side of its face will contact a tooth on the ring gear. This results in reverse bending of the gear teeth on the planetary gears. Since the gears are operated at bending stresses less than those that may cause bending fatigue in carburized gear steels, this effect is ignored. However, for purposes of contact fatigue, Eq. (91) takes into account that both sides of the gear tooth are stressed, where $\eta_{10_{t1}}$ is the L_{10} life in millions of stress cycles of a planet tooth meshing with the

sun gear, and η_{10t_2} is the L_{10} life in millions of stress cycles of a planet tooth meshing with the ring gear.

The calculated gear lives are summarized in Table 10 and shown in the Weibull plots of Fig. 25(b). The gear system predicted L_{10} life is 16,680 hr [87].

Gearbox system life

The L_{10} lives of the individual bearings and gears that make up a rotating machine are calculated for each condition of their operating profiles. For each component, the resulting lives from each of the operating conditions are combined using the linear damage (Palmgren-Langer-Miner) rule, Eq. (14).

The resultant lives of each of the gearbox components are then combined to determine the calculated system L_{10} life using the two-parameter Weibull distribution function (Eq. (17)) for the bearings and gears comprising the system and strict-series system reliability (Eq. (49a)) as follows:

$$\frac{1}{L_{sys}^e} = \sum_{i=1}^n \frac{1}{L_n^e} = \left(\frac{1}{L_1^e} + \frac{1}{L_2^e} + \dots + \frac{1}{L_n^e} \right) \quad (95a)$$

The resultant system lives previously discussed for the bearings and gears are shown in Figs. 25(a) and (b), respectively, and summarized in Table 10.

The system Weibull slope using strict-series reliability (Eq. (95a)) where each component or combinations of multiple bearings and gears have different Weibull slopes is not intuitively obvious. If at each time sequence the probability of survival of each component is multiplied together, the system reliability at that time and, hence, the probability of failure can be determined. When these values are plotted on Weibull paper, a Weibull slope (shape parameter) can be determined for the system life distribution using a least squares fit. When this is done, it is found that the system Weibull slope approximates that of the lowest lived component in the system. For the gearbox, where the subscripts B and G relate the L_{10} life to a bearing or a gear, respectively,

$$\frac{1}{L_{sys}^e} = \left(\frac{1}{L_{B_1}^e} + \frac{1}{L_{B_2}^e} + \dots + \frac{1}{L_{B_n}^e} \right) + \left(\frac{1}{L_{G_1}^e} + \frac{1}{L_{G_2}^e} + \dots + \frac{1}{L_{G_n}^e} \right) \quad (95b)$$

The lowest lived components in the gearbox are the roller bearings. As a result, the Weibull slope assumed for the planetary gear spherical roller bearings is

assumed to be the Weibull slope e of the entire gearbox system in Eq. (95b). The individual system and combined system lives are shown in the Weibull plots of Fig. 25(c) and summarized in Table 10. The predicted L_{10} life of the gearbox is 774 hr [87].

Gearbox field data

The application of the Lundberg-Palmgren model [9] to predict gearbox life and reliability needs to be benchmarked and verified under a varied load and operating profile. The cost and time to laboratory test a statistically significant number of gearboxes to determine their life and reliability is prohibitive. A practical solution to this problem is to benchmark the analysis to field data. Fortunately, these data were available for the commercial turboprop gearbox used in this study.

No two gearboxes are expected to operate in exactly the same manner. Flight variables include operating temperature and load. Small variations in operational load can result in significant changes in life. Hence, the accuracy of the calculations is dependent on how close the defined mission profile is to actual flight operation.

The condition of the gearboxes is monitored, and they are removed from service upon the detection of a perceived component failure. At the time of removal, the gearboxes are functional. The removal precludes secondary damage; that is, the damage is limited to the failed component.

Field data were collected for 64 new commercial turboprop gearboxes. From these field data, the resultant time to removal of each gearbox is presented in the Weibull plot of Fig. 26. The failure index was 59 out of 64. That is, 59 out of the 64 gearboxes removed from service were considered failed. For these data, there was no breakdown of the cause for removal or the percent of each component that had failed. The resultant L_{10} life from the field data was 5627 hr and the Weibull slope e was 2.189. Using the Lundberg-Palmgren method (above), the predicted L_{10} life was 774 hr and the Weibull slope e was 1.125. The field data suggest that the L_{10} life of the gearbox was underpredicted by a factor of 7.56 [87].

Reevaluation of bearing load-life exponent p

Although errors in the assumed operating profile of the gearbox may account for the difference between actual and predicted life, it is suggested that using the Lundberg-Palmgren equations results in a life prediction that is too low for the bearings.

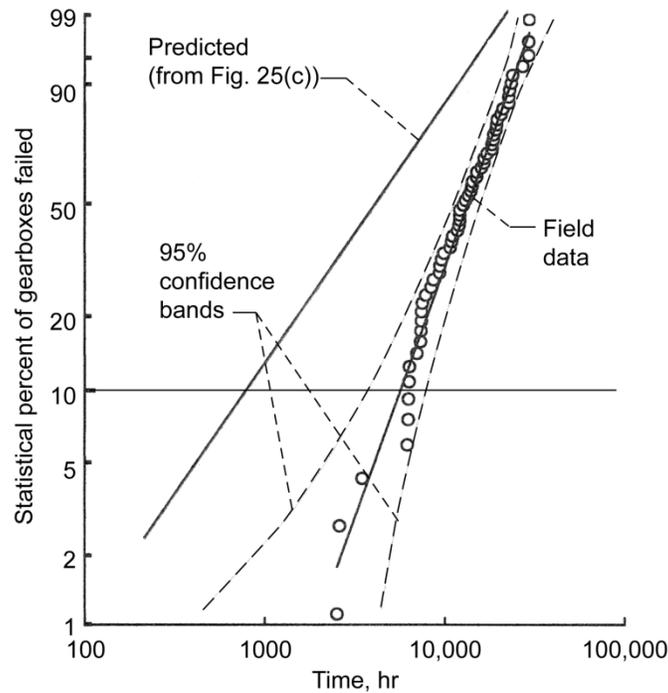


Figure 26. Weibull plot of field data for lives of turboprop gearboxes compared with predicted lives using Lundberg-Palmgren life model. Failure index, 59 out of 64 [87].

With reference to Eq. (48), in their 1952 publication [10], Lundberg and Palmgren proposed a load-life exponent $p = 10/3$ for roller bearings, where one raceway has point contact and the other raceway has line contact. The $10/3$ load-life exponent has been incorporated in the ANSI/ABMA and ISO standards first published in 1953 [12 to 14]. Their assumption of point and line contact may have been correct for many types of roller bearings in use at that time. However, it is no longer the case for most roller bearings manufactured today, and most certainly not for cylindrical roller bearings. The analysis employed for the bearing life calculations used a value of $p = 3$ for ball bearings and $p = 4$ for roller bearings. Poplawski, Peters, and Zaretsky [91,92] suggest the use of $p = 4$ for ball bearings and $p = 5$ for roller bearings [87].

From Eq. (95b), assuming that the bearing system, which is the shortest lived component in the gearbox and has the same Weibull slope as that of the gearbox system (i.e., $e = 2.189$),

$$\frac{1}{L_{sys}^e} = \frac{1}{L_B^e} + \frac{1}{L_G^e} \quad (96a)$$

and substituting in the known values

$$\frac{1}{(5627)^{2.189}} = \frac{1}{L_B^{2.189}} + \frac{1}{(16,680)^{2.5}} \quad (96b)$$

and solving for the bearing system life, results in a value of

$$L_B = 5627 \text{ hr} \quad (96c)$$

From Lundberg-Palmgren [9], the predicted bearing system life is

$$L_B \sim \left(\frac{C_D}{P_{eq}} \right)^4 \sim 774 \text{ hr} \quad (97a)$$

then,

$$\left(\frac{C_D}{P_{eq}} \right) \sim 5.27 \quad (97b)$$

Calculating a revised value for the load-life exponent p for the gearbox bearings based on the actual bearing system life of 5627 hr,

$$\left(\frac{C_D}{P_{eq}} \right)^p \sim (5.27)^p \sim 5627 \text{ hr} \quad (98a)$$

Solving for load-life exponent p ,

$$p = 5.2 \quad (98b)$$

Referring to Eq. (43c), for line contact (roller bearing) the Hertz stress-life exponent $n = 2p$. From Eq. (98b), $n = 10.4$ for line contact for the turboprop gearbox data. Referring to the Zaretsky model for line contact, Eq. (62b), the shear stress-life exponent is

$$c = n - \frac{1}{e} \quad (99a)$$

and where $n = 10.4$ and $e = 2.189$,

$$c = 10.4 - \frac{1}{2.189} = 9.943 \quad (99b)$$

Using the value of c from Eq. (99b) in Eq. (62a) for point contact where e is assumed to equal 1.11, then

$$\begin{aligned} n &= c + \frac{2}{e} = 9.943 + \frac{2}{1.11} \\ &= 11.74 \end{aligned} \quad (100a)$$

and for point contact,

$$p = \frac{n}{3} = \frac{11.74}{3} = 3.91 \quad (100b)$$

The apparent load-life exponent p for the roller bearings is equal to 5.2 and correlates with the Zaretsky model. Were the roller bearing lives to be recalculated using a load-life exponent $p = 5.2$, the predicted L_{10} life of the gearbox would be equal to the actual life obtained in the field, 5627 hr. It should be noted that if an exponent $p = 5$ were used, the predicted L_{10} life of the gearbox would be 4065 hr. This result suggests a strong reliance of the predicted bearing life upon the load-life exponent p . The values of the load-life exponent p for roller bearings equal to $10/3$ from the ANSI/ABMA and ISO standards [12,14] and 4 from computer codes may provide predicted roller bearing lives that are too conservative for design purposes. The use of the ANSI/ABMA and ISO standards load-life exponent of $10/3$ to predict roller bearing life is not reflective of modern roller bearings. It will underpredict bearing lives.

Appendix A – Derivation of Weibull distribution function

According to Weibull [6,7] and as presented in [93] (also see [91]), any distribution function can be written as

$$F(X) = 1 - \exp[-f(X)] \quad (A1)$$

where $F(X)$ is the probability of an event (failure) occurring and $f(X)$ is a function

of an operating variable X . Conversely, from Eq. (A1) the probability of an event not occurring (survival) can be written as

$$1 - F(X) = \exp[-f(X)] \quad (\text{A2a})$$

or

$$1 - F = \exp[-f(X)] \quad (\text{A2b})$$

where $F = F(X)$ and $(1 - F) = S$, the probability of survival.

If there are n independent components, each with a probability of the event (failure) not occurring $(1 - F)$, the probability of the event not occurring in the combined total of all components can be expressed from Eq. (A2b) as

$$1 - F^n = \exp[-nf(X)] \quad (\text{A3})$$

Equation (A3) gives the appropriate mathematical expression for the principle of the weakest link in a chain or, more generally, for the size effect on failures in solids. The application of Eq. (A3) is illustrated by a chain consisting of several links. Testing finds the probability of failure F at any load X applied to a “single” link. To find the probability of failure F_n of a chain consisting of n links, one must assume that if one link has failed the whole chain fails. That is, if any single part of a component fails, the whole component has failed. Accordingly, the probability of nonfailure of the chain $(1 - F_n)$, is equal to the probability of the simultaneous nonfailure of all the links. Thus,

$$1 - F_n = (1 - F)^n \quad (\text{A4a})$$

or

$$S_n = S^n \quad (\text{A4b})$$

Where the probabilities of failure (or survival) of each link are not necessarily equal (i.e., $S_1 \neq S_2 \neq S_3 \neq \dots$), Eq. (A4b) can be expressed as

$$S_n = S_1 \cdot S_2 \cdot S_3 \cdot \dots \quad (\text{A4c})$$

This is the same as Eq. (18) of the main text.

From Eq. (A3) for a uniform distribution of stresses σ throughout a volume V ,

$$F_v = 1 - \exp[-Vf(\sigma)] \quad (\text{A5a})$$

or

$$S = 1 - F_v = \exp[-Vf(\sigma)] \quad (\text{A5b})$$

Equation (A5b) can be expressed as follows:

$$\ln \ln \left[\frac{1}{S} \right] = \ln f(\sigma) + \ln V \quad (\text{A6})$$

It follows that if $\ln \ln (1/S)$ is plotted as the ordinate and $\ln f(\sigma)$ as the abscissa in a system of rectangular coordinates, a variation of volume V of the test specimen will imply only a parallel displacement but no deformation of the distribution function. Weibull [6] assumed the form

$$f(\sigma) = \left[\frac{\sigma - \sigma_u}{\sigma_\beta - \sigma_u} \right]^e \quad (\text{A7})$$

Where e is the Weibull slope, σ is a stress at a given probability of failure, σ_u is a location parameter below which stress no failure will occur, and σ_β is the characteristic stress at which 63.2 percent of the population will fail, Eq. (A6) becomes

$$\ln \ln \left[\frac{1}{S} \right] = e \ln(\sigma - \sigma_u) - e \ln(\sigma_\beta - \sigma_u) + \ln V \quad (\text{A8})$$

If the location parameter σ_u is assumed to be zero, and V is normalized whereby $\ln V$ is zero, Eq. (A8) can be written as

$$\ln \ln \left[\frac{1}{S} \right] = e \ln \left[\frac{\sigma}{\sigma_\beta} \right] \quad \text{where } 0 < \sigma < \infty \text{ and } 0 < S < 1 \quad (\text{A9})$$

Equation (A9) is identical to Eq. (17) of the main text.

The form of Eq. (A9) where σ_u is assumed to be zero is referred to as “two-parameter Weibull.” Where σ_u is not assumed to be zero, the form of the equation is referred to as “three-parameter Weibull.”

Appendix B – Derivation of strict series reliability

As discussed and presented in [91] and [93], G. Lundberg and A. Palmgren [9] in 1947, using the Weibull equation for rolling-element bearing life analysis, first derived the relationship between individual component lives and system life. The following derivation is based on but is not identical to the Lundberg-Palmgren analysis.

Referring to Fig. 1(a), from Eq. (A9) in Appendix A, the Weibull equation can be written as

$$\ln \ln \left[\frac{1}{S_{\text{sys}}} \right] = e \ln \left[\frac{L}{L_{\beta}} \right] \quad (\text{B1})$$

where L is the number of cycles to failure at a given system reliability S_{sys} , and L_{β} is the characteristic life at which 63.2 percent of the population has failed.

Figure 27 is a sketch of multiple Weibull plots where each Weibull plot represents a cumulative distribution of each component in the system. The system Weibull plot represents the combined Weibull plots 1, 2, 3, and so forth. All plots are assumed to have the same Weibull slope e [91]. The slope e can be defined as follows:

$$e = \frac{\ln \ln \left[\frac{1}{S_{\text{sys}}} \right] - \ln \ln \left[\frac{1}{S_{\text{ref}}} \right]}{\ln L - \ln L_{\text{ref}}} \quad (\text{B2a})$$

or

$$\frac{\ln \left[\frac{1}{S_{\text{sys}}} \right]}{\ln \left[\frac{1}{S_{\text{ref}}} \right]} = \left[\frac{L}{L_{\text{ref}}} \right]^e \quad (\text{B2b})$$

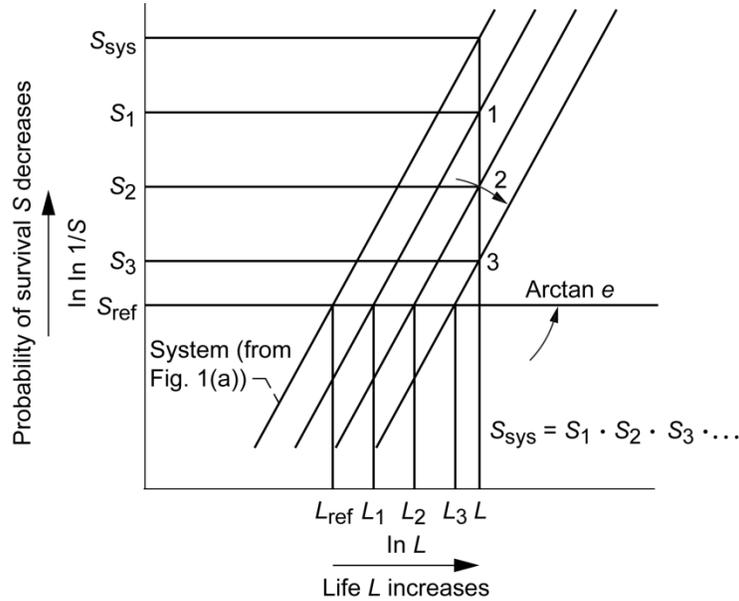


Figure 27. Sketch of multiple Weibull plots where each numbered plot represents cumulative distribution of each component in system and system Weibull plot represents combined distribution of plots 1, 2, 3, etc. (all plots are assumed to have same Weibull slope e) [91].

From Eqs. (B1) and (B2b),

$$\ln \left[\frac{1}{S_{\text{sys}}} \right] = \left[\ln \frac{1}{S_{\text{ref}}} \right] \left[\frac{L}{L_{\text{ref}}} \right]^e = \left[\frac{L}{L_{\beta}} \right]^e \quad (\text{B3})$$

and

$$S_{\text{sys}} = \exp - \left[\frac{L}{L_{\beta}} \right]^e \quad (\text{B4})$$

where $S_{\text{sys}} = S$ in Eq. (B1). For a given time or life L , each component or stressed volume in a system will have a different reliability S . From Eq. (A4c) for a series reliability system

$$S_{\text{sys}} = S_1 \cdot S_2 \cdot S_3 \cdot \dots \quad (\text{B5})$$

Combining Eqs. (B4) and (B5) gives

$$\exp\left[-\left(\frac{L}{L_\beta}\right)^e\right] = \exp\left[-\left(\frac{L}{L_{\beta 1}}\right)^e\right] \times \exp\left[-\left(\frac{L}{L_{\beta 3}}\right)^e\right] \times \dots \quad (\text{B6a})$$

$$\exp\left[-\left(\frac{L}{L_\beta}\right)^e\right] = \exp\left\{-\left[\left(\frac{L}{L_{\beta 1}}\right)^e + \left(\frac{L}{L_{\beta 2}}\right)^e + \left(\frac{L}{L_{\beta 3}}\right)^e + \dots\right]\right\} \quad (\text{B6b})$$

It is assumed that the Weibull slope e is the same for all components. From Eq. (B6b)

$$-\left[\frac{L}{L_\beta}\right]^e = -\left\{\left[\frac{L}{L_{\beta 1}}\right]^e + \left[\frac{L}{L_{\beta 2}}\right]^e + \left[\frac{L}{L_{\beta 3}}\right]^e + \dots\right\} \quad (\text{B7a})$$

Factoring out L from Eq. (B7a) gives

$$\left[\frac{1}{L_\beta}\right]^e = \left[\frac{1}{L_{\beta 1}}\right]^e + \left[\frac{1}{L_{\beta 2}}\right]^e + \left[\frac{1}{L_{\beta 3}}\right]^e + \dots \quad (\text{B7b})$$

From Eq. (B3) the characteristic lives $L_{\beta 1}, L_{\beta 2}, L_{\beta 3}$, etc., can be replaced with the respective lives L_1, L_2, L_3 , etc., at S_{ref} (or the lives of each component that have the same probability of survival S_{ref}) as follows:

$$\left[\ln \frac{1}{S_{\text{ref}}}\right] \left[\frac{1}{L_{\text{ref}}}\right]^e = \left[\ln \frac{1}{S_{\text{ref}}}\right] \left[\frac{1}{L_1}\right]^e + \left[\ln \frac{1}{S_{\text{ref}}}\right] \left[\frac{1}{L_2}\right]^e + \left[\ln \frac{1}{S_{\text{ref}}}\right] \left[\frac{1}{L_3}\right]^e + \dots \quad (\text{B8})$$

where, in general, from Eq. (B3)

$$\left[\frac{1}{L_{\beta}} \right]^e = \left[\ln \frac{1}{S_{\text{ref}}} \right] \left[\frac{1}{L_{\text{ref}}} \right]^e \quad (\text{B9a})$$

and

$$\left[\frac{1}{L_{\beta 1}} \right]^e = \left[\ln \frac{1}{S_{\text{ref}}} \right] \left[\frac{1}{L_1} \right]^e, \text{ etc.} \quad (\text{B9b})$$

Factoring out $\ln (1/S_{\text{ref}})$ from Eq. (B8) gives

$$\left[\frac{1}{L_{\text{ref}}} \right]^e = \left\{ \left[\frac{1}{L_1} \right]^e + \left[\frac{1}{L_2} \right]^e + \left[\frac{1}{L_3} \right]^e + \dots \right\}^{1/e} \quad (\text{B10})$$

or rewriting Eq. (B10) results in

$$\left[\frac{1}{L} \right]^e = \sum_{i=1}^n \left[\frac{1}{L_i} \right]^e \quad (\text{B11})$$

Equation (B11) is identical to Eqs. (47) and (49a) of the text.

Appendix C – Contact (Hertz) stress

The contact (Hertz) stresses at the respective races of a bearing are a function of the bearing geometry, the normal load at the contact, and the elastic properties of the bearing materials. Jones [29] relates the Hertz contact theory for the stresses of nonconforming bodies in contact for both ball and roller bearings. From Jones, the following relations for the maximum Hertz stresses S_{max} at the inner and outer races of ball bearings can be derived:

1. For deep-groove ball bearings with a radial load only,

$$S_{\max_{or}} = \frac{K \left[\frac{2}{D_{or}} + \frac{4}{d} - \frac{1}{f_{or}d} \right]^{2/3} P_{\max}^{1/3}}{\mu\nu} \quad (\text{C1a})$$

for the outer race, and

$$S_{\max_{ir}} = \frac{K \left[\frac{2}{D_{ir}} + \frac{4}{d} - \frac{1}{f_{ir}d} \right]^{2/3} P_{\max}^{1/3}}{\mu\nu} \quad (\text{C1b})$$

for the inner race, where

$$P_{\max} = W_r \frac{5}{Z} \quad (\text{C2})$$

2. For angular-contact ball bearings with a thrust load only,

$$S_{\max_{or}} = \frac{K \left[\frac{2 \cos \beta}{d_e + d \cos \beta} + \frac{4}{d} - \frac{1}{f_{or}d} \right]^{2/3} P_N^{1/3}}{\mu\nu} \quad (\text{C3a})$$

for the outer race, and

$$S_{\max_{ir}} = \frac{K \left[\frac{2 \cos \beta}{d_e - d \cos \beta} + \frac{4}{d} - \frac{1}{f_{ir}d} \right]^{2/3} P_N^{1/3}}{\mu\nu} \quad (\text{C3b})$$

for the inner race, where

$$P_N = \frac{W_t}{Z} \frac{1}{\sin \beta} \quad (\text{C4})$$

3. For roller bearings with a radial load only

$$S_{\max_{or}} = K \sqrt{\frac{P_{\max}}{\ell} \left(\frac{-2}{D_{or}} + \frac{2}{d} \right)} \quad (\text{C5a})$$

for the outer race, and

$$S_{\max_{ir}} = K \sqrt{\frac{P_{\max}}{\ell} \left(\frac{2}{D_{ir}} + \frac{2}{d} \right)} \quad (\text{C5b})$$

for the inner race, where

$$P_{\max} = W_r \frac{4}{Z} \quad (\text{C6})$$

for

D_{or}	outer-race diameter
D_{ir}	inner-race diameter
d	ball diameter
d_e	pitch diameter
f_{or}, f_{ir}	outer- and inner-race conformity, respectively
W_r	radial load
P_{\max}	normal load on maximum loaded ball
μ, ν	transcendental functions [29]
K	constant based on elastic properties of bearing steel
P_N	normal ball load
W_t	bearing thrust load
β	contact angle
ℓ	effective roller length

For bearing steel on bearing steel, $K = 1.58 \times 10^{-3}$ for S_{\max} in GPa and $K = 23.58$ for S_{\max} in ksi. The values for the transcendental functions μ and ν vary with conformity f and are found in [29]. Values of their product $\mu\nu$ can be found in Table 4.

References

1. Stribeck, R., 1900. Reports from the central laboratory for scientific investigation. Translation by H. Hess, 1907, *ASME Trans.*, **29**, 420–466.
2. Goodman, J., 1912. Roller and ball bearings. *Proc. Inst. Civil Engrs.*, **189**, 82–166.
3. Colvin, F.H., and Stanley, F.A., 1914. *American Machinist Handbook and Dictionary of Shop Terms*, 2nd ed., McGraw-Hill, New York, pp. 381–386.
4. Palmgren, A., 1924. Die Lebensdauer von Kugellagern (The Service Life of Ball Bearings). *Zeitschrift des Vereines Deutscher Ingenieure*, **68**, 14, 339–341 (NASA TT–F–13460, 1971).
5. Palmgren, A., 1945. *Ball and Roller Bearing Engineering*, 1st ed., Translation by G. Palmgren and B. Ruley, SKF Industries, Philadelphia, PA.
6. Weibull, W., 1939. A statistical theory of the strength of materials. *Ingeniors Vetenskaps Adademien* (Proc. Royal Swedish Academy of Engr.), 151.
7. Weibull, W., 1939. The phenomenon of rupture. *Ingeniors Vetenskaps Adademien* (Proc. Royal Swedish Academy of Engr.), 153.
8. Thomas, H.R., and Hoersch, V.A., 1930. Stresses due to the pressure of one elastic solid upon another with special reference to railroad wheels. Bulletin 212, Engineering Experimental Station, Univ. of Illinois, Urbana.
9. Lundberg, G., and Palmgren, A., 1947. Dynamic capacity of rolling bearings. *Acta Polytechnica Mechanical Engineering Series*, **1**, 3, Stockholm, Sweden.
10. Lundberg, G., and Palmgren, A., 1952. Dynamic capacity of roller bearings. *Acta Polytechnica Mechanical Engineering Series*, **2**, 4, Stockholm, Sweden.
11. Hertz, H., 1881. Uber die beruehrung elastischer koerper (On contact of elastic bodies). *Gesammelte Werke (Collected Works)*, 1, Leipzig, Germany.
12. ISO 281:1990(E), 1990. Rolling bearing – Dynamic load ratios and rating life. International Organization for Standards, Geneva.
13. ANSI/AFBMA 9–1990, 1990. Load rating and fatigue life for ball bearings. The Anti-Friction Bearing Manufacturers Association, Washington, DC, USA.
14. ANSI/AFBMA 11–1990, 1990. Load rating and fatigue life for roller bearings. The Anti-Friction Bearing Manufacturers Association, Washington, DC, USA.
15. Zaretsky, E.V., 1997. *Tribology for Aerospace Applications*, STLE SP–37, Society of Tribologists and Lubrication Engineers, Park Ridge, IL, USA.
16. Ertel, A.M., 1939. Hydrodynamic lubrication based on new principles. *Prikadnaya Matematika I Mechanika*, **3**, 2 (in Russian).
17. Grubin, A.N., 1949. Fundamentals of the hydrodynamic theory of lubrication of heavily loaded cylindrical surfaces. *Investigation of the Contact Machine Components*, Kh. F. Ketova, ed., translation of Russian Book No. 30, Central Scientific Institute of Technology and Mechanical Engineering, Moscow (Available from Dept. of Scientific and Industrial Research, Great Britain, Trans. CTS–235, and from Special Libraries Association, Chicago. Transl. R-3554).
18. Zaretsky, E.V., 1992. *STLE life factors for rolling bearings*, STLE SP–34, Society of Tribologists and Lubrication Engineers, Park Ridge, IL, USA.

19. Styri, H., 1951. Investigators of rolling bearings at SKF Industries, Inc., Review of current and anticipated lubricant problems in turbojet engines. NACA Subcommittee on Lubrication and Wear, NACA RM 51D20, pp. 30–39.
20. Zaretsky, E.V., 1998. A Palmgren revisited – A basis for bearing life prediction. *Lubrication Engineering, J. STLE*, **54**(2), 18–24.
21. Anderson, W.J., 1964. Fatigue in rolling-element bearings. Advanced Bearing Technology, E.E. Bisson and W.J. Anderson, ed., NASA SP–38, pp. 371–450.
22. Langer, B.F., 1937. Fatigue failure from stress cycles of varying amplitude. *ASME J. Applied Mechanics*, **59**, A160–A162.
23. Miner, M.A., 1945. Cumulative damage in fatigue. *ASME J. Applied Mechanics*, **12**(3), A159–A164.
24. Johnson, L.G., 1964. *The Statistical Treatment of Fatigue Experiments*. Elsevier Publishing Co., Amsterdam, The Netherlands.
25. Tallian, T.E., 1962. Weibull distribution of rolling contact fatigue and deviations therefrom. *ASLE Trans.*, **5**(1), 183–196.
26. Weibull, W., 1951. A statistical distribution of wide applicability. *J. Applied Mechanics*, **18**(3), 292–297.
27. Weibull, W., 1962. Efficient methods for estimating fatigue life distribution of rolling bearings. *Rolling Contact Phenomenon*, (ed.: J.B. Bidwell), Elsevier, New York, NY, USA, pp. 252–265.
28. Zaretsky, E.V., Poplawski, J.V., and Peters, S.M., 1996. Comparison of life theories for rolling-element bearings. *STLE Trans.*, **39**(2), 237–248.
29. Jones, A.B., 1946. New departure – Analysis of stresses and deflections, vols. 1 and 2, New Departure Div., Gen. Motors Corp., USA.
30. Palmgren, A., 1959. *Ball and Roller Bearing Engineering*, 3rd ed., Translation by G. Palmgren and A. Palmgren, SKF Industries, Philadelphia, PA, USA.
31. Zaretsky, E.V., Poplawski, J.V., and Root, L.E., 2007. Reexamination of ball-race conformity effects on ball bearing life. *Trib. Trans.*, **50**(3), 336–349.
32. Ioannides, E., and Harris, T.A., 1985. A new fatigue life model for rolling bearings. *J. Tribol., Trans. ASME*, **107**(3), 367–378.
33. Tosha, K., Ueda, D., Shimoda, H., and Shimizu, S., 2008. A study on P–S–N curve for rotating bending fatigue test for bearing steel. *Tribology Trans.*, **46**(4), 166–172.
34. ASME Tribology Division, 2003, Life Ratings for Modern Rolling Bearings: A Design Guide for the Application of International Standard ISO 281/2, *American Society of Mechanical Engineers*, New York, NY, USA.
35. ISO 281:1990/Amd 1 and 2:2000, 2002. *Rolling Bearings – Dynamic Load Ratings*. International Organization for Standardization, Geneva, Switzerland.
36. ISO 281:2007, 2007. *Rolling Bearings – Dynamic Load Ratings and Rating Life*. International Organization for Standardization, Geneva, Switzerland.
37. Ioannides, E., Bergling, G., and Gabelli, A., 1999. An analytical formulation for the life of rolling bearings. *Acta Polytechnica Scandinavica*, Mechanical Engineering Series, **137**, Finland.
38. Zaretsky, E.V., 1987. Fatigue criterion to system design, life and reliability. *AIAA Journal of Propulsion and Power*, **107**(3), 76–83.
39. Zaretsky, E.V., 1994. Design for life, plan for death. *Machine Design*, **66**(15), 55–59.

40. Vlcek, B.L., Hendricks, R.C., and Zaretsky, E.V., 2003. Determination of rolling-element fatigue life from computer generated bearing tests. *Tribology Trans.*, **46**(4), 479–493.
41. Zaretsky, E.V., Anderson, W.J., and Parker, R.J., 1962. The effect of contact angle on rolling-contact fatigue and bearing load capacity. *ASLE Trans.*, **5**(1), 210–219.
42. Takata, H., 1992. Fatigue life theory of rolling bearings considering the fatigue life of rolling elements. *Japanese J. Tribology*, **37**(12), 1605–1621.
43. Zaretsky, E.V., Poplawski, J.V., and Root, L.E., 2008. Relation between Hertz stress-life exponent, ball-race conformity, and ball bearing life. *Tribol. Trans.*, **51**(2), 150–159.
44. Parker, R.J., and Zaretsky, E.V., 1972. Reevaluation of the stress life relation in rolling-element bearings. NASA TN D-6745.
45. Styri, H., 1951. Fatigue strength of ball bearing races and heat treated 52100 steel specimens. *Proc. American Society for Testing and Materials*, **51**, 682–700.
46. Cordiano, H.V., Cochran, E.P., Jr., and Wolfe, R.J., 1956. A study of combustion resistant hydraulic fluids as ball bearing lubricants. *Lubrication Engineering*, **12**(4), 261–266.
47. McKelvey, R.E., and Moyer, C.A., 1963. The relation between critical maximum compressive stress and fatigue life under rolling contact. Paper 1, presented at the Institution of Mechanical Engineers Symposium on Fatigue in Rolling Contact, Mar. 28, 1963.
48. Townsend, D.P., Coy, J.J., and Zaretsky, E.V., 1978. Experimental and analytical load-life relation for AISI 9310 steel spur gears. *J. Mechanical Design*, **100**(1), 54–60.
49. Barwell, F.T., and Scott, D., 1956. Effect of lubricant on pitting failure of ball Bearings. *Engineering*, **182**, 4713, 9–12.
50. Butler, R.H., and Carter, T.L., 1957. Stress-life relation of the rolling-contact fatigue spin rig. NACA TN 3930.
51. Baughman, R.A., 1958. Experimental laboratory studies of bearing fatigue. ASME paper 58-A-235.
52. Scott, D., 1958. Lubricants at higher temperatures: assessing the effects on ball bearing failures. *Engineering*, **185**, 4811, 660–662.
53. Utsmi, T., and Okamoto, J., 1960. Effect of surface roughness on the rolling fatigue life of bearing steels. *J. Japan Society of Lubrication Engineers*, **5**(5), 291–296.
54. Greenert, W.J., 1962. The toroid contact roller test as applied to the study of bearing materials. *J. Basic Engineering*, **84**(1), 181–191.
55. Valori, R.R., Sibley, L.B., and Tallian, T.E. 1965. Elastohydrodynamic film effects on the load-life behavior of rolling contacts. ASME paper 65-LUBS-11.
56. Schatzberg, P., and Felsen, I.M., 1969. Influence of water on fatigue-failure location and surface alteration during rolling-contact lubrication. *J. Lubrication Technology*, **91**(2), 301–307.
57. Zaretsky, E.V., and Parker, R.J., 1967. Discussion to paper “On competing failure modes in rolling contact” by T.E. Tallian, *ASLE Trans.*, **10**(4), 436–437.
58. Lorosch, H.K., 1982. Influence of load on the magnitude of the life exponent for rolling bearings. *Rolling Contact Fatigue Testing of Bearing Steels* (ed.: J.J.C. Hoo), ASTM STP-771, American Society for Testing and Materials, Philadelphia, PA, USA, pp. 275–292.

59. Zwirlein, O., and Schlicht, H., 1982. *Rolling Contact Fatigue Testing of Bearing Steels* (ed.: J.J.C. Hoo), ASTM STP-771, American Society for Testing and Materials, Philadelphia, PA, USA, pp. 358-379.
60. Bamberger, E.N., and Signer, H., 1976. Endurance and failure characteristics of main-shaft jet engine bearing at 3×10^6 DN. *J. Lubrication Technology*, **98**(4), 580-585.
61. Parker, R.J., Zaretsky, E.V., and Bamberger, E.N., 1974. Evaluation of load-life relation with ball bearings at 500 °F. *J. Lubrication Technology*, **96**(3), 391-397.
62. Bamberger, E.N., and Zaretsky, E.V., 1971. Fatigue lives at 600 °F of 120-millimeter-bore ball bearings of AISI M-50, AISI, M-1, and WB-49 Steels. NASA TN D-6156.
63. Bamberger, E.N., Zaretsky, E.V., and Anderson, W.J., 1970. Effect of three advanced lubricants on high-temperature bearing life. *J. Lubrication Technology*, **92**(1), 23-33.
64. Zaretsky, E.V., Anderson, W.J., and Bamberger, E.N., 1969. *Rolling-Element Bearing Life from 400 to 600 °F*. NASA TN D-5002.
65. Townsend, D.P., Bamberger, E.N., and Zaretsky, E.V., 1976. A life study of ausforged, standard forged, and standard machined AISI M-50 spur gears. *J. Lubrication Technology*, **98**(3), 418-425.
66. Koistinen, D.P., 1964. The generation of residual compressive stresses in the surface layers of through-hardened steel components by heat treatment. *ASM Trans.*, **57**, 581-588.
67. Stickels, C.A., and Janotik, A.M., 1980. Controlling residual stresses in 52100 bearing steel by heat treatment. *Metallurgical Trans.* **11**(3), 467-473.
68. Almen, J.O., 1962. Effects of residual stress on rolling bodies. *Rolling Contact Phenomena* (ed.: J.B. Bidwell), Elsevier, pp. 400-424.
69. Gentile, A.J., Jordan, E.F., and Martin, A.D., 1965. Phase transformations in high-carbon, high-hardness steels under contact loads. *AIME Trans.*, **233**(6), 1085-1093.
70. Gentile, A.J., and Martin, A.D., 1965. The effect of prior metallurgically induced compressive residual stress on the metallurgical and endurance properties of overload-tested ball bearings. ASME paper 65-WA/CF-7.
71. Scott, R.L., Kepple, R.K., and Miller, M.H., 1962. The effect of processing-induced near-surface residual stress on ball bearing fatigue. *Rolling Contact Phenomena* (ed.: J. B. Bidwell), Elsevier, pp. 301-316.
72. Naisong, X., Stickels, C.A., and Peters, C.R., 1984. The effect of furnace atmosphere carbon potential on the development of residual stresses in 52100 steel. *Metallurgical Trans. A*, **15**(11), 2101-2102.
73. Jones, A.B., 1947. Metallurgical observations of ball bearing fatigue phenomena. *Symposium on Testing of Bearings*, American Society for Testing and Materials, Philadelphia, PA., USA, pp. 35-52.
74. Carter, T.L., 1960. A Study of Some Factors Affecting Rolling-Contact Fatigue Life. NASA TR R-60.
75. Akaoka, J., 1962. Some considerations relating to plastic deformation under rolling contact. *Rolling Contact Phenomena* (ed.: J.B. Bidwell), Elsevier, pp. 266-300.
76. Zaretsky, E.V., Parker, R.J., and Anderson, W.J., 1969. A study of residual stress induced during rolling. *J. Lubrication Technology*, **91**(2), 314-319.

77. Zaretsky, E.V., Parker, R.J., Anderson, W.J., and Miller, S.T., 1965. Effect of Component Differential Hardness on Residual Stress and Rolling-Contact Fatigue. NASA TN D-2664.
78. Foord, C.A., Hingley, C.G., and Cameron, A., 1969. Pitting of steel under varying speeds and combined stresses. *J. Lubrication Technology*, **91**(2), 282–293.
79. Cioclov, D., 1969, Discussion (pp. 290–291) of Foord, C. A., Hingley, C. G., and Cameron, A. 1969, “Pitting of Steel Under Varying Speeds and Combined Stresses,” *J. Lubrication Technology*, **91**(2), 282–293.
80. Coe, H.H., and Zaretsky, E.V., 1987. Effect of interference fits on roller bearing fatigue life. *ASLE Trans.*, **30**(2), 131–140.
81. Czyzewski, T., 1975. Influence of a tension stress field introduced in the elastohydrodynamic contact zone on rolling-contact fatigue. *Wear*, **34**(2), 201–214.
82. Zaretsky, E.V., Poplawski, J.V., and Miller, C.R., 2001. Rolling bearing life prediction – Past, present, and future. In *Proc. of the International Tribology Conf. Nagasaki, 2000*, Vol. 1, Japanese Society of Tribologists, Tokyo, Japan, pp. 101–107.
83. Vlcek, B.L., Hendricks, R.C., and Zaretsky, E.V., 2003. Determination of rolling-element fatigue life from computer generated bearing tests. *STLE Tribology Trans.*, **46**(3), 479–493.
84. Harris, T.A., 1995, “Final report-establishment of a new rolling bearing contact life calculation method. U.S. Naval Air Warfare Center, Aircraft Division Trenton, Contact No. N68335-93-C-0111.
85. Harris, T.A., and McCool, J.J., 1996. On the accuracy of rolling bearing fatigue life prediction. *ASME J. Tribology*, **118**(2), 297–310.
86. Lewicki, D.G., Black, J.D., Savage, M., and Coy, J.J., 1986. Fatigue life analysis of a turboprop reduction gearbox. *J. Mech. Trans. Autom. Des.*, *Trans. ASME*, **108**(2), 225–262.
87. Zaretsky, E.V., Lewicki, D.G., Savage, M., and Vlcek, B.L., 2007. Determination of turboprop reduction gearbox system fatigue life and reliability. *Tribol. Trans.*, **50**(4), 507–516.
88. Coy, J.J., Townsend, D.P., and Zaretsky, E.V., 1975. Analysis of Dynamic Capacity of Low-Contact-Ratio Spur Gears Using Lundberg-Palmgren Theory. NASA TN D-8029.
89. Coy, J.J., and Zaretsky, E.V., 1975. Life Analysis of Helical Gear Sets Using Lundberg-Palmgren Theory. NASA TN D-8045.
90. Coy, J.J., Townsend, D.P., and Zaretsky, E.V., 1983. An Update on the Life Analysis of Spur Gears. In *Advanced Power Transmission Technology*, NASA, Cleveland, OH, USA, NASA CP-2210, pp. 421–434.
91. Poplawski, J.V., Peters, S.M., and Zaretsky, E.V., 2001. Effect of roller profile on cylindrical roller bearing life prediction – Part I: Comparison of bearing life theories, *STLE Tribology Trans.*, **44**(3), 339–350.
92. Poplawski, J.V., Peters, S.M., and Zaretsky, E.V., 2001. Effect of roller profile on cylindrical roller bearing life prediction – Part II: Comparison of roller profiles, **44**(3), 417–427.
93. Melis, M.E., Zaretsky, E.V., and August, R., 1999. Probabilistic analysis of aircraft gas turbine disk life and reliability. *J. Propulsion and Power, Trans. AIAA*, **15**, 658–666.



Research Signpost
37/661 (2), Fort P.O.
Trivandrum-695 023
Kerala, India

Recent Developments in Wear Prevention, Friction and Lubrication, 2010: 137-157
ISBN: 978-81-308-0377-7 Editor: George K. Nikas

3. Laser surface texturing and applications

Izhak Etsion

*Yeshayahu Winograd Chair in Fluid Mechanics and Heat Transfer
Mechanical Engineering Department, Technion – Israel Institute of Technology, Israel*

Abstract. Surface texturing has emerged in the last decade as a viable option of surface engineering resulting in significant improvement in load capacity, wear resistance, friction coefficient etc. of tribological mechanical components. Various techniques can be employed for surface texturing but Laser Surface Texturing (LST) is probably the most advanced so far. LST produces a very large number of micro-dimples on the surface and each of these micro-dimples can serve either as a micro-hydrodynamic bearing in cases of full or mixed lubrication, a micro-reservoir for lubricant in cases of starved lubrication conditions, or a micro-trap for wear debris in either lubricated or dry sliding. The present article reviews the current effort being made world wide on laser surface texturing and the potential of this technology in various tribological applications.

1. Introduction

Surface texturing as a means for enhancing tribological properties of mechanical components is well known for many years. Perhaps the most familiar and earliest commercial application of surface texturing is that of cylinder liner honing. Today surfaces of modern magnetic storage devices are commonly textured and surface texturing is also considered as a means for overcoming adhesion and stiction in MEMS devices. Fundamental research work on various forms and shapes of surface texturing for tribological applications is carried out worldwide and various texturing techniques are employed in these studies including machining, ion beam texturing, etching techniques and laser texturing. Of all the practical micro-surface patterning methods it seems that laser surface texturing (LST) offers the most promising concept. This is because the laser is extremely fast and allows short processing times, it is clean to the environment and provides excellent control of the shape and size of the texture, which allows realization of optimum designs. By controlling energy density, the laser can safely

process hardened steels, ceramics, and polymers as well as crystalline structures. Indeed, LST is starting to gain more and more attention in the Tribology community as is evident from the growing number of publications on this subject. LST produces a very large number of micro-dimples on the surface (see Fig. 1) and each of these micro-dimples can serve either as a micro-hydrodynamic bearing in cases of full or mixed lubrication, a micro-reservoir for lubricant in cases of starved lubrication conditions, or a micro-trap for wear debris in either lubricated or dry sliding.

The pioneering work on LST started at Technion in Israel as early as 1996 [1, 2]. At about the same time work on laser surface texturing was done in Germany but unfortunately, most of it is published in the German language and hence, is not even referenced in English archive journals. A few exceptions are papers coming from the group lead by Geiger at the University of Erlangen-Nuremberg e.g. [3, 4].

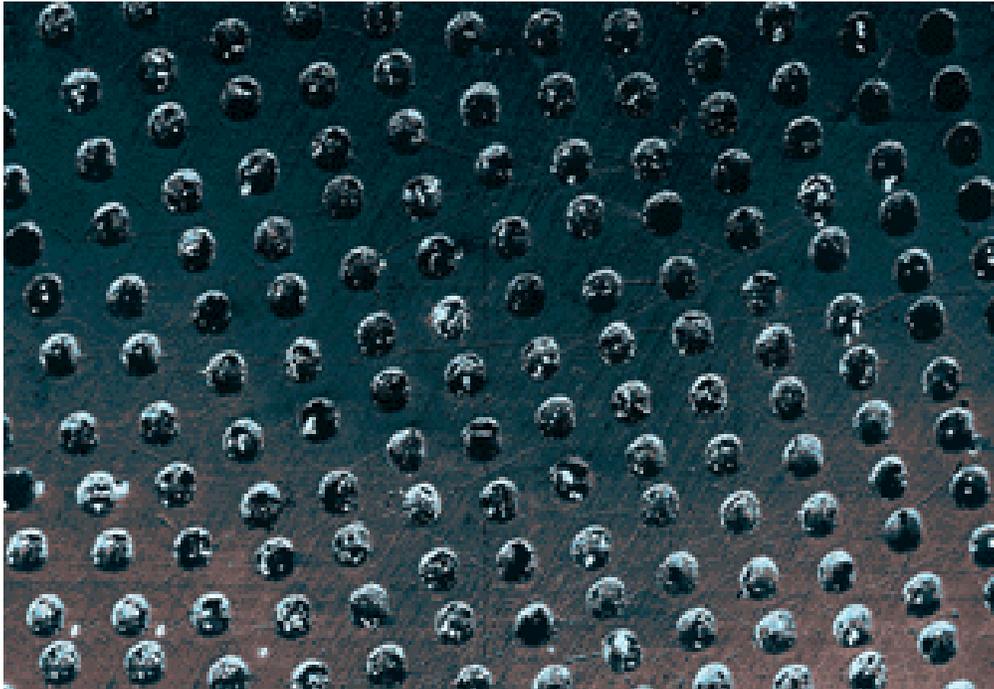


Figure 1. LST regular micro-surface structure in the form of micro-dimples.

This group uses an excimer laser with a mask projection technique, a mask is illuminated with the laser beam and its geometrical information is projected onto the textured surface. This method was applied to a punch, used in a backward cup extrusion process for the production of rivets, and showed a substantial increase of up to 169% in cold forging tool life. These as well as many other papers on LST are described in a review of the state of the art of LST covering this subject until 2005 [5]. In the next sections the work that was done on LST prior to 2005 will be

described briefly followed by the new developments since 2005. Laser surface texturing has been used in the magnetic storage industry [6, 7] mainly to prevent stiction during start up. This issue will not be dealt with in the present review. Instead, the potential of LST in enhancing Tribological performance during continuous operation will be described.

2. LST prior to 2005

Laser was used at Tohoku University, Japan [8] to texture SiC surfaces for studying the effect of LST on the transition from hydrodynamic to mixed lubrication regime. An extensive research work on laser surface texturing was done at the Institute of Applied Physics of the University of Bern in Switzerland utilizing Q-switched Nd:YAG but mostly femtosecond lasers [9–13]. A fundamental research work on LST was carried out at Argonne National Lab. in the USA to study the effect of LST on the transition from boundary to hydrodynamic lubrication regime [14].

By far most of the work on LST prior to 2005 was done on dynamic seals. The earlier simple modeling [1] and experiments [2] of LST in mechanical seals were followed by more in-depth theoretical and experimental studies [15]. It was found that the actual shape of the micro dimple does not play a significant role and that the most important parameter for optimum load capacity is the ratio of the dimple depth over diameter. A high stiffness of the fluid film below a clearance of 1 μm and a very good agreement between theory and experiment was shown in [15]. Further testing of actual seals in water [16] showed dramatic reduction of up to 65% in friction torque. Similar results of lower friction and face temperature with laser textured seal face were found in East China University of Science and Technology [17] where textured SiC rings were tested against Carbon rings in oil. In all these cases full LST was used meaning that the texturing covered the full width of the sealing dam. It was found that with full LST the reduction in friction torque is gradually diminishing at higher sealed pressures. To overcome the poor performance at high pressures a special treatment was developed that enhances hydrostatic effects in high-pressure seals [18]. This treatment consists of applying higher density LST over a portion of the sealing dam adjacent to the high-pressure side and leaving the remaining portion non-textured (partial LST). The textured portion provides an equivalent larger gap so that the end result is a converging seal gap in the direction of pressure drop, which produces hydrostatic effect. A Standard commercial seal that is rated to a maximum pressure of 11 bar could be easily operated up to 23 bar when textured with the partial LST providing high pressure sealing capability that is substantially greater than that of the standard

non-textured one. Another study [19], on both full and partial LST seals demonstrated the potential positive effect of micro-surface texturing on reducing breakaway torque and blister formation in carbon-graphite mechanical seal faces. The LST advantages are not limited to liquid lubrication only, and dry gas seals can benefit from LST as well [20, 21].

Laser surface texturing for other lubricated applications was also investigated prior to 2005. This was done mainly for piston rings [22, 23] where optimum texturing parameters for minimum friction force were found for full LST rings showing a potential reduction of about 30 percent compared to non-textured rings under full lubrication conditions. The use of laser texturing in the form of micro-grooves on cylinder liners of internal combustion engines was presented at the 14th International Colloquium Tribology in Esslingen Germany [24] showing lower fuel consumption and wear. This technique called “laser honing” is commercially available from the Gehring Company in Germany [25].

Analysis of LST in hydrodynamic thrust bearings of the simplest form of parallel sliding disks [26] has shown the potential of LST in this application. It was found that partial LST can improve substantially the load carrying capacity of these simple bearings and make them comparable to more sophisticated tapered or stepped sliders. Test results in water [27] showed that textured bearings operated with a clearance that is about 3 times larger and friction that is about 3 times smaller than non-textured bearings.

Laser texturing is also used extensively in metal forming as a mean for a secondary hydrodynamic lubrication mechanism which is called micro-pool or micro-plastic hydrodynamic lubrication [28].

The potential benefit of LST in providing micro-traps for wear debris in dry contacts subjected to fretting has been demonstrated in [29] and [30]. The results in [29] showed that the escape of oxide wear debris into the LST micro-dimples resulted in up to 84% reduction in electrical contact resistance of textured fretting surfaces compared to the case with non-textured surfaces. Eventually, the dimples may fill-up with wear debris but the useful life of the LST device would be substantially prolonged. The potential effect of LST on fretting fatigue life was demonstrated in [30] showing improved fretting fatigue resistance and almost doubled fretting fatigue life.

3. LST since 2005

In the period from 2005 through 2007, a growing number of publications on surface texturing appeared in the literature, inspired by the LST development prior to that period. The validity of the Reynolds equation when applied to textured

features that have large aspect ratio (the ratio of depth over diameter or width) was questioned and several studies were made to resolve this problem. The Navier-Stokes (NS) equations were solved [31], using a commercial CFD code, for two geometries, cylindrical and spline, of infinitely long single groove in parallel sliding relative to a smooth wall in the presence of incompressible fluid. It was found that fluid inertia was the main contributor to load carrying capacity, which increases with increasing depth and width of the groove. Above a certain aspect ratio a vortex appears in the groove and the load carrying capacity saturates. The groove also reduces the friction between the sliding parallel walls. Similar technique was used in [32] to study the effect of a single rectangular groove for the case of an infinitely long linear convergent slider bearing, and of a 2D pocket for the case of a square bearing pad. It was found that cavitation at the pocket inlet occurs only at very low bearing convergence ratios. The closed pocket can reduce the friction coefficient both at high and low convergence ratios due to its effect on load carrying capacity or on friction force. The pressure distribution and load carrying capacity for a single 3D dimple representing the LST and facing a parallel surface was studied in [33]. Both the full NS equations (using a commercial CFD code) and the Reynolds equation were solved for the case of a compressible fluid at no sliding but with a pressure differential to simulate a hydrostatic gas seal. Comparison between the two solution methods illustrates that in spite of potential large differences in local pressures the differences in load carrying capacity are small for realistic geometrical parameters of LST. Hence, the Reynolds equation can be safely used for most LST applications.

One of the main problems in theoretical modeling of surface texturing effects is the need to deal with a very large number of textured dimples or other features that may consume large computing times. Hence, homogenization techniques may be very helpful in easing this burden. A mathematical analysis based on combination of homogenization techniques and perturbation analysis was presented in [34] to study the effect of periodic textures on the static characteristic of infinitely wide convergent thrust bearings. A multiscale method for modeling surface texture effects in a mixed lubrication journal bearing model was presented in [35]. The local (micro) flow effects for a single surface pocket were analyzed using the NS equations and flow factors were derived that can then be added to the macroscopic smooth flow problem that is modeled by the 2D Reynolds equation. The analysis also accounts for pocket squeeze effect due to surface deformation.

Texturing optimization was studied in [36] using numerically generated textured surfaces technique that was named "virtual texturing". The method was used for preliminary exploration of the relationship between a dimpled texture design, typical of LST, and the mixed lubrication characteristic for a counterformal contact. An interesting optimization technique was presented in [37] where the

Reynolds equation was solved for several micro-textured slider bearing configurations. The dimples were square and arranged in a square pattern. It was shown that non uniform texture provides better performance than the commonly used uniform one.

General experimenting with LST became also popular in the period since 2005. Unfortunately, most of these general experiments were not based on previous theoretical findings and were therefore performed using a trial and error approach when attempting to identify optimum LST parameters. Different test configurations were used in these experiments with different materials and lubricants. Mostly Nd:YAG laser was used for these general experiments [38-41] but also a much shorter pulse femtosecond laser [42]. In [38] a cylinder was reciprocated along its axis against smooth or textured flat plates in distilled water. In [39] a pin-on disk machine was used in unidirectional sliding. The pin was a steel ball with a flat contact area to simulate conformal contact. The disks were polished, ground, and textured with various LST parameters, and tests were performed with both low and high viscosity oils. LST was observed to expand the range of hydrodynamic lubrication regime in terms of load and sliding speed for both high and low viscosity lubricants. Furthermore, a substantial reduction of the friction coefficient in boundary lubrication regime was obtained with LST compared to untextured surfaces under similar operation conditions. In [40] a steel ball was used in oscillating linear sliding against a steel disk that was either polished or textured with different LST parameters. Tests were performed with three different metal working oils. The tests in [41] were performed with a reciprocating flat pin against ceramics and steel plates in distilled water. The plates were textured with different grooves and dimples. The femtosecond laser [42] produces clean texture without the common raised material at the edge of textured features caused by the Nd:YAG laser, which in many cases, require post LST lapping to remove the raised bulges. However, much larger number of pulsed incidents is needed to obtain about 10 μm deep features compared to the Nd:YAG laser. It therefore, took some 25-30 minutes to texture an area of $8 \times 8 \text{ mm}^2$ while with the Nd:YAG laser the required time would be at least an order of magnitude shorter. The tests in [42] were performed with a reciprocating cylinder on flat surfaces that were either smooth or textured and lubricated with oil. A common finding in all these experiments that were carried out in Germany, USA, Finland and Switzerland [38-42] is that textured surfaces have improved tribological performance in terms of friction and wear compared to untextured surfaces under the same test conditions.

A very interesting technique based on interfering laser beams [43] allows minimization of the textured surface features down to the micrometer or even sub-micrometer scale. This laser interference direct structuring can therefore produce

lateral feature sizes from sub-micron up to several micrometers by making minor changes in the optical system. What is even more interesting in this technique is the ability to combine topographic texturing of surfaces with micro structural changes for a hierarchical order that can further improve tribological performance. Some applications like centrifuge cast die mold, computer hard disk and automotive engine block, where this technique was implemented are mentioned in [43].

Some additional techniques other than LST were also described in the literature since 2005 as options for surface texturing. These include: Cr-N coating with a randomly crater-like topography [44], diamond embossing tool to create large array of small indents in metallic surfaces [45], and impulse indentation where special ending act as hammers to form oil pockets on metal surfaces [46].

4. Applications

Like in the period prior to 2005 several applications of LST were considered in the years to follow. These include mainly automotive applications such as in-cylinder friction reduction, various types of bearings, seals, elasto-hydrodynamic (EHD) lubrication, magnetic storage and a few others. These applications will be described next.

4.1. Automotive

The early work on LST piston rings [22, 23] considered full width texturing but very soon it was realized that partial texturing may be more beneficial. The difference between these two types of LST is demonstrated in Fig. 2 and the rationale for the better performance of partial LST is fully described in Ref. [26]. Basically, in the full width LST the area density of the dimples is relatively small and each dimple acts individually as a micro-hydrodynamic bearing with negligible interaction between neighboring dimples. In the partial LST case the dimple area density is higher and the dimples act collectively to form an equivalent step bearing with higher load carrying capacity and much better performance under high pressure differential.

Both theoretical modeling [47] and experimental verification [48] of the concept of partial LST piston rings were done on relatively simple flat face "piston ring" specimens. The LST parameters for the experiments [48] were: dimple diameter of about 80 μm , dimple depth of about 8 μm , area density of 10% for full LST, and 50% for partial LST. It was shown in [47] that in partial LST an optimum textured

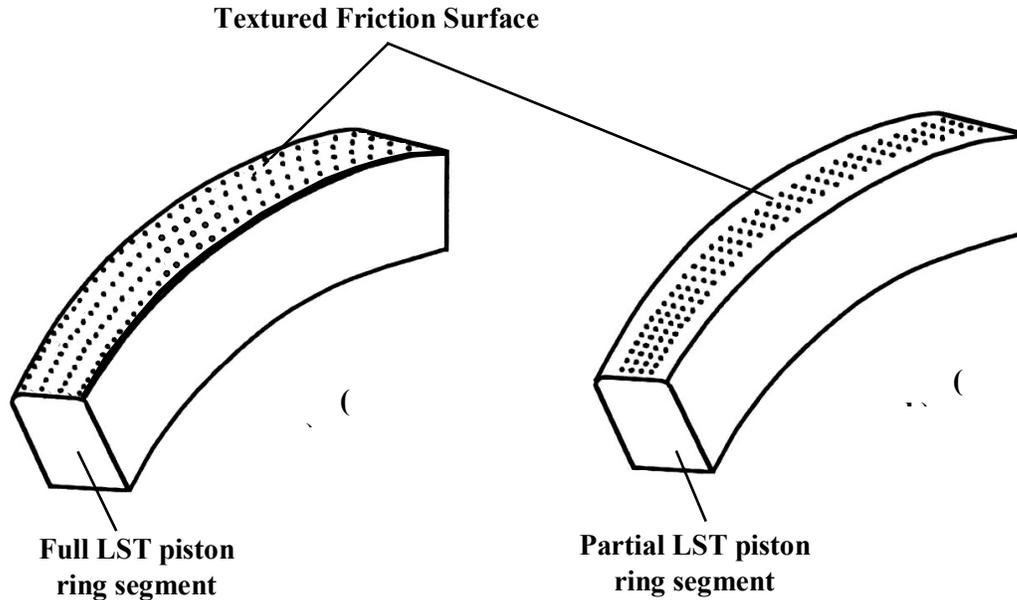


Figure 2. Segments of piston rings: (a) fully textured; (b) partially textured.

portion (the ratio between the width of the textured portion to the total ring width) of 0.6 holds for a wide range of LST parameters and operating conditions regardless of the position of this textured portion on the ring face. Hence, a total textured portion of 0.6 was applied to the partial LST specimens symmetrically at their ends. As expected it was found that the LST has a substantial effect on friction reduction compared to the un-textured reference case. The average friction obtained with the full LST was about 40 to 45 percent lower than in the reference case at low speeds around 500 RPM, and 23 to 35 percent lower at higher speeds around 1200 RPM. These percentage differences between the average friction in the un-textured and full LST cases were almost independent of the external normal load. The results clearly showed additional reduction in friction that can be obtained with partial LST over that of the full LST case as was predicted in [47]. This additional reduction varies from 12 to 29 percent depending on the load and speed.

Some preliminary real firing engine tests that were performed with LST barrel shape rings showed very little friction reduction compared to same un-textured rings. It seems that the barrel shape, which presumably was arrived at by trial and error experience over many years, is not a good candidate for LST. The crowning of the ring face by itself provides strong hydrodynamic effect that masks the weaker hydrodynamic effect of the surface texturing especially at high speeds. Indeed, a more appropriate comparison between the performance of non-textured

barrel shape and optimum partial LST cylindrical shape rings, which was performed on a laboratory reciprocating test rig [49], showed a friction reduction of up to about 25 percent with partial LST cylindrical face rings.

A 4-cylinder, Ford Transit, naturally aspirated, 2,500 cc Diesel engine was used at the Ben Gurion University in Israel to test the effect of the LST as applied to the upper set of rings in a firing engine [50]. The rings outer diameter was 93.7 mm and their nominal width was 2.5 mm.

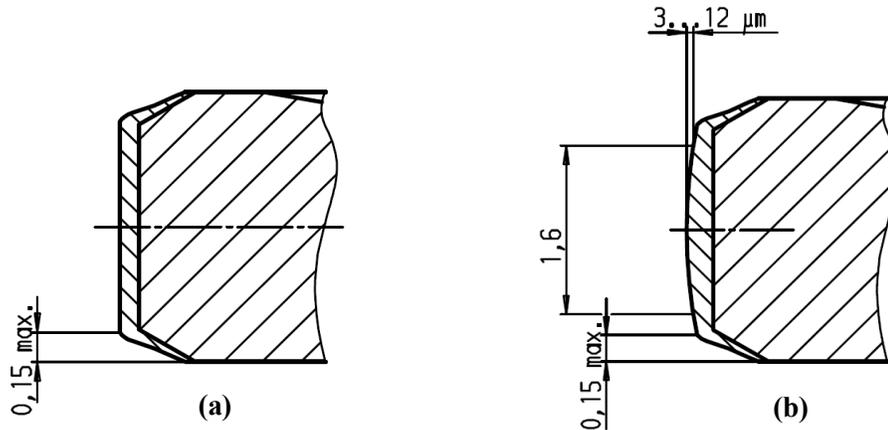


Figure 3. Cross sections of cylindrical (a) and barrel shape (b) Cr coated piston rings.

The peripheral faces of the rings were coated with a Chrome base coating that forms the ring profile in contact with the cylinder liner. Figure 3(a) shows a cross section of a ring with a cylindrical face profile to which partial laser texturing was applied. Figure 3(b) shows a ring with a barrel face profile that is a series production ring and was used as the baseline without texturing. In addition, cylindrical face rings identical to these shown in Fig. 3 but without the Chrome coating were also obtained for texturing.

The laser texturing was applied at both axial ends of the cylindrical face rings with a total textured portion of 0.6. Figure 4 shows a 3D optical profilometer scan of the partial LST cylindrical face ring with the Cr coating. The dimples are located symmetrically along the circumference of the ring on both ends of its width, leaving the central portion of the ring width un-textured. Note also from Fig. 4 that the laser texturing results in bulges of raised material around the rim of the dimples. From previous test rig tests it was found that these bulges are easily removed during the first few reciprocating cycles and hence, no special post LST process is needed to remove them prior to testing.



3-Dimensional Interactive Display

Date: 12/02/2007

Time: 09:42:49

Surface Stats:

Ra: 2.81 μm Rq: 3.40 μm Rt: 20.07 μm

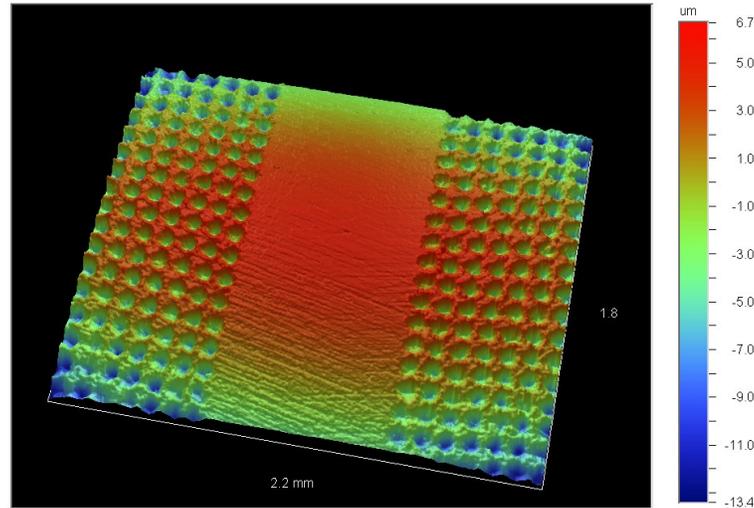
Measurement Info:

Magnification: 2.53

Measurement Mode: VSI

Sampling: 3.31 μm

Array Size: 670 X 470



Title: Subregion

Note: X offset:37 Y offset:0

Figure 4. Partial LST Cr coated cylindrical face piston ring.

A comparison between the performance of the reference un-textured conventional barrel shape rings and optimum partial LST cylindrical shape rings with and without the Cr coating is shown in Fig. 5. Clearly the laser treated rings are superior to the baseline reference rings over the entire range of engine speeds. The partial LST piston rings exhibited up to 4 percent lower fuel consumption at 1800 RPM, which corresponds to the maximum torque of the engine.

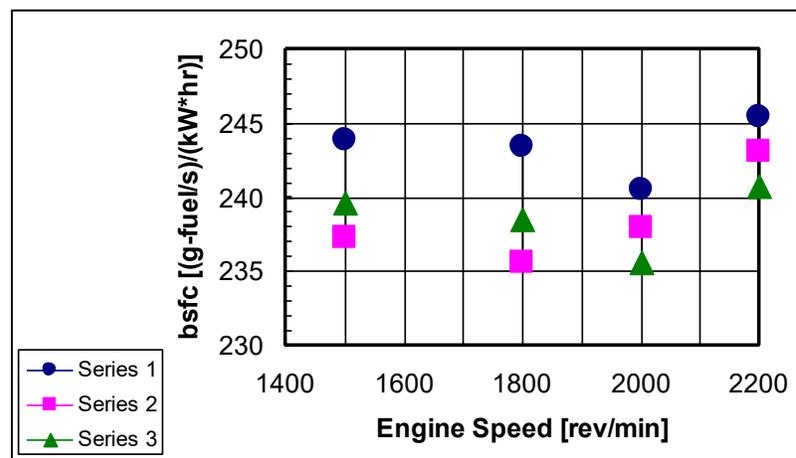


Figure 5. Engine specific fuel consumption vs. engine speed. Series 1: Barrel, chrome coated, baseline ring, Series 2: Flat, chrome coated, laser treated ring, Series 3: Flat, no chrome, laser treated ring.

Other in-cylinder components that were studied for the effect of surface texturing are the cylinder liner [51] and the piston pin [52]. The aim of the research in [51] was to undertake a comparative study between standard cross-hatched (honed) super finish cylinder liners for high-performance engines and those of identical material construction but with surface laser-etched profiles to retain a lubricant film through entrapment. The groove patterns and their interspacing and depth were optimized through a number of numerical simulation studies in order to maximize film thickness at reversal positions. A 449 cm³ four-stroke, single-cylinder engine was used to test the concept. Interestingly a 4.5 percentage gain in torque was obtained with the laser etched pattern cylinder in Ref. [51], which is very similar to the percentage gain in fuel consumption reported in [50]. Here too the maximum benefit was obtained at the pick torque. Scuffing resistance of piston pin provided by laser surface texturing in comparison with CrN and diamond-like carbon (DLC) coatings and a base-line standard piston pin was studied in Ref. [52]. Scuffing inception could be obtained with low viscosity base oil. In this case, all the treated pins performed better than the standard one, with the laser surface texturing offering the best performance. The search for better texturing in reciprocating sliding suitable for in-cylinder application is still going on as can be seen for example in Ref. [53] where the influence of surface topography on lubricant film thickness has been investigated for reciprocating sliding of patterned plane steel surfaces against cylindrical counter-bodies under conditions of hydrodynamic lubrication.

Inspired by previous finding of the function of surface texturing as micro-traps for wear debris, e.g. [29], a study was carried out on the effect of surface texturing in disk brakes [54]. In order to prevent the dispersion of particles into the surrounding environment a surface texturing in the form of radial microgrooves on the disk was utilized to trap wear particles immediately after their formation. The microgrooves entrapped wear particles from the brake pad/disk sliding interface and also reduced the total wear mass.

4.2. Bearings and seals

An impressive amount of work was done on the effect of surface texturing on hydrodynamic bearings. Several theoretical studies, that were inspired by the earlier publications [26, 27], can be found in Refs. [34] and [55] to [61].

The effect of periodic texture on the static characteristic of thrust bearing was studied in [34] where a question was raised whether a uniform texture can improve the tribological performance of a general non-parallel thrust bearing. The authors concluded that the optimum texture for best performance of such bearings is no texture at all. The same conclusion was stated in [55] where the same technique of

[37] was applied to both incompressible and compressible lubricants. It should be noted however, that in both [34] and [55] cavitation was completely neglected and hence, the conclusion mentioned above is questionable. Indeed, several other studies involving the effect of texturing in convergence film cases do not agree with that conclusion of [34] and [55]

The idea of "inlet suction" was studied in [56] for parallel thrust bearing containing a single pocket near the inlet to the bearing. It was shown that cavitation in this pocket causes lubricant to be "sucked" into the bearing through the inlet land and thus provides load support. This idea of inlet suction was extended in [57] to a linear convergent thrust bearing. It was shown, contrary to the conclusion in [34] and [55], that a textured bearing with realistic cavitation in the pocket performs better than an untextured bearing even with convergent film up to a certain convergence ratio. Only above that convergence ratio the texturing beneficial effect vanishes.

Another theoretical study on partial texturing of parallel thrust bearing is presented in [58]. Here the optimum geometrical parameters of square shaped micro-dimples which give the best tribological performance of the bearing in terms of load capacity and friction coefficient were sought. An analysis of surface textured air bearing sliders with rarefaction effects is presented in [59]. Both full and partial LST were studied and the optimum LST geometry for best tribological performance of a single row of dimples was found. Here again, like in [57], it was shown that textured slider bearing perform better than untextured bearings even with a small convergence film that can reach 350 μ rad in the case of partial texturing. It seems that the conclusion in [34] and [55] regarding the inefficiency of convergent textured bearing was only partly correct and textured bearings do in fact perform better than untextured ones not only in parallel sliding but also with small convergence as well.

The performance of a journal bearing having a smooth journal and a textured bush with square shape micro-dimples was studied theoretically in [60]. Two different cavitation models were used, the Reynolds model and the Elrod and Adams mass conservation model. It was claimed by the authors that, unlike in most untextured bearing configurations, a mass conserving cavitation model is crucial when evaluating the performance of micro-textured bearings. In another theoretical study on textured journal bearings [61] a finite difference numerical model was used to solve the Reynolds equation for a bearing with spherical dimples. It was found that an appropriate re-partition of textures on the bearing surface improves the performance of the bearing. The dimples used in [61] were of relatively large diameter in the order of 1 mm with depth of the order of several μ m, and the bearing eccentricity was at least 0.6. With such eccentricity the film thickness

convergence is substantial and once again puts in question the conclusion of [34] and [55].

Experimental studies on textured thrust and journal bearings are described in Refs. [62] to [65]. The effect of surface texturing on the performance of tilting pad thrust bearings was studied in [62]. The working faces of six pads from a 228.6 mm outer diameter bearing were textured by milling channels of less than 10 μm in depth into the Babbitt surface. Although no significant change in collar and pad temperature could be observed the textured bearing showed a tendency to exhibit lower power loss, and the inlet and outlet film thicknesses of the textured pads were larger than those for the plain Babbitt pads. The influence of shaft surface texture on the pressure development of journal bearing was investigated experimentally in [63]. Unfortunately, it is not very clear from this paper how the texture affects the bearing performance. It seems that the textured shaft had circumferential grooves, which would not provide the expected hydrodynamic effect of a conventional texturing but this is not clearly specified in the paper. A very well designed and explained experiment with textured journal bearings is described in [64]. The friction characteristic of a journal bearing with different dimpled bushings was investigated. The bearing has a steel shaft and a bronze bushing that was dimpled using two techniques; machining by indenting carbide ball shape burs, and chemical etching. The dimples are rather large with diameters of 2 and 4 mm and depth ranging from 0.13 to 1.04 mm. With proper dimple size, shape and depth a clear benefit of the textured bearing in lowering friction was demonstrated when low viscosity oil was used. A mechanical indentation technique, which the authors named "percussive burnishing", was used to texture the inner diameter of a cylindrical block made of bronze that was loaded against a rotating steel ring [65]. This block-on-ring configuration was used to simulate a journal bearing and study the effect of textured oil pockets on the bearing wear resistance. It was found that the textured oil pockets may increase wear resistance under mixed lubrication condition.

An interesting concept of combining two types of texture geometries is presented in [66] for a configuration simulating parallel thrust bearing or mechanical face seal. A large 350 μm diameter circular dimples with low area density of 4.9% in combination with 4% area density small square dimples having side length of 40 μm were textured onto a flat surface of SiC disk by lithography and reactive ion etching. The textured disk was tested in water while in relative rotation against a SiC ring under gradually increasing normal loading. It was found that the combined texturing performs better than each of the large or small dimples alone in terms of load capacity.

Two theoretical studies [67] and [68] were devoted to analyze the effect of partial LST on hydrostatic gas seals similar to the previous work [18] that was

done on liquid seals. The seal efficiency in terms of the ratio of load capacity [67] or gas film stiffness [68] over gas leakage was maximized in these studies by optimization of the texturing geometry. In [67] it was found that a textured portion of 0.5 provides the best efficiency for load capacity compared to the optimum 0.7 value that was found in [68] for the best efficiency of gas film stiffness.

4.3. Elastohydrodynamic lubrication

Differently from the previous applications where the pressure in the fluid film did not cause any deformation of the mating surfaces in relative sliding, some work was also done to evaluate the effect of surface texturing in elasto-hydrodynamic lubrication (EHL), where elastic deformation of the surfaces is important.

An extensive study on the effect of a single dimple on the EHL between a sphere and a disk is presented in [69]. This study includes both experimental and theoretical results with good correlation of the two. For the experiments a 52100 steel ball having a 25 mm diameter was tested with an EHL tribometer against a silica disk with a 60 mm diameter. The rotational speed of the ball and the disk are independently controlled to obtain different slide-to-roll ratios. Isolated circular dimples with a diameter varying from 20 to 120 μm and depth from 0.2 to over 100 μm were produced on the ball surface by a femtosecond pulse laser. Under the test conditions the contact radius between the ball and the disk was maintained at 136.5 μm . It was found that in pure rolling conditions the micro-dimple does not induce any significant variation compared to a smooth ball. When sliding is introduced the film thickness may decrease or increase depending on the dimple depth where shallower dimples are the ones with the positive effect, moving the transition between EHL and boundary lubrication towards more severe operating conditions. The slide-to-roll ratio is an important parameter too, showing greater effect of the texturing when the disk is moving faster than the ball. Very similar effects are reported in [70] and [71] where an array of several dimples, produced by micro-indentation of the ball surface instead of just one single dimple as in [69], were passing through the EHL contact. The depth of the micro-dents in [70] varied between 1100 and 1900 nm and their diameter at the ball surface between 90 and 120 μm . An attempt to change the micro-dents depth by polishing the sphere surface changed the micro-dents diameter as well. In the second paper by the same authors [71] the depth of the micro-dents was smaller to begin with, varying between 513 and 1453 nm, but apparently not small enough to show the absolute positive affect of shallower dimples the depth of which according to [70] should be less than 500 nm.

A theoretical study [72] reports virtual texturing and simulation of a group of textured surfaces in a lubricated concentrated contact. The focus of the study is on selecting the best texture distribution patterns for best lubrication performance. The area density of the texture was about 10% and the depth of all the textured features was 3 μm . The geometrical configuration consisted of a textured crowned steel cylinder with radii of curvature 21.5 and 700 mm inside a smooth hollow aluminum cylinder with inside radius of 22.5 mm. Hence, the simulation is of a conformal concentrated contact which is different from the non-conformal concentrated contact cases described in [69] – [71]. It was concluded that narrow short grooves perpendicular to the motion direction seems to be the best choice.

4.4. Magnetic storage

Laser surface texturing was used in magnetic storage mainly to reduce adhesion and stiction at start up e.g. [7]. However, the hydrodynamic lubrication provided by the dimples can be also beneficial during the flying phase of the recording device. The effect of LST on both hard disk sliders and magnetic tapes was studied in [73], [74], and [59]. A theoretical investigation on friction reduction between a magnetic tape and its guide is presented in [73]. It was shown that the friction coefficient can be minimized by creating micro-dimples on the cylindrical surface of the guide (see Fig. 6).

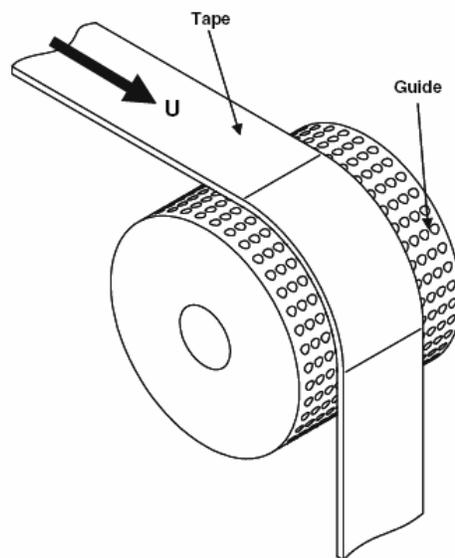


Figure 6. Tape moving over a laser surface textured guide.

The dimples enhance the formation of an air bearing and reduce the friction coefficient between the tape and the guide due to increased spacing. A parametric analysis was performed in [73] to find the optimum LST geometry. It was found that the optimum dimple aspect ratio (depth over diameter) for maximum average air bearing pressure is 0.006 and the best area density is between 0.1 and 0.3. The model results in [73] confirmed previously observed experimental results [74] with two different magnetic tapes and several guide types. Lower friction coefficient was observed in these tests when the performance of LST guide was compared with that of an un-textured commercial guide over a wide range of sliding speeds from 1 to 8 m/s. The LST guide provided earlier take-of speed at 1 m/s with up to 50 percent reduction in friction coefficient at that speed compared to the commercial guide.

A numerical model was developed in [59] to analyze surface textured air bearing sliders in hard disk drives. The effect of the texture on the steady state flying characteristics as well as the flying height modulation, and pitch and roll motion of an actual magnetic recording slider excited by a disturbance on the disk were evaluated. It was found that textured sliders show better dynamic performance compared to untextured sliders in terms of stiffness and damping.

4.5. Miscellaneous

Several other potential applications of surface texturing can be found in experimental studies presented in [75] to [79]. A study of applying micro-texture to cast iron surfaces to improve the tribological properties of reciprocating sliding guideways of machine tools is presented in [75]. It was clearly shown in [75] that the texture can be beneficial or detrimental depending on the texture geometry. For example, groove pattern texture led to a higher friction coefficient while circular dimple pattern texture led to a lower friction coefficient compared to untextured surfaces. In [76] an attempt was made to use surface texturing for improved lubrication at high pressure and low sliding speed of roller/piston in hydraulic motors. The simulated piston surface was textured using embossing tools that generate micro-grooves pattern. It was found that the friction level was only marginally influenced by the textures. An important conclusion that can be drawn from both [75] and [76] is that thorough theoretical modeling, to optimize the texturing geometry, is required to ensure successful texturing. The experimental trial and error approach can easily lead to poor performance with wrong conclusions regarding the benefit of surface texturing.

A somewhat different potential application for surface texturing is described in [77] and [78] where textured dimples are used as reservoirs for solid lubricant. In [77] a UV laser beam was used to LST the surface of hard TiCN coatings with area

density between 0.5% and 50%. Solid lubricants based on MoS₂ and graphite were then applied by burnishing and sputtering to the textured surfaces. Friction tests against steel balls indicated an optimum area density of about 10% resulting in an order of magnitude longer life of the solid lubricant on the textured surfaces compared to that on untextured ones. A very similar concept is presented in [78] where the surface texturing was performed by using pulsed air arc treatment. Here too friction tests were performed against steel balls showing between 1.75 to 3 times longer life of the solid lubricant. Finally a new Electrolytic plasma technology EPT is claimed by the authors of [79] of being capable to combine the benefits of LST in texturing with coating and alloying of surfaces.

5. Summary

A review of surface texturing, and more specifically laser surface texturing (LST), has revealed the potential of this technology in improving tribological performance of various mechanical components over a wide range of different operating conditions. The micro-dimples produced on the surface by a pulsating laser beam can act as micro-hydrodynamic bearings in cases of full or mixed lubrication with either incompressible or compressible lubricants. These dimples can serve as micro-reservoirs for lubricant in cases of starved lubrication conditions, in EHL, and for solid lubricants, and they can also provide micro-traps for wear debris in either lubricated or dry sliding.

Many theoretical and experimental studies on surface texturing were performed by a large number of researchers with various types of texturing geometries and with different texturing technologies. These researchers come from many countries around the world as shown in Table 1. The Table presents the distribution of researchers by countries of origin and references corresponding to work done in the period from 2005 through 2007. In almost all these studies surface texturing showed a beneficial effect on the tribological performance. Some of the experimental studies involve limited trial and error approach which not always was able to produce the expected benefit. On the other hand whenever a thorough theoretical modeling was performed with extensive parametric analysis to optimize the texturing geometry, success during following experiments was inevitable. Of all the various technologies used for surface texturing like, for example, machining, embossing, ion beam texturing, etching etc the LST is probably the most advanced so far. It is friendly to the environment, can be used on almost any material, is very precise and can be incorporated in production lines for fast processing. LST is finding its way into different applications and will probably become a widely accepted technology of surface engineering.

Table 1. Distribution of researchers by countries of origin and by references corresponding to work done in the period from 2005 through 2007.

Country	References
Algeria	61
Argentina	34, 38, 55, 60
Brazil	60
Czech Republic	70, 71
Finland	40*
France	34, 37, 55, 60, 61, 69*
Germany	38*, 40*, 41*
Greece	40*
Israel	33, 39* 47, 48*, 49*, 50*, 52*, 59, 67, 68, 73, 74, 78
Japan	66, 72, 75
Netherlands	35
Poland	46, 65
Sweden	31, 40*, 45, 62, 76
Switzerland	40*, 42*, 44
Turkey	63
UK	32, 44, 51*, 53, 56, 57, 58, 62
USA	36, 39*, 43*, 54, 59, 64, 72, 73, 74, 77, 79

* indicates experimental work involving LST

References

1. Etsion, I., and Burstein, L. 1996. *Tribology Transactions*, 39, 677.
2. Etsion, I., Halperin G., and Greenberg, Y. 1997. *Proc. 15th Int. Conf. on Fluid Sealing*, BHR Group, Maastricht, 3.
3. Geiger, M., Roth, S., and Becker, W. 1998. *Surface and Coatings Technology*, 100-101, 17.
4. Geiger, M., Popp, U., and Engel, U. 2002. *Annals of the CIRP*, 51, 231.
5. Etsion, I. 2005. *ASME J. Tribology*, 127, 248.
6. Ranjan, R., Lambeth, D.N., Tromel, M., Goglia, P., and Li, Y. 1991. *J. Applied Physics*, 69, 5745.
7. Zhou, L., Kato, K., Vurens, G., and Talke, F.E. 2003. *Tribology Int.*, 36, 269.
8. Wang, X., Kato, K., Adachi, K., and Aizawa, K. 2001. *Tribology Int.*, 34, 703.
9. Kononenko, T.V., Garnov, S.V., Pimenov, S.M., Konov, V.I., Romano, V., Borsos, B., and Weber, H.P. 2000. *Appl. Phys. A*, 71, 627.
10. Dumitru, G., Romano, V., Weber, H.P., Haefke, H., Gerbig, Y., and Pflüger, E. 2000. *Appl. Phys. A*, 70, 485.

11. Dumitru, G., Romano, V., Weber, H.P., Sentis, M., and Marine, W. 2002. *Appl. Phys. A*, 74, 729.
12. Dumitru, G., Romano, V., Weber, H.P., Pimenov, S., Kononenko, T., Hermann, J., Bruneau, S., Gerbig, Y., and Shupegin, M. 2003. *Diamond and Related Materials*, 12, 1034.
13. Dumitru, G., Romano, V., Weber, H.P., Sentis, M., and Marine, W. 2003. *Appl. Surface Science*, 205, 80.
14. Kovalchenko, A., Ajayi, O., Erdemir, A., Fenske, G., and Etsion, I. 2004. *Tribology Trans.*, 47, 299.
15. Etsion, I., Kligerman, Y. and Halperin, G. 1999. *Tribology Trans.*, 42, 511.
16. Etsion, I. 2000. *Proc. 17th Int. Pump Users Symposium*, 17.
17. Yu, X.Q., He, S., and Cai, R.L. 2002. *J. Materials Processing Technology*, 129, 463.
18. Etsion, I., and Halperin, G. 2002. *Tribology Trans.*, 45, 430.
19. Pride, S., Folkert, K., Guichelaar, P., and Etsion, I. 2002. *Lubrication Engineering*, 58, 16.
20. Kligerman, Y., and Etsion, I. 2001. *Tribology Trans.*, 44, 472.
21. McNikel, A., and Etsion, I. 2004. *ASME J. Tribology*, 126, 788.
22. Ronen, A., Etsion, I., and Kligerman, Y. 2001. *Tribology Trans.*, 44, 359.
23. Ryk, G., Kligerman, Y., and Etsion, I. 2002. *Tribology Trans.*, 45, 444.
24. Golloch, R., Merker, G.P., Kessen, U., and Brinkmann, S. 2004. *Proc. 14th International Colloquium Tribology*, Jan 13-15, Esslingen, Germany, 321.
25. Gehring GmbH & Co. KG, (web: <http://www.Gehring.de>).
26. Brizmer, V., Kligerman, Y., and Etsion, I. 2003, *Tribology Trans.*, 46, 397.
27. Etsion, I., Halperin, G., Brizmer, V., and Kligerman, Y. 2004. *Tribology Letters*, 17, 295.
28. Lo, S.W., and Wilson, W.R.D. 1999. *ASME J. Tribology*, 121, 731.
29. Varenberg, M., Halperin, G., and Etsion, I. 2002. *Wear*, 252, 902.
30. Volchok, A., Halperin, G., and Etsion, I. 2002. *Wear*, 253, 509.
31. Sahlin, F., Glavatskih, S.B., Almkvist, T., and Larsson, R. 2005. *ASME J. Tribology*, 127, 96.
32. Brajdic-Mitidieri, P., Gosman, A.D., Ioannides, E., and Spikes, H.A. 2005. *ASME J. Tribology*, 127, 803.
33. Feldman, Y., Kligerman, Y., Etsion, I., and Haber, S. 2006. *ASME J. Tribology*, 128, 345.
34. Buscaglia, G.C., Ciuperca, I., and Jai, M. 2005. *ASME J. Tribology*, 127, 899.
35. de Kraker, A., van Ostayen, R.A.J., van Beek, A., and Rixsen, D.J. 2007. *ASME J. Tribology*, 129, 221.
36. Wang, Q.J., and Zhu, D. 2005. *ASME J. Tribology*, 127, 722.
37. Buscaglia, G.C., Ausas, R.F., and Jai, M. 2006. *Inverse Problems in Science and Engineering*, 14, 365.
38. Schreck, S., and Zum Gahr, K.H. 2005. *Appl. Surface Science*, 247, 616.
39. Kovalchenko, A., Ajayi, O., Erdemir, A., Fenske, G., and Etsion, I. 2005. *Tribology Int.*, 38, 219.
40. Andersson, P., Koskinen, J., Varjus, S., Gerbig, Y., Haefke, H., Georgiou, S., Zhmud, B., and Buss, W. 2007. *Wear*, 262, 369.
41. Zum Ghar, K.H., Mathieu, M., and Brylka, B. 2007. *Wear*, 263, 920.

42. Dumitru, G., Romano, V., Gerbig, Y., Weber, H.P., and Haefke, H. 2005. *Appl. Phys. A*, 80, 283.
43. Daniel, C., and Dahotre, N.B. 2006. *Advanced Engineering Materials*, 8, 925.
44. Gerbig, Y.B., Ahmed, S.I.U., Chetwynd, D.G., and Haefke, H. 2006. *Tribology Int.*, 39, 945.
45. Pettersson, U., and Jacobson, S. 2006. *Tribology Int.*, 39, 695.
46. Koszela, W., Pawlus, P., and Galda, L. 2007. *Wear*, 263, 1585.
47. Kligerman, Y., Etsion, I., and Shinkarenko, A. 2005. *ASME J. Tribology*, 127, 632.
48. Ryk, G., Kligerman, Y., and Etsion, I. 2005. *Tribology Trans.*, 48, 583.
49. Ryk, G., and Etsion, I. 2006. *Wear*, 261, 792.
50. Etsion, I., and Sher, E. 2009. *Tribology Int.*, 42, 542.
51. Rahnejat, H., Balakrishnan, S., King, P.D., and Howell-Smith, S. 2006. *Proc. ImechE, Part D: J. Automobile Engineering*, 220, 1309.
52. Etsion, I., Halperin, G., and Becker, E. 2006. *Wear*, 261, 785.
53. Costa, H.L., and Hutchings, I.M. 2007. *Tribology Int.*, 40, 1227.
54. Mosleh, M., and Khemet, B.A. 2006. *Tribology Trans.*, 49, 279.
55. Buscaglia, G.C., Ciuperca, I., and Jai, M. 2007. *J. Math. Anal. Appl.*, 335, 1309.
56. Olver, A.V., Fowell, M.T., Spikes, H.A., and Pegg, I.G. 2006. *Proc. IMechE Part J: J. Engineering Tribology*, 220, 105.
57. Fowel, M., Olver, A.V., Gosman, A.D., Spikes, H.A., and Pegg, I. 2007. *ASME J. Tribology*, 129, 336.
58. Brahamani, R., Shirvani, A., and Shirvani, H. 2007. *Tribology Trans.*, 50, 401.
59. Murthy, A.N., Etsion, I., and Talke, F.E. 2007. *Tribology Letters*, 28, 251.
60. Ausas, R., Ragot, P., Leiva, J., Jai, M., Bayada, G., and Buscaglia, G.C. 2007. *ASME J. Tribology*, 129, 868.
61. Tala-Ighil, N., Maspeyrot, P., Fillon, M., and Bounif, A. 2007. *Proc. IMechE Part J: J. Engineering Tribology*, 221, 623.
62. Glavatskih, S.B., McCarthy, D.M.C., and Sherrington, I. 2005. *Tribology Trans.*, 48, 492.
63. Sinanoglu, C., Nair, F., and Karamis, M.B. 2005. *J. Materials Processing Technology*, 168, 344.
64. Lu, X., and Khonsari, M.M. 2007. *Tribology Letters*, 27, 169.
65. Galda, L., Koszela, W., and Pawlus, P. 2007. *Tribology Int.*, 40, 1516.
66. Wang, X., Adachi, K., Otsuka, K., and Kato, K. 2006. *Appl. Surface Science*, 253, 1282.
67. Feldman, Y., Kligerman, Y., and Etsion, I. 2006. *Tribology Letters*, 22, 21.
68. Feldman, Y., Kligerman, Y., and Etsion, I. 2007. *ASME J. Tribology*, 129, 407.
69. Mourier, L., Mazuyer, D., Lubrecht, A.A., and Donnet, C. 2006. *Tribology Int.*, 39, 1745.
70. Krupka, I., and Hartl, M. 2007. *Tribology Int.*, 40, 1100.
71. Krupka, I., and Hartl, M. 2007. *ASME J. Tribology*, 129, 502.
72. Ren, N., Nanbu, T., Yasuda, Y., Zhu, D., and Wang, Q. 2007. *Tribology Letters*, 28, 275.
73. Raeymaekers, B., Etsion, I., and Talke F.E. 2007. *Tribology Letters*, 2007, 28, 9.
74. Raeymaekers, B., Etsion, I., and Talke F.E. 2007. *Tribology Letters*, 2007, 27, 89.
75. Nakano, M., Korenaga, A., Miyake, K., Murakami, T., Ando, Y., Usami, H., and Sasaki, S. 2007. *Tribology Letters*, 28, 131.

76. Pettersson, U., and Jacobson, S. 2007. *Tribology Int.*, 40, 355.
77. Voevodin, A.A., and Zabinski, J.S. 2006. *Wear*, 261, 1285.
78. Moshkovith, A., Perfiliev, V., Gindin, D., Parkansky, N., Boxman, R., and Rapoport, L. 2007. *Wear*, 263, 1467.
79. Gupta, P., Tenhundfeld, G., Daigle, E.O., and Ryabkov, D. 2007. *Surface and Coating Technology*, 201, 8746.



Research Signpost
37/661 (2), Fort P.O.
Trivandrum-695 023
Kerala, India

Recent Developments in Wear Prevention, Friction and Lubrication, 2010: 159-196
ISBN: 978-81-308-0377-7 Editor: George K. Nikas

4. Unification of friction and wear

Michael D. Bryant

Mechanical Engineering, University of Texas at Austin, Austin, Texas, 78712-0292, USA

Abstract. Friction and wear are often treated as unrelated, distinct phenomena. In fact, friction and wear are macroscopic manifestations of common physical processes operative at sliding interfaces. Friction and wear are related through the physics of these processes, which are thermodynamically irreversible and dissipative, since energy loss is almost always involved. This chapter attempts to unify friction and wear, by focusing on the dissipative processes found at sliding interfaces. Prominent mechanisms of friction and wear are first reviewed, with goal of identifying the associated dissipative processes. The laws of thermodynamics are then reviewed, with a focus on entropy generation and the first and second laws. This is followed by review of the Degradation-Entropy Generation theorem, which relates degradation of any form to the irreversible entropy generated by responsible dissipative processes. The theorem is applied to sliding interfaces and a tribological control volume, to relate wear to responsible interfacial dissipative processes. Friction force is also related to entropy generated by interfacial dissipative processes. Next, dissipative processes associated with friction and wear, and found at sliding interfaces, are identified and reviewed, and the entropy generated by these dissipative processes is presented. Finally, conclusions regarding friction and wear, relations to dissipative processes, are presented.

1. Introduction

Friction and wear have classically been treated as unrelated, distinct phenomena [1, 2, 3]. This view persists. Even Tribology often classifies friction and wear separately. In 1965, Rabinowicz [3] suggested why: “Friction is usually classified as a branch of Physics, or of Mechanical Engineering. Wear is often considered to be part of Metallurgy.” In reality, friction and wear are manifestations of the same interfacial physics and events, and consequently, friction and wear often depend on the same phenomenological variables. A vivid example of this interdependency is Hwang *et al.*'s [4] measurements of friction coefficient with simultaneous photography of wear particle formation and detachment. Both friction and wear are

associated with, or induced by, dissipative irreversible processes that dissipate mechanical power, reorder the material structures of the bodies, and generate entropy. Indeed, friction force usually provides the energy that induces wear.

This was recognized by Czichos [5] who advocated a systems approach to tribology. Systems approaches [6, 7] typically emphasize balance or conservation of energy, power flows, mass, and other fundamental quantities of physical systems. Czichos [5] cited various “planes”—functional, work, thermal, and material—for these balance and conservation applications to tribology. Of particular interest to this article, Czichos identified dissipative processes of friction and wear, and states on page 42, “The processes on and between the material planes are likely to involve entropy changes, generally the production of entropy...”.

This chapter will review friction and wear, with focus on showing dependencies and relationships. Past reviews [8, 9, 10, 11] of friction and wear have generally focused on physical mechanisms of friction and wear. The theme of this review is to unify friction and wear, by identifying and highlighting the dissipative irreversible processes germane to both. This chapter will review friction, wear, the thermodynamic underpinnings between friction force and degradation by wear, and the more important associated dissipative processes common to both. The goal is to show that by identifying, studying, and understanding these underlying irreversible processes, friction and wear can be related, and predicted.

2. Friction

Friction force [12] is a non-conservative and dissipative reaction force between two bodies in relative motion, tangential to the interface between the bodies, with force direction always opposing motion. Suh [13] stated “the external work done by the friction force must be equal to the sum of the internal energy increase and the energy dissipated at and near the interface.” Friction, a principal cause of power loss, converts usable work into heat and produces entropy. Friction is dissipative in the thermodynamic sense [14]. Dissipative processes, driven by energy of friction force, induce irreversible changes in surface morphology and composition. Dissipative processes associated with friction and wear include hysteresis associated with asymmetries of adhesion [15], movement of dislocations and plastic deformation from surface plowing, surface and subsurface fracture of surfaces [13], strain hardening of metallic surfaces during sliding, phase changes induced by friction heating or mechanical stresses, and mixing or transfer of material. Friction force can arise from many complex and diverse effects, including surface adhesion followed by surface rupture or tearing, surface and near surface deformation, surface gouging, surface plowing, asperity interaction, film

cracking and rupture, and lubricant shear, among others. Friction can depend on numerous phenomenological variables. Dominant factors [13] include kinematics of the surfaces in contact; applied external loads and/or displacements; environmental conditions such as temperature, pressure, humidity, chemistry, and presence of lubricants; surface topography and roughness, surface geometry, and environment; materials of slider (first body) and counter-surface (second body); material properties of these bodies; deformations of slider and counter-surface; and presence of surface films, lubricants (solid, liquid and gas), and third bodies [16] (such as entrapped wear debris or accumulated material from surface films, or transferred material from one body to another).

In reference to the Stribeck curve for lubrication [17], friction has been classified as dry, where surfaces slide without intervening lubricants, allowing direct contact between asperities; boundary lubricated, where surfaces contact but thin films adhered to one or both surfaces protect underlying material; mixed film, where lubricant films partially separate surfaces, and reduce surface contact; and hydrodynamic wet film, where pressures developed from viscous shearing of fully developed lubricant films separate surfaces, and prevent surface contact. Friction force is largest for dry friction, and reduces with boundary lubrication, mixed film lubrication, and hydrodynamic lubrication in that order. The focus of this chapter is dry and boundary lubricated friction.

Classical models dating to DaVinci and Coulomb found friction independent of apparent area of contact and proportional to normal force. The Coulomb-Amontons friction law

$$F = \mu N \quad (1)$$

relates friction force F to normal force N , via the coefficient of friction μ . Friction force often depends at least weakly on sliding speed v , the difference between surface velocities. Observations of dry sliding identify a static friction force F_s and coefficient μ_s , applicable for $v = 0$, and a kinetic friction force F_k and coefficient μ_k , valid for large v . Since usually $F_s \geq F_k$, continuity suggests $F = F(v)$ to be monotonically decreasing. Ovcharenko *et al.* [18] observed static friction force at the inception of sliding for lightly to heavily loaded steel and copper spheres sliding on sapphire. Static friction coefficient depended on real contact area, normal force, and deformation mode. Kinetic friction, which is dissipative and involves power loss, induces heating, higher temperatures, material changes, stress, cracking, fatigue, and morphological changes of surfaces. Attempts to classify kinetic friction force [10, 9, 13] identified mechanisms of adhesion, plowing, surface fatigue, tribochemical actions, delamination [13], and effects of surface films [10]. Suh [13] estimated friction force for sliding of metals via mechanics of

asperity adhesion, shearing and plowing. These mechanisms assumed irreversible plastic deformations coupled friction force and energy dissipation. For any system, multiple mechanisms of friction force are simultaneously present. Which type of friction dominates depends on conditions. The total friction force, friction coefficient [13, 9], and work of friction

$$F = \sum_j F_j, \quad \mu = \frac{F}{N} = \sum_j \mu_j, \quad W^\mu = \sum_j W_j^\mu \quad (2)$$

are the sum over all friction components. Conditions can cause one or more components of friction in the sums to dominate. For low speed dry sliding of metals, friction force is often dominated by adhesion and plowing components. The remainder of this section will introduce types of friction force and associated dissipative processes.

2.1. Adhesion component of friction

Bodies pressed into contact by forces oriented along the normal to the interface tend to adhere upon separation. Indeed, after removing a compressive force, a tensile force normal to the surface is often needed for full separation. Adhesion can be due to “cold welding” from a variety of bonding types, such as chemisorption caused by strong atomic bonding (metallic, ionic or covalent), physisorption due to intermolecular forces such as van der Waals forces [19, 20], or hydrogen bonding. Adhesion influences normal and tangential contact forces between bodies [21, 20, 22, 23], especially at length scales of order of nanometers to microns. Surface films affect adhesion. Adhesion can be strengthened or inhibited, depending on the type and nature of the film. Very strong adhesion requires bare surfaces. Wipe of surface films or tearing off of material on surfaces exposes underlying atoms to stronger metallic, ionic or covalent bonding with atoms on the opposing surface.

Bodies pressed into contact develop a real area of contact considerably smaller (tens to hundreds of times smaller) than the apparent area of contact [24, 25]. The deformations that develop the real area of contact can be elastic for lightly loaded contacts with plasticity index $\psi = (E'/H)\sqrt{\sigma_s/R'} \leq 0.6$, but for contacts with $\psi \geq 1$ or $0.6 < \psi < 1$ with heavy load, asperity deformations are principally plastic [24, 25], and the real area of contact $A_r \approx N/H$. Here N is the contact force normal to the contact surface, σ_s is the standard deviation of the surface heights, usually assumed to have Gaussian distribution,

$1/E' = (1 - \nu_1^2)/E_1 + (1 - \nu_2^2)/E_2$ is the equivalent elastic modulus and $1/R' = 1/R_1 + 1/R_2$ is the equivalent curvature between contacting asperities of bodies 1

and 2, and H is the Meyer hardness [26, 27]. After removing the contact loads, attractive forces at the interface tend to maintain contact [25]. Indeed, tensile forces, albeit small, are often required to fully separate the surfaces. These surface attractions, due to intermolecular forces [19, 20] are most pronounced on length scales of order of hundreds of nanometers or less. The overall effects of these multiple attractive forces can be lumped into surface energy γ , the work per unit area needed to create the surface, defined in section 5.4.1. During contact, the adhesive forces tend to augment the real area of contact. Johnson [21] reviewed adhesion, and how it enlarges contact area. The presence of a condensed liquid film increases the pull-off force needed to separate bodies [21, 20, 23].

Adhesive effects increase friction force, discourage separation of contact areas, and change dynamic response [22]. Bowden and Tabor's [8] early model of adhesion friction envisioned asperities on opposing surfaces contacting, cold welding (or adhering), and rupturing due to tangential forces imposed during sliding. The tangential force $F = SA_r \approx SN/H$, where the junction adhesive shear strength S depends on the thermodynamic work of adhesion, see section 5.4.1. The coefficient of friction for adhesion $\mu_a = F/N = S/H \approx 0.2$ [8, 13], since hardness H for metals is five to six fold the shear strength S . Surface films, which interfere with asperity to asperity contact, tend to reduce adhesion and adhesion friction. Suh [13] improved Bowden and Tabor's estimate using better models of asperity plasticity. Straffelini [28] extended the adhesive junction strength theory, to include plastic deformation, chemical reactions and other irreversible phenomena occurring at the asperities during adhesion and shear.

2.2. Plowing coefficient of friction

Asperities of a harder surface, or hard abrasive particles trapped between surfaces (associated with abrasive wear), can plow a softer counter-surface during sliding. The front of a "plow" digs a groove or furrow and displaces material to the sides of the wear track. For ductile metals, the principal dissipative process is plastic deformation. Suh [13] viewed an asperity as a rigid cone with aperture angle $\phi = \pi/2 - \theta$ plowing a v-shaped groove of width w and angle 2ϕ through a rigid perfectly plastic counter body. With tangential and normal forces $F = Pw^2[\tan \theta + (s/P) \sec \theta]/4$ and $N = P\pi w^2/8$ equilibrating pressure P and shear stress s on the front of the cone, the plowing coefficient of friction $\mu_p = F/N = 2/\pi(\tan \theta + s/P \sec \theta)$. The first term was derived by Bowden and Tabor [8]. With $10^\circ \leq \theta \leq 40^\circ$ estimating the range of typical asperity slopes, pressure P estimated by hardness H which is five to six times shear strength S , then for s/S between zero and one, $0.1 \leq \mu_p \leq 0.6$.

Komvopoulos *et al.* [29, 30, 31, 32] studied friction and wear of like metals sliding under boundary lubrication and dry [33] conditions. For aluminum, copper, and chromium specimens, pin on disk tests produced grooves typical of plowing. Since the specimens were initially smooth and clean, the authors hypothesized that wear particles shed from the surfaces formed, became entrapped within the sliding interface, strain hardened, and functioned as plows. Analytical estimates of friction coefficient [33] and wear volume, derived from plasticity based models of microscopic “plows” cutting out material, compared well with their measurements of friction and wear. Inspection of surfaces also suggested little adhesion. Surface films tend to avoid adhesive junction welding, moderate stress fields beneath sliders, and avoid surface damage due to plastic damage from plowing [31, 32]. These films can be self-forming, such as surface oxides [31], or applied hard TiN coatings [32]. Removal or disruption of these films permitted plowing [31, 32] with concomitant high friction and wear.

2.3. Other components of friction force

Various other components of friction exist, related to dissipative interfacial processes with dependence on distance slid x . For example, an asperity ratchet mechanism of friction force [34, 15] has interacting roughness on opposing surfaces with asperity geometries that permit asperities on one surface to climb up and fall down asperities on the other surface via different process paths, forming a hysteresis. The third body, see section 5.1 is a principal site of dissipation. Material entrapped between sliding first bodies is heavily worked by the mechanical energy dissipated by friction force.

3. Wear

Wear is loss of material from surfaces of bodies [2, 3]. Most wear requires rubbing, but wear can also be induced by non-rubbing processes such as erosion or thermally induced fracture. Sliding wear is caused by rubbing between bodies in relative motion. Wear is usually defined as the volume of material removed from a surface, but other forms of surface damage can be classified as wear, such as movement of material over a surface (due to plastic deformation or detachment), transfer of material from one surface to another, and cracking of surfaces. In all cases, wear is an irreversible process that involves permanent movement of atoms and long term degradation of the surface and underlying body. Wear can depend on numerous phenomenological variables, including geometry and kinematics of bodies, relative motion such as vibration [35, 36, 37] between bodies, force or load, sliding speed, distance slid, power dissipated, material combinations,

environment (temperature, humidity, vibrations, chemistry, atmosphere, contaminants, surface deposits, etc.), and presence of third body materials, among others. Wear removes material from surfaces, which compromises tolerances and fits of machine components. Wear debris can migrate and contaminate other parts of a machine.

Wear is commonly measured in terms of mass loss (via weight loss), volume loss, or surface recession (reduction in length of a specimen of constant cross section). When mass loss is too difficult or inconvenient to measure, wear can be estimated as amount of surface damage such as size of wear scar geometry, or estimated by the amount the wear debris shed from a surface. Wear measurements can exhibit large variability. Results from different locations and/or times can vary as much as 50 to 1000 percent. Causes of variability include slight differences in environmental conditions (e.g., temperature and humidity), kinematics of the test machine [3], and vibrations [35, 38, 36, 37], among others. Since most sliding wear is mild to moderate, tests can require several hours to days to accumulate enough wear for a reliable measurement. Consequently, wear measurement are usually expensive relative to other material measurements.

Sliding wear occurs by many different physical mechanisms [8, 2, 9, 13, 10, 39]. Tabor [10] sorted wear of metals and ceramics into mechanisms of adhesion, grooving and cutting, fracture or cracking (more important for brittle ceramics), friction heating, and chemical effects, and wear of polymers into mechanisms of adhesion, grooving and cutting, and heat induced surface softening and melting. Recently, Kato [39] grouped wear into mechanical, chemical and thermally induced forms, and identified abrasive, adhesive, flow, fatigue, corrosive, melt, and diffusive wear modes, with dominant wear processes of fracture, plastic flow, liquid flow, dissolution, oxidation, and evaporation. Kato identified "physical adsorption, chemical adsorption, tribochemical activation, tribofilm formation, oxidation and delamination, oxidation and dissolution, oxidation and gas formation, phase transition, recrystallization, crack nucleation and propagation, and adhesive transfer and retransfer" (of material to and from bodies) as principal physics and chemistry of wear. Suh [13] classified wear into adhesive, abrasive, surface fatigue, tribochemical, and delamination modes. Each of these forms of wear involves a different physio-chemical mechanism with a different set of dissipative processes operative at the interface. Ludema [40, 41] noted that wear models formulated before 1995 involved many diverse wear mechanisms, many different physical principles, and many variables.

Similar to equation (2), the total wear w can be viewed [39, 13] as a sum

$$w = \sum_j w_j \quad (3)$$

of components from multiple wear mechanisms w_j present and operative at a sliding interface. Conditions (e.g., speed, load, environment) can favor one mechanism of wear to dominate the sum, and marked changes in wear rate can be observed. The conditions that trigger transitions in wear rates are of particular interest to machine designers. Wear maps identify operating and environmental conditions for severe, moderate, and low wear. Wear maps plot curves of constant wear rate in a multi-axis space, where the axes indicate the intensity of physical variables such as speed, temperature, load, and humidity. Wear maps associate levels and/or modes of wear with physical conditions. Boundaries between regions graphically identify the conditions for transitions in wear. Hsu and Shen [42] developed wear maps for ceramics. Maeda *et al.* [43] developed wear maps for rubber, and showed that scratch wear of rubber depended on adhesion, deformation, and energy dissipated by friction. Deformation depended on viscoelastic properties, tearing depended on static strength properties. Briscoe *et al.* [44, 45] reviewed a body of work involving lubricated polymer nano-coatings. Different materials, films, and lubricants were compared, and assessed. Measurements of wear and friction involved use of scratch tests. Deformation mechanisms were analyzed.

The remainder of this section will overview prominent mechanisms of wear, and identify the principal dissipative processes involved.

3.1. Adhesive wear

Adhesive wear is usually present, generally mild, and usually dominant at low speed. Suh *et al.* [30] attributed sliding wear of like metals to plowing and abrasive wear, but their experiments with initially clean surfaces likely generated hard abrasive particles via adhesive wear. Surfaces adhere through attraction from van der Waals forces, and sometimes stronger metallic bonding if surfaces are bare. During sliding, asperities on opposite surfaces approach, interfere, and bond or weld [8] to each other. The adhesive bonding can be caused by many effects discussed in section 5.4.1. Further sliding breaks the bonds, and can detach a wear particle of size 10 to 100 microns from the surface. To model sliding wear, Holm [26] and Archard [46, 47], each working independently, expressed the volume of material lost to wear

$$W = \frac{K}{H} NL \quad (4)$$

to the normal force N pressing the bodies together, the distance slid L , the diamond pyramid hardness number (DPN) H , and a dimensionless wear coefficient K that depends on material selection and environment [47]. Hardness H balances the

dimensions of wear volume w with phenomenological variables N and L . Typical values of wear coefficient K range from 10^{-5} to 10^{-3} for metals [3, 48]. Although Holm and Archard derived equation (4) by considering the probabilistic interaction of atoms, asperities, or contact regions between the sliding opposing surfaces, note that the product NL has dimensions of energy. If equation (1) is substituted into equation (4),

$$w = \frac{K}{\mu H} FL = K' FL = K' W \quad (5)$$

where $K' = K/(\mu H)$. Equation (5) emphasizes the relation between wear and friction force. Implicit in equation (5) is an energy statement involving work $W = FL$ of friction force. Equation (4) was formulated for adhesive wear, but describes other mechanisms such as abrasive wear. Equation (4) can be revised to be consistent with other measurements of wear. For surface recession (loss of length from worn slider), divide w by the cross section area A of the slider.

During adhesive wear, asperity junctions adhere. Forces induced by subsequent sliding dissipate energy by rupturing and/or plastic deforming the adhered regions. These actions can tear off strips of material which can transfer to the counter surface, reattach to the slider, join the third body [16], or exit the sliding interface to become wear particles [49]. Spalling, galling, scuffing and seizure [10, 50] are severe forms of adhesive wear with large loss of material. These modes often involve changes in thickness, material structure, and composition of surface layers. Suh *et al.* [51] found a basic change in surface material composition of aluminum and cast iron discs (run against pins of similar material) at the onset of scuffing. The surface composition changed after protective coatings deteriorated, exposing underlying disc material to adhesion, seizure, and severe wear.

3.2. Abrasive wear

Abrasive wear is moderate to severe. Abrasive particles originating from the environment or generated by wear processes [30] can become entrapped within a sliding interface. As the slider moves over the counter surface, some of these abrasive particles become plows, cutters, or rollers, which can gouge, scratch, or indent the surfaces [52]. The much larger wear particles dug from the surface results in grooves, and a wear coefficient K , see equation (4), typically tens to thousands of times higher than adhesive wear. Abrasive wear can be modeled by equation (4), or more accurately by the Preston-Rabinowicz [3] equation, which is similar to Archard's law, equation (4). Metallic wear debris entrapped within the

interface join the third body. As proposed by Godet [16, 53], third body accommodates the velocity discontinuity generated by the slip velocity of sliding between the bodies. As such, the particles are subjected to forces from rolling, shear, and compression. The wear debris particles can work-harden due to plastic deformations imposed on the particles, as part of the third body. These hardened particles can become abrasive particles, causing an abrupt change in wear rate. Since abrasion involves mechanisms of plowing and machining, plastic deformation and creation of new surface are the major dissipative processes.

3.3. Fretting wear

Fretting involves rubbing between bodies with small amplitude, low frequency oscillatory motions [54]. Fretting degrades components through surface wear and structural fatigue. Fretting wear of metals involves intense plastic deformations near the surface of the wearing body, sometimes accompanied with corrosion or oxidation of material. Rise of bulk temperatures are usually moderate, of order of tens of degrees Kelvin [55]. Temperature affects friction and fretting wear [54]. The fretting process is gradual, with the fretted component in state of equilibrium or quasi-equilibrium [56]. Mohrbacher *et al.* [57], Huq and Celis [58], and later Fouvry *et al.* [59, 60, 61, 62] developed an energy based model for fretting wear. Exhaustive measurements by these groups showed the Archard wear law, equation (4) to be an inappropriate descriptor of their data. Mohrbacher *et al.* [57] found that the wear volume, measured by measuring the geometry of the wear scar beneath the fretting contact, when plotted versus total friction energy dissipated, produced a straight line. This observation was confirmed by others [59, 60, 61, 62, 58, 63, 64] for fretting wear of various coatings on substrates. Models of fretting wear, when based on accumulated friction energy dissipated [65, 66, 62, 67], can predict the time evolution of the wear scar profile. The principal dissipation modes for fretting wear are plastic deformation and adhesion. This is sometimes supplemented by corrosion and oxidation.

Bryant, Khonsari and Ling's [68] application of the Degradation-Entropy Generation theorem to fretting wear predicted a dependence on friction energy dissipated, consistent with exhaustive measurements of Huq and Celis [58], Fouvry *et al.* [59, 60, 61, 62], and Mall *et al.* [63, 64], who collectively found Archard's wear law to be a poor predictor of fretting wear. Fretting wear involves an irreversible process in a critically stable state near equilibrium, but in the process of transitioning to a new equilibrium state [56]. Instead of assuming a steady process or an equilibrium process, Bryant, Khonsari and Ling [68] used Dai *et al.*'s [56] thermodynamic condition, which perturbed about the equilibrium point, set the first

variation of entropy to zero, and expressed the system state in terms of the second variation.

3.4. Corrosion wear

Corrosive wear is a mechanism wherein a corrosion [69] film forms on the surface of a slider, and subsequent sliding removes the film. Removal of the film exposes underlying parent material to new corrosion. Oxidational wear is a form of corrosive wear [70, 71], especially in steels. The corrosion film, of different composition from the slider's parent material, has different mechanical behavior. Friction, hardness, and resistance to wear can change, depending on conditions. Corrosion films tend to protect underlying parent material from environmental and mechanical duress. For example, oxide films on aluminum discourage further surface oxidation [26, 69]. If the film has sufficient mechanical strength to not fail under stress imposed by sliding, the film and corrosion can reduce wear. Mechanically weak films, with low shear strength, can easily scale or flake off, and increase wear. Corrosive wear can involve mechanisms of adhesive, abrasive, and delamination wear. As with any reaction, surface corrosion is controlled by supply of reactants and activation energy. Reactant supply is controlled by oxygen pressure (which controls oxygen concentration), and diffusion of metal ions and oxygen to the metal-corrosion film interface [69, 70]. The reaction rate affects the wear rate through the thickness and composition of the film. Friction heat can supply energy to drive reactions and encourage thicker films. Temperature influences composition, phase, and thickness of the surface oxide film. Thermodynamic conditions determine the corrosion reaction rate and products, which indirectly influences wear. Tendency to corrode is controlled by Gibbs free energy, which drives the direction of a reaction.

3.5. Fatigue wear

Fatigue wear is a process wherein repeated application of subcritical stress – stresses below the failure limit or material strength – moves dislocations through the stressed body, until they pile up and initiate tiny cracks. Continued application of the stress causes the cracks to propagate, or extend. After hundreds to possibly hundreds of millions of cycles of the repeated stress, a volume of material defined by the crack propagation path detaches from the body, forming a wear particle. Delamination wear [13] involves subsurface cracks a millimeter or so beneath the surface, propagating parallel to the surface, and driven by repeated passing of contact loads. Under repeated application of the load, the cracks propagate until, upon reaching a critical length, the cracks propagate up towards the surface. When the cracks intersect with the surface, strips peel off or delaminate from the surface. Pitting wear, a result of

rolling contact fatigue, is a fatigue process wherein cracks form under repeated application of rolling loads, usually associated with rolling element bearings or gear teeth contacts. Cracks initiate near or at the surface, and propagate into the material body at angles of about 20° to the contact surface, opposite the rolling direction. As the cracks lengthen, the crack propagates parallel to the surface. Eventually, the crack propagation direction turns, and the crack intersects the surface, resulting in an arrowhead shaped wear particle with a concomitant pit in the surface.

3.6. Thermal and thermomechanical effects on wear

Thermal effects can accelerate wear and cause severe surface damage. Rubbing of bodies dissipates heat power $Q = Fv$ from friction force F and relative velocity (e.g, slip velocity) between surfaces v . Friction heat can exert many different effects on sliders and lubricants. Higher temperatures can change material phase, which alters material properties. For example, metals can soften or melt, carbon graphite mixtures can harden, and thermal and electrical conductivities tend to diminish. Often, these altered properties reduce heat transfer from hot regions. Concentrated contact, called thermoelastic instability, hot spotting, thermal mounding, or thermal distress, is a thermomechanical instability that results in much higher wear. Rough surfaces segment the contact into discrete islands of contact within the apparent area of contact. As sliding commences, friction dissipation, which occurs at the islands of contact, is concentrated. Resulting elevated temperatures expand the regions about the contact islands, causing the spots to expand and grow outward toward the interface. Differences in spot geometry, contact pressure, and local friction cause some spots to heat and grow faster, which tends to separate lesser growing spots. The loads carried by the lesser spots transfer to the still-connected spots, inducing more intense conditions (forces, heating, temperatures, and thermal expansions) on the still-connected spots. This process continues until the slider runs on only a few spots, with temperatures of several hundreds to thousands of degrees kelvin and stresses at plastic limits. Under these extreme conditions, wear coefficients can double or increase tenfold or more.

4. Unifying friction and wear

Friction and wear, manifestations of the same interfacial physics and dissipative processes, are intimately related. Non-conservative friction force dissipates power and generates irreversible entropy. Wear is surface degradation driven by interfacial dissipative processes, including those associated with friction. This

section considers the thermodynamics, energy, entropy, and dissipative processes related to friction and wear.

4.1. Godet's third body

A major advance towards unifying friction and wear was Godet's [16, 53] introduction of the third body concept, which considers the "other" material entrapped within the interface between two sliding first bodies. The third body consists of contaminants and particles from the environment, material and wear debris shed from the first bodies, and films on the first body surfaces [4]. For sliding, a difference in velocities must exist between the bodies, which seems to create a discontinuity in velocity across the interface. Godet [16, 53] questioned the existence of this discontinuity, and suggested that material within the interface – the third body – must accommodate the discontinuity. Godet likened the third body's velocity accommodation effect to a lubricant film between two surfaces sheared by relative tangential motions. Here the flow velocities vary continuously across the film to accommodate the slip velocity across the bodies. To accommodate the discontinuous slip between sliding first bodies, Godet [16] showed that the third body must experience severe strains imposed by intense shear and compression during sliding.

The third body can serve as a media to transmit forces and prevent contact between first bodies. Since velocity accommodation subjects the third body to severe strains, the third body is a principal site of intense energy dissipation such as plastic work. Iordanoff *et al.*'s [72, 73] computer model of third body effects and kinematics suggested that the third body can behave like a fluid or solid. Flow of third body material influences friction and wear. Friction force, increased particle size due to clumping of particles, density of third body material, and third body and surface flows were found dependent on the amount of adhesion between small particles within the third body. Richard *et al.*'s [74] 3D simulations assumed sliding first bodies to be rigid, single layers of particles and third body particles to be 5 micron radius spheres. From particle motions driven by first body motions, macroscopic friction and rheology of the third body were derived. The model varied adhesion between particles and only allowed dissipation in the third body. Under different regimes of adhesion force, Richard observed four possible flow patterns of third body particles: fluid-like, with easy shear and mixing between particles; semifluidic, with some shear and mixing, but also with some particles moving in groups; elasto-plastic, with some particles moving only slightly, but others moving in groups with irreversible displacements; and elastic, with reversible motions and deformations. Richard found the dynamic friction coefficient proportional to the power dissipated. Experiments with many sub-

millimeter sized balls sheared between tangentially moving parallel plates verified their particle kinematics.

Iordanoff *et al.* [72] reviewed constitutive laws of third bodies, identified kinematic modes of third body particles for velocity accommodation, and surveyed possible fluid-like rheological constitutive models to describe third body behavior under imposed boundary conditions. Jang and Khonsari [75] treated powder lubrication as a granular media, formulated equivalent continuity, momentum, and energy equations into an equivalent Reynold's equation relating solid lubricant "pressure" to film thickness, and applied boundary conditions relevant to bearings to describe composite behavior of solid particles entrapped between sliders. Solid lubricants, a form of granular media subject to laws of thermodynamics [76], are a site of energy dissipation, including adhesion and plastic deformations. Particle to particle interactions manifest equivalents of temperature, pressure, and other thermodynamic states. Terrell and Higgs' [77] model of the motions of particles in abrasive slurries for chemo-mechanical polishing of silicon wafers considered the slurry dilute, to decouple the fluid pressures and flow (described by Reynold's equation) from particle motions.

The third body can be considered a principal locus of dissipative processes and entropy production associated with sliding. The literature sampled above suggests that sliding subjects third body material and particles to intense loads, strains, severe plastic deformation [4], and mixing. In addition, Singer [78] identified film formation and mixing or recycling of interfacial material and wear debris as important dissipative processes associated with third bodies.

4.2. Energy considerations for friction and wear

Friction and wear are related through energy dissipated by dissipative processes. Rymuza [79] considered friction as an energy transformation process: "Friction is a process that transforms the external mechanical energy to the energy of internal processes. The synergism of processes during friction leading to formation of dissipative structures is a characteristic feature of tribological systems." Rymuza's [79] proposed coefficient of "friction energy losses," the ratio of dissipated energy to input energy, reflects friction's dissipative nature. Rearrangement and disordering of material structures by work of dissipative irreversible processes causes wear. Friction force usually provides the energy for wear, although other sources can contribute, such as electrical dissipation during sliding wear of electrical brushes [80, 81]. Past models of wear reflected this energy dependence, albeit implicitly. Early models [9] related sliding wear in brakes as a ratio $J_w = w/W$ of wear volume lost w to work $W = FL$ of friction force F , where L is distance slid. Note the similarity to equation (5). Another energy

related approach, Pv factor [17] with dimensions of power per unit surface area, implicitly describes friction power dissipation. Here P is the nominal pressure or load per projected bearing area, and v is the tangential component of velocity between the bearing surfaces. Some models employ Pv factor to predict wear, especially plastics [2]. Significant wear of boundary lubricated bearing surfaces can occur if a limit on Pv factor is exceeded [2]. Pv factor can predict wear of carbon brushes [81, 82]. Bayer [2] mentions that wear of cutting tools can be modeled with Pv factor. Several authors believe Pv factor appeared in their wear models because friction heating induces high temperatures that influence material response and wear. Material removal rate, which is surface recession per unit time, is proportional to Pv factor [83] for chemo-mechanical polishing of silicon wafers. This relation for abrasive wear [3] was first discovered by Preston. Ramalho reviewed energy dissipation methods for wear [84] and showed these methods could predict sliding wear, in addition to fretting wear. Several authors [56, 62, 59, 57, 64] showed fretting wear of oscillating contacts roughly proportional to the energy dissipated by friction force. Larbi *et al.* [85] related wear of steels to friction energy dissipated, for adhesive, abrasive, and oxidative wear. In a recent review, Briscoe and Sinha [86] identified many factors that influence polymer wear, all involved through dissipative processes.

Equation (5) suggests that wear should depend on the amount of friction energy dissipated. Uetz and Fohl [87] and Scherge *et al.* [88] noted that wear is driven by dissipation of friction, and proposed an energy balance to predict wear. Shakhvorostov *et al.* [89] monitored wear, friction, and bulk temperature for oil lubricated sliding of steel pins on a cast iron bushing, and split the power dissipated into heat carried by conduction, material lost by wear, and material transformations [88] near the surface. The amount of wear depended on the energy dissipated. Abdel-Aal [90, 91] concluded that wear aids dispersal of thermal energy dissipated by friction, since materials have limits on rate of dissipation of externally applied energy. Abdel-Aal found that the amount of wear correlated to the amount of energy dispersed, dispersal was temperature dependent, and transitions in wear rate and mechanisms depended on heat dissipation. Abdel-Aal constructed heat dissipation capacity (HDC) and specific rate of heat dissipation (SRHD) to measure heat dissipation ability. Wear rate of steels and other alloys correlated to these measures. Abdel-Aal reasoned that failure involved thermal distress from buildup of heat at the rubbing interface, and mentioned that his measures were related to entropy flow and entropy generation.

4.3. Thermodynamics of friction and wear

A thermodynamic analysis of a system focuses on exchanges of energy, entropy, and materials across the boundary of a control volume, which encloses a body of interest. A thermodynamic analysis captures the physics within the control volume, through an approach that applies conservation laws over the control volume. Because the perspective is global, a complete description of the microscopic physics occurring within the control volume is not necessary to predict system behavior, and can even be ignored. This is especially convenient when modeling systems with physics on microscopic to molecular length scales, but with effects that manifest on macroscopic length scales. Thermodynamic approaches have explained behavior of gases, liquids and solids. Since tribology has physics that occur over length scales from nanometers to millimeters, with concomitant effects that manifest on length scales from microns to meters [21] in components such as bearings, gears, brakes, and seals, a thermodynamic approach can be useful. In a series of papers Klamecki [92, 93, 94, 95] applied the laws of thermodynamics to a body undergoing wear. Klamecki [92] considered the entropy production during wear, and how the second law constrained wear. Entropy production consisted of components due to material deformation, mass loss, surface creation, and heat transfer. Klamecki [93] showed that mechanical friction interactions must be caused by dissipative processes, and assessed the entropy produced by friction about equilibrium states. In [94], Klamecki thermodynamically analyzed a sliding situation, to assess stability and entropy generation. Klamecki [95] considered sliding of metals, the work dissipated by plastic deformation about the sliding interface, and the entropy produced. In a three part paper, Zmitrowicz [96, 97, 98] applied Rational Thermodynamics [99, 100] to formulate governing equations for two bodies in sliding contact with a third body [16] in the interface. Rational Thermodynamics insures equations are consistent with conservation laws of physics and laws and assumptions of thermodynamics. The first paper [96] developed conservation equations – mass, momentum, angular momentum, energy, and entropy – for two bodies in contact with an interfacial layer or third body in between. Boundary constraints (conditions) imposed by contact between bodies were also formulated. Zmitrowicz then [97] considered constitutive laws governing behavior of materials of the sliding bodies and the interfacial layer or third body, and especially rules governing thermoelastic deformation and heat transfer over sliding contacts. Here the second law and nonnegative entropy restricted signs of material constants. Zmitrowicz finally [98] formulated constitutive laws for friction force, generation of friction heat, wear, and conduction of heat through contacts and third bodies. Again the second law

insured material constants and constitutive relations consistent with thermodynamics.

Thermodynamics is underpinned by conservation laws and balance statements. The first law of thermodynamics

$$dE = dQ - dW + \sum_k \eta_k dN_k \quad (6)$$

balances energy changes over a control volume. Here dE refers to changes in internal energy, dQ refers to heat transfer over the control volume (conduction and radiation) not involving transport of matter, dN_k refers to change of mole number or molar mass N_k (measured in moles) of species k , and $-dW$ accounts for work done on the environment by the control volume. Increments of work

$$dW = PdV + d\bar{W} = PdV + \sum_j F_j d\xi_j \quad (7)$$

involves products of generalized forces F_j and generalized displacements ξ_j , and contributions from changes in volume V induced by pressure P . Equation (6) also involves the chemical potentials for mass species k ,

$$\eta_k = \frac{\partial E}{\partial N_k} \quad (8)$$

where the partial derivatives holds all independent variables except N_k constant. Other definitions of chemical potentials replace E in equation (8) with the other thermodynamic energies, such as Gibbs free energy G , see equations (13). Change of mole number N_k , expressed as

$$dN_k = dN'_k + dN_k^e \quad (9)$$

balances increments of molar mass dN_k^e exchanged across the control volume, with increments dN'_k created or consumed as products or reactants by reactions inside the control volume. The second law of thermodynamics [101]

$$0 \leq dS' = dS - dS^e \quad (10)$$

equates changes of entropy dS within a control volume to entropy flow dS^e exchanged across the control volume and to production of irreversible entropy dS' by dissipative processes inside the control volume. The second law, equation (10), also demands non-negative production of irreversible entropy dS' . Equality in

equation (10) is restricted to reversible exchange of heat across the control volume, wherein entropy production dS' is zero. The entropy flow term

$$dS^e = \frac{dQ + \sum \eta_k dN_k^e}{T} \quad (11)$$

accounts for entropy transported across the control volume by heat and mass flows. In the absence of mass flow ($dN_k^e = 0$) such as for a closed system, $dS^e = dQ/T$. Substituting equations (9) and (7) into equation (6), grouping and substituting in that result according to equation (11), applying equation (10) to the overall result, and finally expressing derivatives in terms of time t yields

$$T \frac{dS'}{dt} = T \frac{dS}{dt} - \frac{dE}{dt} - \frac{d\bar{W}}{dt} - P \frac{dV}{dt} + \sum \eta_k \frac{dN_k'}{dt} \quad (12)$$

Equation (12) gives the irreversible entropy produced by a dissipative process operating at temperature T . The left side of equation (12), and thus the right side, is non negative. With solids and fluids, the internal energy E is often replaced with enthalpy H , Helmholtz free energy Φ , or Gibbs free energy G as follows:

$$H = E + PV, \quad \Phi = E - TS, \quad G = E - TS + PV \quad (13)$$

Equations (6)-(7), (9), (10), and (11) contain sixteen variables E , Q , W , \bar{W} , η_k , N_k , N_k' , N_k^e , S , S' , S_e , T , P , V , F_j , and ξ_j . An equation of state for the relevant thermodynamic energy (e.g., $E = E(S, V, N_k, \xi_j)$) supplies differential relations that render constitutive laws for η_k , P , T , and F_j . Equation (8) for the chemical potentials is an example. Boundary conditions for heat Q (as a temperature or heat flux condition), work \bar{W} (as a displacement ξ_j or force F condition), volume V , and mass flow N_e generate four conditions. The stoichiometry of chemical reactions, see equation (32), generates conditions for dN_k' . Overall, the variables exceed the number of equations by two, requiring two more conditions for closure. The system's dynamic state can impose conditions on the relevant thermodynamic energy and system entropy. Classical irreversible thermodynamics [102, 99] deals with systems at or near equilibrium, or at steady state. At equilibrium, entropy S maximizes and production of entropy ceases ($dS' = 0$). At a steady state, changes in entropy and energy over the control volume vanish, fulfilled by

$$dE = 0, \quad dS = 0 \quad (14)$$

Tribology involves systems not in equilibrium with motion and energy at the interface. However, many tribology systems can be considered near equilibrium, or operating at steady state. Many sliding devices running at a steady speed can be approximated as steady state, without large error in analysis [103]. Fretting and fretting wear involves low frequency oscillating motions, with stable slowly varying temperatures [55], suggesting a system near equilibrium. Here the entropy and relevant thermodynamic energy [101] can be expanded via a Taylor series about the equilibrium point S_e . For example

$$\Delta S = S - S_e = \delta S + \frac{1}{2} \delta^2 S + \dots \quad (15)$$

With the equilibrium condition, equation (14), applied to the first order term, $\delta S = 0$, the higher order terms in equation (15) govern behavior. Dai Zhendong *et al.* [56] treated fretting wear as an irreversible process in a critically stable state near equilibrium, in the process of transitioning to a new equilibrium state. Dai predicted changes induced by fretting using values for states near the local equilibrium point, obtained through the second order term in equation (15). For continuums, equations (6) to (15) have continuum counterparts [99, 96, 97, 98].

Friction force and wear manifest from irreversible dissipative processes

$$p_j = p_j(\zeta_k^j), \quad \zeta_k^j = \zeta_k^j(t) \quad (16)$$

operative at the sliding interface. These processes transform work into heat, and generate entropy. Here j indexes the process energy p_j , which depends on a set of time dependent phenomenological variables ζ_k^j indexed by k . For example, fracture driven by friction force draws energy to form new crack surface, and materials sheared between sliding surfaces derive energy from friction force. Degradation is a consequence of dissipative irreversible processes that disorder a system. The dissipative processes can be directly linked to thermodynamic entropy, or associated thermodynamic energies. Feinberg and Widom [104, 105] related material or component parameter degradation to Gibbs free energy, and predicted change in the system characteristic results with a log-time aging behavior versus time. The emerging field of damage mechanics [106] constructs a “damage parameter,” but also tracks entropy and thermodynamic energies [107, 106], to obtain closure of the physics equations needed to numerically estimate damage. Entropy can quantify the behavior of irreversible tribological processes such as friction and wear [92, 93, 94, 95, 96, 97, 98]. The amount of fretting wear of specimens in oscillating fretted contacts was shown [56, 62, 59, 57, 64] roughly proportional to the energy dissipated by friction force. To treat fretting wear, Dai

Zhendong *et al.* [56] equated the perturbed entropy flow to the perturbed entropy production, and solved for wear as the mass flux component of entropy flow. Doelling *et al.* [103] proposed a thermodynamic degradation paradigm, wherein degradation by wear under boundary lubricated conditions was related to the entropy generated during sliding. Later, Ling, Bryant and Doelling [108] analyzed the irreversible entropy produced by sliding, for application to wear.

To relate degradation to entropy production, Bryant, Khonsari and Ling [68] formulated and proved a “Degradation-Entropy Generation” theorem (DEG theorem), applicable to a generalized material degradation. Concomitant with material degradation, the second law of thermodynamics, equation (10), asserts that irreversible entropy dS' must be produced. The irreversible entropy produced, and the associated degradation must both depend upon the same dissipative processes $p_j = p_j(\zeta_k^j)$ and phenomenological variables ζ_k^j . The theorem relates degradation w to the irreversible entropy S_j' generated by the array p_j of dissipative processes. Application of the chain rule yields the rate of entropy generation

$$\frac{dS'}{dt} = \sum_{j,k} \left(\frac{\partial S'}{\partial p_j} \frac{\partial p_j}{\partial \zeta_k^j} \right) \frac{d\zeta_k^j}{dt} = \sum_j \frac{dS_j'}{dt} = \sum_{j,k} X_k^j J_k^j \quad (17)$$

expressed as a sum of products of generalized thermodynamic forces $X_k^j = (\partial S' / \partial p_j)(\partial p_j / \partial \zeta_k^j)$ and generalized thermodynamics fluxes or flow $J_k^j = d\zeta_k^j / dt$. Index j indicates individual dissipative processes, and the sum over j accumulates the entropy generated by the array of processes. Notation $\sum_{j,k}$ indicates a double sum. The term $\partial S' / \partial p_j = 1/T_j$ in X_k^j was recognized as the inverse of a temperature T_j related to p_j , suggesting that $F_k^j = \partial p_j / \partial \zeta_k^j$ is a generalized force conjugate to the generalized flow $J_k^j = d\zeta_k^j / dt$, an energy concept from system dynamics [6, 7]. By recognizing that degradation and entropy generation *must depend on the same dissipative processes and phenomenological variables* through the dissipative processes, again using the chain rule, Bryant *et al.* [68] expressed the rate of degradation

$$\begin{aligned} \frac{dw}{dt} &= \sum_{j,k} \left(\frac{\partial w}{\partial p_j} \frac{\partial p_j}{\partial \zeta_k^j} \right) \frac{d\zeta_k^j}{dt} = \sum_{j,k} \frac{\partial w / \partial p_j}{\partial S' / \partial p_j} \left(\frac{\partial S'}{\partial p_j} \frac{\partial p_j}{\partial \zeta_k^j} \right) \frac{d\zeta_k^j}{dt} = \sum_{j,k} B_j X_k^j J_k^j = \sum_{j,k} Y_k^j J_k^j = \\ &\sum_j B_j \frac{dS_j'}{dt} \end{aligned} \quad (18)$$

as products of generalized degradation forces $Y_k^j = (\partial w / \partial p_j)(\partial p_j / \partial \zeta_k^j)$ and the same J_k^j as in equation (17). The outer equalities equate the degradation rate dw/dt to a linear combination of the rates of entropy production dS'_j/dt of each of the dissipative processes p_j . In equation (18), the degradation coefficient

$$B_j = \frac{\partial w / \partial p_j}{\partial S' / \partial p_j} = \frac{dw}{dS'} \Big|_j \quad (19)$$

is a material property which, like other material properties, must be measured. The notation $|_j$ indicates that the right side of equation (19) must be measured with dissipative process j active. Although equation (17) is written in the form of the entropy generation formula of classical irreversible thermodynamics, limited to near-equilibrium and stationary thermodynamic processes [102], the theorem proof [68] expressed degradation and entropy generation in terms of the partial derivative X_k^j , J_k^j and Y_k^j , and made no assumptions on the thermodynamic state. Thus the theorem is also valid for processes far from equilibrium, the purview of Extended Irreversible Thermodynamics [100].

Wear

By enclosing the sliding surface and near surface regions of a wearing body with a tribological control volume, as shown in figure 1, Bryant *et al.* [68] then applied the DEG theorem to sliding wear and fretting wear, induced by friction force.

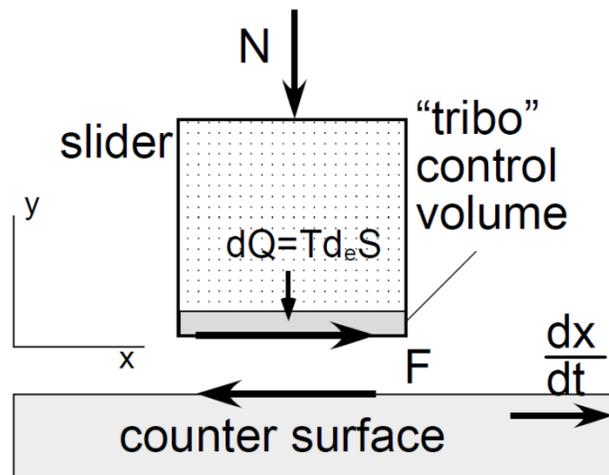


Figure 1. Thermodynamic analysis of slider on counter-surface, with tribo control volume.

Equation (17) was applied to the power dissipated by friction force F . The theorem [68] resulted in a thermodynamic based derivation of the Archard-Holm wear law (4). Sliding wear at steady speed $v = dx/dt$, driven by friction force F , was treated as a stationary thermodynamic process, wherein $dS = dE = 0$, as in equation (14). In addition, the energy ηdN^e transported out of the control volume by wear debris was neglected in the first law, equation (6), compared to the work $-dW = Fdx$ and heat exchange dQ terms. By applying equation (18) to the “tribo” control volume of figure 1, Bryant *et al.* [68] found thermodynamic force $X = F/T$, thermodynamic rate $J = dx/dt$, wear-degradation force $Y = BX$, and volumetric rate of wear per unit time to be

$$\frac{dw}{dt} = YJ = BXJ = B \frac{F}{T} \frac{dx}{dt} \quad (20)$$

Here T is the temperature at the contact between the rubbing bodies. For convenience, subscripts and superscripts were dropped. For sliding of an insulated copper rider against a steel counter surface under boundary lubricated conditions, Doelling *et al.* [103] measured wear, friction and normal forces, and temperatures at and about the sliding surface. From the temperatures was derived entropy flow dS^e/dt . Because the sliding system is steady, the entropy flow equated to the entropy generated $dS' = Fdx/T$ in the “tribo” control volume. From measurements of slider wear w arose estimates of slope dw/dS' , which is degradation coefficient B in equation (19). By comparing equation (20) to equation (4), Bryant *et al.* [68] related the wear coefficient $K = B\mu H/T$ to the degradation coefficient B , contact temperature T , friction coefficient μ , and hardness H of the wearing body. The wear coefficient thus calculated compared to within one to two percent of wear coefficients measured by Rabinowicz [48], under same conditions. Of interest is that Rabinowicz’s wear coefficient $K = 10^{-4}$, calculated via equation (4), arose from measurements of wear, forces and distance, whereas Doelling *et al.* [103] obtained the wear coefficient from measurements of wear and temperatures. The theorem also predicted the empirically observed dependence of fretting wear on friction energy dissipated in fretted specimens [62, 59, 57, 64]. Following Dai *et al.* [56], this formulation treated fretting wear as an irreversible thermodynamic process, in a critically stable state near equilibrium, but in the process of transitioning to a new equilibrium state. The formulations were similar to equations (17) through (20), except that the term Fdx/dt in equation (20), the product of power conjugates F and dx/dt [6, 7], was identified as the rate of work dW/dt dissipated in the fretted contact, similar to equation (5).

The theorem formulates degradation (and wear as a special case) in terms of the dissipative processes associated with the sliding interface, and suggests a

form for the generalized force $F_k^j = \partial p_j / \partial \zeta_k^j$. When the dissipative process p_j is mechanical, F_k^j contributes a component to friction force, consistent with equation (2). Consistent with equation (3), equation (18) suggests that total wear is the sum of components from different mechanisms. These components can be related to the dissipative processes p_j germane to the mechanisms, via the energetics embedded in the process entropy generation terms dS_j'/dt . In equation (18), dS_j'/dt acts as an “influence function” and brings the influence of the dynamics of process p_j to the degradation wear rate. Because entropy generation dS_j'/dt of dissipative processes p_j appears important, the next section will review principal dissipative processes associated with friction and wear, and concomitant entropy generation.

Friction force

This section relates friction force to the dissipative processes at the sliding interface that manifest friction force. Recall that equation (12), an amalgam of the first and second laws, balances entropy production with the dissipation from work crossing the tribological control volume, and changes in thermodynamic states. With rate of external work $-d\bar{W}/dt = Fv + N\dot{\alpha}$ arising from friction force F and sliding velocity v , and normal force N and rate of normal approach $\dot{\alpha}$ between surfaces, equation (12) gives

$$T\dot{S} - \dot{E} + Fv + N\dot{\alpha} - P\dot{V} + \sum \eta_k \dot{N}'_k = T \sum_j \dot{S}'_j \geq 0 \quad (21)$$

where “dot” denotes time derivative. Equation (21) relates the power dissipated by contact forces to entropy production and change of states. The nonnegative entropy production asserted by equation (21), along with independence of v from $\dot{\alpha}$, demands constancy of direction between friction force F and sliding velocity v . If expressions describing the entropy produced by the various dissipative processes are substituted into the right side, friction force can be estimated. To simplify the formulation, and to stress the connection of friction to dissipative processes, assume steady sliding conditions (see equations (14)), no reactions or negligible influence of reaction energy on friction force, and interfacial compressions principally due to normal force, i.e., $N\dot{\alpha} \approx P\dot{V}$. Equation (21) then simplifies to

$$F \approx \frac{T}{v} \sum_j \dot{S}'_j \quad (22)$$

Similar to equation (2), the total friction force results from a sum over friction components arising from various dissipative mechanisms operative at the sliding interface. Understanding and quantitatively describing the entropy produced by these processes or mechanisms is germane towards understanding and modeling friction. Equation (22) suggests that a quantitative description of friction force can be based on the entropy production by the associated dissipative friction mechanisms. This is also true for wear, as equations (18) and (20) suggest. Describing the entropy produced by the various dissipative processes operative at a sliding interface is the purview of the next sections.

4.4. Dissipative processes of friction and wear

Almost all processes of nature are dissipative and irreversible. Friction force is the macroscopic manifestation of dissipative processes operating on molecular to micron scales [109]. A fundamental question is *how* the organized motion of bodies, including their surface atoms, transform into the disordered thermal motions of atoms characteristic of irreversible adhesive friction. Starting in the 1990s, many investigations of the dissipative processes and mechanisms of friction have been conducted [109, 110, 111, 112]. These dissipative processes include phenomena such as conversions [110], irreversible losses due to adhesion [21, 112], plastic deformation and mechanisms in metals [113, 114], and motions of molecular chains adhered to surfaces, during collisions between chains caused by sliding [115]. Discussed in this section are dissipative processes prominent to tribology, and the associated entropy generation. Nosonovsky [15] mentions that tribological processes are in principle reversible, but become irreversible and dissipative due to hysteresis mechanisms caused by differences in length scales of the forward and reverse segments of the mechanism. Straffelini [28] extended the thermodynamic work of adhesion, plastic deformation, and other irreversible effects such as back transfer of material within a sliding contact, and showed coefficient of friction dependent on thermodynamic work of adhesion.

4.4.1. Adhesion

Adhesion is a hodgepodge of many types of bonding forces [15] operating on microscopic scales. Although these microscopic forces are in principle reversible, many factors cause adhesion to be irreversible and dissipative. Singer [109] and Krim [111] observed that friction due to adhesion usually involves dissipative processes operating on nano to atomic length scales. Pure van der Waals interactions between individual molecules are reversible, however, in large groups, motions of molecules during surface separation can couple to other allowed

vibration modes. Since the sequence of motions that originally joined the surfaces do not play backwards, the “acoustic loss” [21] process becomes irreversible. Hirano [112] formulated dissipation due to friction in terms of atomic potentials between contacting surfaces, and explained static friction force via “atomistic locking” of surfaces. Dynamic friction force was explained via movement of atoms on the surfaces, wherein kinetic energy of the slider, upon separation of interacting atoms of opposing surfaces, transfers to translational vibration modes of surface atoms. These phonon vibrations can couple into other modes associated with internal degrees of freedom. Since even a small surface has trillions of atoms, each with many internal degrees of freedom, Hirano concluded this diffusion to be irreversible, since the probability of backup of energy through the various modes into the original motion is nil. This dissipative process can also permanently displace atoms from old equilibrium positions to new positions, altering the material structure. With simple models, Streater [116] demonstrated dynamic instabilities that coupled friction power into heat loss. Harrison and Perry [115] studied friction dissipation mechanisms for boundary lubricated sliding. Nosonovsky [15] proposed a hysteresis mechanism for adhesion, wherein a dissymmetry occurs between joining and in separating of adhesion forces during sliding.

Classical theory of adhesion [20] defines an interface surface energy γ as the work per unit area to create new surface area. The energy of two solid bodies in contact over an area A is $E_c = E_o + \gamma_{12}A$, where γ_{12} is the surface energy of the interface and E_o is the baseline energy. After separating the surfaces, the energy of the two solids is $E_s = E_o + \gamma_1A + \gamma_2A$, where γ_1 and γ_2 are the surface energies to create and isolate the new free surfaces 1 and 2. To separate the two surfaces, work $W_{sc} = E_s - E_c$ must be expended. The energy per unit area to separate the surfaces is

$$\gamma = \frac{W_{sc}}{A} = \gamma_1 + \gamma_2 - \gamma_{12} \quad (23)$$

Surface energies of 1 to 3 J m⁻² are observed for clean metals, 0.1 to 0.5 J m⁻² for ionic crystal ceramics, and less than 0.1 J m⁻² for molecular crystals and polymers [20]. When two solids join and form a single crystal, $\gamma_{12} = 0$, and $\gamma_1 = \gamma_2 = \gamma_o$, giving $\gamma = 2\gamma_o$. Considering equations (12), (13), and (23) under isothermal conditions, the entropy generated by separating the surfaces against the forces of adhesion can be estimated as

$$\Delta S'_s = \frac{\gamma}{T} A_s \quad (24)$$

where T is the local temperature of the media, and A_s is the amount of free surface produced. The separation work W_{sc} in equation (23) originates from free energy, see equation (13), from work crossing the system boundary dW , from changes in the free energy $d\Phi$, or combinations thereof.

4.4.2. Plastic deformation and viscous dissipation

For dry or poorly lubricated sliding of metals, a major portion of the work dissipated by rubbing involves plastic deformation [95] and related dissipative effects [110]. The energy dissipated by plastic deformation is a major contributor to friction force [114]. Plastic flow in ductile metals occurs when the metal is overloaded in shear. Motions of dislocations drastically reduce the loads needed to cause permanent deformation [117, 118, 13]. Sliding imposes tractions onto the sliding surfaces, which can subject the sliding bodies and interfacial third body material to intense shears. Plastic deformation occurs through large surface and subsurface plastic deformation imposed by surface tractions, and surface plowing [13] observed during abrasion and asperity deformation.

For copper sliding against steel, Tsuya [119] observed intense plastic deformations in a 10 to 50 μm thick skin layer on the softer copper, and attributed the plastic dissipation in this “severely deformed region” as a principal mechanism of friction force. Rigney *et al.* [113, 120, 114] and Heilmann and Rigney [121] equated the friction energy to the plastic work dissipated, in good agreement with Tsuya’s measured friction force. Kennedy [122] measured near-surface deformations due to sliding, using microscopic observation of the contact region, and compared these values with finite element viscoplastic estimates of the high rate plastic strains of a moving contact. Kennedy suggested the thickness of the severely deformed region was less than 300 μm , for the loads in his experiment.

For metals, plastic dissipation transforms the material state and structure in the near surface regions about the asperities. Dislocations appear, move, and annihilate, effecting permanent change. Plowing removes and relocates material from one or both sliding bodies. Suh and Komvopoulos *et al.* [13, 29, 33, 32] macro and mesoscale observations identified plowing and concomitant plasticity as the dominant cause of friction and wear for sliding of metals. At smaller scales, Schmid measured forces during micro plowing of aluminum, with an atomic force microscope [123, 124, 125], and reached similar conclusions. After significant plastic flow, a mechanically mixed layer can develop within a wearing body [126], with sublayers atop the substrate that include mixed material, fragmentation, fracture, and flow. Material is usually under compression during sliding. Biswas [126] observed that instabilities within this layer governed transition to higher wear

rates. Instabilities can occur at different combinations of load and velocity, and appear to depend on temperature and strain rate.

The energy dissipated by large plastic deformation, and the entropy production can be obtained from the dissipation function of continuum mechanics [127]

$$Y = \int \sigma : \frac{d\epsilon}{dt} \approx \tau_f |\dot{\epsilon}^*_{\max}|, \quad dS' = \frac{Y}{T} \quad (25)$$

where σ is the stress tensor, $d\epsilon/dt$ is the rate of strain tensor, and the colon denotes tensor product. The form of equation (25) is valid for solids and fluids. For plastic work dissipated in solids [118], σ is associated with maximum shear, and can be approximated by the flow stress in shear τ_f , defined as the true stress needed to maintain plastic deformation at a particular true shear strain ϵ^* . Theorems can supply upper and lower bounds to the plastic work [118], if the stress field or velocity fields, respectively are known.

For sliding of ductile metals on hard metals, power loss occurs mainly in the severely deformed region, via intense localized plastic work. Rigney [113] proposed a simple estimate $Y \approx \bar{E}(\omega hs)$ of the plastic work dissipated, where $\bar{E} \approx \tau \bar{\epsilon}$ is the average deformation energy dissipated per unit volume of the severely deformed region, s is the distance slid, h is the thickness of the region, and ω is the width of the sliding track. Within the severely deformed region, τ was estimated as the yield strength in shear, and $\bar{\epsilon}$ is the average strain in that region. Equation (25) suggests the entropy produced for steady sliding can be estimated as

$$dS' = \frac{\bar{E}(\omega hds)}{T} \quad (26)$$

4.4.3. Abrasion and cutting

Abrasion, associated with plowing and adhesion friction and abrasive wear, see section 3, cuts grooves into surfaces, which forms “chips” or wear particles. Here a cutting edge moves at a depth below the surface, and severs material. Cutting or machining of metals involves three dissipative events: (1) intense plastic deformation in a shear zone directly in front of the cutting edge, (2) adhesion of the chip—material severed from the work piece—to the cutting tool, and (3) plastic deformation and fracture of the chip as the chip flows off the cutting tool. The cutting tool contains the cutting edge, which moves through and cuts the work piece. Abrasion, a form of cutting, is characterized by several dissipative processes operating together. Specific cutting energy u_s is the energy needed to machine a unit volume of material from a workpiece [128]. Specific cutting energy, measured for machining and grinding of most metals, implicitly lumps all the dissipative

processes associated with cutting of metals. Specific grinding energy u_g measures the energy to grind a unit volume of material from a workpiece. Grinding is a machining process wherein very many small abrasive grains on a grinding wheel collectively cut and remove material from a workpiece. In addition to the dissipative processes of cutting, grinding and thus u_g includes energy dissipated by fracture and wear of abrasive grains, and adhesive friction due to rubbing of the wheel's binding material against the workpiece. Since specific cutting or grinding energy represents the losses during cutting, the entropy produced per unit volume can be estimated by

$$\delta S' = \frac{u_i}{T} \quad (27)$$

where u_i is the specific energy of grinding or cutting, depending on the conditions of the abrasion.

4.4.4. Fracture

Fracture is an irreversible process wherein material, continuous within a body, separates and forms new free surface. Bodies overstressed can fracture, especially brittle materials. A crack has crack tips that intensify stresses about the crack tips. These elevated stresses can rupture material, extending the crack. For a crack of length a in a two dimensional (2D), linearly elastic brittle material at uniform temperature T , Rice [129] derived the irreversible entropy generated by the crack extending an amount da . Rice enclosed the cracked body with a control volume, formulated equation (12) in terms of Helmholtz free energy Φ , see equation (13), for the special case of constant temperature, no material loss, and no chemical reactions or phase changes, and considered $\Phi(\Delta_k, a, T) = U(\Delta_k, a) + 2\gamma_0 a$ to consist of surface energy $2\gamma_0 a$ and elastic strain energy $U = U(\Delta_k, a)$ dependent on the displacements Δ_k applied at the boundaries of the control volume, and the crack length a . Since Rice's formulation was 2D, all quantities were viewed on a per thickness basis. The work of isothermal separation of surfaces, see equation (23), involves the surface energy $\gamma = 2\gamma_0$ introduced in equation (23) of section 5.4.1. Since the summation terms in the incremental Helmholtz free energy $d\Phi = \sum_k (\partial U / \partial \Delta_k) d\Delta_k + (\partial U / \partial a) da$ equates to the work $-dW$ performed by forces $P_k = \partial U / \partial \Delta_k$ applied at the boundary, Rice solved for the entropy produced during crack extension as

$$\frac{dS'_{cr}}{dt} = \frac{G - 2\gamma_0}{T_{cr}} \frac{da}{dt} \quad (28)$$

In equation (28), $G = -\partial U/\partial a$ is the energy release rate, $2\gamma_o$ is the reversible work of separating surfaces, per unit area, and T_{cr} is the temperature of the cracked material at the crack tip. Equation (28), valid for 2D plane strain, can be extended to three dimensions (3D) by integrating the right side of equation (28) over the 3D curve through the material that defines the crack tip or crack front. Comparing equation (28) to equation (17) gives thermodynamic force $X = (G - 2\gamma_o)/T_{cr}$ and thermodynamic rate $J = da/dt$, for fracture. To apply equation (28) to tribological systems such as delamination wear or pitting, see sections 4.5 and 4.5, the strain energy U must be expressed in terms of the loads applied to the cracked body.

4.4.5. Phase changes

Phase changes morph a material to another state or structure, which alters material properties and behavior, including tribological behavior. Energy is absorbed and entropy is generated by the change. Phase changes that can be induced by sliding include melting, solidification, recrystallization, and vaporization. With change of structure, material properties such as strength, elastic modulus, and conductivity can change, altering tribological behavior. The entropy generated by a phase change [101] is

$$\Delta S = \frac{\Delta H}{T_{phase}} \quad (29)$$

where ΔH is the change in enthalpy – the latent heat absorbed or shed during the phase change – and T_{phase} is the temperature associated with the phase change. For transitions from solid to liquid, and vice versa, T_{phase} is the melting temperature, and ΔH and ΔS are the molar enthalpy and molar entropy of fusion. For transitions from liquid to gas, T_{phase} is the boiling temperature, and ΔH and ΔS are the molar enthalpy and molar entropy of vaporization. Solids can change material structures by transforming to a different crystal structure, for example, the iron changes from an alpha body centered cubic structure to the gamma face centered cubic structure at 914°. The entropy produced during a structural phase change is also given by equation (29). The rate of a phase change is governed by the reaction kinetics, which depends on the type of transformation.

4.4.6. Chemical reactions

Chemical reactions alter the composition of materials, which changes the material properties and tribological behavior of sliding bodies. Several forms of sliding wear involve chemical reactions and concomitant change of material

properties and behavior due to altered composition of interfacial components. During oxidational wear of steel [71], hardness of surface layers are reduced, which equation (4) suggests should alter wear rate. For a chemical reaction with stoichiometric equation



involving reactants R_i , $i = 1, \dots, m_r$, and products P_j , $j = 1, \dots, m_p$, in stoichiometric proportions κ_i and λ_j , Kondepudi and Prigogine [101] give the entropy produced as

$$\frac{dS'_r}{dt} = \frac{A}{T} \frac{d\xi}{dt}, \quad A = \sum_{i=1}^{m_r} \kappa_i \eta_i - \sum_{j=1}^{m_p} \lambda_j \eta_j \quad (31)$$

The chemical affinity A for the reaction depends on the chemical potentials η_k of reactants and products, see equation (8), and coefficients κ_i and λ_j in the stoichiometric equation (30). Reactants contribute positive terms to A and products contribute negative terms. Here T is the temperature of the reaction. The extent of the reaction ξ is defined such that

$$d\xi = \frac{dN'_i}{-\kappa_i} = \dots = \frac{dN'_j}{\lambda_j} \quad (32)$$

where terms arising from reactants have negative signs, terms arising from products have plus signs, and the prime on the mole numbers N'_r and N'_p indicate changes in molar mass due to chemical reactions, see the text following equation (9). Comparing equation (31) to equation (17), thermodynamic rate $J = d\xi/dt$ and thermodynamic force $X = A/T$. If equation (30) is the resultant of several successive reaction steps, then the affinity of equation (31) is the resultant affinity, and equation (31) gives the entropy production if the temperature T is maintained throughout the steps.

Formation of corrosion films on surfaces affect friction and wear [71]. Oxidation of copper to cuprous oxide, is a three step process that involves formation of copper ions, transfer of electrons to oxygen, and combination. For this reaction, stoichiometry $4 \text{ Cu} + \text{O}_2 \rightarrow 2 \text{ Cu}_2\text{O}$ summarizes the three step process, affinity $A = 2\eta_{\text{Cu}_2\text{O}} - 4\eta_{\text{Cu}} - \eta_{\text{O}_2}$, and, extent of reaction is $d\xi = -dN'_{\text{Cu}}/4 = -dN'_{\text{O}_2} = dN'_{\text{Cu}_2\text{O}}/2$.

4.4.7. Diffusion

Diffusion migrates molecules, ions, or particles of a species from regions of higher concentration to regions of lower concentration. Diffusion is driven by concentration differences or gradients, which changes the thermodynamic state, thermodynamic energies, and chemical potentials, see equations (6), (13), and (8). For diffusion from region 1 of higher concentration, to region 2 of lower concentration, the mass gained by region 2 must equal the mass lost from region 1, i.e., $dN_2 = -dN_1$. For entropy generation, diffusion is viewed as a “reaction” [101], with material in regions of higher concentration being “reactants” and material transported to regions of lower concentration being “products”. Rate of entropy production is then governed by equations (31) and (32), but with chemical potentials η_1 and η_2 principally determined by concentrations of species, and equation (32) defined by mass conservation. For diffusion, T in equation (31) is the temperature of the diffusion media. For the simple example cited, equation (32) becomes $d\xi = dN_2 = -dN_1$, affinity of equation (31) becomes $A = \eta_1 - \eta_2$, and rate of entropy production $dS'/dt = [(\eta_1 - \eta_2)/T]d\xi/dt$.

4.4.8. Mixing of materials

When materials of initially separate but different species mix, entropy is generated in the less organized mixture. Mixing pertinent to tribology occurs principally in the third body or about the interface, see section 5.1, and can involve mixing of macroscopic solid particles and/or solution of solid or fluid constituents. For particles, the irreversible entropy produced when n initially pure constituents having volumes V_α , $\alpha = 1, 2, \dots, n$ are mixed is the difference between the sum $S_i = \sum_{\alpha=1}^n S_\alpha$ of the entropies S_α of the initial states and the entropy S_f of the final mixed state [130]

$$\delta S'_{mix} = S_f - S_i = \sum_{\alpha=1}^n kN_\alpha \ln \frac{V}{V_\alpha} \quad (33)$$

where $k = 1.38066 \times 10^{-23}$ J/K is Boltzmann’s constant, N_α is the number of particles of species α , and final volume V contains the mixture. For liquids and gases, $V = \sum_{\alpha=1}^n V_\alpha$, but packing factors may be needed to estimate V for solids or particles. Since $V/V_\alpha \geq 1$, the entropy produced is positive. Equation (33) was derived for mixing of ideal gases initially segregated into separate volumes, V_α , but was extended by Muller and Weiss [130] to mixing of other pure “particles”. For tribology, this could include third body debris, lubricant powders, or liquid dispersions. Derivation of equation (33) applied Boltzmann’s formula $S = k \ln(\omega)$, where ω is the probability of occurrence of a

micro-state. Since the probability of a specific particle (e.g., gas molecule or wear particle) being in volume V_α – and not other parts of V – is V_α/V , the probability of N_α independent particles being in V_α is $\omega_\alpha = (V_\alpha/V)^{N_\alpha}$. Boltzmann's formula renders $S_\alpha = kN_\alpha \ln(V/V_\alpha)$. The final fully mixed state, with very many micro-state manifestations and probability near unity, has vanishing S_f .

The molar entropy \bar{S}'_{sol} of a solution of n initially pure constituents, expressed in terms of molar fractions $x_\alpha = N_\alpha/N$, was presented in [101] as

$$\delta\bar{S}'_{sol} = -R \sum_{\alpha=1}^n x_\alpha \ln(x_\alpha) \quad (34)$$

where R is the universal gas constant. Here \bar{S}'_{sol} is the entropy produced per mole N . The ideal gas law suggests different gases at same temperature and pressure have same ratio of volume to number of moles, giving $x_\alpha \approx V_\alpha/V$. Multiplying and dividing the right side of equation (33) by Avogadro's number N_o , and dividing both sides by N renders equation (34), since $R = kN_o$ and $N_\alpha = N_\alpha/N_o$.

The entropy production rates can be obtained from equations (33) and (34) by dividing those entropy differences by the time required for mixing.

4.4.9. Heat transfer

Transferring an amount of heat dQ from a body at a higher temperature T_h to a body at a lower temperature T_l generates entropy. The entropy of the hotter body reduces by dQ/T_h , and the entropy of the cooler body increases by dQ/T_l . The system is isolated, giving entropy flow $dS^e = 0$ from the exterior. The entropy change is $dS = dQ/T_l$. Via equation (10), the entropy generated (in the cooler body) by the transfer of heat dQ is

$$dS'_Q = dQ \left(\frac{1}{T_l} - \frac{1}{T_h} \right) \quad (35)$$

With reference to equation (18), $J = dQ/dt$ and $X = 1/T_l - 1/T_h$. Heat flows from regions of high to low temperature. For temperature differences $T_h - T_l$, Fourier heat conduction gives $J = dQ/dt = \beta(T_h - T_l)$, where $\beta = AK/\ell$ is the thermal conductance, consisting of thermal conductivity K , cross section area A , and length ℓ .

5. Summary

Friction and wear mechanisms were reviewed, analyzed, and related in terms of their associated thermodynamics, and energy losses and entropy produced by

common dissipative processes. Friction force relates to the entropy generated by those dissipative processes, see equation (22). Wear rate was expressed by the Degradation-Entropy Generation theorem as a linear combination of entropy generation of the dissipative processes, for example see equations (18) and (20). Dissipative processes of interest to dry sliding were introduced, and the irreversible entropy generated by these processes were quantitatively presented, see section 5.4. In consideration of equations (22), (17) and (18), the expressions for entropy generation presented in section 5.4 for the dissipative processes operative at a sliding interface, and *common to both friction and wear*, could result in *unified* friction and wear models.

6. Acknowledgement

The author gratefully acknowledges support from the U.S. Office of Naval Research, MURI Grant N00014-04-1-0599 RQM and the Accenture Endowed Professorship in Manufacturing Systems Engineering, University of Texas at Austin.

References

1. Blau, P. J. 1996. *Friction Science and Technology*. Marcel Dekker, Inc., New York.
2. Bayer, R. G. 1994. *Mechanical Wear Prediction and Prevention*. Marcel-Dekker, Inc., New York, NY, USA.
3. Rabinowicz, E. 1965. *Friction and Wear of Materials*. John Wiley & Sons, Inc., New York, NY, USA.
4. Hwang, D., Kim, D., and Lee, S. 1999. *Wear*, **225-229**, 427.
5. Czichos, H. 1978. *Tribology, A Systems Approach to the Science of Friction, Lubrication and Wear*. Elsevier Scientific Publishing Company, New York, NY, USA.
6. Karnopp, D. C., Margolis, D. L., and Rosenberg, R. C. 1990. *System Dynamics, A Unified Approach*. Wiley Interscience, New York, NY, USA; 2nd edition.
7. Brown, F. T. 2001. *Engineering System Dynamics – A Unified Graph Centered Approach*, Marcel-Dekker, New York, NY, USA.
8. Bowden, F. P. and Tabor, D. 1950. *The Friction and Lubrication of Solids*. Clarendon Press, Oxford, UK.
9. Kragelskii, I. 1965. *Friction and Wear*. Butterworth and Co., Bath, UK.
10. Tabor, D. 1987. *Tribology-Friction, Lubrication, and Wear – Fifty Year On*. Institution of Mechanical Engineers, London, UK; volume 1 of C245, pp. 157-172.
11. Tabor, D. 1995. *Tribol. Int.*, **28**, 7.
12. Blau, P. J. 2001. *Tribol. Int.*, **34**, 585.
13. Suh, N. P. 1986. *Tribophysics*. Prentice-Hall, Englewood Cliffs, NJ, USA.
14. Bejan, A. 1988. *Advanced Engineering Thermodynamics*. John Wiley & Sons, Inc., New York, NY, USA.

15. Nosonovsky, M. and Bhushan, B. 2008. *Multiscale Dissipative Mechanisms and Hierarchical Surfaces: Friction, Superhydrophobicity and Biomimetics*. Springer-Verlag, Berlin, Germany.
16. Godet, M. 1984. *Wear*, **100**, 437.
17. Halling, J. 1978. *Principles of Tribology*. Macmillan Press Ltd, Hong Kong.
18. Ovcharenko, A., Halperin, G., and Etsion, I. 2008. *ASME J. Tribol.*, **130**, 021401.
19. Ruths, M. and Israilachvili, J. N. 2008. *Surface Forces and Nanorheology of Molecularly Thin Films*. Springer-Verlag, Berlin, Germany; 2nd edition, pp. 417-497.
20. Maugis, D. 1999. *Contact, Adhesion and Rupture of Elastic Solids*. Springer, Berlin, Germany.
21. Johnson, K. L. 1994. *The Mechanics of Adhesion, Deformation, and Contamination in Friction*. Elsevier, New York, NY, USA, pp. 21-33.
22. Shi, X. and Polycarpou, A. A. 2006. *ASME J. Tribol.*, **128**, 841.
23. Xue, X. and Polycarpou, A. A. 2007. *J. Colloidal and Interface Science*, **311**, 203.
24. Greenwood, J. and Williamson, J. 1966. *P. Roy. Soc. Lond. A Mat.*, 300.
25. Johnson, K. L. 1985. *Contact Mechanics*. Cambridge University Press, Cambridge, UK.
26. Holm, R. 1946. *Electric Contacts Handbook*. Almquist and Wiksells, Stockholm, Sweden, section 40.
27. Tabor, D. 1948. *P. Roy. Soc. Lond. A Mat.*, **192**, 247.
28. Straffelini, G. 2001. *Wear*, **249**, 78.
29. Komvopoulos, K., Saka, N., and Suh, N. P. 1985. *ASME J. Tribol.*, **107**, 452.
30. Komvopoulos, K., Suh, N. P., and Saka, N. 1986. *Wear*, **107**, 107.
31. Komvopoulos, K., Saka, N., and Suh, N. P. 1986. *ASME J. Tribol.*, **108**, 502.
32. Komvopoulos, K., Saka, N., and Suh, N. P. 1987. *ASME J. Tribol.*, **109**, 223.
33. Komvopoulos, K., Suh, N. P., and Saka, N. 1986. *ASME J. Tribol.*, **108**, 301.
34. Bhushan, B. 2002. *Introduction to Tribology*. John Wiley & Sons, Inc., New York, NY, USA; 3rd edition.
35. Bryant, M. D., Tewari, A., and Lin, J. W. 1995. *IEEE Transactions on Components, Hybrids and Manufacturing Technology-Part A*, **18**, 375.
36. Bryant, M. D., York, D., and Tewari, A. 1998. *Wear*, **216**, 60.
37. Bryant, M. D. and York, D. 2000. *ASME J. Tribol.*, **122**, 374.
38. Lin, J. W. and Bryant, M. D. 1996. *ASME J. Tribol.*, **118**, 116.
39. Kato, K. 2002. *P. I. Mech. Eng. J – J. Eng. Tribol.*, **216**, 349.
40. Ludema, K. C. 1995. *J. Korean Soc. Tribol. Lubr. Eng.*, **11**, 10.
41. Meng, H. C. and Ludema, K. C. 1995. *Wear*, **181-183**, 443.
42. Hsu, S. M. and Shen, M. C. 1996. *Wear*, **200**, 154.
43. Maeda, K., Bismarck, A., and Briscoe, B. J. 2005. *Wear*, **259**, 651.
44. Briscoe, B. J., Evans, P. D., Pelillo, E., and Sinha, S. K. 1996. *Wear*, **200**, 137.
45. Briscoe, B. J., Pelillo, E., and Sinha, S. K. 1996. *Polym. Eng. Sci.*, **36**, 2996.
46. Archard, J. F. 1953. *J. Appl. Phys.*, **24**, 981.
47. Archard, J. F. 1980. *Wear theory and mechanisms*. Am. Soc. Mech. Eng., New York, NY, USA, pp. 35-80.
48. Rabinowicz, E. 1980. *Wear theory and mechanisms*. Am. Soc. Mech. Eng., New York, NY, USA, 486.
49. Yamamoto, T. and Buckley, D. H. 1983. *Tribol. Trans.*, **26**, 277.

50. Lee, Y.-Z. and Ludema, K. C. 1990. *Wear*, **138**, 13.
51. Suh, A. Y., Patel, J. J., Polycarpou, A. A., and Conry, T. F. 2006. *Wear*, **260**, 735.
52. Hutchings, I. M. 2002. *Proc. I. Mech. Eng. J – J. Eng. Tribol.*, **216**, 55.
53. Godet, M. 1990. *Wear*, **136**, 29.
54. Waterhouse, R. B. 1984. *Wear*, **100**, 107.
55. Szolwinski, M. P., Harish, G., Farris, T. N., and Sakagami, T. 1999. *ASME J. Tribol.*, **121**, 11.
56. Zhendong, D., Shenrong, Y., and Qunji, W. M. X. 2000. *J. Nanjing Univ. Aeronautics & Astronautics*, **32**, 125.
57. Mohrbacher, H., Blanpain, B., Celis, J. P., Roos, J. R., Stals, L., and Stappen, M. V. 1995. *Wear*, **188**, 130.
58. Huq, M. Z. and Celis, J. P. 2002. *Wear*, **252**, 375.
59. Fouvry, S., Kapsa, P., Zahouani, H., and Vincent, L. 1997. *Wear*, **203-204**, 393.
60. Fouvry, S. and Kapsa, P. 2001. *Surf. Coat. Tech.*, **138**, 141.
61. Fouvry, S., Liskiewicz, T., Kapsa, P., Hannel, S., and Sauger, E. 2003. *Wear*, **255**, 287.
62. Fouvry, S., Paulin, C., and Liskiewicz T., 2007. *Tribol. Int.*, **40**, 1428.
63. Lee, H., Mall, S., Sanders, J. H., and Sharma, S. K. 2005. *Tribol. Lett.*, **19**, 239.
64. Magaziner, R. S., Jain, V. K., and Mall, S. 2008. *Wear*, **264**, 1002.
65. Gallego, L. and Nelias, D. 2006. *ASME J. Tribol.*, **128**, 476.
66. Gallego, L., Nelias, D., and Jacq, C. 2007. *ASME J. Tribol.*, **129**, 528.
67. Liu, Y., Xua, J.-Q., and Mutoh, Y. 2008. *Int. J. Mech. Sci.*, **50**, 897.
68. Bryant, M. D., Khonsari, M. M., and Ling, F. F. 2008. *P. Roy. Soc. Lond. A Mat.*, **464**, 2001.
69. Uhlig, H. H. and Revie, R. W. 1985. *Corrosion and Corrosion Control*, John Wiley & Sons, Inc., New York, NY, USA.
70. Stott, F. H. 1998. *Tribol. Int.*, **31**, 61.
71. Quinn, T. 2002. *Tribol. Int.*, **35**, 691.
72. Iordanoff, I., Berthier, Y., Descartes, S., and Heshmat, H. 2002. *ASME J. Tribol.*, **124**, 725.
73. Iordanoff, I., Seve, B., and Berthier, Y. 2002. *ASME J. Tribol.*, **124**, 530.
74. Richard, D., Iordanoff, I., Berthier, Y., Renouf, M., and Fillot, N. 2007. *ASME J. Tribol.*, **129**, 829.
75. Jang, J. Y. and Khonsari, M. M. 2005. *P. Roy. Soc. Lond. A Mat.*, **461**, 3255.
76. Herrmann, H. 1993. *J. de Physique II*, **3**, 427.
77. Terrell, E. J. and Higgs, C. F. 2007. *ASME J. Tribol.*, **129**, 933.
78. Singer, I. L. 1998. *Mat. Res. Bull.*, **23**, 37.
79. Rymuza, Z. 1996. *Wear*, **199**, 187.
80. Bryant, M. D. 1991. *IEEE Transactions on Components, Hybrids and Manufacturing Technology*, **14**, 71.
81. El-Refaie, A. M. F., Aziz, M. M. A., Khorshid, S. A. Y., and Elzahab, E. E. M. A. 2000. *IEEE T. Energy Conver.*, **15**, 176.
82. Hu, Z., Chen, Z., Xia, J., and Ding, G. 2008. *Wear*, **265**, 336.
83. Zhou, C., Shan, L., Hight, J. R., Danyluk, S., Ng, S. H., and Paszkowski, A. J. 2002. *Tribol. Trans.*, **45**, 232.
84. Ramalho, A. and Miranda, J. 2006. *Wear*, **260**, 361.

85. Larbi, A. B. C., Cherif, A., and Tarres, M. 2005. *Wear*, **258**, 712.
86. Briscoe, B. J. and Sinha, S. K. 2002. *Proc. I. Mech. Eng. J – J. Eng. Tribol.*, **216**, 401.
87. Uetz, H. and Fohl, J. 1978. *Wear*, **49**, 253.
88. Scherge, M., Shakhvorostov, D., and Pohlmann, K. 2003. *Wear*, **255**, 395.
89. Shakhvorostov, D., Pohlmann, K., and Scherge, M. 2004. *Wear*, **257**, 124.
90. Abdel-Aal, H. A. 2003. *Wear*, **255**, 348.
91. Abdel-Aal, H. A. 2005. *Wear*, **259**, 1372.
92. Klamecki, B. E. 1980. *Wear*, **58**, 325.
93. Klamecki, B. E. 1980. *Wear*, **63**, 113.
94. Klamecki, B. E. 1982. *Wear*, **77**, 115.
95. Klamecki, B. E. 1984. *Wear*, **96**, 319.
96. Zmitrowicz, A. 1987. *Wear*, **114**, 135.
97. Zmitrowicz, A. 1987. *Wear*, **114**, 169.
98. Zmitrowicz, A. 1987. *Wear*, **114**, 198.
99. Haupt, P. 1993. *Thermodynamics of Solids*. Springer-Verlag, New York, NY, USA, pp. 66-138.
100. Lebon, G. 1993. *Extended Thermodynamics*. Springer-Verlag, New York, NY, USA, pp. 139-204.
101. Kondepudi, D. and Prigogine, I. 1998. *Modern Thermodynamics From Heat Engines to Dissipative Structures*, John Wiley & Sons, Inc., New York, NY, USA.
102. Jou, D., Casas-Vasquez, J., and Lebon, G. 1996. *Extended Irreversible Thermodynamics*. Springer-Verlag, Berlin Heidelberg, Germany; 2nd edition.
103. Doelling, K. L., Ling, F. F., Bryant, M. D., and Heilman, B. P. 2002. *J. Appl. Phys.*, **88**, 2999.
104. Feinberg, A. A. and Widom, A. 1996. *IEEE T. Reliab.*, **45**, 28.
105. Feinberg, A. A. and Widom, A. 2000. *IEEE T. Reliab.*, **49**, 136.
106. Voyiadjis, G. 1999. *Advances in Damage Mechanics*, Elsevier, New York, NY, USA.
107. Bhattacharya, B. and Ellingwood, B. 1999. *Int. J. Solids Struct.*, **36**, 1757.
108. Ling, F. F., Bryant, M. D., and Doelling, K. L. 2002. *Wear*, **253**, 1165.
109. Singer, I. L. 1994. *J. Vac. Sci. Technol.*, **12**, 2605.
110. Robbins, M. O. and Krim, J. 1998. *Mat. Res. Bull.*, **23**, 23.
111. Krim, J. 2002. *Surface Science*. **500**, 741.
112. Hirano, M. 2006. *Surface Science Reports*. **60**, 159.
113. Rigney, D. A. and Hirth, J. P. 1979. *Wear*, **53**, 345.
114. Rigney, D. A. and Hammerberg, J. E. 1998. *Mat. Res. Bull.*, **23**, 32.
115. Harrison, J. A. and Perry, S. S. 1998. *Mat. Res. Bull.*, **23**, 27.
116. Streator, J. L. 1994. A molecularly based model of sliding friction. Elsevier, New York, NY, USA, Tribology Series **27**, pp. 173-183.
117. Weertman, J. and Weertman, J. R. 1964. *Elementary Dislocation Theory*, Macmillan, London, UK.
118. Calladine, C. R. 1985. *Plasticity for Engineers*. Halsted Press, John Wiley & Sons, Inc., Berlin Heidelberg, Germany.
119. Tsuya, Y. 1976. Microstructures of wear, friction, and solid lubrication. Technical Report 81, Mech. Eng. Lab., Igusa Suginami-ku, Tokyo, Japan.

120. Rigney, D. A., Chen, L. H., Naylor, M. G. S., and Rosenfield, A. R., 1998, *Wear*, **100**, 195.
121. Heilmann, B. P. and Rigney, D. A. 1981. *Wear*, **72**, 195.
122. Kennedy, F. E. 1989. *Key Eng. Mat.*, **33**, 35.
123. Hector, L. G. and Schmid, S. R. 1998. *Wear*, **215**, 247.
124. Schmid, S. R. and Hector, L. G. 1998. *Wear*, **215**, 257.
125. Opalka, S. M., Hector, L. G., Schmid, S. R., Reich, R. A., and Epp, J. M. 1999. *ASME J. Tribol.*, **121**, 384.
126. Biswas, S. K. 2002. *Proc. I. Mech. Eng. J – J. Eng. Tribol.*, **216**, 357.
127. Frederick, D. and Chang, T. S. 1965. *Continuum Mechanics*. Allyn and Bacon, Inc., Boston, MA, USA.
128. Kalpakjian, S. 1991. *Manufacturing Processes for Engineering Materials*. Addison-Wesley, Reading, MA, USA; 2nd edition.
129. Rice, J. R. 1978. *J. Mech. Phys. Solids*. **26**, 61.
130. Muller, I. and Weiss, W. 2005. *Entropy and Energy – A Universal Competition*. Springer-Verlag, Berlin Heidelberg, Germany.



Research Signpost
37/661 (2), Fort P.O.
Trivandrum-695 023
Kerala, India

Recent Developments in Wear Prevention, Friction and Lubrication, 2010: 197-226
ISBN: 978-81-308-0377-7 Editor: George K. Nikas

5. Tribofilms – On the crucial importance of tribologically induced surface modifications

Staffan Jacobson and Sture Hogmark

*The Tribomaterials Group at Ångström Laboratory, Uppsala University, Box 534,
SE-751 21 Uppsala, Sweden*

Abstract. The interface between two solid bodies in sliding contact is an extreme environment. The local conditions typically result in substantial changes of the composition and properties of the mating surfaces. These tribologically induced surface modifications have decisive effects on the tribological performance of very different mechanical components and tools. They include topography changes (smoothing or roughening), formation of micro-cracks, phase transformations, deformation hardening, formation of oxides, formation of solid films by reactions with lubricant additives, transfer of material from the counter surface, and so on. The thickness of these layers and films range from atomic monolayers to tens of micrometres. Due to these modifications, it is *not* the original materials but the strongly modified layers that provides the wear resistance and friction level of face seals, brake pads, cutting tools, rock drills and so on.

This chapter gives an overview of this crucially important area. Due to its complexity and enormous range, this is done by presenting illustrative examples covering different materials combinations from a wide range of tribological applications.

1. Introduction

The interface between two solid bodies in sliding contact is an extreme environment. The local pressures, stresses and temperatures are high enough to break the bonds of even the strongest materials. Under such tribological contact the materials deform plastically and fracture on the micro scale. Atoms, molecules or larger particles are transferred between the two surfaces and to the surfaces from the environment or lubricant. The outermost layers become mixed, chemical reactions take place between the elements involved and new compounds are formed. This inferno of activities on the atomic scale and upwards often results in

the formation of films or layers of completely new materials on the surfaces, so-called tribofilms, as illustrated in Fig. 1.

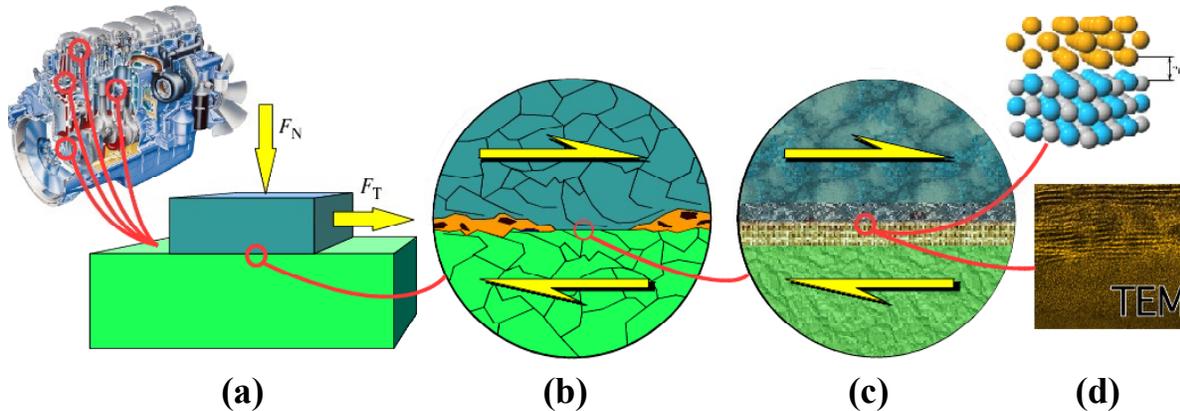


Figure 1. The real contact between tribological components (a) takes place on the surface asperity level, often within areas just a few μm wide (b). The local pressures in these micro contacts are typically several GPa. The materials may change properties down to considerable depth due to various phase transformations and deformation hardening. The outermost tribofilm (c) typically becomes nano-crystalline or amorphous and only 10-50 nm thick. The relative movement between the bodies is typically accommodated through shear deformation within this thin film, which of course results in extreme shear rates. This thin film has a decisive influence on the friction level and wear rate. It may today be studied using high resolution TEM and be predictable using atomic scale theory (d).

Hence, a significant characteristic of almost any type of dry or boundary lubricated tribological component is that the tribological properties will change dramatically during use, as a consequence of the modification of the original interface. The change may be very rapid, and thus often not even noticed, and in other cases noticeable as a slow running-in resulting in better performance of a component, a vehicle, etc. Very often the new surface materials exhibit totally different friction and wear properties compared to the original bulk materials [1–6]; the friction may be reduced or increased, the wear resistance may be better or worse.

The modifications include topography changes (smoothing or roughening), formation of micro-cracks, phase transformations, deformation hardening, formation of oxides, formation of solid films by reactions with lubricant additives, transfer of material from the counter surface, and so on. The thickness of these layers and films range from atomic monolayers (e.g. hydrogen termination of diamond surfaces) to tens of μm (e.g. plastic deformation of metals).

Recent developments of surface analysis techniques and high-resolution electron microscopy have revealed very thin solid films on surfaces where they previously could not be detected or analyzed. These films are often nano crystalline and partly amorphous and in the thickness range of 10–50 nm [6–11]. This chapter presents information obtained by scanning electron microscopy SEM, focused ion beam preparation (FIB) for transmission electron microscopy (TEM), scanning transmission electron microscopy (STEM), and electron energy loss spectroscopy (EELS).

Since the tribological properties of tools, wear parts, mechanical components and whole systems such as vehicles, are determined by these modified surfaces rather than by the original, they deserve attention and careful assessment. Without knowledge about how these surface layers are formed and how they modify the tribosystem, it is not possible to predict the friction and wear properties of a material in a given tribological situation. Further, the development of new materials, coatings and lubricants will be much more efficient if it is clearly understood that the surfaces will always be modified during use, and that the development should be focused on optimizing the properties after this modification.

This chapter gives an overview of this crucially important area – tribologically induced surface modifications. Due to its complexity and enormous range, it is not possible to give a full coverage. We rather try to provide insights by presenting selected illustrative examples covering very different materials combinations and a selection from a wide range of tribological applications.

1.1. Definition of tribofilm

Due to their complexity and multifaceted nature, these phenomena have been given many names such as transfer films, built-up layers, third bodies [12], tribolayers, tribosintered layers, selective transfer layers, self-organizing surface films [5], etc. The relations between these terms and their exact definitions have not been very clear. In the present chapter we apply the term *tribofilm* very broadly, to cover *all tribologically modified surface layers*.

2. Mechanisms of tribofilm formation and surface modifications

One way of organizing the tribofilm phenomena is presented in Figure 2. Here, they are classified based firstly on whether they are of the surface transformation or deposition type. For the first type of tribofilm, the outermost region of the original material is transformed due to diffusion, plastic deformation or just

frictionally heat treated. Such tribofilms have no sharp interface towards the underlying material.

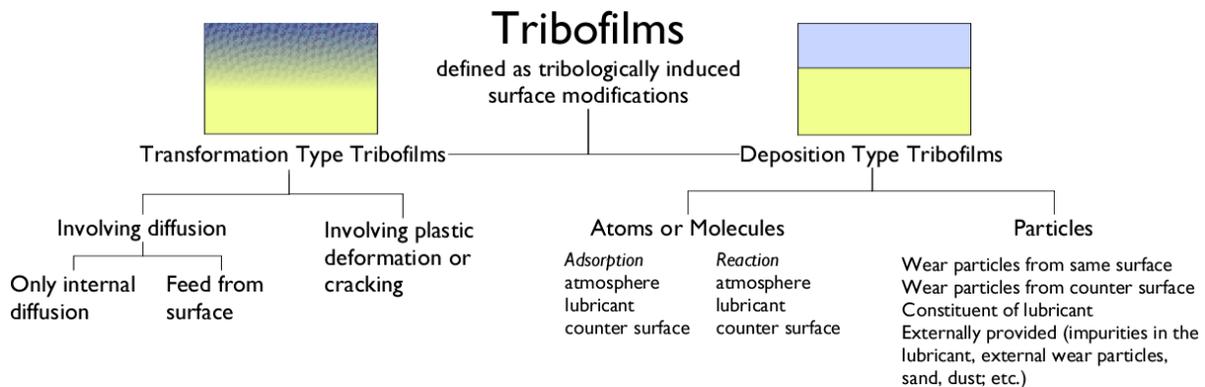


Figure 2. Classification of tribofilms based on whether they are formed by transformation of an existing surface layer or by formation on top of the original surface. Also combinations of these two sets of mechanisms are common.

The deposition type of tribofilms is formed on top of an existing surface by pick-up of particles, adsorption of molecules, chemical reactions, etc. These processes lead to a relatively well-defined interface, and correspondingly phenomena such as adhesion and delamination become more important here. Of course some tribofilms do not fall very clearly within one or the other type, and may involve both diffusion and pick up of particles, etc.

The *transformation type tribofilms* include two subgroups.

- (1) Films formed by transformation of the original surface involving diffusion and possibly chemical reaction and phase transformations. The diffusing atoms may originate from within the material itself or being fed from the interface, originating from the mating surface or surrounding media. This diffusion leads to local changes of composition and phase (such as second phase precipitation or dissolution), segregation, recrystallisation, grain growth, etc.
- (2) Films formed by transformation of the original surface without any material transfer or diffusion, but based on plastic deformation and crack formation. Such mechanisms lead to deformation hardening, contact fatigue, sub surface crack generation, texture formation, topography changes, grain refinement, etc. Frictional heating may cause martensitic transformations in the surface of carbon steels without diffusion across the interface.

The *deposition type tribofilms* also include two subgroups;

- (1) Films formed from individual atoms or molecules, where these are fed from the countersurface, from the lubricant or the environment and may just

adsorb on the surface, or may react chemically on the surface or with the surface atoms. Both are common mechanisms of many lubricant additives.

- (2) Films formed from larger particles that stick to or become integrated in the surface. These may be wear debris originating from the surface or the countersurface, or may be externally provided via the lubricant, etc. The particles may simply become entrapped in the interface and pressed into the surface. They may just be loosely mechanically bonded or become intimately integrated in the structure through cold-welding, sintering and similar mechanisms.

3. A short overview of typical tribofilms

Tribofilms are generated in countless applications and appear in numerous forms. A presentation of some of the most common tribofilms is given below.

3.1. Tribocorrosion films

Tribocorrosion can be said to be a special family of tribofilm formation mechanisms. Here, thin films are formed by corrosive chemical reactions with components from the environment. Typically, a combination of transformation and deposition type tribofilm is generated, which may give some protection against wear, but may also be much less wear resistant than the original surface. During wear, synergistic or antagonistic effects of mechanical wear and corrosion may result in removal rates that are much lower or much higher than the sum of the individual mechanisms [13, 14].

3.2. Films on oxide ceramics

During wear of oxide ceramics, at least two types of tribofilms are often reported [2, 15-21]. One is a very thin amorphous film, the other, which is often thicker, is partly crystalline or nano crystalline and is composed mainly from compacted or sintered wear debris [22]. The small thin rolls often found on ceramics worn in the mild regime have been proposed to be formed by small removed patches of the amorphous type tribofilm, which have become curled due to residual stresses within the film [22]. Typically, oxide ceramics form deposition type tribofilms.

3.3. Tribofilms on low-friction coatings

Many of the popular low-friction coatings rely on the formation of a tribofilm. The detailed mechanisms and chemistry behind the formation and low-friction properties have been elegantly studied using in situ optical microscopy and Raman spectroscopy using a transparent slider [23]. By this approach, Singer et al. demonstrated that the three low-friction coatings investigated (amorphous Pb-O-Mo, DLC and annealed boron carbide) all were lubricated by thin tribofilms.

Recently several low-friction coating concepts have been developed, which all rely on formation of different tribofilms to provide the low friction [24-30]. It has been demonstrated that the low-friction performance often found for diamond in humid atmospheres is dependent on the formation and replenishment of a molecularly thin film [31–33]. So, contrary to its reputation of being a low-friction material, sliding between naked diamond surfaces *always* gives high friction. The friction forces are high enough to destroy the surfaces. It is only when the surfaces are masked by a layer of adsorbed or reacted molecules or atoms that the friction becomes low.

This layer readily forms in most atmospheres; the water content in air of normal humidity is sufficient. Adsorbed water molecules will mask the strong bonds, i.e. avoid covalent bonding between the contacting surfaces, and thus keep the friction low [34–36].

3.4. Tribofilms from ZDDP and other lubricant additives

Perhaps the most well-known and most studied tribofilms are those formed on ferrous metals in the presence of lubricants with the zinc dialkyldithiophosphate ZDDP additive [37–39]. This, the most successful lubricant additive in motor oils, was introduced over 60 years ago. It is still unbeaten as former of protective films on steel surfaces in combustion engines. These films are of the deposition type.

3.5. Transformed and transferred layers ("friction layers") on metals

In many cases metals can be considered to be smart materials; they respond to severe deformation by strengthening their structure in the surface layer. As is well known, the properties of metallic materials may alter considerably due to plastic deformation, oxidation and phase transformation. In fact, all mechanisms known for strengthening of bulk crystalline materials [40] can be activated locally in the surface layer during tribological contact. This includes both transformation type tribofilm formation in the form of deformation hardening (dislocation strengthening), grain refinement (Hall-Petch strengthening, also due to

deformation), hard phase generation (e.g. martensitic transformations in steel), solute strengthening (alloying by diffusion), and deposition type tribofilm formation in the form of particle strengthening (e.g. oxide particle intermixing), composite strengthening (pick-up of counter material or wear particles), etc.

The success of metals as tribological materials is partly explained by the fact that these changes often act to improve the wear resistance [41, 42]. Selected examples of hardness and structure changes in structural steels subjected to severe abrasion have been presented in [43] and [44].

4. Illustrative examples

4.1. A unique tribofilm of the transformation type accounts for the low-friction behavior of Stellites

Stellites – a family of Co based chromium alloyed metals – are well known for their good low-friction and anti-galling behavior in heavily loaded, unlubricated contacts. These properties make Stellites very popular as hardfacing materials for components in demanding tribological applications such as heavy duty valves, turbines, spindles, etc. [45]. In dry contacts they typically generate friction coefficients of the order of 0.20–0.25, and show a very good resistance against material pick-up and galling.

This performance has often been attributed to beneficial thermal and chemical/corrosive properties. However, with the aid of modern analytical techniques, Persson has recently revealed the mechanism behind this behavior [45]. Self-mated Stellite 21 surfaces exposed to dry reciprocating sliding, exhibited a phase transformation from bcc to hcp down to a depth of about 100 μm . Also, a substantial hardness increase was detected in this layer. In the most superficial ≈ 10 nm layer, the hcp basal planes were further more lined up parallel to the surface (see Fig. 3a). The latter process offers easy dislocation glide parallel to the sliding direction. In this way, the ideal situation of a hard, load bearing layer (i.e. minimized real contact area) with an easily sheared surface film is created. The whole procedure is repeated as soon as the superficial layer is worn off. The fcc to hcp transformation occurred also at lower contact pressures, but no alignment of the basal planes along the sliding direction was observed (cp. Fig. 3b), and the friction coefficient was higher around 0.4.

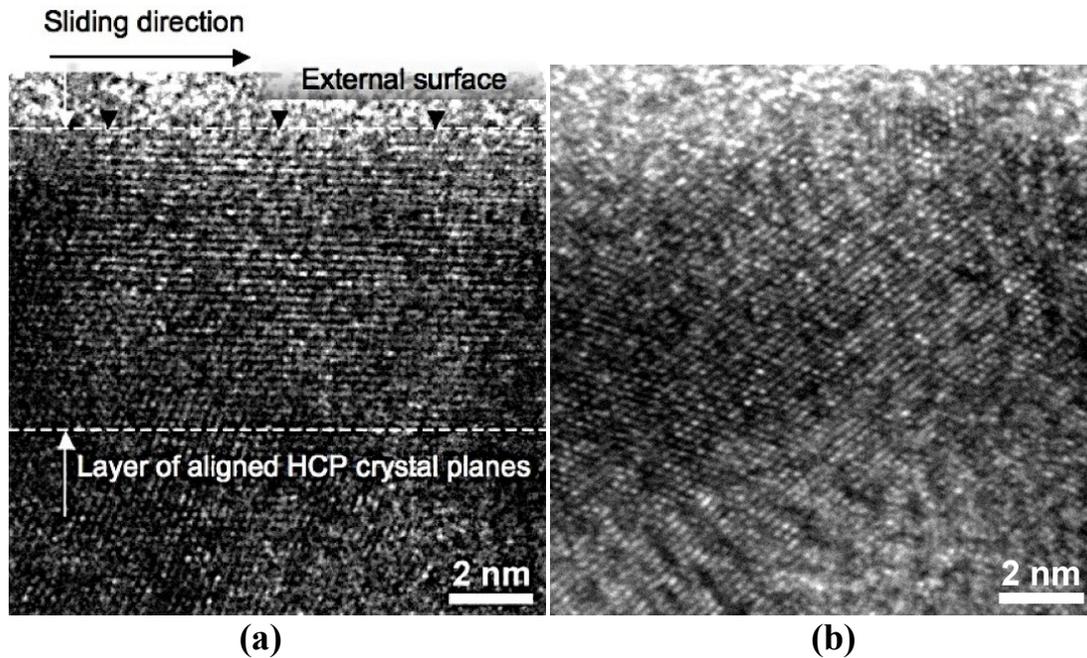


Figure 3. High-resolution TEM micrographs of cross-sectioned surface layers of Stellite 21. (a) Surface deformed by heavily loaded dry sliding. Notice the key to the low-friction behavior – the formation of the 10 nm thin, easily sheared top layer within which the hcp basal planes have become oriented parallel to the worn surface. (b) Ground surface. As in (a), the outermost atomic layers have completely transformed from fcc to hcp. However, unlike the case of the heavily loaded sliding contact no preferential orientation can be observed [45]. (Stellite 21: in wt% 27 Cr, 0.25 C, 5.5 Mo, 1.5 Si, 1.0 Mn, 2.5 Ni, < 3 Fe, bal. Co.)

Conclusion

Three separate transformation type tribofilm mechanisms explains the unique tribological behavior of Stellite. Firstly, the stress induced deformation hardening and secondly phase transformation of the Co matrix from the fcc structure of the bulk to hcp in the surface layer. Thirdly, in the outermost layers the shear stress acts to align the hcp structure with its easy-to-shear basal planes parallel to the sliding interface.

This model as illustrated in Fig. 4 [45, 9] explains the intrinsic low-friction and anti-galling properties of the Stellites in heavily loaded sliding contacts. The oxide layer does not play any major role in the mechanisms described, indicating that Stellites would function equally well in water, vacuum and other oxygen free environments.

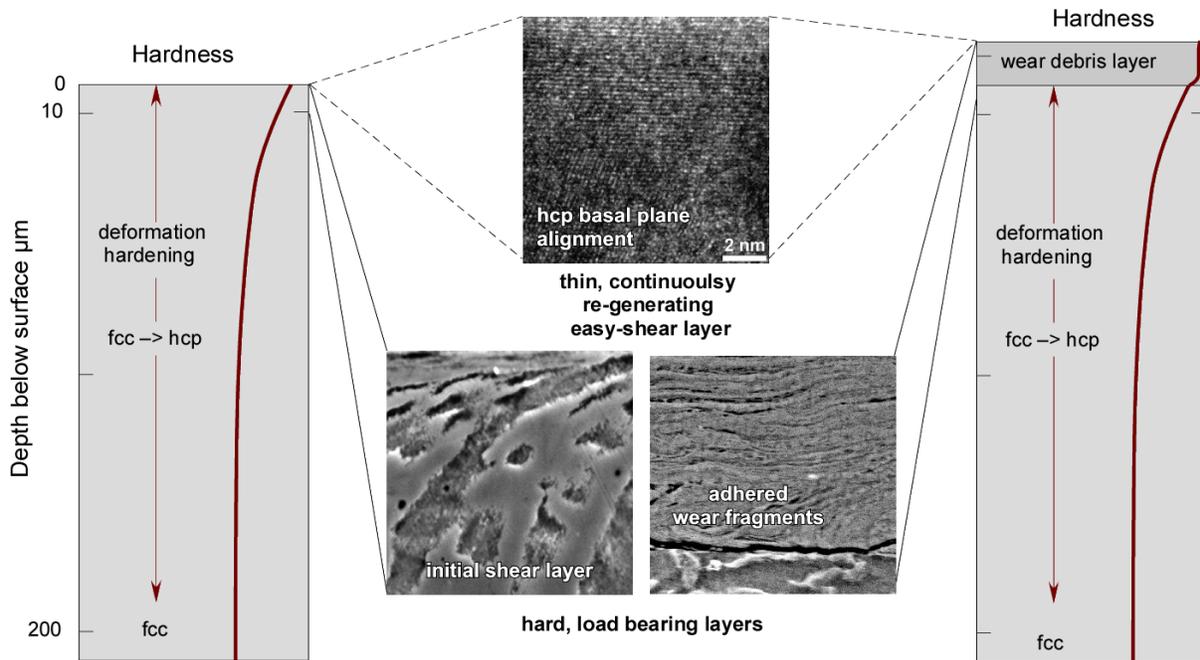


Figure 4. Schematic view of the self-generating low-friction system of Stellites typical of highly loaded sliding contacts. The hardness is increased down to substantial depth and the ultra-thin (just a few atomic layers) easily sheared aligned hcp layer continuously regenerates at the surface (left graph). The easy-shear layer forms equally well on surfaces involving a layer of accumulated wear debris (right graph) [45, 9].

4.2. Galling on forming tools is caused by problematic tribofilm formation

Plastic forming of austenitic stainless steel has proven to be one of the most demanding tooling operations. The reason is the extreme tendency of this type of work material to adhere to the tool. Lumps of steel adhered to the tool surface will cause damage to the next products to be formed. This process of adhesion and subsequent damage is usually named galling.

It is well documented that the primary reason behind most of the galling problems in industrial forming is related to a rough tool topography [46,47]. For sufficiently smooth tools, the general belief has been that adhesion occurs when any oxide or lubricant is removed from the interface such that the work and tool materials are making metallic contact on the atomic level.

However, recent studies in a Load Scanner test [48] revealed an oxidic interlayer between adhered stainless steel and the tool [8]. It was even shown that, on some areas of the sliding track, no *metal* from the stainless steel work piece had

become transferred but the tool surface was covered with *oxide* transferred from the stainless steel.

One conclusion is that the adhesion of the stainless steel oxide (mainly CrO) to the tool steel is at least as strong as the adhesion to its parent steel, another is that metallic contact is not an absolute prerequisite to adhesion and galling.

Previously unpublished results presented in figures 5-7 reveal that there is an oxidic interlayer between the tool and transferred work material also for industrial tools that have been used to form a large number of austenitic stainless steel sheet components.

The tool is made of a powder, metallurgical, cold-working steel with the nominal composition (wt%) of 9V, 4.5W, 5 Cr and 4 Mo. It was hardened and tempered to about 60 HRC. The stainless steel is of the AISI 304 type (18.5 Cr and 10 Ni). Areas where stainless steel had adhered were identified in the SEM, and specimens for TEM-studies were prepared using a focused ion beam (FIB) instrument, see Fig. 5.

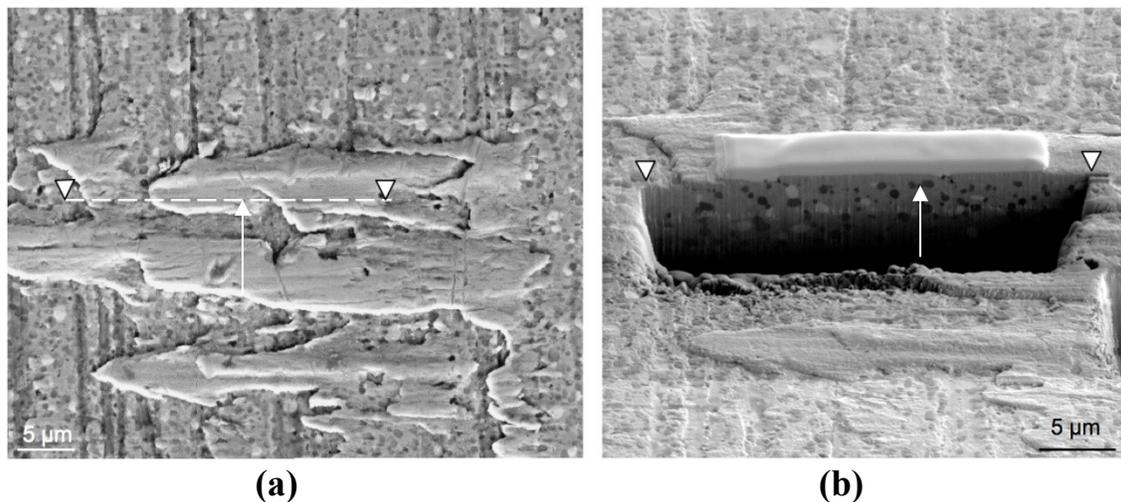


Figure 5. (a) Appearance of PM tool steel surface where some patches of austenitic stainless steel have adhered during sheet forming. The direction of sheet sliding during forming was from left to right. (b) A crater in the tool surface formed by the FIB along the dashed line in (a). A similar crater was subsequently formed slightly above the dashed line to leave a thin disk of tool steel with adhered stainless steel as the TEM specimen. Thus, this specimen is perpendicular to the tool surface and contains the interface between tool and work material. The arrows in (a) and (b) indicate the location of the TEM pictures in figures 6 and 7.

TEM studies (figures 6 and 7) reveal an oxidic interlayer (20-60 nm thick) between the tool steel and the adhered austenitic stainless steel. It is made up of a combination of Fe, Cr and V oxides, and contains carbon from the forming oil.

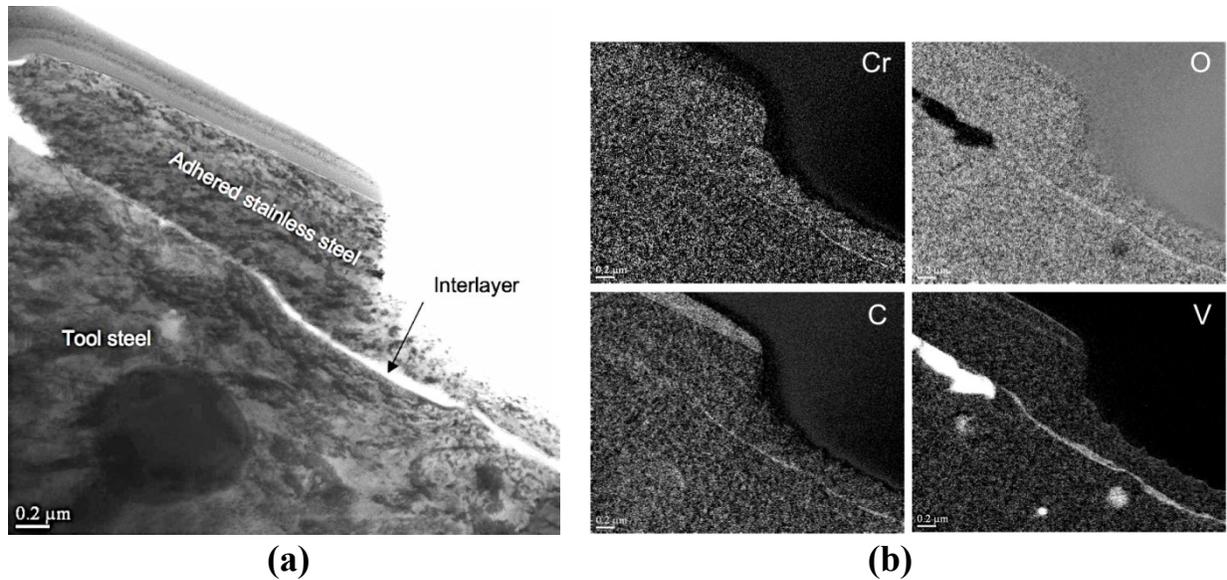


Figure 6. (a) TEM cross-section micrograph showing austenitic stainless steel adhering to the tool steel. The light contrast of the interlayer between the tool steel and the stainless steel is due to its amorphous structure in combination with its high concentration of light elements (mainly O and C). (b) Elemental mapping using electron energy loss spectroscopy (EELS) of the same area as in (a). Note the slightly reduced magnification.

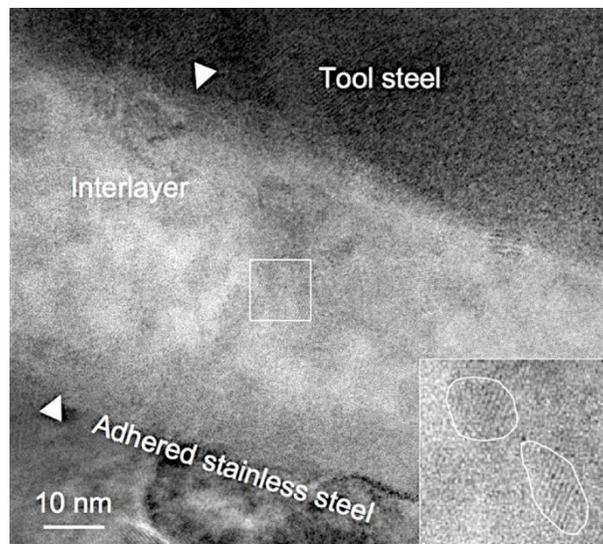


Figure 7. High resolution TEM of the interlayer between adhered stainless steel and tool steel shown in Fig. 6. Note that the structure is in the form of an amorphous matrix with nanometer-sized crystallites (encircled in the magnified insert).

Conclusion

Formation of a tribofilm of the deposition type is responsible for the galling phenomenon. It is made up of two layers; one thin very fine grained or amorphous oxide layer next to the tool surface followed by a heavily deformed fine grained layer of stainless steel. The strength of the thin oxide layer is obviously higher than the cohesive strength of the austenitic stainless steel. Usually, the strong galling tendency of austenitic stainless steels is attributed to their combination of strong ability for deformation hardening, their thin natural oxide layer (of the order of 5-10 nm) and their relatively low heat conductivity (about half of that of carbon steel). However, we have shown that the formation of an extremely strong and adhesive oxidic tribofilm may be the primary explanation. This oxide obviously acts as a very strong weld or glue between the tool steel and the stainless steel. It is composed of elements (fragments) from both tool and work material (Fe, V, Cr) and from reactions with the environment (O and C).

4.3. Tribofilms protect metal cutting tools

Machining of steels can be very demanding for the cutting tool. This is illustrated by the wear mechanisms activated when cutting a H14 type of hot working tool steel with a cemented carbide (CC) tool [44, 7]. The SEM and FIB technique was used to make cross-sections through the rake face of a cutting insert. After having identified an area where some work material (tool steel) had adhered, a TEM specimen was produced by ion beam milling, see Fig. 8. After cutting loose the TEM specimen and preparing it by further ion beam thinning, STEM imaging was performed – see figures 9 to 11. In this mode, EDS was used to determine the elemental composition of interesting features around the tool steel – CC tool interface, cp. Fig. 10. The structure of the interface was best studied by conventional TEM. The WC grains, Co-binder and adhered steel could be resolved – see Fig. 11. It is revealed that the wear of CC occurs by a mechanism involving formation of a deposition type of tribofilm. (In the metal cutting community usually denoted built-up layer.) The film is continually deposited and removed by plastic shear. Although being formed by tool steel – normally a much weaker material than the CC – the shearing tribofilm is strong enough to gradually remove the Co-binder of the CC material. Subsequently, the hard WC grains also become fragmented and removed as part of the shearing tribofilm – see Fig. 10. The strength of the tribofilm is primarily due to its extremely fine grain size, cp. Fig. 11.

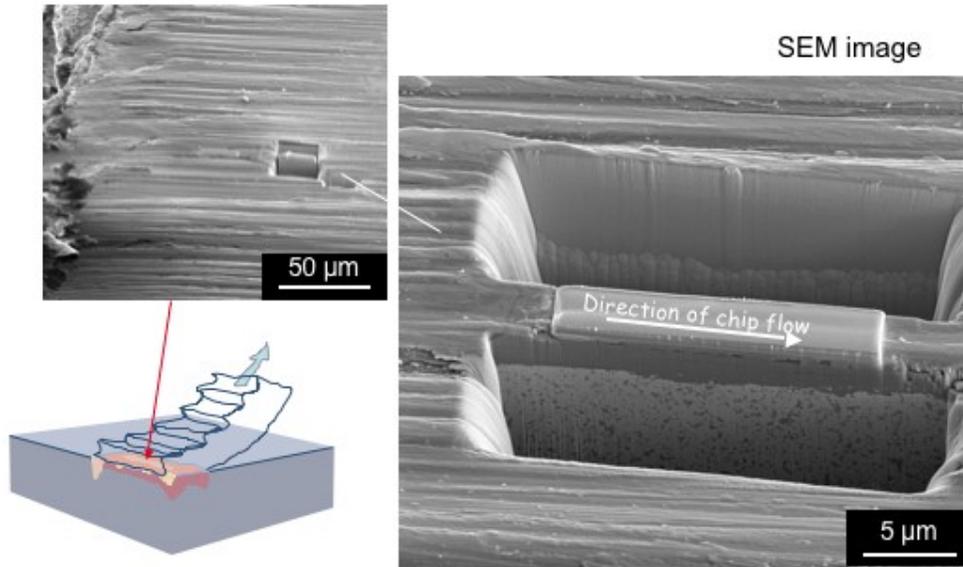


Figure 8. SEM micrographs of the rake face of a cemented carbide CC cutting tool showing (in two magnifications) where a cross-section is made using the FIB-SEM. The tool has been cutting in carbon steel, some of which is adhered to the rake face.

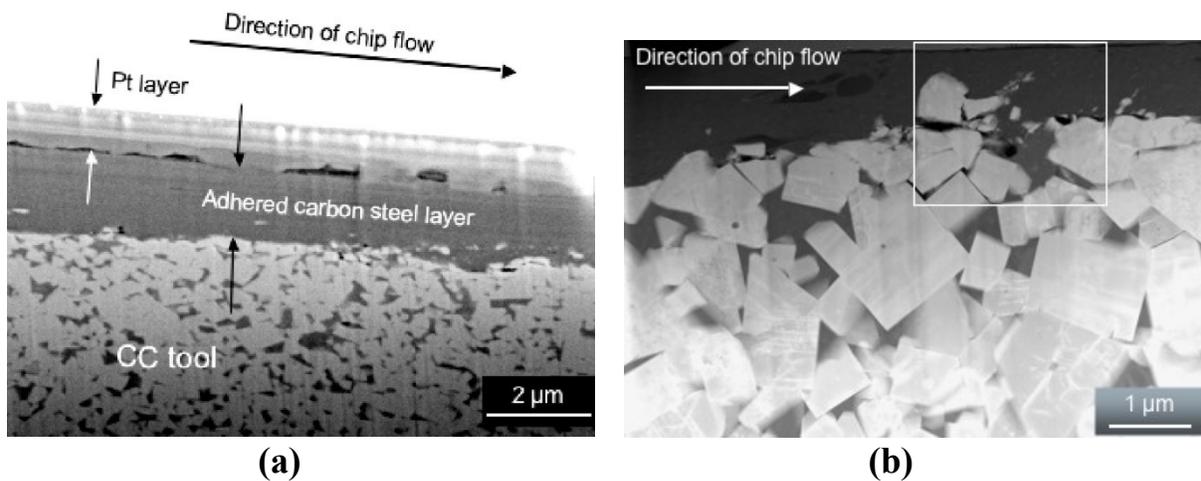


Figure 9. Close-up of the cross-section in Fig. 8. (a) The top layer of Pt is deposited to protect the external surface of the tribofilm (mainly consisting of adhered steel) during the preparation. SEM (b) STEM micrograph of the tribofilm/CC interface region. (CC = WC grains in a matrix of Co.) The light particles are WC grains of the CC, the intermediate grey material is the Co binder, and the dark grey layer is the transferred carbon steel. The area for further TEM studies (Fig. 10) is indicated by the white rectangle.

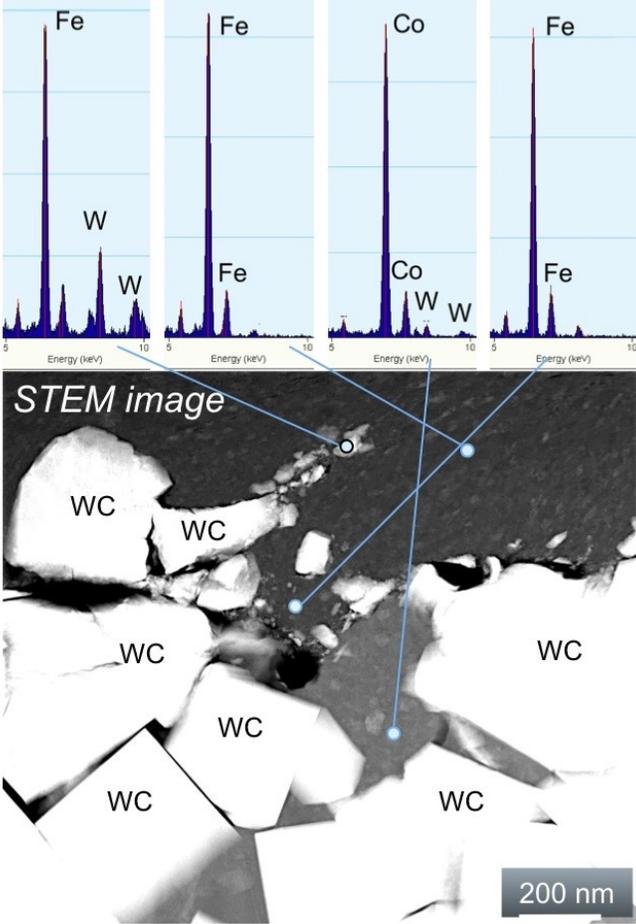


Figure 10. STEM micrograph and selected EDS spectra of the tribofilm/CC interface region. Note the entrainment of steel into the cemented carbide structure, the fragmentation of the outermost WC grains and the inclusion of the small fragments into the shearing adhered steel layer, see also Fig. 11.

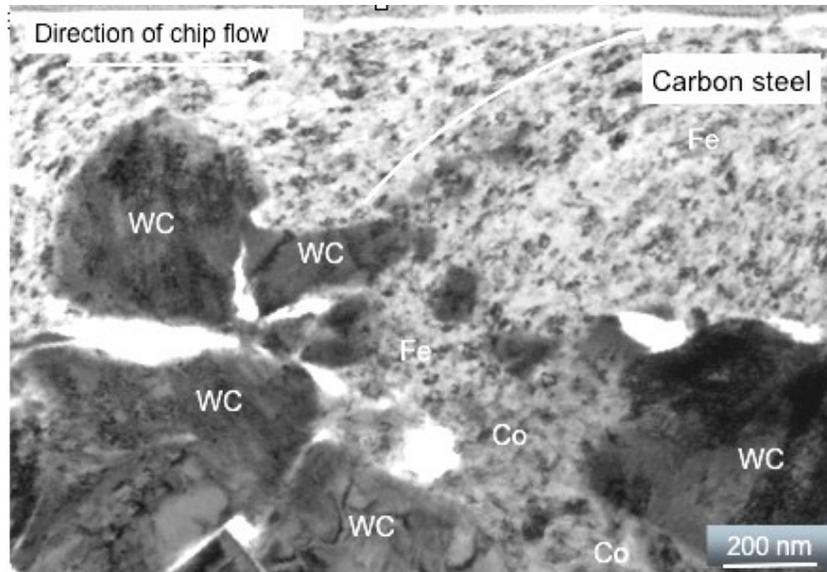


Figure 11. TEM micrograph revealing the crystallographic structure of the interface region between the adhered work material (carbon steel) and the CC tool material. White areas represent voids in the CC material formed by the deformation involved in the removal of the solid WC grains. (Same area as in Fig. 10.)

Some metallic materials such as aluminium and titanium alloys and austenitic stainless steel are generally regarded as being difficult to machine since they easily form strong built up layers. For the stainless steels, this has led to the development of so called free-machining steels, i.e. grades that have been alloyed to contain a small proportion of, for instance, inclusions of manganese sulphide (MnS). (Formed from Mn and S added to the steel.) During cutting, a thin tribofilm of MnS will adhere to the tool and act as a solid lubricant, and protect the tool from wear. However, free-machining grades have a lower corrosion resistance and slightly reduced ductility as compared to their non-free machining equivalents because of the presence of these non-metallic inclusions.

Conclusion

Tribofilms associated with metal cutting are usually of the deposition type. They are generally believed to protect the cutting tool from wear even if they contribute to the wear mechanisms. However, if they grow in thickness to form built up edges, they may cause chipping of the cutting edge and also degrade the surface roughness of the cut product.

4.4. Tribofilm formation on self-mated ceramic face seals – optimum recycling of wear debris

Face seals are used in many harsh applications, for instance in submersible pumps where an electric motor has to be sealed from the outside media. If this media contains hard particles such as sand, it is not possible to use rubber seals. Instead, seals in the form of two disks with central holes are used (see insets in figures 12 and 13). The seals are pressed against each other face to face and have to be flat and smooth to effectively prevent e.g. water to leak from the high-pressure side to the low-pressure side during the rotation [2, 20, 21, 49]. Typical materials in such components are cemented carbide (CC), alumina (Al_2O_3) and silicon carbide (SiC).

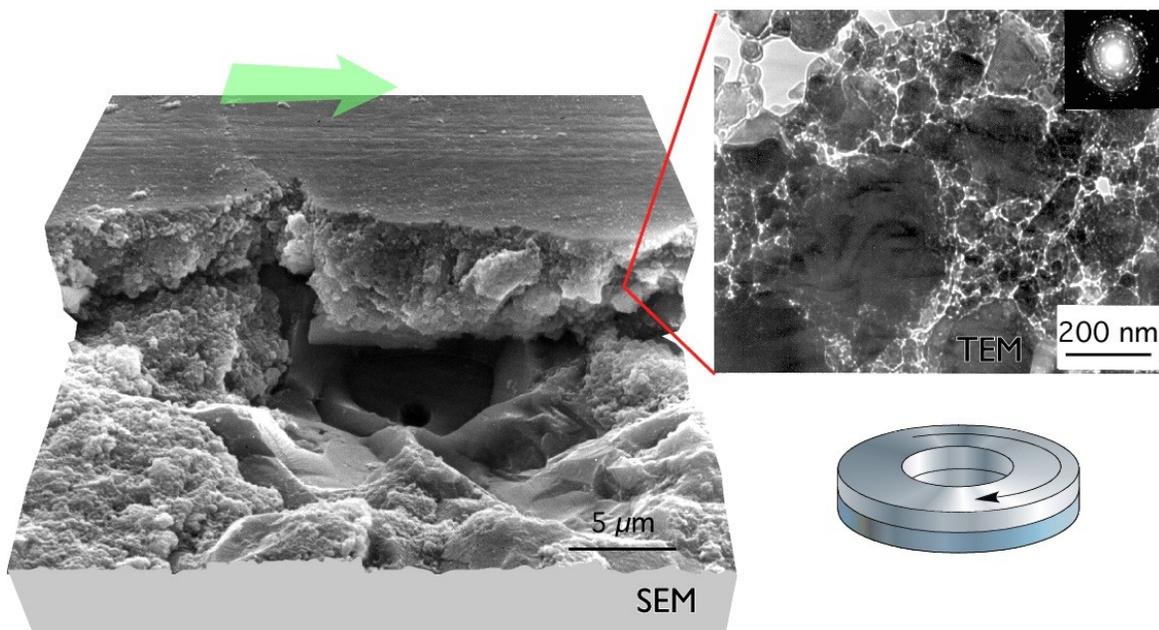


Figure 12. Tribofilm formation on a ceramic (alumina) face seal. The wear debris is to a high degree retained in the closed contact between the two plane seals, which give a high probability to become recycled as a wear protective tribofilm. The TEM micrograph reveals a tribosintered deposition type of film consisting of relatively loosely bonded alumina fragments [2, 20, 21]. The arrow indicates the sliding direction of the counter body.

Figure 12 shows a detail of the surface of an alumina face seal. During use, a 5 μm thick film of alumina debris has formed. A small part of this almost fully covering film is present in the upper part of the figure. Note that the top surface of the tribofilm is much smoother than the worn bulk surface beneath (lower part of

the figure). This reveals a positive effect from this type of film formation (less leakage). Another obvious benefit is wear protection.

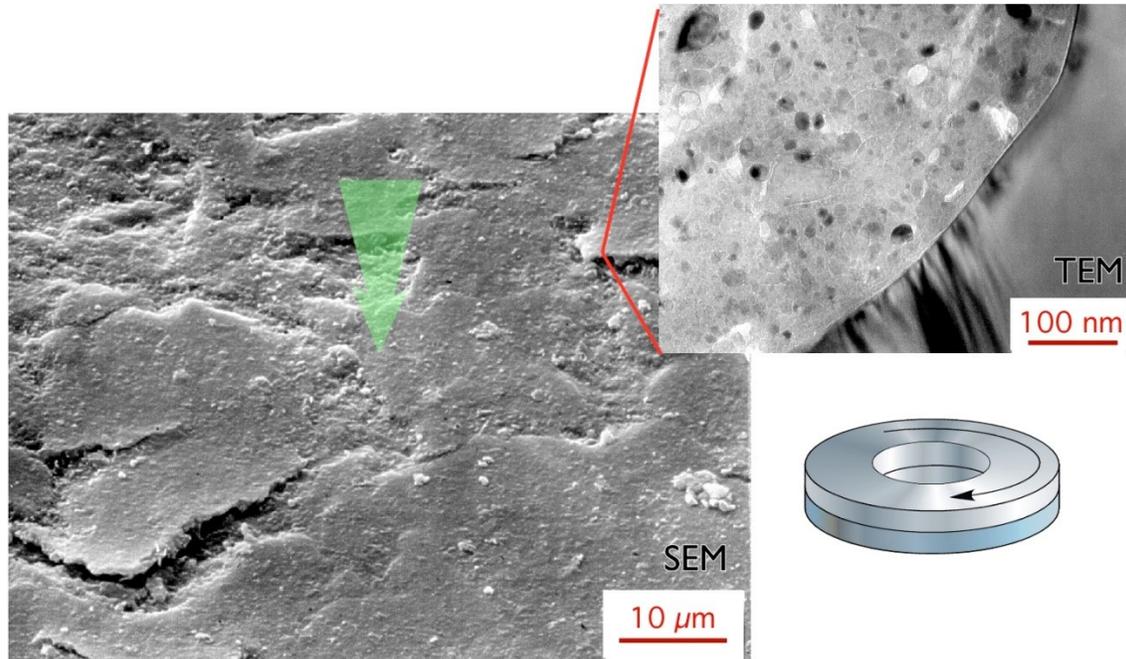


Figure 13. Tribofilm formation on a ceramic (SiC) face seal exposed to sliding in dry air. SiC against SiC results in a film of SiC nano-crystals embedded in a dense SiO₂ matrix, firmly bonded to the underlying unworn SiC grain [2, 20, 21]. Left SEM picture of the top surface, showing an almost entirely covering tribofilm. Top right TEM picture showing the dense structure of the tribofilm and the intimate bond towards the SiC grain. The arrow indicates the sliding direction of the counter body.

Figure 13 shows a similar set of micrographs representing a face seal of SiC that has been running under conditions identical to those of the alumina seal. Also here a deposition type tribofilm is formed. It is smooth and covers almost the entire nominal contact area. Compared to that of alumina, it is thinner and denser. The TEM micrograph shows the interface between the tribofilm (upper left) and the SiC material (lower right). The interface is very sharp. The structure is revealed to contain nano sized SiC particles (dark grains in the image) embedded in a matrix of amorphous SiO₂. This means that the SiC debris have gradually oxidized during the tribosintering process [20,21]. The positive effect of tribofilm formation is here threefold; smoothening (reduced leakage), wear protection and friction reduction. Typically, the dry friction level associated with this SiO₂-based tribofilm is of the order of 0.2 to 0.3.

Conclusion

The tribofilms from these two examples of self-mated ceramic materials are of the deposition type. During running-in, they effectively improve the function of the seals by smoothening the contact surfaces, and they also protect the surface from further wear. For the SiC seals the films also contribute to a significantly reduced friction, which means a reduced energy loss through the seal.

4.5. Tribofilm formation on drive elements of micromotors provides the necessary grip

Ultrasonic motors are often very attractive in applications where miniature size, high speed, good precision and low power consumption are essential features. Traditional motor technology uses gearboxes or lead screws and nuts to accomplish linear movement. Such miniaturized systems are complicated to manufacture and assemble, while piezoelectric motors based on direct friction drive systems are well adapted to miniaturization.

The principle of one such ultrasonic piezoelectric motor is shown by the insert in Fig. 14. When appropriate drive signals are applied to the piezoelectric elements, they oscillate in such a way that the attached cylindrical drive pads describe an elliptical motion. To achieve the linear motion, the drive pads transfer the movement to the drive rail by gripping and pushing it forward during each half cycle. The rail is pushed approximately 1 μm per cycle, at a frequency of 96 kHz. The total stroke is in this case 6 mm. The pads and rail consist of alumina, selected based on its relatively low wear rate and relatively high coefficient of friction.

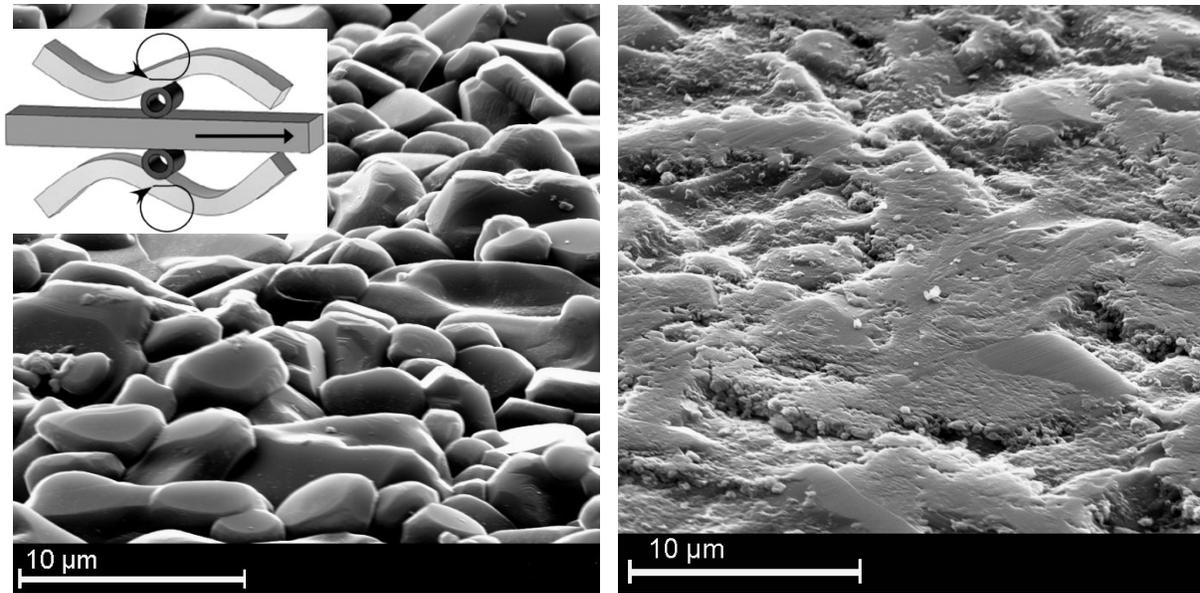


Figure 14. Appearance of the alumina drive rail from the friction drive system of a linear piezoelectric micromotor. Left: the as sintered surface before use. Right: after about 350,000 strokes of the rail. Inserted: schematic of the drive system consisting of a driven central alumina rail and two piezoelectric vibrating beams that transfer their vibrating movement to a linear translation of the rail via two hollow alumina drive pads [16].

The as sintered surfaces initially wear very rapidly, resulting in flattening of the most protruding alumina grains, and correspondingly high production of wear debris. This debris rapidly fills the cavities between the grains and is efficiently squeezed and worked to form a dense tribofilm – see Fig. 15. The film becomes mostly amorphous and partly nano crystalline, and develops a very intimate bond to the unworn grains [15, 16]. Unprotected edges of the film will easily become worn off again, so even if the initially rapid wear rate of the ceramic soon levels off, there will be some circulation of debris that first form tribofilms, then fracture again into debris, participate in the new film formation, and so on.

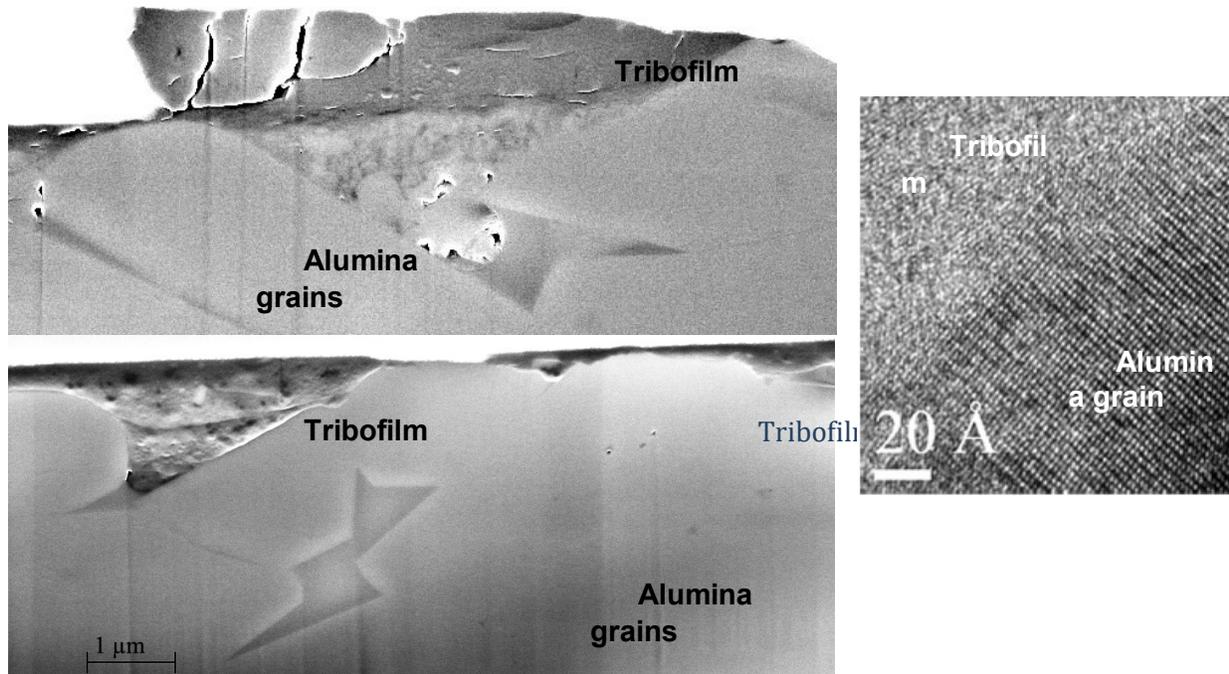


Figure 15. Cross sections of a drive rail after about 200,000 strokes. A relatively thick tribofilm consisting of sintered or compacted wear debris fills out the cavities between the unworn alumina grains. A thinner tribofilm partly covers the top parts of the alumina grains that have become worn flat [16]. Top: tribofilm with several cracks indicating that this unprotected part of the tribofilm is repeatedly worn and rebuilt (SEM). Bottom: tribofilm stabilized by the protecting surrounding alumina grains (left) and thin film partly covering worn area of the alumina grains (right) (SEM). Right: high resolution TEM illustrating the virtually perfect bond between the mainly amorphous tribofilm filling out a cavity and the neighboring crystalline alumina grain (TEM).

Conclusion

In this case the tribofilm gives some wear protection, and furthermore it provides a higher friction coefficient than the original surface. The high friction coefficient is necessary for the proper operation of the friction driven motor [15, 16].

4.6. Complex tribofilms bring the car to a halt

The disc against pad interface of automotive brakes is a fascinating world of intense tribological activity. Surfaces are wearing and reforming, local contact areas are born, grow, mature and die while legions of micron sized and nanometer

sized particles rush along the constantly reshaping maze formed between the irregular pad and relatively flat disc – see Fig. 16.

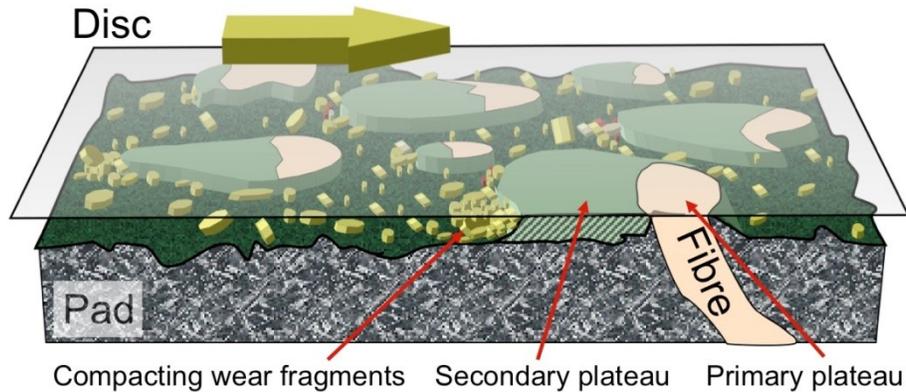


Figure 16. Schematic of the contact situation between brake disc and pad according to the plateau model. The disc sliding direction is from left to right. Protruding hard phase constituents, i.e. primary plateaus of the pad are white, compacted debris in the form of secondary plateaus are grey. A constant flow of wear debris in the gap between pad and disc wears the lower regions of the pad through three body abrasion and supply the secondary plateaus with new material. Occasionally secondary plateaus break down, releasing heavily deformed particles back to the flow of wear debris [50].

The brake must deliver a relatively high and totally reliable coefficient of friction, regardless of speed, temperature, pressure, presence of water, contaminants, etc. It must deliver this friction without too much wear, without seizure, while causing minimum vibrations and squeal [51].

Primarily based on an unconventional study involving direct observation of the contact interface through a glass disc, Eriksson et al. [52] could describe the buildup and breakdown of the contact plateaus formed in the gap between disc and pad. Wear debris was observed compacting against hard phases, e.g. steel fibers and quartz particles, of the brake pad forming flakes of agglomerated particles. Based on this and a series of other studies, a model for the formation of the contact interface was presented [50, 53]. The mechanism proposed to cause the formation and growth of contact plateaus is described by the plateau model, illustrated in Fig. 16. It involves a pad surface in sliding contact against a disc, where protruding hard phases, called the primary plateaus, initially take up the major part the load and shear stress in the contact. Against these primary plateaus wear debris is compacted, forming secondary plateaus. The lower regions surrounding the plateaus wears by three body abrasion at the same wear rate as the plateaus.

A secondary plateau (Fig. 17) was studied more closely by using a SEM equipped with a field emission gun (FEG) electron source. It was shown that the

outermost surface of the secondary plateau has been sintered, forming a hard amorphous or nanocrystalline layer [50]. Further down in the plateau the grains get coarser and less compacted. A diffuse structure was spotted at or just below the surface, with particle sizes down to 5-10 nm, cp. Fig. 18. When looking at the detached secondary plateau from below, considerably larger grains were found, with a typical size range of 0.1 to 1 μm . It was concluded that the secondary plateau of the standard brake pad of a Volvo passenger car mainly consisted of iron oxide with the iron to oxygen ratio around 4/5. Small amounts of sulphur and copper were found scattered over the surface except on the primary plateaus. Surprisingly, despite the abundant supply of carbon-containing materials, almost no carbon was detected on the contact plateaus.

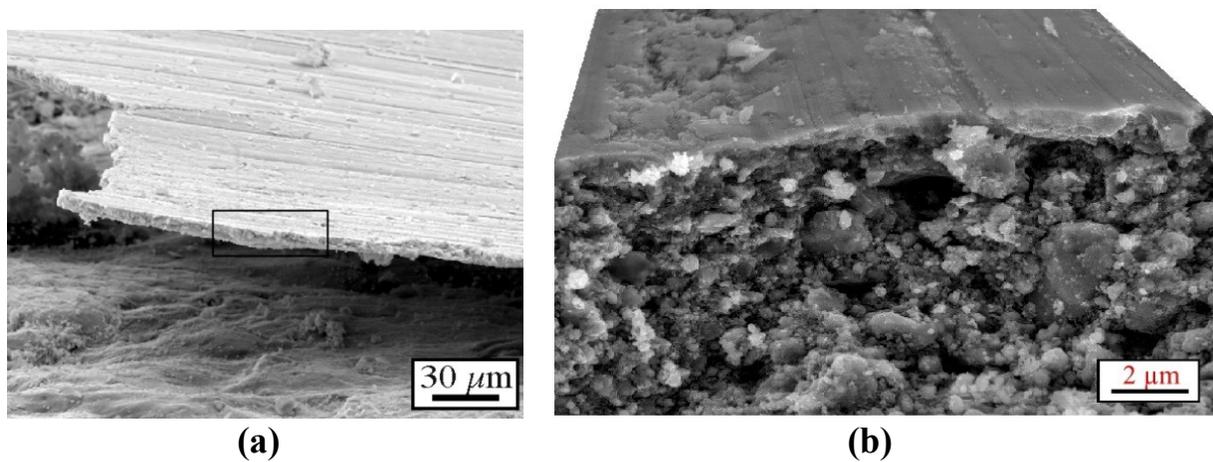


Figure 17. Partly detached secondary plateau. (a) Overview showing its thin flake like shape. (b) Detail showing the structure of relatively porous partly tribosintered debris topped by a very dense mostly amorphous top layer.

Obviously, it takes several consecutive steps to form the complex and dynamically responding tribofilm on the pads. A superficial very thin film covers both the primary and secondary plateaus. It is the flow properties of this film that account for the friction resistance.

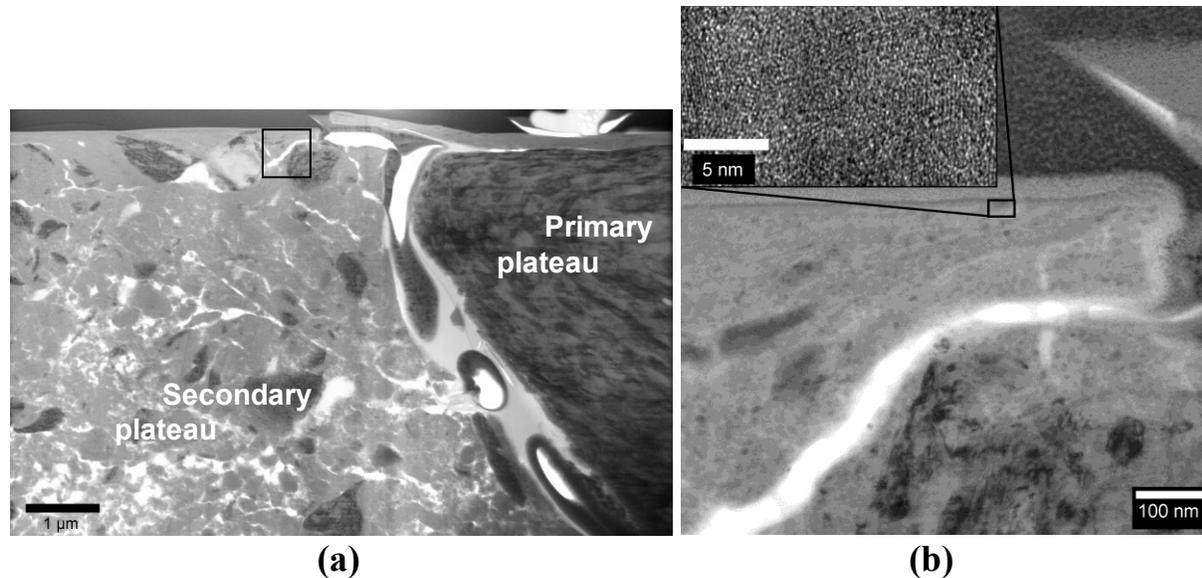


Figure 18. Cross-section of the transition area from secondary plateau (compacted debris on the left) to a primary plateau based on a steel fiber (right). The disc has been sliding against the upper surface from left to right. (TEM in bright field mode.) (a) Both primary and secondary parts of the contact plateau are covered by a thin tribofilm (medium grey, dense but broken above the crack separating the two parts). A gradient in degree of compaction in the secondary plateau is very clear, with the most loosely compacted debris at the bottom. The area shown in (b) is indicated by a black frame. (b) Detail indicating a flow in the sintered iron oxide following the shape of an iron particle. Inset: high-resolution image showing the dense amorphous character of the most superficial sintered material.

The superficial film is distinguished by being:

- nanocrystalline, with a typical grain size of around 5 nm;
- fully dense;
- very thin (we have typically noted 50–200 nm, Österle reported 100 nm [10]);
- very hard compared to the underlying secondary plateau and the medium hardness of the pad, and harder than the disc [50], and
- forming on both the primary and secondary plateaus, and
- having a composition dominated by iron oxides (Österle has in some cases noted magnetite to be the prominent phase of iron oxide [10].)

Conclusion

The tribofilms formed on automotive brake pads are of a complex deposition type. The highly dynamic formation process involves compaction of wear debris

and tribosintering forming a very thin (50–200 nm) glassy or nanocrystalline top layer. It is this vanishingly thin top layer of the tribofilm that accommodates the relative movement and provides the required stable coefficient of friction $\mu \approx 0.4$ –0.5. Despite the complex composition of a typical organic pad, including some 30 ingredients, this active top layer is dominated by nanocrystalline iron oxide.

4.7. The dramatic surface modification of rock drill buttons

One of the most demanding tribological applications is percussive rock drilling. The best materials to resist this situation of high impact against rock materials are coarse-grained WC/Co cemented carbides (CC). For a long time, these materials were primarily optimized to resist impact and abrasive wear. However, in his Ph.D. thesis, Ulrik Beste showed that the surface of the CC drill bits typically transforms into a totally different material before being removed [11, 35, 54, 56].

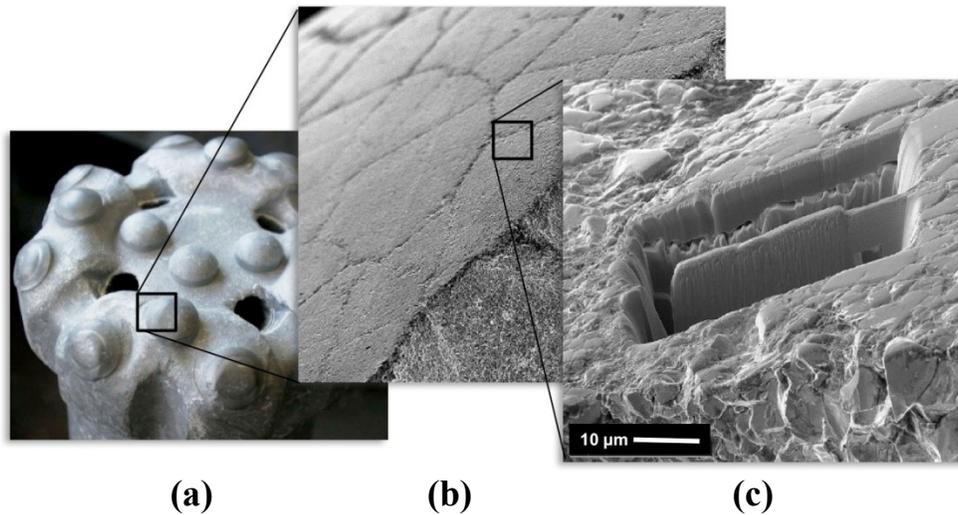


Figure 19. Appearance of a used percussion rock drill bit and the location of the studied cross-section. (a) The drill bit with its arrangement of attached CC drill buttons. (b) Close-up of one drill button, here with fracture cross-section to reveal modified surface layer (dark). (c) Detail showing the location of the TEM-sample prepared by the use of a FIB instrument.

Small buttons of CC are the active elements in crushing and fragmenting rock in percussive rock drills – see Fig. 19. Such buttons were sectioned using the FIB and studied by SEM and TEM. In conflict with the prevailing view that the temperature is too low to melt any rock material, a very thin layer of re-solidified rock material was found on top of the CC surface – see Fig. 20(a). TEM studies including energy dispersive X-ray spectroscopy (EDS) suggest that even below the outer surface, the

Co-matrix of the CC is gradually intermixed with rock material before it is removed – see figures 20 and 21. The rock mineral and the CC have formed a completely new composite material; a WC hard-phase kept together by an amorphous, Co-enriched binder phase of quartz. This tribofilm constitutes a unique form of cemented carbide, which to a large extent controls the wear rate of the button. Figure 21 shows the intimate, atomic level combination of the new matrix and the WC phase.

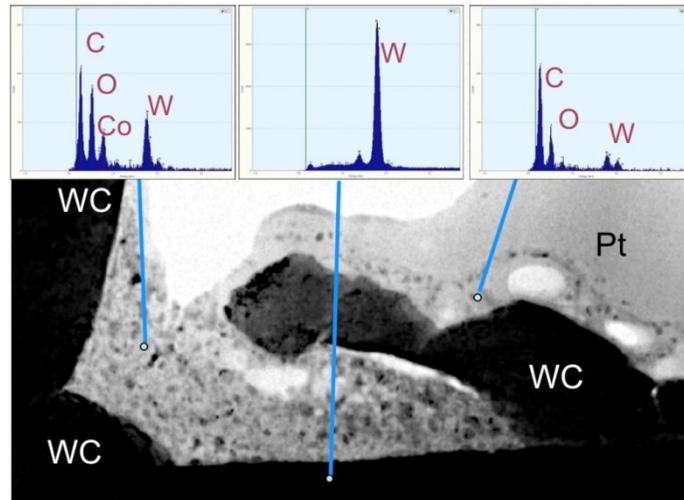


Figure 20. TEM micrograph and EDS spectra of indicated spots over a cross-section through the surface of a CC drill button used to drill in a quartz-containing rock mineral. The CC grain structure and a superficial layer of re-solidified rock material, including solidification pores can be distinguished. (Pt is used to protect the external surface during specimen preparation.)

Conclusion

The surface layer of CC drill buttons transforms from the original WC/Co structure into a new composite. The intermixed layer is probably the modification which is most significant for the performance of the drill button. Here, the carbide phase is relatively intact while the binder consists of a mixture of cobalt and rock material. It should be expected that this composite exhibits properties significantly different from the original CC, due to the following alterations:

- a less ductile and possibly harder binder;
- probably an increased level of compressive stresses in the surface layer;
- probably wedge effects operating to widen cracks.

The rock penetration phenomena motivate special considerations in the development of new cemented carbide grades for rock drilling.

- Wear resistant CC materials should be designed to either obstruct intermixed layer formation or to generate such layers with high wear resistance.
- CC materials with improved fracture resistance should be developed to avoid rock material penetration.

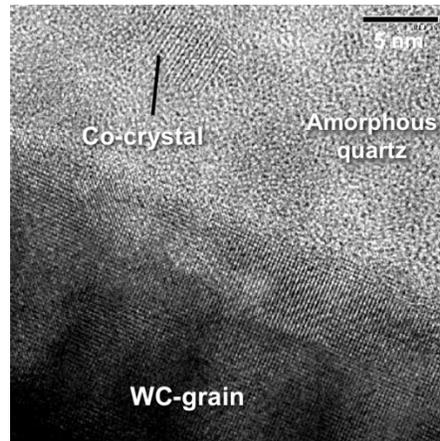


Figure 21. High-resolution TEM of the interface between a WC grain and an quartz rich tribofilm. An area of re-crystallized Co within the amorphous quartz is indicated. The WC-grain shows no indication of having been molten.

4.8. Low-friction WS₂ film formed by selective transfer and chemical reactions

An example of positive tribofilm formation is obtained with a type of DLC coatings in boundary lubricated sliding against uncoated steel. Here, friction coefficients down to 0.05 have been recorded. In this specific example a film containing large amounts of WS₂ is continually formed on the uncoated steel surface, following a somewhat more complex route of formation. The DLC coating is of the WC/C type and the oil is a polyalphaolefin (PAO) additivated with compounds containing sulphur. A smooth nanocrystalline tribofilm is generated on the steel surface by chemical reactions between W extracted from the coating and S extracted from the additive compounds (see Fig. 22). Crystals of WS₂ were identified by X-ray photoelectron spectroscopy (XPS) and high-resolution TEM – see Fig. 23 [27].

Recently Stavlid et al. showed theoretically that alloying DLC coatings with Mo could also give low friction, due to formation of MoS₂ [24]. These films reduce the already low boundary lubricated friction by 40% or more. By including theory and extended experimental and surface analytical work, we are currently working to transfer these exciting lab results into industrially useful solutions.

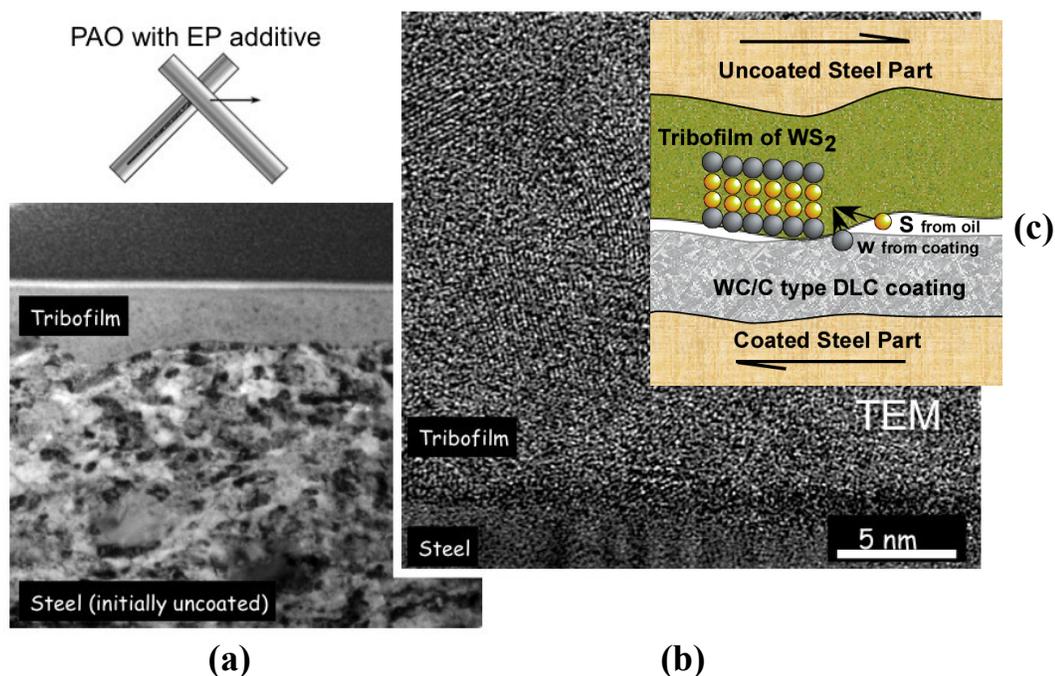


Figure 22. Cross-section of the tribofilm formed on a ball-bearing steel sample after rubbing against a W-doped DLC and (a) overview over the uncoated steel part after formation of the tribofilm, (b) HR-TEM of the tribofilm material, (c) simplified principle of the mechanism of formation [24].

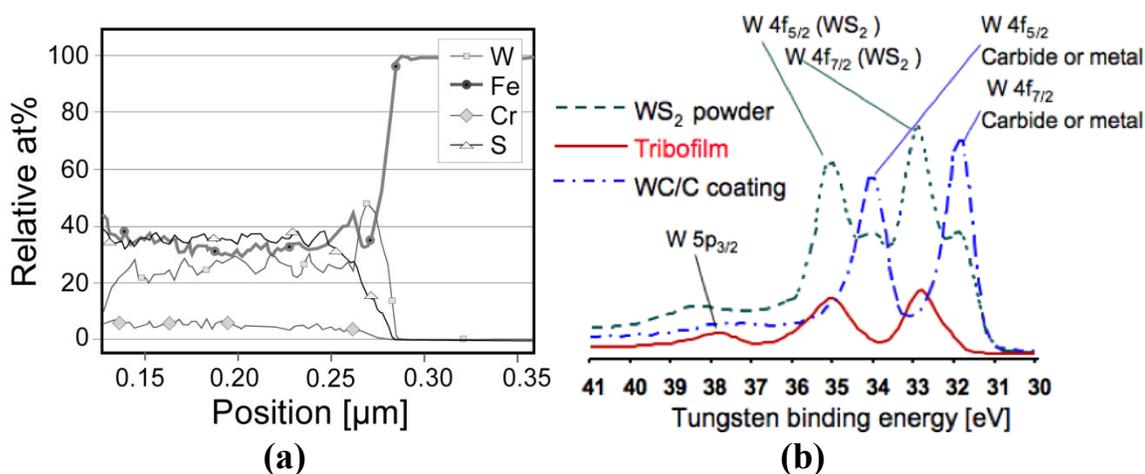


Figure 23. Chemical and depth profile analysis of the tribofilm in Fig. 22. (a) EDS elemental depth profile across the interface to the steel reveals high amounts of W, S and Fe in the film. (b) XPS spectrum comparing the composition of the tribofilm with the DLC coating (WC/C) and a reference WS_2 powder [24, 27].

Conclusion

A deposition type tribofilm forms on steel components in lubricated sliding against DLC coatings containing W or Mo. This film is formed by chemical reactions between wear products from both steel, DLC and sulphur containing additives in the lubricant. The film smoothens the contacting surfaces and reduces the friction considerably in boundary lubricated contacts.

5. Concluding remarks and key messages

As elucidated by the selected examples, tribofilms (using the broad definition of tribologically induced surface modifications) have decisive effects on the tribological performance of very different mechanical components and tools. It is *not* the original material that provides the wear resistance and friction level of the face seals, the brake pads, the cutting tools, the rock drill and so on. It is the tribofilms that constitute the real active surface layers of the different systems, and it is those films that will provide the properties to the components.

The list of examples could be made much longer, e.g. to include some of the very important and well investigated tribofilms formed by EP additives and friction modifiers in heavily loaded lubricated contacts, or by including the many forms of surface transformation of metals in sliding contact. Yet, the key messages would remain the same.

To meet the rapidly increasing tribological requirements of modern technology, we must extend our ability to optimize the use of tribofilms in the development of materials, coatings, surface finishing techniques and lubricants. This requires that we:

- Include analysis of tribofilms and surface modifications in all investigations of tribological performance and all evaluation of tribological tests. Their importance have often been highly underestimated. Researchers, mechanical designers and material developers have repeatedly been led to draw the wrong conclusions since their models of the tribological materials have been too simplistic.
- Appreciate the fact that the outermost, active part of the surface will invariably become modified, and hence systematically develop tribological materials and coatings that are tuned to achieve optimum properties after this modification.

Acknowledgements

Our present and former Ph.D. students and colleagues Fredrik Svahn, Lars Hammerström, Nils Stavlid, Ernesto Coronel, Daniel Persson (now Maglione), Magnus Hanson, Ulrik Bestes, Adam Blomberg, Mikael Olsson, Johanna Olofsson, Jun Lu, Fredrik Lindberg and Stefano Rubino are gratefully acknowledged for offering their micrographs and results for the illustrations in this chapter.

References

1. Hogmark, S., Jacobson, S., and Vingsbo, O. 1992. Surface damage. In *Friction, Lubrication and Wear Technology*, ASM Handbook 18, (Ed.: P. Blau), ASM International, pp. 176-183.
2. Blomberg, A. 1993. *Friction and wear of ceramics*. Uppsala University, Sweden.
3. Rigney, D.A., Fu, X.Y., Hammerberg, J.E., Holian, B.L., and Falk, M.L. 2003. *Scripta Mater.*, **49**, 977.
4. Rigney, D.A. 2000. *Wear*, **245**, 1.
5. Totten, G.E., and Fox-Rabinovich, G.S. (Ed.). 2007. *Self-Organisation During Friction*. CRC Press, USA.
6. Jacobson, S., and Hogmark, S. 2009. *Wear*, **266**, 370.
7. Coronel, E. *Solving Problems in Surface Engineering and Tribology by Means of Analytical Electron Microscopy*. 2005. Thesis No. 12, Uppsala University. ISSN: 1651-6214.
8. Hanson, M., Stavlid, N., Coronel, E., and Hogmark, S. 2007. *Wear*, **264**, 781.
9. Persson, D., Jacobson, S., and Hogmark, S. 2008. A physical model for the superior tribological performance of Stellites in highly loaded sliding contacts, *Wear*; in press.
10. Österle, W., Urban, I. 2006. *Tribology International*, **39**, 401.
11. Beste, U., Coronel, E., Jacobson, S. 2006. *Int. J. Refractory Met. Hard Mater.*, **24**, 168.
12. Berthier, Y. 1996, Maurice Godet's third body. In: Dowson, D., et al. (eds.), *The third body concept: interpretation of tribological phenomena*, Proc. 22nd Leeds-Lyon symposium on tribology, Elsevier, Amsterdam, The Netherlands.
13. Wood, R.J.K. 2007. *J. Phys. D: Appl. Phys.*, **40**, 5502.
14. Fischer, A., and Mischler, S. (Ed.) 2006, Tribocorrosion: fundamentals, materials and applications, *J. Phys. D: Appl. Phys.*, 39.
15. Olofsson, J., Jacobson, S., and Johansson, S. 2007. Analysis of the tribofilm formation on the friction drive surfaces of a piezoelectric motor. In *STLE/ASME Int. Joint Tribology Conf.*, San Diego, California, USA.
16. Olofsson, J., Lindberg, F., Johansson, S., and Jacobson, S. 2009. On the role of tribofilm formation on the alumina drive components of an ultrasonic motor. *Wear*, **267**, 1295.
17. Erickson, L.C., Blomberg, A., Hogmark S., and Bratthall, J. 1993. *Tribology International*, **26**, 83.
18. Ajayi, O.O., and Ludema, K.C. 1990. *Wear*, **140**, 191.
19. Adachi, K., and Kato, K. 2000. *Wear*, **245**, 84.
20. Andersson, P., and Blomberg, A. 1993. *Wear*, **170**, 191.
21. Blomberg, A., Hogmark, S., and Lu, J. 1993. *Tribology International*, **26**, 369.
22. Rainforth, W.M., 2004. *J. Mat. Science*, **39**, 6705.
23. Singer, I.S., Dvorak, S.D., Wahl, K.J., and Scharfb, T.W. 2003. *J. Vac. Sci. Technol. A*, **21**, 5.
24. Stavlid, N. 2006. On the formation of low-friction tribofilms in Me-DLC – Steel sliding contacts, Doctoral thesis, Uppsala University, Sweden.

25. André, B., and Jacobson, S. 2008. Comparisons between commercial low-friction coatings and emerging coating concepts in ball-on-disc tests – coefficient of friction, tribofilm formation and surface damage. In *NORDTRIB 2008*, Tampere, Finland.
26. Wilhelmsson, O., Rasander, M., Carlsson, M., Lewin, E., Sanyal, B., Wiklund, U., Eriksson, O., and Jansson, U. 2007. *Functional Materials*, **19**, 1611.
27. Podgornik, B., Hren, D., Vizintin, J., Jacobson, S., Stavlid, N., Hogmark, S. 2006. *Wear*, **261**, 32.
28. Lindqvist, M., and Wiklund, U. 2009. Tribofilm formation from TiC and nanocomposite TiAlC coatings, studied with Focused Ion Beam and Transmission Electron Microscopy. *Wear*, **266**, 988.
29. Lindquist, M. 2008, Self Lubrication on the Atomic Scale: Design, Synthesis and Evaluation of Coatings. Thesis No. 391, Uppsala University, Sweden. ISSN: 1651-6214.
30. Polcar, T., Evaristo, M., and Cavaleiro, A. 2007. *Vacuum*, **81**, 1439.
31. Bowden, F., and Hanwell, A.E. 1966. *Proc. Royal Soc. A*, **295**, 233.
32. Bowden, F., and Tabor, D. 1964. *The Friction and Lubrication of Solids. Part II*. Oxford University Press, Oxford, England.
33. Gardos, M.N. 1999. *Surf. Coat. Technol.*, **113**, 183.
34. Andersson, J., Duda, L., Schmitt, T., and Jacobson, S. 2008. *Tribology Letters*, **32**, 31.
35. Andersson, J., Erck, R.A., Erdemir, A. 2003. *Surf. Coat. Technol.*, **163–164**, 535.
36. Andersson, J. Microengineered CVD diamond surfaces – tribology and applications, 2004. Ph.D. thesis (Comprehensive summaries of Uppsala Dissertations from the Faculty of Science and Technology 977), Uppsala University, Sweden.
37. Martin, J.M., Grossiord G., Le Mognea, T., and Igarashi, J. 2000. *Wear*, **245**, 107.
38. Spikes, H. 2004. *Tribology Letters*, **17**, 469.
39. Hsu, S.M., and Gates, R.S. 2001. Boundary lubrication and boundary lubricating films. In *Modern Tribology Handbook* (Ed.: B. Bhushan), CRC Press, USA, pp. 455-492.
40. Callister Jr., W.D. (Ed.) 2003. *Materials Science and Engineering – An Introduction*. Wiley, USA.
41. Hogmark, S., Söderberg, S., and Vingsbo, O. 1979. Surface deformation and frictional heating of steel during machining and laboratory testing. *Proc. 3rd Int. Conf. on Mechanical Behaviour of Materials*, Cambridge, England, pp. 621-631.
42. Eklund, L.-H., and Hogmark, S. 1982. *Scand. J. Metall.*, **11**, 226.
43. Vingsbo, O., and Hogmark, S. 1981. Wear of steels. In *Fundamentals of Friction and Wear of Materials* (Ed.: D.A. Rigney), ASM: Ohio, USA, pp. 373–408.
44. Hogmark, S., Jacobson, S., and Coronel, E. 2007. *Tribologia* (Finnish Journal of Tribology), **26**, 3.
45. Persson, D. 2005. On the mechanisms behind the tribological performance of Stellites. Thesis No. 129, Uppsala University, Sweden. ISSN: 1651-6214.
46. Podgornik, B., Sandberg, O., and Hogmark, S. 2004. *Surface and Coatings Technology*, **184**, 338.
47. Schedin, E. Micro-mechanisms of sheet-tool contact during sheet metal forming, 1992. Ph.D. thesis, Royal Institute of Technology, Stockholm, Sweden.
48. Hogmark, S., Jacobson, S., and Wänstrand, O. 1999. A new universal test for tribological evaluation. In *Proc. 21st IRG-OECD Meeting* in Amsterdam, The Netherlands.
49. Hogmark, S., Olsson, M., and Blomberg, A. 1992. *Journal of Hard Materials*, **3**, 153.
50. Eriksson, M., and Jacobson, S. 2000. *Tribology International*, **33**, 817.
51. Eriksson, M. 2000. Friction and contact phenomena of disc brakes related to squeal. Comprehensive summaries of Uppsala Dissertations, Faculty of Science and Technology, Sweden, 537.
52. Eriksson, M., Lord, J., and Jacobson, S. 2001. *Wear*, **249**, 272.
53. Eriksson, M., Bergman, F., and Jacobson, S. 2002. *Wear*, **252**, 26.
54. Beste, U. 2004. On the nature of cemented carbide wear in rock drilling. Comprehensive summaries of Uppsala Dissertations, Faculty of Science and Technology, Sweden, 964.
55. Beste, U., and Jacobson, S. 2007. *Wear*, **264**, 1129.
56. Beste, U., Hogmark, S., and Jacobson, S. 2007. *Wear*, **264**, 1142.



Research Signpost
37/661 (2), Fort P.O.
Trivandrum-695 023
Kerala, India

Recent Developments in Wear Prevention, Friction and Lubrication, 2010: 227-262
ISBN: 978-81-308-0377-7 Editor: George K. Nikas

6. Transient phenomena in elastohydrodynamic lubrication

Romeo P. Glovnea

School of Engineering and Design, University of Sussex, England

Abstract. Tribological contacts of machine components working in the elastohydrodynamic regime always experience transient conditions due to variation of speed, load, geometry, or vibrations generated within the contacts or transmitted from the environment. Given the fact that the contacting bodies and lubricant, together with the other components of the mechanisms or machine, form a dynamic system comprising of springs and dampers, this will respond to dissipate the instability caused by the variation of the above parameters. At the lubricating film level, this is done by constant-amplitude, or dampened oscillations of its thickness.

This chapter is a review of experimental research carried out in the past decade into the non-steady state elastohydrodynamic lubrication, including transient loading, sudden variation of entrainment speed and variation of micro-geometry.

1. Introduction

The behaviour of elastohydrodynamically-lubricated contacts in steady-state conditions is well understood both from a theoretical and an experimental point of view. However, in real applications, *elastohydrodynamic (EHD)* contacts rarely experience perfect, steady-state conditions, and are instead subjected to transitory variations of load, geometry or velocity of the contacting surfaces. Often more than one of these parameters varies at the same time, making the prediction of the film thickness a very difficult task, even if those variations are known, which is not usually the case. For example in gear and cam mechanisms, all these parameters vary during a working cycle, while in rolling element bearings the load changes when the rolling element enters the loading zone. The bearings from stepper motors are subjected to sudden variations of speed in a repetitive, start/stop succession. Rolling elements bearings are inherently subjected to vibrations caused by dynamic unbalances of the shaft they support, surface irregularities or simply transmitted from the surrounding. As well as these examples, all EHD lubricated

contacts experience transient conditions during start and stop of the machinery of which they are part.

Steady state elastohydrodynamic films can be easily achieved in laboratory conditions, where working parameters (load, velocity, geometry, temperature) are kept constant, and the contact can be isolated from surrounding vibrations. Steady state conditions give an estimate of the effect of lubricant composition and properties upon film thickness and friction force in the contact, which are mainly used in two ways: one is to validate the theoretical models of the behaviour of EHD films, and the second is to extend or generalise results obtained in idealised conditions to practical situations. Controlled variation of one of the working parameters mentioned above allows the evaluation of its effect upon the EHD film formation under imposed transient conditions.

In recent years transient phenomena in elastohydrodynamic lubrication have received particular attention from both a theoretical and experimental point of view. A literature review of the experimental work regarding transient EHD carried out until the mid-1990s can be found in [1].

In this chapter, experimental contributions to the understanding of transient phenomena will be reviewed, with the focus on those published over the past decade. It must be mentioned that significant work and progress has also been made on the numerical simulation of the behaviour of elastohydrodynamic films in transient conditions. In fact theoretical and numerical approaches have often preceded experimental findings. It is not the intention of this review to cover the vast literature concerning numerical/transient EHD lubrication however some references will be made when the numerical work was performed to validate certain experimental results.

Transient events in elastohydrodynamic lubrication can be grouped into three main categories:

- variation of load, including impact loading;
- variation of contacting surfaces geometry, including surface roughness;
- variation of surface velocity, including squeeze, entrainment, and a combination of the two.

2. Background

2.1. Mechanisms of elastohydrodynamic lubrication

Many machine elements and mechanisms transmit relative motion and forces between surfaces which make theoretical contact in a point or along a line. In

practice the contact takes place on a very small surface, which makes the contact pressure considerably large, even at relatively low loads. Assuming elastic bodies in contact, which is often the case with steel components, under typical loads, the deflection of these surfaces is predicted by the Hertz theory of elastic contact [2]. The elastic deflection of the contacting surfaces is one of the three important phenomena which are involved in the elastohydrodynamic lubrication. Another is the hydrodynamic effect, which is responsible for the “lift” or load carrying capacity of the bearing and it is described by the Reynolds equation [3].

$$\frac{dp}{dx} = 6U\eta \left(\frac{h - h_c}{h^3} \right) \quad (1)$$

Where p is the pressure, U is the average velocity of the surfaces in direction x of the flow, h is the current film thickness and h_c is the film at the position where pressure is maximum. The third mechanism required is the variation of the fluid’s viscosity with pressure. Putting these three phenomena together in an analysis of an exquisite simplicity and elegance, Grubin and Vinogradova were able to reveal the mechanisms of elastohydrodynamic lubrication [4] of linear contacts (this is also attributed to Ertel [5]). They considered that the film is completely flat and that the shape of the deformed surfaces in the inlet of the contact is given by the Hertz theory. It follows that the separation between the surfaces in the convergent inlet is the sum of the constant lubricant film thickness and the Hertzian elastic deformation (h_s), as seen in Fig. 1.

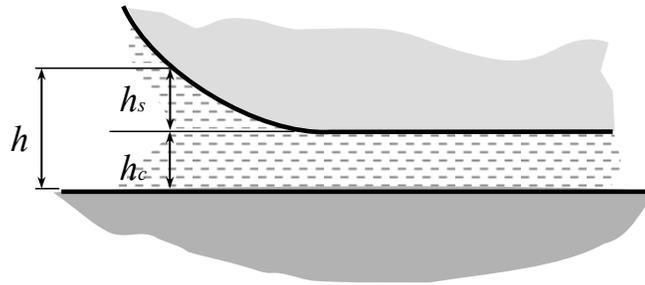


Figure 1. The geometry of the EHD contact inlet.

Assuming an exponential law for the dependence of the viscosity on the pressure and that the pressure remains Hertzian inside the contact, the Reynolds equation can be integrated to give, after some manipulations, the thickness of the EHD film.

$$h_c = 1.93(\alpha\eta_0 U)^{3/4} R^{3/8} W^{-1/8} E'^{1/8} \quad (2)$$

In this relation, R is the reduced radius of curvature in the direction of flow ($R = [(1/R_1) + (1/R_2)]^{-1}$), W is the load per unit length of contact and E' is the reduced elastic modulus of the materials ($E' = [(1 - n_1^2)/E_1 + (1 - n_2^2)/E_2]^{-1}/2$).

Although approximate, Grubin's solution gives an insight into the physics of the elastohydrodynamic lubrication and has been fully supported by exact numerical solutions [6]. The latter have also revealed that the film is not completely flat inside the contact, but shows constrictions at the exit and toward the sides of the contact. Experimental evidence supported these findings [7], as illustrated in Fig. 2 for a point contact.

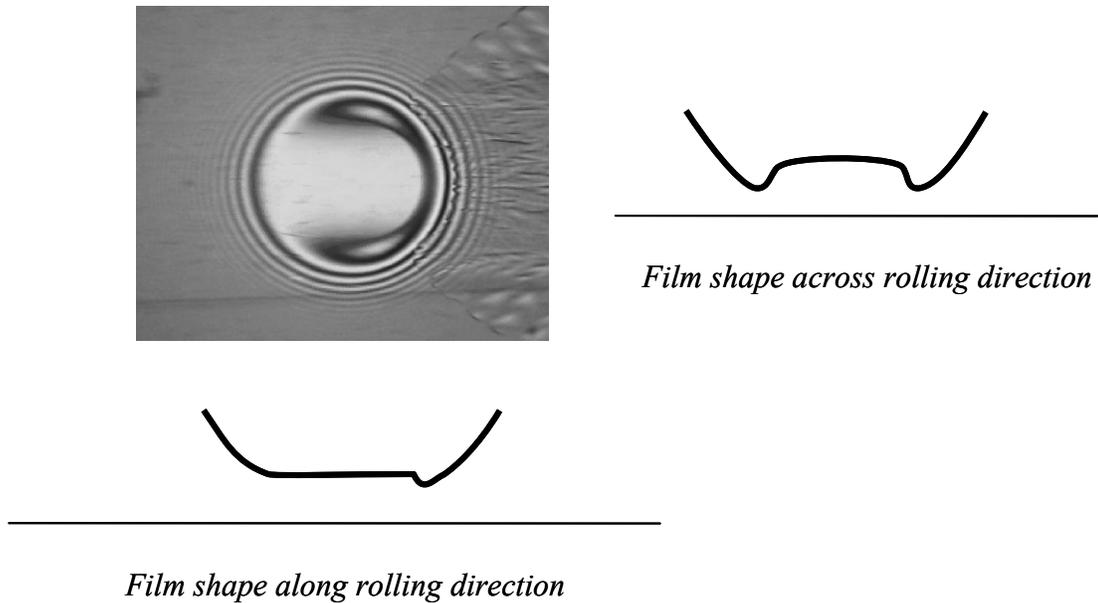


Figure 2. Film shape in a point contact obtained by optical interferometry.

2.2. EHD film thickness measurement methods

2.2.1. Electrical methods

Historically, electrical methods (resistive and capacitive), have preceded optical methods for studying film formation in elastohydrodynamic contacts. Usually the electrical circuit measures both resistance and capacitance, the former being used to detect full film conditions (complete separation between the contacting bodies) and the latter to measure the thickness of film, e.g. Crook [8], Archard and Kirk [9] and MacCarthy *et al.* [10]. The electrical resistance method was also used

relatively extensively for the study of mixed lubrication regime (Furey [11], Tallian *et al.* [12] Guanteng *et al.* [13]). The main advantage of electric methods is that the elastohydrodynamic contact can be formed between two bodies made out of steel, i.e. similar to the contacts found in most machine components that operate in the elastohydrodynamic regime. There are, however, some considerable disadvantages to the electrical methods. The capacitance of the contact depends of the shape of the bodies, which can only be presumed inside the contact, where large local deformations occur. The electrical methods also require knowledge of the permittivity and resistivity of the lubricant, whose variation with pressure is not precisely known. Additionally these methods only give average values over the contact area and offer no indication of the local shape of the film.

2.2.2. Interference of light

Many phenomena specific to light, reflection, refraction, interference, etc., can be satisfactorily explained by the classical wave theory. Some of these phenomena make light a versatile tool for experimental research in a wide range of fields. In this paragraph some terms and concepts characteristic to physical optics will be briefly defined. A light wave can be represented by a sine or cosine function, as in equation (3).

$$y = A \sin \frac{2\pi}{\lambda} (x - vt) \quad (3)$$

This is the equation of a transverse wave, which moves at velocity ‘v’ in direction +x, [14]. ‘A’ is the amplitude and ‘λ’ the wavelength of the light, as seen in Fig. 3.

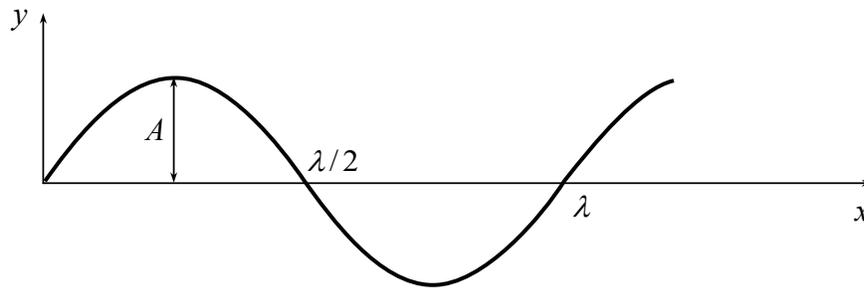


Figure 3. A transversal sine wave expressed by equation (3) at $t = 0$.

It is convenient to express the equation of simple harmonic waves in terms of the angular frequency ω .

$$y = A \sin\left(\omega t - \frac{2\pi}{\lambda} x\right) \quad (4)$$

The quantity in the parentheses in equation (3) is known as *phase* and expresses the position of the wave at a certain time t . In practical terms what is important is the phase difference (δ) between two beams of light when they reach a certain point.

$$\delta = \frac{2\pi}{\lambda} \Delta \quad (5)$$

where Δ is the path difference. When travelling through different media the velocity of the light waves is altered according to the refractive indexes of the media. The *optical path* is therefore defined as the product between the geometrical path and the refractive index of the medium. It follows that the phase difference can be written as:

$$\text{Phase difference } \delta = \frac{2\pi}{\lambda} \times \text{optical path difference} = \frac{2\pi}{\lambda} \Delta \quad (6)$$

Whenever a beam of light travelling through a medium with a certain refractive index arrives at the boundary with another medium with different refractive index, it is partly *reflected* and partly transmitted (*refracted*). It is important to mention that the reflection of light at any boundary is accompanied by a phase change.

When two beams of light of equal wavelength, arrive at a certain point in space a phenomenon of superposition (combination) takes place, with the net result being a change of the amplitude and intensity of the resulted light. The modification of the intensity obtained by the superposition of two (or more) light beams is called *interference* [14]. In practice, interference can be obtained, from a single beam, by *wave front* or *amplitude division* [15]. The latter takes places when a beam of light arrives at the interface between two media, being both reflected and transmitted. As this is the phenomenon exploited in the measurement of thin elastohydrodynamic films, it will be briefly explained in the following paragraphs.

Consider a beam of light which hits a thin plate of thickness h , at a certain angle of incidence, θ . The refractive index of the material, of which the plate is made, is

n . At the interface between air and the plate the beam will be partly reflected, path ADF , and partly transmitted, path AB , as seen in Fig. 4. At C , the beam is transmitted, path CF' , and reflected. This internal reflection can take place several times; a phenomenon known as multiple reflection. It should be noted that the amplitudes of the reflected rays become negligible after the second internal reflection. Moreover, for the purpose of the calculation of the path difference only the first two rays are needed.

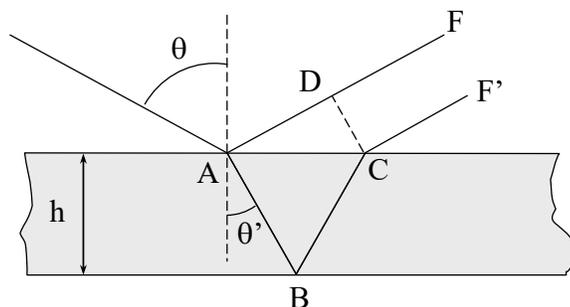


Figure 4. Interference by amplitude division. Calculation of path difference.

As seen by division of amplitude, the initial ray results in two rays: AF and CF' . There is a phase difference between these two rays resulting from the path difference between them and from the advance in phase produced by the reflection at A . The latter is equal to $\lambda/2$, for air-glass reflection. The former is the difference between the path ABC , travelled by the transmitted ray, in the medium with index of refraction n and the path AD travelled, in air, by the reflected ray. D is the foot of the perpendicular from C to the reflected beam.

$$\Delta = 2nAB - AD + \frac{\lambda}{2} \quad (7)$$

Distances AB and AD are found from simple geometrical considerations.

$$AB = \frac{h}{\cos \theta'} \quad AD = AC \sin \theta = 2h \sin \theta \tan \theta'$$

The path difference becomes:

$$\Delta = 2nh \cos \theta' + \frac{\lambda}{2} \quad (8)$$

The intensity of the light at a point on the surface of the plate is given, up to a constant, by:

$$I \approx \cos^2 \frac{\delta}{2} \quad (9)$$

Replacing the phase difference according to expression (6) the intensity becomes:

$$I \approx \cos^2 \frac{\pi \Delta}{\lambda} = \cos^2 \left[\pi \left(\frac{2nh}{\lambda} \cos \theta' + \frac{1}{2} \right) \right]$$

It is obvious that what an observer will see are bright and dark fringes. Assuming the refractive index of the medium constant, the path difference depends only on the thickness of the plate. The condition to obtain a bright fringe is given by:

$$\frac{2nh}{\lambda} \cos \theta' + \frac{1}{2} = N \quad (10)$$

where $N = 0, 1, 2, \dots$ is the fringe order. A dark fringe is obtained for

$$\frac{2nh}{\lambda} \cos \theta' = N \quad (11)$$

For normal incidence, $\cos \theta' = 1$, and a dark fringe corresponds to a thickness $h = N\lambda/2n$. Whenever we pass from a dark fringe to another, the thickness of the plate changes by $\lambda/2n$.

2.2.3. Dispersion of light; spectrometry

It was shown above that a beam of light is partly reflected and partly refracted at the separation boundary between two media with different refractive index. The refraction of white light is also accompanied by a separation into its component wavelengths [14], phenomenon known as *dispersion* and due to the variation of the refraction index with wavelength. Dispersion of light was discovered by Newton in 1665, when he passed a beam of light from the sun through a glass prism. The phenomenon is illustrated in Fig. 5.

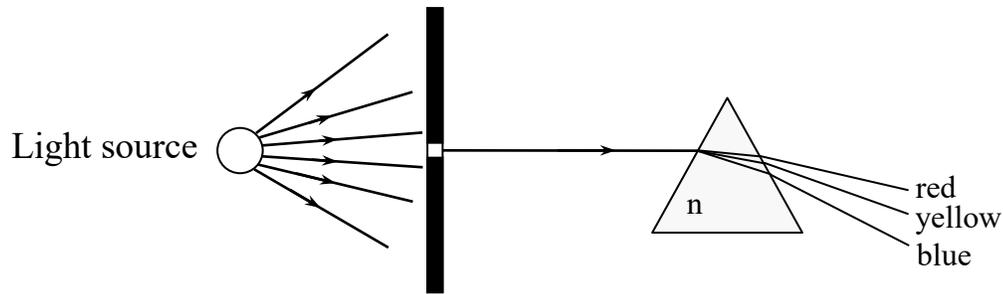


Figure 5. Dispersion of light through a prism.

The phenomenon of dispersion of light in its wavelength components is exploited in *spectrometers*, instruments which measure the relative amount of radiation of each wavelength. Dispersion of light in a spectrometer can be achieved by refraction in a prism or by interference by wave front division. This kind of interference is obtained passing the light through slits positioned at regular intervals, as is simply illustrated by the well-known Young's experiment, schematically seen in Fig. 6.

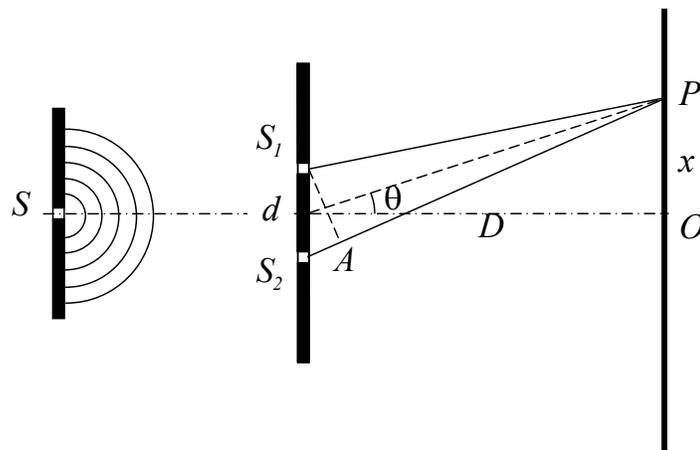


Figure 6. Interference by wave front division. Determination of the path difference.

When arriving at the pinholes (or slits) the wave front generated at S will extend beyond the regions directly exposed to the oncoming wave, phenomenon known as *diffraction*. Assuming that the waves have the same phase at S_1 and S_2 at a certain point P on a plane situated at distance D they will superimpose, the intensity of the resulted wave depending on the phase difference between them, according to equation (9). The phase difference depends on the optical path difference, which approximating the sinus with its argument, can be written as:

$$\Delta = d \sin \theta = d \frac{x}{D}$$

Substituted into equation (6) the phase difference can be obtained.

$$\delta = \frac{2\pi}{\lambda} \Delta = \frac{2\pi}{\lambda} \frac{xd}{D} \quad (12)$$

Replacing this phase difference into equation (9) it can be seen that the intensity has a maximum whenever the path difference is an integral multiple of λ . Consequently we have:

$$\frac{xd}{D} = 0, \lambda, 2\lambda, 3\lambda, \dots = N\lambda$$

With other words, the position of bright fringes is given by:

$$x = N\lambda \frac{D}{d} \quad (13)$$

The intensity is minima, i.e. zero, when $\delta = \pi, 3\pi, 5\pi, \dots$. Introducing this result in (12) the condition for a dark fringe occurrence becomes:

$$\frac{xd}{D} = \frac{\lambda}{2}, \frac{3\lambda}{2}, \frac{5\lambda}{2}, \dots = \left(N + \frac{1}{2}\right)\lambda \Rightarrow x = \left(N + \frac{1}{2}\right) \frac{D}{d} \lambda \quad (14)$$

N is the fringe order, as in the previous paragraph.

In spectrometers the dispersion of light is obtained with *diffraction gratings*, which are optical elements equivalent in action to a number of parallel, equidistant slits (14). It is beyond the purpose of this chapter to enter into details of the construction and theory of ruled diffraction gratings however, it can be mentioned that usually they consist of equally spaced parallel grooves, formed on a reflective coating and deposited on a substrate. Fig. 7 shows the spectra obtained by a diffraction grating. For purpose of clarity only the negative orders are shown.

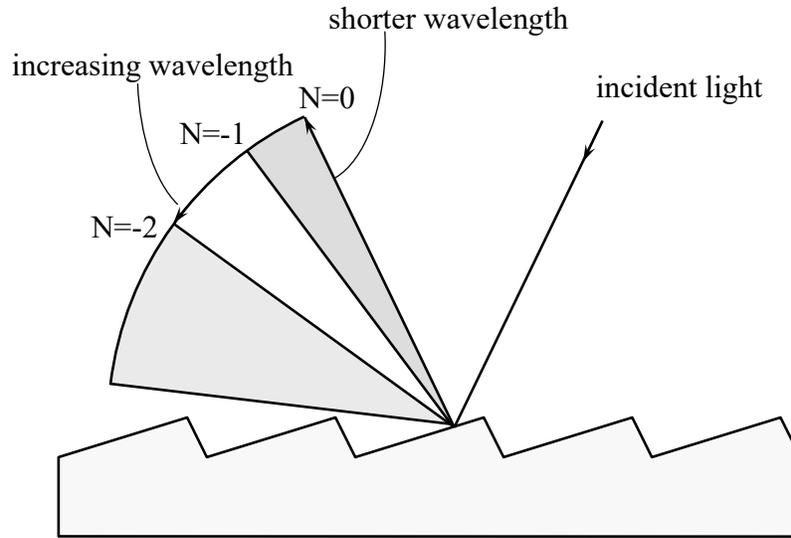


Figure 7. Dispersion of light by a diffraction grating.

2.2.4. EHD film thickness measurement by optical interferometry

For research purposes, electrical methods of measuring film thickness in EHD contacts have gradually lost ground in the favour of the optical interferometry technique, which eliminates most of the disadvantages mentioned before and allows the mapping of the whole contact with a sub-micron resolution (Cameron and Gohar [7], Foord *et al.* [16]). Developed in the early sixties, this method employs a flat, transparent disc (usually glass or sapphire), loaded against a shiny steel ball or roller. The contacting flat surface is coated with a semi-reflective metallic layer (chromium) such that any incident light shone onto the contact (usually monochromatic), is twice reflected, firstly at the glass-metal layer interface and secondly at the ball/roller surface as seen in Fig. 8.

Upon recombination the path difference between the reflected rays results in either constructive or destructive interference. The main limitation of the method in this form is that it cannot distinguish between films thinner than approximately a quarter of the wavelength of the light used.

The refinement that overcame this limitation was the addition of a solid spacer layer on top of the semi-reflective chromium layer. Made out of silica (SiO_2) the spacer layer has a refractive index close to that of mineral oils, which makes it act as a ‘solid oil’, increasing the separation between the surfaces of the contact and thus allowing measurements of films theoretically of any thickness. The benefits of the spacer layer were fully exploited in ultra-thin film optical interferometry method (UTFI), which uses white light and a spectrometer to disperse it into its

component wavelengths as shown in Fig. 9. The resolution of this technique is of the order of nanometres [17, 18].

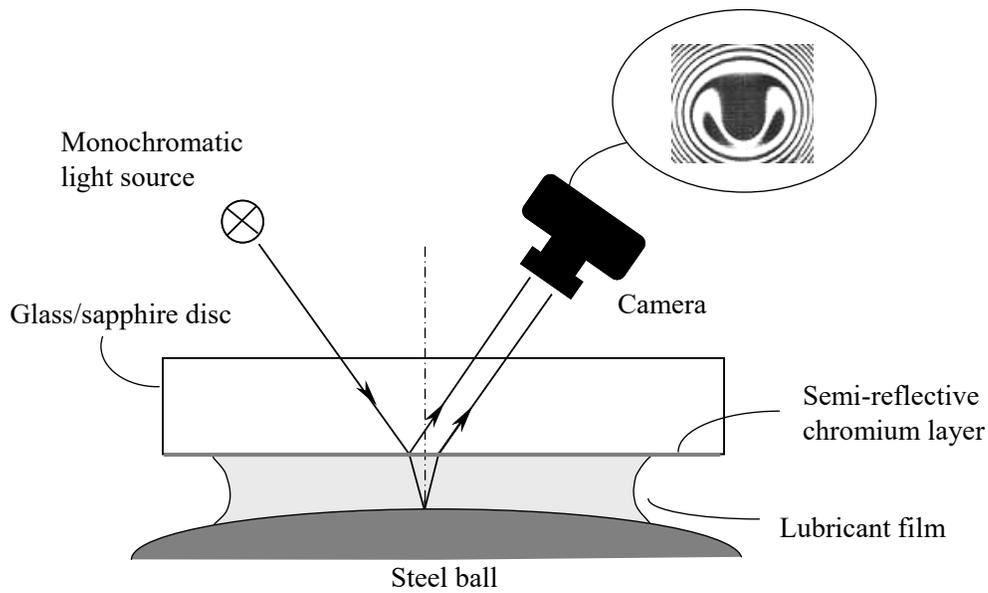


Figure 8. Principle of optical interferometry.

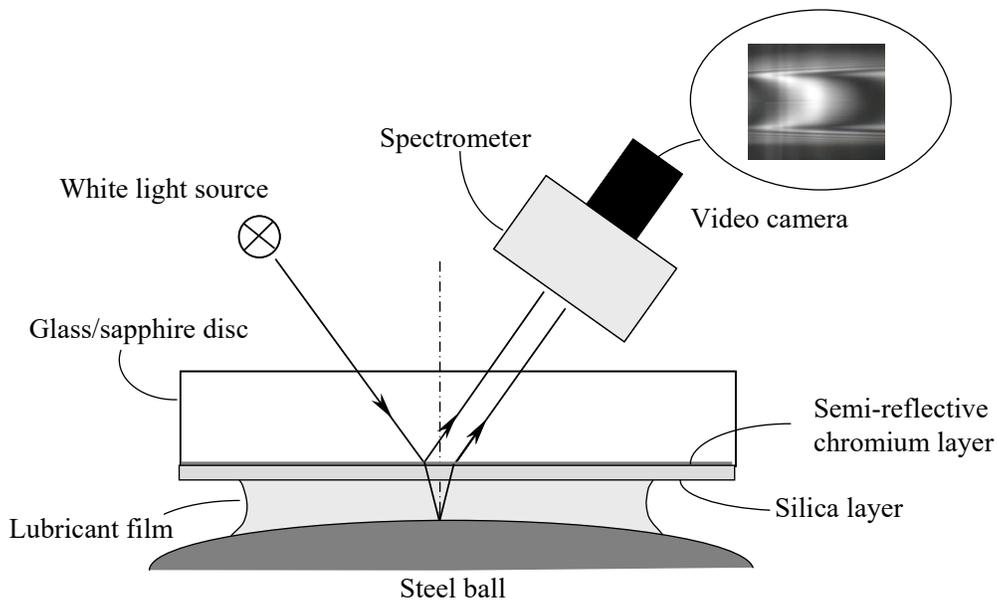


Figure 9. Ultrathin film interferometry technique.

2.3. Early experimental work on transient EHD lubrication

As electrical methods were historically first used for the study of EHD films, it is natural that they were employed in the early study of transient phenomena [19-21]. Vichard [20] carried out major theoretical and experimental studies into the effect of load variation upon EHD film thickness. He used the electrical capacitance method to measure the film thickness in the contact formed between a cam and a follower and compared with theoretical predictions. With the development of optical interferometry, and the increased attention to non-steady effects in elasto-hydrodynamic lubrication, this method started to be used for transient studies as well. Sanborn and Winer reported a study on the effect of transient loading upon the film thickness in pure sliding EHD contacts [22]. The interferometric fringes of the contact were recorded using high speed video during rapid change of the applied load. The researchers concluded that film thickness during rapid loading in a sliding/rolling experiment could be predicted from the steady-state behaviour; however, it should be noted the speed of load variation was limited by the acquisition speed of the camera used. Notable early experimental work on transient EHD lubrication using optical interferometry is also due to e.g. Hoglund and Jacobson [23], Ren *et al.* [24], Larson and Lundberg [25] and Sugimura *et al.* [26], among others. To be noted that while in earlier experimental work, employing electrical methods, mainly line contacts (characteristic to spur gears and cams) were studied. Although in optical interferometry both line and point contact can be obtained, it is much simpler to use a ball against a flat arrangement, in order to avoid alignment problems. It follows that all experimental work described in this chapter refers to point contacts. It is only reasonably to assume that the basic phenomena found in transient lubrication of point contacts apply in a similar fashion to line contacts.

3. Behaviour of EHD films in transient conditions

3.1. The effect of dynamic variation of load upon the behaviour of EHD films

It is known that load change has little influence on steady-state EHD film thickness. For this reason experimental studies on this aspect of non-steady EHD lubrication are less numerous than those on speed or micro-geometry; however they are not necessarily less important. Once formed, the elasto-hydrodynamic film is very stiff, and inside the contact the film thickness will not change instantaneously as a response to a variation of the load. The thickness of the film is

established by the conditions at the inlet, where the pressure is relatively low and the lubricant responds to external perturbations. A sudden increase of the load both increases the contact area and creates a squeeze effect in the lubricant at the periphery of the contact. When both squeeze and entrainment are present the squeeze of the lubricant will create a perturbation that subsequently travels relatively unchanged across the contact, creating pressure fluctuations that can affect the fatigue life of the contact.

Two types of non-steady phenomena will be considered in this part: transient loading and vibrations.

3.1.1. Transient loading

Studies in this direction have focused on the effect of transient loading upon film thickness and friction force. Any elastohydrodynamic contact can be seen as a dynamic system, consisting of springs and dampers connected together. It would be expected that any attempt to pull this system from a steady state will inevitably lead to dampened oscillations until the system reaches another equilibrium state. This has been proved to be true for a sudden change of applied load and also for rapid variation of speed, as will be shown later in this chapter.

Wijnant *et al.* [27] applied a load step of 45 N to 165 N, on a pure rolling contact and monitored the film thickness variation with optical interferometry and high speed imaging. They observed oscillations in the film thickness, which dampened after about 22 milliseconds. They performed a parallel numerical simulation of the experiments, but did not attempt a quantitative comparison of the experiment and numerical analysis. No film thickness variation is given for the experiments, but the numerical simulation indicates that the maximum amplitude of the film thickness is about 30 per cent of the initial, steady state central value. One conclusion which is clear from their experiments is that the combined damping of the system, including the EHD film, must be small. Similar results are reported by Kilali *et al.* [28] and Sakamoto *et al.* [29]. Both showed the film thickness along the central plane of the contact following a sudden variation of load, but did not capture the variation of central film thickness with time, from which the damping characteristics of the system could have been extracted. No film thickness oscillations were found by Kaneta *et al.* [30], in a pure rolling contact subjected to a sudden load pulse of about six times the initial load magnitude. However, as in previous cases, they found that a crescent shaped fluid entrapment is formed in the inlet and transported through the conjunction at the entrainment speed. The maximum value of the thickness in the entrapped region was found to increase with the loading rate. This suggests that the formation of oscillations in the film thickness is related not only to the characteristics of the

elastohydrodynamic contact itself, but also on the dynamic behaviour of the whole assembly.

In pure squeeze (i.e. no entrainment motion) the behaviour of the system is opposite, as shown by Chu *et al.* [31]. The impact load exhibits large oscillations until it stabilises at the final value, but the film thickness falls nearly continuously, with only negligible variation.

It can be therefore concluded that load fluctuations only translate into film thickness oscillations when both entrainment and squeeze are present, the latter to create film perturbations in the inlet and the former to convect these through the EHD conjunction.

3.1.2. Friction coefficient under transient loading

Evidence of the frictional behaviour of transient loaded contacts is given by Workel *et al.* [32, 33] and Arhstrom [34]. Both researchers used a bouncing-ball type apparatus to obtain transient loading conditions but employed different methods to measure the normal and transversal forces during impact. Workel and co-workers focussed on a traction fluid while Arhstrom employed two mineral and one polyalphaolefin, amongst other oils, in his experiments. It is difficult to directly compare these sets of results as different types of lubricants are employed, however both show a clear (albeit small), tendency of a decreasing friction coefficient with an increasing mean contact pressure. This result is consistent with those obtained in a disc machine by Evans and Johnson [35] for the same traction fluid used by Workel *et al.*, but not for a mineral oil, for which Evans and Johnson's results conversely show an increase in the traction coefficient with an increase in the mean contact pressure.

3.1.3. Vibrating EHD contacts

This case is probably more relevant to practical applications, as almost all machine components encounter vibrations in their working life and the way the lubricant film transmits or eventually dampens these vibrations influences the functioning of the whole assembly of which the machine component is part of. Vibrations caused by a change of the load or entrainment velocity have been reported in various studies, as shown throughout this chapter, but systematic work dedicated specifically to the effect of vibrations upon the behaviour of elastohydrodynamic contacts is sparse. Glovnea and Spikes [36] have shown that oscillations transverse to the main rolling motion create ripples through the lubricant film thickness which propagate along the instantaneous entrainment

direction, as seen in Fig. 10. The film thickness profile along the instantaneous entrainment direction is also shown.

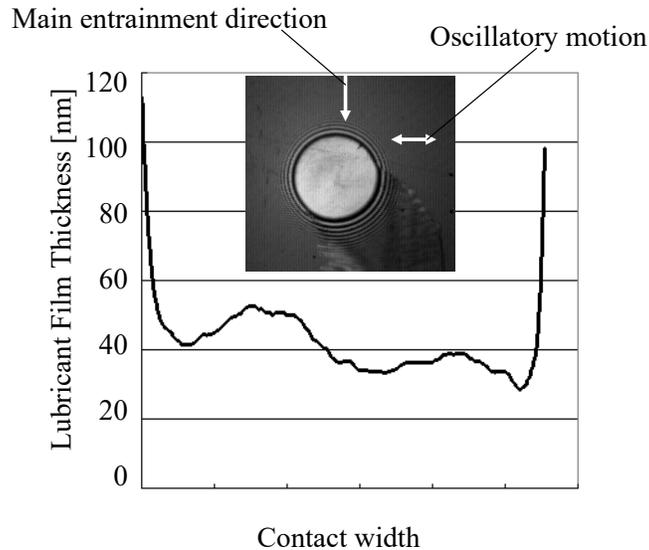


Figure 10. Ripples through film thickness due to lateral oscillations.

Ciulli and Bassani [37] did not induce vibrations, but instead recorded the effect of the vibrations caused by the imprecision of their instrument upon the film thickness and friction force. They found a direct correlation between load fluctuations due to the dynamics of the system and the friction force measured.

3.2. Effect of speed variation upon EHD film thickness

The entrainment speed enters steady-state EHD film thickness formulae with an exponent having a value of about 0.7 for line contacts [4, 6], and 0.67 for point contacts [38]. Experiments carried out for point contacts using optical interferometry have shown values between 0.6 and 0.7 which makes it one of the parameters with the strongest influence upon the film thickness [39, 40]. The EHD film's thickness is much smaller in comparison to its width or length, which makes the inertia forces due to acceleration small relative to viscous and pressure forces. Thus in most practical applications inertial forces in the lubricant film, due to the variation of entrainment speed, are negligible [41, 42]. It can hence be stated that the formation of the film is governed by entrainment and squeeze. Entrainment directly determines the passage of the lubricant through the conjunction, as it is known that once the thickness is established by the conditions at the inlet, in order to maintain the continuity of flow the lubricant travels along the contact with its

thickness unchanged. The squeeze effect occurs when one solid body drops or rebounds on a lubricated surface, or when the two solid surfaces approach or separate rapidly due to a variation of the entrainment speed.

Research into the effect of entrainment speed variation has addressed five particular types of motion:

- speed ramp (start from rest or step of entrainment speed);
- sudden stop (shut down or collapse);
- unidirectional variation of entrainment speed;
- repetitive start-stop;
- reversal of entrainment.

The steady-state EHD film thickness equations predict that, during start from rest under constant acceleration the film must exhibit a wedge-like shape until the lubricant completely separates the two surfaces. However, experimental evidence on the behaviour of the elastohydrodynamic film during starting from rest has shown that this type of behaviour is not always obeyed. Depending on its properties and the acceleration, the lubricant either forms a wedge shaped film, or travels through the conjunction as a front of almost constant thickness until it reaches the outlet of the contact and a complete film is established over the whole contact area [43, 44].

Figures 11 and 12 compare the theoretical, steady state and measured transient film profile at certain time intervals after start of motion from rest. The entrainment speed is increased from 0 to 0.2 m/s at an average acceleration of 5 m/s^2 [44]. A wedge-like film was only formed for low viscosity and pressure/viscosity coefficient oils, and under low accelerations.

When a stepped front of lubricant is formed, the thickness of this step is larger than the film thickness corresponding to the entrainment speed reached at the outlet of the contact. For high viscosity and pressure/viscosity oils the ratio of the thickness of the first front to that corresponding to a steady state thickness can be as large as 3:1. As the entrainment speed increases further, a second stepped shaped front forms as soon as the first front exits the contact.

If the acceleration is sufficiently high, the time needed for the speed to reach its final value is equal to or less than the time needed for the first front to arrive at the exit of the contact. In this case the two fronts overlap and the film thickness overshoot results in a series of dampened oscillations about the thickness corresponding to the final, steady state entrainment speed, as illustrated in Fig. 13.

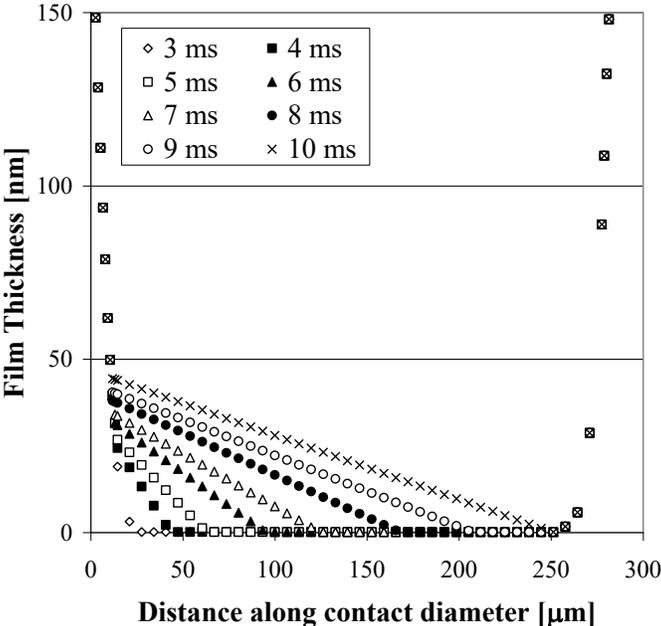


Figure 11. Theoretical film build-up at sudden start of motion.

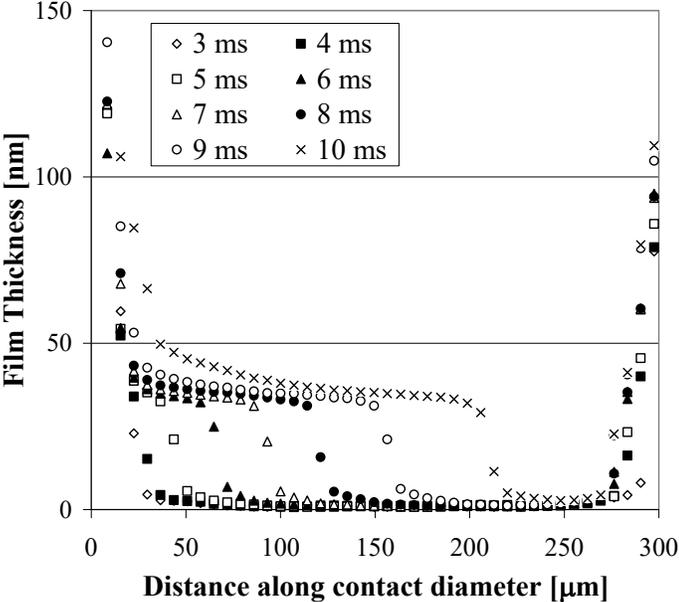


Figure 12. Measured film build-up at sudden start of motion.

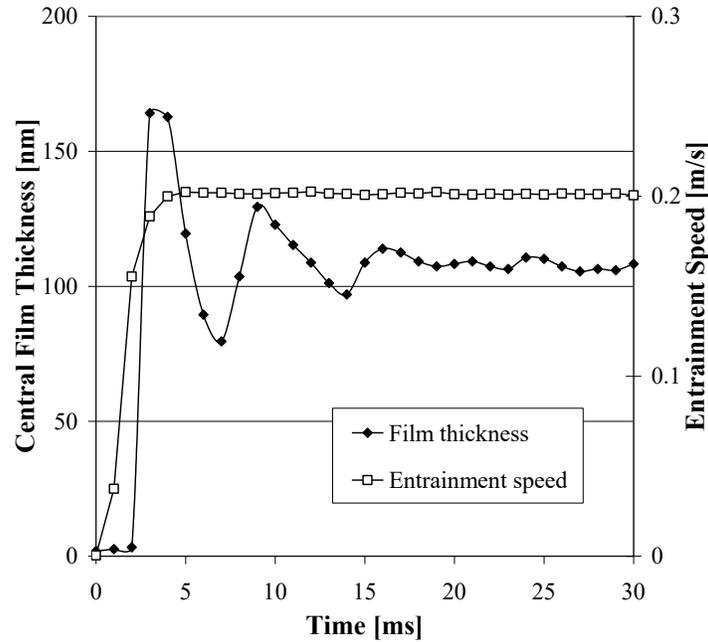


Figure 13. Central film thickness during sudden start of motion.

The acceleration in this case is 50 m/s^2 , for which the time needed for the first front to travel through the contact is about 3.3 milliseconds. This is close to the theoretical time of 4 milliseconds required by the entrainment speed to achieve its final value. The variation of the entrainment speed is also seen in Fig. 13. Similar oscillations of the film thickness were observed when a sudden step was applied to the entrainment speed of a contact running at steady state conditions [45].

This behaviour is analogous to the behaviour of any system incorporating a spring and dashpot which is suddenly taken out of its steady-state condition. Numerical analysis of the dynamics of the start-up process of a similar system revealed induced oscillations in the film thickness; however those were of much smaller amplitude and larger frequency than the values measured [46]. A possible explanation given by Glovnea and Spikes [44] is based on the work of Kettleborough [47] on film thickness oscillations induced in rapidly-accelerating, hydrodynamic slider bearings. This analysis assumes a delay in the build-up of the hydrodynamic pressure due to the finite time needed for the momentum transfer across the fluid film. Of course, in the case of an elastohydrodynamic film, the situation is more complex, involving the elastic deformations of the surfaces, the transverse compliance of the solid, contacting surfaces, the squeeze effect in the film, and the dynamic behaviour of the loading system. A thorough theoretical analysis including all these factors has not been attempted.

In pure sliding tests, with a moving glass disc and a stationary ball, an intriguing phenomenon takes place, as reported by Kaneta [43]. During the film build up phase, a deep, conical-shaped dimple forms in the central area of the contact. For the particular conditions and lubricant used, the maximum film thickness was observed to be about 60 per cent larger than the final, steady state thickness. The time evolution of the central film thickness also shows dampened oscillations of the central film thickness, but of a different origin from those reported earlier [44, 45]. During the dimple formation the minimum thickness of the film was not observed at the sides of the contact (as is the case for steady state conditions), but at the contact exit. No notable dimple was formed in pure rolling or pure sliding with the ball moving and the disc stationary. Kaneta [43] concludes that the dimple is the result of solidified oil in the centre of the contact, thus slip may occur between the lubricant and the boundary surfaces. It is also worth mentioning that such dimples form more readily with oils that have a relatively low viscosity, but a large pressure/viscosity coefficient. The fact that the pressure/viscosity coefficient is the lubricant property with strongest influence upon film thickness in transient conditions supports the results obtained during sudden halting of motion, as described later.

The situation opposite to that described above is that of sudden decrease of entrainment speed under constant deceleration (sometimes referred to as “contact shutdown”). Studies of the EHD film behaviour under controlled deceleration have revealed that two distinct stages of the film behaviour can be observed. During the first stage both entrainment and squeeze are present, with the latter becoming more dominant as the film thickness decreases. In the second stage entrainment is negligible and the film collapses at the periphery of the contact entrapping the lubricant inside [48, 49].

Central and minimum film thickness variation during collapse is compared to film thickness predicted by steady state behaviour in Fig. 14, obtained for the deceleration of 200 m/s^2 and an initial entrainment speed of 0.5 m/s . Relative film is calculated by dividing the actual value to the initial, steady-state thickness. The figure shows that during the first stage the central film thickness is significantly larger than that predicted by the steady state theory. The origin of this enhanced film is based on two mechanisms: one is squeeze, and the other is the time lag between the film formation at the entrance and the time when the lubricant passes the centre of the contact where the thickness was measured.

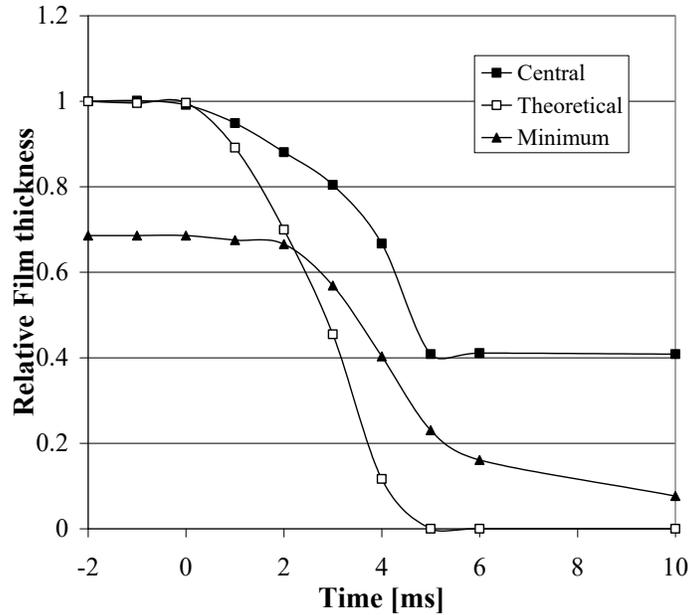


Figure 14. Film thickness variation during deceleration.

Sugimura *et al.* [50] deduce a simple equation for the central film thickness under acceleration, based on the continuity of flow, which takes into account the time of passage of the lubricant through the EHD conjunction.

$$h(u, a) = h_s(u) \left(1 - 0.67 a \xi b / u^2 \right) \quad (15)$$

In this equation, h_s is the steady state film thickness, u is the entrainment speed, a is the acceleration, b is the contact radius, and ξ is a non-dimensional parameter expressing the upstream distance where the film thickness is established. This formula compared well with experimental results obtained earlier by the authors, in accelerating/decelerating tests [51], when ξ took a value between 4 and 9.

A semi-analytical solution developed for film thickness during sudden halting of motion, which incorporates both entrainment and squeeze is given in [42]. This solution was based on the assumption that the inlet geometry does not change significantly during the deceleration phase, an assumption proved experimentally to be correct over about 85 per cent of the stopping time. This behaviour was also observed by Sugimura *et al.* [51]. From Reynolds' theory, the pressure gradient equation takes the form:

$$\frac{\partial p}{\partial x} = 6\eta U \frac{h - h_0}{h^3} + 12\eta \dot{h}_0 \frac{x - x_0}{h^3} \quad (16)$$

This equation can be treated in a fashion similar to that used by Grubin [4] to yield, after some transformations, a central film thickness of the form:

$$h_o = h_s \left(1 - 6.522 \frac{b}{h_o} \frac{\dot{h}_o}{U} \right) \quad (17)$$

This is a Bernoulli-type differential equation, which can be solved analytically to give a central film thickness expressed by the equation:

$$h_o(t) = \frac{h_s}{e^{\left(-\frac{1}{6.522 b} \int_0^t U dt \right)} \left\{ \frac{h_s}{h_i} + \frac{U^{8/11}}{6.522 b} \int_0^t \left[U^{3/11} e^{\left(\frac{1}{6.522 b} \int_0^t U dt \right)} \right] dt \right\}} \quad (18)$$

where h_i is the initial film thickness, whilst the other parameters have the same meaning as in equation (1). If a power relationship between the film thickness and the entrainment velocity is assumed, equation (17) can be transformed into:

$$h_o = h_s \left(1 - 4.74b \frac{a_o}{U^2} \right) \quad (19)$$

This is exactly the same as equation (15), if ζ is equal to 7.11.

The formation of fluid entrapment had been observed previously, but mainly for normal approach, where the film formation is governed by pure squeeze motion. The main difference between the two cases is that in rapid deceleration the load is practically constant, while in pure squeeze the impacting load exceeds many times the falling body's weight.

Tests carried out for a range of working parameters have shown that the central thickness of the entrapped fluid has little dependence on the initial entrainment speed. It instead depends largely on the viscosity and pressure/viscosity coefficient of the lubricant. The latter two parameters also seem closely related to the shape of the entrapment. The curves in Fig. 15 were recorded one second after the entrainment had completely ceased. For a low viscosity and pressure/viscosity coefficient lubricant the entrapment has a bell shape, indicating significant radial fluid flow both during approach and after the complete halting of motion, as indicated by profile 1 in Fig. 15. Profile 2 was obtained for a lubricant with a relatively low, atmospheric pressure, viscosity and a very large pressure/viscosity

coefficient. In this case the relatively modest viscosity ensures a quick collapse of the film at the periphery and a rapid encapsulation of the lubricant inside the contact. The large pressure/viscosity coefficient results in an effective viscosity of about 1 GPa inside the contact, according to the Barus formula. Consequently the core of the lubricant inside the contact escapes slowly, once the entrapment is formed. This combination results in a thick plug of fluid entrapped which shrinks from the edges whilst its thickness remains nearly unchanged in the centre of the contact area. The other two lubricants show consistent behaviour. Profile 3 was obtained for a large viscosity and relatively low pressure/viscosity coefficient, whilst profile 4 indicates a lubricant with large viscosity and relatively large pressure/viscosity coefficient.

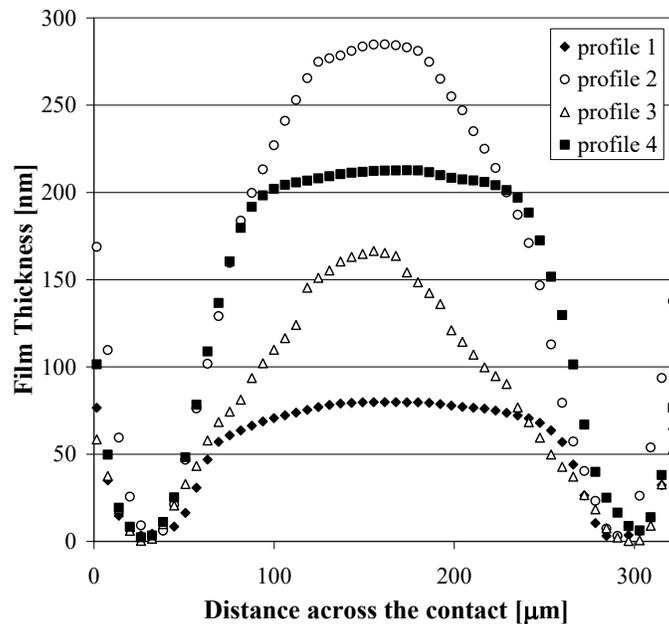


Figure 15. Shape of entrapped lubricant during halting of motion.

Further evidence supporting the dependence of the thickness of the entrapped film on the pressure viscosity coefficient and the viscosity of the lubricant, during sudden halting of motion, has been provided by Ohno and Yamada [52]. They have shown that the maximum thickness of the entrapped film follows a power law dependence of the product between the pressure viscosity coefficient, the viscosity and the acceleration, $(\alpha\eta a)^{0.74}$. Although these tests were carried out at relatively low accelerations of about 0.1 m/s^2 , this relationship has been confirmed for a wider range of accelerations and lubricants as reported by Glovnea and Spikes [48].

Numerical simulations of the contact during deceleration, under the same conditions as those in [49], have been carried out and demonstrated the same qualitative behaviour and relatively good quantitative correlation with the experimental results [53, 54].

A combination of the previous cases is repetitive start/stop motion. This is governed, as explained above, by entrainment and squeeze; however the frequency of the speed variation plays an important role [51, 55, 56]. At low frequencies the film collapses partially or completely, depending on lubricant properties. At higher frequencies, the squeeze effect and entrapment formed prevent the direct contact between the bounding surfaces, which helps avoid seizure of the surfaces during rapid start under load.

Experiments with cyclical variation of the entrainment speed can be of three types: unidirectional variation (pulsating cycle), offset velocity oscillation and reversal of entrainment (alternating cycle). The time lag due to the fluid passage and the squeeze at high frequencies are the two phenomena that govern the film thickness [51, 57]. The latter will also depend on the lubricating properties of the oil, i.e. viscosity and piezo-viscosity coefficient. Figure 16 shows the ratio between measured and theoretical steady state central film thickness for two oils with similar viscosity but with the pressure/viscosity coefficient for oil 2 about 30 per cent higher than that of oil 1.

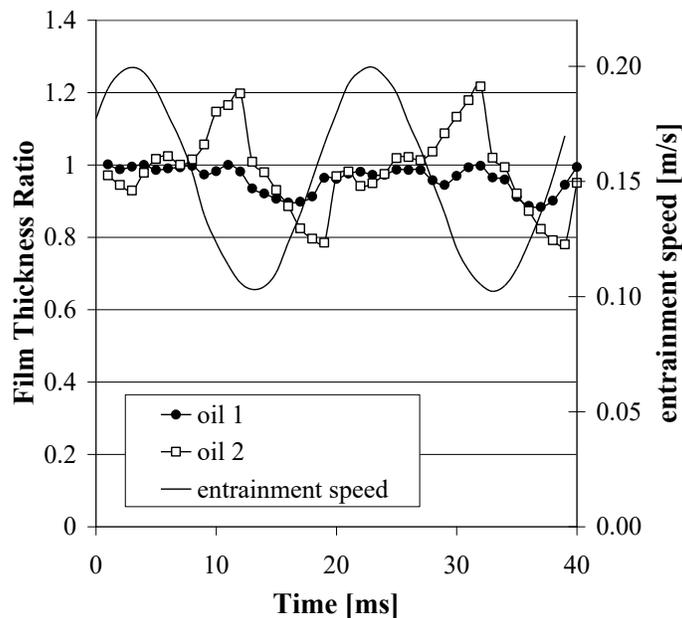


Figure 16. Relative central film thickness during pulsating velocity cycle.

The disc is driven at a constant velocity of 0.1 m/s while the ball's velocity varies sinusoidally between 0.1 m/s and 0.3 m/s, at 50 Hz frequency. As was observed in sudden stop tests, the oil with the larger pressure/viscosity coefficient shows a larger variation from the theoretical profile. This could be attributed to large fluctuations of the film thickness in the inlet, due to the variation of entrainment speed and squeeze. These fluctuations then travel through the contact eventually reaching the centre where the film thickness is measured.

In pulsating speed tests, the central film thickness reaches minimum values some time after the moment of zero entrainment. This delay has been shown to be proportional to the acceleration. Furthermore, unlike the entrainment speed, the central film thickness never falls to zero during the cycle of motion.

The basic behaviour observed in start/stop or pulsating motion is also found in reciprocating motion. This is because, as in the previous cases, the film thickness formation is controlled by squeeze and entrainment [58-61]. The most notable difference between reciprocating motion and unidirectional motion is that in the former case the film recovery is a slower phenomenon attributed to lubricant starvation when reversal of entrainment takes place. Based on a combined experimental and numerical study, Izumi *et al.* show that care must be taken in considering fluid replenishment after reversal when modelling reciprocating motion [60].

The combination of continuity of flow and squeeze of the film can preserve a film thick enough to completely separate the solid surfaces even if the entrainment speed falls to zero. At the same time these effects cause large variations of the film thickness, which induce significant fluctuations of the local pressure over the contact area, which will likely have a negative effect upon the fatigue life of the contact.

3.3. Effect of micro-geometry upon EHD film behaviour

Early measurements of film thickness in systems with varying geometry such as cams and gears were carried out using the electrical capacitance method [19, 20]. In addition to its limited precision this method also does not have the ability to map the contact thickness over its entire area. Optical methods have this ability but it is extremely difficult to implement them into a varying geometry system since, in these systems, the contact point does not remain stationary during the working cycle. This difficulty, together with the development of numerical models of EHD lubrication regime and their improved ability of simulating transient events, means there are no experimental results on the behaviour of EHD films with variable macro-geometry reported during the past two decades. Conversely the effect of

micro-geometry (roughness) on the behaviour of EHD films has been intensely studied, not only theoretically but also experimentally.

There are two approaches to the study of asperity behaviour in EHD lubrication: one is to use real, randomly rough surfaces; the other is to use artificially manufactured roughness. The advantage of the latter approach is that the shape of the un-deformed asperities is well known and a direct comparison to their behaviour inside the contact can be easily made. This helps in the understanding of asperities compression and micro-EHD film formation and enables the development of general rules which can be then applied to more complex, real-roughness systems. These asperities can be obtained by sputtering chromium on the contacting surfaces, (usually on the steel ball). Profiles of such features are shown in figures 17(a) and 17(b).

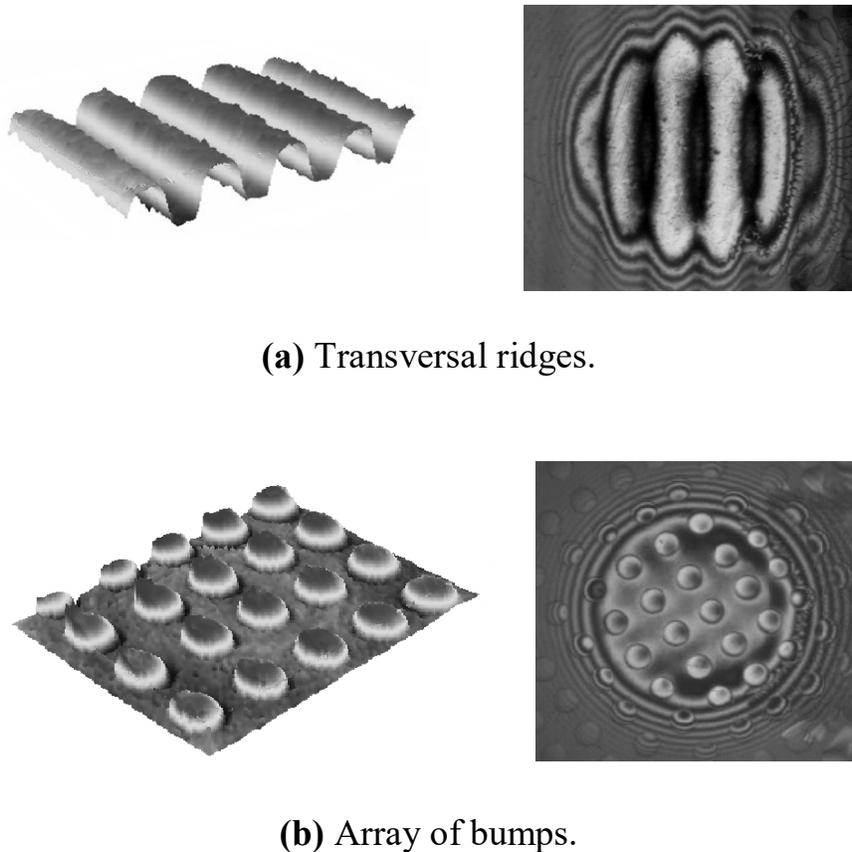


Figure 17. Model roughness used in the study of mixed lubrication.

In this chapter, experimental evidence of the lubricant film thickness in contacts during the passage of simulated roughness is presented. The main types of micro-geometry features approached in various studies are:

- ridges, which can be longitudinally, transversally or obliquely oriented relative to the rolling direction [61-72];
- array of bumps [61, 64, 73-75];
- grooves (longitudinal or transversal) [76];
- single or arrays of circular dents [77-81].

3.3.1. Influence of asperities on local EHD film features

The existence of micro-elasto-hydrodynamic film at the crests of the asperities has been revealed in many studies [62, 64, etc.]. Detailed experiments by Choo *et al.* [72, 78] show that for both circular-bump asperities and transverse ridges a local “horse-shoe” shape film forms with an orientation opposing that of the macro contact. That is, the minimum film thickness is found at the trailing edge of the asperity and not at the leading edge, as is the case of the smooth, macro-EHD contact. The shape of the micro-EHD film formed by transverse ridges and circular bumps can be observed in figures 17(a) and 17(b).

Similar behaviour was observed by Kaneta *et al.* for transversely oriented ridges in pure rolling conditions [68, 71] and was attributed to the micro-squeeze action of the leading edge of the asperity, which forms a divergent conjunction just ahead of the asperity.

It can be considered that, when approaching the high-pressure region at the inlet of the contact, an asperity is compressed elastically generating at same time a perturbation of the film thickness. As the flow inside the contact is dominated by the Couette mechanism, both the compressed asperity and the film perturbation subsequently travel along the contact unchanged, at speeds that can be equal or not, depending on the slide/roll ratio. The film perturbation moves at the average speed of the surfaces, while the asperity itself moves at the speed of the surface to which it is attached. This makes the local shape of the perturbed film depend strongly on the geometry of the asperity and on the slide/roll ratio. By comparing the behaviour of the film in which the featured surface moved slower or faster than the smooth surface, Kaneta and co-workers concluded that film thickness fluctuations also depend on the wavelength of the asperities [71].

The presence of sliding makes the local shape of the film depend also on the rheology of the lubricant, as high shear rates make the lubricant shear-thin and consequently the local effective viscosity is smaller than in non-sliding conditions. Venner *et al.* [67] showed experimentally a lower asperity deformation than was theoretically predicted. This could only be explained by taking into account non-Newtonian effects.

It has also been observed experimentally that in pure rolling conditions, a dimple, or fluid entrapment forms both in front of and behind the asperities,

travelling unchanged through the contact [64, 65, 70, 72]. A similar effect has been observed by Ehret *et al.* with square shaped asperities [74] and Kaneta and Nishikawa for a single circular bump [64]. These variations in the film, in conjunction with the flattening of the asperities, give rise to large fluctuations in the local pressure, with potential implications upon the fatigue resistance of the materials. In rolling/sliding contacts these large pressure fluctuations are amplified by the fact that the film perturbation and the deformed asperity (responsible for the perturbation in the film) travel at different speeds, thus they overlap during passage through the contact.

3.3.2. Thickness of the micro-elastohydrodynamic film

The thickness of the film at the peak of asperities, which are rounded and not step-like, is dominated by the radius of curvature of the crest of the asperity. The ratio of the overall, central and crest film thicknesses, based on the Hamrock and Dowson equation for EHD film thickness is given by the relationship [38]:

$$\frac{h_{\text{asperity}}}{h_c} = \left[\frac{(R'_x)_{\text{asperity}}}{(R'_x)_{\text{macro}}} \right]^{1.134-X} \left[\frac{\left(1 - 0.61e^{-0.75(R'_y/R'_x)^{0.64}}\right)_{\text{asperity}}}{\left(1 - 0.61e^{-0.75(R'_y/R'_x)^{0.64}}\right)_{\text{macro}}} \right] \quad (20)$$

In this expression X is an exponent, which takes a value of 0.67 in the Hamrock and Dowson equation, and varies between 0.64 and 0.67 for the two lubricants tested in [72, 73]. In the case when the asperities are sputtered on a ball $R'_{x \text{ ball}} = R'_{y \text{ ball}}$ and the denominator of the second square bracket becomes $(1 - 0.61e^{-0.75})$. Moreover when the asperities are circular, the second bracket becomes unity and equation (20) simplifies to:

$$\frac{h_{\text{asperity}}}{h_c} = \left[\frac{(R'_x)_{\text{asperity}}}{(R'_x)_{\text{ball}}} \right]^{1.134-X} \quad (21)$$

It follows that smooth surface equations can be used to estimate film thickness at the crest of asperities as long as the latter are not step shaped.

3.3.3. Asperity height recovery

Theoretical studies predict that under pure rolling conditions the asperities recover their original, undeformed height completely as overall film thickness

increases [86, 87]. In other words the ratio between the deformed and undeformed height (known as ‘amplitude ratio’), varies between zero (asperities completely flattened) and unity (asperities fully recovered), with the film thickness. This is supported to some extent by experimental results [66, 67, 71]. At the same time, experiments carried out over an extended range of the speed parameter have shown departures from the theoretical predicted behaviour. The first observation is that the amplitude ratio never reaches zero even in static conditions, i.e. there is not absolute conformity between the surfaces. The second is that the amplitude ratio can exceed unity, with the increase of the film thickness, to fall again at greater film thicknesses [65, 72]. The increase of the amplitude of the asperities can be explained if the behaviour of a single, transversal ridge is considered [70].

As seen in Fig 18, thicker film regions form in front of the asperity, and of lesser magnitude behind. This is as a result of the perturbation of the inlet geometry by the asperity or the formation of a fluid entrapment. Similar effects have been observed by Guanteng *et al.* [65], Kaneta [68], Felix-Quinonez [69] in transversal ridges and Felix-Quinonez for flat, square shaped asperities [75]. If the in-contact asperity height is calculated as the difference between the film thickness at the tip and that at the valley, it would result an enhanced height, as reported in [72].

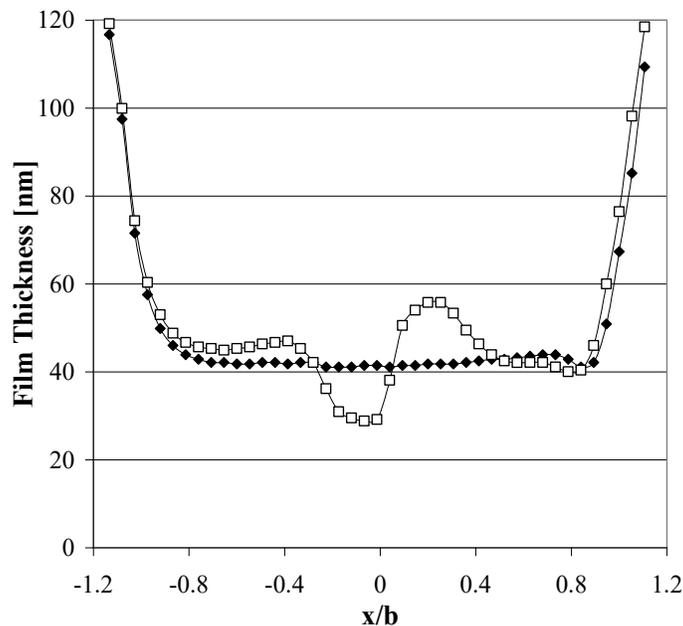


Figure 18. Fluid entrapment associated with a transverse ridge.

Subsequent decrease of the amplitude with the increased film thickness is more difficult to explain, but it is believed to be caused by increasing compliance of the rough surfaces due to the dynamics of the lubrication process [65, 72].

3.3.4. Influence of dents on local EHD film features

Whilst artificial ridges or bumps are just a tool in the understanding of the lubrication phenomena associated with real surface roughness, deliberately induced dents or groves are seen as a method of improving the tribological behaviour of lubricated machine components [77-79]. Although this method is proven in conformal contacts, it is not clear yet whether they are as effective in non-conformal, high pressure contacts. Recent experimental studies have focused on the effect of the depth of the features, pressure distribution and on the validation of numerical analyses [80-85].

Kaneta and Nishikawa [64] have observed a reduction of film thickness for a longitudinal groove, which they attribute to a reduction in the inlet pressure and to side leakage. Side leakage does not appear to occur in the case of circular dents, however depending on their depth, they distort considerably the pressure distribution and the convergence at the inlet region of the contact. Deep cavities are associated with a decrease in local film thickness while shallow ones have the opposite effect. Mourier *et al.* performed a comparative study of the behaviour of circular dents and concluded that a highly viscous oil forms inside the cavity [80]. When sliding is present, this volume of fluid is forced out of the cavity and elastically deforms the contacting surfaces.

Krupka and Hartl [82-85] show that the presence of deep, conical-shaped dents result in local film thickness reduction downstream, when the textured surface moves slower than the opposing, smooth surface. This effect diminishes with a reduction in the depth of the features and has not been observed in the case of the featured surface moving faster than the smooth surface.

Using Raman spectroscopy Vergne and Ville [81] measured the pressure distribution over a circular dent, in a pure sliding contact. They found that the presence of the dent strongly distorts the smooth-surface pressure distribution, generating pressure peaks of a magnitude twice that of the maximum Hertzian pressure.

These studies give a qualitative estimation of the behaviour of the lubricant and tendencies of the film thickness around the features, allowing future optimisation of micro-textured surfaces in elastohydrodynamic contacts.

4. Conclusions

Significant work, both theoretical and experimental, has been carried out during the past decade on the effect of transient conditions upon the behaviour of elastohydrodynamic films. Experimental studies have given us a better insight into the response of the lubricant film alone and lubricant-film/contacting-bodies system to variations of entrainment speed, geometry and load.

There are a series of findings which are quite clear now, regarding the behaviour of the elastohydrodynamic film in non-steady conditions.

- During pure impact loading, the dynamic load exhibits dampened oscillations, however these do not translate into oscillations of the film thickness.
- Impact loading on a steady-state contact causes fluctuations of the film thickness. These oscillations are of constant amplitude, in case of forced force fluctuations, or dampened in case of a step load.
- The film inside the contact is very stiff and does not respond instantaneously to a variation of the load. Nevertheless squeeze effect at the inlet, where the pressure is relatively low, causes perturbations of the film which subsequently travel through the EHD conjunction. It follows that impact loading only translates into film thickness fluctuations when both entrainment and squeeze are present, the latter to create film perturbations in the inlet and the former to convect these through the EHD conjunction.
- The effect of forced and free vibrations is still little studied and needs more attention in the future, in order to elucidate the effect of vibration parameters and lubricant properties upon elastohydrodynamic film behaviour.
- Sudden increase of the entrainment speed, under constant load, generates dampened oscillations of the film thickness. Currently there is no quantitative agreement between theoretical and numerical results, which invites further research into the nature of those oscillations and the effect of various parameters upon their characteristics.
- During sudden halting of motion the film collapses in two stages. In the first stage the film thickness rapidly falls, while the geometry of the conjunction shows little change. In the second stage the central film thickness stays nearly constant, while the contact closes at all sides, entrapping a plug of lubricant inside. Detailed experimental studies have revealed the effect of lubricant properties upon the thickness and the shape of the fluid entrapment.

- Analytical and semi-analytical solutions to the film behaviour under rapid deceleration, based on time of fluid passage and on a combined entrainment/squeeze effects, led to similar results.
- Experimental studies on the effect of the micro-geometry (*roughness*) of the surfaces upon the elastohydrodynamic contacts, have confirmed the existence of micro-elastohydrodynamic films. The geometry of the film, at the tip of the asperities, is reminiscent of the macro-contact horse shoe shape, but oriented towards the trailing edge of the contact.
- It has been found that for round asperities the thickness of the micro-elastohydrodynamic film can be calculated with the relationships known for macro-contacts.
- Flattened inside the contact, the roughness recovers completely its original height as the overall film thickness increases. It has also been found that with further increase of the film, the height of the asperities inside the contact, measured from their tip to the bottom of the valleys, can exceed the initial, un-deformed height. This was explained by the formation of fluid entrapments in the valleys, due to the perturbation of the inlet geometry as an asperity approaches the high pressure zone.
- The trend of an increase of the asperity height with the overall film thickness does not continue at larger films; to the contrary, at a certain value of the speed parameter, the height of the asperities decreases. Although this behaviour has been attributed to the increasing compliance of the rough surfaces due to the dynamics of the lubrication process, more investigations are required to elucidate this phenomenon.

Regarded as a dynamic system, the lubricating film and the contacting bodies react when they are forced out of a steady state condition. They respond by generating oscillations in the film thickness, of constant amplitude in case of repetitive perturbations, or of decaying amplitude in case of singular perturbing event. From this point of view, a sudden variation of the applied load, or a step in the entrainment speed have similar effects upon the film thickness. The behaviour of the elastohydrodynamic film in transient conditions is governed by entrainment and squeeze mechanisms, which act to maintain the dynamic balance of the external load and the pressure-generated reaction.

References

1. Larson, R. 1996. PhD Thesis, Luleå University of Technology; ISSN 0348-8373.
2. Johnson, K.L. 1985. *Contact Mechanics*. Cambridge University Press, Cambridge, UK.
3. Cameron, A. 1966. *The Principles of Lubrication*. Longmans, UK.

4. Grubin, A.N., and Vinogradova, I.E. 1949. Central Scientific Research Institute for Technology and Mechanical Engineering, Book No. 30, Moscow, 1949. D.S.I.R Trans. No. 337.
5. Cameron, A. 1985. *Tribology Int.*, **18**, 92.
6. Dowson, D., and Higginson, G.R. 1977. *Elasto-hydrodynamic Lubrication*. Pergamon, London, UK.
7. Cameron, A., and Gohar, R. (1966). *Proc. Royal Soc. London*. **A291**, 520.
8. Crook, A.W. 1958. *Phil. Trans. Royal Soc. London A*, **250**, 387.
9. Archard, J.F., and Kirk, M.T. 1960. *Phil. Trans. Royal Soc. London A*, **261**, 532.
10. McCarthy, D.M.C., Glavatskih, S.B., and Sherrington, I. 2005. *IMEchE J. Engineering Tribology*, **219**, 179.
11. Furey, M.J. 1961. *ASLE Trans.*, **4**, 1.
12. Tallian, T.E., Chiu, Y.P., Huttenlocher, D.F., Kamenshine, J.A., Sibley, L.B., and Sindlinger, N.E. 1964. *ASLE Trans.*, **7**, 109.
13. Guanteng, G., Olver, A.V., and Spikes, H.A. 1999. The advancing frontier of engineering tribology, STLE/ASME, (eds: Q. Wang, J. Nethzel and F. Sadeghi), 64.
14. Jenkins, F.A., and White, H.E. 1957. *Fundamentals of Optics*. McGraw-Hill Kogakusha, Ltd, Tokyo, Japan.
15. Francon, M. 1966. *Optical Interferometry*. Academic Press, New York, NY, USA.
16. Foord, C.A., Hamman, W.C., and Cameron, A. 1968. *ASLE Trans.*, **11**, 31.
17. Johnston, G.J., Wayte, R.C., and Spikes, H.A. 1991. *Tribology Trans.*, **34**, 187.
18. Glovnea, R.P., Forrest, A.K., Olver, A.V., and Spikes, H.A. 2003. *Tribology Letters*, **15**, 217.
19. MacConochie, I.O., and Cameron, A. 1960. *ASME J. Basic Eng.*, **82D**, 29.
20. Vichard, J.P., 1971. *J. Mech. Eng. Sci.*, **13**, 173.
21. Safa, M.M.A., and Gohar, R. 1986. *ASME J. Tribology*, **108**, 372.
22. Sanborn, D.M., and Winer, W.O. 1971. *ASME J. Lubrication Technology*, **93**, 262.
23. Hoglund, E., and Jacobson, B. 1986. *ASME J. Tribology*, **108**, 571.
24. Ren, N., Zhu, D., and Wen, S.Z. 1991. *Tribology International*, **24**, 225.
25. Larsson, R., and Lundberg, J. 1998. *STLE Tribology Trans.*, **41**, 489.
26. Sugimura, J., and Spikes, H.A. 1997. *Elastohydrodynamics '96* (Ed. D. Dowson *et al.*), pp. 91-100 (Elsevier).
27. Wijnant, Y.H., Venner, C.H., Larsson, R., and Eriksson, P. 1999. *ASME J. Tribology*, **121**, 259.
28. El Kilali, T., Perret-Liaudet, J., and Mazuyer, D. 2004. *Proc. 30th Leeds-Lyon Symposium on Tribology* (Elsevier, ISBN: 9780444517067), 409.
29. Sakamoto, M., Nishikawa, H., and Kaneta, M. 2004. *Proc. 30th Leeds-Lyon Symposium on Tribology* (Elsevier, ISBN: 9780444517067), 391.
30. Kaneta, M., Ozaki, S., Nishikawa, H., and Guo, F. 2007, *IMEchE J. Engineering Tribology*, 221, 271.
31. Chu, H.M, Lee, R.T., and Chiou, Y.C. 2004. *IMEchE J. Engineering Tribology*, **218**, 503.
32. Workel, M.F., Dowson, D., Ehret, P., and Taylor, C.M. 2000. *IMEchE J. Mechanical Engineering Science*, **214**, 309.

33. Workel, M.F., Dowson, D., Ehret, P., and Taylor, C.M. 2001. *IMEchE J. Engineering Tribology*, **215**, 211.
34. Ahrstrom, B.O. 2001. *Tribology International*, **34**, 809.
35. Evans, C.R., and Johnson, K.L. 1986. *Proc. IMechE*, **200**(C5), 303.
36. Glovnea, R.P., and Spikes, H.A. 2005. World Tribology Congress III, Washington DC, USA.
37. Ciulli, E., and Bassani, R. 2006. *IMEchE J. Engineering Tribology*, **220**, 319.
38. Hamrock, B.T., and Dowson, D. 1981. *Ball Bearing Lubrication: The Elastohydrodynamics of Elliptical Contacts*. J. Wiley, New York, NY, USA.
39. LaFountain, A.R., 1999. PhD thesis, University of London.
40. Glovnea, R., Olver, A.V., and Spikes, H.A. 2005. *Tribology Trans.*, **48**, 328.
41. Chang, L. 2000. *Tribology Trans.*, **43**, 116.
42. Glovnea, R.P., and Spikes, H.A. 2001. *ASME J. Tribology*, **123**, 262.
43. Kaneta, M. 1999. *Lubrication at the Frontier*. Elsevier Science B.V, p. 25.
44. Glovnea, R.P., and Spikes, H.A. 2001. *IMEchE J. Engineering Tribology*, **215**, 125.
45. Glovnea, R.P., and Spikes, H.A. 2003. *Lubrication Science*, **15-4**, 311.
46. Popovici, G., Venner, C.H., and Lugt, P.M. 2004. *ASME J. Tribology*, **126**, 258.
47. Kettleborough, C. F. 1974. *J. Mech. Eng. Sci.*, **16**, 357.
48. Glovnea, R.P., and Spikes, H.A. 2000. *STLE Tribology Trans.*, **43**, 731.
49. Glovnea, R.P., and Spikes, H.A. 2001. *ASME J. Tribology*, **123**, 254.
50. Sugimura, J., Okumura, T., Yamamoto, Y., and Spikes, H.A. 1999. *Tribology International*, **32**, 117.
51. Sugimura, J., Jones, W.R. Jr., and Spikes, H.A. 1998. *ASME J. Tribology*, **120**, 442.
52. Ohno, N., and Yamada, S. 2007. *IMEchE J. Engineering Tribology*, **222**, 279.
53. Zhao, J., and Sadeghi, F. 2003. *ASME J. Tribology*, **125**, 76.
54. Holmes, M.J.A., Evans, H.P., and Snidle, R.W. 2003. *Tribology and Interface Engineering Series*, **41**, 79.
55. Sugimura, J., 2002. *J. Jap. Soc. Trib.*, **47**, 752.
56. Glovnea, R.P., Spikes, H.A., and Jones, W.R. 2002. *J. Synthetic Lubrication* (Leaf Coppin), 191.
57. Glovnea, R.P., and Spikes, H.A. 2004. *Transient Processes in Tribology* (Elsevier), **43**, 401.
58. Jolkin, A., Larsson, R., and Ehret, P. 2000. *Proc. Int. Tribology Conf.*, Nagasaki, 313.
59. Glovnea, R.P., and Spikes, H.A. 2002. *Tribology Trans.*, **45**, 177.
60. Izumi, N., Tanaka, S., Ichimaru, K., and Morita, T. 2004. *Transient Processes in Tribology* (Elsevier), **43**, 565.
61. Wang, J., Hashimoto, T., Nishikawa, H., and Kaneta, M. 2005. *Tribology International*, **38**, 1013.
62. Guanteng, G., Cann, P.M., Spikes, H.A., and Olver, A.V. 1999. *Proc. 1998 Leeds Lyon Symposium on Tribology*, Elsevier Tribology Series 36 (Ed. D. Dowson), 175.
63. Kaneta, M., and Nishikawa, H. 1999. *Lubrication at the Frontier* (Ed.: D. Dowson *et al.*), Elsevier, 185.
64. Kaneta, M., and Nishikawa, H. 1999. *IMEchE J. Engineering Tribology*, **213**, 371.
65. Guanteng, G., Cann, P.M., Olver, A.V., and Spikes, H.A. 2000. *ASME J. Tribology*, **122**, 65.

66. Venner, C.H., Kaneta, M., and Lubrecht, A.A. 2000. *Thinning Films and Tribological Interfaces* (Ed: D. Dowson *et al.*), Elsevier, 25.
67. Venner, C.H., Kaneta, M., Nishikawa, H., and Jacod, B. 2000. *Proc. Int. Tribology Conf.*, Nagasaki, Japan, 631.
68. Kaneta, M., Tani, N., and Nishikawa, H. 2003. *Tribological Research and Design for Engineering Systems*, (Ed: D. Dowson *et al.*), Elsevier, 101.
69. Felix-Quinonez, A., Ehret, P., and Summers, J.L. 2003. *ASME J. Tribology*, **125**, 275.
70. Glovnea, R.P., Choo, J.W., Olver, A.V., and Spikes, H.A. 2003. *ASME J. Tribology*, **125**, 275.
71. Kaneta, M., Nishikawa, H., and Matsuda, K. 2006. *Proc. IUTAM Symposium*, Springer Netherlands, 189.
72. Choo, J.W., Olver, A.V., and Spikes, H.A. 2007. *Tribology International*, **40**, 220.
73. Choo, J.W., Glovnea, R.P., Olver, A.V., and Spikes, H.A. 2003. *ASME J. Tribology*, **125**, 533.
74. Ehret, P., Felix-Quinonez, A., Lord, J., Larsson, R., and Marklund, O. 2001. *Proc. Int. Tribology Conf.*, Nagasaki, Japan, 478.
75. Felix-Quinonez, A., Ehret, P., and Summers, J.L. 2005. *ASME J. Tribology*, **127**, 51.
76. Yagi, K., Kyogoku, K., and Nakahara, T. 2004. *Transient Processes in Tribology* (Elsevier), **43**, 429.
77. Etsion, I., Kligerman, Y., and Halperin, G. 1999. *Tribology Trans.*, **42**, 511.
78. Wang, X., Kato, K., Adachi, K., and Aizawa, K. 2003. *Tribology International*, **36**, 189.
79. Kovalcenko, A., Ajayi, O., Erdemir, A., Fenske, G., and Etsion, I. 2005. *Tribology International*, **38**, 219.
80. Mourier, L., Mazuyer, D., Lubrecht, A.A., and Donnet, C. 2006. *Tribology International*, **39**, 1745.
81. Vergne, P., and Ville, F. 2006. *Proc. IUTAM Symp.*, Springer Netherlands, 201.
82. Krupka, I., and Hartl, M. 2006. *Tribology International*, **40**, 1100.
83. Krupka, I., and Hartl, M. 2007. *ASME J. Tribology*, **129**, 502.
84. Krupka, I., and Hartl, M. 2007. *STLE Tribology Trans.*, **50**, 488.
85. Krupka, I., Hartl, M., Urbanec, L., and Cermak, J. 2007. *IMEchE J. Engineering Tribology*, **221**, 635.
86. Lubrecht, A.A., and Venner, C.H. 1999. *IMEchE J. Engineering Tribology*, **213**, 397.
87. Hooke, C.J., and Venner, C.H. 2000. *IMEchE J. Engineering Tribology*, **214**, 439.



Research Signpost
37/661 (2), Fort P.O.
Trivandrum-695 023
Kerala, India

Recent Developments in Wear Prevention, Friction and Lubrication, 2010: 263-278
ISBN: 978-81-308-0377-7 Editor: George K. Nikas

7. On the Stribeck curve

Michael M. Khonsari¹ and E. Richard Booser²

¹*Dow Chemical Endowed Chair in Rotating Machinery, Louisiana State University,
Department of Mechanical Engineering, Baton Rouge, LA 70803, USA;*

²*Engineering Consultant, Vero Beach, FL 32966, USA*

Abstract. Different regimes of lubrication in sliding at low surface speeds are often characterized by use of the so-called Stribeck curve. This chapter covers the changes in frictional processes encountered in passing from boundary lubrication at the start of motion, to mixed film lubrication at the lowest surface speeds, and finally to lift-off with development of a complete separating fluid film. Particular attention is given to analysis of asperity involvement in load support and friction in the mixed lubrication regime. In the final transition to complete lift-off, definition is also given to contributions by hydrodynamic fluid-film factors. Based on a historical perspective, related wear, wear-in, stick-slip, and lubricant additive effects are also reviewed.

1. Introduction

1.1. The Borromean rings

While this well-known design of three interlinked circles comes in many shapes and forms, importance of the structure is in its wholeness and unity: together the rings are inseparable, but the structure falls apart if any ring is detached. Within the context of tribology, each Borromean ring represents a major lubrication regime: boundary, mixed, and hydrodynamic (Fig. 1).

1.2. Lubricant regimes and tribological components

For many years, tribology researchers have focused on these individual regimes, with less attention to the whole structure. From one view point, this is not surprising: most applications involving conformal contacts – thrust bearings, journal bearings, hydrostatic bearings and the like – are primarily confined to the

hydrodynamic circle; while non-conformal applications – ball and roller element bearings, gears, and cam-followers – are concentrated contacts with surface asperity interactions belonging to the mixed lubrication ring, leaving the third ring to boundary lubrication.

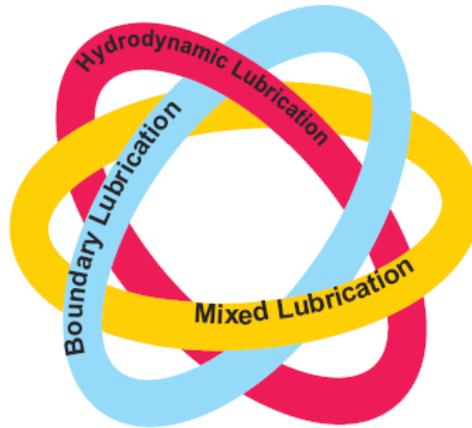


Figure 1. The Borromean rings.

A tribological component, however, does not always operate in a single regime. Take a wet clutch in an automotive transmission, for example. It has a series of disks – alternating metallic separator disks and disks with porous and sand-paper-like friction material in close proximity in a bath of automatic transmission fluid (ATF).

At the onset of engagement, the clutch operates in the hydrodynamic regime, with the disk surfaces completely separated by ATF. The first stage of engagement begins when external pressure is applied to push the disks toward each other and hydrodynamic pressure is developed in the ATF as a result of squeeze action that supports most of the applied load. This period lasts only a fraction of a second, typically on the order of 0.01 s.

As the engagement process progresses, the disk separation gap drops and surface asperities come into direct contact with intimate interaction between disk surfaces. As the engagement continues, the film thickness is further reduced and so is the relative speed between the separator and friction disks. During this stage, the friction-lining material undergoes elastic deformation, as in the mixed lubrication regime, while the relative speed continues to drop.

Eventually, boundary lubrication prevails: relative speed between the disks becomes nil, and disks lock at the conclusion of the engagement. It follows, therefore, that in a full engagement cycle, the lubrication regime undergoes a transition from hydrodynamic, to elastohydrodynamic and mixed, and finally to

boundary lubrication – all within an engagement period lasting on the order of 1 second.

Clearly, therefore, the notion that a tribological component commonly works in any one regime does not hold. To understand the full extent of operating characteristics necessitates first-hand knowledge of all the lubrication regimes, i.e. the whole of the Borromean Rings and not an individual segment.

2. The origin and geniuses of Stribeck curve

The most interesting aspects and perhaps most challenging to predict are the processes that occur at interfaces between these three regimes. Further understanding of these transitions involves one of the most elementary concepts in the science of tribology: “the Stribeck curve” shown in Fig. 2.

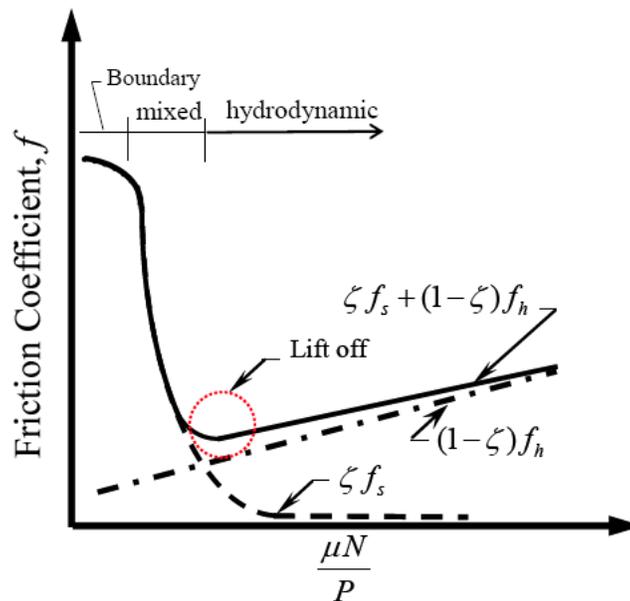


Figure 2. Illustration of the Stribeck curve.

The purpose of this article is not a historical perspective, nor a review of all relevant published papers. The interested reader can refer to Dowson [1] for the history and to Jacobson [2], Lu & Khonsari [3] as well as Wang *et al.* [4] for discussion on recent research developments and associated technical papers. Nonetheless, it is appropriate to relate the original work and discuss where current research seems to be heading.

In 1902, Richard Stribeck reported the results of a series of experimental tests conducted on a journal bearing [5] affirming that friction when plotted as a

function of speed gradually drops, passes through a minimum, and then increases. While the premise of this work trailed that of Hirn [6] by over half a century, the friction-speed characteristic is almost universally referred to as the Stribeck curve, in part due to the work of Gumbel [7] who summarized the results of Stribeck in a single curve.

There are several variations of Stribeck curve shown in Fig. 2. While the ordinate is always the coefficient of friction, the abscissa can be speed, Hersey number $\mu N/P$ – a dimensionless parameter akin to the Sommerfeld number S , which groups speed N [1/s], projected load P [N/m²], viscosity μ [N·s/m²], and the square of the inverted internal bearing clearance ratio $(D/C_d$ or $R/C_r)^2$ – or the film parameter, Λ , which is film thickness divided by the composite surface roughness, $h/\sqrt{\sigma_1^2 + \sigma_2^2}$. For thrust bearings, an analogous dimensionless number $K = \frac{\mu L U}{m^2 W}$

may be defined, where $m = \frac{h_1 - h_2}{B}$ with B representing the length of the slider bearing; h_1 and h_2 denoting the film thickness at the inlet and outlet of the slider, respectively.

Physically, one can postulate that the friction coefficient is a combination of two sources: the solid friction f_s and the viscous friction f_h expressed as follows.

$$f = \zeta f_s + (1 - \zeta) f_h \quad (1)$$

where parameter ζ varies between 0 and 1 corresponding to fully hydrodynamic and dry friction, respectively.

As shown in Fig. 2, the Stribeck curve involves three distinctive features: (1) boundary lubrication at the left; mixed film lubrication where friction drops with increasing speed N up to a transition lift-off point; (2) a transition lift-off point where sufficient hydrodynamic fluid film pressure first develops with increased speed (higher S) to completely support the bearing load; and, finally, (3) a third zone having a complete separating fluid film with properties defined by viscous hydrodynamic behavior.

If load W is maintained constant in the hydrodynamic regime, increasing speed (or Sommerfeld number) translates to increasing film thickness h . Similarly, if the speed is maintained constant, increasing load results in a reduction in the film thickness, as one would intuitively expect. The following variations can be expected in this hydrodynamic regime where film thickness increase is generally proportional to the square root of speed over load:

- load-film thickness: $W \propto 1/h^2$,

- friction force: $F \propto 1/h$, and
 - friction coefficient: $f = F/W \propto h$
- where \propto denotes proportionality.

It follows, then, that coefficient of friction in the hydrodynamic regime increases directly with factors that tend to increase the film thickness, e.g., speed, viscosity, and inversely with load. The Hersey or the Sommerfeld number lump all of these relations into a single dimensionless variable of the form $\mu N/P$. This direct proportionality prevails for fairly large operating speeds so long as the flow remains laminar. Changes in the behavior of friction coefficient are, however, expected during transition to turbulent flows.

3. Mixed lubrication regime

Mixed lubrication regime is the mode where both the hydrodynamics of the flow and the surface roughness actively play a role: the load is partially carried by the pressure generated in the fluid and partially by the asperities, so that the total load is a combination of the two:

$$W_T = W_h + W_A \quad (2)$$

Compared to hydrodynamic lubrication, this mixed lubrication regime is far more difficult to model, especially the boundary lubrication zone where viscous forces do not play a major role and friction coefficient generally remains constant. The film parameter becomes minute, surface asperities are in intimate contact, and influence of the lubricant and its additives involve details of molecular proportions. Nevertheless, recent analytical developments and experimental tests have provided significant insight into behavior in this regime.

Investigating efforts toward modeling and understanding the characteristics of mixed lubrication have captured the attention of many researchers since the 1970s. The interested reader may refer to papers by Christensen [8], Tallian [9], Johnson *et al.* [10], Tsao [11], Jacobson [12], Yamaguchi and Matsuoka [13], Spikes and Oliver [14], Spikes [15].

It what follows, we briefly describe the ideas put forward by Glenick and Schipper [16] and extended by Lu *et al.* [17] to mixed lubrication in conformal contacts. The underpinning idea here is the so-called load-sharing concept developed by Johnson *et al.* [10] to determine the contribution of the asperities and hydrodynamic both in terms of load-carrying capacity and friction force.

Letting γ_1 and γ_2 represent the scaling factors for hydrodynamic and asperities, the total load balance equation can be expressed as follows.

$$W_T = \frac{W_T}{\gamma_1} + \frac{W_T}{\gamma_2} \quad (3)$$

Or, simply:

$$1 = \frac{1}{\gamma_1} + \frac{1}{\gamma_2} \quad (4)$$

where γ_1 and γ_2 are two unknowns that depend on the surface characteristics and film thickness.

Similarly, the friction force is the sum of the hydrodynamic friction and the asperity interaction friction:

$$F_T = F_h + F_A \quad (5)$$

Once the friction force and the total load are determined, the friction coefficient can be easily calculated using:

$$f = \frac{F_T}{W_T} \quad (6)$$

The hydrodynamic friction (traction) force depends on the lubricant viscous behavior. That is:

$$F_h = \int_A \tau_h dA \quad (7)$$

where τ_h represents the fluid shear stress and A is the area. Therefore, the lubricant constitutive equation is needed to proceed. One must note, of course, that Newtonian assumption may not be applicable, particularly in the view of limiting shear stress associated with high pressure rheological behavior.

The friction force due to the asperity interaction can be evaluated using the following expression:

On the Stribeck curve

$$F_A = \sum_{i=1}^N \int_{A_{c_i}} \tau_{A_i} dA_{c_i} \quad (8)$$

where τ_{A_i} represents the shear stress associated with the contacting of a pair of asperities, A_c is the area of the asperities in contact, and N is the number of asperities involved.

The parameter τ_{A_i} can be conveniently related to the contacting asperity friction coefficient by noting that

$$\tau_{A_i} = f_{A_{c_i}} P_{c_i} \quad (9)$$

where P_{c_i} is the central contact pressure of an asperity pair, which can be estimated using, for example, the relationships developed by Greenwood and Williamson [18] given below.

$$p_{c_i} = \frac{2}{3} n \beta \sigma_s \sqrt{\frac{\sigma_s}{\beta} E' F_{3/2} \left(\frac{h_c - d_d}{\sigma_s} \right)} \quad (10)$$

where n , β , σ represent the parameters associated with density, average radius, and standard deviation of asperities, respectively. E' is the equivalent modulus of elasticity, h_c denotes the film thickness, and d_d is the distance between mean plane passing through the asperity summits and the mean plane through the surface height, approximately $1.15 \sigma_s$. The general function $F_{3/2}$ is defined as follows.

$$F_{3/2} \left(\frac{h(x)}{\sigma_s} \right) = \frac{1}{\sqrt{2\pi}} \int_{\frac{h(x)}{\sigma_s}}^{\infty} \left(s - \frac{h(x)}{\sigma_s} \right)^{\frac{3}{2}} e^{-\frac{1}{2}s^2} ds \quad (11)$$

The coefficient of friction between different asperity pairs is often uniform, so that substituting Eq. (9) into Eq. (8) yields the following expression of the total friction force due to summation of all asperities:

$$F_A = f_c \sum_{i=1}^N p_{c_i} dA_i \quad (12)$$

Friction factors f_c for lubricated contacts are rarely available in the literature since test results become distorted not only by the hydrodynamic lift associated with viscous effect of the lubricant at all but very slow speeds, but also by any surface films generated by oil additives. In general, lubricated friction factor for the soft Babbitt alloy in bearing liners running with a steel shaft will range from about 0.10 to 0.15 (as compared with 0.17 to 0.25 as a common starting coefficient of friction), bearing bronzes from 0.20 to 0.25, and steel-on-steel varies more widely from about 0.30 to 0.50.

Asperity contact pressure given in Eq. (10) can be further related to the maximum Hertzian pressure using expression developed by Gelinck and Schipper [19]. The result is:

$$\frac{2}{3}n\beta\sigma_s\sqrt{\frac{\sigma_s}{\beta}}E'F_{3/2}\left(\frac{h_c-d_d}{\sigma_s}\right)=p_h\left[1+\left(a_1n^{a_2}\sigma_s^{a_3}W^{a_2-a_3}\right)^{a_4}\right]^{\frac{1}{a_4}} \quad (13)$$

Substituting $\frac{E'}{\gamma_2}$ for E' , $\frac{F_T}{\gamma_2}$ for F_T , and $n\gamma_2$ for n into Eq. (13), yields the following dimensionless equation in terms of γ_2

$$\frac{2}{3}n'\sigma_s'\sqrt{\sigma_s'}F_T'F_{3/2}\left(\frac{h_c'-d_d'}{\sigma_s'}\right)=\left[1+\left(a_1n'^{a_2}\sigma_s'^{a_3}W^{a_2-a_3}\gamma_2^{a_2}\right)^{a_4}\right]^{\frac{1}{a_4}}\frac{1}{\gamma_2} \quad (14)$$

where $n' = nR\sqrt{\beta R}$; $\sigma_s' = \frac{\sigma_s}{R}$; $F_T' = \sqrt{\frac{2\pi BR'E'}{F_T}}$; $d_d' = \frac{d_d}{R}$.

For line contact, the EHL film thickness equations involve the following dimensionless parameters:

- dimensionless load parameter: $w = \frac{W_T}{E'RB}$
- dimensionless speed parameter: $u = \frac{\mu_o U}{E'RB}$
- dimensionless material parameter: $G = \alpha E'$

For a specified load and speed, the central film thickness can be determined using an elastohydrodynamic (EHL) analysis or readily available film thickness equations for “smooth surfaces,” such as those developed by Moes [20].

$$h'_c u^{\frac{1}{2}} = \left[(\gamma_1)^{\frac{s}{2}} \left(H_{RI}^{\frac{7}{3}} + (\gamma_1)^{-\frac{14}{15}} H_{EI}^{\frac{7}{3}} \right)^{\frac{3}{7^s}} + (\gamma_1)^{-\frac{s}{2}} \left(H_{RP}^{-\frac{7}{2}} + H_{EP}^{-\frac{7}{2}} \right)^{-\frac{2}{7^s}} \right]^{s^{-1}} (\gamma_1)^{\frac{1}{2}} \quad (15)$$

where

$$s = \frac{1}{5} \left(7 + 8e^{\left(-2\gamma_1^{-\frac{2}{5}} \frac{H_{EI}}{H_{RI}} \right)} \right); \quad H_{RI} = 3M^{-1}; \quad H_{EI} = 2.621M^{\frac{1}{5}}; \quad H_{RP} = 1.287L^{\frac{2}{3}};$$

$$H_{EP} = 1.311M^{\frac{1}{8}}L^{\frac{3}{4}}; \quad M = Wu^{-\frac{1}{2}}, \quad L = Gu^{\frac{1}{4}}; \quad H_c = h'_c u^{-\frac{1}{2}}, \quad h'_c = \frac{h_c}{R}$$

Note that in above equation, Johnson's load-sharing concept has been introduced into the formulation by substituting E'/γ_1 for the equivalent modulus E' and W_T/γ_1 for the total load W_T .

Equations (4), (14) and (15) can be combined to determine h_c , γ_1 and γ_2 . Then, Eq. (4) gives the total load W_T . The knowledge of h_c would enable computing the viscous friction and asperity friction in order to determine the total friction coefficient.

Figure 3 shows the results of simulations and experimental work of Lu *et al.* [17] for different oil inlet temperatures. This figure clearly indicates the usefulness of this modeling scheme for mixed lubrication. This concept was recently extended to analyze the lubrication behavior of grease-lubricated journal bearings [21]. There, the viscosity of the grease base oil was used to define a representative Sommerfeld number and a corresponding Stribeck curve for grease-lubricated bearings was developed. Remarkably, the mixed lubrication analysis prediction yielded good agreement with experimental measurements.

Extension of the mixed lubrication analysis to starved line contacts was made by Faraon and Schipper [22]. With this powerful tool, one can efficiently handle very difficult problems involving mixed lubrication with high degree of confidence. An application of this method to spur gears has been recently made by Akbarzadeh and Khonsari [23]. Gear surfaces are often several fold rougher than ball and roller element bearings and contact pressures are typically much higher than journal bearings. Further complications arise because the load and contact pressure change along the line of action during the engagements of gear teeth.

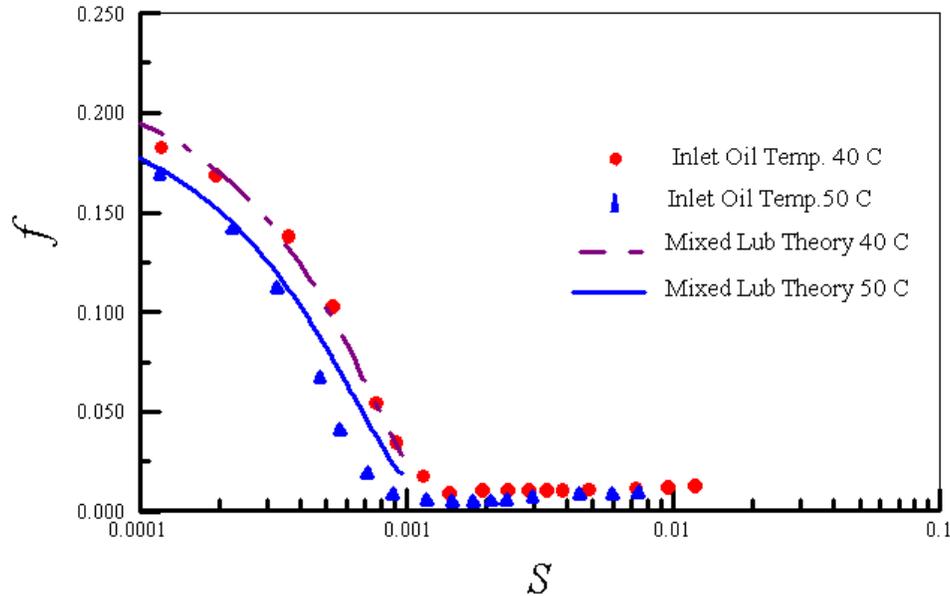


Figure 3. Comparison of predicted and measured friction coefficient as a function of Sommerfeld at two different oil temperatures [17].

In what follows, we confine our attention to applications involving journal bearings using approximate expressions that lend themselves to rapid estimation of the friction.

4. Application to journal bearings

4.1. Prediction of lift-off speed

As shown in the Stribeck curve, coefficient of friction reaches its minimum at the low oil-film thickness corresponding to the “lift-off” speed at which a full fluid supporting film first forms on start up. Above this speed, surface asperities have essentially no influence on friction; below lift-off speed, asperity peaks on the bearing and shaft surfaces begin to interact to restrain shaft motion with increased friction.

This lift-off speed can be observed in many electric motors and other rotating machines by monitoring their coast down. The otherwise steady pattern of deceleration from friction and windage is broken as friction increases when full-film support is first lost [24]. For a journal bearing, this transition speed N then will define the lift-off Sommerfeld Number ($S_o = \mu N/P(C_r/R)^2$) when including oil viscosity μ at the bearing temperature, radial load P , and clearance ratio (C_r/R), where C_r represents the radial clearance. With several bearings in question,

increase in load and lower viscosity (higher temperature or less viscous oil) may help define the results for an individual bearing.

Lift-off speed N_o can alternatively be calculated directly from Reynolds hydrodynamic equation modified as follows for low speeds and small oil film thickness h_o [3].

$$N_o = \left(\frac{h_o}{C_r} \right) P \left[4.678 \mu \left(\frac{R}{C_r} \right)^2 \left(\frac{L}{D} \right)^{1.044} \right]^{-1} \quad (16)$$

This h_o can be used with assumption that lift-off occurs when the ratio λ of oil film thickness rises to 3 times the RMS journal finish, σ_j , when mating with a soft bearing surface commonly polished to a similar finish σ_b by initial operation. This corresponds to a composite film parameter at break-away $\lambda = 3 = h_o / (\sigma_j^2 + \sigma_b^2)^{0.5}$. Asperity interaction can be expected as λ drops below 3 and operation then enters the mixed film region of the Stribeck curve.

Equation (16) can be written in terms of the lift-off Sommerfeld number and the minimum film thickness, h_o [3]:

$$S_o = \frac{h_o}{4.678 C_r (L/D)^{1.044}} \quad (17)$$

where $h_o = \lambda (\sigma_j^2 + \sigma_b^2)^{0.5}$. Note that Eq. (17) is derived by fitting an appropriate curve to the numerical results of Reynolds equation for a journal bearing operating under high eccentricity ratios. While Eq. (17) provides useful information, the prediction is only an estimate. Uncertainties arise as to polishing of asperities during the running in period.

Lubricant additives can also have a significant influence on the film thickness and corresponding lift-off speed. For example, experience with a lead Babbitt bearing surface loaded on a steel journal shows that 10 percent addition of sulfurized lard oil may increase the lubricant film thickness by two fold and at the same time reduce the friction coefficient by a third at lift-off. On the other hand, no effect on lift-off was obtained with a phosphate type anti-wear additive despite formation of a phosphate surface layer on the steel journal. Static coefficient of friction was 0.14 with both additives.

4.2. Extension to hydrodynamic operation

In hydrodynamic lubrication, the Reynolds equation provides a solution for the pressure and friction coefficient [25]. It is possible to develop an expression for a coefficient friction directly by appropriate curve fitting expression. Assuming that laminar flow prevails, the following expression can be derived for plane journal bearings [26].

$$f = (C_r / R) \left(0.43431 + \frac{0.33771}{(L/D)^2} + 19.32261S \right) \quad (18)$$

Equation (18) is best suited for eccentricity ratios in the range of $0.1 < \varepsilon < 0.9$. For very high eccentricity ratios, i.e. just after lift-off, Eq. (18) should be modified to read:

$$f = (C_r / R) \frac{1}{0.047932 \frac{0.097555}{L/D} + \frac{0.13721}{\sqrt{S}}} \quad (0.9 \leq \varepsilon \leq 0.99) \quad (19)$$

This equation is valid for the eccentricity range of $0.9 \leq \varepsilon \leq 0.99$. For $L/D = 1$, Eq. (18) should be valid for $S > 0.023$ and Eq. (19) for $S < 0.023$.

Figure 4 shows the prediction of friction coefficient using equations (18) and (19) along with numerical solutions of Reynolds equation over a wide range of Sommerfeld number. The results are indicative of the usefulness of these relations, particularly for journal bearings with aspect ratios of 1.

5. Related operating effects

The evolving definition of the Stribeck curve aids in understanding a variety of effects in bearings running at low speeds. The general performance map of Fig. 2 and the earlier discussion especially guides considerations of the following.

5.1. Higher friction

With operation at speeds below lift-off, friction typically rises by some 100-fold as rotation slows through the mixed-film zone. This translates into need for higher torque in a turning gear or other power source to maintain slow-speed rotation during cool-down or warm-up of a turbine, in creep speed of a ship or train, in

operations of heavily-loaded construction and earth-moving equipment, or in deep-drilling operations involved in oil and gas exploration where the rotational speed is typically below 200 rpm.

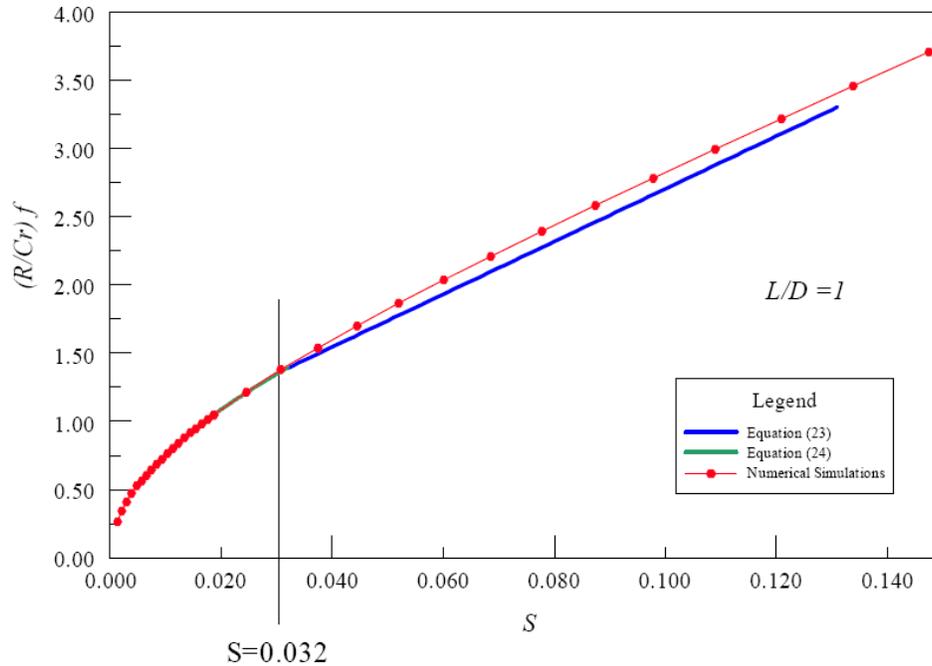


Figure 4. The behavior of friction coefficient in hydrodynamic regime.

5.2. Wear and wear-in

As more-and-more load is carried by metallic surface asperities at slowing speeds in the mixed lubrication zone, a proportionate increase is to be expected in wear rate. If design constraints permit, this deleterious effect can be reduced by applying higher lubricant viscosity, reducing contact unit load, or increasing rotation speed to lower the lift-off point. Use of antiwear and extreme-pressure lubricants may also reduce wear by forming protective surface layers on bearings.

Despite the general difficulty and uncertainty in predicting wear rates of asperity contacts, Maru and Tanaka [27] have demonstrated some relation of wear coefficients with friction coefficients, particularly in the beginning stages of the mixed-film regime.

Initial wear-in at mild operating conditions can provide a limited smoothing of the finish of both bearing and shaft surfaces. As shown in Fig. 5, this useful action moves the lift-off speed and the mixed lubrication region in the Stribeck curve to the left to reduce, or even possibly eliminate, the support needed from

contact of solid asperities [24]. Grease lubrication and some adhering oil-additive films produce a similar effect in lowering the lift-off speed.

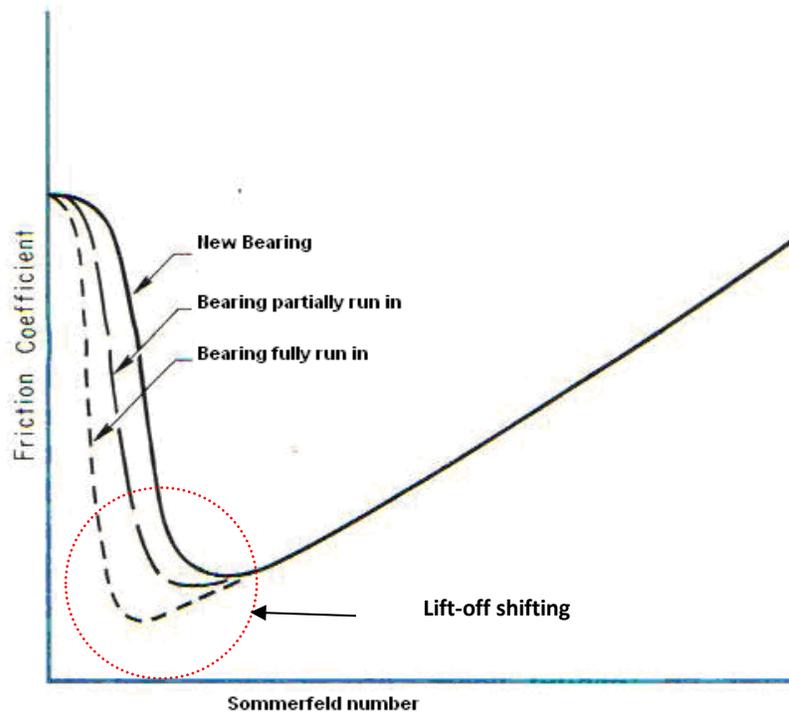


Figure 5. Behavior of Stribeck curve during the wearing in process (adapted from Elwell and Booser [24]).

5.3. Stick-slip chatter

When the shaft surface speed drops below lift-off, asperity peaks begin to interlace, and shaft rotation is restrained with the increased friction. If torsional elastic energy is then accumulated in a slowed or halted flexible shaft, it may then finally jump ahead. As the wind-up torque is then relieved, the shaft may stick again in a repeating stick-slip vibration or “chatter”.

Long, limber shafts in turbines, compressors, oil-well drill shafts, and water-lubricated ship stern-tube bearings; and shaft face seals are susceptible to this chatter. A similar phenomenon may take place in satellite tracking devices swinging back and forth under ultra-slow speeds in ball bearings as they zero in on a target [28]. The hysteresis effects that are brought out by oscillatory motion—with a wide spectrum of amplitude, frequency, acceleration/deceleration effects that may be involved—add largely unexplored dimensions to the behavior of the classical Stribeck curve that await further research.

6. Concluding remarks

The Stribeck curve models low-speed bearing friction in two primary regions. On initial start-up in a mixed region, bearing load support by asperity contact is promptly supplemented by fluid film pressure that continually increases with increasing speed. When increasing speed finally generates sufficient oil film pressure to lift-off and support the total load, asperity contact is eliminated and a hydrodynamic region develops.

The mixed film region involves a complex pattern of load support involving wear and significantly higher friction. Asperity contact involved can be modeled by the Greenwood and Williamson [18] who formulated relations involving density, average radius and elastic modulus. Following lift-off into the hydrodynamic region, a direct expression is available for coefficient of friction from curve fitting of numerical solutions for Reynolds equation.

Accurate and reliable prediction of the performance of tribological components requires a systematic research strategy beyond focusing solely on the individual lubrication regimes. While the steady-state operation of many tribological components may dwell largely in a single lubrication regime, practically all machines go through entire the Stribeck curve when starting from rest. Henceforth, research should be devoted to the full spectrum of operating conditions. To this end, challenging questions remain to be at the interfaces between different lubrication regimes, from boundary to mixed and transition from there to hydrodynamic. This is an important step toward achieving the goals of virtual tribology design to provide the enabling computerized technology that minimizes expensive experimental testing.

Acknowledgements

The authors wish to thank the Drs. Lu, Jang, and Mr. Akbarzadeh (all from the LSU Center for Rotating Machinery) for their assistance and insightful discussion during the course of this investigation.

References

1. Dowson, D. 1998. *History of Tribology*, 2nd Edition, Professional Engineering Publishing, London and Bury St Edmunds, UK.
2. Jacobson, B. 2003. *Tribology International*, **36**, 781.
3. Lu, X., and Khonsari, M.M. 2005. *Tribology Letters*, **30**, 299.
4. Wang, Y., Wang, Q.J., Lin, Ch., and Shi, F. 2006. *Tribology Trans.*, **49**, 52.

5. Stribeck, R. 1902. Kugellager für beliebige Belastungen, *Zeitschrift des Vereines deutscher Ingenieure*, 46(37), 1341 (part I), 46(38), 1432 (part II), 46(39), 1463.
6. Hirn, G. 1854. *Bulletin of the Industrial Society of Mulhouse*, **26**, 188.
7. Gümbel, L. 1916. *Mbl. berl. BezVer. dt. Ing. (VDI)*, 109.
8. Christensen, H. 1972. *Proc. Institution of Mechanical Engineers*, **186**, 421.
9. Tallian, T.E. 1972. *Wear*, **21**, 49.
10. Johnson, K.L., Greenwood, J.A., and Poon, S.Y. 1972. *Wear*, **19**, 91.
11. Tsao, Y.H. 1975. *Trans. ASLE*, **18**, 90.
12. Jacobson, B. 1990. *Wear*, **136**, 99.
13. Yamaguchi, A., and Matsuoka, H. 1992. *ASME J. Tribology*, **114**, 116.
14. Spikes, H.A. and Olver, A.V. 2003. *Lubrication Science*, **16**, 3.
15. Spikes, H.A. 1997. *Lubrication Science*, **9**, 2.
16. Gelinck, E.R.M., and Schipper, D.J. 1999. *ASME J. Tribology*, **121**, 449.
17. Lu, X., Khonsari, M.M., and Gelinck, E.R.M., 2006. *ASME J. Tribology*, **128**, 789.
18. Greenwood, J.A., and Williamson, J.B.P. 1966. *Proc. Royal Society of London A*, **295**, 300.
19. Gelinck, E.R.M., and Schipper, D.J. 2000. *Tribology International*, **33**, 175.
20. Moes, H. 1992. *Wear*, **159**, 57.
21. Lu, X., and Khonsari, M.M. 2007. *ASME J. Tribology*, **129**, 84.
22. Faraon, I.C., and Shipper, D.J. 2007. *ASME J. Tribology*, **129**, 181.
23. Akbarzadeh, S., and Khonsari, M.M. 2008. *ASME J. Tribology*, **130**, 021503.
24. Elwell, R.C., and Booser, E.R. 1972. *Machine Design*, issue 10, 129.
25. Khonsari, M.M., and Booser, E.R. 2008. *Applied Tribology – Bearing Design and Lubrication*, 2nd Edition, John Wiley & Sons, Inc, UK.
26. Jang, J.Y., and Khonsari, M.M. 2004. *J. Engineering Tribology*, **218**, 355.
27. Maru, M.M., and Tanaka, D.K. 2007. *J. Brazilian Soc. Mech. Sci. and Engrg.*, **29**, 55.
28. Khonsari, M.M., and Booser, E.R. 2008. *Machine Design*, issue 6, 80.



Research Signpost
37/661 (2), Fort P.O.
Trivandrum-695 023
Kerala, India

Recent Developments in Wear Prevention, Friction and Lubrication, 2010: 279-314
ISBN: 978-81-308-0377-7 Editor: George K. Nikas

8. Surface characterization, adhesion measurements and modeling of microelectromechanical systems

Xiaojie Xue* and Andreas A. Polycarpou

Department of Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, 1206 West Green Street, Urbana, Illinois 61801, USA

Abstract. Adhesion, also referred to as stiction, is a common failure mechanism occurring in microelectromechanical systems (MEMS) during device fabrication and operation and has become the main barrier to the advancement and wide commercialization of MEMS and miniature devices. To investigate the adhesion behavior of MEMS devices, experimentally and theoretically, different small scale interfacial experiments and continuum-based models have been developed and applied with varying success to explain such phenomena. In this work, a beam-peel-based MEMS experimental setup was designed and built to measure the adhesion energy between MEMS microcantilevers and a substrate at different humidity levels. The microcantilever arrays were separated from the substrate and their fixed ends were directly attached to a piezoactuator to control the beam displacement with sub-nanometer accuracy. The experimental setup was sealed in a chamber with precise humidity control. An in-situ interferometer was used to measure the beam deflection and crack length during the peel test. To examine the effects of surface roughness and relative humidity on adhesion energy, different surface pairs were measured at humidity levels ranging from 40% to 92%. Before testing, the microcantilevers and substrates were scanned using an Atomic Force Microscope (AFM). Surface roughness parameters and the exact probability density function of the asperity heights were extracted and directly entered into a statistical-based roughness model. An Extended-Maugis-Dugdale (EMD) single-asperity meniscus model considering both asperity deformation and solid surface interaction was coupled with the Pearson surface statistical model to develop an improved elastic asymmetrical surface meniscus model. The model compared favorably with the experimental data.

* Currently with Analog Devices Inc.

Correspondence/Reprint request: Professor Andreas A. Polycarpou, Department of Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, 1206 West Green Street, Urbana, Illinois 61801, USA
E-mail: polycarp@illinois.edu

Introduction

By integrating mechanical elements, sensors, actuators, and electronics on a common silicon substrate, Micro-Electro-Mechanical Systems (MEMS) reduce cost, size, weight, and power consumption while increasing performance, production volume, and functionality by orders of magnitude [1]. However, due to the relatively high compliance of micromachined components, large surface-to-volume ratios, and small surface separation distances, surface forces such as van der Waals, capillary, and hydrogen bonding, become dominant over body forces or restoring forces (stiffness of the freestanding microstructures) causing suspended MEMS structures to deflect towards the substrate and even permanently adhere. This phenomenon is called *adhesion*, which is sometimes referred to as *stiction* in the MEMS literature.

Adhesion can occur during the final stage of the micromachining process (release adhesion), during normal operation (in-use adhesion) [2], or while parts are stored (dormancy adhesion). Release adhesion is very common in sacrificial surface micromachining as this type of manufacturing requires the chemical wet-etch of the sacrificial layer and several liquid rinse steps before the device becomes operable. When the device is taken out of the final rinse solution, the remaining liquid trapped between the compliant MEMS structures and the surface below results in strong capillary force pulling the suspended structures down to the substrate. In-use adhesion occurs during normal operation of MEMS devices and can be caused by liquids, especially water from the environment that condenses on the device surfaces [3], mechanical shock, or “over range” input signals. Moreover, some MEMS applications, e.g. a microengine [4], a digital-micro-display (DMD) [5], accelerometer [6] and RF MEMS switch [7], involve surface contact and sliding, which may cause adhesion failures and hinder normal operation of the devices. Dormancy adhesion relates to an adhesion failure that occurs while the device is stored and may be caused by environmental conditions such as humidity [8].

Clearly adhesion is a common failure mechanism occurring in MEMS devices, a major reliability concern, and a barrier to widely commercializing MEMS devices. In order to increase the MEMS fabrication yield and long-term durability, significant efforts have been made to investigate adhesion behavior experimentally and theoretically, e.g., [9-11], to alleviate adhesion problems in MEMS devices, e.g. [12-15], and to repair adhesion-failed devices, e.g. [16, 17].

Adhesion and interaction between a single asperity and a flat surface have mainly been examined by surface force apparatus (SFA) [18] and atomic force microscopy (AFM) [19]. Since adhesion behavior is directly related to the surface energy, a more convenient parameter, *adhesion energy*, is usually used to evaluate

the interaction between two approaching surfaces. A number of techniques have been explored to experimentally study the adhesion behavior between two rough surfaces. The cantilever beam array technique presented by Mastrangelo and Hsu [20] is the most widely used method, in which the adhesion energy is determined by the length of the shortest adhered beam. de Boer and Michalske [21] expanded on this approach and developed a fracture mechanics model to determine the adhesion energy. Using the cantilever array technique, DelRio *et al.* [22] experimentally investigated the adhesion as a function of relative humidity and surface roughness. Jones *et al.* [23] presented experiments and models to examine the adhesion of microcantilevers by means of mechanical actuation. Leseman *et al.* [24] developed an experimental technique termed “beam-peel-test” to measure the effects (under laboratory ambient conditions) of deionized (DI) water and isopropyl alcohol (IPA) on release adhesion, where the microcantilevers were peeled from the substrate using a piezoelectric actuator and the crack length was measured to calculate the adhesion energy.

In addition to cantilever beam array methods, other techniques have also been explored to investigate the in-plane adhesion in MEMS, for example the doubly clamped cantilever beam array presented by Legtenberg *et al.* [25] and the symmetric folded-beam butterfly suspension developed by Alley *et al.* [26]. Moreover, besides the contact between the bottom surface of MEMS structures and the ground surface (in-plane contact), many MEMS applications such as the microengine involve contact between vertical surfaces (sidewalls). A sidewall adhesion tester was developed by Ashurst *et al.* [27] to measure the sidewall adhesion.

Adhesion forces between two rough surfaces in close proximity or contacting each other may arise due to van der Waals, capillary, and electrostatic forces. In a dry environment, the van der Waals interaction usually determines the adhesion behavior of microstructures. Based on the Greenwood and Williamson (GW) statistical roughness model [28], several continuum mechanics-based adhesive models have been proposed based on asperity interactions at various separation regimes. Chang *et al.* [29] developed an adhesion model (CEB model) for rough surfaces under dry contact conditions by combining the IDMT (improved Derjaguin-Muller-Toporov) model [30] with an extended GW model that included elastic-plastic contacts. Kogut and Etsion [31] developed a rough surface adhesion model (KE model) based on their improved single-asperity (finite element-based) model [32]. An advantage of both CEB and KE models is that they are valid for elastic-plastic contact conditions. However, a major drawback of these models is that since they are based on the IDMT adhesion model, they are only valid for a limited range of low adhesion parameter values. Combining the Extended-Maugis-Dugdale (EMD) model [33] and the ISBL (improved sub-boundary lubrication)

model [34], Shi and Polycarpou [35] proposed an elastic-plastic hybrid “dry” adhesion model for rough contacting surfaces which is valid for elastic-plastic contacts and for the entire range of adhesion parameter values.

At high relative humidity (RH) values, the meniscus force due to capillary condensation becomes dominant. Incorporating different single-asperity models into the Greenwood-Williamson statistical model [28], several meniscus models have been presented to predict the effects of surface roughness/texture, lubricant film thickness, and environmental humidity. Based on Israelachvili’s single-asperity capillary model [36], Li and Talke [37] explored the effect of humidity on the adhesion force for multi-asperity contacts using the GW statistical model. Assuming liquid volume conservation, Tian and Matsudaira [38] presented an improved meniscus model for head-disk interfaces (HDIs) covered with a uniform liquid layer. Gao *et al.* [39] proposed yet another model to calculate the adhesion force between a sphere and a lubricated flat surface based on energy change considerations. Taking into account the local redistribution of molecularly thin lubricant, Gui and Marchon [40] proposed a meniscus-based static friction (stiction) model for HDIs. Interestingly, despite the numerous improvements in the abovementioned models, they do not specifically consider the change of the projected meniscus area due to the solid asperity deformation when contacting with the flat solid surface. The meniscus force between the contacting asperities was treated as a constant value independent of the spherical interference. Therefore, these multi-asperity meniscus models are valid only for high relative humidity levels ($> 70\%$) and significantly underestimate the adhesion energy (or force) at intermediate and low humidity levels.

To fully analyze the adhesive contact interaction between two rough surfaces, a thorough understanding and modeling of single asperity/sphere adhesion contact is necessary. Fogden and White [41] extended the Hertz theory to analyze the elasticity of contacting spheres in the presence of capillary condensation. Zhang and Nakajima [42] analyzed the nanometer deformation of the sphere caused only by the Laplace pressure when “just contacting” with the flat surface. Xue and Polycarpou [43] developed a meniscus model for a deformable sphere on a rigid flat surface covering a large range of interference values from non-contact to a fully plastic contact. Using the extended-Maugis-Dugdale (EMD) theory, Xue and Polycarpou [44] presented a single-asperity capillary meniscus model considering asperity deformation due to both contact and adhesive forces.

A different modeling approach to predict adhesive forces in microsystems in the presence of molecularly thin lubricants has also been proposed using contact-mechanics based adhesion models. Incorporating the KE dry contact model [31] with the original sub-boundary lubrication (SBL) adhesion model [45], an alternative model termed improved SBL (ISBL) model [46] was presented to take

into account the presence of molecularly thin lubricant. However, in this model, the very thin lubricant strongly and uniformly adheres to the surface and the meniscus formation is energetically unfavorable. Thus, this model is not applicable for MEMS devices at high humidity levels and cannot capture the adhesion behavior in the presence of mobile lubricant in magnetic storage HDIs.

Since humidity is an important factor that greatly affects the meniscus force and adhesion behavior, the beam-peel-based experimental setup was improved to measure the adhesion energy between MEMS microcantilevers and different substrates at different humidity levels. The method allows measuring and comparing various substrates of different materials and surface topographies at controlled humidity levels using the same microcantilever arrays. Before testing, both the underside of the microcantilevers and the substrates were scanned using an AFM. The surface topographies were analyzed and surface roughness parameters including root-mean-square roughness (R_q), skewness (S_{sk}), kurtosis (S_{ku}), areal density of asperities (η), and average asperity radius (R) were extracted. Also, the exact probability density function of the asperity heights was generated using the measured roughness parameters. The Extended-Maugis-Dugdale (EMD)-based single-asperity meniscus model [44] considering both the asperity deformation and solid surface interaction was coupled with the Pearson surface statistical model to develop an improved elastic asymmetrical surface meniscus model, which compared favorably with the experimental data.

1. MEMS surface characterization

All real surfaces are rough at the microscopic and submicroscopic scales. Contact between two rough surfaces occurs at discrete asperity summits. Surface forces are strong functions of surface properties and separation distance. To accurately analyze these interfacial forces and understand a system's overall behavior, it is necessary to investigate the micro-topography and mechanical properties of the opposing surfaces, especially surface roughness, and how it influences adhesion. A polycrystalline silicon (polysilicon) film of a few micrometers in thickness is the primary structural material for MEMS applications. For improved reliability of MEMS devices, different efforts have been made to characterize the polysilicon mechanical properties including surface roughness effects using different fabrication processes and surface treatments [47-50].

1.1. Sample fabrication

The microcantilevers used in this study were fabricated by Sandia National Laboratories using the 4-layer SUMMiT IVTM process [51]. It begins with single

side polished (SSP) n-type silicon wafers on which a 630 nm oxide layer is thermally grown. An 800 nm thick, low-stress, silicon nitride film is then deposited using low-pressure chemical vapor deposition (LPCVD), which is followed by depositing a 300 nm thick n-doped LPCVD polysilicon layer (referred to as *P0*) on the nitride layer. Then, a 2.0 μm thick sacrificial oxide layer is deposited in an LPCVD furnace on the *P0* and patterned to provide an anchor for the microcantilevers. Two structural polysilicon layers, *P1* and *P2*, are deposited and patterned to create microcantilevers that are approximately 2.6 μm thick. During the process known as “release etching,” the silicon dioxide layers between the polysilicon layers were completely etched away by soaking the sample in 49% hydrofluoric acid (HF) for 15 minutes, and then supercritical CO_2 drying was used to obtain free-standing microcantilever arrays. To obtain a hydrophilic surface, the microcantilevers were then treated by oxygen plasma and the water contact angle on the treated surface was measured and found to be almost 0 degrees. After that, the die was scored and fractured along a line just beyond the anchor of the microcantilevers, generating a test sample with microcantilevers protruding freely beyond the anchor point, as shown in the optical and SEM images of Fig. 1. The number at the top of the beam arrays shows the beam length, 1500 μm , the one at the bottom indicates the beam width, 30 μm , and the numbers at the left represent the spacing between beams in the array, 2-100 μm . The release steps described above are also listed in Table 1.

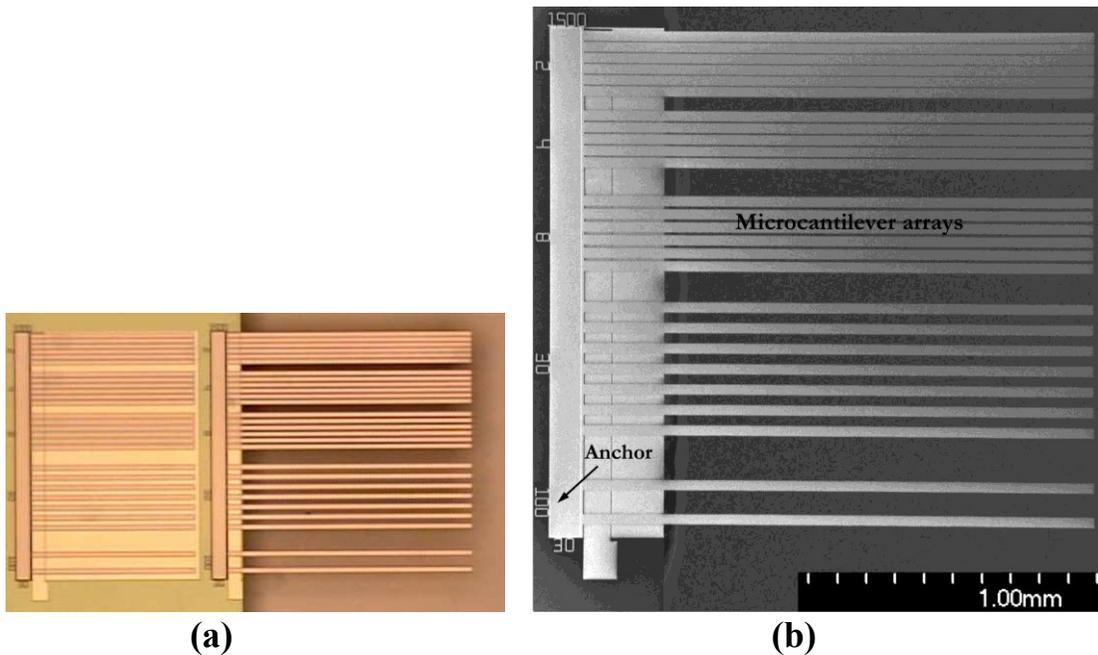


Figure 1. Array of 1500 μm polysilicon microcantilevers. (a) optical image; (b) SEM image.

Table 1 Microcantilever release procedure and sample preparation.

Step	Experimental Procedure
1	Acetone rinse for 30 min
2	IPA rinse for 10 min
3	DI water rinse for 10 min
4	49% hydrofluoric acid for 15 min
5	4:1 CH ₃ OH:H ₂ O rinse for 10 min
6	Pure CH ₃ OH rinse for 10 min
7	Soak in pure CH ₃ OH
8	Supercritical CO ₂ dry
9	Oxygen plasma treatment
10	Separate microcantilevers from substrate

1.2. AFM measurements

In the LPCVD deposition process, parameters such as temperature, pressure, source gas dilution, and deposition time all greatly affect the structure and crystal growth of polysilicon which determine the mechanical and electrical properties of the film [52]. Since the surface roughness is an important geometrical surface property that directly affects adhesion and contact between two surfaces, in this study microcantilevers fabricated by two different runs were used to perform the experiments. The thicknesses of the microcantilevers from the first and second runs were measured to be 2.60 μm and 2.55 μm , respectively. Also, three types of substrates were used in this study, the first being non-doped single crystal silicon wafer, the second being polysilicon film deposited by LPCVD at a temperature of 585 $^{\circ}\text{C}$ for one hour and then annealed for 10 minutes, and the third being a rougher polysilicon substrate fabricated by LPCVD at 590 $^{\circ}\text{C}$ and annealed for 10 minutes. Both the microcantilevers and the substrates were treated by oxygen plasma before the peel experiments.

Due to the dependence on processing conditions and grain growth of polysilicon films, the surface topographical evolution and the roughness of the top and bottom surfaces in micro-fabricated MEMS are known to vary significantly [53]. To more accurately predict the adhesion behavior, the bottom surfaces of the microcantilevers and the substrates were measured using a tapping mode AFM with a root-mean-square noise resolution of 0.05 nm. For each surface, a sample area of 5x5 μm^2 was scanned at four different locations, which ensured at least 200 grains in each image. The data were repeatable and were digitized and exported for further processing. Using a graphics user interface program that extracts numerous

topographical parameters including the Birmingham-14 parameters [54], the following parameters were extracted: standard deviation of surface heights (R_q), skewness (S_{sk}), kurtosis (S_{ku}), areal density of the asperities (η), and average asperity radius (R). The reduced skewness, S_{sk} , is a measure of the symmetry of the surface height distribution and the kurtosis, S_{ku} , is a measure of the concentration of the surface height distribution curve. For a Gaussian distribution, $S_{sk} = 0$ and $S_{ku} = 3$. Note that the skewness and kurtosis of the asperity heights are the same as those of the surface heights, which has been experimentally verified [55]. The standard deviation of the asperity heights σ is related to the standard deviation of surface heights by [56]

$$R_q^2 = \sigma^2 + \frac{3.717 \times 10^{-4}}{\eta^2 R^2} \quad (1)$$

The validity Eq. (1), which was derived for a Gaussian distribution of asperity heights, has also been confirmed numerically for different roughness measurements [57].

Figure 2 depicts representative AFM images for different microcantilever bottom surfaces and substrates. It can be seen that the bottom surface of Run *B* is rougher than that of Run *A*, which could be caused by the LPCVD or the annealing processes. Based on the AFM measurements, the roughness parameters of the individual surfaces were extracted and are listed in the top part of Table 2. The R_q value of Run *B* is more than 2 times higher than that of Run *A*. Clearly, both types of beam surfaces display non-Gaussian distribution of the surface heights. An inaccurate prediction will be made if a Gaussian distribution is assumed for modeling and simulation, as discussed in section 3.3. For the substrates, it is clearly seen that the silicon wafer is much smoother than the LPCVD deposited polysilicon film, with 0.17 nm R_q for the silicon wafer, 5.5 nm R_q for polysilicon deposited at 585 °C, and 7.1 nm for polysilicon deposited at 590 °C. Again, similar to the case of the microcantilevers, the polysilicon substrate shows a non-Gaussian distribution. Note that a 392 nm high-pass filter was performed to remove the low-frequency components (waviness) and a 98 nm low-pass filter was used to eliminate the atomic noise, as discussed in [54]. Also, as discussed in [54], other ranges around these values could be used and the results would be similar.

Based on the AFM measurements, we see that for all five substrates, the areal asperity density is between 20.9 to 111.5 asperities per μm^2 . In the experiment, the beam width is 30 μm and only the data point with contact length no less than 200 μm was used for the calculation. Therefore, the total number of asperities in the

contact area is at least 1.2×10^5 , which assures the feasibility of using a statistical model to calculate the adhesion force and adhesion energy.

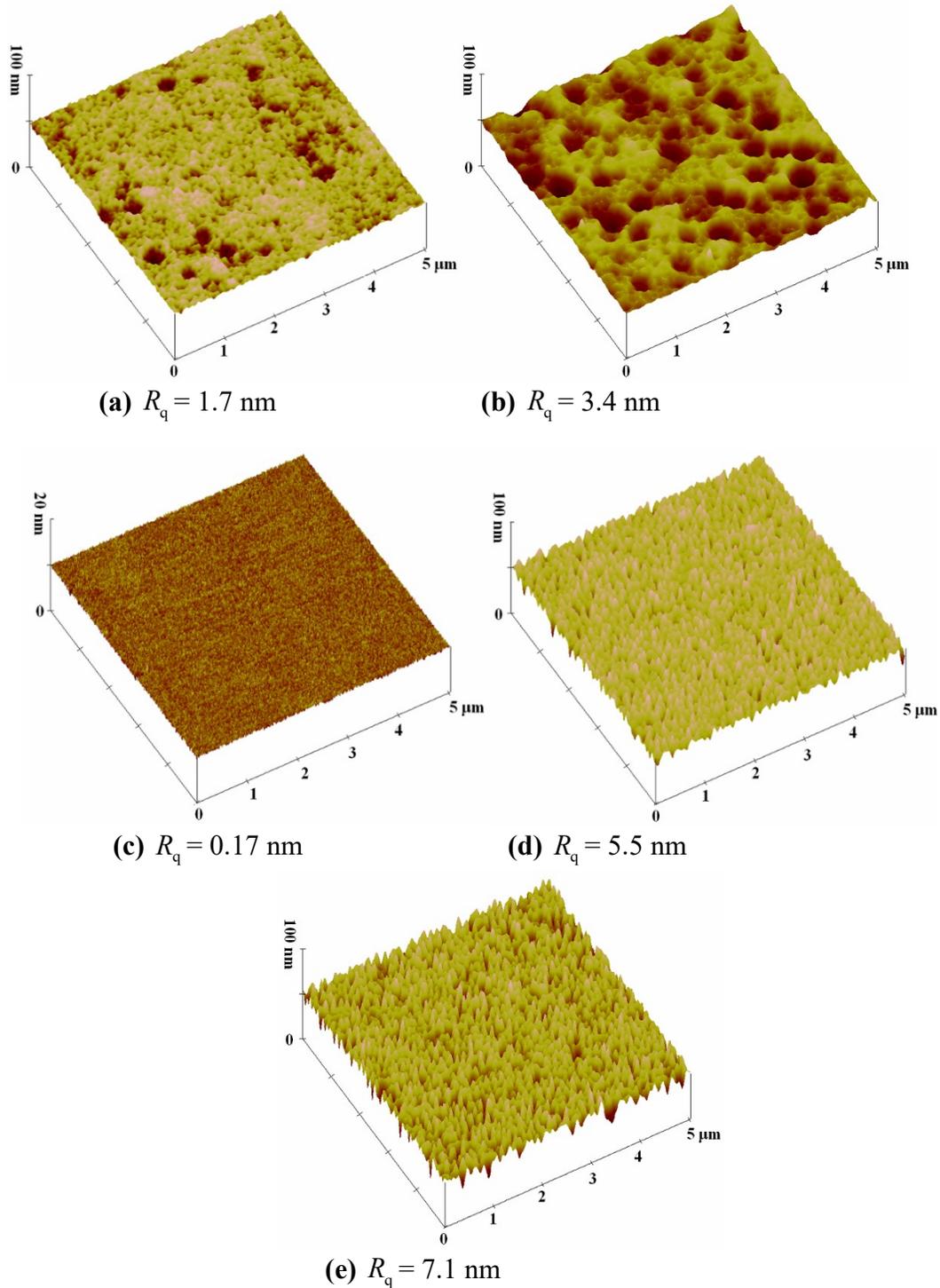


Figure 2. AFM images of bottom surfaces of microcantilevers fabricated by (a) Run *A* (b) Run *B*, (c) silicon wafer, (d) polysilicon film deposited by LPCVD at 585 °C, and (e) polysilicon film deposited by LPCVD at 590 °C.

Table 2. Measured individual and extracted combined surfaces roughness parameters.

Individual surface	Bottom surface of microcantilever		Substrate		
	1. Run <i>A</i>	2. Run <i>B</i>	3. Si	4. Polysilicon (LPCVD 585)	5. Polysilicon (LPCVD 590)
R_q (nm)	1.7	3.4	0.17	5.5	7.1
R (μm)	0.41	0.84	3.0	0.12	0.09
η ($1/\mu\text{m}^2$)	67.9	20.9	111.5	48.0	49.4
σ (nm)	1.5	3.3	0.16	4.2	5.5
S_{sk}	-0.78	-0.76	-0.31	-1.1	-0.73
S_{ku}	4.6	3.9	3.1	5.1	3.3
κ	0.38	-59.9	-0.74	1.6	-0.44
Surface type	IV	I	I	VI	I
Equivalent Rough Surface	Surface Pair				
	<i>a.</i> 1 vs 3	<i>b.</i> 2 vs 3	<i>c.</i> 1 vs 4	<i>d.</i> 1 vs 5	
R_q (nm)	1.7	3.4	5.7	7.3	
R (μm)	0.41	0.81	0.11	0.08	
η ($1/\mu\text{m}^2$)	68.4	22.3	49.0	50.0	
σ_{ab} (nm)	1.5	3.2	4.5	5.7	
S_{sk}	-0.77	-0.77	-0.86	-0.66	
S_{ku}	4.6	3.8	4.8	3.2	
κ	0.38	-5.2	0.49	-0.41	
Surface type	IV	I	IV	I	

R_q = standard deviation of surface heights;

R = average asperity radius;

η = areal density of asperities;

σ = standard deviation of asperity heights;

σ_{ab} = standard deviation of asperity heights of the equivalent rough surface *ab*;

S_{sk} = skewness;

S_{ku} = kurtosis; κ = parameter in the Pearson system of frequency curves.

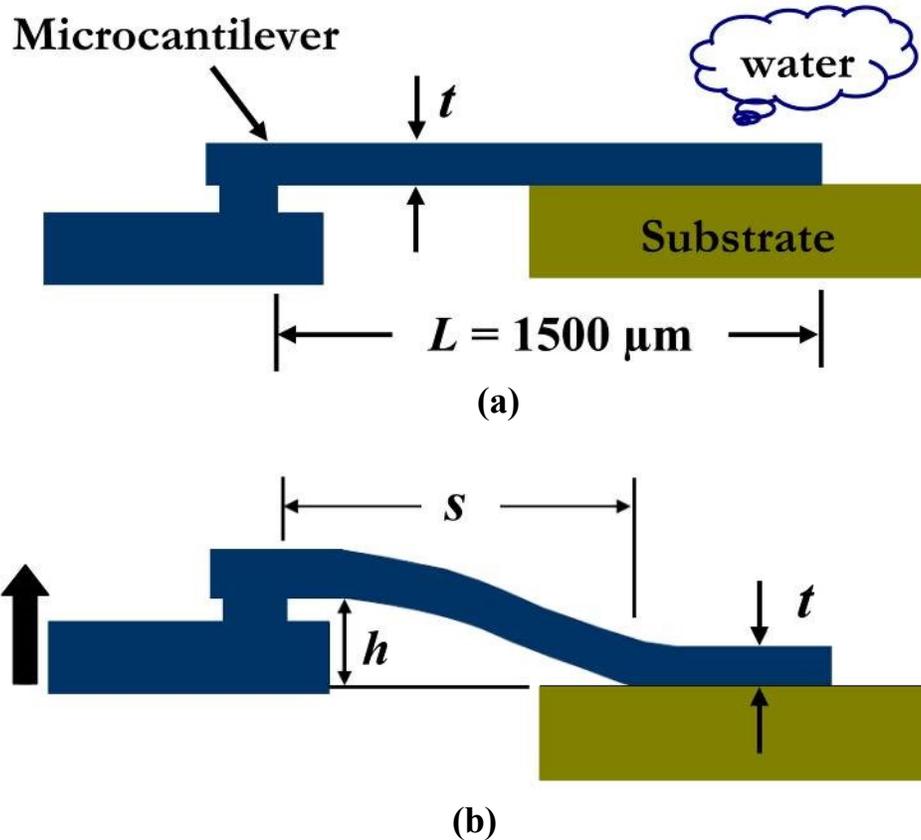


Figure 3. Schematic representation of experimental principle of the microcantilever peel test. (a) Microcantilever is aligned parallel to and tangent with the substrate, (b) as the anchor of the microcantilever is raised, the free-end of the beam sticks to the substrate, showing s-shape deformation.

2. Adhesion energy measurements

2.1. Experimental method

Figure 3 shows a schematic of the experimental principle: a series of microcantilevers are aligned to be parallel to the substrate and then lowered until just making contact with the substrate, as shown in Fig. 3(a). In the presence of capillary condensation, strong capillary forces arise between the bottom beam surface and the substrate. Note that the beams are very compliant as their dimensions are $1500 \mu\text{m}$ long, $2.6 \mu\text{m}$ thick, and $30 \mu\text{m}$ wide. When the fixed end (the anchor) of the beams is raised gradually above the substrate, the free-end of the beams will be sticking to the substrate as shown with the s-shape deflection in Fig. 3(b), in which t is the thickness of the beam, L is the length of the beam, h is the height of the beam above the substrate, and s is the crack length, i.e. the portion

of the beam that is above the substrate (not stuck). de Boer and Michalske [21] modeled the adhered microcantilever substrate system using fracture mechanics. In this model, the crack driving force is given by the strain energy release rate, G

$$G = -\frac{1}{w} \frac{dU}{ds} \quad (2)$$

where w is the width of the cantilever, and U is the strain energy. The crack resistance is the adhesion energy between the microcantilever and the substrate. The adhesion energy, Γ , is determined when the strain energy release rate equals the crack's resistance to propagation

$$G = \Gamma \quad (3)$$

For s-shaped beams, the elastic strain energy is given by [58]

$$U = \frac{6h^2 EI}{s^3} \quad (4)$$

where E is the Young's modulus of the cantilever beam, and I the moment of inertia of the cantilever. If the crack propagates without external work, the strain energy release rate G is given by

$$G = \frac{3}{2} \frac{h^2 Et^3}{s^4} \quad (5)$$

The above equations were derived assuming small strains arising from small beam deflections. Previous work [59] has shown that large deflection theory applies only when the end-deflection exceeds 30% of the beam length. As the end-deflections in the experiments performed in this work were always less than or equal to 10% of the beam length, the use of linear beam theory is justified.

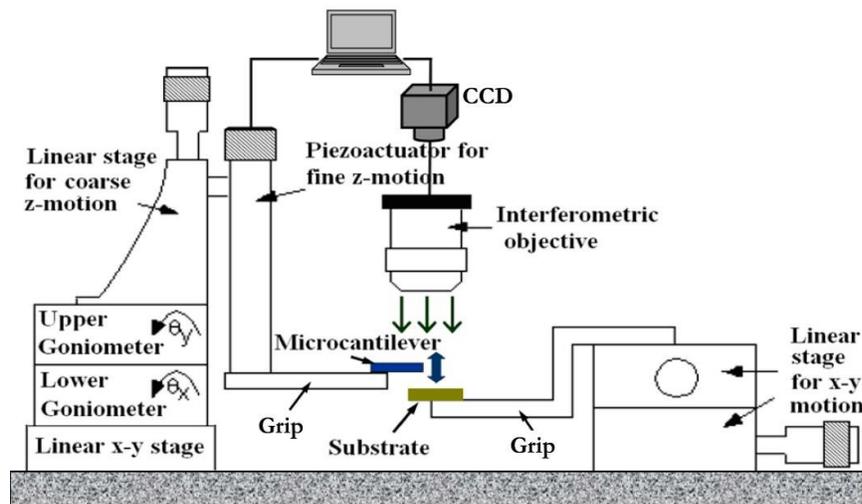
2.2. Experimental setup and procedure

Figure 4 depicts the schematic and actual photographs of the experimental apparatus. The microcantilever beam arrays are attached to a grip that is fixed to a piezoactuator (Physik Instrumente P845.60). The piezoactuator is connected to linear stages that allow for coarse movement in the X-Y directions as well as the

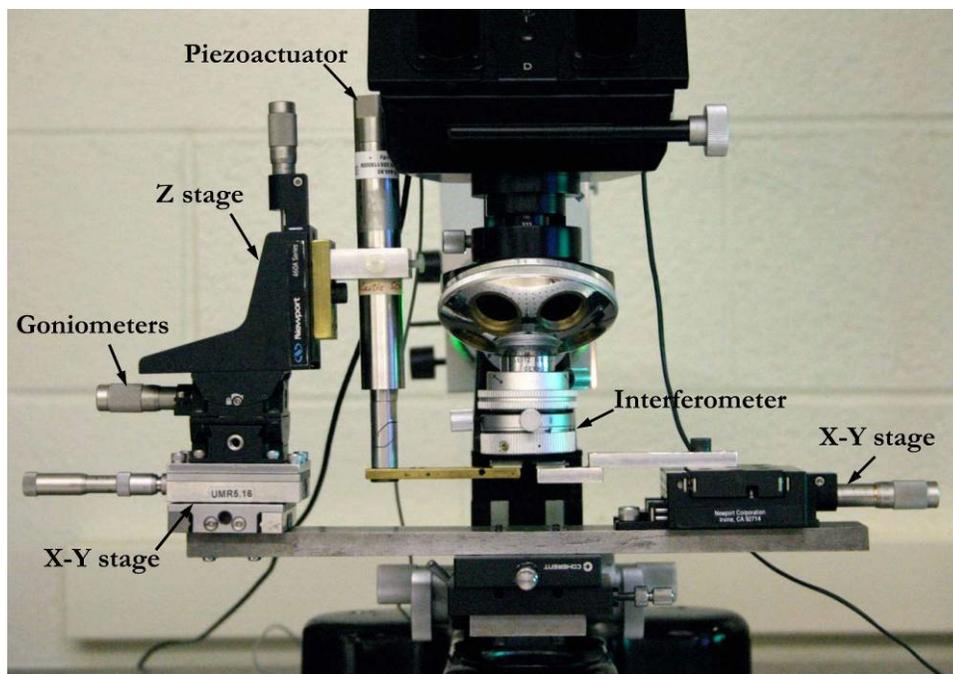
vertical Z direction. A computer connected to the piezoactuator accurately controls the vertical displacement of the microcantilevers with respect to the substrate. The piezoactuator has a vertical displacement range of 90 μm with a resolution of 50 nm. The substrate is supported by another custom grip that is attached to a pair of stacked linear stages, and the X-Y stages are used to position the substrate below the microcantilever arrays. Two goniometers, mounted on the left between the linear Z-stage and the stacked X-Y stage, allow tilt rotation in two directions, which is essential to the proper alignment of the microcantilever beams to the contact substrate. A Michelson interferometer is attached to the optical microscope (Leitz/Leica Camera AG) to generate alternating dark and light fringes, which are captured by a CCD camera (Panasonic GP-KR222). Since environmental humidity affects the contact and adhesion between the beam and substrate, the experimental setup is sealed in a chamber with humidity control and the entire apparatus is placed on a vibration isolation table.

Figure 4(c) shows the humidity control system. The N_2 tank has three outlets, with one of them feeding dry N_2 directly into the plastic mixing box. The other two dry N_2 gas streams flow through the glass flask filled with water and thus pick up humidity. All three N_2 gas streams are mixed in the mixing box and the temperature and humidity of the mixed N_2 are monitored by a digital hydrothermometer. By adjusting the N_2 flow rate for each stream using the flowmeters, the humidity of the mixed N_2 that finally flows into the humidity chamber can be precisely controlled. The in-situ humidity and temperature in the chamber during the beam-peel-test are recorded by another digital hydrothermometer. For an intermediate humidity level, for example 70%, it takes up to one hour to reach a steady state in the humidity chamber.

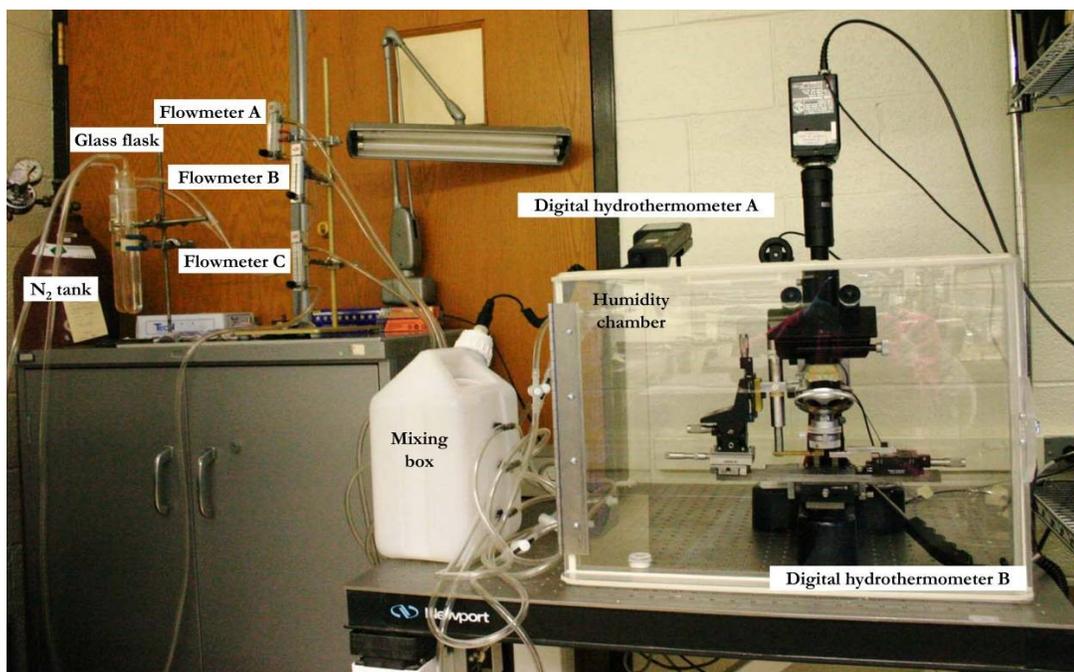
Figure 4



(a)



(b)



(c)

Figure 4. Experimental set-up. (a) Schematic, (b) close-up photograph, (c) photograph of humidity control system.

The proper alignment of the set-up is critical for the testing. As shown in Table 3, where the step-by-step experimental procedure is described, first the substrate is

positioned under the interferometer's objective. Then, the objective is adjusted to be parallel to the substrate until no fringes are observed on the substrate. Then, the substrate is moved away and the microcantilevers are brought into the field of view of the microscope. Goniometers are used to align the beam array parallel to the focal plane (no apparent fringes). At this point, the substrate, the beam array, and the objective focal plane should be parallel to each other. After alignment, the substrate is positioned about $80\ \mu\text{m}$ below the beams using the linear Z stage. The piezoactuator is then used to lower the beam gradually by $100\ \text{nm}$ decrements until the bottom surface of the beam is touching the substrate, which can be identified by the disappearance of the fringes on the microcantilevers, as shown in the actual interferometric image of Fig. 5(a) ($h = 0\ \mu\text{m}$). The scale of the image is given by the two parallel diagonal mark lines ($100\ \mu\text{m}$ marker).

Table 3. Experimental procedure for microcantilever peel experiments.

Step	Experimental Procedure
1	Align substrate parallel to objective
2	Align microcantilevers parallel to objective
3	Bring microcantilevers into contact with the substrate
4	Keep the constant humidity level for at least 12 hours to reach steady-state
5	Raise microcantilevers and measure the crack length when crack propagation occurs

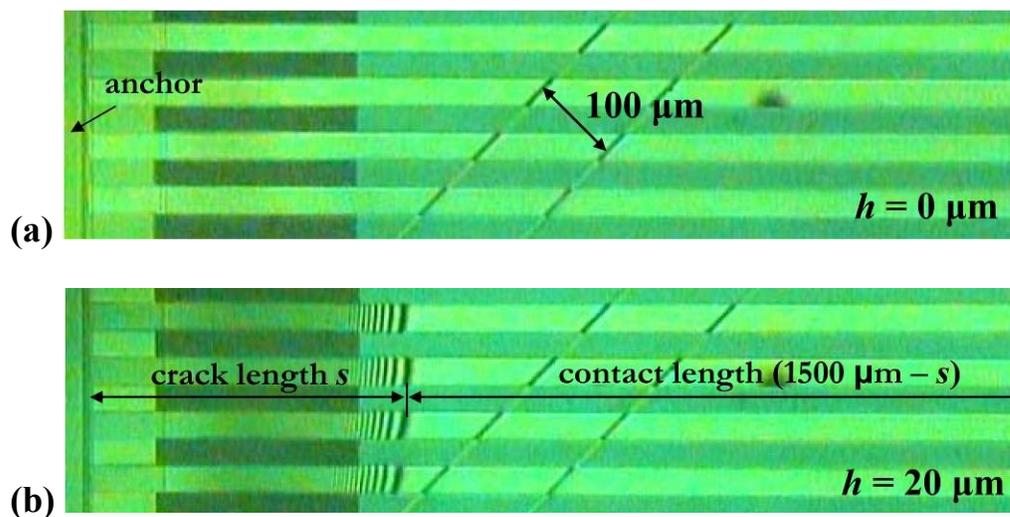


Figure 5. Interferometric images when (a) microcantilevers are contacting with the substrate, $h = 0\ \mu\text{m}$, (b) microcantilevers are $20\ \mu\text{m}$ higher than the substrate, $h = 20\ \mu\text{m}$ (the contact length is equal to the length of the microcantilever minus the crack length s).

Typically, at a specified RH level, the beam array will take twelve or more hours of exposure in the chamber to reach equilibrium. Once the equilibrium is reached, the piezoactuator is used to raise the anchor of the beam in 100 nm increments. Due to the high compliance of the beam, the adhesion force between the bottom surface and the substrate will keep the free end of the beam into contact with the substrate over a considerable length, as shown in the schematic of Fig. 3(b). The Michelson interferometric attachment illuminates the samples with green (548 nm) light and provides alternating dark and light fringes that account for an out-of-plane deflection of 274 nm per fringe. A CCD camera is used to capture the interferometric images. The captured fringe density is directly related to the deflection profile of the microcantilever and the crack length can be readily determined from the images, as shown in Fig. 5(b). Equations (3) and (5) are then used to calculate the corresponding adhesion energy between the microcantilevers and the substrate. The microcantilevers are continuously raised until the beam breaks free from the substrate.

In the experimental set-up of this work, since the substrate and the microcantilever arrays are two separate parts, the microcantilevers and the substrate can be treated separately and different substrates can be tested and compared using the same microcantilevers at different humidity levels. Even though in this work the focus was on roughness effects, one could readily test substrates with coatings or other surface modifications.

2.3. Experimental results

Peel tests were performed on four different surface pairs at controlled humidity levels ranging from about 40% to 95%. Surface pair *a* is microcantilever *A* on silicon wafer substrate, pair *b* is microcantilever *B* on silicon wafer, pair *c* is microcantilever *A* on LPCVD deposited polysilicon substrate at 585 °C, and pair *d* is microcantilever *A* on polysilicon deposited at 590 °C. From geometrical considerations, the static contact between two rough surfaces can be modeled as the contact between an equivalent rough surface and a rigid flat surface [28]. The surface roughness parameters of the equivalent rough surface for each contact pair are listed in the bottom part of Table 2, and were extracted using the method described in [53]. It is seen that the surface roughness of the equivalent surfaces is dominated by the rougher surface of the contacting pair. For example, the roughness parameters of surface pair *b* are very close to those of microcantilever *B*.

Figure 6 depicts interferometric images taken for surface pair *a* (microcantilever *A* on silicon wafer substrate) at humidity levels of 90%, 80%, 70% and 45% when the microcantilever was positioned 20 μm above the substrate. It can be seen that as the humidity decreases, the adhesion energy decreases, which

results in the increase of the measured crack length. For example, when the humidity is 90%, the crack length is only 366 μm , which corresponds to an adhesion energy of 99.32 mJ/m^2 . When the humidity decreases to 45%, the crack length increases to 732 μm and the corresponding adhesion energy is only 6.21 mJ/m^2 .

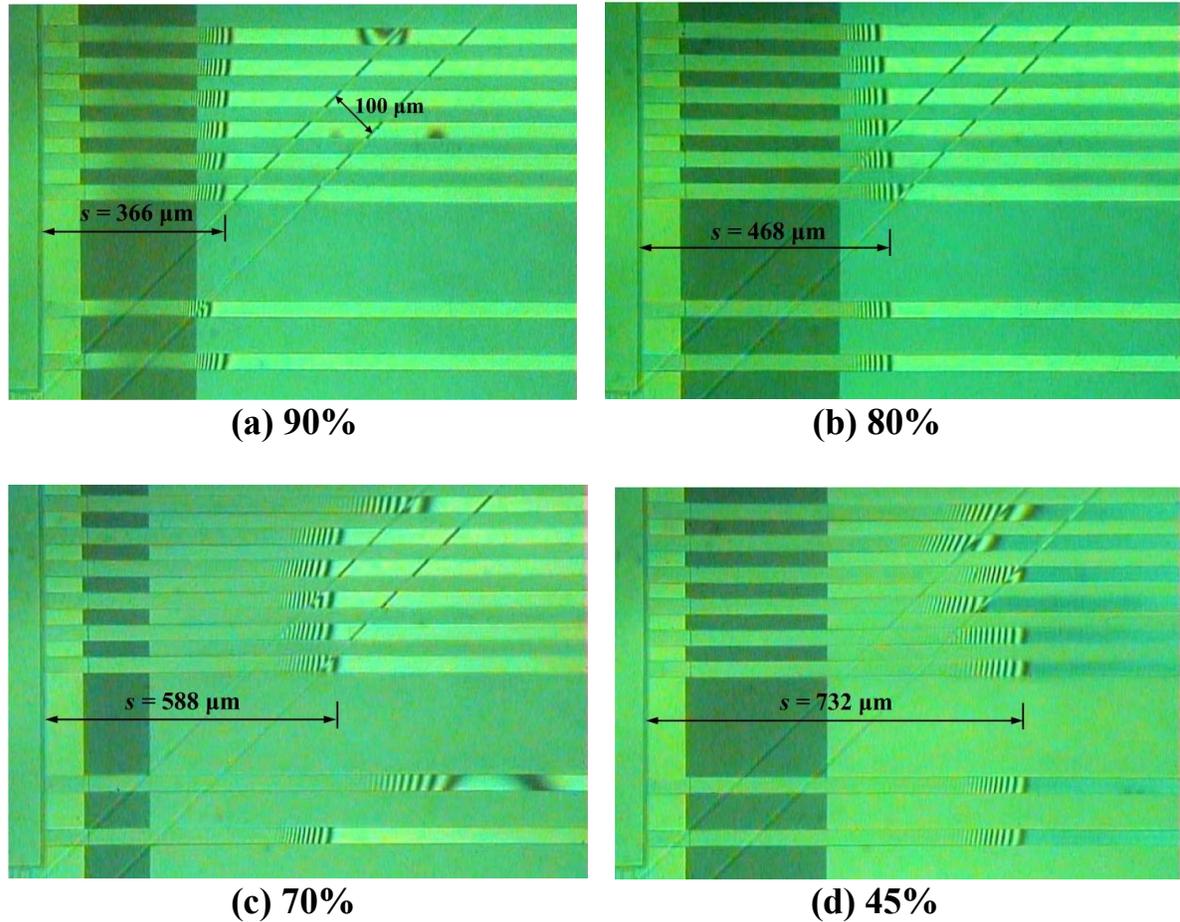


Figure 6. Interferometric images when microcantilevers are 20 μm higher than the substrate for surface pair *a* at humidity levels of (a) 90%, (b) 80%, (c) 70%, and (d) 45% (*s* is the crack length).

Figure 7 depicts the average adhesion energy values measured on surface pairs *a*, *b*, and *c* for *RH* levels ranging from 44% to 92%. The error bars represent plus and minus one standard deviation of the measured adhesion energies when the beams were at different heights above the substrate. It is seen that for all surface roughness pairs, the adhesion energy between the bottom microcantilever surface and the substrate increases with increasing *RH* value. The smoother the surfaces are (pairs *a* and *b*), the higher the measured adhesion energy and also the lower the

dependence on RH . For example, when $RH = 80\%$, the adhesion energy for surface pair c ($\sigma_{ab} = 5.7$ nm) is only 0.47 ± 0.09 mJ/m², while the adhesion energy of surface pair a ($\sigma_{ab} = 1.7$ nm) is much higher at 37.36 ± 5.59 mJ/m². For surface pair d , due to its relatively high roughness, it was difficult to get s-shaped beams (i.e. pair d exhibited very low adhesion energy values), and no data is shown in this work.

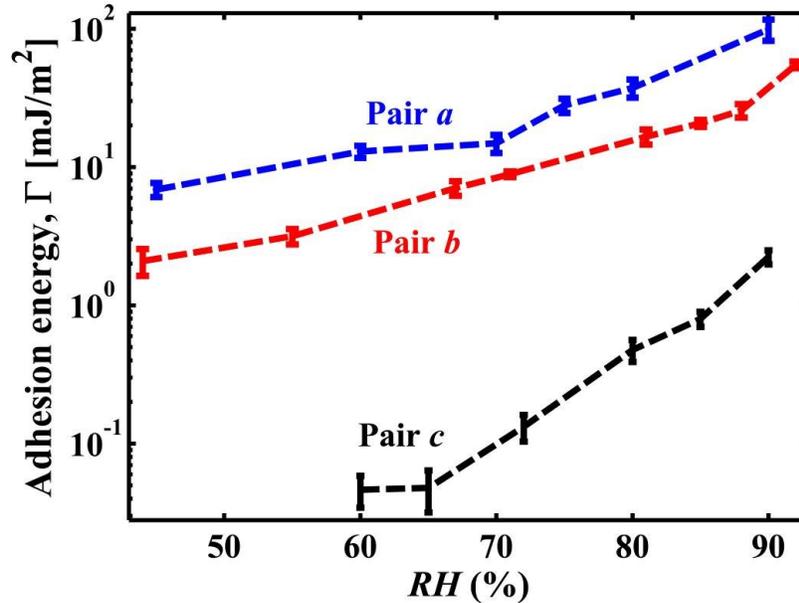


Figure 7. Average adhesion energy measured on surface pairs a ($\sigma_{ab} = 1.5$ nm), b ($\sigma_{ab} = 3.2$ nm), and c ($\sigma_{ab} = 4.5$ nm) for RH ranging from 44% to 92% (parameters shown in Table 2, σ_{ab} = standard deviation of asperity heights of the equivalent rough surface).

3. Modeling

The static contact between two rough surfaces can be modeled as the contact between an equivalent rough surface and a flat rigid surface. In the presence of capillary condensation, menisci form around the asperities due to surface tension effects, as shown schematically in Fig. 8. By integrating the contribution of each asperity, the adhesion force and the contact load between the two rough surfaces can be determined. To calculate the single-asperity capillary force, the surface asperities can be treated as a deformable sphere of radius R contacting with a flat surface [28].

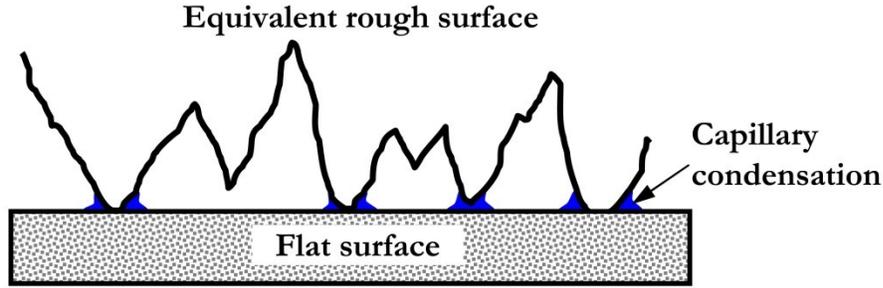


Figure 8. Schematic of a rough surface in contact with a rigid flat surface in the presence of capillary condensation.

3.1. Simplified single-asperity meniscus model

In most surface meniscus models that appear in the literature, the classical single-asperity equations presented by Israelachvili [36] are widely used, which do not take into account the spherical deformation caused by the adhesive force and repulsive contact force, i.e.

$$f_{m_simplified} = \begin{cases} 4\pi R\gamma_{lv} \cos\theta \left(1 + \frac{\omega}{h_{lk}}\right) & \omega < 0 \\ 4\pi R\gamma_{lv} \cos\theta & \omega \geq 0 \end{cases}, \quad (6)$$

where γ_{lv} is the surface tension of the liquid-air interface, θ is the contact angle, ω is the interference, h_{lk} is the liquid film thickness estimated by the Laplace-Kelvin equation,

$$h_{lk} = \frac{2\gamma_{lv}V \cos\theta}{R_g T \ln(RH)}. \quad (7)$$

V is the liquid molar volume, R_g is the gas constant, and T is the absolute temperature. For water at room temperature, $\gamma_{lv}V/(R_g T) = 0.54$ nm.

3.2. Improved EMD-based single-asperity meniscus model

Maugis [60] and Johnson [61] pointed out that the Laplace pressure acting in a meniscus area is a “perfect” example of the Dugdale model. Xue and Polycarpou [44] developed an improved EMD-based single-asperity meniscus model, in which

the EMD theory is adopted to analyze the single-asperity adhesion behavior in the presence of capillary condensation. Instead of the approximation to the Lennard-Jones potential, the Dugdale stress is taken as the Laplace pressure acting on the wetted area. This improved EMD-based, single-asperity meniscus model considering both asperity deformation and solid surface interaction is coupled with the Pearson system of frequency curves to predict the contact and adhesion between two rough surfaces at different relative humidity levels.

3.2.1. Non-contacting asperity ($\omega < 0$)

For a non-contacting sphere ($\omega < 0$) as shown in Fig. 9(a), the pressure inside the liquid is lower than that outside the liquid and the pressure difference Δp acting on the circular wetted area of radius c is given by the Laplace equation

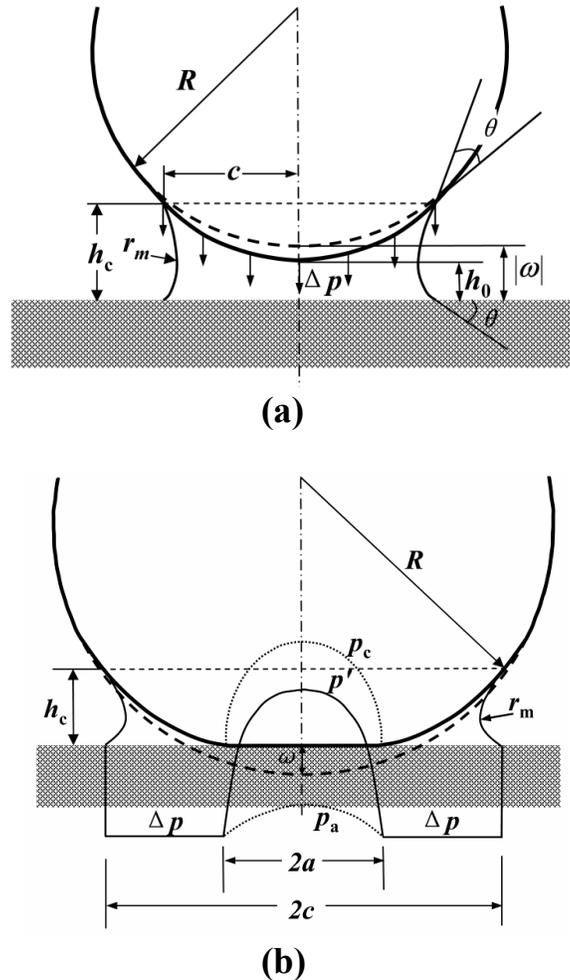


Figure 9. Schematic of a deformable elastic sphere and a rigid flat surface. (a) non-contacting $\omega < 0$, (b) contacting $\omega > 0$.

$$\Delta p = \frac{\gamma_{lv}}{r_m} = \frac{2\gamma_{lv} \cos \theta}{h_{lk}}, \quad (8)$$

where r_m is the meniscus radius. Assuming the displacement is negative in the departing direction, the radius of the projected meniscus area, c , can be obtained by solving the equation below:

$$\frac{c^2}{2R} + \frac{2\Delta p c (\pi - 2)}{\pi E_r} + h_0 - h_{lk} = 0, \quad (9)$$

where h_0 is the normal separation at the center. E_r is the reduced elastic modulus defined by $1/E_r = (1 - \nu_1^2)/E_1 + (1 - \nu_2^2)/E_2$, where E_1 , E_2 , ν_1 , and ν_2 are Young's moduli and Poisson's ratios of the sphere and the flat surface, respectively. The Laplace pressure acts on the wetted area πc^2 , pulling the sphere towards the substrate with an attractive force

$$f_m = \pi \Delta p c^2. \quad (10)$$

The thick dashed line in Fig. 9(a) shows the original spherical profile and the solid line represents the actual deformed shape. It can be seen that the sphere deforms towards the surface due to the attractive capillary force. The approach of the sphere is given by

$$\omega = h_0 + \frac{2c\Delta p}{E_r}. \quad (11)$$

3.2.2. Contacting asperity ($\omega > 0$)

Figure 9(b) shows a sphere contacting with a rigid surface ($\omega > 0$), assuming the liquid is expelled from the contact region, allowing solid-to-solid contact. The constant Dugdale stress Δp (which is taken as the Laplace pressure) acts on the wetted area $a < r < c$ up to a maximum separation h_{lk} beyond which it falls to zero. The induced Hertzian compressive pressure $p_c(r)$ and the tensile adhesion pressure $p_a(r)$ act inside the contact zone [60] and are given by:

$$\text{for } r < a, \begin{cases} p_c(r) = \frac{2E_r \sqrt{a^2 - r^2}}{\pi R} \\ p_a(r) = \frac{2}{\pi} \Delta p \tan^{-1} \left(\frac{c^2 - a^2}{a^2 - r^2} \right)^{1/2}. \end{cases} \quad (12)$$

When $r = a$, the stress is equal to Δp , which ensures stress continuity at the crack tip.

In the MD model [60], the “work of adhesion” w , defined as the external work done to separate a unit area of the adhering surfaces, is determined by the Lennard-Jones surface potential; whereas, in the current model, the effective work of adhesion is given by the value of the meniscus adhesion assuming no solid-solid interaction across the liquid (outside the contact region),

$$w = h_c \Delta p = 2r_m \cos \theta \frac{\gamma_{lv}}{r_m} = 2\gamma_{lv} \cos \theta. \quad (13)$$

The adhesion parameter, λ , can then be expressed as:

$$\lambda = \Delta p \left(\frac{9R}{2\pi w E_r^2} \right)^{1/3} = \frac{1}{r_m} \left(\frac{9R \gamma_{lv}^2}{4\pi \cos \theta E_r^2} \right)^{1/3}. \quad (14)$$

Letting $m = c/a$, the MD theory gives [60]

$$\frac{\lambda a^{*2}}{2} \left[\sqrt{m^2 - 1} + (m^2 - 2) \tan^{-1} \sqrt{m^2 - 1} \right] + \frac{4\lambda^2 a^*}{3} \left[\sqrt{m^2 - 1} \tan^{-1} \sqrt{m^2 - 1} - m + 1 \right] = 1, \quad (15)$$

where a^* is the normalized contact radius

$$a^* = a \left(\frac{4E_r}{3\pi w R^2} \right)^{1/3}. \quad (16)$$

One could also define a normalized asperity approach as:

$$\omega^* = \omega \left(\frac{16E_r^2}{9\pi^2 w^2 R} \right)^{1/3} = a^{*2} - \frac{4}{3} a^* \lambda \sqrt{m^2 - 1}. \quad (17)$$

The adhesion force due to the Laplace pressure outside the area of contact is obtained as

$$f_m = \Delta p \pi (c^2 - a^2) \text{ or } \frac{f_m}{\pi w R} = \frac{\pi}{2} \lambda a^{*2} (m^2 - 1). \quad (18)$$

For small λ values (as for example for the case of high humidity levels and small stiff spheres) the MD model becomes a ‘‘DMT-like’’ contact [62] where capillary condensation dominates the adhesion force. At low humidity levels and large compliant spheres, λ is large and the model predicts a ‘‘JKR-type’’ contact [63]. Note that Fogden and White [41] defined the transition from JKR-type to DMT-type contact using a similar adhesion parameter, which is related to the inverse of the parameter λ .

In the case of large λ values, the contribution of the adhesion force inside the contact region ($r < a$) is also important and needs to be considered. Integrating the adhesion stress gives the adhesion force inside the contact region:

$$f_{s-s} = \int_0^a 2\pi r p_a(r) dr = 4\Delta p \int_0^a r \tan^{-1} \sqrt{\frac{c^2 - a^2}{a^2 - r^2}} dr. \quad (19)$$

The total adhesion force is obtained as

$$f_a = f_m + f_{s-s} = 2\Delta p a^2 \left(m^2 \tan^{-1} \sqrt{m^2 - 1} + \sqrt{m^2 - 1} \right). \quad (20)$$

Integrating the compressive stress gives the contact force:

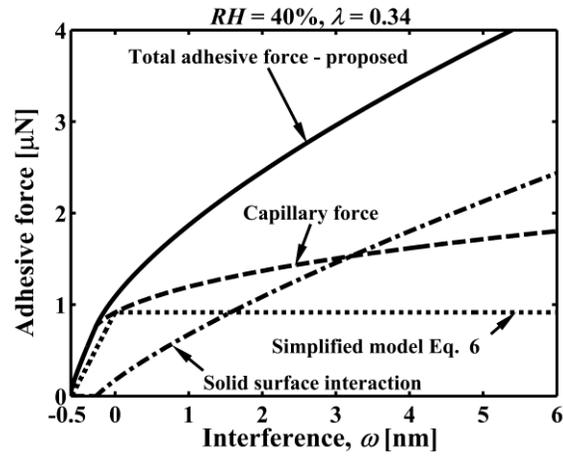
$$p = \int_0^a 2\pi r p_c(r) dr = \frac{4E_r a^3}{3R}. \quad (21)$$

Note that in this study, the effect of solid surface energy on the adhesion force is ignored and a constant γ_{lv} due to water is assumed.

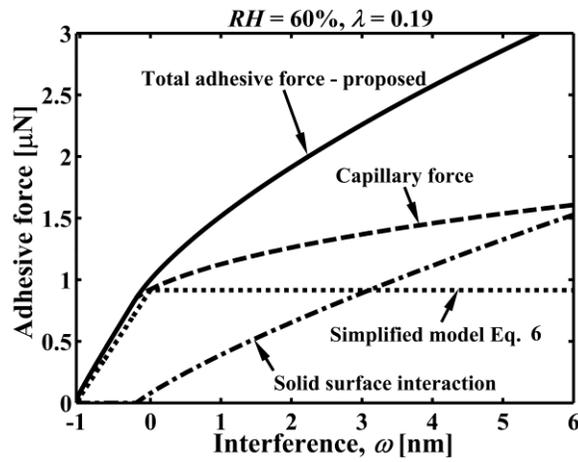
3.2.3. Effect of adhesion parameter λ

Figure 10 depicts adhesive forces as a function of interference at varying humidity levels of 40%, 60%, and 85% for a 2 μm sphere radius on a rigid flat (the reduced modulus is 43 GPa and the contact angle is 60°). Note that in this case the liquid thickness is estimated by Eq. (7). The simplified model predicts a constant adhesive force for contacting spheres ($\omega > 0$) independent of the relative humidity and sphere interference as it ignores the spherical deformation and the contribution from inside the contact zone, while the EMD-based meniscus model shows an increase of the adhesive force with the spherical interference. At low humidity levels, λ is larger and the EMD-based model shows a JKR-like contact. As shown in Fig. 10(a), for non-contacting conditions ($\omega < 0$), due to the spherical deformation, the induced capillary force is higher than that predicted by Eq. (6) and the sphere will touch the solid surface when they are brought into close proximity. When the sphere makes contact with the substrate ($\omega > 0$), the adhesive force due to solid surface interaction inside the contact region arises and increases with increasing interference (contact area). Meanwhile, the capillary force increases with interference due to the spherical deformation. Thus, the total adhesive force (capillary force plus solid surface force) is higher than that predicted by the simplified model (Eq. (6)) in which case the maximum force is reached at the onset of contact ($\omega = 0$) and a constant value is assumed thereafter (for the contacting asperities ($\omega > 0$)).

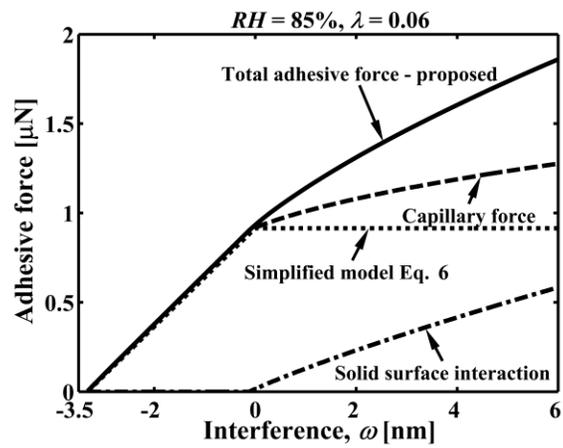
As humidity increases, the λ value becomes smaller and the EMD-based model shows more of a DMT-like contact behavior. As shown in Fig. 10(c), for the non-contacting sphere ($\omega < 0$) the adhesion force difference caused by the spherical deformation is very minor, and for the contacting case, the contribution of the adhesion force from inside the contact zone is much smaller than that at the low humidity level. It is also seen in Fig. 10, that for contacting asperities ($\omega > 0$), lower relative humidity results in higher meniscus forces at the same interference. This is because the Laplace pressure acting on the sphere surface by the condensed liquid is lower at the higher humidity, and the increase of the wetted area is not as fast as that at the lower humidity.



(a)



(b)



(c)

Figure 10. Adhesive forces versus interference for a sphere on a rigid flat surface ($R = 2 \mu\text{m}$, $\theta = 60^\circ$) at varying relative water vapor pressures of: (a) 0.4, (b) 0.6, and (c) 0.85.

3.3. EMD-Pearson meniscus surface model

By incorporating the single-asperity meniscus model in the GW statistical model [28] the adhesion force and adhesion energy between two rough surfaces can be obtained. Note that in this work we use a statistical surface model, which is justified based on the roughness measurements presented above. The basic single-asperity meniscus model presented above is generally applicable and could also be used with other surface roughness models, such as the fractal model [64]. For a rough surface with multi-asperity contact interface, the net adhesion force arises from the individual asperity contributions. According to the GW model, the surface topography of the equivalent rough surface can be characterized by three parameters, namely σ , η , R , and that the asperity heights z follow a probability density function. In most meniscus surface models, a Gaussian (symmetric) distribution is usually assumed, given by

$$\phi(z) = \frac{1}{\sigma_{ab} \sqrt{2\pi}} e^{-0.5(z/\sigma)^2}. \quad (22)$$

However, as clearly seen in Table 2, MEMS surfaces do not typically exhibit a Gaussian distribution. To describe an asymmetric non-Gaussian surface, two statistical parameters, the skewness and kurtosis, representing the asymmetry and flatness of the asperity height distribution are used. Several different types of non-Gaussian height distributions have been proposed including the Weibull distribution [57] and the Pearson system of frequency curves [65]. Although the Weibull distribution is mathematically attractive and can represent asymmetric distributions, the effects of kurtosis and skewness are coupled and cannot be studied independently. By generating the exact probability density function of the asperity heights using the measured four moments, i.e., mean, standard deviation, skewness, and kurtosis, the Pearson system of frequency curves [65] can be used to study the effect of skewness and kurtosis independently.

Table 4 lists the probability density functions of different types of Pearson surfaces [65]. The Pearson parameter, κ , is used to determine the type of the surface distribution:

$$\kappa = \frac{S_{sk}^2 (S_{ku} + 3)^2}{4(2S_{ku} - 3S_{sk}^2 - 6)(4S_{ku} - 3S_{sk}^2)}. \quad (23)$$

Table 4. Probability density functions using the Pearson system of frequency curves.

Type		Criterion	Equations with origin at the mean
Main types	I	$-\infty < \kappa < 0$	$\phi(z) = y_0 \left(1 + \frac{z}{A_1}\right)^{m_1} \left(1 - \frac{z}{A_2}\right)^{m_2} \quad (-A_1 < z < A_2)$
	IV	$0 < \kappa < 1$	$\phi(z) = y_0 \left(1 + \left(\frac{z-v}{a} - \frac{v}{r}\right)^2\right)^{-m} \exp\left(-v \tan^{-1}\left(\frac{z-v}{a} - \frac{v}{r}\right)\right)$
	VI	$1 < \kappa < \infty$	$\phi(z) = y_0 \left(1 + \frac{z}{A_1}\right)^{-q_1} \left(1 + \frac{z}{A_2}\right)^{q_2}$
Transition types	Normal	$\kappa = 0, S = 0, K = 3$	$\phi(z) = y_0 \exp\left(-\frac{z^2}{2\sigma^2}\right)$
	II	$\kappa = 0, S = 0, K < 3$	$\phi(z) = y_0 \left(1 - \frac{z^2}{a^2}\right)^m \quad (-a < z < a)$
	VII	$\kappa = 0, S = 0, K > 3$	$\phi(z) = y_0 \left(1 + \frac{z^2}{a^2}\right)^{-m}$

Equations for curve fitting parameters ($y_0, A, A_1, A_2, m, m_1, m_2, a, v, r, q_1, q_2$) can be found in [65].

Integrating the capillary force at each surface asperity, the meniscus force between two rough surfaces can be obtained as

$$F_m = A_n \eta \int_{d-h_k}^{\infty} f_m \phi(z) dz, \tag{24}$$

where A_n is the nominal area of contact at a rough interface and $\phi(z)$ is the Pearson distribution given in Table 4.

To calculate the total adhesion force at low and intermediate *RH* levels, in addition to the contribution of capillary condensation, the solid surface interaction inside the contact zone is also included as follows

$$F_{s-s} = A_n \eta \int_{d-h_k}^{\infty} f_{s-s} \phi(z) dz, \tag{25}$$

and the total adhesion force between two contacting rough surfaces is obtained by

$$F_a = F_m + F_{s-s}. \quad (26)$$

At equilibrium, an external force F separating the surfaces must be applied

$$F = P - F_a, \quad (27)$$

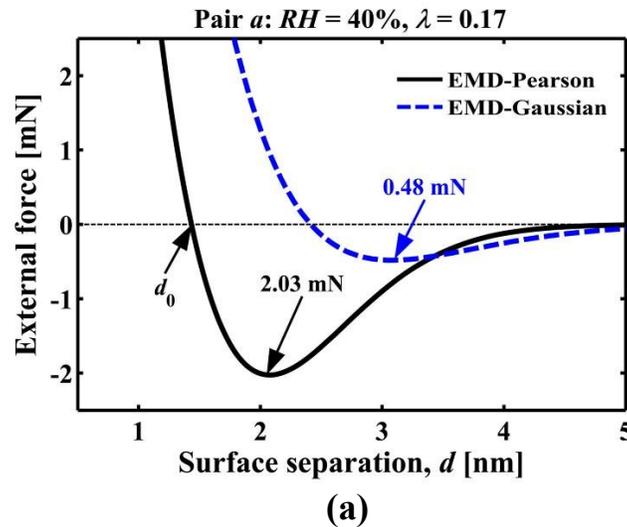
where P is the asperity contact deformation force calculated by

$$P = A_n \eta \int_{d-h_k}^{\infty} p \phi(z) dz, \quad (28)$$

and the pull-off force is defined as the minimum value of the external force.

Figure 11 depicts the external force as a function of surface separation for surface pair a at RH of 40%, 70%, and 85%. It is seen that when the humidity increases from 40% to 85%, the corresponding pull-off force also increases. In addition, at the high humidity level, the force estimated using a Gaussian distribution (13.2 mN) is close to that predicted by the Pearson distribution (15.45 mN). As the humidity decreases, the Gaussian distribution greatly underestimates the pull-off force. For example, at $RH = 40\%$, the Gaussian distribution predicts 0.48 mN at $d = 3.1$ nm, while the Pearson distribution predicts 2.03 mN at $d = 2.05$ nm.

Figure 11



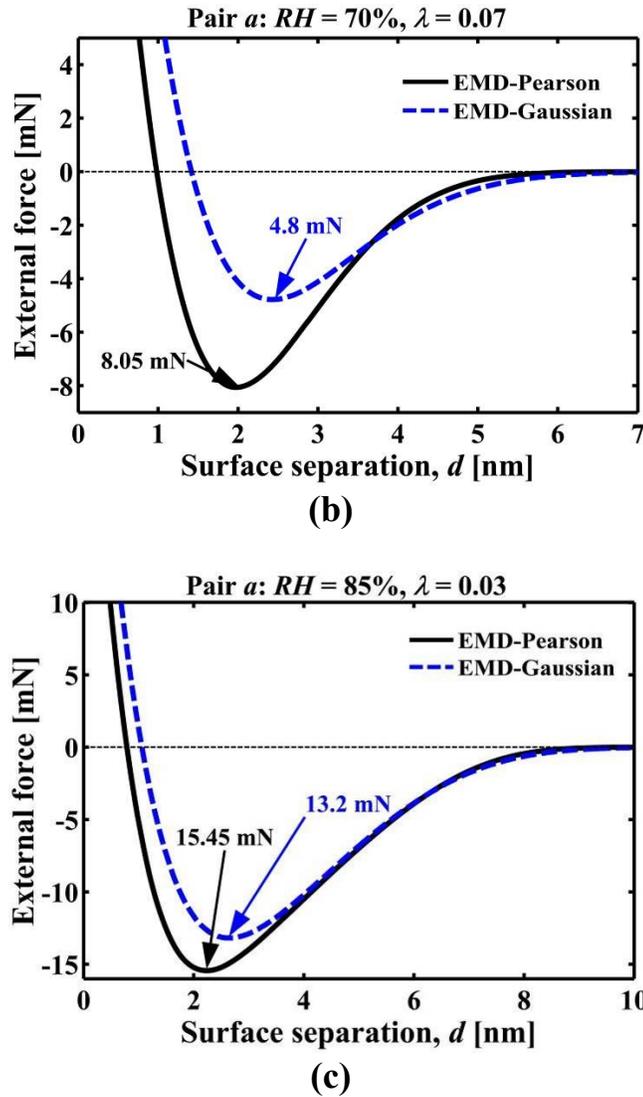


Figure 11. External force versus surface separation for surface pair *a* at humidity levels of (a) 40%, (b) 70%, and (c) 85%. d_0 is the equilibrium point at zero applied load, λ is the adhesion parameter given by Eq. (14).

To examine the effects of using an inaccurate asperity distribution, the histogram of the surface heights for the bottom surface of microcantilever *A* was extracted using the measured AFM data as shown in Fig. 12. Note that in surface pair *a*, since the microcantilevers are much rougher than the silicon substrate, the adhesion force is primarily determined by the bottom surface of the microcantilevers. The fitted Pearson and Gaussian distributions are also shown in the figure. It is seen that the distribution generated using the Pearson system matches the actual surface better than the Gaussian distribution. This is because the Pearson system uses more fitting parameters which capture both the asymmetry

and flatness of the surface heights, while the symmetric Gaussian distribution simply fits the data with $S_{sk} = 0$ and $S_{ku} = 3$.

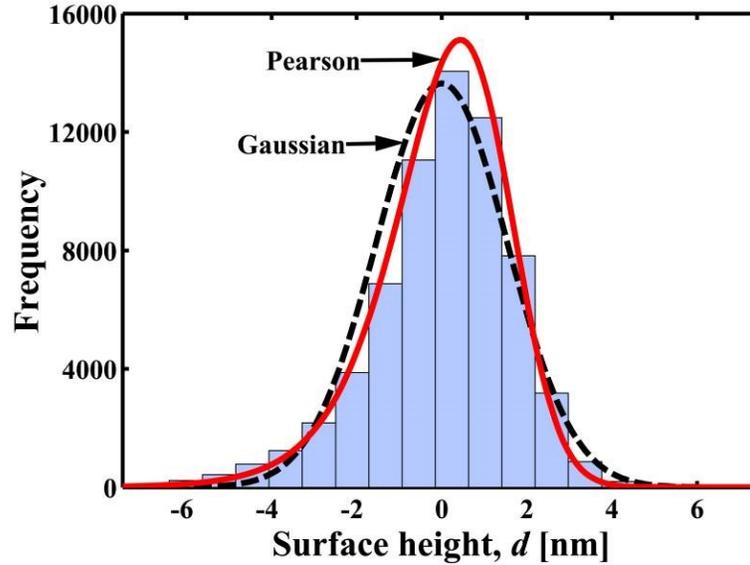


Figure 12. Histograms with Gaussian and Pearson distribution fits of the surface roughness of bottom surface of microcantilever *A*.

At low humidity levels, the thickness of the condensed liquid is small and only the highest asperities on the surface affect the adhesion behavior. The Gaussian distribution predicts more asperities contact with the substrate (compared to the actual surface) and thus overestimates the surface separation at contact and underestimates the pull-off force. At high *RH*, the condensed liquid is thicker and the contact and adhesion behavior of the surface is determined by the majority of asperities, thus the difference caused by the distribution is not as significant as at the low humidity, as shown in Fig. 11.

3.4. Comparison between model and experiments

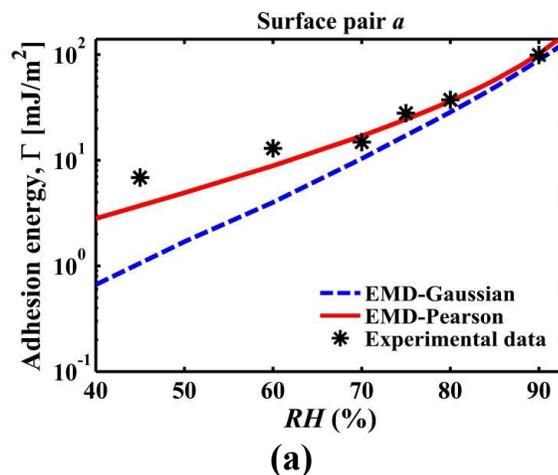
To calculate the adhesion energy per unit area (mJ/m^2) Γ from the modeled predictions, the force-displacement curves shown in Fig. 11 were integrated from the equilibrium point d_0 (at zero applied load) to infinity,

$$\Gamma = \frac{1}{A_n} \int_{d_0}^{\infty} F(d) dd, \quad (29)$$

where F is the total external force given by Eq. (27).

Figure 13 shows measured and modeled adhesion energy Γ as a function of RH for three of the surface pairs. The solid line represents the EMD-based Pearson meniscus model and the dashed line is the EMD-based-Gaussian meniscus model. It is seen that by generating the exact asymmetric surface topography, the Pearson-based model matches the experimental data very well from low humidity to high humidity for surface pairs a and b , while the Gaussian-based model greatly underestimates the adhesion energy at low RH . For example, for surface pair b (shown in Fig. 13(b)), when $RH = 44\%$, the measured energy is 2.09 mJ/m^2 and the predicted energy using the correct Pearson model is 0.93 mJ/m^2 while the Gaussian-based model gives an unrealistic low value of 0.0015 mJ/m^2 . Also, it is seen that both models work better for smoother surfaces and higher humidity levels. As the surface roughness increases, the difference between model and experimental results increases. This is possibly because in the improved EMD single-asperity meniscus model, the solid surface interaction within the contact zone was calculated by integrating the induced adhesion stress within the contact region due to the Laplace pressure instead of the surface energy of the solid surface [44]. Also note that in this study, we assume a constant γ_{lv} due to water. In fact, at very low humidity levels, the surface tension of the liquid could be altered by the solid surface, as shown for example in the case of molecularly thin perfluoropolyether lubricants in magnetic storage [66]. However, no similar analysis exists for water vapor.

Figure 13



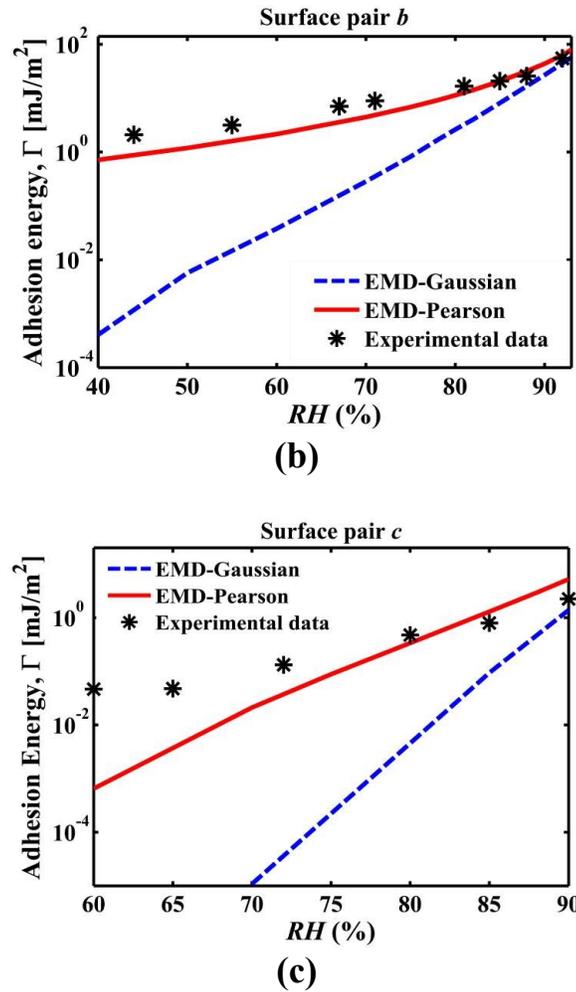


Figure 13. Comparison between measured and modeled adhesion energies Γ versus RH for (a) surface pair *a*, (b) surface pair *b*, (c) surface pair *c*.

The good agreement between the experimental and model results for low roughness and high humidity conditions demonstrates that statistical representation of the surfaces enables accurate predictions of adhesion energy for these conditions. Since the proposed model is based on the generalized EMD model, which is valid for a wide range of adhesion parameter values, it is applicable to other material systems, including soft materials (large adhesion parameter values). Through incorporating the effect of solid surface energy and also considering the plastic asperity deformation, the applicability of the model would likely be extended to higher surface roughness at lower humidity.

4. Summary

MEMS design and fabrication have significantly improved, yet adhesion and contact are major reliability concerns and thus an obstacle to widespread use of such miniature devices. In this chapter, MEMS-scale beam-peel-test experimental set-up was developed and microcantilever beam arrays fabricated by different runs were used to measure the adhesion energy of substrates with different roughness values at controlled humidity levels from 40% to 92%. It was found that the adhesion energy and pull-off force increased with increasing RH and decreasing surface roughness. The adhesion behavior was determined by the rougher surface of the contact pair. A EMD-based meniscus single-asperity adhesive model was coupled with the Pearson system of frequency curves to develop an improved asymmetrical surface meniscus model, which considers the asperity deformation and the adhesion contribution of surface interaction. The model works well from low humidity to high humidity range. Both the beam bottom surface and the substrates were scanned using AFM before the experiments and the extracted surface roughness parameters were used to generate the exact surface topography and input into the model for simulation. The model and the experimental data were favorably compared. The Pearson-based meniscus surface model matches with the experimental data very well from low RH to high RH for both surface pairs a and b , while the Gaussian-based model greatly underestimates the adhesion energy at the low RH . The proposed and validated continuum-based adhesive contact model is only a function of the surface roughness and material properties of the interacting surfaces, and does not include empirical coefficients. Moreover, the model is valid for a wide range of adhesion parameter values, covering the practical range of application of MEMS and other miniature devices.

Acknowledgements

This work was supported by the National Science Foundation under grant number CAREER CMS-0239232. The authors acknowledge Dr. Leslie M. Phinney, Sandia National Laboratories, for providing the MEMS samples and valuable discussions. The authors acknowledge Professor Thomas J. Mackin, California Polytechnic State University, for initial discussions in regards to the experimental set-up. AFM measurements were conducted in the Center for Microanalysis of Materials, University of Illinois, which is partially supported by the U.S. Department of Energy under grant DEFG02-91-ER45439.

References

1. Bustillo, J.M., Howe, R.T., and Muller, R.S. 1998. *Proc. IEEE*, **86**, 1552.
2. Maboudian, R. 1998. *Surf. Sci. Rep.*, **30**, 207.
3. de Boer, M.P., Clews, P.J., Smith, B.K., and Michalske, T.A. 1998. *Mater. Res. Soc. Proc.*, **518**, 131.
4. Tanner, D.M., Miller, W.M., Eaton, W.P., Irwin, L.W., Peterson, K.A., Dugger, M.T., Senft, D.C., Smith, N.F., Tangyunyong, P., and Miller, S.L. 1998. *Proc. IEEE 36th Annu. Int. Reliability Physics Symp.*, p. 26.
5. van Kessel, P.F., Hornbeck, L.J., Meier, R.E., and Douglass, M.R. 1998. *Proc. IEEE*, **86**, 1687.
6. Hartzall, A. and Woodilla, D. 1999. *Proc. IEEE 37th Annu. Int. Reliability Physics Symp.*, pp. 202-205.
7. Majumder, S., Mcgruer, N.E., Adams, G.G., Zavracky, P.M., Morrison, R.H., and Krim, J. 2001. *Sensor Actuat. A*, **93**, 19.
8. Kaajakari, V., Kan, S.H., Lin, L., Lal, A., and Rodgers, M. 2000. *Proc. SPIE*, 4180, 60.
9. Komvopoulos, K. 1996. *Wear*, **200**, 305.
10. Maboudian, R., and Carraro, C. 2004. *Annu. Rev. Phys. Chem.*, **55**, 35.
11. Xue, X., Polycarpou, A.A., and Phinney, L.M. 2008. *J Adhes. Sci. Technol.*, **22**, 429.
12. Alley, R.L., Mai, P., Komvopolous, K., and Howe, R.T. 1993. *Proc. 7th Int. Conf. Solid-State Sens. Actuators, Transducers '93, Japan*, p. 288.
13. Abe, T., Messner, W.C., and Reed, M.L. 1995. *Proc. IEEE Micro Electro Mechanical Systems Conf., Amsterdam, The Netherlands*, p. 94.
14. Dugger, M.T. 2005, *Proc. World Tribology Congress III, Washington D.C.*, p. 711.
15. Gao, D., Carraro, C., Howe, R.T., and Maboudian, R. 2006. *Tribology Lett.*, **21**, 226.
16. Rogers, J.W. and Phinney, L.M. 2001. *J. Microelectromech. Syst.*, **10**, 280.
17. Gogoi, B.P., and Mastrangelo, C.H. 1995. *J. Microelectromech. Syst.*, **4**, 185.
18. Christenson, H.K. 1988. *J. Colloid Interface Sci.*, **121**, 170.
19. Mate, C.M., McClelland, G.M., Erlandsson, R., and Chiang, S. 1987. *Phys. Rev. Lett.*, **59**, 1942.
20. Mastrangelo, C.H., and Hsu, C.H. 1992. *Proc. IEEE Solid-State Sens. Actuator Workshop, Hilton Head*, p. 208.
21. de Boer, M.P., and Michalske, T. 1999. *J. Appl. Phys.*, **86**, 817.
22. DelRio, F.W., Dunn, M.L., Phinney, L.M., Bourdon, C.J., and de Boer, M.P. 2007. *Appl. Phys. Lett.*, **90**, 163104.
23. Jones, E.E., Begley M.R., and Murphy, K.D. 2003. *J. Mech. Phys. Solids*, **51**, 1601.
24. Leseman, Z.C., Carlson, S., Xue, X., and Mackin, T.J. 2006. *Proc. ASME Intl Mech. Eng. Congress and Expo.*, Paper No. IMECE2006-14498, Chicago, IL.
25. Legtenberg, R., Tilmans, A.C., Elders, J., and Elwenspoek, M. 1994. *Sensor Actuat. A*, **43**, 230.
26. Alley, R.L., Cuan, G.J., Howe, R.T., and Komvopoulos, K. 1992. *Proc. IEEE Solid-State Sens. Actuator Workshop, Hilton Head*, pp. 202-207.
27. Ashurst, W.R., de Boer, M.P., Carraro, C., and Maboudian, R. 2003. *Appl. Surf. Sci.*, **212-213**, 735.

28. Greenwood, J.A., and Williamson, J.B.P. 1966. *Proc. Roy. Soc. Lond. Ser. A*, **295**, 300.
29. Chang, W.R., Etsion, I., and Bogy, D.B., 1988. *J. Tribology*, **110**, 50.
30. Muller, V.M., Yushchenko, V.S., and Derjaguin, B.V. 1983. *J. Colloid Interface Sci.*, **92**, 92.
31. Kogut, L., and Etsion, I. 2004. *J. Tribology*, **126**, 34.
32. Kogut, L., and Etsion, I. 2003. *J. Colloid Interface Sci.*, **261**, 372.
33. Shi, X., and Polycarpou, A.A. 2005. *J. Colloid Interface Sci.*, **281**, 449.
34. Lee, S.C. 2004, *Microtribodynamics of Sub-Five Nanometers Flying Head-Disk Interfaces in Magnetic Storage*, Ph.D. Thesis, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.
35. Shi, X., and Polycarpou, A.A. 2005. *J. Colloid Interface Sci.*, **290**, 514.
36. Israelachvili, J. 1985, *Intermolecular and Surface Forces*, Academic Press, San Diego.
37. Li, Y., and Talke, F.E. 1990. *Tribology and Mechanics of Magnetic Storage Systems*, B. Bhushan (Ed.) (STLE, Park Ridge, IL) Vol. VII, *STLE Special Publications SP-29*, pp. 79-84.
38. Tian, H., and Matsudaira, T. 1993. *ASME J. Tribology*, **115**, 28.
39. Gao, C., Tian, X., and Bhushan, B. 1995. *Tribology Trans.*, **38**, 201.
40. Gui, J., and Marchon, B. 1995. *J. Appl. Phys.*, **78**, 4206.
41. Fogden, A., and White, L.R. 1990. *J. Colloid Interface Sci.*, **138**, 414.
42. Zhang, B., and Nakajima, A., 1999. *J. Colloid Interface Sci.*, **211**, 114.
43. Xue, X., and Polycarpou, A.A. 2008. *J. Appl. Phys.*, **103**, 023502.
44. Xue, X., and Polycarpou, A.A., 2007. *J. Colloid Interface Sci.*, **311**, 203.
45. Stanley, H.M., Etsion, I., and Bogy, D.B. 1990. *ASME J. Tribology*, **112**, 98.
46. Suh, A.Y., Lee, S.C., and Polycarpou, A.A. 2006. *ASME J. Tribology*, **128**, 801.
47. Hegde, R.I., Paulson, W.M., and Tobin, P.J. 1995. *J. Vac. Sci. Technol. B*, **13**, 1434.
48. Sundararajan, S., and Bhushan, B. 2001. *J. Vac. Sci. Technol. A*, **19**, 1777.
49. Xue, X., and Phinney, L.M. 2003. *Proc. SPIE*, 4980, 130.
50. Xue, X., Koppaka, S.B., Phinney, L.M., and Mackin, T.J. 2004, *SEM International Congress & Exposition on Experimental & Applied Mechanics*, Paper No. 305, pp. 1-7.
51. Sniegowski, J.J., and de Boer, M.P. 2000. *Annu. Rev. Mater. Sci.*, **30**, 299.
52. Voutsas, A.T., and Hatlis, M.K. 1992. *J. Electrochem. Soc.*, **139**, 2659.
53. Xue, X., Phinney, L.M., and Polycarpou, A.A. 2008. *Microsyst Technol.*, **14**, 17.
54. Suh, A.Y., and Polycarpou, A.A. 2006. *Wear*, **260**, 538.
55. Yu, N., and Polycarpou, A.A. 2004. *ASME J. Tribology*, **126**, 225.
56. McCool, J.I. 1987. *ASME J. Tribology*, **109**, 264.
57. Yu, N., and Polycarpou, A.A. 2002. *ASME J. Tribology*, **124**, 367.
58. Rogers, J.W., Mackin, T.J., and Phinney, L.M. 2002. *J. Microelectromech. Syst.*, **11**, 512.
59. Schennum, S. 1995, *An Experimental and Theoretical Investigation of Delamination Resistance in Glass Fiber Reinforced Polyestere Laminates*, M.S. Thesis, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.
60. Maugis, D. 1992. *J. Colloid Interface Sci.*, **150**, 243.
61. Johnson, K.L., 1998. *Tribology Int.*, **31**, 413.
62. Derjaguin, B.V., Muller, V.M., and Toporov, Yu.P. 1975. *J. Colloid Interface Sci.*, **53**, 314.

63. Johnson, K.L., Kendall, K., and Roberts, A.D. 1971. *Proc. R. Soc. London Ser. A*, **324**, 301.
64. Komvopoulos, K. 2003. *J. Adhesion Sci. Technol.*, **17**, 477.
65. Elderton, W.P., and Johnson, N.L. 1969. *System of Frequency Curves*, Cambridge University Press, Cambridge, UK.
66. Tyndall, G.W., Leezenberg, P.B., Waltman, R.J., and Castenada, J. 1998. *Tribology Letters*, **4**, 103.