E

E2 – 17β-Estradiol

► Modified UHMWPE for the Hip Joint (Particle Filled and Reinforced)

Ecofriendly Textiles

► Sustainable Technology for Tribological Textiles

Edge Crack

► Rigid Punch Problem with a Crack

EDSC – Equivalent Discrete Spherical Convolution Considering Spherical/Aspheric Geometry

► Geometry of Spherical/Aspheric Bearings

EDSC – Equivalent Discrete Spherical Convolution for Joint Simulation

► Biotribological Joint Simulation System

EDSC – Equivalent Discrete Spherical Convolution for Spherical-Bearing Friction Prediction

► Friction Prediction for Spherical Bearings

EDSC – Equivalent Discrete Spherical Convolution for Spherical-Bearing Lubrication Analysis

► Lubrication Theory for Spherical Bearings

EDSC – Equivalent Discrete Spherical Convolution for Spherical-Bearing Wear Modeling

► Wear Modeling of Spherical Bearings

Effect of Lubrication on Fuel Economy in IC Engines

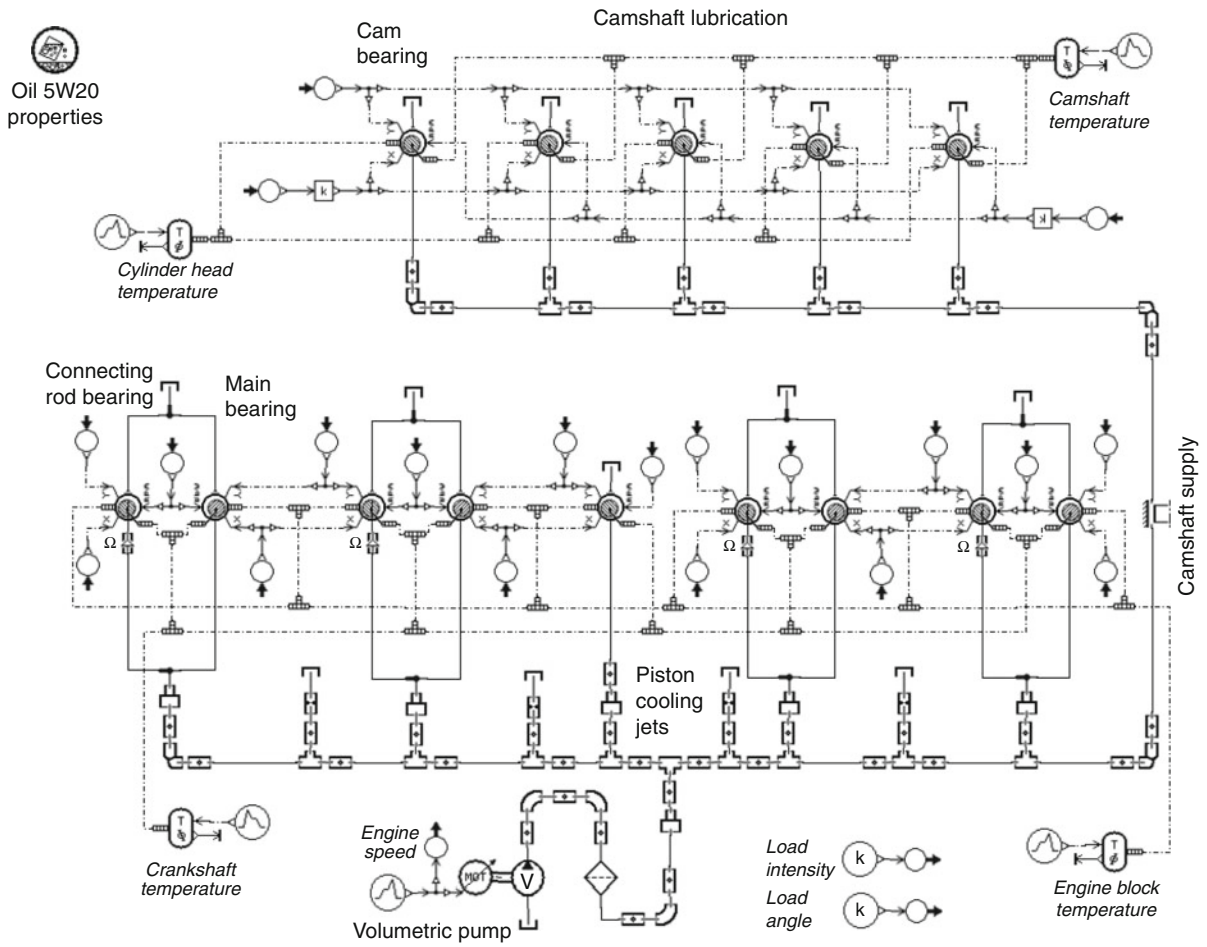
William Wei Tao
Chrysler Group, Auburn Hills, MI, USA

Definition

The main functions of engine lubrication are to reduce friction, cool the engine, limit wear on moving parts, and protect against corrosion. Lubrication primarily affects friction, which impacts fuel economy in engines. Engine fuel economy is related to engine power losses. Better fuel economy can be achieved by minimizing power losses. The main cause of power losses is friction, followed by auxiliary units and lubricant losses.

Rising fuel costs and the need to reduce emissions have led to increased demand on lubricants to improve fuel economy. Lubricant formulations can provide a beneficial reduction in engine friction, thus improving fuel economy. The new approach in engine oil development is to reduce friction losses from both elastohydrodynamic lubrication and hydrodynamic lubrication regimes by

Q.J. Wang, Y.-W. Chung (eds.), *Encyclopedia of Tribology*, DOI 10.1007/978-0-387-92897-5,



Effect of Lubrication on Fuel Economy in IC Engines, Fig. 1 Typical engine lubrication system (AMESim model)

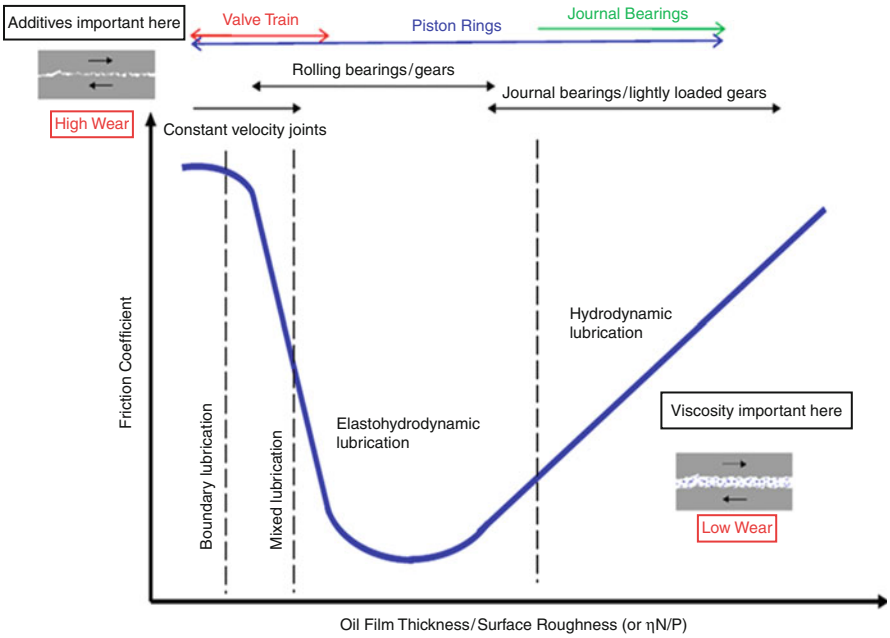
tailoring the viscosity characteristics of the base oil and the chemistry of the friction additives.

Scientific Fundamentals

Friction is a major cause of energy losses in engines. Engine oils reduce friction by separating metal pairs to prevent Hertzian contact (hydrodynamic lubrication), and use friction modifiers to reduce wear when metal-to-metal contact occurs (boundary lubrication). Oil viscosity and friction modifiers directly influence engine fuel economy. Low viscosity results in better fuel economy, but its application is limited by engine oil consumption and durability concerns.

The main purpose of the engine lubrication system is to provide a specific lubricant at the correct temperature and pressure to all moving parts of the engine. The lubrication oil is pulled from the sump into the oil pump, then forced through an oil filter and pressure-fed to the engine block through the main oil gallery to the main bearings. From the main bearings, the oil passes through the rotating passages in the crankshaft and onto the big-end bearings of the connecting rod. The cylinder walls and piston-pin bearings are lubricated by oil being dispersed by the rotating crankshaft or squirters. The camshaft bearings, roller finger followers, variable valve timing systems, and hydraulic lash adjusters are fed by the oil supply passage in the cylinder head. The other bleed supplies the timing chains or gears on the camshaft drive. The excess oil then drains back to the sump, where the heat is dissipated to the surrounding air either directly or heat exchanger. The lubricating system of a typical engine is shown in Fig. 1.

The lubrication system serves several functions essential to the safe and reliable operation of an engine. These functions are:



Effect of Lubrication on Fuel Economy in IC Engines, Fig. 2 Influence of lubricant properties on friction (R. I. Taylor, © Shell Research Limited. "Tribology & Energy Efficiency" Robert Ian Taylor, Shell Global Solutions (UK). Donald Julius Groen Prize Lecture. Institute of Mechanical Engineers, London, 3rd Dec 2008)

- Lubricating the bearings (main, rod, cam, balance shaft bearings, etc.);
- Supplying pressure to activate/deactivate hydraulic systems (hydraulic lash adjusters, variable timing systems, actuators, cylinder cut-off, etc.);
- Cooling of the bearings, pistons, and cylinder walls;
- Supplying a squeeze film between the piston rings when they stop moving at the top dead center (TDC) and the bottom dead center (BDC);
- Removing contaminants from the lubricant.

Key Applications

Lubrication plays a more and more important role in increasing engine performance, improving fuel economy, and reducing emissions. There are many state-of-art engine technologies that have been introduced in engine development to deliver significant fuel efficiency gains with reduction of parasitic losses, including as variable valvetrain (cam phasing, dual cam phaser, and coordinated cam phaser, continuously variable valve lift, discrete cam profile switching, discrete variable valve lift, and camless), homogeneous charge compression ignition combustion, cylinder deactivation, gasoline direct injection, and turbocharging/downsizing.

The majority of the frictional losses in an engine are attributable to the three main tribological categories: piston assembly, journal bearings, and valvetrain. These components have been classified as boundary/mixed, elastohydrodynamic, and hydrodynamic lubrication regimes based on their tribological characteristics, as shown in Fig. 2. The piston assembly is lubricated hydrodynamically, although mixed/boundary friction occurs at top and bottom dead centers (TDC and BDC), where loads are high and speeds are low. The journal bearings are lubricated hydrodynamically, except during cold start, which is in the elastohydrodynamic regime. For piston assembly and journal bearings, the most effective way of reducing friction losses is to reduce the viscosity of the lubricant.

The valvetrain predominantly operates in the mixed/boundary lubrication regime. The oil film thickness between cam and tappet is determined using elastohydrodynamic lubrication theory. Typically, the oil film thickness is substantially lower than the surface roughness of the metallic components due to high contact pressure and local metal surface elastic deformation. Under these circumstances, an effective way to reduce friction is to use friction modifiers.

Reducing engine frictional losses and improving fuel efficiency can be achieved by lowering the dynamic viscosity of the base lubricant in the hydrodynamic lubrication regime; by lowering the alpha coefficients of the lubricants in the elastohydrodynamic regime; and by using combinations of solid additives in boundary/mixed lubrication.

Mechanical friction contributes to about 10% of the combustion heat. Efforts have been made to improve thermodynamics and minimize mechanical losses in engines. The improvement in thermodynamic efficiency also creates very high cylinder pressures, which increases piston ring friction and wear. The low sliding speed around the top dead center results in a thin lubricant film and significant metallic contact, which increases the friction coefficient dramatically. High peak pressure that occurs around TDC ultimately intensifies this phenomenon.

It has been reported that the piston assembly accounts for 45–65% (Pesic et al. 2005) of total engine friction. Improvements in design of pistons, application of new coatings, and lubrication improvements directly lead to improvement in fuel economy.

Besides the lubricant considerations, other factors also need to be addressed. Engine design has a significant influence on the effectiveness of the lubricant in reducing fuel consumption. For example, a lubricant containing a friction modifier is effective in reducing fuel consumption, but, due to the contact difference, the same lubricant is not so effective in reducing fuel consumption in an engine that uses a roller follower valvetrain system.

In the development of high fuel efficient engines, reducing boundary friction can increase fuel economy by 1%; using low-viscosity lubricant increases fuel economy by 0.5%; reducing boundary/asperity friction enables the use of lower-viscosity fluids to gain fuel savings of 3–5%. However, lower viscosity fluids may cause higher wear and thus require more wear-resistant or durable materials and interfaces. The potential disadvantage of moving to lower-viscosity lubricants is the thinner oil film that is expected to exist between lubricated contact surfaces within the engine. Bearing durability is also an area of concern.

The lubrication system in engine development is an essential element for any engine design. However, it is desirable to achieve maximum fuel economy during the engine design stage while all the factors discussed in this context are collaboratively optimized.

Cross-References

► Lubrication Regimes
► Oil Life

References

G. Fenske, Impact of Friction Reduction Technologies on Fuel Economy, Society of Tribologists and Lubricated Engineers, Chicago Chapter, Mar 2009

P.M. Lee, M. Priest, *Influence of Gasoline Engine Lubricant on Tribological Performance, Fuel Economy and Emission* (IME/Chandos, London/Oxford, 2007), pp. 235–246. ISBN 1843344513

R. Pesic, A. Davinic, S. Veinovic, Methods of tribological improves and testing of piston engines, compressors and pump. Tribol. Ind. **27** (1&2), pp. 38–47 (2005)

R.I. Taylor, R.C. Coy, Improved fuel efficiency by lubricant design: A review, *Published in the IMechE Proceedings, Part J, Journal of Engineering Tribology*, 2000

R.I. Taylor, *Tribology & Energy Efficiency* (IME, London, 2009)

Effect of Plastic Deformation on EHL Characteristics

► Plasto-Elastohydrodynamic Lubrication (PEHL)

Effect of Roughness on Lubrication

► Average Reynolds Equation

Effect of Roughness Orientation on EHL Film Thickness

► Roughness Effect on Elastohydrodynamic Lubrication

Effect of Surface Finish on EHL Performance

► Roughness Effect on Elastohydrodynamic Lubrication

Effect of Surface Roughness on EHL Characteristics

► Roughness Effect on Elastohydrodynamic Lubrication

EHD Lubrication Subjected to Line Contacts

► Lubrication Regimes – Line Contacts

EHL

► Elastohydrodynamic Lubrication (EHL)

EHL Considering the Effect of Deterministically Modeled Surface Roughness

► Deterministic Models of Rough Surface EHL

EHL Film

► EHL Film Thickness Behavior

EHL Film Thickness Behavior

Roland Larsson
Division of Machine Elements, Luleå University of Technology, Luleå, Sweden

Synonyms

EHL film

Definition

EHL film thickness is the separation between surfaces of an elastohydrodynamically lubricated contact. The thickness is normally described by a minimum value, h_{min}, which is the smallest separation and thus also the most critical. The variation of the thickness inside the contact zone is called the *film thickness profile.*

Scientific Fundamentals

Elastohydrodynamic lubrication (EHL) occurs when lubricant film pressure is high enough to significantly influence the shape of the film thickness profile. EHL conditions are typically obtained in nonconformal contacts such as ball on cylinder (elliptical contact), ball on ball (circular contact), and cylinder on cylinder (line contact). Refer to "► Elastohydrodynamic Lubrication (EHL)" for more general description of EHL.

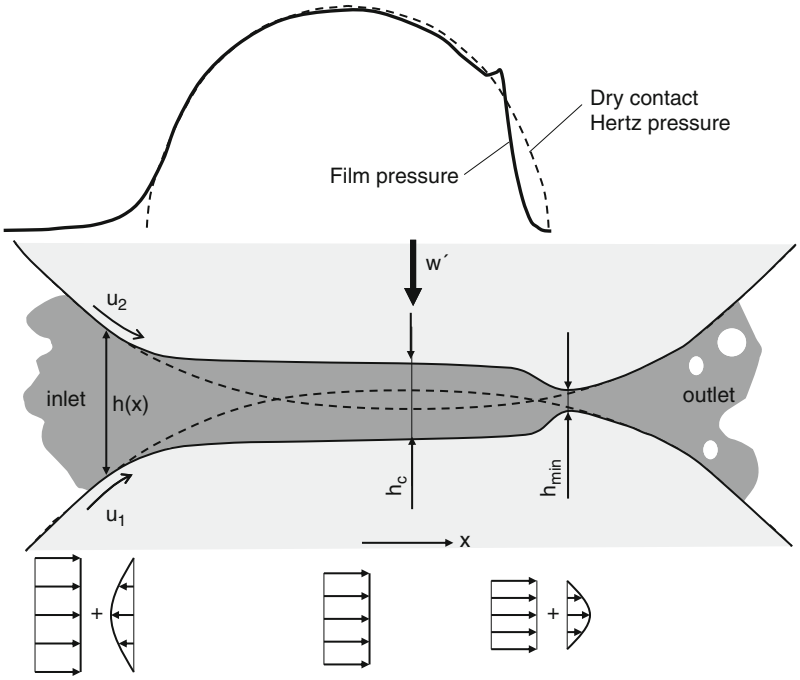
Film thickness is one of the primary properties of EHL since it describes how well surfaces are separated and thus how well the lubricant film protects them from wear and high friction. Figure 1 shows a typical film thickness profile for a line contact. Note that the scale in y-direction is strongly exaggerated, the film thickness is typically $O(10^3)$ times smaller than the length of the contact. When surfaces start to move they will force lubricant into the inlet zone and a small amount will enter the high pressure zone of the contact (see Fig. 1). The surfaces are deformed by the high pressure and the contact becomes more conformal. Pressure rapidly reaches 100–4,000 MPa and the viscosity of normal lubricants, such as mineral oils, will increase dramatically. The viscosity increase near the inlet (at pressures in the range 0–300 GPa) of the contact is well described by Barus equation (Barus 1893):

$$\eta = \eta_0 \exp(\alpha p) \tag{1}$$

where α is the pressure-viscosity coefficient that normally is in the range 15–30 GPa^{-1}, η_0 is the dynamic viscosity under atmospheric conditions. At 300 MPa the viscosity will have increased by a factor 90–8,000 and it will be more and more difficult to squeeze out the lubricant from the contact. At even higher pressure the lubricant will undergo a phase transition into a solid glassy state. A common way to exemplify and illustrate how thin the film is in comparison to its length and width is to compare it with a sheet of printing paper. It is then easy to understand that it is very difficult to force the thin sheet of glassy lubricant out from the contact. The physics and mathematics of this mechanism is explained in next section.

Flow Continuity Controls Film Thickness

Surfaces are deformed when two nonconformal bodies such as two cylinders are brought in contact. A flat narrow contact zone is formed and contact width and pressure can be predicted by using the Hertz contact theory. The contact pressure can easily reach the GPa-range for steel surfaces. This flat-deformed zone will remain also when a lubricant is supplied and cylinders start to roll. The lubricant is *entrained* into the flat region and the film thickness profile in Fig. 1 is formed. The flow velocity



EHL Film Thickness Behavior, Fig. 1 Typical film thickness profile for an EHL line contact. Flow velocity profiles (Couette + Poiseuille flow) at three different positions are shown in the bottom of the figure

profiles at three different positions along the film are shown in Fig. 1. The flow has two components, Couette flow (surface driven flow) and Poiseuille flow (pressure driven flow). Due to continuity requirements, the total flow must be the same at all three positions. This can be described mathematically as (Hamrock 1994):

$$q(x) = -\frac{\rho h^3}{12\eta}\frac{\partial p}{\partial x} + \frac{u_1 + u_2}{2}(\rho h) = const. \qquad (2)$$

where $h(x)$ is film thickness profile, $p(x)$ is the film pressure, $\rho(x)$ is density, and u_i surface velocities. The first term in (2) is the pressure gradient driven Poiseuille flow and the second the velocity driven Couette flow. The ratio between Poiseuille and Couette flow becomes normally very small in the central region of the contact. This can be shown by an order of magnitude analysis for a typical case where $h = O(10^{-6})$, $\eta = O(10^3)$, $\partial p/\partial x = O(10^{13})$, and $u_1 = u_2 = O(1)$. The ratio will thus be $O(10^{-3})$ and often even smaller. The Reynolds equation (see "► Reynolds Equation") is derived as the continuity equation:

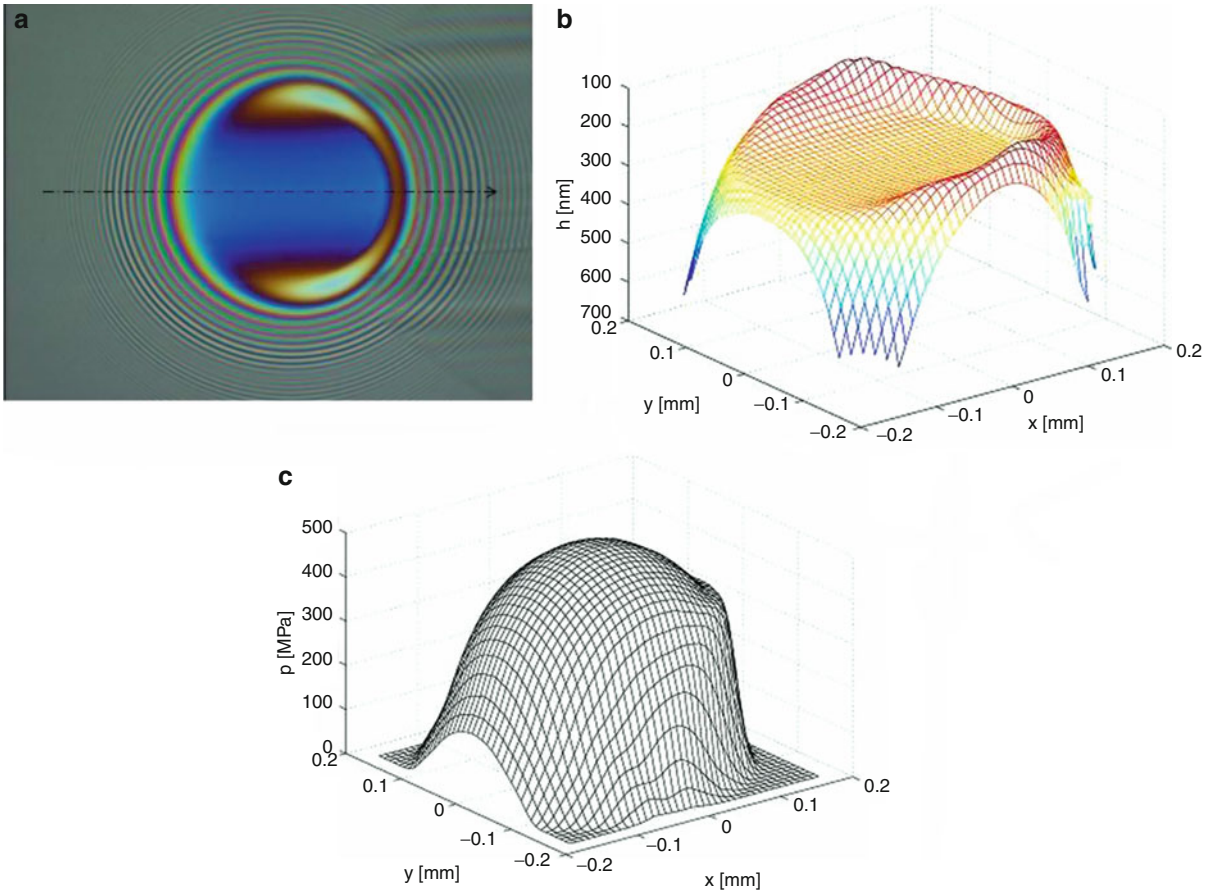
$$\frac{\partial q}{\partial x} = -\frac{\partial}{\partial x}\left(\frac{\rho h^3}{12\eta}\frac{\partial p}{\partial x}\right) + \frac{u_1 + u_2}{2}\frac{\partial(\rho h)}{\partial x} = 0 \qquad (3)$$

Due to the dominance of the Couette term it can be reduced to

$$\frac{\partial(\rho h)}{\partial x} = 0 \qquad (4)$$

in the center of the contact. The effect of (4) can be seen in Fig. 1 as an almost constantly thick lubricant film in the center region. An oil is relatively incompressible and density of the lubricant, ρ, does not vary more than a few percent at pressures below 1 GPa and the equation can thus be written $h \approx$ constant. At higher pressure and with more compressible lubricants there will be a 10–30% increase of density, and film thickness will consequently be reduced with the same amount.

Further to the right, just before the outlet of the contact, there is a sudden decrease of the film thickness. This is the position where h_{min} occurs. This flow constriction occurs in order to maintain continuity of flow. The Poiseuille flow is significantly larger near the inlet and outlet due to the high pressure gradients. In the inlet it will counteract the Couette flow, while it will be in the same direction as the Couette flow in the outlet, and the gap must therefore be closed to balance flow into and out from the contact. This means that the



EHL Film Thickness Behavior, Fig. 2 EHL point contact, (**a**) interferogram, (**b**) film thickness map, (**c**) pressure distribution

entrained amount of lubricant controls the film thickness both in the central part and at the gap closing. Heating of the lubricant in the inlet zone will give rise to lower viscosity and it will be more difficult to entrain the fluid into the contact. The same effect will occur if the lubricant slips at the surfaces or if the lubricant has a non-Newtonian shear thinning behavior already at the relatively low pressure and low shear rates in the inlet zone.

Side leakage may occur in the three-dimensional case (i.e., there is a possibility for the lubricant to escape sideways). The film formation mechanisms are, however, the same as in the line contact case. Figure 2a shows the results from an interferometric study of a ball in rolling contact with a transparent disc (see "► Optical Interferometry" for more details). Each color represents a certain film thickness. As in the line contact case there is a large proportion of the contact where film thickness is constant (blue area in Fig. 2a). Along the outlet and along the sides, however, there is a horseshoe-shaped constriction that prevents the lubricant from leaking out and balances the flows. The minimum film thickness will occur at two symmetrical positions along this horseshoe-shaped ridge (see Fig. 2b). At lower loads the two h_{min} positions will move closer to the centerline or even coincide at the centerline. Figure 2c shows the pressure distribution required to obtain the film thickness profile shown in Fig. 2b.

Physical Parameters Controlling Film Thickness

The most important physical parameters that control film thickness in isothermal EHL are described in Table 1. Thermal effects are discussed later.

In order to reduce the number of parameters, a Buckingham π-theorem analysis can be used to identify four dimensionless groups. One of them is D and the other

EHL Film Thickness Behavior, Table 1 Main EHL parameters (numbers 1 and 2 refers to lower and upper surface)

Average lubricant entrainment velocity, $\tilde{u} = (u_1 + u_2)/2$ (m/s)
Reduced radius of curvature in the direction of motion, R_x (m)
Effective elastic modulus of the surfaces, $E' = \left[\left(1 - \nu_1^2\right)/E_1 + \left(1 - \nu_2^2\right)/E_2\right]^{-1}$ (N/m^2)
Dynamic viscosity of lubricant under atmospheric conditions, η_0 (Pas)
Lubricant pressure-viscosity coefficient, α (GPa^{-1})
External load, w (load per unit width for line contacts) (N) or (N/m)
Ratio of reduced radii of curvature, $D = R_x/R_y$

three are U, W, and G, which are defined as (Dowson and Higginson 1966):

Dimensionless speed parameter: $U = \frac{\tilde{u}\eta_0}{E' R_x}$

Dimensionless load parameter: $W = \frac{w}{E' R_x^k}$ where $k = 1$ for line contact and 2 for elliptical contact

Dimensionless material parameter: $G = \alpha E'$

Hamrock and Dowson (1981) derived an empirical expression describing the influence of the dimensionless numbers on the dimensionless minimum and central film thickness:

$$H_{\min} = 3.63\, U^{0.68} G^{0.49} W^{-0.073} \left[1 - \exp\left(-0.68 D^{-2/\pi}\right)\right] \tag{5}$$

$$H_c = 2.69\, U^{0.67} G^{0.53} W^{-0.067} \left[1 - 0.61 \exp\left(-0.73 D^{-2/\pi}\right)\right] \tag{6}$$

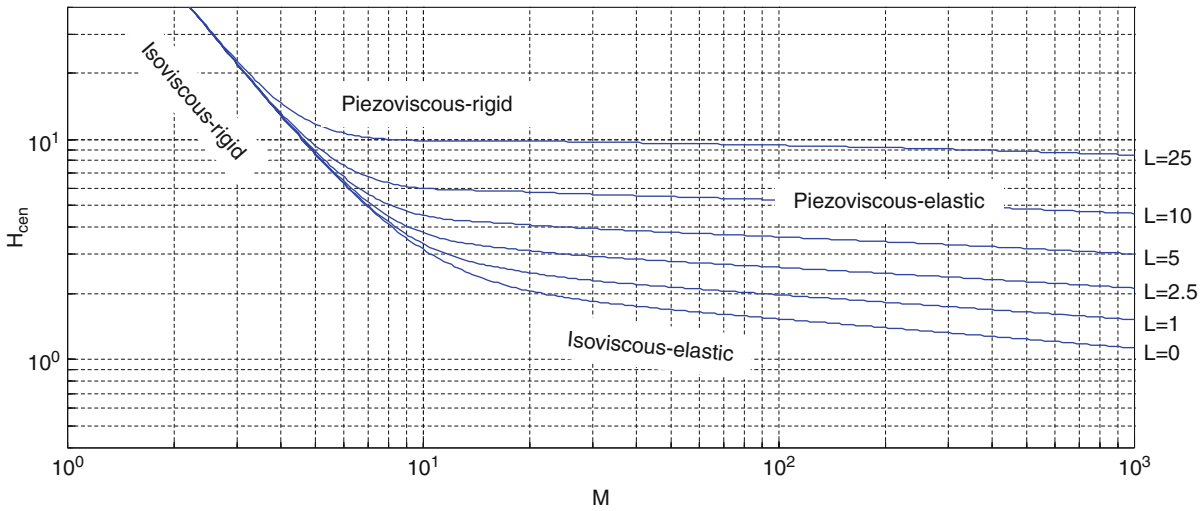
where $H_{\min} = h_{\min}/R_x$ and $H_c = h_c/R_x$. These expressions where obtained using regression analysis of a large number of data sets (H, U, W, G, D) obtained from numerical solutions of the EHL governing equations, see "► EHL Governing Equations". The reader must be aware of that these empirical expressions are only valid within a limited parameter range. Please refer to "► Film Thickness Formulas: Point Contacts" for more information about film thickness formulae.

Even if expressions (5) and (6) are approximate, they show the effect of the different physical parameters. As expected, the entrainment velocity has a large effect on film thickness, the exponent is 0.67–0.68. The higher the velocity, the more lubricant will be entrained and the thicker the film will become. The viscosity has the same exponent and will thus have as large influence on film thickness as the velocity. As explained above, the film thickness is controlled by the inlet conditions and the balance between Poiseuille and Couette flows. The Poiseuille flow will counteract the entrainment of lubricant and will decrease when viscosity increases, see (3). The pressure-viscosity coefficient has also a high exponent, 049–0.53, and the explanation is the same as for η_0. The higher the pressure-viscosity coefficient becomes, the higher the viscosity will be at the inlet where pressure starts to rise.

The load has a relatively small influence on film thickness as the exponent is around −0.07. This means that the film thickness will decrease with only 5% if the load is doubled. This can be explained by the growth of the contact area when load increases. The lubricant inside the contact has an enormous viscosity and cannot be squeezed out easily. When the contact grows it will be even more difficult and the surfaces are wrapped around the viscous lubricant in the contact center.

The total exponent for the effective elastic modulus is around −0.10, which means that it has a relatively small influence. This is a little bit surprising since EHL is defined by the occurrence of elastic deformation. At very high E', the surfaces can be seen as rigid and there will be no influence on film thickness at all. The deformation will increase when E' becomes smaller and the contact becomes more conformal and more lubricant can be entrapped. The film thickness will thus increase and that explains the negative sign of the exponent for E'. The pressure required to deform the surfaces will decrease as the elastic modulus becomes smaller. The so-called "soft-EHL" regime occurs when E' is so small that the maximum fluid film pressure does not cause a significant increase of the fluid viscosity. This regime is also called the *isoviscous-elastic* regime and will be further discussed below. The influence of E' on film thickness will increase as the surfaces become less rigid and (5) and (6) are not valid for this regime. In the "hard-EHL" regime, when E' still is relatively high, there is a balance between two effects and this explains the small number of the exponent. When E' is decreasing the contact becomes larger and it will be more difficult to squeeze out the lubricant from the contact and the film thickness tends to grow. The maximum pressure and thus also the viscosity will, on the other hand, be lower and the film thickness tends to decrease.

The reduced radius of curvature in the direction of motion, R_x, has relatively large influence on film thickness. The total exponent is around 0.46. As R_x increases the contact becomes more conformal and more lubricant will be entrapped in the inlet zone. Film thickness will, therefore, increase as radius increases.



EHL Film Thickness Behavior, Fig. 3 EHL regimes and dimensionless film thickness and its variation with M and L (D = 0.4)

The ellipticity, represented by D, of the contact zone influences the film thickness in such way that it becomes thinner as D increases. The contact will become longer but narrower as D increases. The negative effect of lubricant side leakage will thus cause the film thickness to decrease.

EHL Regimes

So far, the discussion has mainly concerned the hard EHL regime where both pressure-viscosity and elastic deformation have significant effects on film thickness. There are three additional regimes of EHL depending on the degree of elastic deformation and the effect of the increase in viscosity with pressure (Hamrock 1994). By using so-called optimum similarity analysis it can be shown that only two dimensionless groups are required to accurately describe how film thickness varies for a given contact geometry, D. The two Moes parameters, M and L, will be used here (Moes 1992). For an elliptical contact they are defined as:

$$M = \frac{W}{(2U)^{3/4}} \tag{7}$$

$$L = G(2U)^{1/4} \tag{8}$$

The regimes can be defined from the value of M and L (see Fig. 3):

Isoviscous-rigid regime ($L = 0$ and small M): This is the regime where there is no effect of elastic deformation and the viscosity can be taken as a constant. Consequently, this is not an EHL regime but is normally included to make the regime map complete.

Piezoviscous-rigid regime (large L and small M): This is a case where surface deformations are insignificant but pressure is high enough to influence viscosity. This form of lubrication may be found in some piston ring–cylinder liner contacts.

Isoviscous-elastic regime ($L = 0$ and large M): This is the soft-EHL regime mentioned above. A typical example of this type of lubrication can be found in elastomer lip seal contacts.

Piezoviscous-elastic regime (large L and large M): This is the hard-EHL regime. The film formation mechanisms, as described in the previous section, rely on elastic deformation of surfaces in order to make them more conformal and on the increase in viscosity with pressure to keep the lubricant inside the contact. Typical applications are rolling element bearings, cams, and gears.

Nijenbanning et al. (1994) used numerical simulations to solve the EHL equations (see "► Stochastic Models for Rough Surface EHL") for different contact geometries and for a large number of combinations of M and L. The regime diagram in Fig. 3 shows dimensionless central film thickness

$$H_{cen} = \frac{h_{cen}}{R_x\sqrt{2U}} \tag{9}$$

for an elliptical contact where $D = R_x/R_y = 0.4$ in this case, where R_y is the reduced radius of the contacting bodies in the direction perpendicular to rolling/sliding motio. The diagram looks the same for other values of D but with somewhat different levels of H_{cen}.

Thermal Effects

Thermal effects occur when the lubricant is sheared, especially at high slide/roll ratios, and when the lubricant is rapidly compressed when it enters the contact region. Again it is important to study the inlet zone. The viscosity will drop if the lubricant temperature increases due to shearing and compression. Film thickness will consequently be smaller. The viscosity-temperature relationship will, therefore, become important. The lower the gradient $d\eta/dT$ is, the less influence the temperature increase will have. The specific heat capacity, ρc_p, has a similar effect. The higher it is, the more heat is required to increase lubricant temperature. Thermal conductivity of the lubricant, λ, and the bounding surfaces, λ_s, will also play an important role. The higher the conductivities are, the cooler film. Thermal effects in EHL have been studied intensively but there are only a few regression formulae available. Gupta et al. (1992) presented the following thermal correction factor for the isothermal film thickness:

$$C_T = \frac{1 - 13.2(p_h/E')Br^{0.42}}{1 + 0.213(1 + 2.23A^{0.83})Br^{0.64}} \leq 1 \qquad (10)$$

where Brinkman's number is defined as

$$Br = \left(-\frac{d\eta}{dT}\right)\frac{\tilde{u}^2}{\lambda} \qquad (11)$$

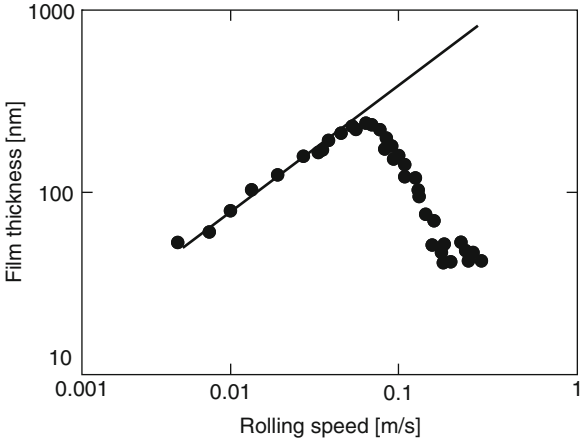
and slide/roll ratio is defined as the relative difference in velocity for surface 1 and 2:

$$A = \frac{u_2 - u_1}{\tilde{u}} \qquad (12)$$

The thermally corrected film thickness is obtained by multiplying the isothermal film thickness with the correction factor. Please note that expression (10) is only valid within a limited range but describes qualitatively the influence of the different parameters.

Starvation

Starvation is an effect that may limit the formation of a full EHL film thickness. Starvation occurs when the inlet of the contact is not fully flooded with lubricant. As shown in Fig. 1, the film pressure starts to increase well outside the Hertzian contact zone and this cannot occur if the gap is not filled with lubricant. Starvation has been studied theoretically by adjusting the position of the point where pressure starts to increase from zero (Hamrock and Dowson 1977) and later by using mass conserving cavitation boundary conditions for reformation zones (Chevalier et al. 1998). Optical interferometry has been applied in order to study the phenomena experimentally.

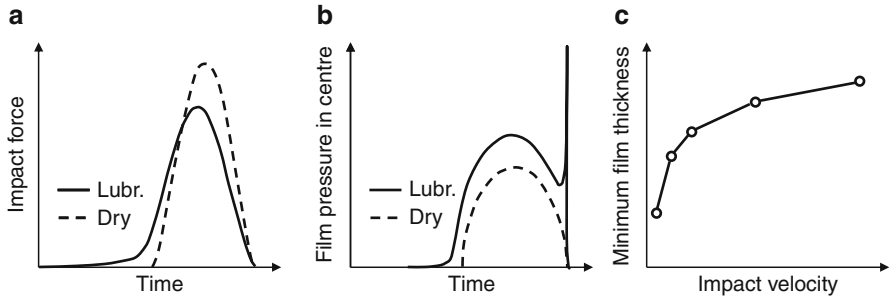


EHL Film Thickness Behavior, Fig. 4 Central film thickness versus entrainment speed in a starved EHL contact (Reproduced from Damiens et al. 2004)

Figure 4 shows how film thickness increases as velocity increases, but at a certain point it declines and the film becomes thinner (Chevalier et al. 1998). Starvation typically occurs in grease-lubricated contacts.

Dynamic Effects

There are many different types of dynamics effects. In fact, most EHL contacts operate under dynamic conditions. The contact between gear teeth is, for example, under continuous variation of entrainment speed, reduced radius, slide/roll ratio, and load. EHL contacts in cam mechanisms and rolling element bearings also operate under rapidly changing conditions. Dynamic conditions also occur on the local micro scale in a sliding stationary loaded contact since asperities travelling at different speed collide with each other. All these examples cannot be covered here and dynamic conditions will be exemplified by the impacting ball case (Dowson and Wang 1994; Larsson and Höglund 1994). If a perfectly smooth ball is dropped onto a smooth lubricated surface it is almost impossible to penetrate the film. When the ball impacts the lubricated surface, it will trap some of the lubricant in the contact center and, due to the increasing pressure, its viscosity will become high and very difficult to squeeze out since there is no sliding motion. Figure 5a shows that the lubricant film causes damping and the maximum impact force is lower than for the corresponding nonlubricated case. But the pressure at the contact center becomes higher in the lubricated case than in the dry case (see Fig. 5b). That is due to the extremely viscous lubricant that cannot escape from the center. The faster this process is, the more



EHL Film Thickness Behavior, Fig. 5 (**a**) Impact force versus impact time, dry and lubricated contact, (**b**) film pressure at the contact center versus impact time, (**c**) minimum film thickness versus initial impact velocity (Reproduced from Larsson and Höglund 1994)

lubricant will be entrapped and this explains why the minimum film thickness actually becomes thicker as the impact velocity increases (see Fig. 5c).

Non-smooth Surfaces

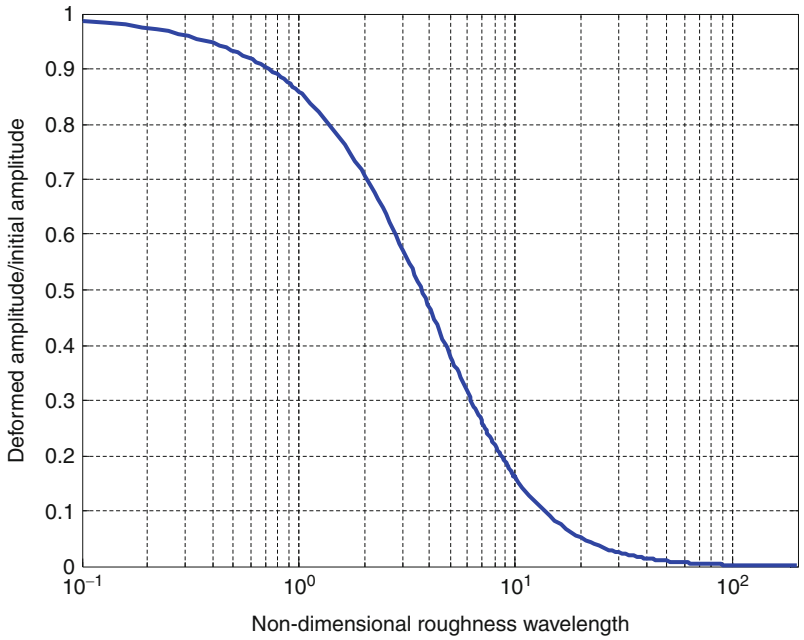
Film thickness behavior of nonsmooth surfaces is a very active research field. If the mechanisms of film formation in rough surfaces contacts can be understood, it will be possible to determine the optimum surface topography. It would then be feasible to obtain transition from mixed lubrication to full film lubrication at as low speed as possible.

The film parameter is the ratio between the smooth surface minimum film thickness, obtained from, for example, (5), and the composite *r.m.s.* surface roughness height:

$$\Lambda = \frac{h_{\min}}{\sqrt{S_{q1}^2 + S_{q2}^2}} \tag{13}$$

This equation is used in the classical method to estimate how smooth the surfaces have to be in order to obtain full film lubrication. Full film lubrication is normally said to occur if Λ is greater than 3 and boundary lubrication occurs if Λ is less than 1 (Hamrock 1994). Mixed lubrication occurs in the intermediate Λ-interval. The severity of each lubrication problem can therefore be predicted in a very efficient and rapid way, since it is relatively easy to calculate the film parameter. There are, however, obvious limitations with the film parameter theory. One of the more important ones is the fact that most modern elastohydrodynamically lubricated machine components seem to operate without problems at values of Λ less than 3 or even less than 1 (Cann et al. 1994). The explanation is that the surface roughness is flattened inside the contact and the film parameter should be based on the "in-contact" roughness height instead. The flattening effect is a complex interaction between surface deformation and flow of the highly viscous lubricant. The analytical work by Morales-Espejel and Greenwood (1994) showed that the film formation in line contacts with transversal roughness is made up of two effects. The first one dominates in rolling contacts and causes large deformation of the asperities as they enter the EHL contact region. The asperities are more or less totally flattened and the "true" film parameter is thus very much increased. For that reason, full film lubrication can be obtained even if the theoretical film parameter is less than one. When some degree of sliding is superimposed, the other effect becomes more important. This effect is caused by the entrainment of lubricant to the contact. The entrainment is controlled by the mean velocity of the surfaces and the film thickness at the inlet to the contact region. When one of the surfaces is rough and moves at another velocity than the other surface, there will be a fluctuating entrainment of lubricant. A great deal of lubricant is entrained when a valley enters the contact and less lubricant is entrained when an asperity passes the inlet. The fluctuating entrainment causes a film thickness variation that moves at the mean velocity. The asperity will therefore move at another speed than the film thickness disturbance it causes. The practical consequence is that analyses of rough surface EHL require full transient solutions of the coupled EHL equations.

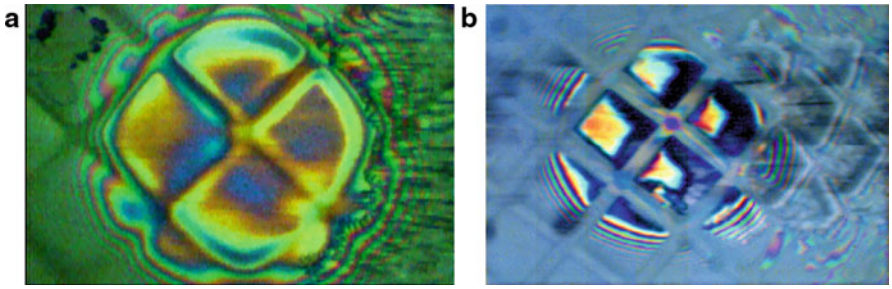
Lubrecht and Venner have in a number of papers (Venner and Lubrecht 1994, 1996; Lubrecht and Venner 1999) presented results from numerical simulations of rough surface EHL. One can learn much from their analyses, even if they studied a model problem with single wavelength sinusoidal roughness positioned on only one of the surfaces. Figure 6 shows their "amplitude reduction diagram," which can be useful for better predictions of lubrication conditions. The curve in the diagram shows



EHL Film Thickness Behavior, Fig. 6 The ratio between the deformed "in-contact" waviness amplitude and the initial "out-of-contact" waviness amplitude versus the nondimensional roughness wavelength (Lubrecht and Venner 1999)

the ratio between the deformed "in-contact" asperity amplitude and the undeformed "out-of-contact" asperity amplitude. The parameter on the horizontal scale describes the roughness wavelength, but takes also lubricant parameters and running conditions into account. Note that small wavelength roughness remains almost undeformed while waviness with a large wavelength becomes totally flattened. The curve can be applied to real surfaces if the roughness is dominated by one wavelength. A deformed roughness height can be obtained and then also a true film parameter. Masen et al. (2002) also tried to apply the curve to a case without dominant wavelengths. Instead they applied the asperity deformation theory on each individual wavelength component, obtained by using Fourier analysis. Then they superimposed all the deformed wavelengths on each other and obtained a deformed roughness. This is a promising method, but since there is a strong coupling between flow and deformation, one must investigate whether such a superposition can be allowed. Another obstacle is that only one surface is assumed to be rough. Both surfaces have roughness heights of the same order of magnitude in many tribological contacts. In the case of pure rolling it is possible to just add the roughness of both surfaces to arrive at a combined effective surface roughness. The combined roughness will, however, change transiently as soon as there is a relative velocity difference between the surfaces.

The advances in numerical computation methods and the increasing computer power have inspired some researchers to study real contacts with real measured roughness on both surfaces. Zhu and Ai (1997) have presented such a simulation where they studied the effect of roughness obtained by different machining processes. They also studied the effect of the roughness orientation. They found that the machining process (i.e., the surface texture and roughness orientation) did not have a significant influence on the average film thickness. The effect on the pressure peak amplitudes was, however, greater. They also showed that the flattening of asperities depends on the roughness wavelength. Interesting in their model is the ability to simulate breakdown of the lubricant film. The transition between mixed and full film lubrication was modeled by solving the EHL governing equations where film occurs and by solving the contact mechanics problem where metal-metal contact occurs, see for example (Hu and Zhu 2000) and (Ren et al. 2009) for more information about the model. The methodology presented by Zhu and co-workers has caused a debate about whether this technique is correct from a numerical analysis point of view; see references (Venner 2005) and (Zhu 2007) to learn more about the two standpoints.



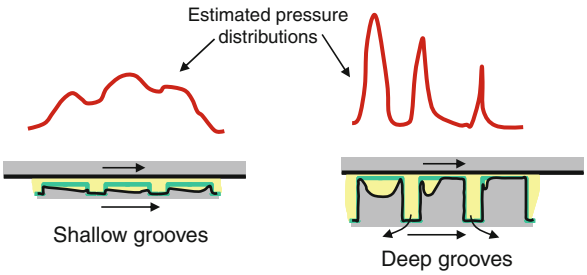
EHL Film Thickness Behavior, Fig. 7 Transient passage of a square pattern during sliding motion, (**a**) shallow grooves (**b**) deep grooves (Ehret et al. 2000)

The problem with numerical analyses of film formation between surfaces with real roughness is the need for a very dense mesh. In order to resolve the smallest roughness irregularities one might need 10^{10} grid points, and due to the transient nature of the problem one needs to solve it in approximately 10^5 time steps (Lubrecht and Venner 1999). This is still impossible, even with the most powerful computers. One can, however, question whether it is necessary to take the smallest irregularities into account. It is a simplification to disregard the smallest wavelengths, since it is known that those are very difficult to deform and thus also might also cause contact with the other surface. There are, on the other hand, many other simplifications that might cause larger deviations from the real behavior. The assumptions of a Newtonian fluid model and isothermal conditions are two such simplifications. It is also known that, even if it was possible to make an exact calculation of the film thickness and contact pressure, this would be of little value, since new surfaces, with new topography, enter the contact region all the time. The most reasonable use of analyses with real roughness is therefore to try to understand the coupling between the machining process or statistical surface roughness parameters and the film formation. There is, however, a great need for experimental validation of the numerical results. Optical interferometry can be used for that purpose.

Kaneta and his co-workers (1992, 1993; Kaneta and Nishikawa 1994) have utilized optical interferometry in the study of rough surface EHL. They demonstrated how well-defined surface irregularities manufactured on the smooth ball surface could be used in order to learn more about film formation mechanisms. They used bumps, ridges, waviness, and dents. They were the first to observe how the effect of single surface irregularities travelled at the mean surface speed, while the irregularity itself moved at another speed. They also studied the effect of waviness and the effect of the waviness orientation. A review of their work is presented by Kaneta and Nishikawa (1994).

A similar technique was adopted by Ehret et al. (2000). They applied a pattern of squares on the ball. Two different patterns were used, one with squares separated by shallow grooves and one with squares separated by deep grooves. A high-speed camera was used to capture single squares as they passed through the contact. Figure 7a shows an interferogram from the experiments with shallow grooves. The square pattern is still visible, but it is deformed and it can be shown that the lubricant films underneath each square interact with their neighboring squares. Figure 7b shows that deep grooves will make the squares dissociated from each other. Each square can be seen as a single asperity lubricated by the oil entrapped at the inlet of the contact. The difference in groove depth makes a great difference. The total load is carried by both squares and grooves in the case of shallow grooves, while only the squares carry the load in the case of deep grooves. The pressure on each square will therefore be higher in the case of deep grooves. Figure 7b also shows that each square is gradually emptied as the entrapped lubricant is forced into the deep groove. The film collapses and direct contact between the surfaces occurs in the dark areas along the right edge of each square. Figure 8 shows schematically the pressure distributions and how a single square is emptied as it moves through the contact. The conclusion to be drawn is that grooves or valleys play an important role in film formation. The film may collapse if excessively deep valleys are present and if these valleys are connected to the low pressure regions outside the contact.

Jacobson (1990, 2002) has presented an elegant theory about film breakdown due to asperities penetrating the lubricant film. In this theory sliding motion is seen as the detrimental factor. An asperity entering the EHL contact will be flattened already in the inlet and it will entrap a small amount of lubricant in the same way as described



EHL Film Thickness Behavior, Fig. 8 Schematic explanation of the difference between shallow and deep grooves

under section "Dynamic effects." This small amount of lubricant will, due to its high viscosity, remain between the surfaces the whole time it takes to travel from the contact's inlet to its outlet. If some small portion of sliding motion is superimposed there is a gradual emptying of the asperity contact as shown experimentally in Fig. 7b. Breakdown of the film will occur if the asperity contact is emptied, before it reaches the outlet and contact between the small asperities on the asperity contact surfaces occurs. Such small asperities will not be flattened and will penetrate the film if it is very thin. By making reasonable assumptions Jacobson showed that no breakdown occurs for ball bearing contacts if:

$$\frac{N(u_1 - u_2)}{u_2} < 2.66 \qquad (14)$$

where N is the number of (larger) asperities across the contact region and u_2 is the velocity of the slower surface. Equation (14) must not be used as a general rule but may act as a good rule of thumb.

The experimental study of EHL in with two real rough metallic surfaces requires other measurement techniques. Guangteng et al. (2000) used electric contact resistance measurements with good results. They were able to show how the lift-off speed (i.e., the transition between mixed and full film lubrication) varied with the surface roughness. The same technique was adopted by Masen et al. (2002), who showed that the lift-off speed was affected by the process used in the machining of the surfaces. A stone-honed surface, for example, gave a lift-off at lower speeds than ground or cross-honed surfaces. Lord and Larsson (2008) studied the influence of surface finishing method on running-in and lift-off. They showed that a chemically deburred surface with long smooth asperities gave a much less severe running-in in comparison to a ground surface. This is well in line with the theoretical findings on rough surface EHL.

More details about film formation in non-smooth contacts can be found in "► Stochastic Models for Rough Surface EHL", "► Deterministic Models of Rough Surface EHL."

Key Applications

Typical applications of the EHL film thickness behavior is in the design of rolling element bearings and gear transmissions. With good understanding about the mechanisms of film formation it is possible to optimize lubricant, component size, and surface finishing method for each specific application. By choosing a proper surface finishing method it will be possible to obtain full fluid film lubrication even if the surfaces have relatively high surface roughness height and lubricant viscosity is low.

Cross-References

- ► Deterministic Models of Rough Surface EHL
- ► EHL Governing Equations
- ► Elastohydrodynamic Lubrication (EHL)
- ► Film Thickness Formulas: Point Contacts
- ► Optical Interferometry
- ► Stochastic Models for Rough Surface EHL

References

C. Barus, Isothermals, isopiestics, and isometrics relative to viscosity. Am. J. Sci. **45**, 87–96 (1893)

P. Cann, E. Ioannides, B. Jacobson, A.A. Lubrecht, The lambda ratio – a critical re-examination. Wear **175**, 177–188 (1994)

F. Chevalier, A.A. Lubrecht, P.M.E. Cann, F. Colin, G. Dalmaz, Film thickness in starved EHL point contacts. ASME J. Tribol. **120**, 126–132 (1998)

B. Damiens, C.H. Venner, P.M.E. Cann, A.A. Lubrecht, Starved lubrication of elliptical EHD contacts. ASME J. Tribol. **124**, 105–111 (2004)

D. Dowson, D. Wang, An analysis of the normal bouncing of a solid elastic ball on an oily plate. Wear **179**, 29–38 (1994)

D. Dowson, G.R. Higginson, *Elastohydrodynamic Lubrication, the Fundamentals of Roller and Gear Lubrication* (Pergamon, Oxford, 1966)

P. Ehret, A. Felix-Quinonezi, J. Lord, R. Larsson, O. Marklund, Experimental analysis of micro-elastohydrodynamic lubrication conditions, in *Proceedings of the International Conference on Tribology*, ITC2000, Nagasaki, 2000

J.A. Greenwood, G.E. Morales-Espejel, The behavior of transverse roughness in EHL contacts. Proc. Inst. Mech. Eng. J. Eng. Tribol J **208**, 121–132 (1994)

G. Guangteng, P.M. Cann, A.V. Olver, H.A. Spikes, Lubricant film thickness in rough surface, mixed elastohydrodynamic contacts. ASME J. Tribol. **122**(n1), 65–76 (2000)

P.K. Gupta, H.S. Cheng, N.H. Forster, Viscoelastic effects in MIL-L-7808-typelubricant. Part 1: analytical formulation. STLE Tribol. Trans. **35**(2), 269–274 (1992)

B.J. Hamrock, *Fundamentals of Fluid Film Lubrication* (McGraw-Hill, New York, 1994)

B.J. Hamrock, D. Dowson, Isothermal elastohydrodynamic lubrication of point contacts, Part IV – starvation results. J. Lubr. Technol. **99**, 15–23 (1977)

B.J. Hamrock, D. Dowson, *Ball Bearing Lubrication – The Elastohydrodynamics of Elliptical Contacts* (Wiley-Interscience, New York, 1981)

Y.Z. Hu, D. Zhu, A full numerical solution to the mixed lubrication in point contacts. ASME J. Tribol. **122**, 1–9 (2000)
B. Jacobson, Mixed lubrication. Wear **136**(1), 99–116 (1990)
B. Jacobson, Nano-meter film rheology and asperity lubrication. ASME J. Tribol. **124**, 595–599 (2002)
M. Kaneta, H. Nishikawa, Local reduction in thickness of point contact EHL films caused by transversely oriented moving groove and its recovery. ASME J. Tribol. **116**, 635–639 (1994)
M. Kaneta, T. Sakai, H. Nishikawa, Optical interferometric observations of the effect of a bump on point contact EHL. ASME J. of Trib. **114**, 779–784 (1992)
M. Kaneta, T. Sakai, H. Nishikawa, Effects of surface roughness on point contact EHL. STLE Tribol. Trans. **36**(4), 605–612 (1993)
R. Larsson, E. Höglund, Elastohydrodynamic lubrication at impact loading. ASME J. Tribol. **116**, 770–776 (1994)
J. Lord, R. Larsson, Film-forming capability in rough surface EHL investigated using contact resistance. Tribol. Int. **41**, 831–838 (2008)
A.A. Lubrecht, C.H. Venner, Elastohydrodynamic lubrication of rough surfaces. Proc. Inst. Mech. Eng. J. Eng. Tribol. J **213**, 397–404 (1999)
M.A. Masen, C.H. Venner, P.M. Lugt, J.H. Tripp, Effects of surface microgeometry on the lift-off speed of an EHL contact. STLE Tribol. Trans. **45**(1), 21–30 (2002)
H. Moes, Optimum similarity analysis with applications to elastohydrodynamic lubrication. Wear **159**, 56–66 (1992)
G. Nijenbanning, C.H. Venner, H. Moes, Film thickness in elastohydrodynamically lubricated elliptic contacts. Wear **176**(2), 217–229 (1994)
N. Ren, D. Zhu, W.W. Chen, Y. Liu, Q.J. Wang, A three-dimensional deterministic model for rough surface line contact EHL problems. ASME J. Tribol. **131**, 1–9 (2009)
C.H. Venner, EHL film thickness computations at low speeds: risk of artificial trends as a result of poor accuracy and implications for mixed lubrication modeling. Proc. Inst. Mech. Eng. J. Eng. Tribol. J **219**, 285–290 (2005)
C.H. Venner, A.A. Lubrecht, Numerical simulation of a transverse ridge in a circular EHL contact under rolling/sliding. ASME J. Tribol. **116**, 751–761 (1994)
C.H. Venner, A.A. Lubrecht, Numerical analysis of the influence of waviness on the film thickness of a circular EHL contact. ASME J. Tribol. **118**, 153–161 (1996)
D. Zhu, On some aspects in numerical solution of thin-film and mixed EHL. Proc. Inst. Mech. Eng. J. Eng. Tribol. J **221**, 561–579 (2007)
D. Zhu, X. Ai, Point contact EHL based on optically measured three-dimensional rough surfaces. ASME J. Tribol. **119**(3), 375–384 (1997)

EHL for an Elastic Layer Bonded to a Rigid Body

► EHL of Coated Bodies

EHL for Rolling Element Bearings

► Rolling Bearing Lubricants

EHL Formulation

► EHL Governing Equations

EHL Governing Equations

Roland Larsson
Division of Machine Elements, Luleå University of Technology, Luleå, Sweden

Synonyms

EHL formulation; Equations for EHL

Definition

EHL governing equations are required to model the fluid–solid interaction between a lubricant and the bounding surfaces of an elastohydrodynamically lubricated contact

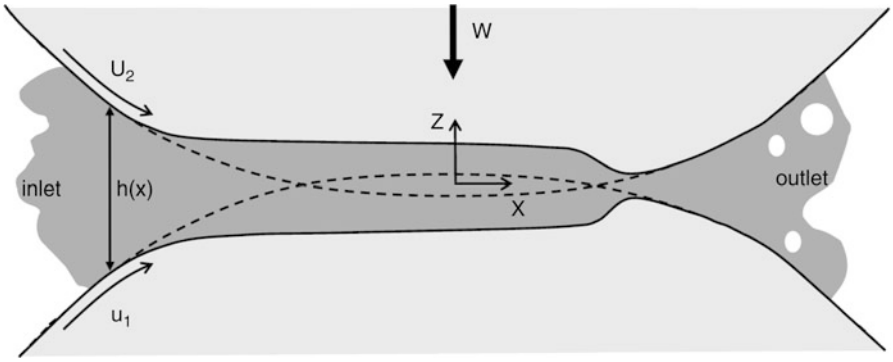
Scientific Fundamentals

Hydrodynamic lubrication is obtained when the gap between moving contact surfaces is filled with a fluid and is wedge shaped (i.e., the gap becomes narrower in the direction of motion). The fluid is entrained into the constriction between surfaces and a hydrodynamic pressure is formed in order to maintain fluid continuity. The wedge dimensions, the fluid film profile, play an important role in determining the distribution and magnitude of pressure in the fluid. This pressure may become large enough to deform the surfaces and thus also change the film profile. When the film profile has changed sufficiently to influence film pressure there is an interaction between fluid and solid and *elastohydrodynamic lubrication* (EHL) has occurred. See "► Elastohydrodynamic Lubrication (EHL)."

The equations required to model EHL will be described in two steps. First the simplest case, with a smooth cylinder in rolling/sliding contact with another cylindrical surface, and then the more general case.

Line Contact with Newtonian Fluid Model

Figure 1 shows the typical features of the film profile, the film thickness $h(x)$, between two smooth (surface roughness is assumed to be negligible) cylindrical surfaces moving at speeds of u_1 and u_2. The dashed lines indicate the original undeformed shapes. The fluid is forced into the gap by the moving surfaces and a constriction is created at the outlet in order to maintain continuity of



EHL Governing Equations, Fig. 1 Typical film profile for an EHL line contact

flow. High fluid pressure, $p(x)$, will occur due to the constriction and this pressure will be high enough to carry the load per unit length, w'. The thin film approximation is normally valid for this case and the continuity of flow can be accurately modeled by the *Reynolds equation* (Reynolds 1886) (see "► Reynolds Equation"):

$$\frac{\partial}{\partial x}\left(\frac{\rho h^3}{12\eta}\frac{\partial p}{\partial x}\right) = \frac{u_1 + u_2}{2}\frac{\partial(\rho h)}{\partial x} \tag{1}$$

where η and ρ denote fluid viscosity and density, respectively. The fluid film pressure is obtained from solutions of this partial differential equation. The film thickness profile can be described according to expression (2). The commonly used parabolic approximation has been applied to represent the original cylindrical geometry.

$$h(x) = h_0 + \frac{x^2}{2R} + \delta(x) \tag{2}$$

where $R^{-1} = R_1{}^{-1} + R_2{}^{-1}$ and R_1, R_2 are the radii of the cylinders, h_0 is a constant, and δ is the deformation of the surfaces. More correctly, δ describes the variation (in the x-direction) of the deformation, not the total deformation of the surfaces. The deformation is obtained from

$$\delta(x) = -\frac{4}{\pi E'}\int_{\text{inlet}}^{\text{outlet}} p(x')\ln|x - x'|d'; \tag{3}$$

Equation (3) describes the elastic deformation of two elastic half space surfaces in contact. It is valid only if the deformations are local near the contact zone (Johnson 1985). Other ways to model deformation are required for thin shell surfaces or cases where the global deformation of the bodies influences the film profile. The effective elastic modulus is defined by the elastic modulii and Poisson's ratios for the two surfaces:

$$E' = 2\left[\frac{\left(1 - v_1^2\right)}{E_1} + \frac{\left(1 - v_2^2\right)}{E_2}\right]^{-1} \tag{4}$$

The viscosity's variation with pressure can be modeled by the exponential Barus equation (Barus 1893):

$$\eta = \eta_0 \exp(\alpha p) \tag{5}$$

where α is the pressure-viscosity coefficient, a fluid property. The density's variation with pressure is modeled by the Dowson-Higginson expression (Dowson and Higginson 1966):

$$\rho = \rho_0\left(1 + \frac{0.6 \times 10^{-9} p}{1 + 1.7 \times 10^{-9} p}\right) \tag{6}$$

where ρ_0 is the density at atmospheric (ambient) pressure. The final equation describes load balance. The applied load must be balanced by the fluid film pressure:

$$w' = \int_{\text{inlet}}^{\text{outlet}} p dx \tag{7}$$

Equations (1) to (7) cannot be solved analytically and this nonlinear problem must be solved numerically (Dowson and Higginson 1966). The numerical problem and its solutions are described in (Venner and Lubrecht 2000).

The boundary conditions to (1) are normally $p(inlet) = p(outlet) = 0$ and $p \geq 0$ for all x. The position of the inlet must be set sufficiently far from the contact center, normally 2–4 times the contact half-width.

General Case with Special Attention to Elliptical Contacts

In the general case the following effects must be taken into account:

- Non-Newtonian rheology
- Non-smooth (rough) surfaces

- Contact of arbitrarily shaped three-dimensional bodies
- Heating of lubricant and surfaces
- Time-dependent problems

The general equations are derived here and exemplified by the special case of ellipsoid contact bodies. The first step is to define a general constitutive equation:

$$\tau_x = \eta_x \frac{\partial u}{\partial z} \tag{8}$$

$$\tau_y = \eta_y \frac{\partial v}{\partial z} \tag{9}$$

where u and v are the velocity fields in x- and y-directions and η_x and η_y are effective viscosities of any rheological model and, therefore, functions of shear rate, shear stress, and pressure, temperature. The effective viscosities in the x- and y-directions are uncoupled. A less general, still realistic model would be to couple η_x and η_y by writing

$$\tau_e = \frac{\eta(p,T)}{S(I_2)} I_2 \tag{10}$$

where η is the viscosity under zero shear stress/rate while S describes the non-Newtonian behavior and

$$\tau_e = \sqrt{\tau_x^2 + \tau_y^2} \tag{11}$$

$$I_2 = \sqrt{\left(\frac{\partial u}{\partial z}\right)^2 + \left(\frac{\partial v}{\partial z}\right)^2} \tag{12}$$

where I_2 is the second invariant of the strain rate tensor. For the Newtonian model, $S = 1$.

Now assume that the thin film approximation holds. The momentum equations become:

$$\frac{\partial p}{\partial x} = \frac{\partial \tau_x}{\partial z} \tag{13}$$

$$\frac{\partial p}{\partial y} = \frac{\partial \tau_y}{\partial z} \tag{14}$$

By combining (8) and (13) and integrating twice, the following expression for velocity will be obtained:

$$u(x,y,z) = \frac{\partial p}{\partial x} \int_0^h \frac{z}{\eta_x} dz + C_1 \int_0^h \frac{dz}{\eta_x} + C_2 \tag{15}$$

where the constants can be obtained from boundary conditions:

$$u(x,y,0) = u_1 \quad u(x,y,h) = u_2 \tag{16}$$

Here it assumed that the lubricant sticks to the surfaces. This might be wrong in some cases and stress boundary conditions should be applied instead. The final expressions for velocity and shear strain rate are:

$$\begin{aligned} u(x,y,z) = &\frac{\partial p}{\partial x}\left[f_{x1}(z) - f_{x0}(z)\frac{f_{x1}(h)}{f_{x0}(h)}\right] \\ &+ \frac{f_{x0}(z)}{f_{x0}(h)}(u_2 - u_1) + u_1 \end{aligned} \tag{17}$$

$$\frac{\partial u}{\partial z}(x,y,z) = \frac{1}{\eta_x}\left[\frac{\partial p}{\partial x}\left(z - \frac{f_{x1}(h)}{f_{x0}(h)}\right) + \frac{(u_2 - u_1)}{f_{x0}(h)}\right] \tag{18}$$

where

$$f_{x0}(z) = \int_0^z \frac{dz}{\eta_x} \quad f_{x1}(z) = \int_0^z \frac{z}{\eta_x} dz \tag{19}$$

By following the same route, velocity and shear strain rate in the y-direction will be:

$$\begin{aligned} v(x,y,z) = &\frac{\partial p}{\partial y}\left[f_{y1}(z) - f_{y0}(z)\frac{f_{y1}(h)}{f_{y0}(h)}\right] \\ &+ \frac{f_{y0}(z)}{f_{y0}(h)}(v_2 - v_1) + v_1 \end{aligned} \tag{20}$$

$$\frac{\partial v}{\partial z}(x,y,z) = \frac{1}{\eta_y}\left[\frac{\partial p}{\partial y}\left(z - \frac{f_{y1}(h)}{f_{y0}(h)}\right) + \frac{(v_2 - v_1)}{f_{y0}(h)}\right] \tag{21}$$

where

$$f_{y0}(z) = \int_0^z \frac{dz}{\eta_y} \quad f_{y1}(z) = \int_0^z \frac{z}{\eta_y} dz \tag{22}$$

The mass flow rates per unit length are obtained from:

$$\dot{m}_x = \int_0^h \rho u dz \quad \dot{m}_y = \int_0^h \rho \mathrm{v} \mathrm{d}z \tag{23}$$

By inserting (18) and (20) into (23), the expressions for mass flow rate will become:

$$\begin{aligned} \dot{m}_x &= \frac{\partial p}{\partial x}\left[G_{x1} - G_{x0}\frac{f_{x1}(h)}{f_{x0}(h)}\right] + \left(u_2 - u_1\right)\frac{G_{x0}}{f_{x0}(h)} + u_1 \rho_m h \\ \dot{m}_y &= \frac{\partial p}{\partial y}\left[G_{y1} - G_{y0}\frac{f_{y1}(h)}{f_{y0}(h)}\right] + \left(v_2 - v_1\right)\frac{G_{y0}}{f_{y0}(h)} + v_1 \rho_m h \end{aligned} \tag{24}$$

where

$$\begin{aligned} G_{x0} &= \int_0^h \rho f_{x0}(z) dz \;\; G_{x1} = \int_0^h \rho f_{x1}(z) dz \\ G_{y0} &= \int_0^h \rho f_{y0}(z) dz \;\; G_{y1} = \int_0^h \rho f_{y1}(z) dz \end{aligned} \tag{25}$$

$$\rho_m = \frac{1}{h}\int_0^h \rho dz \tag{26}$$

The next step is to apply the continuity equation:

$$\frac{\partial \rho}{\partial t} + \frac{1}{\partial x}(\rho u) + \frac{1}{\partial y}(\rho v) + \frac{1}{\partial z}(\rho w) = 0 \tag{27}$$

The continuity equation can be integrated across the film to:

$$\frac{\partial \dot{m}_x}{\partial x} + \frac{\partial \dot{m}_y}{\partial y} + \frac{\partial}{\partial t}(\rho_m h) = 0 \tag{28}$$

This is the generalized Reynolds equation (Dowson 1962; Peiran and Shizhu 1990) and can be rewritten to a more familiar form by combining (24) and (28):

$$\frac{\partial}{\partial x}\left(\varepsilon_x \frac{\partial p}{\partial x}\right) + \frac{\partial}{\partial y}\left(\varepsilon_y \frac{\partial p}{\partial y}\right) = \frac{\partial c_x}{\partial x} + \frac{\partial c_y}{\partial y} + \frac{\partial}{\partial t}(\rho_m h) \tag{29}$$

where

$$\begin{aligned} \varepsilon_x &= G_{x0}\frac{f_{x1}(h)}{f_{x0}(h)} - G_{x1} \quad \varepsilon_y = G_{y0}\frac{f_{y1}(h)}{f_{y0}(h)} - G_{y1} \\ c_x &= (u_2 - u_1)\frac{G_{x0}}{f_{x0}(h)} + u_1 \rho_m h \\ c_y &= (v_2 - v_1)\frac{G_{y0}}{f_{y0}(h)} + v_1 \rho_m h \end{aligned} \tag{30}$$

The integrals (19), (22), and (25) can be solved directly for the special case with Newtonian lubricant, no motion in the y-direction, no thermal effects and $\eta_x = \eta_y = \eta$. The Reynolds equation becomes:

$$\frac{\partial}{\partial x}\left(\frac{\rho h^3}{12\eta}\frac{\partial p}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\rho h^3}{12\eta}\frac{\partial p}{\partial y}\right) = \frac{(u_2 + u_1)}{2}\frac{\partial(\rho h)}{\partial x} + \frac{\partial}{\partial t}(\rho h) \tag{31}$$

which can be compared with the stationary line contact version, (1).

A general expression for film thickness can be written as:

$$\begin{aligned} h(x, y, t) = h_{00}(t) + g(x, y, t) - Z_1(x, y, t) \\ - Z_2(x, y, t) + \delta(x, y, t) \end{aligned} \tag{32}$$

where h_{00} is a function of time only, g describes the original shape of the contact bodies, Z_1 and Z_2 describe surface roughness height of both surfaces, and δ the combined deformation of surfaces. The shape function g for ellipsoidal contact bodies becomes:

$$g(x, y) = \frac{x^2}{2R_x} + \frac{y^2}{2R_y} \tag{33}$$

where $R_x^{-1} = R_{x1}^{-1} + R_{x2}^{-1}$ and $R_y^{-1} = R_{y1}^{-1} + R_{y2}^{-1}$ are the reduced radii of curvature in both directions. The deformation can be obtained from any structural mechanics model, for example, a finite element model. For ellipsoidal contact bodies it is, however, common to assume an elastic half-space model of the same type as (3). In two dimensions it is (Johnson 1985):

$$\delta(x, y) = \frac{2}{\pi E'}\int\int \frac{p(x', y')dx'dy'}{\sqrt{(x - x')^2 + (y - y')^2}} \tag{34}$$

Rheological models, such as non-Newtonian (shear dependency) and viscoelasticity models, are described elsewhere. It is, however, worth mentioning a number of Newtonian viscosity models that are more general than (5). The relationship between viscosity and pressure is normally described by an exponential expression

$$\eta = \eta_0 e^{f(p)} \tag{35}$$

where $f(p)$ is αp in (5) and

$$f(p) = (\ln \eta_0 + 9.67)\left[-1 + \left(1 + 5.1 \times 10^{-9} p\right)^Z\right] \tag{36}$$

for the more accurate Roelands model (Roelands 1966) where Z is a constant. The Roelands model depicts a less exponential increase in viscosity with pressure than the Barus model, but Bair (Bair 1993) has shown that this is true only up to approximately 0.5 GPa. Above 0.5 GPa, the viscosity increases even more rapidly with the pressure than described by the Barus model. Bair et al. (1998) describe an isothermal free volume model that is also accurate at very high pressures:

$$f(p) = Bc\left(\frac{1}{V/V_0 - c} - \frac{1}{1 - c}\right) \tag{37}$$

where

$$V/V_0 = 1 - \frac{1}{K'_0 + 1}\ln\left(1 + p\frac{(1 + K'_0)}{K_0}\right) \tag{38}$$

and K_0 is the bulk modulus, K'_0 is the pressure rate of change of the bulk modulus, B is a constant and $c = V_{occ}/V_0$, the ratio between the occupied volume and the volume at zero pressure. This model has physical foundations and it also couples viscosity and compressibility to each other. The disadvantage is the requirement that several additional constants must be determined for each single lubricant. A free volume model that also takes temperature into account was suggested by Yasutomi et al. (1984). Eight constants have to be determined in order to use that model, but when this is accomplished, it gives

a good description of the viscosity-pressure-temperature relationship. There are a few more viscosity models but is important to remember that EHL film thickness is dominated by the entraining effect in the inlet zone, where pressure is still relatively low and the chosen viscosity model does not influence the average central film thickness very much. Viscosity-pressure-temperature data for real lubricants of different type can be found in (Larsson et al. 2000). The same source presents alternatives to the density-pressure relationship, (6). A thourough description of lubricant rheology is given by Bair (Bair 2009) See "► Temperature and Pressure Dependence of Viscosity."

There is also a need for a more general expression for the force balance. Inertia may play an important role in the time-dependent case and Newton's second law can be applied as:

$$m\ddot{z} = \int\int p(x, y, t)dxdy - w(t) \tag{39}$$

where m is the mass of the body that impacts the other (not moving) body and w is the applied load.

For the thermal problem, the governing equation, the energy equation, reads:

$$c_p \frac{\partial}{\partial t}(\rho T) + c_p \frac{\partial}{\partial x}(\rho u T) + c_p \frac{\partial}{\partial y}(\rho v T) + c_p \frac{\partial}{\partial z}(\rho w T) = \frac{\partial}{\partial z}\left(\lambda \frac{\partial T}{\partial z}\right) + \tau_{xz}\frac{\partial u}{\partial z} + \tau_{yz}\frac{\partial v}{\partial z} + \varepsilon T\left[\frac{\partial p}{\partial t} + u\frac{\partial p}{\partial x} + v\frac{\partial p}{\partial y}\right] \tag{40}$$

where $T(x,y,z)$ is the three-dimensional temperature field inside the oil film, c_p is the lubricant specific heat, τ_{xz} and τ_{yz} are shear stresses in the film parallel to the surfaces, and ε is the thermal expansion coefficient. The same equation must also be solved for the bodies in order to take the heat conduction into account; for more details see "► Thermal EHL theory."

It is important to note that the generalized Reynolds equation (Dowson 1962; Peiran and Shizhu 1990), (29), must be applied when the full effect of a three-dimensional fluid film temperature field $T(x,y,z)$ should be taken into account. Temperature will influence lubricant viscosity and density.

Finally, a few words about the validity of the equations presented above. The fundamental requirement is that the thin film assumption holds. That means that the hydrodynamic wedge must not be too steep. A useful measure is the ratio between film thickness and wedge length. When this number is too big there is a need to use Stokes or even Navier–Stokes equations to model the flow in the lubricant film. This situation occurs, for example, when short wavelength surface roughness must be taken into account. For more information, see, for example, (Odyck and Venner 2003) and (Almqvist and Larsson 2004).

Key Applications

Typical applications of the EHL governing equations are numerical simulations of the behavior of line and point contacts under different operation conditions; see "► EHL, Full Numerical Solution Methods" for details. These simulations have helped researchers to understand the film formation mechanisms, that is, how the film profile varies under different conditions. The understanding about film formation is important since friction and wear are strongly influenced by the thickness of the lubricating film. Surface roughness has a large effect on film formation and if the roughness is properly chosen it is possible to promote transition to full film lubrication. Simulations based on solutions of the EHL governing equations help to determine positive features of the surface's roughness. Contact pressure and its distribution are also obtained from simulations. Pressure is important in order to determine the subsurface stresses and to predict wear and failure risk. Lubricant shear stress, and thus also contact friction, may be obtained from (8) and (9). It is, however, very important to apply realistic non-Newtonian rheological and thermal models, otherwise friction will be enormously overestimated.

Cross-References

► EHL, Full Numerical Solution Methods
► Elastohydrodynamic Lubrication (EHL)
► Reynolds Equation
► Temperature and Pressure Dependence of Viscosity
► Thermal EHL Theory

References

T. Almqvist, R. Larsson, Some remarks on the validity of Reynolds equation in the modeling of lubricant film flows on the surface roughness scale. ASME J. Tribol. **126**(4), 703–710 (2004)

S. Bair, A note on the use of Roelands equation to describe viscosity for EHD Hertzian zone calculations. ASME J. Tribol. **115**(2), 334 (1993)

S. Bair, Rheology and high-pressure models for quantitative elastohydrodynamics. Proc. Instn. Mech. Engrs., J. Eng. Tribol. **223**, 617–628 (2009)

S. Bair, M. Khonsari, W.O. Winer, High pressure rheology of lubricants and limitations of the Reynolds equation. Tribol. Int. **31**(10), 573–586 (1998)

C. Barus, Isothermals, isopiestics, and isometrics relative to viscosity. Am. J. Sci. **45**, 87–96 (1893)

D. Dowson, A generalised Reynolds equation for fluid film lubrication. Int. J. Mech. Eng. Sci. **4**, 159–170 (1962)

D. Dowson, G.R. Higginson, *Elastohydrodynamic Lubrication, the Fundamentals of Roller and Gear Lubrication* (Pergamon, Oxford, 1966)
K.L. Johnson, *Contact Mechanics* (Cambridge University Press, Cambridge, UK, 1985)
R. Larsson, P.O. Larsson, E. Eriksson, M. Sjöberg, E. Höglund, Lubricant properties for input to hydrodynamic and elastohydrodynamic lubrication. Proc. Instn. Mech. Engrs., J. Eng. Tribol. **214**, 17–27 (2000)
D.E.A. Odyck, C.H. Venner, Stokes flow in thin films. ASME J. Tribol. **125**, 121–134 (2003)
Y. Peiran, W. Shizhu, Generalized Reynolds equation for non-Newtonian thermal elastohydrodynamic lubrication. ASME J. Tribol. **112**(4), 631–636 (1990)
O. Reynolds, On the theory of lubrication and its application to Mr. Beauchamp Tower's experiments, including an experimental determination of the viscosity of olive oil. Philos. Trans. R. Soc **177**, 157 (1886)
C.J.A. Roelands, *Correlational Aspects of The Viscosity-Temperature-Pressure Relationship Of Lubricating Oils* (Druk, V.R.B, Groningen, the Netherlands, 1966)
C.H. Venner, A.A. Lubrecht, *Multi-Level Methods in Lubrication* (Elsevier, Amsterdam, 2000)
S. Yasutomi, S. Bair, W.O. Winer, An application of a free volume model to lubricant rheology. ASME J. Tribol. **106**(2), 291–303 (1984)

EHL History (Elastohydrodynamic Lubrication)

Dong Zhu[1], Q. Jane Wang[2]
[1]State Key Laboratory of Mechanical Transmission, Chongqing University, Chongqing, People's Republic of China
[2]Department of Mechanical Engineering and Center for Surface Engineering and Tribology, Northwestern University, Evanston, IL, USA

Synonyms

History of EHL development; History of elastohydrodynamics

Definition

Elastohydrodynamic lubrication (EHL) is a mode of fluid-film lubrication in which hydrodynamic action is significantly enhanced by surface elastic deformation and lubricant viscosity increases due to high pressure. This article briefly reviews its history of development and overlooks its future prospect.

Scientific Fundamentals

Background and Early Studies (1886–1940s)

It has long been well known that lubrication is vital to efficiency and durability of machine elements, and one of the best ways to improve performance and life is to form lubricant films that separate the contacting surfaces. Various types of oils and greases have been used for the purpose of lubrication for thousands of years. However, in the past, lubrication mechanisms were not well understood. Prediction of lubrication performance was first successful for conformal contact problems. In 1886, O. Reynolds published his milestone theoretical lubrication analysis based on journal bearing experiments. The Reynolds equation was derived and it has been the foundation of hydrodynamic lubrication theory ever since. Good agreement was obtained between analyses and experiments for some conformal contact components, such as fluid-film bearings, in which the hydrodynamic pressure is low, typically on the order of 1∼10 MPa or less.

Attempts were made later to extend the application of hydrodynamic lubrication theory to non-conformal contact components, such as gears and rolling element bearings, in which the maximum Hertzian contact pressure may reach 1∼4 GPa. One of the remarkable early studies was by Martin, who in 1916 published his lubrication analysis for line contacts in spur gears. A pair of gear teeth was simplified to two parallel, smooth, rigid cylinders lubricated by an incompressible isoviscous Newtonian fluid. Using a simplified Reynolds equation, he derived an expression of loading capacity. It was found, however, that the predicted lubricant film thickness between gear teeth was extremely small, disagreeing with engineering observations. This somewhat discouraged further effort over the next 10–20 years. But, indeed, Martin's work was a good beginning in the study of lubrication for non-conformal contact components.

Beginning in the 1930s, researchers tried to improve the hydrodynamic lubrication analyses for non-conformal contacts by including either the effect of localized elastic deformation of the two surfaces or that of the lubricant viscosity increase due to high pressure. In 1949, Grubin published his paper on "Fundamentals of the Hydrodynamic Theory of Lubrication of Heavily Loaded Cylindrical Surfaces," which was the first to take into account simultaneously the effects of both elastic deformation and viscosity increase. It is said that Grubin's theory was based on Ertel's preliminary results obtained as early as in 1939. Due to the lack of computing power at that time, a simplified approximate solution was developed based mainly on the following two assumptions:

1. The shape of the elastically deformed cylindrical bodies in a heavily loaded lubricated contact is the same as that produced in a dry contact;
2. The hydrodynamic pressure approaches infinity at the inlet border of the Hertzian contact zone.

The geometric shape of the gap could be calculated, therefore, by an analytical solution from the Hertzian theory for dry contacts. The viscosity also approaches infinity at the inlet border according to the following pressure-viscosity relationship:

$$\eta = \eta_o e^{\alpha p} \tag{1}$$

Based on the above, Grubin numerically calculated the integral of a simplified Reynolds equation in the inlet zone and then curve-fitted his results in a range of central film thickness value reasonable for practical applications. The following expression was then successfully derived for predicting the lubricant film thickness:

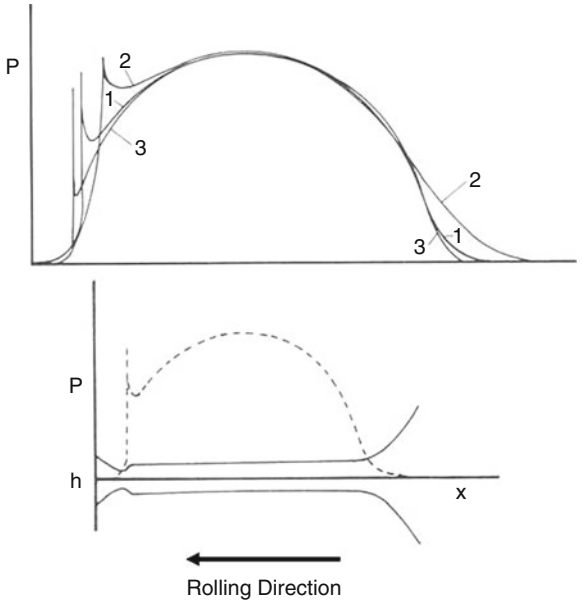
$$H_c = h_c/R_x = 1.95(G^* U^*)^{8/11}/W^{*1/11} \tag{2}$$

The Ertel-Grubin theory well describes the basic characteristics of line contact EHL, e.g., a constant film thickness in the contact zone, and a pressure distribution close to Hertzian. Although the assumptions were heroic and the theory was approximate, the film thickness results predicted by (2) have been found to be in reasonably good agreement with experimental data especially under heavy loading conditions.

Establishment of Fundamental EHL Theories (1950–1970s)

Full numerical solutions having the above-mentioned two assumptions removed for line contact problems were given much attention in 1950s and1960s. The first successful solution was published in 1951 by Petrusevich, who presented three cases in detail for different speeds but the same load, demonstrating all the typical EHL characteristics for the first time, including a nearly constant film thickness and an EHL pressure distribution close to the Hertzian over the majority of contact zone, a film constriction downstream near the outlet, and, especially, a high pressure spike at the outlet side immediately before the film constriction, which was later named the "Petrusevich spike" (see Fig. 1). Also, the three film thickness values he presented were close to those predicted by currently used formulae developed much later. Based on his limited results, he somehow derived a film thickness formula, which quite correctly reflected the relationship between the film thickness and the speed, but showed a small film thickness increase with increasing load, which appeared to be difficult for people to understand at that time.

Shortly after Petrusevich, Dowson and Higginson presented their milestone paper, "A Numerical Solution to the Elastohydrodynamic Problem," in 1959. They developed a new solution approach, called inverse solution, to overcome difficulties associated with slow numerical convergence. It appeared to be able to handle heavily loaded cases and get a converged solution within a small number of calculation cycles, although the computation was not fully automatic. A curve-fitting formula for predicting line contact EHL minimum film thickness was presented by Dowson and Higginson in 1961. It can be expressed as follows:



EHL History (Elastohydrodynamic Lubrication), Fig. 1 Line contact EHL solutions by Petrusevich (1951)

$$H_m = h_m/R_x = 1.6\, G^{*0.6} U^{*0.7} W^{*(-0.13)} \tag{3}$$

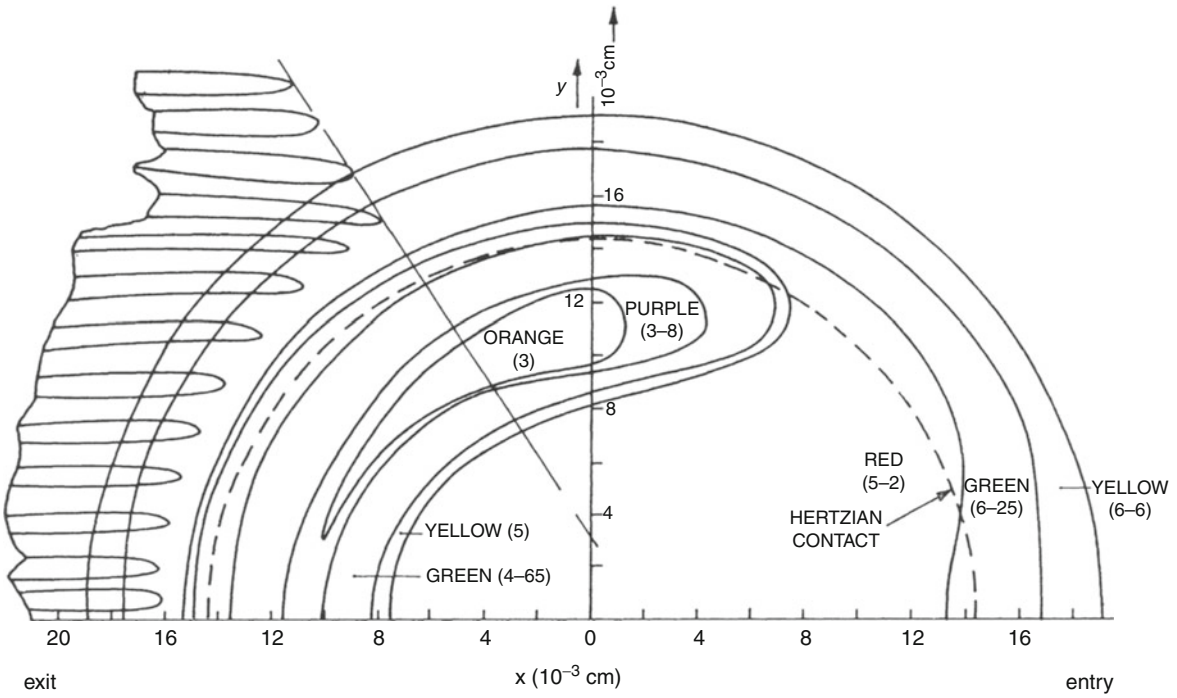
This formula was further modified by Dowson in 1965 as follows (see Dowson and Higginson 1966):

$$H_m = h_m/R_x = 2.65\, G^{*0.54} U^{*0.7} W^{*(-0.13)} \tag{4}$$

Later, central film thickness formulae were also derived based on full numerical solutions. The following is one by Dowson and Toyoda (1978):

$$H_c = h_c/R_x = 3.06\, G^{*0.56} U^{*0.69} W^{*(-0.10)} \tag{5}$$

In these formulae, four dimensionless parameters are used for line contact problems: speed parameter $\boldsymbol{U}^*$, load parameter $\boldsymbol{W}^*$, materials parameter $\boldsymbol{G}^*$, and film thickness parameter $\boldsymbol{H}$ (see Nomenclature, below, for their definitions). Note that for line contacts there should be only three independent parameters, and the four given above are actually interrelated. However, these parameters are easy to use, making physical sense explicitly, so they have been widely accepted. Some other dimensionless



EHL History (Elastohydrodynamic Lubrication), Fig. 2 Optical interference fringes by Gohar and Cameron (1966), for an EHL circular contact

parameter groups have also been used, however, they will not be discussed here.

Parallel to the analytical studies, experimental investigations also showed fruitful results. Early studies since 1950s were focused mainly on the line contact EHL film thickness measurements on disc/roller machines with capacitance technique (e.g., Crook 1961–1963; Dyson et al. 1966) and x-ray transmission method (Sibley and Orcutt 1961). The basic trends in the EHL were confirmed experimentally, e.g., the film thickness is significantly affected by the rolling speed, but the load effect is nearly negligible. Measured film thickness results were found to be in reasonably good agreement with (2–5). In addition, EHL pressure distribution was measured with thin-film transducers applied onto disc specimens by Kannel (1966), and Hamilton and Moore (1971), and others. The Petrusevich spike was observed experimentally.

Note that the capacitance technique was used to measure an average or central film thickness, while the x-ray could give approximate estimates of minimum film thickness. No detailed information about the shape of EHL film could be provided until optical interferometry, which was originally developed by Gohar and Cameron (1963–1966) (a sample optical contour map is shown in Fig. 2), and further modified by Foord et al. (1969–1970), and others. With a super-finished steel ball against a glass disc, one could observe the film thickness distribution through interference fringes under a microscope. A remarkable new finding was that, in such circular contact, the film constriction takes on a horseshoe shape and the minimum film thickness is actually located on two sides away from the centerline. Due to its great accuracy and capability to provide detailed mapping of film thickness, optical interferometry has been a major experimental means in fundamental EHL research since then. Its limitations include requirements of the use of super-finished transparent disk and highly reflective ball, etc.

Simplified inlet analyses of Grubin's type for point contact problems were developed by Archard and Cowking (1966), and Cheng (1970), about 15–20 years later than that by Grubin for line contacts. Full numerical solutions for point contacts did not appear until 1975–1976, more than 10 years behind the successful experimental studies and 20–25 years later than the full solutions of line contact. This is because additional computing capacity needed for point contacts demanded significantly more powerful digital computers, which were not widely available to researchers earlier. Ranger et al. (1975), presented the first full solution from a straightforward iterative procedure, numerically demonstrating the typical

point contact EHL characteristics and confirming experimental observations from optical interferometry for the first time. It was questionable, however, that their results showed an increasing film thickness with increasing load. It should be noted that Petrusevich and Ranger et al. were the first to present the full numerical solutions in line and point contacts, respectively, but full credit was not given to them for the same reasons: Both studies showed slight film thickness increase with increasing load in the parameter ranges they analyzed. Today it is understood that the film thickness may first increase then decrease, if the load is continuously increasing over an extended wide range.

Shortly after Ranger et al. Hamrock and Dowson (1976a, b, 1977a), published a series of papers systematically investigating the effects of speed, load, materials properties, contact ellipticity, and lubricant starvation on central and minimum film thicknesses in elliptical contacts through full numerical solutions. The following curve-fitting formulae were derived for point contact problems (1977a):

$$H_c = 2.69G^{*0.53}U^{*0.67}W^{*-0.067}\left(1-0.61e^{-0.73k}\right) \quad (6)$$

$$H_m = 3.63G^{*0.49}U^{*0.68}W^{*-0.073}\left(1-e^{-0.68k}\right) \quad (7)$$

These formulae use dimensionless parameters nearly the same as those in (2, 3, 4, 5), except that the load parameter is slightly different and a parameter of contact ellipticity, $\boldsymbol{k = b/a}$, is added to take into account the effect of point contact geometry. Comparative studies were conducted later by different researchers. It was found by Koye and Winer (1981), that the discrepancies in film thickness between formula predictions and optical interferometry results were about 30% as an average under given testing conditions. In addition to (2, 3, 4, 5, 6, 7), there have been some other formulae published, and comparison of their accuracy is a complicated topic. Generally, however, these formulae have been found to be practically acceptable in engineering practice, because the film thickness is dominated by the lubricant entraining action in the inlet zone, where the gap is still large and the effects of thermal and non-Newtonian behaviors and surface roughness are often still limited. Therefore, the isothermal analyses developed in early years based on the Newtonian fluid and smooth surface assumptions are in many cases still acceptable.

Significant Development on Various Subjects

The development of EHL theory and practice has been prosperous and many significant contributions have been presented since 1960s. A brief review such as this one can only give a snapshot, citing a very small portion of published papers and focusing on a limited number of topics as follows.

Starvation Effect

Insufficient lubricant supply due to various reasons may cause a condition called "starvation," which may significantly affect EHL film formation. Early attempts began with optical interferometry experiments, because the lubricant supply could be readily quantified under a microscope with the distance from the meniscus inlet boundary of EHL film to the center of contact. Pioneer studies include those by Wedeven et al. (1971), and Chiu (1974), who well defined the starvation problem, and used the meniscus position as a criterion of starvation severity. Early analytical investigations were conducted by Wolveridge et al. (1971), for line contacts and Hamrock and Dowson (1977b), for elliptical contacts, and others, using the inlet distance as an input parameter in predicting the film thickness. Various film thickness reduction formulae have been obtained through curve-fitting based on either experimental or analytical results. It was found that the basic trends from those formulae are in good agreement, but quantitative differences still exist. One of the reasons for the differences is probably that, in numerical simulations, a straight line inlet boundary was assumed while, in the experiments, the boundary was found to be more complicated. More recently efforts have been made in order to consider realistic conditions, e.g., the analytical studies by Chevalier et al. (1998), and others.

Thermal EHL

Thermal behavior is an important subject, as significant temperature increase can negatively affect EHL performance due to reduced viscosity, and possibly lead to EHL film breakdown and early failure. Pioneering studies on thermal EHL in line contacts were presented by Cheng and Sternlicht (1964) and Dowson and Whitaker (1966), coupling the energy equation with other EHL equations to solve for temperature variations in the film. A full solution of thermal EHL in point contacts was later presented by Zhu and Wen (1984). It was found that the temperature increase could be significant in the contact zone, but relatively small in the inlet zone, where entraining action actually dominates the EHL film formation. That is why the effect of temperature rise caused by sliding on film thickness is often limited, except at extremely high speeds, which may result in significant heating due to increased lubricant shear rate and possible reverse flows in the inlet zone. The thermal reduction of film thickness due to inlet

heating was studied by Cheng (1967), Murch and Wilson (1975), and others, mainly through simplified inlet solutions of Grubin's type. A more comprehensive prediction was later presented by Gupta et al. (1992):

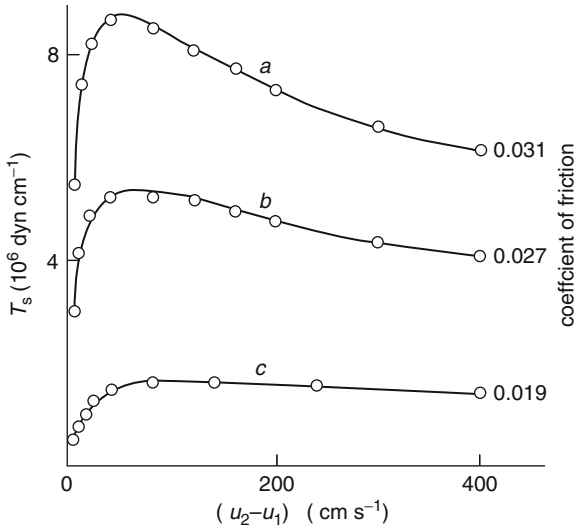
$$C_T = \frac{1 - 13.2\left(\frac{P_h}{E'}\right)L^{0.42}}{1 + 0.213(1 + 2.23S^{0.83})L^{0.64}} \qquad (8)$$

where C_T is the thermal reduction factor for film thickness, $L = \eta_o \beta U^2/k_f$, the thermal loading parameter, and $S = (u_2 - u_1)/U$, the slide-to-roll ratio Improved viscosity and density models have recently been employed, and the prediction further modified by Kumar et al. (2010).

Temperature measurement in a tiny EHL contact is a challenging task. There have been mainly two techniques: (1) measurement with thin-film transducers deposited onto specimens in a rolling-sliding contact (Orcutt 1965; Kannel and Bell 1972; and others); and (2) detection of infrared radiation on a device similar to that of the optical interferometry (Turchina et al. 1974 and others).

Friction/Traction in EHL

EHL friction, sometimes called traction, is of great importance as it is directly associated with machine component performance, efficiency, and energy consumption. For hydrodynamic lubrication, in which pressure is relatively low and lubricant film is thick so that lubricant shear strain rate in the film is low, commonly used industrial lubricants can be considered as Newtonian fluids and friction prediction is relatively simple. For the EHL, however, the frictional mechanism becomes more complicated. Early experimental studies (e.g., by Crook 1963, as shown in Fig. 3) revealed that measured friction is usually much lower than predicted with Newtonian fluid models. A new concept of limiting shear stress was then established in further studies by Plint (1967–1968), and Johnson and Cameron (1967–1968), and others. Basically, in the inlet zone of an EHL contact, where the entraining action actually dominates the lubrication formation, the pressure is relatively low and the gap is large, so that the shear rate is still low and the Newtonian models may still be acceptable. That is why the Newtonian models can be used successfully to solve for EHL film thickness. However, a vast majority of sliding friction is generated in the contact zone, where the lubricant probably passes through in a fraction of a mini-second, the pressure is high, and the film thickness tiny, resulting in a lubricant shear strain rate possibly as high as $10^7 \sim 10^8$ 1/s. Under such conditions the lubricant can no longer be considered as Newtonian. When the sliding increases, the shear stress may increase but cannot go beyond the limit. The limiting shear stress is found to be a property of lubricant, and also a function of pressure and temperature.



EHL History (Elastohydrodynamic Lubrication), Fig. 3 EHL friction measured by Crook (1963), as a function of load and sliding speed. Load (10^7 dyn/cm): (**a**) 20, (**b**) 15, (**c**) 7.5. T_s: measured traction/friction

Modeling friction should describe the non-Newtonian viscous-elastic characteristics of lubricants, considering lubricant shear due to both elastic and viscous behaviors under the high-pressure high-shear transient conditions stated above. The following Maxwell model is so far widely accepted:

$$\dot{\gamma} = \dot{\gamma}_e + \dot{\gamma}_v = \frac{1}{G}\frac{d\tau}{dt} + F(\tau) \qquad (9)$$

The viscous term in (9) can be expressed as follows according to Johnson and Tevaarwerk (1977):

$$F(\tau) = \frac{\tau_L}{\eta}\sinh\left(\frac{\tau}{\tau_L}\right) \qquad (10)$$

or by Bair and Winer (1978):

$$F(\tau) = -\frac{\tau_L}{\eta}\ln\left(1 - \frac{\tau}{\tau_L}\right) \qquad (11)$$

Recently, a modified Carreau model has also been used due to its convenience in application with the shear stress as an independent variable (see Bair and Khonsari 1996). More efforts on modeling EHL friction are ongoing. Fortunately, in engineering practice friction is often relatively easy to evaluate experimentally. Based on friction test data, one can estimate the limiting shear stress of a lubricant under given conditions of contact pressure and temperature.

More Rheology-Related Issues

Defining lubricant rheology under EHL conditions is a challenging task, because it is difficult, if not impossible, to reproduce such transient high-pressure, high-shear strain rate conditions in a laboratory outside the tiny EHL contact zone. Viscosity at the inlet may noticeably affect the EHL film thickness analysis while lubricant shear characteristics determine the traction and thermal behaviors of an EHL interface. Although transient lubricant properties under high pressure, high shear rate, and varying temperature are difficult to obtain, certain empirical relationships to correlate viscosity with pressure and temperature, as well as to describe other physical phenomena such as "shear thinning," have assisted the advancement of EHL modeling. A thorough review of rheology research is beyond the scope of this article. Readers may find in-depth coverage of rheology for EHL from Bair (2007). Here, only several commonly used viscosity equations are listed.

Barus (1893), reported the viscosity data of marine glue as a function of average pressure in a linear model. However, the exponential relation (1), more well-known in tribology as the Barus equation, is widely used in EHL analyses due to its simplicity and capture of a certain nonlinear pressure-viscosity behavior. This exponential relationship was confirmed with various mineral oils in the 1930s and 1940s, and also well documented for more than 40 lubricants in an ASME report, 1953. According to Cameron (1966), the "Barus" law describes the viscosity behavior quite well up to a pressure range of 200–400 MPa at low temperatures. Beyond this limit, Cameron proposed a power function, employing two constants, θ and n, to gain more flexibility:

$$\eta = \eta_o(1 + \theta p)^n \tag{12}$$

In the meantime, several viscosity models were presented by Roelands in (1966). The following Roelands equation is often seen as an improved pressure-viscosity relationship used in EHL analyses with z as the pressure-viscosity index:

$$\eta = \eta_0 e^{\left[(\ln \eta_0 + 9.67)\left(\left(1+\frac{p}{p_0}\right)^z - 1\right)\right]} \tag{13}$$

For the pressure in excess of 1 GPa, (1, 12, 13) might considerably overestimate the viscosity. Doolittle (1951), explored the relationship between viscosity and the fractional free volume using an exponential function and developed the first free-volume viscosity model, showing that the resistance to flow depends on the relative volume of molecules present per unit of free volume. Based on this, improved free-volume viscosity models were developed, and the one presented by Cook et al. (1993), is given below.

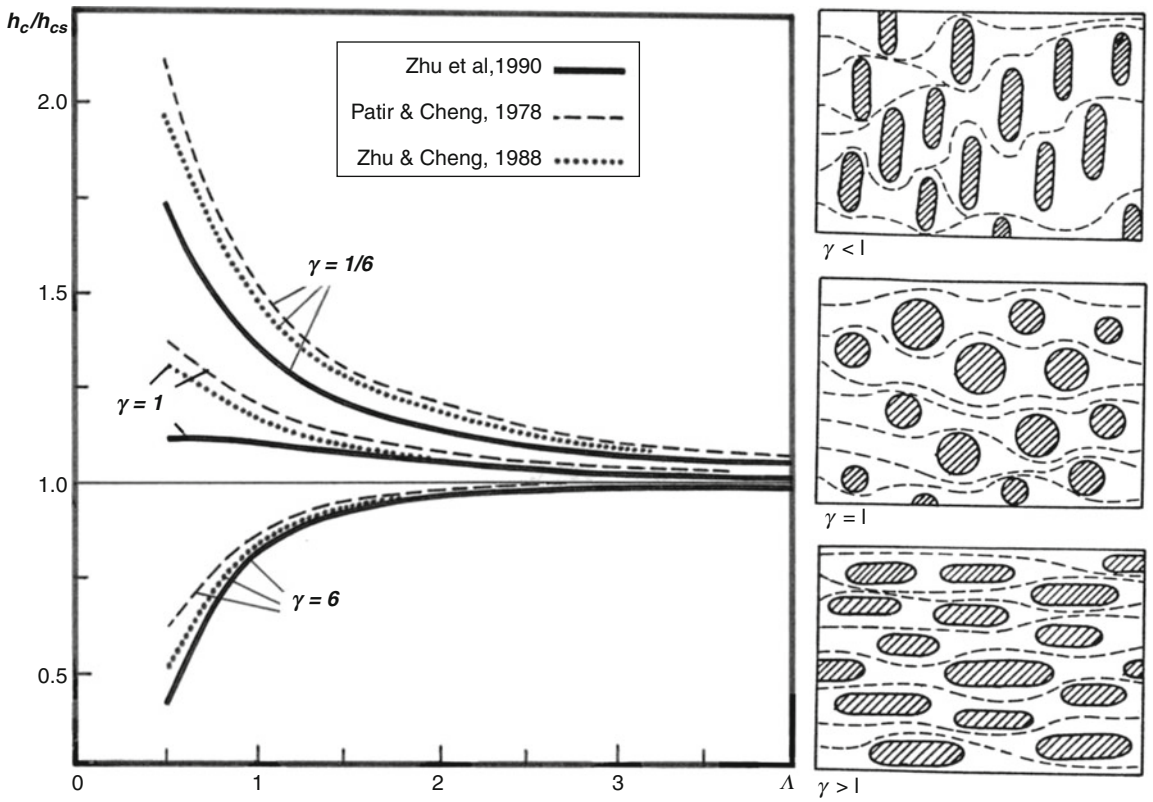
$$\eta = \eta_o \exp\left\{ B \frac{V_{occ}}{V_0} \left[\frac{1}{\frac{V}{V_0} - \frac{V_{occ}}{V_0}} - \frac{1}{1 - \frac{V_{occ}}{V_0}} \right] \right\} \tag{14}$$

where V is the volume, V_0, the volume at ambient pressure, V_{occ} the occupied volume, and B the Doolittle parameter. So far, the free-volume model seems to be able to yield EHL simulation results closest to experimental data.

Roughness Effect

In engineering practice, no surface is ideally smooth, and roughness is often of the same order of magnitude as, or greater than, the film thickness estimated by the smooth surface EHL theory. Effects of surface roughness and topography, therefore, ought to be taken into account in engineering applications. Great efforts have been made since 1970s, and there have been mainly two types of rough surface EHL models: stochastic and deterministic. Early studies mainly employed stochastic models, using a few statistic parameters to describe the surface and lubrication characteristics. Among various models published, the one by Patir and Cheng (1978a), for line contacts has enjoyed wide recognition. It employed an average Reynolds equation derived by Patir and Cheng (1978b), for hydrodynamics and a load-compliance relation given by Greenwood and Tripp (1970–1971), based on a simplified stochastic contact model. Obtained solutions showed that transverse roughness may lead to a significant increase in film thickness, while longitudinal roughness may cause a film reduction (see Fig. 4). This approach was extended to point contacts by Zhu and Cheng (1988).

The stochastic models, however, could only estimate the overall performance and approximate average values. Parameter variations and localized details, such as the maximum/minimum values (which may be critical for studies on lubrication breakdown and failures), were missed. Recently, more attention has been given to deterministic approaches. Early deterministic models employed mainly artificial roughness, such as sinusoidal waves and irregularities of simple geometry. More realistic two-dimensional machined or random roughness was used by Venner (1991), and Kweh et al. (1992), and others. However, full-scale point contact EHL solutions utilizing digitized three-dimensional machined roughness did not appear until Xu and Sadeghi (1996), Zhu and Ai (1997), and others. Since, in reality, the rough surface topography is usually three-dimensional, although the macro contact geometry may be two-dimensional, a three-dimensional



EHL History (Elastohydrodynamic Lubrication), Fig. 4 Roughness orientation effect predicted by stochastic models (From Zhu et al. 1990)

line contact deterministic model is needed. It has recently been developed by Ren et al. (2009).

Generally, the effects of surface roughness and orientation on the EHL film thickness predicted by the deterministic models are not as great as those predicted by the stochastic models. For line contacts, the basic trends presented by Ren et al. are the same as those by Patir and Cheng, but quantitatively the influences appear to be relatively mild. For point contacts, the effects of roughness and orientation may be more complicated. For example, in a circular contact, the transverse roughness may possibly yield a thinner film than the longitudinal due to significant lateral flows that can be enhanced by the transverse roughness but may have a negative influence on the EHL film formation. So far there seems to be no systematic study found in literature on this topic over a wide range of operating conditions considering various types of contact geometry and roughness orientation.

The importance of surface roughness effect on the lubrication performance and components\ life was recognized as early as the 1960s, by Dawson (1962), and 1970s, by Tallian (1972), and many others. A parameter called film thickness ratio, or λ ratio, or specific film thickness, was introduced for evaluation of lubrication effectiveness in a rough surface EHL contact. More discussion will be given below.

Improvements of Experimental Techniques

As the lubrication theory and practice were fast advancing, new challenges were imposed on EHL experimental technologies. Extremely thin lubricant films and rough asperity contacts may coexist and they are difficult to measure with the capacitance, electric resistance, and x-ray techniques. Efforts, therefore, have focused more on optical interferometry since the 1980s. Originally, the resolution of the film thickness measurement with manual calibration methods was limited to about a quarter of the wavelength of the light being used to produce interference fringes, which is usually around 110–160 nm. Advancements in computer technologies have enabled significant improvements in different ways by different researchers. The main contributions include the following:

Spacer Layer Imaging Method (SLIM) developed at Imperial College, London (Johnston et al. (1991); Cann et al. (1996); and others). In order to overcome the resolution limitation stated above, a combination of a solid spacer layer, having the same reflective index as that of the oil to be measured, with a spectrum analysis technique enables accurate measurement of very thin lubricant films on the nanometer scale.

Relative Optical Interference Intensity (ROII) Technique developed at Tsinghua University, China, by Luo et al. (1996), and others. A monochrome light is used to produce interference fringes and the lubricant film thickness at a certain location is determined by the relative light intensity between the maximum and minimum within the same order of fringe. The intensity is precisely measured through a digitized image analysis, and the resultant film thickness resolution is claimed to be about 0.5 nm.

Thin Film Colorimetric Interferometry developed at Brno University of Technology, Czech Republic (Hartl et al. 1999). This method incorporates a computer-controlled test apparatus with an extensive imaging process software, so that real-time instantaneous evaluation of film thickness distribution can be successfully conducted through colorimetric Interferometry and the measurement range is about 1–800 nm.

Improvements in Numerical Solution Methods

EHL research has relied heavily on numerical analyses through model-based simulations. The development of EHL theory and practice, therefore, has been dependent largely upon advancements of computer technologies and numerical solution methods. An EHL basic equation system typically includes the Reynolds equation that governs the pressure generation, the film thickness (or gap) equation that describes the geometry between the two surfaces, the elasticity equation, the load balance equation, and those of lubricant rheology such as the viscosity and density as functions of pressure (please refer to "► EHL Governing Equations" for details).

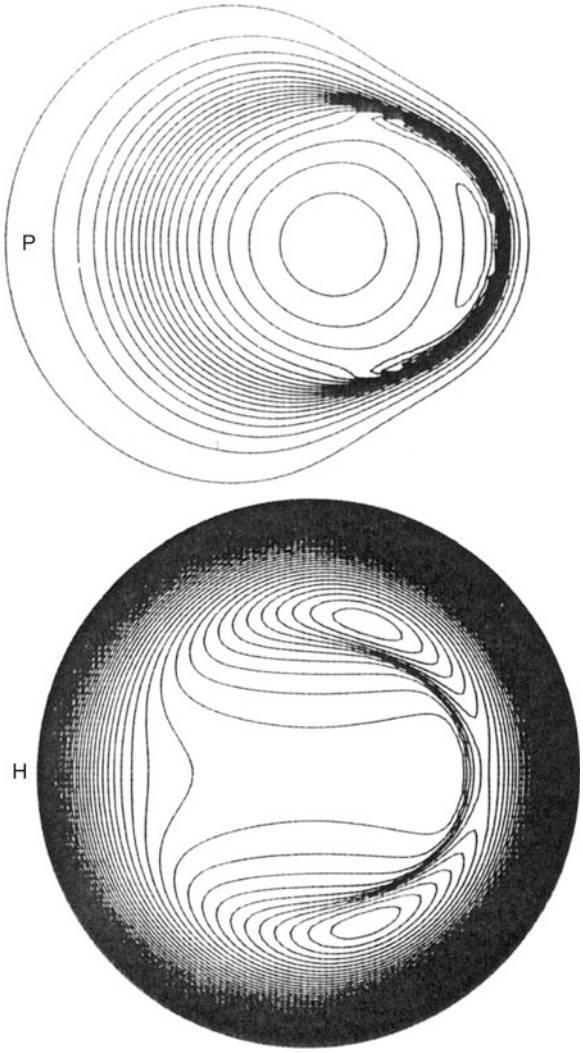
The equation system exhibits very strong nonlinear behaviors, resulting mainly from the surface elastic deformation and much increased viscosity due to pressure. Therefore, the numerical solution has long been a great challenge to researchers. Basically, there have been four major means of solving EHL problems numerically:

1. *Inverse Solution*: This can successfully handle heavily loaded cases and obtain solutions within a small number of calculation cycles. Following Dowson and Higginson's basic approach for line contacts, Evans and Snidle (1981), and Hou et al. (1987), developed inverse solutions for point contact problems. However, the inverse solution procedure requires tricky manual adjustment and domain division from case to case, so it is not fully automatic.
2. *Newton-Raphson Iterative Procedure*: This was first employed by Rohde and Oh (1975), for line contact problems and Oh and Rohde (1977), for point contacts. However, the published solutions were still limited to relatively low Hertzian contact pressures, about 0.15–0.35 GPa. Houpert and Hamrock (1986), and others, improved the numerical approach for line contacts based on Okamura's formulation presented earlier in 1982. They employed a non-uniform computational mesh with much increased mesh density around the inlet of contact zone and also in the vicinity of the pressure spike. Their presented solutions reached a maximum Hertzian pressure as high as 4.8 GPa. However, the Newton-Raphson requires a good initial guess of the pressure distribution and, especially, calculation and storage of several large full matrices in the solution process, which are mainly due to the Jacobean matrix of gradients and its inversion. For point contact problems with an appropriate mesh density, this would demand huge computing time and memory space, so the method has not been extended to the point contact EHL.
3. *Coupled Differential Deflection Method*: This approach was developed by Hughes et al. (2000), for line contact problems and modified by Holmes et al. (2003), for point contacts. A differential form of the deformation equation was derived, with which the effect of pressure appears to be much localized compared with that of the original integral form. As a result, the matrix of pressure-deformation coefficients can be simplified into a bandwidth matrix that enables an elimination solver to be used for line contacts and a coupled iterative technique for point contacts. Based on this, the elastic and hydrodynamic equations can be effectively coupled and solved simultaneously. In this way, the computation is accelerated and the solution better stabilized under heavy loading conditions.
4. *Improved Direct Iterative Approaches*: This approach was the first used in EHL studies, and its main advantages include its simplicity and small memory requirement. However, its success was limited in early years due to its slow convergence and the difficulty of obtaining solutions under practical loading conditions. Great efforts were made from the 1940s through the mid-1980s, but the solutions obtained were still limited to relatively low Hertzian pressures, mostly

below 0.4–0.5 GPa. In order to increase computational speed and also ensure solution convergence, significant efforts have been made, mainly including the following:

(a) *Multi-Grid Method*, originally developed in the areas of computational fluid mechanics, then employed in solving EHL problems first by Lubrecht (1987), followed by Venner (1991) and Ai (1993), and others. It is found that higher frequency errors can be reduced quickly on a finer mesh while lower frequency errors can be removed efficiently on a coarser mesh. Thus, in a multi-grid process the computational mesh is constantly changed with mesh density increasing and decreasing alternately. Through the repeated transitions between coarse and fine meshes, the total error is minimized quickly. By using this method, together with a much reduced dimensionless mesh size $\Delta x/a$ from the previous level of about 0.06–0.20 down to 0.0075–0.03, the solution process appears to be significantly accelerated and numerical accuracy improved in both line and point contact solutions (see Fig. 5 for a sample case).

(b) *Semi-System Approach.* With a direct iterative procedure in each iteration, the coefficient matrix of the discretized Reynolds equation is traditionally constructed only from the pressure flow terms on the left-hand side, and all the other terms on the right are considered as known and calculated using the available pressure from the initial guess or previous iteration. In this way, the solution process may suffer from very slow convergence and numerical instability due to vanishing pressure flow terms under heavy loading and/or low speed conditions. The basic idea of the semi-system approach is to consider the entraining flow term as a function of unknown nodal pressures, so that the construction of the coefficient matrix will utilize not only the pressure flow terms but the entraining flow term as well. Therefore, the diagonal dominance is guaranteed even when the pressure flow becomes extremely weak. This approach was employed by Ai (1993), then by Zhu and Hu (1999), Hu and Zhu (2000), and others. It has been proven that with this approach the solution convergence and stability can be ensured even under extremely severe conditions, and cases with ultra-thin film, zero-film, and rough surface asperity contacts can be handled.



EHL History (Elastohydrodynamic Lubrication), Fig. 5
Multigrid solution of an EHL circular. Contact by Lubrecht (1987)

(c) *Progressive Mesh Densification (PMD) Method.* Morales-Espejel et al. (2005) and Liu et al. (2006a), and others pointed out that, as calculated EHL film thickness is getting smaller, down to nanometer scale or approaching zero, converged solution becomes more significantly dependent on the mesh density and differential schemes used. Zhu (2007), further revealed that the converged film thickness is approaching a limit if mesh density continuously increases, and this asymptotic limit can be readily estimated. Based on this, a progressive mesh densification (PMD)

method has been developed (see Zhu 2007, for details). It appears to be capable of speeding up solution process remarkably while ensuring numerical accuracy, beneficial especially to ultrathin film and mixed EHL.

An important component of the EHL solution is the calculation of surface elastic deformation, which may demand more than 50–70% of total computing time. There have been mainly four types of numerical algorithm for point contact problems: (1) Direct summation with influence coefficients, employed earlier by Hamrock and Dowson (1976a) (via zero-order discretization of the pressure distribution), Ranger et al. (1975) (bilinear discretization), and Zhu and Wen (1984) (biquadratic discretization), and others; (2) Multilevel multi-integration (MLMI) by Lubrecht and Ioannides (1991), and others; (3) Differential deflection method by Evans and Hughes (2000); (4) Discrete convolution and FFT (DC-FFT) originally developed by Liu et al. (2000), and employed in the EHL by Wang et al. (2003).

Thin Film and Mixed EHL with Real Engineering Roughness (Mid-1990s to Present)

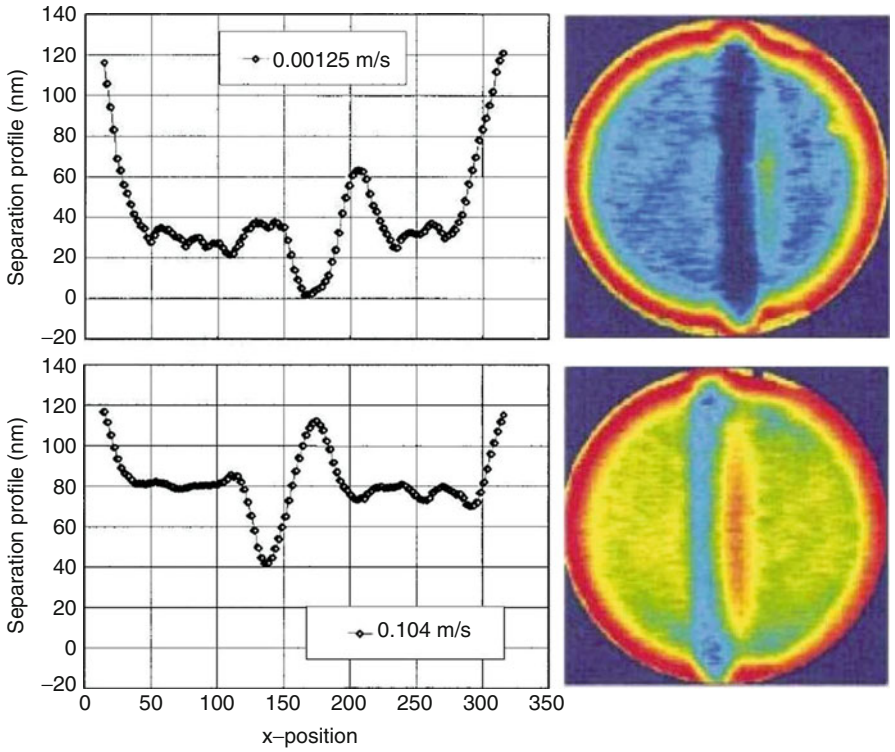
As described above, surface roughness is often of the same order of magnitude as, or greater than, the lubricant film thickness, so that a complete separation between the two rough surfaces is seldom seen in most engineering applications. Mixed EHL (also called partial EHL) is the mode in which both EHL films and surface asperity contacts coexist, and neither can be ignored. Actually, most functional components operate in the mixed lubrication regime. Early stochastic models, such as those by Patir and Cheng (1978a), and Zhu and Cheng (1988), were developed for estimating global performance and parameter average values with no heavy contact and interaction of asperities. Further studies on detailed distributions of pressure, film thickness, and subsurface stresses under severe operating conditions require deterministic solutions with real machined roughness, which were difficult to obtain in the past. The great advancement of computer and information technologies has fueled significant breakthroughs in thin-film and mixed EHL research since the 1990s.

First, improved optical interferometry, described previously, has recently yielded fruitful results in the areas of ultra-thin film and boundary lubrication as well as EHL with textured surfaces. Representative contributions include those by Guangteng and Spikes (1995), Luo et al. (1996), Spikes (2000), Guangteng et al. (2000), Krupka and Hartl (2007), and others. The thin EHL film measurements have come down to a nanometer scale, and the transition from the thin film EHL to boundary lubrication has been a focus of investigation. Also, various types of textured surfaces have been used in the experiments showing asperity contact patterns in mixed EHL. A sample measurement can be found in Fig. 6.

Concurrently, deterministic solutions for mixed EHL have achieved significant progress. Basically, there have been two approaches: the first is to use a unified equation system and solution method for both the lubricated areas and asperity contacts simultaneously, and the second to use separate models for lubrication and contact, respectively. An example of separate solution was presented by Jiang et al (1999), using machined roughness for point contact mixed EHL. The first unified approach for point contacts with machined three-dimensional roughness was published by Zhu and Hu (1999), then by Hu and Zhu (2000). Another unified solution, using a coupled differential deflection method, was obtained by Holmes et al. (2005). The separate approach was also employed by Zhao et al. (2001), and others, mainly for start-up and slow-down problems, in which the boundary conditions between the lubricated and contact areas are relatively easy to handle.

Considering that dry contact is nothing but a special case of lubricated contact under extreme conditions (such as ultra-low viscosity, ultra-low speed, and high pressure concentrated in tiny asperity contact areas), theoretically one should be able to use a unified lubrication equation system to simulate both EHL films and asperity contact simultaneously. Newly developed numerical approaches appear to be capable of simulating the entire transition from full-film and mixed EHL down to dry contact under severe operating conditions (see Fig. 7 for an example). Based on the mixed EHL model by Zhu and Hu, a virtual texturing technology has been developed as a design tool for surface optimization in design practice by Wang and Zhu (2005), and a mixed EHL model for point contacts of coated/layered bodies by Liu et al. (2007), published. Most recently, mixed PEHL (plasto-elastohydrodynamic lubrication) solutions have been presented by Ren et al. (2010–2011), considering the effect of possible plastic deformation on the mixed EHL characteristics.

Comparative studies have been conducted by Felix-Quinonez et al. (2005), Liu et al. (2006b, 2009), Zhu (2007), Wang et al. (2010), Zhu and Wang (2011), and others to validate thin-film and mixed EHL models. Good agreement has been found between model-based



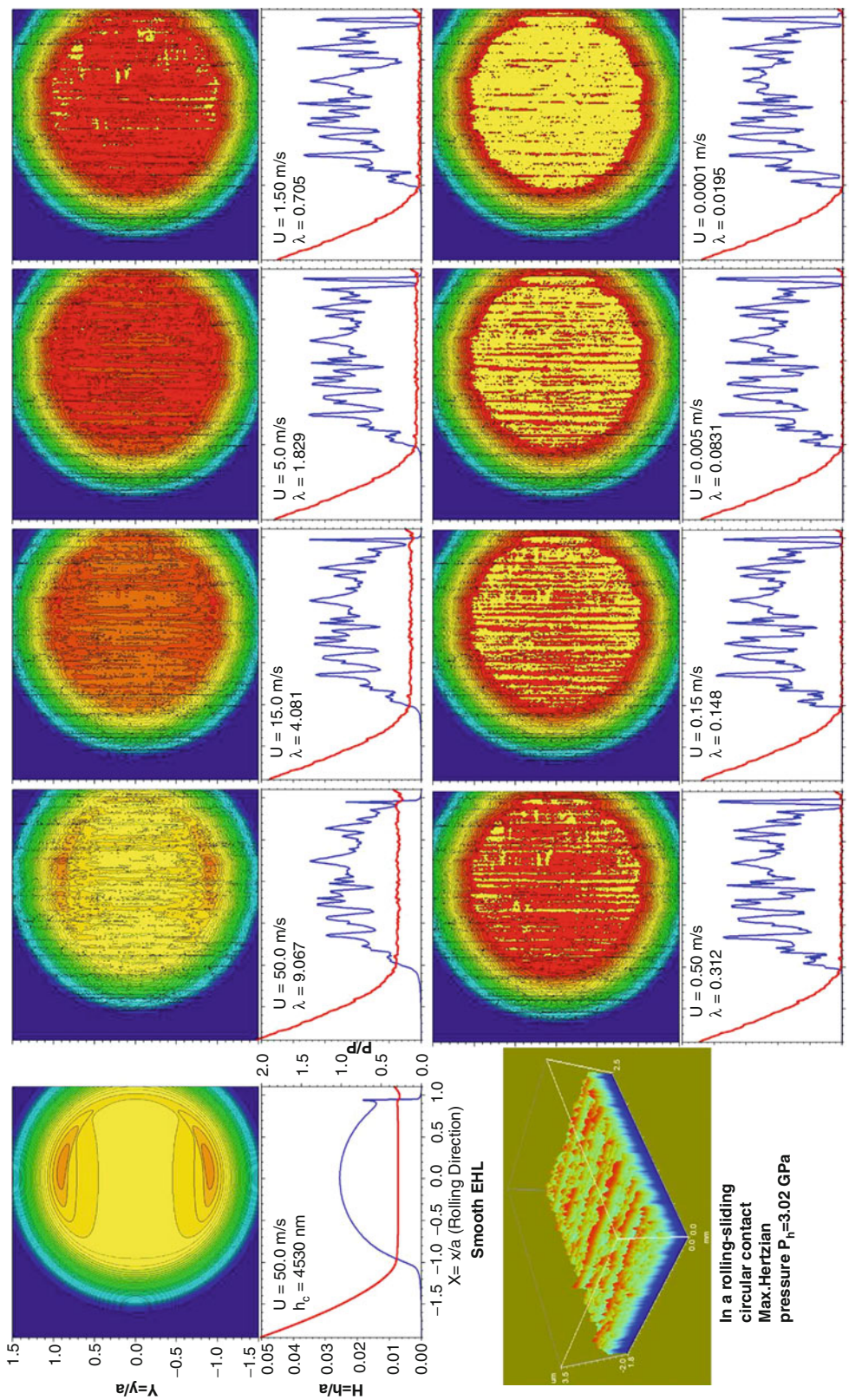
EHL History (Elastohydrodynamic Lubrication), Fig. 6 A transverse ridge passing through an EHL contact, measured by ultrathin-film optical interferometry (Guantang et al. 2000)

simulations and experimental results. Also, as the speed decreases, the mixed EHL results gradually approach those from dry contact analyses. Perfect agreement between the mixed EHL and dry contact solutions has been found when the speed is extremely low. This doubtlessly proves that the mixed EHL models can be used to simulate dry contact interfaces.

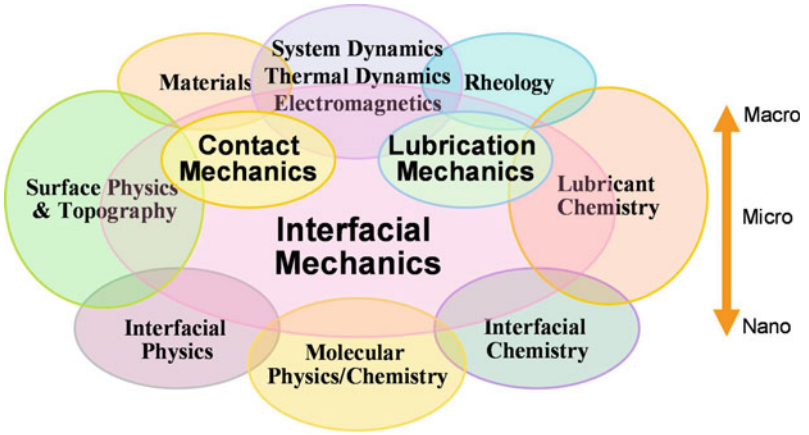
Interfacial Mechanics: A Future Prospect

It has been more than 60 years since the first successful EHL solution was published. The EHL theory and practice have progressed from infancy to maturity. Although tribology research efforts have been dynamically changing their focuses in the last 60 years, the EHL has remained an active field, attracting much attention. There has been enormous achievement both in numerical modeling and experimental investigation. However, much still remains to be done.

In fact, as soon as elastic deformation is introduced into lubrication analyses, elastic dry contact becomes a special case of EHL if air is considered a lubricant. In principle, mixed EHL has included the basic physical elements for modeling contact, or hydrodynamic lubrication, or both together, so that contact mechanics and hydrodynamic lubrication theory, which have traditionally been separate since the 1880s, are being merged. Actually, engineering interfaces usually exhibit multidisciplinary and multi-scale complexity, possibly accomplishing mechanical, physical, chemical, electrical, and biological functions in a natural or designed environment. In early years, interfacial problems had to be greatly simplified with various assumptions, including those of ideally rigid or elastic homogeneous continuum materials, simple geometry, perfectly smooth surfaces or artificial roughness of simple geometry, isoviscous Newtonian fluids, and more, due to the lack of powerful analytical and numerical tools at that time. As research is gradually deepened and widened, some simplification assumptions are inchmeal removed, and macro-scale research based on continuum mechanics are being combined with discrete analyses on a molecular scale. Under the circumstances, the conventional terms, such as "contact," "hydrodynamic," and "elasto-" may no longer be sufficient to characterize interfacial phenomena in further studies, so an integrated, evolving, and more generic concept of



EHL History (Elastohydrodynamic Lubrication), Fig. 7 Unified deterministic model by Zhu and Hu (1999) et al. Being able to simulate the entire transition from full-film and mixed EHL to dry contact with 3-D machined roughness



EHL History (Elastohydrodynamic Lubrication), Fig. 8 Interfacial mechanics and related fields of study

"Interfacial Mechanics" is suggested, which naturally embraces contact and lubrication theories and is a wider advocate for advanced future research in much extended integrated scopes (see Zhu and Wang 2011). It provides a much wider scientific platform for further in-depth investigations of interfacial behaviors and collaborations among multi-scale interdisciplinary studies on interfacial physics, chemistry, and molecular sciences. Figure 8 serves as a pictorial reference, summarizing the above discussion (Zhu and Wang 2011).

Nomenclature

a	Semi-axis of Hertzian contact ellipse in rolling direction, or radius of Hertzian contact circle, or half-width of Hertzian contact for a line contact
b	Semi-axis of Hertzian contact ellipse in the direction perpendicular to rolling
E	Effective elastic modulus
$G^* = \alpha E'$	Dimensionless material parameter
h_c, H_c	Central film thickness, $H_c = h_c/R_x$
h_m, H_m	Minimum film thickness, $H_m = h_m/R_x$
l_e	Effective length of line contact
p_h	Maximum Hertzian pressure
R_x	Effective radius of curvature in x- (rolling) direction
$U^* = \eta_o U/(E'R_x)$,	Dimensionless speed parameter
U	Rolling velocity (or entraining velocity)
W^*	$w/(E'R_x^2)$, or $w/(E'R_x l_e)$ for line contact dimensionless load parameter
w	Load
α	Pressure-viscosity exponent used in (2)
η, η_o	Viscosity and viscosity under ambient condition, respectively

Cross-References

- ▶ Deterministic Models of Rough Surface EHL
- ▶ EHL, Full Numerical Solution Methods
- ▶ EHL Governing Equations
- ▶ Elastohydrodynamic Lubrication (EHL)
- ▶ Stochastic Models for Rough Surface EHL

References

X. Ai, Numerical analyses of elastohydrodynamically lubricated line and point contacts with rough surfaces by using semi-system and multigrid methods. Ph. D. Thesis, Northwestern University, Evanston, 1993

J.F. Archard, E.W. Cowking, Elastohydrodynamic lubrication at point contacts. Proc. Inst. Mech. Engrs. **180**(3B), 47–56 (1966)

ASME Research Committee on Lubrication, in *Viscosity and Density of over 40 Lubricating Fluids of Known Composition at Pressure to 150,000 PSI and Temperature to 425F*, 1953

S. Bair, *High Pressure Rheology for Quantitative Elastohydrodynamics* (Elsevier, Amsterdam, 2007)

S. Bair, M. Khonsari, An EHD inlet zone analysis incorporating the second Newtonian. J. Tribol. **118**, 341–343 (1996)

S. Bair, W.O. Winer, Rheological response of lubricants in EHD contacts, in *Proceedings of the 5th Leeds-Lyon Symposium on Tribology*, Leeds, pp. 162–169, 1978

C. Barus, Isothermals, isopiestics, and isometrics relative to viscosity. Am. J. Sci. **XLV**(266), 87–96 (1893). Third Series

A. Cameron, *The Principles of Lubrication* (Wiley, New York, 1966)

P.M. Cann, J. Hutchinson, H.A. Spikes, The development of a spacer layer imaging method (SLIM) for mapping elastohydrodynamic contacts. Tribol. Trans. **39**, 915–921 (1996)

H.S. Cheng, A numerical solution of the elastohydrodynamic film thickness in an elliptical contact. J. Lubr. Technol. **92**, 155–162 (1970)

H.S. Cheng, B. Sternlicht, A numerical solution for the pressure, temperature, and film thickness between Two infinitely long, lubricated rolling and sliding cylinders, under heavy loads. J. Basic Eng. **87**, 695–707 (1964)

H.S. Cheng, Calculation of elastohydrodynamic film thickness in high speed rolling and sliding contacts. *MTI Technical Report*, MIT-67TR24, 1967

F. Chevalier, A.A. Lubrecht, P.M.E. Cann, F. Colin, G. Dalmaz, Film thickness in starved EHL point contacts. J. Tribol. **120**, 126–133 (1998)

Y.P. Chiu, An analysis and prediction of lubricant film starvation in rolling contact systems. ASLE Trans. **17**, 22–35 (1974)

R.L. Cook, C.A. Herbst, H.E. King Jr., High-pressure viscosity of glass-forming liquids measured by the centrifugal force diamond anvil cell viscometer. J. Phys. Chem. **97**, 2355–2361 (1993)

A.W. Crook, The lubrication of rollers – II. Film thickness with relation to viscosity and speed. Philos. Trans. Roy. Soc. Lond. **A254**, 223–236 (1961)

A.W. Crook, The lubrication of rollers – IV. Measurements of friction and effective viscosity. Philos. Trans. Roy. Soc. Lond. **A255**, 281–312 (1963)

P.H. Dawson, Effect of metallic contact on the pitting of lubricated rolling surfaces. J. Mech. Eng. Sci. **4**, 16–21 (1962)

A.D. Doolittle, Studies in Newtonian flow II the dependence of the viscosity of liquids on free-space. J. Appl. Phys. **22**, 1471–1475 (1951)

D. Dowson, G.R. Higginson, A numerical solution to the elastohydrodynamic problem. J. Eng. Sci. **1**, 6–15 (1959)

D. Dowson, G.R. Higginson, New roller bearing lubrication formula. Engineering **192**, 158–159 (1961)

D. Dowson, G.R. Higginson, *Elastohydrodynamic Lubrication* (Pergamon, Oxford, 1966)

D. Dowson, A.V. Whitaker, A numerical procedure for the solution of the elastohydrodynamic problem of rolling and sliding contacts lubricated by a Newtonian fluid. Proc. Inst. Mech. Engrs. **180**(3B), 119–134 (1966)

D. Dowson, S. Toyoda, A central film thickness formula for elastohydrodynamic line contacts, in *Proceedings of 5th Leeds-Lyon Symposium on Tribology*, Leeds, pp. 60–65, 1978

A. Dyson, H. Naylor, A.R. Wilson, The measurement of film thickness in elastohydrodynamic contacts. Proc. Inst. Mech. Engrs. **180**(3), 119–134 (1966)

A.M. Ertel, Hydrodynamic lubrication based on new principles. *Akad. Nauk. SSSR, Prikladnaya Matematika i Mekhanika* **3**(2), 41–52 (1939) (in Russian)

H.P. Evans, T.G. Hughes, Evaluation of deflection in semi-infinite bodies by a differential method. Proc. Inst. Mech. Engrs. **214**(Part C), 563–584 (2000)

H.P. Evans, R.W. Snidle, Inverse solution of Reynolds equation of lubrication under point contact elastohydrodynamic conditions. J. Lubr. Technol. **103**, 539–546 (1981)

A. Felix-Quinonez, P. Ehret, J.L. Summers, On three-dimensional flat-top defects passing through an EHL point contact: a comparison of modeling with experiments. J. Tribol. **127**, 51–60 (2005)

C.A. Foord, L.D. Wedeven, F.J. Westlake, A. Cameron, Optical elastohydrodynamics. Proc. Inst. Mech. Engrs. **184**(Part 1), 487–503 (1969–1970)

R. Gohar, A. Cameron, Optical measurement of Oil film thickness under elastohydrodynamic lubrication. Nature **200**, 458–459 (1963)

R. Gohar, A. Cameron, The mapping of elastohydrodynamic contacts. ASLE Trans. **10**, 215–225 (1966)

J.A. Greenwood, J.H. Tripp, The contact of Two nominally flat rough surfaces. Proc. Inst. Mech. Engrs. **185**(Part 1), 625–633 (1970–1971)

A.N. Grubin, Fundamentals of the hydrodynamic theory of lubrication of heavily loaded cylindrical surfaces (Central Scientific Research Institute for Technology and Mechanical Engineering, Moscow, 1949). Book No.30, (DSIR Translation), pp. 115–166

G. Guantang, H.A. Spikes, Behavior of lubricants in the mixed elastohydrodynamic regime, in *Proceedings 21st Leeds-Lyon Symposium on Tribology*, Leeds, pp. 479–485, 1995

G. Guantang, P.M. Cann, A. Olver, H.A. Spikes, Lubricant film thickness in rough surface mixed elastohydrodynamic contact. J. Tribol. **122**, 65–76 (2000)

P.K. Gupta, H.S. Cheng, D. Zhu, N.H. Forster, J.B. Schrand, Viscoelastic effects in MIL-L-7808-type lubricant, part I: analytical formulation. Tribol. Trans. **35**, 269–274 (1992)

G.M. Hamilton, S.L. Moore, Deformation and pressure in an EHD contact. Proc. Roy. Soc. Lond. **A322**, 313–330 (1971)

B.J. Hamrock, D. Dowson, Isothermal elastohydrodynamic lubrication of point contacts, part 1 – theoretical formulation. J. Lubr. Technol. **98**, 223–229 (1976a)

B.J. Hamrock, D. Dowson, Isothermal elastohydrodynamic lubrication of point contacts, part 2 – ellipticity parameter results. J. Lubr. Technol **98**, 375–383 (1976b)

B.J. Hamrock, D. Dowson, Isothermal elastohydrodynamic lubrication of point contacts, part 3 – fully flooded results. J. Lubr. Technol. **99**, 264–276 (1977a)

B.J. Hamrock, D. Dowson, Isothermal elastohydrodynamic lubrication of point contacts, part 4 – starvation results. J. Lubr. Technol. **99**, 15–23 (1977b)

M. Hartl, I. Krupka, R. Poliscuk, M. Liska, An automatic system for real-time evaluation of EHD film thickness and shape based on the colorimetric interferometry. Tribol. Trans. **42**, 303–311 (1999)

M.J.A. Holmes, H.P. Evans, T.G. Hughes, R.W. Snidle, Transient elastohydrodynamic point contact analysis using a New coupled differential deflection method, part 1: theory and validation. J. Eng. Tribol. **217**, 289–303 (2003)

M.J.A. Holmes, H.P. Evans, R.W. Snidle, Analysis of mixed lubrication effects in simulated gear tooth contacts. J. Tribol. **127**, 61–69 (2005)

K.P. Hou, D. Zhu, S.Z. Wen, An inverse solution to the point contact EHL problem under heavy loads. J. Tribol. **109**, 432–436 (1987)

L.G. Houpert, B.J. Hamrock, Fast approach for calculating film thicknesses and pressures in elastohydrodynamically lubricated contacts at heavy loads. J. Tribol. **108**, 411–420 (1986)

Y.Z. Hu, D. Zhu, A full numerical solution to the mixed lubrication in point contacts. J. Tribol. **122**, 1–9 (2000)

T.G. Hughes, C.D. Elcoate, H.P. Evans, Coupled solution of the elastohydrodynamic line contact problem using a differential deflection method. J. Mech. Eng. Sci. **214**, 585–598 (2000)

X. Jiang, D.Y. Hua, H.S. Cheng, X. Ai, S.C. Lee, A mixed elastohydrodynamic lubrication model with asperity contact. J. Tribol. **121**, 481–491 (1999)

K.L. Johnson, R. Cameron, Shear behavior of elastohydrodynamic oil films at high rolling contact pressures. Proc. Inst. Mech. Engrs. **182**, 307–319 (1967–1968)

K.L. Johnson, J.L. Tevaarwerk, Shear behavior of EHD oil films. Proc. Roy. Soc. Lond. **A356**, 215–236 (1977)

G.J. Johnston, R. Wayte, H.A. Spikes, The measurement and study of very thin lubricant films in concentrated contacts. Tribol. Trans. **34**, 187–194 (1991)

J.W. Kannel, The measurement of pressure in rolling contacts. Proc. Inst. Mech. Engrs. **180**(3B), 135–142 (1966)

J.W. Kannel, J.C. Bell, A method for estimating of temperatures in lubricated rolling-sliding gear or bearing EHD contacts, in *Proceedings of*

Elastohydrodynamic Lubrication 1972 Symposium, Inst. Mech. Engrs, Paper C24/72, pp. 118–130, 1972

K.A. Koye, W.O. Winer, An experimental evaluation of the Hamrock and Dowson minimum film thickness equation for fully flooded EHD point contacts. J. Lubr. Technol. **103**, 284–294 (1981)

I. Krupka, M. Hartl, The influence of thin boundary films on real surface roughness in thin film, mixed EHD contact. Tribol. Int. **40**, 1553–1560 (2007)

P. Kumar, P. Anuradha, M.M. Khonsari, Some important aspects of thermal elastohydrodynamic Lubrication. J. Mech. Eng. Sci. **224**, 2588–2598 (2010)

C.C. Kweh, M.J. Patching, H.P. Evans, R.W. Sindle, Simulation of elastohydrodynamic contacts between rough surfaces. J. Tribol. **114**, 412–419 (1992)

S. Liu, Q. Wang, G. Liu, A versatile method of discrete convolution and FFT (DC-FFT) for contact analyses. Wear **243**, 101–111 (2000)

Y. Liu, W.W. Chen, D. Zhu, S. Liu, Q.J. Wang, An elastohydrodynamic lubrication model for coated surfaces in point contacts. J. Tribol. **129**, 509–516 (2007)

Y. Liu, Q. Wang, W.Z. Wang, Y.Z. Hu, D. Zhu, Effects of differential scheme and mesh density on EHL film thickness in point contacts. J. Tribol. **128**, 641–653 (2006a)

Y.C. Liu, Q.J. Wang, W.Z. Wang, Y.Z. Hu, D. Zhu, I. Krupka, M. Hartl, EHL simulation using the free-volume viscosity model. Tribol. Lett. **23**, 27–37 (2006b)

Y.C. Liu, Q. Wang, D. Zhu, W.Z. Wang, Y.Z. Hu, Effect of differential scheme and viscosity model on rough surface point contact isothermal EHL. J. Tribol. **131**(044501), 1–5 (2009)

A.A. Lubrecht, 1987, The numerical solution of elastohydrodynamic lubricated line and point contact problems using multigrid techniques, Ph.D. Thesis, University of Twente, The Netherlands

A.A. Lubrecht, E.A. Ioannides, Fast solution of the Dry contact problem and associated surface stress field, using multilevel techniques. J. Tribol. **113**, 128–133 (1991)

J.B. Luo, S.Z. Wen, P. Huang, Thin film lubrication 1. Study on the transition between EHL and thin film lubrication using a relative optical interference intensity technique. Wear **194**, 107–115 (1996)

H.M. Martin, Lubrication of gear teeth. Engineering (London) **102**, 119–121 (1916)

G.E. Morales-Espejel, M.L. Dumont, P.M. Lugt, A.V. Olver, A limiting solution for the dependence of film thickness on velocity in EHL contacts with very thin films. Tribol. Trans. **48**, 317–324 (2005)

L.E. Murch, W.R.D. Wilson, A thermal elastohydrodynamic inlet zone analysis. J Lubr. Technol. **97**, 212–216 (1975)

K.P. Oh, S.M. Rohde, Numerical solution of the point contact problem using the finite element method. Int. J. Numer. Meth. Eng. **11**, 1507–1518 (1977)

H. Okamura, A contribution to the numerical analysis of isothermal elastohydrodynamic lubrication, in *Proceedings of 9th Leeds-Lyon Symposium on Tribology*, Leeds, pp. 313–320, 1982

F.K. Orcutt, Experimental study of elastohydrodynamic lubrication. ASLE Trans. **8**, 381–396 (1965)

N. Patir, H.S. Cheng, Effect of surface roughness orientation on the central film thickness in EHD contacts, in *Proceedings of the 5th Leeds-Lyon Symposium on Tribology*, Leeds, pp. 15–21, 1978a

N. Patir, H.S. Cheng, An average flow model for determining effects of three- dimensional roughness on partial hydrodynamic lubrication. J Lubr. Technol. **100**, 12–17 (1978b)

A.I. Petrusevich, Fundamental conclusions from the contact-hydrodynamic theory of lubrication. Izv. Akad. Nauk SSR (OTN) **2**, 209–233 (1951)

M.A. Plint, Traction in elastohydrodynamic contacts, Proc. Inst. Mech. Engrs. **182**, 300–306 (1967–1968)

A.P. Ranger, C.M.M. Ettles, A. Cameron, The solution of the point contact elastohydrodynamic problem. Proc. Roy. Soc. Lond. **A346**, 227–244 (1975)

N. Ren, D. Zhu, W.W. Chen, Y. Liu, Q.J. Wang, A three-dimensional deterministic model for rough surface line contact EHL problems. J. Tribol. **131**(011501), 1–9 (2009)

N. Ren, D. Zhu, W.W. Chen, Q.J. Wang, Plasto-elastohydrodynamic lubrication (PEHL) in point contacts. J. Tribol. **132**(031501), 1–11 (2010)

N. Ren, D. Zhu, Q.J. Wang, Three-dimensional plasto-elastohydrodynamic lubrication (PEHL) for surfaces with irregularities. J. Tribol. **133**(031502), 1–10 (2011)

O. Reynolds, On the theory of lubrication and its application to Mr. Beauchamp Tower's experiments, including an experimental determination of the viscosity of olive oil. Philos. Trans. Roy. Soc. **177**, 157–234 (1886)

C.J.A. Roelands, Correlational aspects of the viscosity-temperature-pressure relationship of lubricating oils, Ph.D. Thesis, Technische Hogeschool Delft, The Netherlands, 1966

S.M. Rohde, K.P. Oh, A unified treatment of thick and thin film elastohydrodynamic problems by using higher order element methods. Proc. Roy. Soc. Lond. **A343**, 315–331 (1975)

L.B. Sibley, F.K. Orcutt, Elastohydrodynamic lubrication of rolling contact surfaces. ASLE Trans. **4**, 234–249 (1961)

H.A. Spikes, The borderline of elastohydrodynamic and boundary lubrication. J. Mech. Eng. Sci. **214**, 23–37 (2000)

T.E. Tallian, Theory of partial elastohydrodynamic contacts. Wear **21**, 49–101 (1972)

V. Turchina, D.M. Sanborn, W.O. Winer, Temperature measurement in sliding elastohydrodynamic point contacts. J. Lubr. Technol. **96**, 464–471 (1974)

C.H. Venner, Multilevel solution of EHL line and point contact problems, Ph.D Thesis, University of Twente, The Netherlands, 1991

Q. Wang, D. Zhu, Virtual texturing: modeling the performance of lubricated contacts of engineered surfaces. J. Tribol. **127**, 722–728 (2005)

W.Z. Wang, H. Wang, Y.C. Liu, Y.Z. Hu, D. Zhu, A comparative study of the methods for calculation of surface elastic deformation. J. Tribol. **217**, 145–152 (2003)

W.Z. Wang, Y.Z. Hu, Y.C. Liu, D. Zhu, Solution agreement between Dry contacts and lubrication system at ultra-Low speed. J. Tribol. **224**, 1049–1060 (2010)

L.D. Wedeven, D. Evans, A. Cameron, Optical analysis of ball bearing starvation. J. Lubr. Technol. **93**, 349–363 (1971)

P.E. Wolveridge, K.P. Baglin, J.F. Archard, The starved lubrication id cylinders in line contacts, Proc. Inst. Mech. Engrs. **181**, 1159–1169 (1971)

G. Xu, F. Sadeghi, Thermal EHL analysis of circular contacts with measured surface roughness. J. Tribol. **118**, 473–483 (1996)

J.X. Zhao, F. Sadeghi, M.H. Hoeprich, Analysis of EHL circular contact start Up: part I – mixed contact model with pressure and film thickness results. J. Tribol. **123**, 67–74 (2001)

D. Zhu, On some aspects in numerical solution of thin-film and mixed EHL. J. Eng. Tribol. **221**, 561–579 (2007)

D. Zhu, X. Ai, Point contact EHL based on optically measured three-dimensional rough surfaces. J. Tribol. **119**, 375–384 (1997)

D. Zhu, H.S. Cheng, Effect of surface roughness on the point contact EHL. J. Tribol. **110**, 32–37 (1988)

D. Zhu, S.Z. Wen, A full numerical solution for the thermoelastohydrodynamic problem in elliptical contacts. J. Tribol. **106**, 246–254 (1984)

D. Zhu, H.S. Cheng, B.J. Hamrock, Effect of surface roughness on pressure spike and film constriction in elastohydrodynamically lubricated line contacts. Tribol. Trans. **33**, 267–273 (1990)
D. Zhu, Y.Z. Hu, The study of transition from full film elastohydrodynamic to mixed and boundary lubrication, in *The Advancing Frontier of Engineering Tribology, Proceedings of the 1999 STLE/ASME H.S. Cheng Tribology Surveillance*, Orlando, pp. 150–156, 1999
D. Zhu, Q. Wang, Elastohydrodynamic lubrication: a gateway to interfacial mechanics – review and prospect, J Tribol. 133(041001):1–14 (2011)

EHL Modeling Considering Rough Surfaces

► Deterministic Models of Rough Surface EHL

EHL Modeling Considering Surface Roughness

► Stochastic Models for Rough Surface EHL

EHL Numerical Solution

► EHL, Full Numerical Solution Methods

EHL of a Cylinder on an Elastomeric Layer

► EHL of Coated Bodies

EHL of Coated Bodies

Yuchuan Liu
Powertrain CAE Methods, General Motor Corporation, Pontiac, MI, USA

Synonyms

EHL for an elastic layer bonded to a rigid body; EHL of a cylinder on an elastomeric layer

Definition

EHL of coated bodies is the elastohydrodynamic lubrication generated between coated surfaces.

Scientific Fundamentals

An EHL interface in point contacts formed by coating surfaces is illustrated in Fig. 1, where a coated ball is in contact with a coated flat. Different from a conventional EHL problem, each surface is covered by a layer of coating material that has different mechanical properties from that of its substrate. These coatings will alter the surface elastic deformation as well as the asperity contact friction. As a result, EHL film thickness, pressure, substrate stress, and flash temperature will also be changed. Here a stiff coating refers to the coating with a higher Young's modulus as compared to its substrate, while the compliant coating is the opposite.

In point contacts, the steady-state hydrodynamic pressure is governed by the modified Reynolds equation expressed as follows:

$$\frac{\partial}{\partial x}\left(\varphi_x \frac{\rho h^3}{12\eta}\frac{\partial p}{\partial x}\right) + \frac{\partial}{\partial y}\left(\varphi_y \frac{\rho h^3}{12\eta}\frac{\partial p}{\partial y}\right) = u\frac{\partial(\rho h)}{\partial x} \quad (1)$$

Two flow factors, φ_x and φ_y, have been introduced to describe the non-Newtonian rheological behavior of a lubricant.

The gap between two deformed surfaces should be expressed with:

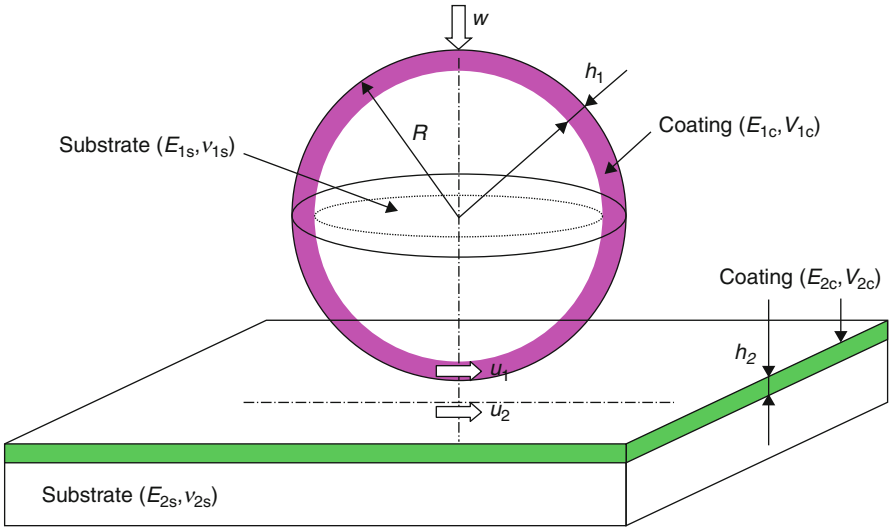
$$h = h_0 + \frac{x^2}{2R} + \frac{y^2}{2R} + d_1(x,y) + d_2(x,y) \quad (2)$$

where d_1 and d_2 are the elastic deformations (normal displacement) of the two surfaces. By removing the terms correlated with the y direction, the above two equations can be used to describe a line contact problem.

In theory, the difference between EHL models for uncoated and coated surfaces is only in the calculation of elastic deformation. If an elastic layer is bonded to a rigid substrate, a simple formula for elastic deformation can be obtained on the basis of the constrained-column model (Johnson 1985):

$$d_i(x,y) = \frac{(1+v_i)(1-2v_i)}{(1-v_i)E_i}p(x,y) \quad (i=1,2) \quad (3)$$

Apparently, this formula cannot be acceptable for a Poisson's ratio $v_i = 0.5$. Dowson and Jin (1989) concluded that this model can be reasonably used to predict the elastic deformation if Poisson's ratio is less than about 0.45 and the ratio of the loading width to layer thickness $\alpha = b/t$ is greater than about 2. A full description of elastic



EHL of Coated Bodies, Fig. 1 Description of the EHL with coated surfaces in a point contact

deformation for the elastic layer bonded to a rigid substrate in line contact is given by Meijers (1968):

$$d_i(x) = \frac{1}{\pi E_i^*}\int_{\Omega} K_i(x-\xi)p(\xi)d\xi \quad (i=1,2) \tag{4}$$

where $E_i^* = E_i/(1-\nu_i^2)$. The Green's function (GF) for this convolution is expressed as:

$$K_i(x) = \frac{1}{\pi E_i^*}\int_0^{\infty} \frac{[(3-4\nu_i)\sinh(2\omega)-2\omega]\cos(x\omega/t)}{\omega[(3-4\nu_i)\cosh(2\omega)+2\omega^2+5-12\nu_i+8\nu_i^2]}d\omega \quad (i=1,2) \tag{5}$$

Similarly, in point contact, the corresponding deformation equation can be found in Jin (2000a).

A rigid substrate is an unrealistic assumption. The substrate can also be elastic. For an uncoated elastic half-surface, the normal displacement by a pressure distribution is controlled by the Boussinesq integration:

$$d_i(x,y) = \frac{1}{\pi E_i^*}\iint_{\Omega} \frac{p(\xi,\zeta)}{\sqrt{(x-\xi)^2+(y-\zeta)^2}}\mathrm{d}\xi\mathrm{d}\zeta \quad (i=1,2) \tag{6}$$

The Green's function in the convolution equation (6) is:

$$g_i(x,y) = \frac{1}{\pi E_i^*\sqrt{x^2+y^2}} \quad (i=1,2) \tag{7}$$

The corresponding frequency response function (FRF) in the frequency domain is:

$$\tilde{\tilde{g}}_i(m,n) = \frac{2}{E_i^*\sqrt{m^2+n^2}} \quad (i=1,2) \tag{8}$$

For a coated elastic half-space surface, i.e., both coating and substrate can be elastically deformed; an explicit Green's function is not available for the normal displacement in the space domain. However, its frequency response function is available in the frequency domain (Nogi and Kato 1997):

$$\tilde{\tilde{g}}_i(m,n) = \frac{2}{E_{ic}^*\omega}\frac{1+4\omega h_i\kappa\vartheta-\lambda\kappa\vartheta^2}{1-(\lambda+\kappa+4\kappa\omega^2h_i^2)\vartheta+\lambda\kappa\vartheta^2} \quad (i=1,2) \tag{9}$$

where,

$$\kappa = \frac{\mu-1}{\mu+(3-4\nu_{ic})} \quad \lambda = 1-\frac{4(1-\nu_{ic})}{1+\mu(3-4\nu_{is})} \quad \omega = \sqrt{m^2+n^2}$$
$$E_{ic}^* = \frac{E_{ic}}{1-\nu_{ic}^2} \qquad \mu = \frac{E_{ic}(1+\nu_{is})}{E_{is}(1+\nu_{ic})} \qquad \vartheta = \exp(-2\omega h_i)$$

The above FRF can be applied in line contact problems by simply replacing $\omega = \sqrt{m^2+n^2}$ with $\omega = |m|$ (Liu et al. 2005b). If multilayer coatings are applied, the equations and algorithms can be found in Elsharkawy and Hamrock (1993) for line contacts and Liu and Wang (2002) for point contacts.

All these equations for elastic deformation are convolution. Therefore, influence coefficients (ICs) are introduced to improve computational efficiency and save memory. If their Green's functions are available in the spatial domain, for example, the (5) and (7), ICs can be directly obtained by a discretization in the spatial domain. The numerical methods for the discretization can be

found in Dowson and Jin (1989) and Jin (2000a). If the convolution can only be expressed in the frequency domain, for example, (9), the corresponding FRF can be first discretized in frequency domain then be transferred into spatial domain to get the ICs by using inverse fast Fourier transform (IFFT). Once the influence coefficients are obtained, the elastic deformation of coated surfaces in EHL can be evaluated using the discrete convolution FFT (DC-FFT) fast algorithm (Liu and Wang 2002) or multilevel multi iIntegration (MLMI) algorithm (Venner and Lubrecht 2000) to speed up the computation.

Key Applications

Coatings are widely used in surface engineering to prevent excessive wear and fatigue, and to prevent corrosion and surface decoration. In the field of biology, cartilage is a natural coating in diarthrodial joints. In transportation and heavy-duty machinery industries, coatings are found on parts under relative motion, such as gears and bearings, which usually work under EHL condition.

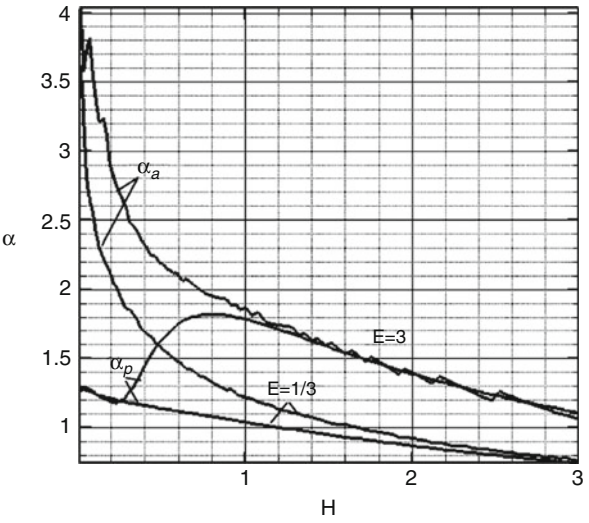
In traditional EHL without coating, the nominal maximum contact pressure and contact half-width are fast evaluated from Hertz theory. However, for coated surfaces the classic Hertz theory is not applicable. Complicated algorithms and iterations are needed to obtain their values. Studies (Liu et al. 2005a, b) on the extension of the Hertz theory to the coated surface contact developed fitting equations based on accurate numerical iterations, and realized the fast evaluation of these parameters for EHL with coated surfaces.

For example, in line contact (Liu et al. 2005b), suppose a coated disk is loaded against an uncoated cylinder. A 10-μm thick boron carbide (B_4C) coating is deposited on the 52,100 steel disk. The material properties are as follows: $E_{1c} = 380$ GPa, $\nu_{1c} = 0.17$, $E_{1s} = 200$ GPa, and $\nu_{1s} = 0.3$ with the modulus ratio, $E = E^*_{1c}/E^*_{1s} = [E_{1c}/(1-\nu^2_{1c})]/[E_{1s}/(1-\nu^2_{1s})] = 1.78$. The cylinder is also made of 52,100 steel and its radius is 10 mm. The applied load is 1 N/mm. Under these conditions, the contact half-width and maximum contact pressure for corresponding uncoated case are $a = 10.764$ μm and $p_h = 59.143$ MPa based on the following classic Hertz equations:

$$a = \sqrt{\frac{4WR}{\pi E^*}} \tag{10}$$

$$p_h = \sqrt{\frac{WE^*}{\pi R}} \tag{11}$$

$$\frac{1}{E^*} = \frac{1}{E^*_1} + \frac{1}{E^*_2} \tag{12}$$



EHL of Coated Bodies, Fig. 2 Representative α_p and α_a values (elastic substrate)

where $E^*_1 = E_{1s}/(1-\nu^2_{1s})$ and $E^*_2 = E_2/(1-\nu^2_2)$. With h = 10 μm, the dimensionless coating thickness ($H = h/a$) is 0.93. The next step is to evaluate two coefficient α_p and α_a by looking up Fig. 2 and calculating the following equation:

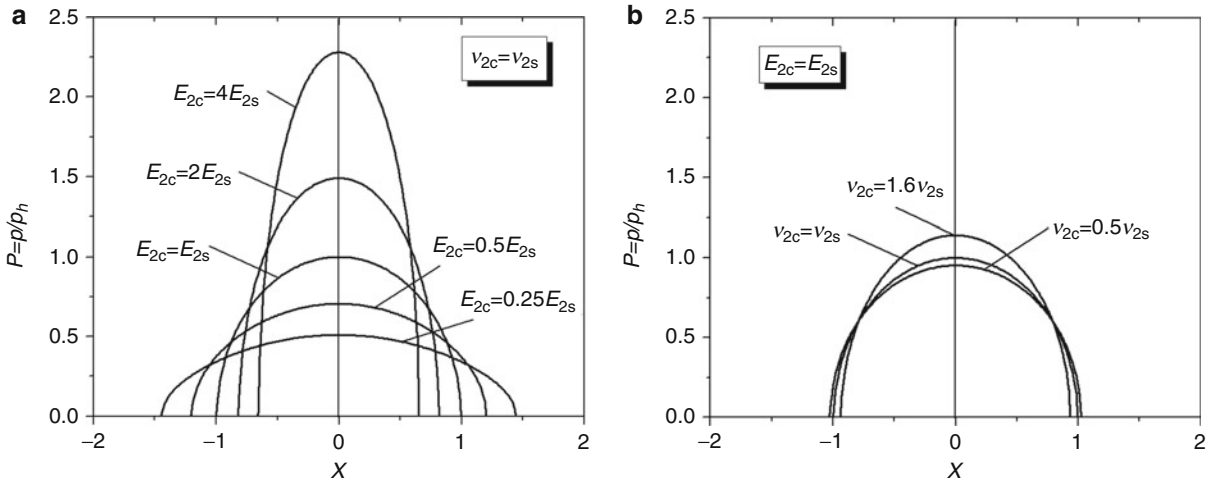
$$\alpha_p(E) = \begin{cases} \alpha_p(1/3) - (1/E - 3)/16 & E \in [1/4, 1/3] \\ 1.25 + [\alpha_p(1/3) - 1.25]/(3E) & E \in [1/3, 1] \\ 1.25 + E[\alpha_p(3) - 1.25]/3 & E \in [1, 3] \\ \alpha_p(3) - (E-3)/12 & E \in [3, 4] \end{cases} \tag{13a}$$

$$\alpha_a(E) = \begin{cases} \alpha_a(1/3) - (1/E - 3)/16 & E < 1 \\ \alpha_a(3) + (E-3)/12 & E > 1 \end{cases} \tag{13b}$$

Due to $E = 1.78 > 1$, a curves in Fig. 2 for $E = 3$ are used to find values $\alpha_p(3) \approx 1.8$ and $\alpha_a(3) \approx 1.88$ in terms of $H = 0.93$. After substitution of these values into (13), one can obtain $\alpha_p(1.78) \approx 1.353$ and $\alpha_a(1.78) \approx 1.778$. Substituting the values into the following equation:

$$E^*_1 = E^*_{1c} \frac{1 - (\lambda + \kappa + 4\kappa\alpha^2 H^2)\exp(-2\alpha H) + \lambda\kappa\exp(-4\alpha H)}{1 + 4\alpha H\kappa\exp(-2\alpha H) - \lambda\kappa\exp(-4\alpha H)} \tag{14}$$

one gets E^*_1 of 324.27 and 293.76 GPa for contact half-width and maximum contact pressure, respectively. The corresponding E^* in (12) are 130.995 and 125.722 GPa. Therefore, the contact half-width and maximum contact



EHL of Coated Bodies, Fig. 3 Effect of coating Young's modulus and Poisson's ratio on contact radius and contact pressure (Liu et al. 2007)

pressure are predicted by (10) and (11): 9.859 μm and 63.260 MPa. The whole process is straightforward without iteration. One may also find the fitting equations for rigid substrate case or the cases in point contact (Liu et al. 2005a, b). In general, as shown in Fig. 3, the coatings with higher Young's modulus or Poisson's ratio will reduce the contact width but increase the contact pressure. The inverse is true for the coatings with lower Young's modulus or Poisson's ratio.

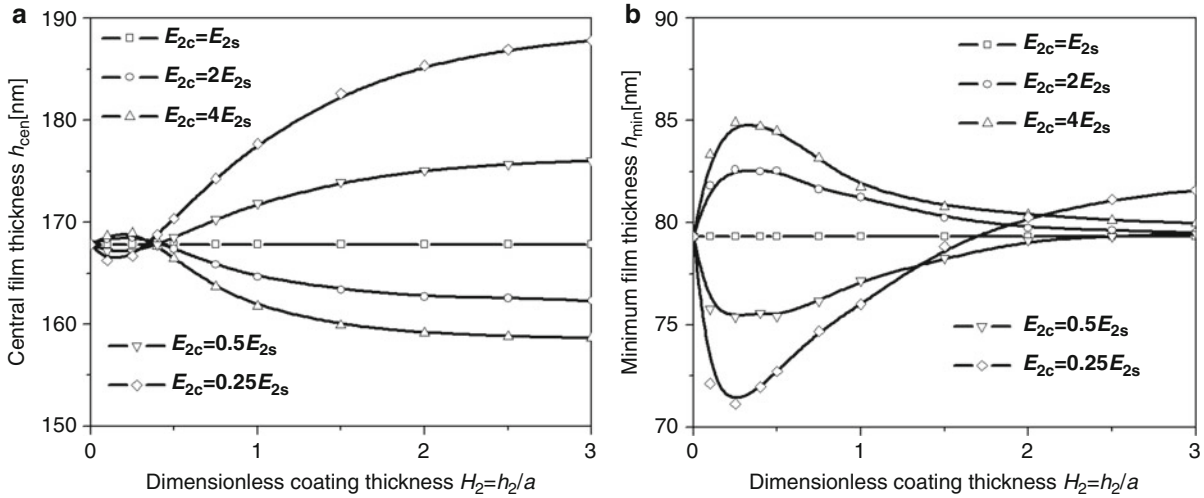
There are two kinds of elastohydrodynamic lubrication. One is the so-called *soft EHL*, working at the elastic isoviscous regime where only surface deformation is taken into account. The typical applications include natural or artificial joint lubrication, lip seals, etc. The other is the so-called *hard EHL*, working at the elastic piezoviscous regime. It covers the lubrications in gears, ball or roller bearings, cams, and rollers. Both surface elastic deformation and lubricant viscosity pressure effect are important in hard EHL.

In soft EHL with coated surfaces, especially for human synovial joints, many researchers have studied the film thickness for the case of an elastic layer firmly bonded with rigid substrate. Hooke (1986) investigated a line contact case using Meijers' equation (4) for an incompressible layer surface ($\nu = 0.5$). For the same problem, Jin et al. (1993) developed the minimum film thickness formulae for two Poisson's ratios of $\nu = 0.4$ and $\nu = 0.5$. These formulae were derived based on simplified elasticity equation, but could be good references.

Jin (2000a, b) numerically studied point-contact soft EHL with coated surfaces. It is found that if the layer thickness is larger than four times the contact radius, the deformation behavior is generally controlled by the elastic layer. Therefore, the film thickness can be predicted from the semi-infinite solid assumption. For the cases with different Poisson's ratio and layer thickness, central and minimum film thickness obtained from numerical simulations were fitted into the equations.

In hard EHL with coated surfaces, Elsharkawy and Hamrock (1994) investigated the coating effect in line contact. The elastic deformation for coated surfaces was computed from a multilayered elasticity analysis described in Elsharkawy and Hamrock (1993). It is found that coating thickness has a significant influence on the minimum film thickness and maximum contact pressure. The pressure peak is higher for stiff coatings than for uncoated surfaces, but the pressure peak disappears for compliant coatings. Both central film thickness and minimum film thickness experience a turning trend as coating thickness increases. For a compliant coating of MoS_2, the central and minimum film thicknesses first decrease to a minimum value then increase. The inverse is true for stiff coatings of TiN. The reason for this was attributed to the viscosity-pressure effect. It is found that in the corresponding isoviscous EHL simulations, the turning trend was removed and replaced by a normal trend: increased film thickness with a compliant coating and reduced film thickness with a stiff coating.

The coating effect was investigated by Liu et al. (2007) in point contacts with hard EHL. The computational method for elastic deformation was based on (9) and DC-FFT fast algorithm (Liu and Wang 2002). It is found



EHL of Coated Bodies, Fig. 4 Effect of coating on central film thickness and minimum film thickness

that a stiff coating tends to increase the contact pressure but decrease the contact radius, as well as the central and minimum film thicknesses. The inverse is true for a compliant coating. However, in the thin film thickness regime, as shown in Fig. 4, the film thickness curves undergo a turning trend that tends to disrupt the general monotonic variation. The observation is similar to that in line contact (Elsharkawy and Hamrock 1994). The turning trend of the minimum film thickness is more significant than that for the central film thickness.

The turning trend in the thin film regime was investigated under different speeds, loads, rheological models, and pressure viscosity coefficients. The results indicate that high speed tends to reduce the turning trend in both the central and minimum film thickness curves. Heavy load tends to reduce the turning trend only for central film thickness. The turning trend of the minimum film thickness is enhanced by load increase. It is verified that the viscosity pressure effect is responsible for the turning trend and tends to counteract the influence of coating on the film thickness.

It is also concluded that a thin stiff coating may be utilized to reduce the friction and wear for parts subjected to conventional EHL in the elastic piezoviscous regime through film thickness enhancement. On the other hand, a thick compliant coating can more significantly affect the EHL performance in the elastic isoviscous regime than does a stiff coating. This conclusion reveals an important approach to improving film thickness and avoiding wear by using a compliant coating in artificial human joints.

Furthermore, a parametric study was conducted to investigate the variation of improvement on minimum film thickness by using a stiff coating in hard EHL (Liu et al. 2008). The various working parameters include dimensionless load parameter $M = [w/(E'R^2)][2\eta_0 u/(E'R)]^{-3/4}$, dimensionless material parameter $L = \alpha E'[2\eta_0 u/(E'R)]^{1/4}$, and coating material parameter $R_2 = E_{2c}(1-\nu_{2s}^2)/[E_{2s}(1-\nu_{2c}^2)]$, as well as coating thickness parameter $H_2 = h_{2c}/a$. Curve-fitting of numerical results indicates that the maximum increase in minimum film thickness, I_{max}, and the corresponding optimal dimensionless coating thickness, H_{2max}, can be expressed in the following forms:

$$I_{max} = 0.769M^{0.0238}R_2^{0.0297}L^{0.1376}\exp(-0.0243\ln^2 L) \quad (15)$$

$$H_{2\,max} = 0.049M^{0.4557}R_2^{-0.1722}L^{0.7611}\exp(-0.0504\ln^2 M - 0.0921\ln^2 L) \quad (16)$$

These formulas can be used to estimate the effect of coatings on film thickness for EHL applications.

Cross-References

- ▶ Contact of Layered Materials
- ▶ EHL Film Thickness Behavior
- ▶ EHL Governing Equations
- ▶ Elastohydrodynamic Lubrication
- ▶ Gear EHL Film Thickness and Wear Risk Analysis
- ▶ Gear Lubrication
- ▶ Gear Surface Treatment
- ▶ Mixed EHL in Gears

References

D. Dowson, Z. M. Jin, The influence of elastic deformation upon film thickness in lubricated bearings with low elastic modulus coatings, in *Proceedings of the 16th Leeds-Lyon Symposium on Tribology* held at Institution National des Sciences Appliquees, Lyon, France, 1989, pp. 263–269

A.A. Elsharkawy, B.J. Hamrock, A numerical solution for dry sliding line contact of multilayered elastic solids. ASME J. Tribol. **115**(2), 237–245 (1993)

A.A. Elsharkawy, B.J. Hamrock, EHL of coated surfaces: Part I – Newtonian results. ASME J. Tribol. **116**(1), 29–36 (1994)

C.J. Hooke, The elastohydrodynamic lubrication of a cylinder on an elastomeric layer. Wear **111**, 83–99 (1986)

Z.M. Jin, Elastohydrodynamic lubrication of a circular point contact for a compliant layered surface bonded to a rigid substrate. Part 1: theoretical formulation and numerical method. Proc. Inst. Mech. Eng. Part J: J. Eng. Tribol. **214**, 267–279 (2000a)

Z.M. Jin, Elastohydrodynamic lubrication of a circular point contact for a compliant layered surface bonded to a rigid substrate. Part 2: numerical results. Proc. Inst. Mech. Eng. Part J: J. Eng. Tribol. **214**, 281–289 (2000b)

Z.M. Jin, D. Dowson, J. Fisher, Minimum and central film thickness formulae for an elastic layer firmly bonded to a rigid cylindrical substrate under entraining motion. Wear **170**, 281–284 (1993)

K.L. Johnson, *Contact Mechanics* (Cambridge University Press, Cambridge/New York, 1985)

S. Liu, Q. Wang, Studying contact stress fields caused by surface tractions with a discrete convolution and fast Fourier transform algorithm. ASME J. Tribol. **124**, 36–45 (2002)

S. Liu, A. Peyronnel, Q.J. Wang, L.M. Keer, An extension of the Hertz theory for three-dimensional coated bodies. Tribol. Lett. **18**(3), 303–314 (2005a)

S. Liu, A. Peyronnel, Q.J. Wang, L.M. Keer, An extension of the Hertz theory for 2D coated components. Tribol. Lett. **18**(4), 505–511 (2005b)

Y. Liu, W.W. Chen, D. Zhu, S. Liu, Q.J. Wang, An elastohydrodynamic lubrication model for coated surfaces in point contacts. ASME J. Tribol. **129**(3), 509–516 (2007)

Y. Liu, D. Zhu, Q.J. Wang, Effect of stiff coatings on EHL film thickness in point contacts. ASME J. Tribol. **130**(3), 031501-1-5 (2008)

P. Meijers, The contact problem of a rigid cylinder on an elastic layer. Appl. Sci. Res. **18**, 353–383 (1968)

T. Nogi, T. Kato, Influence of a hard surface layer on the limit of elastic contact, Part I—analysis using a real surface model. ASME J. Tribol. **119**, 493–500 (1997)

C.H. Venner, A.A. Lubrecht, *Multilevel Method in Lubrication.* Tribology Series, 37, ed. by D. Dowson (Elsevier, Amsterdam/New York, 2000), p. 156

EHL of Joints

► Elastohydrodynamic Lubrication of Natural Synovial Joints

EHL Regime and Point Contact

► Lubrication Regimes: Point Contacts

EHL Regimes and Line Contact

► Lubrication Regimes – Line Contacts

EHL Simulations Based on Deterministic Models of Rough Surfaces

► Deterministic Models of Rough Surface EHL

EHL Solution by FEM

► Finite Element Method for Fluid Film Bearings

EHL with Rough Surface Asperity Contact

► Mixed EHL

EHL, Full Numerical Solution Methods

Wen-zhong Wang
School of Mechanical Engineering, Beijing Institute of Technology, Beijing, People's Republic of China

Synonyms

EHL numerical solution

Definition

Elastohydrodynamic lubrication (EHL) is a lubrication regime where the change in lubricant viscosity and surface elastic deformation are due to high hydrodynamic pressure.

Scientific Fundamentals

EHL is a common lubrication condition in gears, bearings, cams, and traction drivers. Two important features that distinguish it from ordinary hydrodynamic lubrication are high viscosity of the lubricant and predominant elastic deformation of surfaces. These features result from the high pressure caused by the nonconformal concentrated contact. The common governing equations that are used in the theory of elastohydrodynamic lubrication are as follows.

The pressure distribution in the lubricated domain is governed by the Reynolds equation. The general form of this expression for the EHL problem in point contact is given in (1)

$$\frac{\partial}{\partial x}\left(\frac{\rho}{12\eta}h^3\frac{\partial p}{\partial x}\right)+\frac{\partial}{\partial y}\left(\frac{\rho}{12\eta}h^3\frac{\partial p}{\partial y}\right)=V_r\frac{\partial(\rho h)}{\partial x}+\frac{\partial(\rho h)}{\partial t} \tag{1}$$

The film thickness expression, including normal approach (h_0), geometry clearance (h_s), surface roughness (δ), and elastic deformation (v), is as follows:

$$h(x,y,t)=h_0(t)+h_s(x,y)+\delta_1(x,y,t)+\delta_2(x,y,t)+v(x,y,t) \tag{2}$$

The surface elastic deformation $v(x,y,t)$ is calculated from the Boussinesq integration (Johnson 1996)

$$v(x,y,t)=\frac{2}{\pi E'}\iint_{\Omega}\frac{p(\xi,\zeta)}{\sqrt{(x-\xi)^2+(y-\zeta)^2}}\,\mathrm{d}\xi\mathrm{d}\zeta \tag{3}$$

The obtained pressure distribution should balance with the applied load through the following pressure integration over the computation domain:

$$W(t)=\iint_{\Omega}p(x,y,t)\mathrm{d}x\mathrm{d}y \tag{4}$$

Two equations of state relate the physical properties of the lubricant to the pressure and temperature generated inside the contact. The Barus pressure-viscosity relationship is commonly used for viscosity variation.

$$\eta=\eta_0 e^{\alpha p-\gamma(T-T_0)} \tag{5}$$

The Dowson-Higginson pressure-density relationship is used for density variation.

$$\frac{\rho}{\rho_0}=1+\frac{0.6\times10^{-9}p}{1+1.7\times10^{-9}p}+0.0007(T-T_0) \tag{6}$$

Equations (1)–(6) formulate the common isothermal lubrication problem. Solving this set of equations can provide the basic information for product design, such as pressure distribution and lubricant film thickness. It should be noted that the rheological behavior of lubricant in thin film and under high pressure is an issue that still needs to be addressed. For example, there are other pressure-viscosity relationships such as Roelands formula. More recently, the free-volume viscosity model has been used in EHL simulation and shows accurate prediction of the temperature-pressure-viscosity relationship of lubricant (Liu et al. 2006).

The above-mentioned governing equations well formulate the elastohydrodynamic lubrication problems. It is obvious that this set of equations represent strong nonlinearity, therefore, for this strongly nonlinear, coupled system, only solutions for very simple cases can be obtained by analytical deduction. Most solutions must be obtained by means of numerical methods.

Before a certain numerical method is applied for EHL problems, the governing equations should first be nondimensionalized by introducing the nondimensional variables. Usually, the Hertzian contact parameters are used as reference:

$$X=\frac{x}{a},\ Y=\frac{y}{b},\ \bar{t}=\frac{tV_r}{a},\ \bar{V}=\frac{V}{V_r},\ P=\frac{p}{p_{\mathrm{H}}},\ H=\frac{h}{a},\ \bar{\eta}=\frac{\eta}{\eta_0},\ \bar{\rho}=\frac{\rho}{\rho_0},\ \bar{T}=\frac{T}{T_0}$$

Thus, the governing equations can be rewritten as the following nondimensional forms based on this set of reference parameters.

$$\frac{\partial}{\partial X}\left(\varepsilon^x\frac{\partial P}{\partial X}\right)+\frac{\partial}{\partial Y}\left(\varepsilon^y\frac{\partial P}{\partial Y}\right)=\frac{\partial(\bar{\rho}H)}{\partial X}+\frac{\partial(\bar{\rho}H)}{\partial\bar{t}} \tag{7}$$

$$H=H_0+B_xX^2+B_yY^2+\bar{\delta}(X,Y,\bar{t})+V(X,Y,\bar{t}) \tag{8}$$

$$V(X,Y)=\frac{2K_e p_h}{\pi E'}\iint_{\Omega}\frac{P(\bar{\xi},\bar{\zeta})}{\sqrt{(X-\bar{\xi})^2+K_e^2(Y-\bar{\zeta})^2}}\,\mathrm{d}\bar{\xi}\mathrm{d}\bar{\zeta} \tag{9}$$

$$\iint_{\Omega}P\mathrm{d}X\mathrm{d}Y=\frac{2\pi}{3} \tag{10}$$

$$\bar{\eta}=e^{\alpha p-\gamma(T-T_0)} \tag{11}$$

$$\bar{\rho} = 1 + \frac{0.6 \times 10^{-9} p}{1 + 1.7 \times 10^{-9} p} + 0.0007(T - T_0) \qquad (12)$$

where $\varepsilon^x = \left(\frac{a p_H}{12 V_r \eta_0}\right)\left(\frac{\bar{\rho} H^3}{\bar{\eta}}\right)$, $\varepsilon^y = \frac{\varepsilon^x}{K_e^2}$, $B_x = \frac{a}{2R_x}$, $B_y = \frac{aK_e^2}{2R_y}$

Secondly, the governing equations should be converted into a discrete differential equation at each unknown pressure point. There are various schemes available for discretizing Reynolds equation. Usually, the Second-Order Central Scheme for the left-hand side terms and second-order backward scheme for the right-hand terms of Reynolds equation are used; the equation in these schemes becomes:

$$\begin{aligned}
&\frac{1}{\Delta X^2}\left[\varepsilon^x_{i+1/2,j} P_{i+1,j} - (\varepsilon^x_{i+1/2,j} + \varepsilon^x_{i-1/2,j}) P_{i,j} + \varepsilon^x_{i-1/2,j} P_{i-1,j}\right] \\
&+ \frac{1}{\Delta Y^2}\left[\varepsilon^y_{i,j+1/2} P_{i,j+1} - (\varepsilon^y_{i,j+1/2} + \varepsilon^y_{i,j-1/2}) P_{i,j} + \varepsilon^x_{i,j-1/2} P_{i,j-1}\right] \\
&\quad - \frac{3\bar{\rho}_{i,j} H_{i,j} - 4\bar{\rho}_{i-1,j} H_{i-1,j} + \bar{\rho}_{i-2,j} H_{i-2,j}}{2\Delta X} \\
&\quad - \frac{3\bar{\rho}^n_{i,j} H^n_{i,j} - 4\bar{\rho}^{n-1}_{i,j} H^{n-1}_{i,j} + \bar{\rho}^{n-2}_{i,j} H^{n-2}_{i,j}}{2\Delta X} = 0
\end{aligned} \qquad (13)$$

Other governing equations can be straightforwardly discretized into discrete form at each unknown pressure point. It should be noted that other differential schemes, such as the first backward scheme, can also be used to discretize the Reynolds equation.

Finally, a set of algebra equations at each point is constructed in the form of (14), for which certain appropriate numerical methods can be applied to obtain the numerical solutions for pressure and film thickness.

$$A_{i,j} P_{i-1,j} + B_{i,j} P_{i,j} + C_{i,j} P_{i+1,j} = F_{i,j} \qquad (14)$$

Direct Iteration Methods

Gauss-Seidal Iteration

The simplest approach to a full numerical solution to an EHL problem is the straightforward iteration method by which the famous Hamrock-Dowson (H-D) equation was obtained (Dowson and Higginson 1966). This method uses simple Gauss-Seidal (G-S) iteration on the Reynolds equation for pressure distribution. Based on an initial guess of pressure distribution, film thickness can be obtained via the calculation of elastic deformation. Once the film thickness is calculated, new pressure can be obtained by solving the Reynolds equation. In each iteration process, the film thickness is updated from the film thickness equation by using the new pressure distribution. This procedure is continuously repeated. After a number of iterations, the force balance equation (namely, the load equation) is checked, the difference between the calculated load (i.e., the integral of pressure distribution over the entire domain) and the prescribed load is used to adjust the reference film thickness h_0 in the film thickness equation. Therefore, final pressure is obtained through an under-relaxed iteration until the pressure distribution and load balance simultaneously satisfy the corresponding system equations within a prescribed accuracy. In this method, the coefficients $A_{i,j}$, $B_{i,j}$, $C_{i,j}$ appearing in (14) are always obtained from the pressure flow term of Reynolds equation and $F_{i,j}$ is always treated as known variable.

The major advantages of the G-S iterative method are small storage requirement and low computational complexity. However, the disadvantage of this method is the slow asymptotical convergence rate. Unfortunately, the method can only be applied to light loads whose maximum Hertzian pressure does not exceed 0.5 GPa. For heavy loads it is unstable. This is because the G-S relaxation usually updates the pressure guess only from the pressure flow (Poiseuille flow) term, however, under moderate and high load conditions the shear flow term will play a significant role.

Newton-Raphson Iteration: The System Method

The Newton-Raphson method for EHL line contacts was first put forward by (Okamura 1982). This method was later applied for high-pressure cases in which the maximum Hertzian pressure obtained was 4.8 GPa (Houpert and Hamrock 1986). Unlike G-S iteration, in which the unknown variables p_i are relaxed one-by-one (or line-by-line in some two-dimensional problems), the Newton-Raphson (N-R) method linearizes the discrete system equations at a approximation solution P^k based on the Newton method, and then solves simultaneously for all the unknown variables p_i. The newly obtained pressure P^{k+1} serves as a new approximation point at which the aforementioned procedure repeats until the desired convergent accuracy is satisfied. In addition, the reference film thickness h_0 is treated as an unknown variable and can be simultaneously obtained with the pressure; no additional iteration is required for load balance check. For a small number of grid points, the N-R method is rapid with a small number of iterations if the initial pressure guess is in the neighborhood of the real solution, so it is ideal for one-dimensional problems. However, Newton linearization involves calculating and inversing the Jacobian matrix, which consists of the derivatives of all discrete

equations with respect to all the unknown variables. Because film thickness at one point is related to all nodal pressures in the computation domain, the Jacobian matrix of pressure derivatives is a full matrix that takes large memory to store proportional to the square of the number of grid points and need to be inverted, which requires long computing time proportional to the cube of the number of grid points. Additionally, this method is very sensitive to the initial guess. If the initial guess is not good enough, it will suffer convergence difficulty. Due to these limitations, it is not applicable to point contact problems.

Semi-system Method

Most direct iteration methods will suffer numerical instability under heavily loaded conditions. In-depth studies show that the numerical instability of the G-S method results from the update of pressure in the iteration process only through pressure flow (Poiseuille) term. Under heavily loaded conditions, lubricant in the contact zone is pressurized and the viscosity increases almost exponentially with the pressure. Near the contact center, the viscosity of the lubricant can be several orders higher than the ambient value. As a result, the pressure flow term, which represents the lubricant flow due to the hydrodynamic pressure gradient, as expressed by the left terms in the Reynolds equation, becomes much weaker because of the third power of the tiny film thickness and exponentially increased viscosity, and instead, the shear flow term takes over. In conventional G-S iteration, when constructing the iteration scheme, the coefficient matrix always only considers the contribution of pressure flow terms. This will result in the gradual deterioration of the characteristic that the element of leading diagonal of coefficient matrix is dominant, therefore, the convergence criteria based on the dominance theorem are not met and there is an ill-conditioned coefficient matrix.

As mentioned, shear flow term plays a very important role as the load conditions become heavy. Incorporating the contribution of the shear flow term into the coefficient matrix will help to keep the characteristic that elements of the leading diagonal are dominant. Thus, the numerical stability will be regained. This is called a semi-system approach (Ai 1993); its iteration process is as follows.

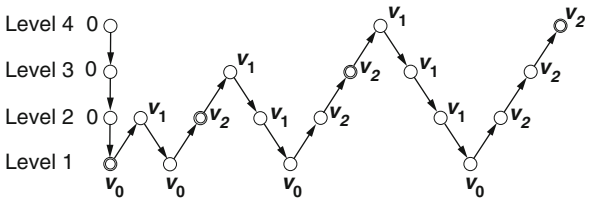
The semi-system method is similar to the direct method. First, the coefficient matrix for the iteration scheme is constructed based on the contribution from both pressure flow and shear flow, in which the dominance of the element of the leading diagonal should be ensured. Given an initial pressure guess, film thickness can be obtained according to (2) and (3) using the numerical methods for elastic deformation. Once film thickness is obtained, discrete equations of Reynolds equation (1) are solved by line relaxation with an under-relaxation factor that results in a new pressure distribution. Then initial pressure and new pressure are compared in order to calculate a new initial pressure. The above loop is repeated until the relative error between two pressures is smaller than a prescribed precision. The modification of normal approach, h_0, is based on load balance and is performed during the pressure iterations. Generally, after a number of pressure relaxations (for example, every four iterations), h_0 should be corrected once to improve convergence speed.

The semi-system method partly combines the simplicity of G-S iteration with the efficiency of the N-R method. However, since the coefficient matrix of the resulting system equations is a full matrix, large storage is required. However, if only the pressure near the points to be solved is extracted to construct the iteration scheme, the coefficient matrix can be simplified into a tri-diagonal matrix (Ai 1993), which significantly reduces the storage requirement and computation complexity.

Multi-grid (MG) Method

The multi-grid (MG) method has been commonly used in computational fluid dynamics for many years. It was implemented in the EHL analysis first by (Lubrecht 1987), then by (Venner 1991) and others. The MG algorithm is described in detail in (Venner and Lubrecht 2000), together with a source code written in C language. The multilevel method is currently the most popular method in EHL analysis. Error analysis is the basis of this method for fast convergence. The error between the initial value and final solution can be divided into high-frequency components and low-frequency components. Errors with high frequency can be removed quickly by relaxation on the fine mesh, while errors with low frequency can be removed quickly by relaxation on the coarse mesh. Thus, during the MG solution process, the computational mesh is constantly changed from level to level, going up and down with either a V-cycle or a W-cycle, in order to accelerate the solution convergence. Through the repeated transitions between coarse mesh and fine mesh, the total error is removed quickly. In addition, EHL behavior at high pressures is dominated by elastic deformation. This behavior leads to two difficulties. First is a loss of coupling in the direction perpendicular to the flow; second is an accumulation of change in the elastic deformation. These two difficulties can be addressed by line relaxation and distribution relaxation, respectively (Venner 1991).

There are several actual schemes and algorithms for carrying out the multi-grid method. The so-called F-cycle



EHL, Full Numerical Solution Methods, Fig. 1 Schedule diagram of F-cycle for the MG method

algorithm for transient problems is precisely described as follows (Ai 1993).

Consider a time-dependent problem with the discrete equation on grid K given by

$$A^K u^K = f^K \tag{15}$$

where K denotes the grid level and the highest level (target grid) in (15), the matrix A is the coefficient matrix obtained from the discretization of the Reynolds equation.

The solution process is started with an initial solution $\hat{u}^{\text{highL}}$.

1. Restrict the initial solution $\hat{u}^{\text{highL}}$ from the finest grid (K = high L) straightforward to the coarsest grid ($K = 1$), obtaining $\hat{u}^K$.
2. Perform v_0 relaxations on the coarsest grid ($K = 1$), yielding $\tilde{u}^K$.
3. Calculate the approximation $\bar{u}^{K+1}$ according to (16), and then set $K = 2$

$$\bar{u}^{K+1} = \hat{u}^K + I_K^{K+1}\left(\tilde{u}^K - I_{K+1}^K \tilde{u}^{K+1}\right) \tag{16}$$

4. Conduct v or w cycle for the coarse grid ($K = 2$), yielding the new approximation $\tilde{u}^K$.
5. Repeat steps 3 and 4 until grid K is up to the finest grid (target grid).

Figure 1 shows the schedule diagram of F-cycle in which four grid levels are used, with the solution on each level improved by a v-cycle algorithm.

Other schemes and algorithms of the MG method can be found in (Venner and Lubrecht 2000; Ai 1993). It is believed that, according to the literature, the computation speed with the MG method may be considerably increased, compared with that from a high-level fixed mesh, if the EHL film is thick and the surfaces are smooth.

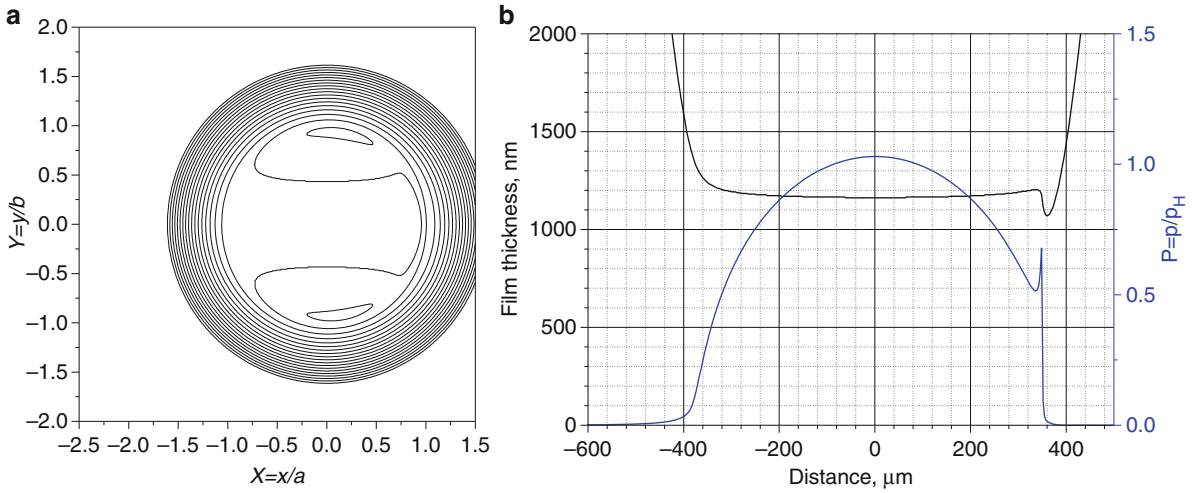
Progressive Mesh Densification (PMD) Method

The progressive mesh densification (PMD) method has been proposed and implemented in the mixed EHL model described in (Hu and Zhu 2000; Zhu 2007). The method utilizes a set of grids to speed up the convergence rate similar to the MG method. In the MG method, the coarse grids serve as correction grids where the compensations or corrections are calculated and then used to improve the solution on the fine grids. However, in the PMD method, the coarse grid is used to generate a good initial guess to the finest grid. A well-known technique to speed up the overall convergence process is to perform some preliminary iterations on a coarse grid and then use the resulting approximation as an initial estimate on the original fine grid. The relaxation on a coarse grid is less expensive because there are fewer unknowns to update and the relaxation on the coarse grid has a marginally improved convergence rate. Generally, with the PMD method the computation is started on a much coarser mesh level, 129 × 129, for example, or even 65 × 65, in order to quickly get the first approximation. As soon as the solution becomes reasonably convergent with a relatively loose convergent criterion, say pressure convergent criterion ER_p and load convergent criterion ER_w smaller than 0.0005–0.001 or so, the computation will be automatically switched to the next higher level. This procedure is repeated until the ultimate solution on the prescribed finest mesh level is reached with a tightened convergence criterion.

Comparing the processes of solution based on PMD method and ML method, although both methods use a set of grids, the PMD method is unidirectional and the ML method is bidirectional. The PMD method starts from a coarse grid and proceeds straightforwardly to the finest grid. The ML method may start from coarse grid or fine grid according to the strategy and always jumps between coarse grid and fine grid during the solving process. This difference of PMD and ML methods results in very different convergent property when handling lubrication problems of rough surfaces and thin film lubrication. As the surface roughness represents different profiles on the coarse grid and fine grid, when the computation is dancing up and down among different mesh levels in the ML method, the computation is actually dealing with different surfaces on different mesh levels and targeting different solutions, the solving process may certainly become unstable. However, in the PMD method, the iteration on a coarse grid just as quickly provides a good approximation for the finest grid, and as the process is unidirectional, the approximation will gradually approach the desired solution.

Coupled Differential Deflection Method

A common numerical strategy to the full numerical solution of EHL, as mentioned above, is a sequential solution



EHL, Full Numerical Solution Methods, Fig. 2 Typical full numerical solution for an EHL problem (p_H = 2.8 GPa, $a = b$ = 0.369 mm, v_r = 3 m/s). (**a**) Film thickness contour plot (**b**) Profiles of pressure and film thickness at $y = 0$

of the Reynolds equation and the elastic deflection equation. That means that, for an initial approximation to the solution, the elastic deflection within the contact is first evaluated to determine the oil film thickness, then the discrete Reynolds equation is solved for a new approximation to the solution. Consequently, the elastic deflection is re-evaluated based on the update pressure approximation. This procedure is repeated until a converged result is obtained that satisfies the Reynolds equation and load balance simultaneously. As the nature of the elastic deformation equation at each point in a computational grid is related to hydrodynamic pressure at all points on the computation grid, which results in a full matrix of pressure-deformation coefficient in numerical solutions, solving the hydrodynamic and elastic deflection equations simultaneously as a coupled pair seems impractical.

Very few efforts have been made to fully couple the Reynolds equation and elastic deformation equation within an EHL solution scheme. A breakthrough to address the issue of problems with many points was developed (Evans and Hughes 2000). A differential form of deformation equation was derived that can be exploited to effectively fully couple the elastic and hydrodynamic equations. When the original integration equation of elastic deformation is transformed into a differential equation, the effect of pressure in the differential form is extremely localized compared with the original integrated form. Thus, the matrix of pressure-deformation coefficient can be simplified into a bandwidth matrix that enables a restricted bandwidth elimination solver to be used for line contacts and a coupled iterative technique for point contacts. Based on this property, the elastic and hydrodynamic equations can be effectively fully coupled and solved simultaneously as a coupled pair. By this coupled differential deformation method, point-contact lubrication problems have been successfully solved (Holmes 2002).

Inverse Iteration Methods

The inverse method was first put forward for line contact (Dowson and Higginson 1959) and later applied for point contact (Evans and Snidle 1981). In the inverse iteration method, the Reynolds equation is integrated to solve for film thickness $\boldsymbol{h}$ for a given pressure distribution $\boldsymbol{P}$, which is completely different from the direct method. The solution process of this method consists of three steps. First, a pressure distribution is assumed. Based on this assumed pressure distribution, a film thickness distribution defined as hydrodynamic film thickness is obtained from the integration form of the Reynolds equation. Then another film thickness, defined as elastic film thickness, is obtained through the film thickness equation for the same pressure distribution. Next, these two film shapes are compared; the difference between two film thicknesses determines a residual function. Finally, the residual function is used to modify the initial assumed pressure distribution. The procedure repeats until the two film shapes agree within a preset convergence criteria. As the integral form of Reynolds equation is a cubic equation of film thickness, there will be three solutions for hydrodynamic film thickness when solving the integral form of Reynolds

equation. The choice of solution requires more physical knowledge and will complicate the programming. The inverse method has been used to handle one-dimensional line contact and point contact under heavy load conditions, but is unstable for light loads.

Key Applications

Tribological Design of Components

For the structure design of components, many advanced technologies have been developed. For example, as FEM/CAD technologies have been well developed with commercial software packages readily available, component structure strength can now be quickly and accurately predicted using computers. Product design cycles have been shortened greatly. It has been found, however, that in reality about 80–90% of component failures fall in the surface failure category. Unfortunately, so far no commercial software is available that can help engineers to accurately analyze surface strength of various components. In fact, modeling tribological problems and predicting lubrication performance, interface performance, strength, and life often appear to be a bottleneck in advanced product design and development. The numerical solution of EHL can be applied in the design of elements in order to provide a good fundamental base for surface failure and friction/efficiency analysis. The early famous Hamrock-Dowson film thickness plays an important role in lubrication analysis of components, which is fitted from full numerical solutions of typical EHL problems. With the development of computers, full numerical solutions are often applied in tribological design such as gears, bearings, and engine frictional pairs under real working conditions and with measured surfaces. Figure 2 shows the typical full numerical solution of an EHL problem.

Nomenclature

a, b	Hertz contact radius in the x and y directions (m)
$D^{i,j}_{k,l}$	Deformation influence coefficient at point (i, j) owing to a unit load acting on positing (k, l)
$A_{i,j}$, $B_{i,j}$, $C_{i,j}$, $F_{i,j}$	Coefficients for discrete Reynolds equation
E'	Effective Young's modulus (Pa)
ER_p	Convergence criterion for pressure
ER_W	Convergence criterion for load
h, H	Film thickness and nondimensional film thickness
h_0, H_0	Normal approach of two stiff surface (m) and dimensionless normal approach, $H_0 = h_0/a$
K_e	Ellipticity, $K_e = b/a$
i, j	Position (x_i, y_j)
R_x, R_y	Reduced radius of curvature in x and y direction
p	Contact pressure (Pa)
p_H	Maximum Hertzian pressure (Pa)
$t, \bar{t}$	Dimensional (s) and nondimensional time
T, T_0	Temperature of lubricant and ambient temperature (K)
v	Normal surface deformation
V_r	$=(V_1 + V_2)/2$, Entrainment velocity, (m/s)
W	Applied load, (N)
x, y, X, Y	Coordinate in the x and y direction and nondimensional counterparts
$\Delta x, \Delta y$	Dimensions of an element in the x and y direction
α	Pressure-viscosity coefficient (Pa^{-1})
γ	Temperature-viscosity coefficient (K^{-1})
η, η_0	Viscosity and ambient viscosity of lubricant (Pa s)
ρ, ρ_0	Density and ambient density of lubricant (kg/m^3)
δ	Surface roughness height
Tidal($\tilde{}$)	New approximation of variables

Cross-References

- ▶ 3D Line Contact EHL
- ▶ Differential Scheme Effect on EHL Solution
- ▶ EHL Governing Equations
- ▶ Lubricant Non-Newtonian Effect on EHL
- ▶ Mesh Density Effect on EHL Solution
- ▶ Mixed EHL
- ▶ Plasto-Elastohydrodynamic Lubrication (PEHL)
- ▶ Point Contact EHL
- ▶ Surface Deformation Calculation for EHL
- ▶ Thermal EHL Theory

References

X.L. Ai, Numerical analysis of elastohydrodynamically lubricated line and point contacts with rough surfaces by using semi-system and multigrid methods. Ph.D. thesis, Northwestern University, IL, 1993

D. Dowson, G.R. Higginson, A numerical solution to elastohydrodynamic problem. J. Eng. Sci. **1**, 6–15 (1959)

D. Dowson, G.R. Higginson, *Elasto-Hydrodynamic Lubrication* (Pergamon, Oxford, 1966)

H.P. Evans, T.G. Hughes, Evaluation of deflection in semi-infinite bodies by a differential method. Proc. Instn. Mech. Eng. C **214**(4), 563–585 (2000)

H.P. Evans, R.W. Snidle, Inverse solution of Reynolds equation of lubrication under point contact elastohydrodynamic conditions. ASME J. Lubr. Tech. **103**(4), 539–546 (1981)
M.J.A. Holmes, Transient analysis of the point contact elastohydrodynamic lubrication problem using coupled solution methods. Ph.D. thesis, Cardiff University, UK, 1993
L.G. Houpert, B.J. Hamrock, Fast approach for calculating film thickness and pressure in elastohydrodynamically lubricated contacts at high loads. ASME J. Tribol. **108**(3), 411–420 (1986)
Y.Z. Hu, D. Zhu, A full numerical solution to the mixed lubrication in point contacts. J. Tribol-T ASME **122**, 1–9 (2000)
K.L. Johnson, Contact Mechanics, Cambridge University Press, 1996
Y. Liu, Q.J. Wang, W. Wang, Y. Hu, D. Zhu, M. Hartl, EHL simulation using the free-volume viscosity model. Tribol. Lett. **23**(1), 27–37 (2006)
A.A. Lubrecht, The numerical solution of elastohydrodynamic lubricated line and point contact problems using multigrid techniques. Ph.D. thesis, University of Twente, The Netherlands, 1987
H.A. Okamura, Contribution to the numerical analysis of isothermal elastohydrodynamic lubrication, tribology of reciprocating engines, *Proceedings of the 9th Leeds-Lyon symposium on Tribology.* Butterworths, Guilford, England. 1982, pp. 313–320
C.H. Venner, Multilevel solution of the EHL line and point contact problems. Ph.D. thesis, University of Twente, The Netherlands, 1991
C.H. Venner, A.A. Lubrecht, *Multilevel Methods in Lubrication* (Elsevier, Amsterdam, 2000)
D. Zhu, On some aspects of numerical solutions of thin-film and mixed elastohydrodynamic lubrication. Proc. Inst Mech. Eng. J. **221**, 561–579 (2007)

EHL-Line Contact

► Film Thickness Formulas: Line Contacts

EHL-Point Contact

► Film Thickness Formulas: Point Contacts

Eigenstrain Method

► Micromechanics for Contact Applications

Elasticity and Creep

► Creep

Elasticity and Creep-Fatigue

► Creep-Fatigue

Elasticity and Fatigue

► Fatigue

Elasticity and Stiffness

► Rolling Bearing Stiffness

Elasticity for Closely Conformal Contact Interface

Q. Jane Wang[1], Shangwu Xiong[2]
[1]Department of Mechanical Engineering and Center for Surface Engineering and Tribology, Northwestern University, Evanston, IL, USA
[2]Department of Mechanical Engineering, Northwestern University, Evanston, IL, USA

Synonyms

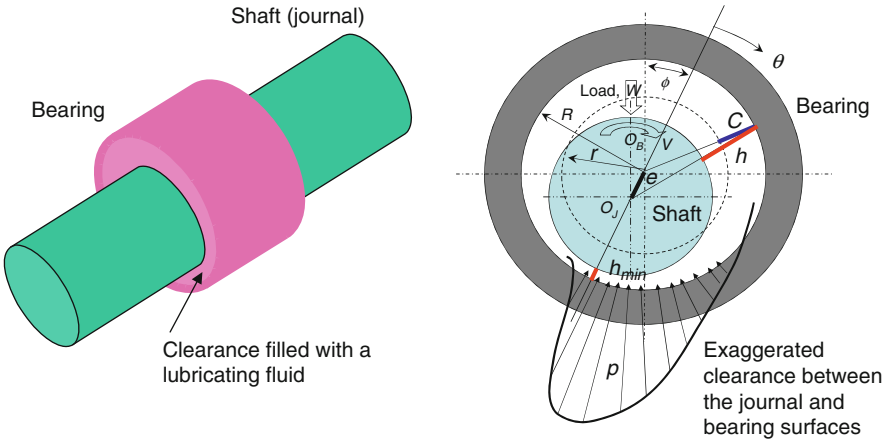
Contact elasticity for journal-bearings

Definition

Elastic solution to the contact problem of closely conformal surfaces for contact pressure and area, surface deformation, normal approach, and subsurface stresses.

Scientific Fundamentals

A closely conformal contact (Fig. 1) usually happens between a pair of convex and concave surfaces of nearly the same radii of curvatures, such as the contact between the outer surface of a journal (shaft, which is convex) and the inner surface of a sleeve bearing (which is concave). The name of the conformal contact is relative to that of non-conformal contact, either the counterformal contact between two convex surfaces or the contact of nominally flat surfaces, for which the surface deformation can be described by the integral of a Green's function. Loading to closely conformal contact elements causes both surface deformation and structural deflection, and, due to the influence of the latter, a conformal contact cannot be



Elasticity for Closely Conformal Contact Interface, Fig. 1 Closely conformal contact of elements between a pair of convex and concave surfaces of nearly the same radii of curvature, such as the outer surface of a shaft and the inner surface of a sleeve bearing. Here, O_J and O_B are the centers and r and R are the radii of the journal and the bearing, respectively, C is the radial clearance, V the speed, W the load, and ϕ the attitude angle. Pressure p and film thickness h are functions of location θ

treated locally. The finite element method is commonly used in the elasticity study of conformal-contact elements. Because the solution to such a conformal contact problem is usually structure-dependent, no general formulas can be derived for either the contact area or the contact pressure distribution except for simplified cases.

Receding Contact of Closely Conforming Solids

Strictly speaking, whether a contact is conformal or not is determined by whether the contact is localized or not and whether the half-space solution method based on Green's functions can be applied. The cross-sectional drawing in Fig. 1 exaggerates the clearance between the inner surface of the bearing (or the bearing bore) and the outer surface of the shaft. The radial clearance, C, measured in the radial direction, is determined by the diameter tolerances specified with the fit of the two. For example, if the bearing bore diameter is $D^{+t_1}_{+0.000}$ with tolerance t_1, and the shaft diameter is $d^{-a}_{-a-t_2}$ with standard deviation a and tolerance t_2, for $D = d$, R = D/2, and r = d/2, the radial clearance is $C = (R + t_1/2) - (r - (a - t_2)/2) = (t_1 + a + t_2)/2$. Such a small dimensional difference makes the two surfaces in a theoretically receding contact, which is so defined that the loaded contact area is completely contained within the unloaded contact area (Johnson 1985). This contact area variation is characteristically different from that of a counterformal contact, where loading causes the contact area to increase. Figure 2 illustrates the contact areas of these two cases.

Issues of Conformal Contact Elasticity

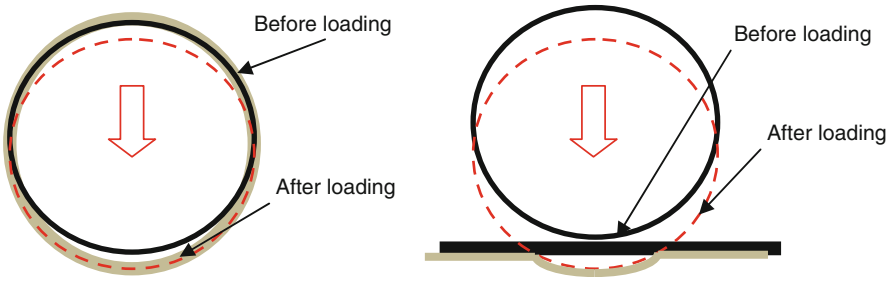
It may be a good idea to review the mechanics of deformation for non-conformal contact first in order to better understand the mechanics for bodies in closely conformal contact. Counterformal contact of two convex bodies usually yields a contact area whose dimension is much smaller than the characteristic size of the contacting bodies. As a result, the contact is highly localized, and the bodies can be treated as half spaces. For plain-strain problems, the normal deformation, du_z, at point x over a half plane caused by a concentrated line load, P, at $x = 0$, can be mathematically expressed by the Flamant solution (Johnson 1985) with E Young's modulus and v Poisson's of the material, and c an integration constant.

$$du_z(x) = -\frac{2(1-v^2)P}{\pi E}\ln|x| + c \qquad (1)$$

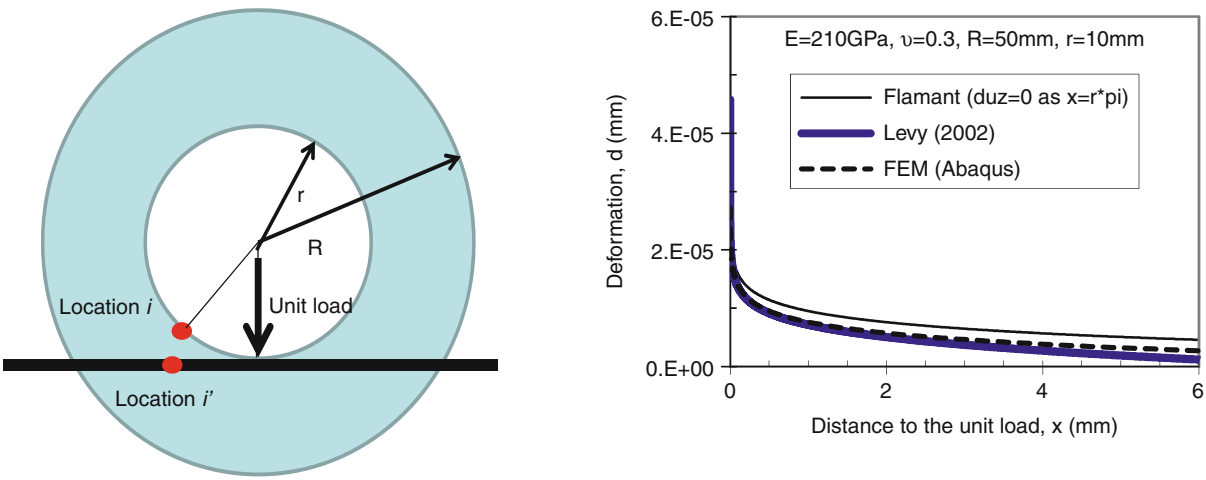
Likewise, the deformation at point (x, y) over a half space caused by point load p in a three-dimensional problem can be expressed as follows by the Boussinesq equation (Johnson 1985).

$$du_z(x,y) = \frac{(1-v^2)}{\pi E}\frac{p}{\sqrt{x^2+y^2}} \qquad (2)$$

The application of Equations (1) and (2) implies that the solid is fixed at the infinite and that the response to a unit load at any location is the same anywhere as long as the distance to the loading point is the same. The latter means a reciprocal condition for the excitation and response. However, concave structures in conformal



Elasticity for Closely Conformal Contact Interface, Fig. 2 Receding (*left*) conformal contact and non-receding counterformal contact. The *solid black* and *dashed circles* are for the shaft (convex) before and after loading. With the former, the mating concave surface deforms along the loading direction, leaving a gap on the unloaded side. With the latter, the convex surface presses to the mating flat upon loading; and the contact area is a dent whose size increases with load



Elasticity for Closely Conformal Contact Interface, Fig. 3 Normal deformations at *i* (on the inner surface of the cylinder) and *i'* (on the surface of the half plane) caused by a unit load, calculated with the finite element method and the Flamant solution for a circular domain (Levy 2002). The outer surface of the cylinder is assumed to be fixed where no deformation is allowed. The Flamant solution for a half plane (1) for the latter is also plotted for reference. Note that the deformation at *i* is affected by the ratio of R/r

contacts may invalidate these conditions because, in most of these cases, the entire element structure and part-specific boundary conditions need to be considered, and generally, the reciprocal condition is not satisfied. The contact of closely conformal solids is not localized.

The difference can be demonstrated with a simple comparison of the cases shown in Fig. 3, which plots the normal deformations at locations *i* (on the cylinder inner surface) and *i'* (on the half plane surface) caused by a unit load, calculated with the finite element method for plain-strain problems. The distances from the load to both points are the same. The outer surface of the cylinder is assumed to be fixed where no deformation is allowed. This boundary still permits the reciprocal condition; however, the structural curvature plays a role. In these cases, E = 210 GPa, $\nu = 0.3$, r = 10 mm, and R = 50 mm. It is shown that the finite element result is notably farther away from those obtained with the classic Flamant solution for a half plane given in (1) at larger x but closer to those by the modified Flamant solution for a circular domain (Levy 2002), which is given as:

$$du_z(\theta) = \frac{P}{4\mu\pi}[(1-2\nu)(\pi-\theta)\sin\theta + (2-2\nu)\ln(1-\cos\theta)\cos\theta] \quad (3)$$

where $\theta = \arccos(x/r)$ as $x \geq 0$ while $\theta = 2\pi - \arccos(x/r)$ as $x < 0$.

The thin-liner model based on the plane-strain assumption (Higginson 1965) can be a reasonable approximation to the radial deformation of a soft and thin compliant shell on a rigid bearing housing, subjected to pressure p, which is given below:

$$du_z(x, y) = (R - r)\frac{(1 + v)(1 - 2v)p}{E(1 - v)} \tag{4}$$

Finite Element Method for Deformation Analysis

Advancing with the rapid development of computers and computation technologies, the finite element method (FEM) and the FEM-based influence-function method have become one of the most popular techniques for the computation of conformal elastic deformations of conformal contact elements. The finite element discretization of a bearing element leads to the following set of linear algebraic equations with **d** for the matrix of displacement at all the nodes, **K** for the stiffness matrix, and **f** for the external nodal forces (Oh and Huebner 1973).

$$\mathbf{Kd} = \mathbf{f} \tag{5}$$

If body forces (e.g., gravity and inertia arising from accelerating) and initial strains are neglected, the components of **f** are zero except for those corresponding to the nodes with prescribed external forces or displacements. For a dry contact problem, **f** is found from the contact and friction constraints and then **d** is obtained by solving (5). When dealing with a lubrication problem, **f** is found from the nodal fluid pressure $\boldsymbol{p}$ while the film thickness in the Reynolds equation is a function of **d**. When using the decoupled direct FEM, once $\boldsymbol{p}$ is solved from the Reynolds equation, one can obtain the elastic deformation of the whole element structure, such as a bearing sleeve, a housing, directly by solving (5). In the fully coupled direct FEM, $\boldsymbol{p}$ can be obtained by solving (5) and the Reynolds equation simultaneously.

Influence Coefficient Method

The analysis for a lubricated journal bearing requires solutions of the Reynolds equation supported by calculations of surface deformation that contribute to the film thickness. Solving the Reynolds equation usually involves iteration, in which the pressure, or the external nodal force term in the above equation, is a changing quantity due to iterative updating, and so is the surface deformation caused by the pressure. Solving (5) each time in the iteration with the finite element procedure can be time consuming. On the other hand, deformation can be solved with the summation method supported by influence coefficients, as shown below with $I(i, j, k, l)$, for the deformation at point (k, l) subjected to force f at point (i, j).

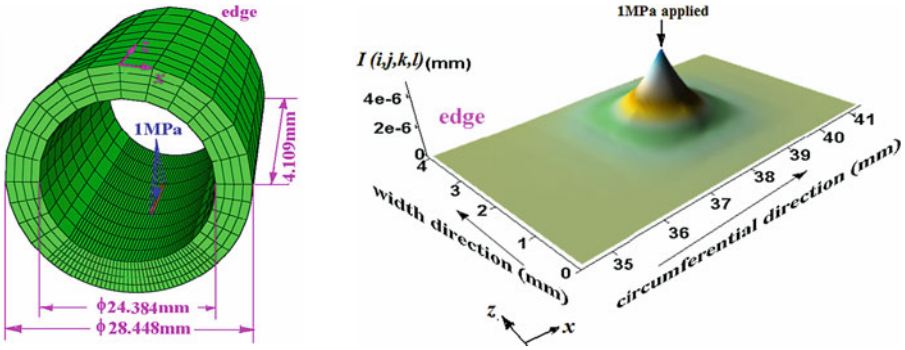
$$u_z(k, l) = \sum_{i,j=1}^{m} I(i, j, k, l) f(i, j) \tag{6}$$

Influence coefficients $I(i, j, k, l)$ can be obtained with the finite element method off the line of the Reynolds equation iteration, to be done once for each solid structure by applying a unit load on each surface node one by one.

Simplified Rapid Computation Methods

The accuracy and efficiency of the solutions for $I(i, j, k, l)$ for a conformal contact problem is affected by mesh density. The following three aspects deserve attention: (a) computational time to obtain $I(i, j, k, l)$ by the 3D FEM; (b) storage requirement of $I(i, j, k, l)$; and (c) computational cost to calculate displacement using (6). If the mesh size is large, the size of the full matrix of $\boldsymbol{I}$ can be a significant burden to storage and computation. Several efficient approximation methods have been suggested to relieve the aforementioned burdens (Dobrica et al. 2008; Xiong et al. 2010). These methods make use of the well-known facts that the displacement, $I(i, j, k, l)$ at loading node (i,j), $I(i,j,i,j)$, has the largest value among those of all other nodes but it drops sharply with distance and becomes negligibly small for most parts of the entire region (Fig. 4) (Xiong et al. 2010).

Moving Grid method. Dobrica et al. (2008) proposed a moving grid method using two different mesh levels (a fine mesh and a coarser mesh) to compute elastic deformation approximately. According to their algorithm, in which local deformation in the neighborhood of each loaded node is calculated with the fine mesh while the deformation outside this neighborhood is computed on the coarse meshes. They defined the "neighborhood" as the set of fine-mesh elements enclosed within the coarse mesh elements adjacent to each loaded point. First, a sweep of the coarse mesh is performed and an equivalent pressure is obtained by interpolation from fine mesh solutions using the area weight averaging technique; the deformation considering pressure from all coarse elements, $\bar{\mathbf{u}}$, is then calculated. Subtracting the deformation caused by the coarse elements, $\mathbf{u}'$, in its neighborhood (e.g., 3×3 coarse elements shaded in Fig. 4) from $\bar{\mathbf{u}}$, one can get subsequently the deformation contributed from the coarse meshes outside the neighborhood (i.e., $\bar{\mathbf{u}}' = \bar{\mathbf{u}} - \mathbf{u}'$). Thereafter, one can interpolate $\bar{\mathbf{u}}'$ linearly from the coarse mesh to the fine one. The displacements contributed from



Elasticity for Closely Conformal Contact Interface, Fig. 4 Elastic deformation caused by a bi-linearly distributed pressure (1 MPa acting at a point (x = 38.302 mm, z = 2.359 mm)). One half of the bearing is analyzed by FEM (Xiong et al. 2010)

the local "neighboring" pressures, $\mathbf{u}^*$, are computed in the second sweep of the domain. Finally, the actual total deformation, u, is obtained by $\mathbf{u} = \bar{\mathbf{u}}' + \mathbf{u}^*$. This technique requires the influence factors for both the fine and course mesh.

Selective fine mesh with selective storage. Xiong et al. (2010) proposed a selective storage technique that reduces the computational cost in (6) and the storage requirement for ***I***. The two main differences compared with the aforementioned method by Dobrica et al. (2008) are (a) only one mesh set (finer) was used and inhomogeneous mesh sizes might also be applied; and (b) a criterion was suggested to select the stored region whilst the values for other points in the remaining region were truncated. The criterion is $|I(i,j,k,l)/I(i,j,i,j)/| < \hat{e}_{min}$ (e.g., $\hat{e}_{min} = 10^{-4}$). However, the computational cost of this approach to obtain ***I*** was the same as that in the conventional approach because the total n_p nodes have to be used in the elastic displacement calculation of each node (i,j). Recently a further developed method, called the selective-fine-mesh with selective storage, has been proposed to tackle the aforementioned computational efficiency problems. A special technique that combines selective fine mesh with selective storage mapping can be used for the cases of cylindrical bearings with uniformly constrained external surfaces or other axisymmetric boundary conditions. These approaches are illustrated in Fig. 6, and the comparisons of the computational cost and the storage requirement are shown in Table 1 (Xiong et al. 2010).

Key Applications

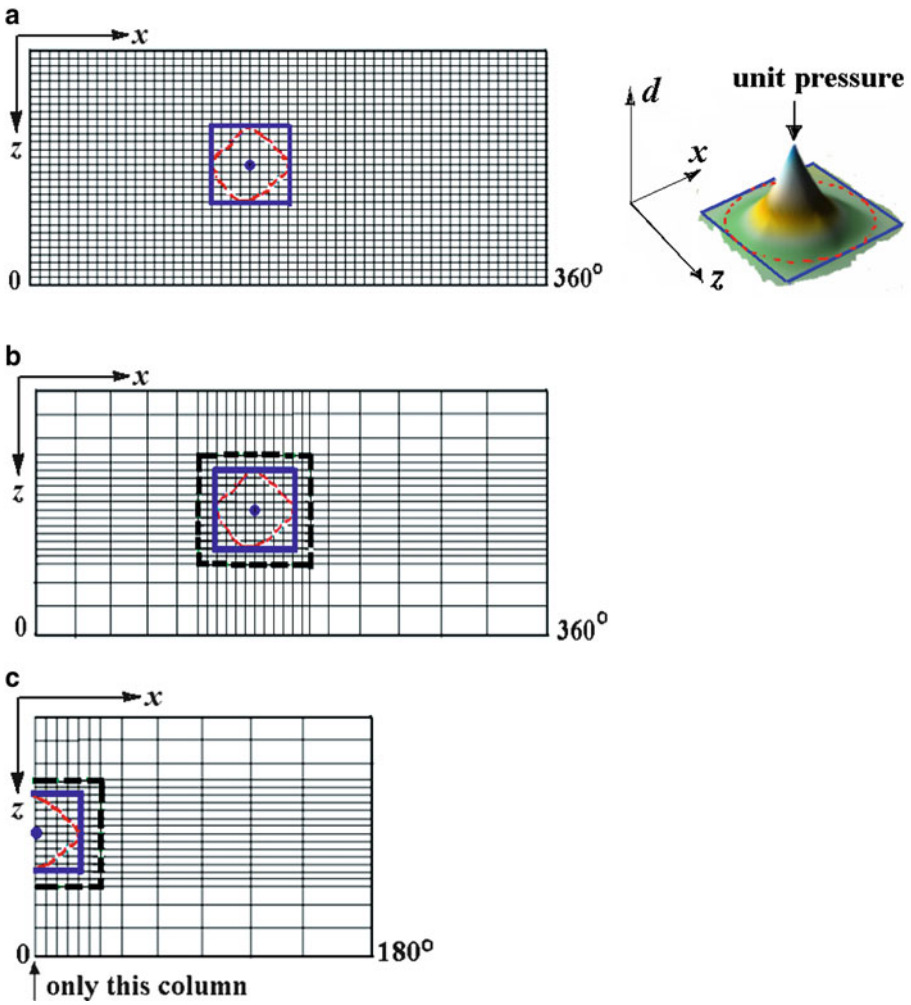
Journal Bearing Elasticity

A common closely conformal-contact system is a journal bearing, as shown in Fig. 1, with the bearing radius, R = D/2 and shaft radius r = d/2, but nominally, D = d. The centers of the bearing and the shaft are at O_B and O_J with an eccentricity, e. The applied load, W, and operation speed, V, as well as the material elastic properties and lubricant rheological properties, further decide the relative position of the shaft center with respect to that of the bearing, which is reflected by the shaft center offset, or the eccentricity, e, the loading angle, or the attitude angle, ϕ, and the distribution of the film thickness, h. The smallest value of the film thickness is referred to as the minimum film thickness, h_{min}, and the ratio of the eccentricity over the radial clearance is named as the eccentricity, $\varepsilon = e/C$. With the assistance of the radial clearance and eccentricity ratio, the film thickness can be expressed as $h = C(1 + \varepsilon \cos\theta)$. The film thickness can be calculated from the deformed average gap, and the flows and asperity contact can be treated in two separate sub-models, namely, a flow model for lubricant hydrodynamics and an off-line contact model for asperity contact pressure. The average Reynolds equation shown below, derived by Patir and Cheng (1978), shows the relationship between film thickness, h, average gap, h_T, and hydrodynamic pressure, p:

$$\frac{\partial}{R\partial\theta}\left(\phi_\theta \frac{\rho h^3}{\eta}\frac{\partial p}{R\partial\theta}\right) + \frac{\partial}{\partial z}\left(\phi_z \frac{\rho h^3}{\eta}\frac{\partial p}{\partial z}\right) = \left(6V\frac{\partial \rho h_T}{R_B \partial\theta}\right) + 6V\sigma\frac{\partial \rho \phi_s}{R_B\partial\theta} \tag{7}$$

where ϕ_θ and ϕ_z are pressure-flow factors, and ϕ_s is the shear-flow factor.

Elasticity affects journal bearing performance through its influence on the film thickness, as shown in the following equation, where h is expressed as the sum of the



Elasticity for Closely Conformal Contact Interface, Fig. 5 Illustration of three approaches: (**a**) selective storage; (**b**) selective fine-mesh with selective storage; and (**c**) combined selective fine-mesh with selective-storage mapping. ($|I(i,j,k,l)/I(i,j,i,j)/| < \hat{e}_{min}$ inside the diamond-like region, *I(i, j, k, l)* for nodes in the rectangular-storage block (*thick solid line*) is stored, and the finer mesh is used inside the *thick dashed* block (Xiong et al. 2010)

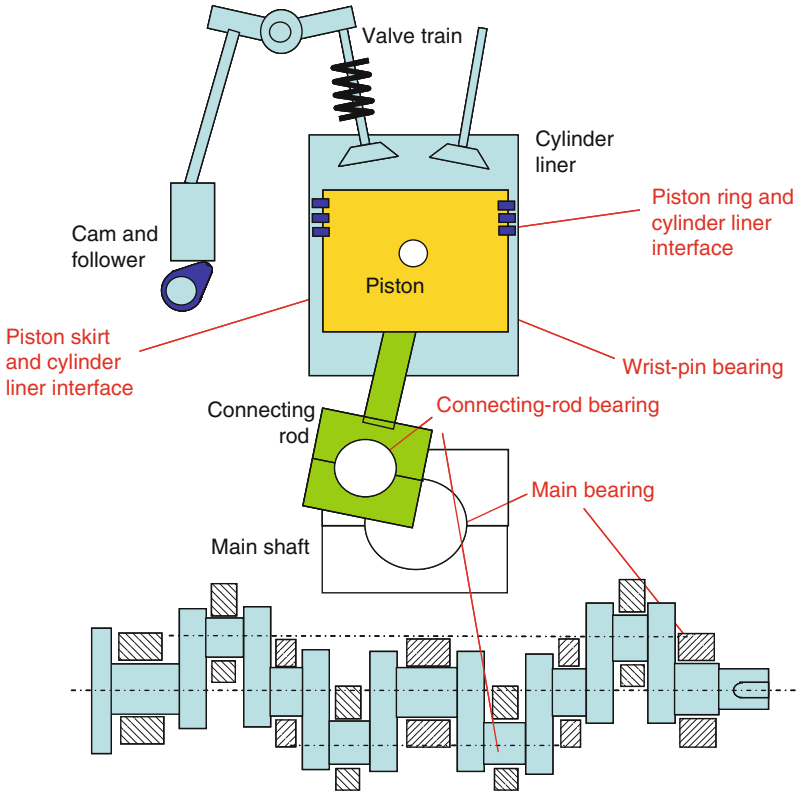
nominal clearance, h_0, the average asperity heights, δ_1 and δ_2, and the thermal-elastic deformation of the journal and the bearing, δ_J and δ_B. Deformation may be considered in two parts: deflection of the center surface (surface formed by the middle points across roughness heights of a three-dimensional topography, analogous to the center-line for a two-dimensional roughness), and deformation of the asperities with respect to the center surface. The latter affects micro flows and should be considered in the average flow (AF) model, while the former is a part of structural deformation and should be analyzed through a macro-scale computation scheme (Shi and Wang 1998).

$$h = h_0(\theta) + \underbrace{\delta_J(\theta, \Delta T, p) + \delta_B(\theta, z, \Delta T, p)}_{\text{to be considered in macro scale}} + \underbrace{\delta_1(x, y, t) + \delta_2(x, y, t)}_{\text{to be considered in the AF model}} \tag{8}$$

Here, the temperature can be determined using the energy equation, and thermal distortion can be solved based on linear thermoelasticity (Pinkus, 1990).

Engine Bearing and Ring Elasticity

The powertrain of an internal combustion engine has several conformal contact bearings, which are the main



Elasticity for Closely Conformal Contact Interface, Fig. 6 Schematic of an internal combustion engine powertrain, which has several conformal contact bearings, such as the main bearings, the connecting rod bearings, the wrist-pin bearings, and the piston, piston ring and cylinder liner interfaces (Modified from Fig. 1 in the paper by Goenka et al. (1992))

Elasticity for Closely Conformal Contact Interface, Table 1 Comparison of different approaches ($\check{N}_p$ (= max($n_{xi} \times n_{zi}$)) is the maximum value of the total number of nodes inside the rectangular-storage blocks shown in Fig. 5, and N_{py} is the total number of nodes in the width direction)

Approach	Total computational time to obtain *I* using 3D FEM	Storage requirement for *I* in 2D mixed EHL	Computational time to calculate *d* for 2D mixed EHL
Conventional summation method	$O(N_p^2)$	$O(N_p^2)$	$O(N_p^2)$
Selective storage (Xiong et al. 2010)	$O(N_p^2)$	$O(N_p \check{N}_p)$	$O(N_p \check{N}_p)$
Selective fine mesh + selective storage (Xiong et al. 2010)	$O(N_p \acute{N}_p)$	$O(N_p \check{N}_p)$	$O(N_p \check{N}_p)$
Selective fine mesh + selective storage + mapping (Xiong et al. 2010)	$O(0.5N_{py} \acute{N}_p)$	$O(N_{py} \check{N}_p)$	$O(N_p \check{N}_p)$

bearing, the connecting-rod bearing, the wrist pin bearing, the piston ring/cylinder liner interface, and the piston skirt/cylinder liner interface, shown in Fig. 6. These bearings allow flexible structure design and complicated motion. The main bearings and the connecting rod bearings are usually combinations of two half bearings that can conveniently facilitate the engine crack shaft assembly. The wrist pin bearings are the links between the small ends of connecting rods and pistons. These engine bearings and interfaces form the links of the

crack-connecting road-piston (slider) system that converts the combustion energy input through the piston reciprocating motion into the engine crack-shaft rotation motion that drives a vehicle. The analyses of the elastic deformation of these components require the use of a method mentioned in this entry. An example can be found in the work by Goenka et al. (1992) and Zhu et al. (1993).

Cross-References

- ► Conformal-Contact Elements and Systems
- ► Elastohydrodynamic Lubrication (EHL)
- ► Hertz Theory: Contact of Cylindrical Surfaces
- ► Hertz Theory: Contact of Spherical Surfaces
- ► Hydrodynamic Journal Bearings
- ► Hydrostatic Journal Bearings

References

M.B. Dobrica, M. Fillon, P. Maspeyrot, Influence of mixed-lubrication and rough elastic-plastic contact on the performance of small fluid film bearings. STLE Tribol. Trans. **51**(6), 699–717 (2008)

P. Goenka, R. Paranjpe, Y-R. Jeng, *Flare: An Integrated Software Package for Friction and Lubrication Analysis of Automotive Engines Part I: Overview and Applications*, SAE International Paper No.920487 (Society of Automotive Engineers, Warrendale, 1992), pp. 47–55

G.R. Higginson, The theoretical effects of elastic deformation of the bearing liner on journal-bearing performance. In *Proceedings of Symposium on Elastohydrodynamic Lubrication*. Instn. Mech. Eng. **180** (Part 3B), 31–38 (1965)

K.L. Johnson, *Contact Mechanics* (Cambridge University Press, Cambridge, 1985)

A.J. Levy, A note on the application of the Flamant solution of classic elasticity to circular domains. ASME J. Appl. Mech. **69**, 856–859 (2002)

K.P. Oh, K.H. Huebner, Solution of the elastohydrodynamic finite journal bearing problem. ASME J. Tribol. **95**(3), 343–351 (1973)

N. Patir, H.S. Cheng, An average flow model for determining effects of three-dimensional roughness on partial hydrodynamic lubrication. ASME J. Lubr. Technol. **100**, 12–17 (1978)

O. Pinkus, *Thermal Aspects of Fluid Film Tribology*, ASME Press, New York (1990)

F.H. Shi, Q. Wang, A mixed-TEHD model for journal-bearing conformal contacts – part I: model formulation and approximation of heat transfer considering asperity contact. ASME J. Tribol. **120**, 198–205 (1998)

S.W. Xiong, C. Lin, Y.S. Wang, W.K. Liu, Q. Wang, An efficient elastic displacement analysis procedure for simulating transient conformal-contact elastohydrodynamic lubrication systems. ASME J. Tribol. **132**(2), 1–9 (2010). Paper No.021502

D. Zhu, H.S. Cheng, T. Arai, K. Hamai, A numerical analysis for piston skirt in mixed lubrication – part II: deformation considerations. ASME J. Tribol. **115**, 115–125 (1993)

Elasticity Theory for Spherical Bearings

Fengcai Wang[1,2], Hanyu Wang[3]
[1]School of Mechanical Engineering, Wuhan University of Science and Technology, Wuhan, Hubei, People's Republic of China
[2]National Research Centre of Bearing Technology (ZWZ), Xi'an Jiaotong University, People's Republic of China
[3]Bradford Grammar School, Bradford, West Yorkshire, UK

Synonyms

Equivalent discrete spherical convolution (EDSC); Equivalent spherical convolution (ESC); Fixed-tracked method (FTM); Smooth spherical inverse filter method (SSIF); Spherical fast fourier transform (SFFT) for elasticity of spherical bearings; Spherical grid data model (SGDM) for elasticity of spherical bearings; Spherical multi-grid technique (SMG)

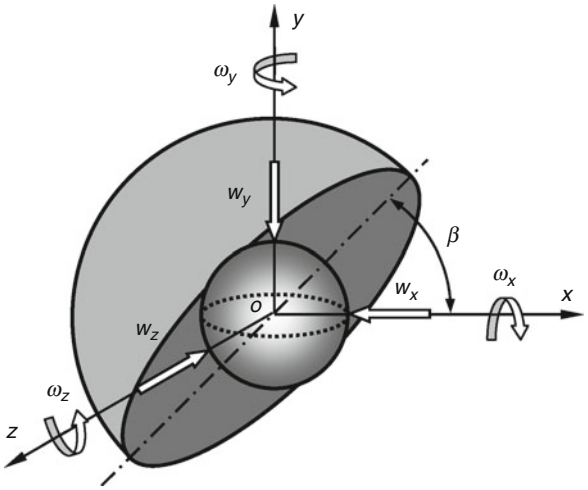
Definition

Elasticity theory for spherical bearings refers to the theoretical method and simulation technique for the elastic deformation of spherical/nonspherical surfaces as well as their application to spherical/aspheric bearings. The theory and method are often required for the study of the tribological phenomena in spherical/aspheric bearings, such as elastohydrodynamic lubrication of hip joints. Although a number of factors that are extensively found in manufacturing, design, and application lead to more complex shapes and features of aspheric bearing geometry such as roughness, non-sphericity, and texture, more advanced elasticity theory for aspheric bearings can still be obtained from the development in the elasticity theory for spherical bearings. In fact, the non-spherical bearing features can be described further on the basis of the spherical grid data model (SGDM), which was developed for the study of tribological problems in spherical bearings. Therefore, the elasticity theory for both spherical and aspheric bearings can be usually placed under the same definition.

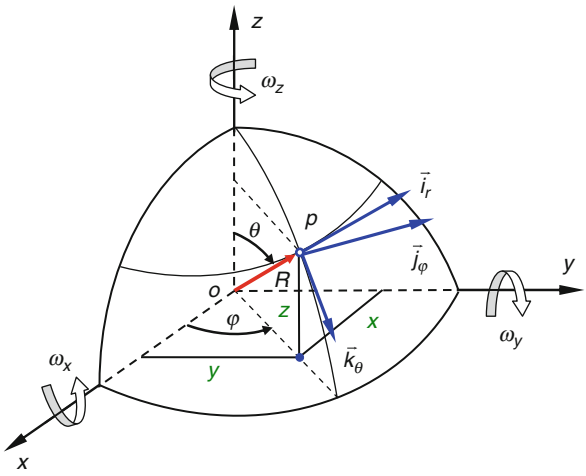
Scientific Fundamentals

Spherical Elastic Problem

The development of elasticity theory for the study of the tribological problems of spherical/aspheric bearing surfaces addressed the more realistic conditions such as bearing



Elasticity Theory for Spherical Bearings, Fig. 1 Schematic diagram of spherical bearing under the three-dimensional loading and motion condition



Elasticity Theory for Spherical Bearings, Fig. 2 Suitable spherical coordinate system with the *z*-polar axis for the study of spherical elasticity theory and spherical bearings

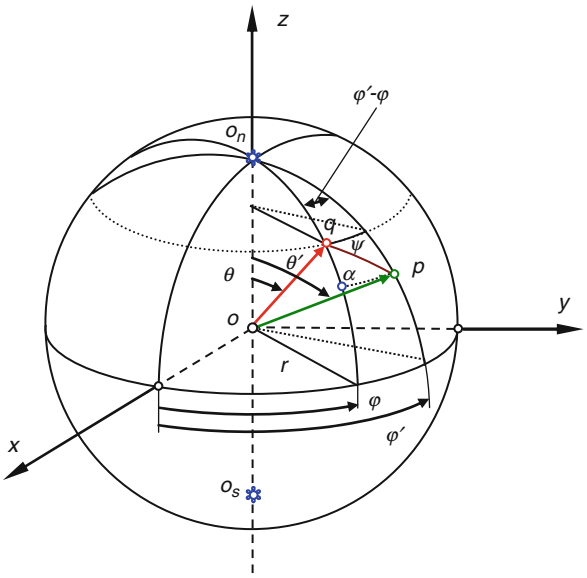
geometry, dynamic loading and transient motion, material property, and combinations thereof, as well as lubricant property, and so on. Therefore, the development of such elasticity theory and methods needs to satisfy the requirements of resolving tribological problems under realistic operating conditions (Wang et al. 2008a, b, 2009a, b, 2010; Wang 2010, 2011).

A spherical bearing is a typical conformal contact problem that can usually work under the three-dimensional dynamic loading and time-dependent motion conditions (Kim et al. 2006). It can be configured using appropriate spherical coordinates such as the *z*-polar system, as shown in Figs. 1 and 2. The spherical cup bearing/bushing may be positioned with an inclination β (Wang et al. 2005a, 2008a). It should be noted that the bearing surface might be perfectly spherical or non-spherical, due to the factors introduced in manufacturing, design, and application. Therefore, the description of both the global and local non-sphericity of bearing geometry in motion with the appropriate mesh-grid has to be considered in the development of the spherical elasticity theory in order to study the tribological mechanism of spherical/aspheric bearing (Wang et al. 2009a, b).

An accurate and efficient evaluation of the elastic deformation is an important aspect in the development of spherical bearing modeling, for example an elastohydrodynamic lubrication analysis of natural/artificial hip joints needs to consider the elastic deformation of the spherical bearing surfaces in the calculation of the fluid film thickness (Wang et al. 2005a, b, 2008a). The finite element method can be employed for the evaluation of elastic deformation of spherical bearing surfaces and has good availability for the various forms of bearings with complex structures such as complex underlying support of the spherical bearing in hip joint replacements. However, the main limitation for its application to the evaluation of elastic deformation is a relatively low density of meshed elements on the bearing surfaces for the corresponding tribological problem. On the other hand, the modeling of conformal spherical contact problems has a strong dependence on the density of mesh grids, for example, the elastic deformation in a full numerical analysis of elastohydrodynamic lubrication of spherical bearing in hip joint implants under the corresponding operational conditions such as viscosity 0.002 Pas needs the mesh density on the order of 256×256 grid points. Therefore, it is necessary to use an effective and efficient numerical method for the deformation calculation, particularly when it is intended to be applied to real tribological problems (Wang et al. 2005a, 2006, 2009a, b).

Deformation on Spherical Surface

The elastic deformation on spherical bearing surface/body can be expressed by the double integration equation of pressure acting on the spherical surface and the



Elasticity Theory for Spherical Bearings, Fig. 3 Spherical distance on a spherical surface in a spherical coordinate system

displacement response function of the spherical bearing as follows:

$$\delta(\psi) = \iint_{\Omega} p(\psi_p)\, S'(\psi - \psi_p) d\Omega \tag{1}$$

where the deformation $\delta(\psi)$ of the spherical surface is described as a function of the spherical distance ψ. In this equation, the variable $p(\psi)$ is the pressure distributions acting on the spherical surface, such as the fluid film pressure or the contact pressure of spherical bearing. The integration kernel $S'(\psi)$ is the displacement response function that refers to the deformation at the point $q(\varphi, \theta)$ on the spherical surface when a unit constant pressure is applied at the point $p(\varphi_\zeta, \theta_\eta)$ with the spherical distance ψ between the two points, as shown in Fig. 3. In fact, although the deformation $\delta(\psi)$ is the function of difference of the spherical distance $\Delta\psi = \psi - \psi_p$, the integration calculation of an elastic problem can not completely be described as a real discrete mathematical convolution in spherical coordinates due to its relation to the latitude coordinate, which is further discussed in detail in (Wang et al. 2004, 2009a, d, 2010; Wang 2010, 2011).

However, the spherical distance can be evaluated as a function of the spherical coordinates (φ, θ) through the triangular function as follows:

$$\sin^2\left(\frac{\psi}{2}\right) = \sin^2\left(\frac{\theta - \theta_\eta}{2}\right) + \sin^2\left(\frac{\varphi - \varphi_\zeta}{2}\right) \sin\theta \sin\theta_\eta \tag{2}$$

Thus, the deformation at a point on the spherical surface can be expressed by

$$\delta(\varphi,\theta) = \iint_{\Omega} p(\varphi_\zeta,\theta_\eta)\, S'(\varphi - \varphi_\zeta, \theta - \theta_\eta, \theta) R^2 \sin\theta_\eta d\varphi_\zeta d\theta_\eta \tag{3}$$

In fact, the deformation calculation of (3) is also not a convolution integration of the pressure and the displacement response function since the deformation expression is not only a function of $\Delta\varphi = \varphi - \varphi_\zeta$ and $\Delta\theta = \theta - \varphi_\eta$ but also a function of the latitude coordinate at a point on the spherical surface θ, that is to say $\delta = \delta(\Delta\varphi, \Delta\theta, \theta)$ (Wang et al. 2009a). Thus, the deformation evaluation of spherical surface under a given pressure distribution can be carried out only by the direct integration rather than by the fast convolution technique (Felix Quinonez et al. 2008, Wang et al. 2008a, b).

Equivalent Spherical Convolution

Mathematically, the convolution operation can be run by a fast convolution algorithm such as the fast Fourier transform technique in the frequent domain. To obtain such convolution integration to evaluate the elastic deformation, a further mathematical transformation of (2) needs to be considered as follows:

$$\sin^2\left(\frac{\psi}{2}\right) = \sin^2\left(\frac{\theta - \theta_\eta}{2}\right) + \sin^2\left(\frac{\varphi - \varphi_\zeta}{2}\right) \left[\sin^2\theta_m - \sin^2\left(\frac{\theta - \theta_\eta}{2}\right)\right] \tag{4}$$

where an average latitude concept for achieving the equivalent convolution integration is defined as $\theta_m = (\varphi + \theta)/2$. Thus, the elastic deformation of a spherical surface can be written as equivalent spherical convolution integration if the average latitude is used as follows:

$$\delta(\varphi,\theta) = \iint_{\Omega} p(\varphi_\zeta,\theta_\eta)\, S'(\varphi - \varphi_\zeta, \theta - \theta_\eta; \theta_m) r^2 \sin\theta_\eta d\varphi_\zeta d\theta_\eta \tag{5a}$$

or

$$\delta(\varphi,\theta) = p(\varphi,\theta) \otimes S'(\varphi,\theta) \tag{5b}$$

where the symbol $\otimes$ represents the convolution calculation. Therefore, the elastic deformation calculation can be carried out by the convolution of the pressure distribution and the corresponding displacement function under the given average latitude of the spherical surface. This integration process for the deformation of a spherical surface/body is called the equivalent spherical convolution model (ESC). Furthermore, the equivalent spherical convolution can be numerically evaluated as follows:

$$\delta_{ij}(\varphi_i, \theta_j) = \sum_{k=1}^{m} \sum_{l=1}^{n} [p_{kl}(\varphi_k, \theta_l) \sin(\theta_l)] S_{kl}^{ij}(\varphi_i - \varphi_k, \theta_j - \theta_l; \theta_m) \tag{6a}$$

or

$$[\delta]_{m\times n} = [p]_{m\times n} \otimes [S]_{m\times n} \tag{6b}$$

where the variables $[\delta]$, $[P]$, and $[S']$ are the overall deformation on spherical surface, the pressure distribution, and the displacement coefficient matrix, respectively. The general methodology for the elastic deformation evaluation in (6a) is the so-called equivalent discrete spherical convolution (EDSC). The EDSC model can be performed by employing a fast numerical method such as the fast Fourier transform technique on a spherical surface in a spherical coordinate system (Wang et al. 2008a, b, 2009a, b). The EDSC model plays a significant role in the deformation evaluation of spherical surfaces and in the corresponding tribological mechanism of spherical/aspheric bearings. The underlying theory of this method forms the basis of efficient computational models for the study of spherical bearings and the biotribology of hip joints (Wang et al. 2009a; Wang 2010, 2011).

Elastic Displacement Response

The elastic displacement response plays an important role in the elastic deformation evaluation of a spherical surface as well as in further contact mechanics calculations of the corresponding conformal spherical contact problem. Therefore, it is necessary to extract the elastic displacement response of the spherical surface by using the finite element method before the EDSC model is employed to evaluate the elastic deformation. The one-dimensional displacement response along either the longitude line a–c or the latitude line a–b through the center of the spherical surface under the action of a unit constant pressure on the element face of the spherical surface can be firstly obtained by the finite element method, as shown in Fig. 4a, b. The deformation response as a function of the spherical distance ψ is subsequently obtained by a curve-fit using a least-squares algorithm as follows (Wang et al. 2004, 2009a):

$$S(\psi) = \sum_{i=1}^{\alpha} \gamma_i s^{\beta - i} \tag{7}$$

where the base function s is expressed as a function of the spherical distance ψ on a unit sphere, $s = \sin(\psi/2)$, while α and β are the length number of the series and the power of the base function, respectively. The coefficient γ can be determined by the least-square fit. Once the displacement response function is obtained, it can be used for the elastic deformation evaluation by the direct numerical integration in (1).

Elastic Displacement Coefficient Matrix

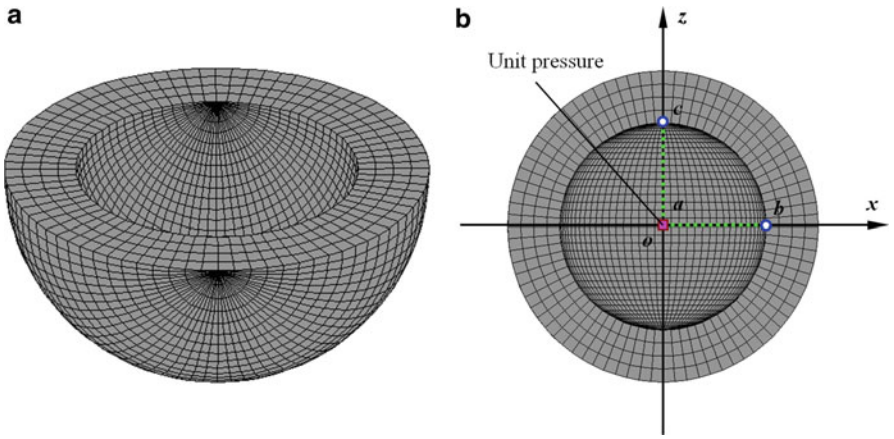
The fast evaluation of the elastic deformation of the spherical surface is usually required for the simulation of the corresponding tribological modeling of spherical bearing. After the displacement response function is obtained by the curve-fit, the displacement response function $S(\psi)$ can be extended by the EDSC model onto the whole solution domain of the spherical surface. Thus, the two-dimensional displacement coefficient matrix on the spherical surface can be expressed as follows:

$$S = [S_{ij}]_{m\times n} \tag{8}$$

The displacement coefficient matrix can then be applied for evaluation of the elastic deformation on a spherical surface and so in further simulations of spherical conformal contact problems, for example, the fast elastic deformation evaluation of spherical surfaces is usually required for the tribological simulations of elastohydrodynamic lubrication and asperity contact mechanics as well as the effect of the non-sphericity of spherical bearings on lubrication (Wang et al. 2005a, b, c, 2006, 2007, 2008a, b, 2009a, b).

Spherical Fast Fourier Transform

An efficient numerical technique for the elastic deformation evaluation is significant to carry out the tribological simulation of spherical bearings with a conformal contact feature because such simulation often needs to be run with a high density of mesh grid points. Since the spherical elastic deformation is mathematically expressed as the equivalent spherical convolution (ESC) or the equivalent discrete spherical convolution (EDSC), it can be efficiently run by the fast convolution technique. Therefore,



Elasticity Theory for Spherical Bearings, Fig. 4 Schematic diagram for extraction of the elastic displacement response of spherical surface: (**a**) 3D finite element model and (**b**) displacement response under an applied unit pressure at the center of spherical surface

this facilitates the application of a fast spherical Fourier transform (SFFT) technique to calculate the deformation of spherical surfaces in the frequency domain by the finite difference method in spherical coordinates as follows:

$$\left[\tilde{\delta}(f_\varphi, f_\theta)\right] = \left[\tilde{p}(f_\varphi, f_\theta)\sin(\theta_m)\right] \bullet \left[\tilde{S}(f_\varphi, f_\theta)\right] \quad (9)$$

where the matrices $[\tilde{\delta}]$, $[\tilde{p}]$, and $[\tilde{S}]$ are the discrete Fourier transform of the elastic deformation $[\delta]$, the fluid film pressure or contact pressure distribution $[p]$, and the displacement coefficient matrix $[S]$, respectively, and θ_m is specified as the mean latitude of solution domain in spherical coordinates. The zero-padding method for elimination of periodic errors caused by the circular convolution property needs to be used when the SFFT technique is applied to the convolution evaluation of the spherical deformation (Wang et al. 2009a). In addition, the deformation evaluation of spherical surfaces can also be performed either by the combination of the EDSC model and the spherical multi-grid technique (SMG) or by a combination of the EDSC model and the SFFT method and the SMG technique (Gao et al. 2009, Wang et al. 2009a, b; Wang 2010, 2011).

Elastic Problem of Non-spherical Surface

As described previously, the non-spherical geometry of real bearings can be caused by a number of factors that extensively exist in manufacturing, in the specific design and in the application. The elastic problem of non-spherical surfaces with macro/micro geometry features relative to the nominal spherical surface can still be evaluated by the EDSC model and the SFFT technique as described above. For such deformation evaluation, the one-dimensional elastic displacement response $S(\psi)$ and the displacement coefficient matrix $[S]_{m\times n}$ can be obtained accordingly by using at first an equivalent/nominal spherical surface as the basic sphere. However, the fixed-tracked method (FTM) may be required for the description of the non-spherical geometry of spherical/aspheric bearing in motion when the time-dependent elastic deformation is considered in the tribological simulation such as the transient elastohydrodynamic lubrication of spherical/aspheric bearing during the operating condition (Wang et al. 2008a, b). As a whole, the elasticity theory for a spherical surface/body can be employed not only for the spherical bearing modeling but also for the aspheric bearing modeling, which can be developed from the former spherical grid data model (SGDM) (Wang et al. 2009a, b).

Both the dynamic and the steady-state loads may lead to the deformation of the bearing surfaces and to further geometrical change of the bearing surface in motion. Therefore, when the deformation $\delta = \delta(\varphi, \varphi)$is considered, the geometry of a spherical/aspheric bearing under either steady-state or transient motion conditions can be expressed by the radial coordinates in a spherical coordinate system as follows:

$$r_i = r_{i0}(\varphi, \theta) + \Delta r_i(\varphi, \theta) + \Theta_i(\varphi, \theta) + \delta_i(\varphi, \theta) \quad (10)$$

where the subscript $i = c$ or b represents either the spherical cup bearing/bushing or the spherical head bearing/spherical bearing surface, the deviations $\Delta r_i(\varphi, \theta)$ of the non-spherical bearing surface relative to a nominal spherical surface, such as ellipsoidal bearing surface, and the item $\Theta_i(\varphi, \theta)$ is the micro-geometrical deviation of the real bearing surface from the nominal

bearing surface, such as roughness and machining marks, respectively (Wang et al. 2009a, b, c, d; Wang 2010, 2011).

Key Applications

Tribology of Spherical Bearings/Joints and Biotribology of Hip joints

The elasticity theory and calculation method of spherical surface/body deformation, including the EDSC model and the SFFT technique as well as other methods such as the FTM method, play a significant role in the development of tribology of spherical/aspheric bearings and biotribology of hip joints as well as dynamics of flexible multi-body system with spherical joints. The general methodology can provide an effective and efficient method for the elastic deformation evaluation of spherical surface and for further study of the tribological mechanism of spherical/aspheric bearings under realistic lubricant properties and operating conditions such as three-dimensional, dynamic loading and motion (Wang et al. 2005a, b, 2008a, b, 2009a, b). The application of the tribological modeling of spherical bearing and biotribology of hip joints can be illustrated by the following examples.

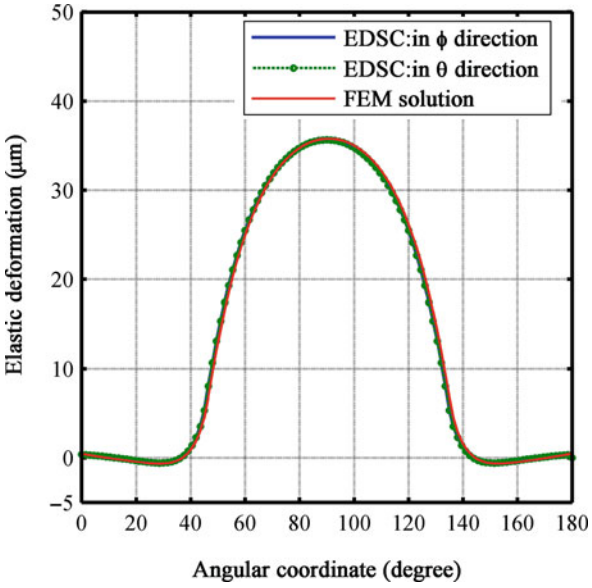
Deformation of Spherical Surface

The fast and accurate evaluation of elastic deformation is one of most important issues in the theoretical simulation of tribological behavior of the conformal spherical/aspheric contact problems in engineering and bioengineering as well as in flexible multi-body systems, for instance the elastic deformation calculation is required to simulate the elastohydrodynamic lubrication of hip joint replacements (Wang et al. 2005a, b, 2008a, b). As an example, the pressure distribution applied on the bearing surface of an acetabular cup in a polyethylene-on-metal hip joint replacement is given as follows:

$$p[\psi(\varphi,\theta)] = p_0 \cdot \left[1-\left(\frac{\psi(\varphi,\lambda)}{\psi(\varphi_0,\lambda_0)}\right)^2\right]^{1/2} \qquad (11)$$

where p_0 is the maximum pressure located at point (φ_0,θ_0) and is typically in the order of 10 MPa. The pressure p is distributed within an angle of $\pi/2$ rad around the polyethylene cup, whose material properties are usually an elastic modulus $E = 1{,}000$ MPa and Poisson ratio $\nu = 0.4$. The spherical distance ψ between two points of (φ,θ) and (φ_0,θ_0) on the spherical cup surface can be calculated as follows (Wang et al. 2009c):

$$\cos(\psi) = \cos\theta\cos\theta_0 + \sin\theta\sin\theta_0\cos(\varphi_0-\varphi) \qquad (12)$$



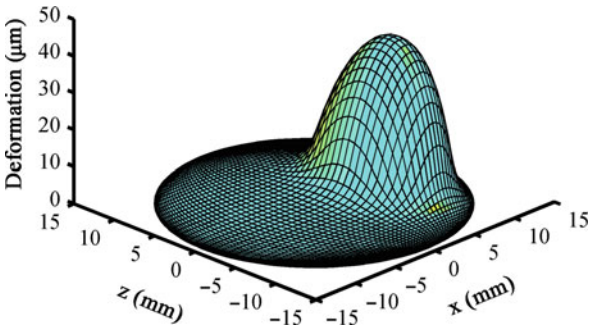
Elasticity Theory for Spherical Bearings, Fig. 5 Comparison of the elastic deformation of a spherical cup bearing in metal-on-polyethylene hip joint replacements predicted by the EDSC model and the SFFT technique and by the finite element method

A mesh density of 128×128 grid points for the elastic deformation evaluation is used to divide the bearing surface in the longitude and latitude directions, respectively. When the maximum pressure point (φ_0,θ_0) is located at the point $(0,0)$ on the spherical cup bearing, there is a good agreement in the elastic deformation predicted by the EDSC model with SFTT technique compared with the finite element method along the longitude and the latitude directions as shown in Fig. 5 (Wang et al. 2004, 2005a, b). When the maximum pressure point (φ_0,θ_0) is moved to the point $\left(\frac{\pi}{4},0\right)$ on the edge of the cup's bearing surface, the elastic deformation distribution on the spherical cup surface is evaluated by the EDSC model and the SFFT technique, as shown in Fig. 6.

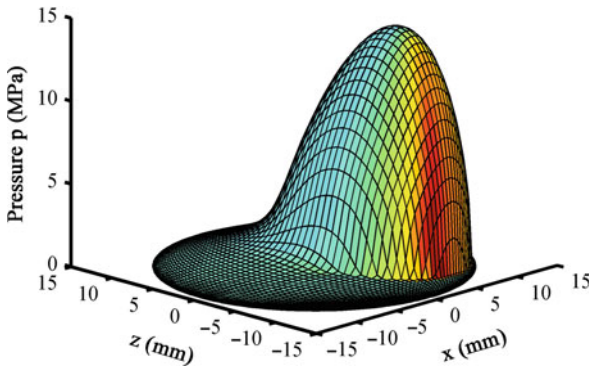
Contact Mechanics

Evaluation of contact pressures is often seen as an inverse problem of the deformation calculation. The contact problem of spherical bearing surfaces can be described by the relationship of the gap $g(\varphi,\theta)$ and the corresponding contact pressure $p_c(\varphi,\theta)$ in the contact area Ω_c as follows:

$$\begin{cases} g(\varphi,\theta) = 0 & (\varphi,\theta) \in \Omega_c \\ p_c(\varphi,\theta) > 0 & (\varphi,\theta) \in \Omega_c \\ g(\varphi,\theta) > 0 & (\varphi,\theta) \notin \Omega_c \\ p_c(\varphi,\theta) = 0 & (\varphi,\theta) \notin \Omega_c \end{cases} \qquad (13)$$



Elasticity Theory for Spherical Bearings, Fig. 6 The elastic deformation distribution on a spherical cup bearing in metal-on-polyethylene hip joint replacements evaluated by the EDSC model and the SFFT technique



Elasticity Theory for Spherical Bearings, Fig. 7 The predicted pressure distributions in the elastohydrodynamic lubrication of MOP hip joint replacements

where the gap $g(\varphi,\theta)$ between the two bearing surfaces consists of the original gap $g_0(\varphi,\theta)$ and the deformation $\delta(\varphi,\theta)$ of the two bearing surfaces due to the contact pressure as follows:

$$g(\varphi,\theta) = g_0(\varphi,\theta) + \delta(\varphi,\theta) \tag{14}$$

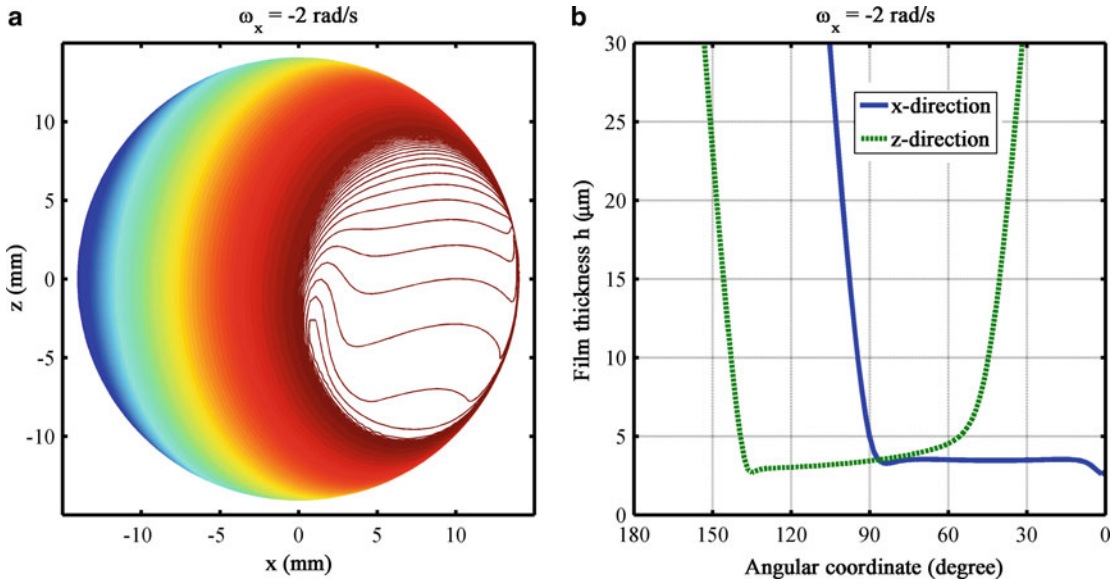
The elasticity theory and calculation method for a spherical surface can be used to obtain the deformation required for the solution of contact mechanics problems of spherical bearings. Since the EDSC model for the elastic deformation of bearing surfaces using a high-density mesh grid can be efficiently run by the SFFT technique, both the macro contact mechanics and micro asperity contact mechanics of spherical/aspheric bearing can be resolved by the general methodology. For example, the contact mechanics of worn bearings in hip joint replacements is evaluated by the smooth spherical inverse filter method (SSIF) based on the equivalent discrete spherical convolution theory. The details of the SSIF method can be found, for example, in the references (Wang et al. 2005c, 2006, 2008b). It should be pointed out that such a method of contact mechanics is able to be further applied to computational wear modeling of spherical bearing or hip joints (Wang et al. 2009c).

Lubrication

The development of an elasticity theory and calculation method for the deformation of spherical surfaces also plays an important role in lubrication modeling and the corresponding tribological behavior prediction of a spherical/aspheric bearing system. For example, the general methodology of the EDSC model and the SFFT technique presented in this section as well as the FTM method have so far become the most significant theory and method for the development of computational biotribology of natural and artificial hip joints, such as the steady and transient elastohydrodynamic lubrication, mixed/boundary lubrication and friction prediction, non-sphericity of bearing geometry and lubrication, wear and lubrication, and so on (Wang et al. 2006, 2008b, 2009d; Wang 2010, 2011). The elastohydrodynamic lubrication simulation of metal-on-polyethylene (MOP) hip joint replacements under the steady-state load and angular velocity condition of $w_y = 2{,}500$ N and $\omega_x = -2$ rad/s shown in Figs. 1 and 2 is taken as an example to describe an application of the general theory and method in lubrication problems. The lubricant film thickness, h, consists of the unreformed gap for the conformal spherical bearing surface couple of hip join implant and the corresponding elastic deformation, δ, due to the hydrodynamic pressure as follows (Wang et al. 2008a, b, 2009a, b):

$$h(\varphi,\theta) = C\left[1 - \varepsilon_x \sin\theta\cos\varphi - \varepsilon_y \sin\theta\sin\varphi - \varepsilon_z\cos\theta\right] + \delta(\varphi,\theta) \tag{15}$$

where the center position of the femoral head is described by the eccentricities, $(\varepsilon_x, \varepsilon_y, \varepsilon_z)$, with reference to the center of the acetabular cup bearing. The material modulus and Poisson ratio of the cup and the head are given as E_1 = 1,000 MPa, v_1 = 0.4, E_2 = 210 GPa, and v_2 = 0.3, respectively. The radii of the cup and the head are 14.1 mm and 14.0 mm, and thus the radial bearing clearance is C = 100 μm. The cup is physiologically positioned at an inclination angle of 45° in an appropriate spherical coordinate system with the z-polar, as shown in Figs. 1 and 2. The lubricant property is chosen to have a typically low viscosity of 0.002 Pas. The elastic deformation of both the cup and the head bearing surfaces



Elasticity Theory for Spherical Bearings, Fig. 8 The predicted film thickness contour (**a**) and the film thickness profiles (**b**) in the elastohydrodynamic lubrication of MOP hip joint replacements

is evaluated by the EDSC model and the SFFT technique through the in-house software developed for computational biotribology of hip joints. A sufficiently high mesh density of 256 × 256 grid points is required for numerically resolving the elastohydrodynamic lubrication problem during the iterative convergence process (Wang et al. 2005a, b, 2006, 2008a, b, 2009a, b). The fluid film pressure distribution is predicted by the general method, as shown in Fig. 7, while the fluid film thickness contour and the film thickness profiles are also given as shown in Fig. 8a, b, respectively.

It should be pointed that the general elasticity theory presented here can be further applied to those of lubrication problems having a non-sphericity in the bearing geometry such as the real spherical bearings and aspheric bearings. For this purpose, the fixed-tracked method (FTM) will be required for the geometry description of the bearing in motion during the operating conditions. The details of the general methodology applied for such lubrication solution have been presented elsewhere (Wang et al. 2009a, b). Furthermore, the general elasticity theory and the corresponding fast numerical technique can be introduced to the friction prediction of spherical/aspheric bearing, possibly working in the mixed/boundary lubrication regime, where both the spherical elastic deformation and the contact mechanics will be addressed during the simulation (Wang et al. 2005c, 2006, 2008b, 2009c, 2010; Wang 2010, 2011).

For specific details of the general elasticity theory and calculation method used for the study of both the tribological mechanism of spherical/aspheric bearing and the computational biotribology of hip joints the reader can refer to the following entries ► Geometry of Spherical/Aspheric Bearings, ► Lubrication Theory for Spherical Bearings, ► Friction Prediction for Spherical Bearings, ► Wear Modeling of Spherical Bearings.

Cross-References

► Contact Mechanics for Spherical/Aspheric Bearing
► Elasticity Theory for Spherical Bearings
► Friction Prediction for Spherical Bearings
► Geometry of Spherical/Aspheric Bearings
► Lubrication Theory for Spherical Bearings
► Wear Modeling of Spherical Bearings

References

A. Felix Quinonez et al., A steady-state elastohydrodynamic lubrication model aimed at natural hip joints with physiological loading and anatomical position. Proc. IMechE. Part J: J. Eng. Tribol. **222**, 503–512 (2008)

L.M. Gao et al., Effect of 3D physiological loading and motion on elastohydrodynamic lubrication of metal-on-metal total hip replacements. Med. Eng. Phys. **31**(6), 720–729 (2009)

B.C. Kim et al., Development of composite spherical bearing. Compos. Struct. **75**, 231–240 (2006)

F.C. Wang et al., Prediction of elastic deformation of acetabular cup and femoral head for lubrication analysis of artificial hip joints. Proc. IMechE. Part J: J. Eng. Tribol. **218**, 201–208 (2004)
F.C. Wang et al., Elastohydrodynamic lubrication modelling of artificial hip joints under steady-state conditions. ASME J. Tribol. **127**(10), 729–739 (2005a)
F.C. Wang et al., Elastohydrodynamic lubrication modelling of spherical metal-on-metal artificial hip joints, in *ASME Proc. WTC2005, World Tribology Congress III*, Washington, WTC2005-63556, pp. 489–490, 2005b
F.C. Wang et al., An integrated experimental and theoretical contact mechanics study of UHMWPE hip implants tested in a hip simulator, in *ASME Proc. World Tribology Congress III*, Washington, pp. 311–312, 2005c
F.C. Wang et al., Lubrication modelling of artificial hip joints, in *IUTAM Symposium on Elastohydrodynamic and Micro-Elastohydrodynamic*, ed. by R.W. Snidle, H.P. Evans. Solid Mechanics and its Applications, vol. 134 (Springer, Dordrecht, 2006), pp. 385–396
F.C. Wang et al., Effect of non-spherical bearing geometry on transient elastohydrodynamic lubrication in metal-on-metal hip joint implants. Proc. IMechE. Part J: J. Eng. Tribol. **221**, 379–389 (2007)
F.C. Wang et al., Transient elastohydrodynamic lubrication of hip joint implants. ASME J. Tribol. **130**(1), p011007 (2008a)
F.C. Wang et al., Lubrication and friction prediction in metal-on-metal hip joint implants. IOP Phys. Med. Biol. **53**, 1277–1293 (2008b)
F.C. Wang et al., Non-sphericity of bearing geometry and lubrication in hip joint implants. ASME J. Tribol. **131**(3), p031201 (2009a)
F.C. Wang et al., Non-spherical bearing geometry and elastohydrodynamic lubrication of hip joint replacements under transient walking conditions, in *World Tribology Congress IV*, Kyoto, 2009b
F.C. Wang et al., Dynamic contact mechanics and wear modelling of hip joint replacements with hard-on-hard material combination under three-dimensional loading and transient motion, in *World Tribology Congress IV*, Kyoto, 2009c
F.C. Wang et al., Non-sphericity of bearing geometry and lubrication in hip joint implants. ASME J. Tribol. **131**(3), p031201 (2009d)
F.C. Wang et al., Tribological modelling of spherical bearing with complex spherical-base geometry and motion, in *Tribology and Design*, Algarve, 2010
F.C. Wang, Dynamic Contact Behaviour and Evolution of Bearing Interface with Spherical-Base Geometry, National Natural Science Foundation of China (NSFC) Report No.10972165, pp. 1–30 (2010)
F.C. Wang, Dynamic Contact Mechanism and Failure of Rolling Bearing, National Key Basic Research Program of China (973 Program) Report No. 2011CB706601 (2011)

Elastic–Plastic Contact

► Contact Elasto-Plasticity

Elastohydrodynamic Lubrication

► Elastohydrodynamic Lubrication (EHL)

Elastohydrodynamic Lubrication (EHL)

Dong Zhu
State Key Laboratory of Mechanical Transmission, Chongqing University, Chongqing, People's Republic of China

Synonyms

EHL; Elastohydrodynamic lubrication; Elastohydrodynamics; Lubrication

Definition

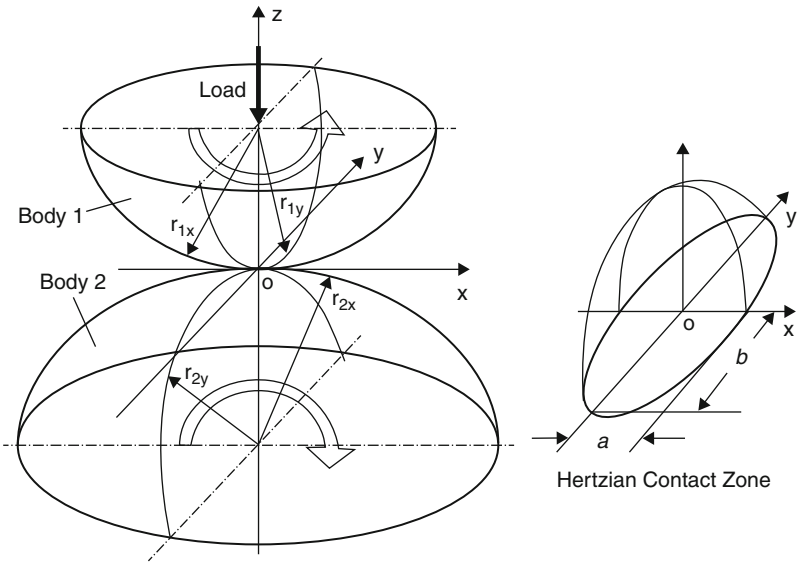
Elastohydrodynamic lubrication (EHL) is a mode of fluid-film lubrication in which hydrodynamic action is significantly enhanced by surface elastic deformation and lubricant viscosity increases due to high pressure.

Scientific Fundamentals

Background and General Description

Power and motion are transmitted through surface contacts in mechanical components. Contact can be generally categorized as conformal contact and non-conformal contact. In conformal contact, the curvatures of two surfaces match each other so that the area of surface interaction is large, typically comparable to dimensions of the machine elements. Conformal contact components include journal and thrust bearings, piston skirt/ring/cylinder liner systems, and many types of joints and seals. Non-conformal contact, on the other hand, is formed by two surfaces whose curvatures do not match. As a result, the contact area is usually small in both principal directions (called point contact), or at least in one direction (line contact), and a localized high pressure concentration may exist at the contact. Non-conformal contact can be found in various gears, rolling element bearings, cam/follower systems, ball screws, vane pumps, metal rolling tools, traction drives and continuously variable transmissions.

Elastohydrodynamic lubrication (EHL) is a relatively new area in the development of lubrication theory and practice. In 1886, O. Reynolds published his milestone theoretical lubrication analysis based on B. Tower's journal bearing experiment. The Reynolds equation has been the foundation of hydrodynamic lubrication theory since then. It has been proven to be satisfactory for lubrication performance prediction of many conformal contact machine elements, in which the hydrodynamic pressure is low, typically on the order of 1∼10 MPa or less. Based on this success, attempts were made to extend the application



Elastohydrodynamic Lubrication (EHL), Fig. 1 An elliptical (point) contact

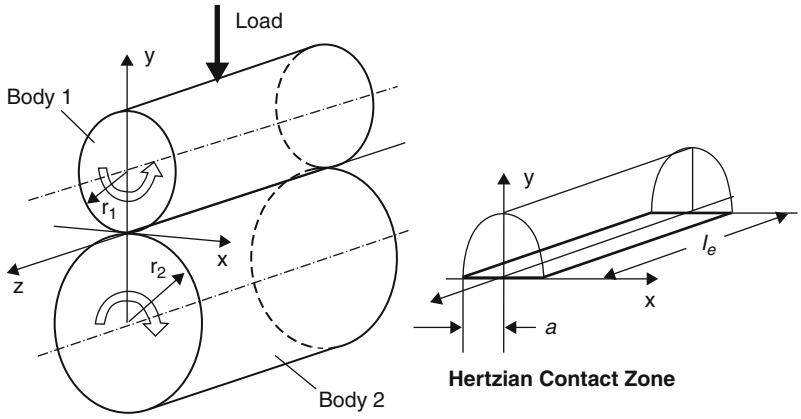
of hydrodynamic lubrication theory to non-conformal contact components (see Martin 1916, and others). It was found, however, that lubricant film thickness in non-conformal contacts predicted by the hydrodynamic lubrication theory is often far below that observed in engineering reality. Since the localized high pressure in non-conformal contacts, often on the order of 0.1∼1 GPa, may cause considerable surface deformation and lubricant viscosity increase, which may play a significant role in lubrication formation, the classic hydrodynamic lubrication theory needs to be modified for those highly stressed non-conformal contact components.

Early efforts to include effects of surface elastic deformation and lubricant pressure-viscosity relationship started in 1930s, but a significant breakthrough was not seen until the middle of the last century. Analyses satisfactorily revealing basic EHL mechanisms by considering hydrodynamics, surface deformation, and pressure-viscosity relationship were successfully made through either simplified solutions (Grubin 1949, and others) or full numerical solutions (Petrusevich 1951; Dowson and Higginson 1959; and others). Some early studies yielded curve-fitting formulae for EHL film thickness prediction, including those by Dowson and Higginson (1961–1965) for line contact problems and Hamrock and Dowson (1976a, b) for point contacts. Those formulae have been widely used in engineering practice. Parallel to the analytical studies, experimental investigations also showed fruitful results, including the EHL film thickness measurements with capacitance technique (Crook 1961–1963; Dyson et al. 1966), optical interferometry (Gohar and Cameron 1963), and other techniques. Good agreement has been observed between the analyses and the testing data. A solid foundation for EHL was laid by these early studies.

EHL has been one of the fastest developing fields of tribology research since the 1960s and 1970s, and today it is still an active and challenging research area that attracts much attention. As computer and information technologies have advanced in the last two decades, EHL solution methods and experimental techniques have significantly improved. The focus of research has shifted to thin-film EHL and mixed EHL, with real engineering roughness from early studies based primarily on a smooth surface assumption. In addition, EHL research is now more directly associated with the study of lubrication breakdown and surface failure as well as surface design optimization.

Point and Line Contacts

In general, a non-conformal contact is formed between two curved surfaces, whose principal radii of curvature can be expressed as r_{1x} and r_{1y} for Body 1 and r_{2x} and r_{2y} for Body 2, as shown in Fig. 1. When a load is applied to the contact, a contact zone is formed due to elastic deformation, and its size and resulting contact pressure can be estimated according to the well-known Hertzian theory. The most generic case is called "elliptical contact," as the contact zone appears to be an ellipse, as illustrated in Fig. 1. If the two semi-axes of the contact ellipse are



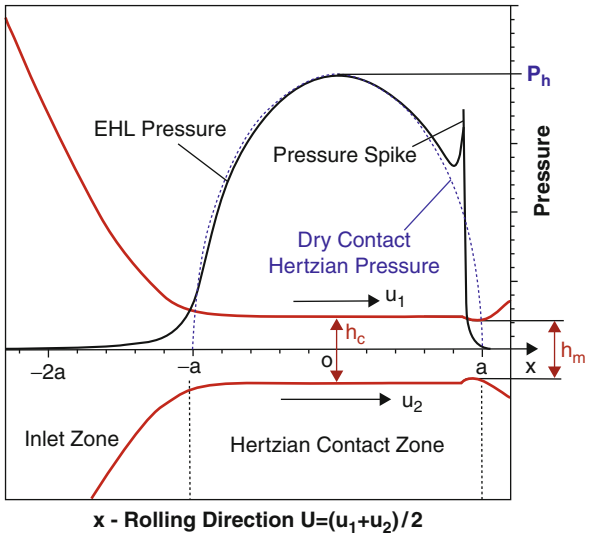
Elastohydrodynamic Lubrication (EHL), Fig. 2 A line contact

equal, $\boldsymbol{a} = \boldsymbol{b}$, it is called "circular contact." If the semi-axis of contact zone in one direction (say $\boldsymbol{b}$ in y-direction) is infinitely long or much greater than that in the other direction, the contact geometry can be simplified as two-dimensional and the Hertzian zone becomes a narrow band area. It is called "line contact," which is often caused by two cylindrical surfaces against each other (see Fig. 2).

When lubricant is provided and the surfaces are in motion, a lubricant film, entirely or partially separating the surfaces, may be formed due to hydrodynamic action, which could be largely enhanced by the viscosity increase and surface deformation due to high pressure. An EHL equation system, consisting of the Reynolds equation for hydrodynamics, an elasticity equation for surface deformation, a geometric equation for contact geometry and film thickness, and lubricant rheology equations for viscosity and density, can be solved through either a simplified approach or a full numerical solution method to predict the EHL characteristics. Please refer to the following entries for details: "► EHL Governing Equations," "► Simplified EHL Solution Methods," and "► EHL, Full Numerical Solution Methods."

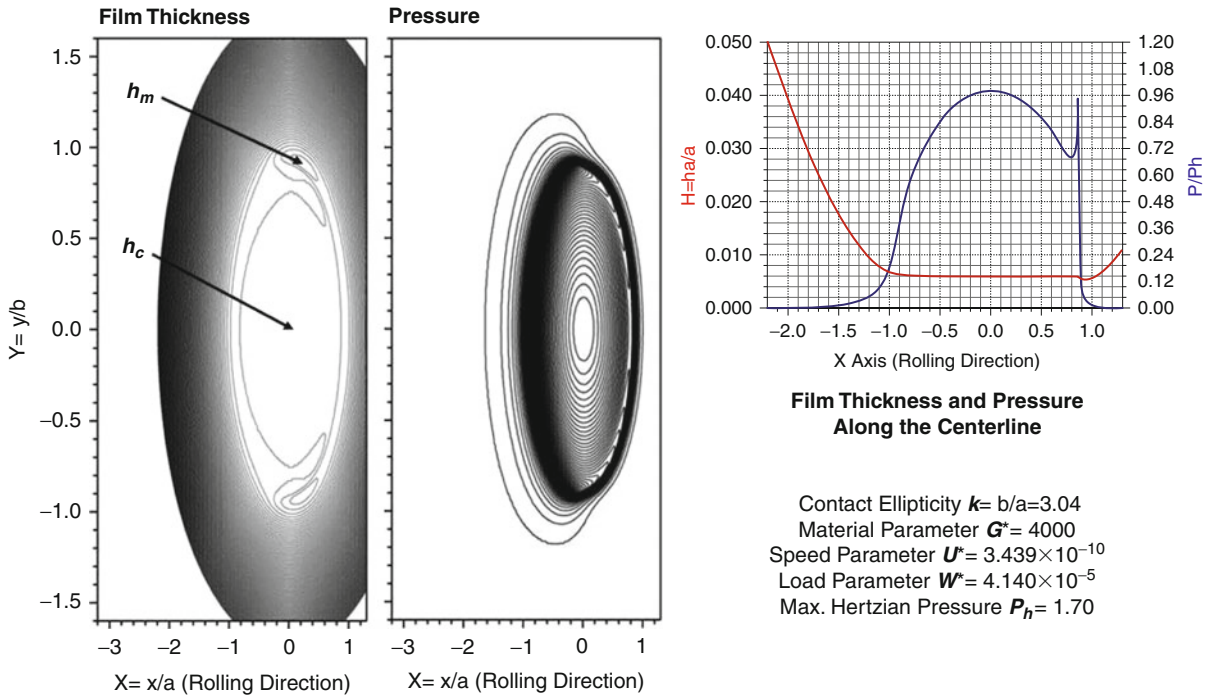
Line Contact EHL Characteristics

Line contact EHL problems can be simplified as two-dimensional because parameter variations along the direction of infinite length of the contact zone are none or negligible. A typical EHL line contact is illustrated in Fig. 3, showing basic EHL characteristics. It can be seen that when one or both surfaces are moving along the x-direction, a nearly constant gap is formed and lubricant filled between the two surfaces. However, close to the outlet the gap narrows to form a constriction (on the right-hand side of the figure). Two characteristic



Elastohydrodynamic Lubrication (EHL), Fig. 3 Line contact EHL

parameters are commonly used to describe the lubrication effectiveness: central film thickness $\boldsymbol{h_c}$ and minimum film thickness $\boldsymbol{h_m}$, as illustrated in the figure. Correspondingly, the EHL pressure distribution is quite close to that of dry contact from the Hertzian theory, except that hydrodynamic pressure exists upstream outside the Hertzian contact zone, and there may be a high pressure spike downstream immediately before the film constriction. The central film thickness and minimum film thickness can be estimated by the empirical formulae mentioned above. Please refer to "► Film Thickness Formulas: Line Contacts" in the present encyclopedia for more details.



Elastohydrodynamic Lubrication (EHL), Fig. 4 EHL solution for an elliptical contact

Point Contact EHL Characteristics

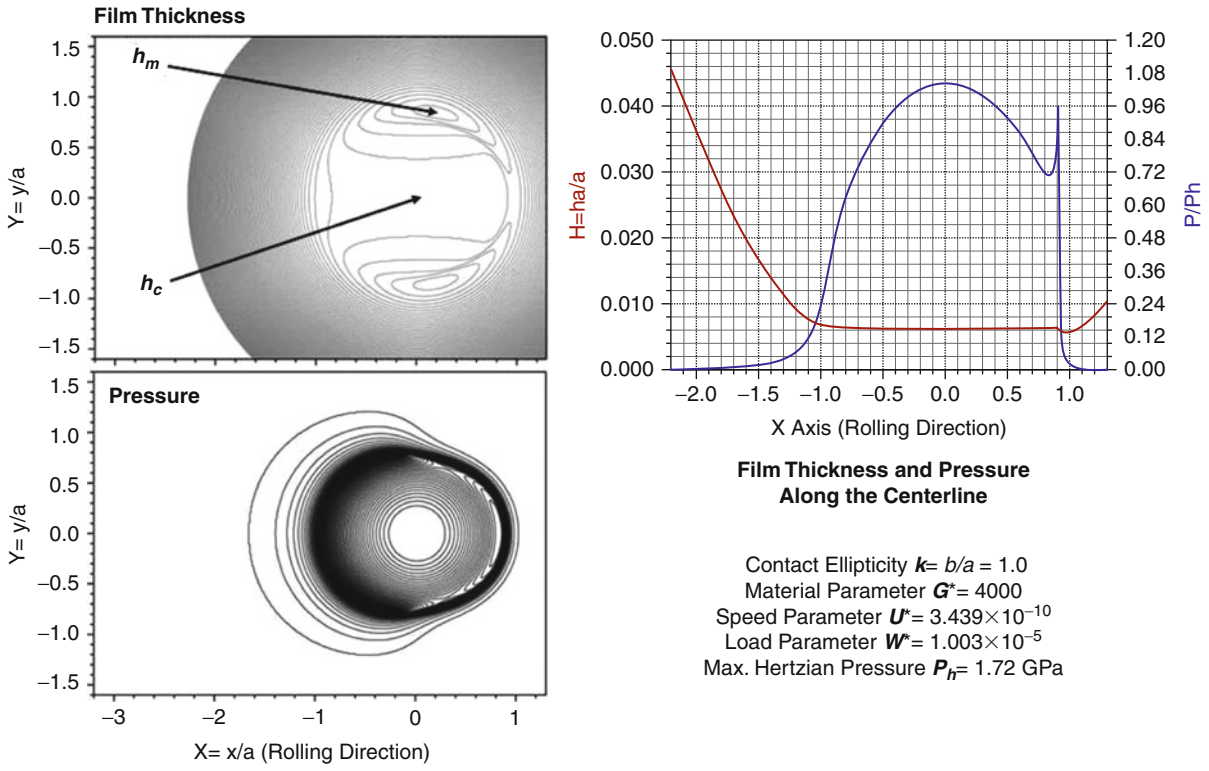
The point contact problem is three-dimensional as film thickness and pressure vary in both principal (***x***- and ***y***-) directions. A typical elliptical contact EHL case is given in Fig. 4, and a circular contact case in Fig. 5, which show basic point contact EHL characteristics.

It is obvious that along the centerline in the direction of motion, the EHL film thickness and pressure variations demonstrate the same characteristics as those in line contacts, e.g., the nearly constant film thickness in most of the contact zone, the film constriction on the outlet side, and the high pressure spike immediately before the constriction. For point contact problems, however, the contact region is finite in the y-direction, and significant lateral flows may have a negative influence on lubricant film formation, especially at locations close to the lateral borders of the contact ellipse. As a result, the film constriction takes a horseshoe shape and the minimum film thickness is found to be on both sides at "earlobes" away from the centerline, as shown in the film thickness contour plots in Figs. 4 and 5. Correspondingly, the pressure spike also takes a horseshoe shape slightly upstream from the film constriction. This typical phenomenon also has been observed from experimental results. Figure 6 shows two sample film thickness contour plots obtained from optical interferometry, in comparison with those from numerical solutions. Good agreement has been found (see Hartl et al. 2005).

The central and minimum film thicknesses can be estimated by using curve-fitting formulae developed by Hamrock and Dowson (1976a, b) and others. Please refer to "► Film Thickness Formulas: Point Contacts" for details.

Speed and Load Effects

It has been found that EHL film thickness is significantly affected by rolling speed (or entraining speed), $U = (u_1 + u_2)/2$, but is relatively insensitive to load. Figure 7 presents a set of EHL solutions demonstrating the effect of speed on the EHL, while Fig. 8 shows that of load. As the speed (and/or viscosity) increases, the lubricant film thickness goes up significantly. On the other hand, if the speed/viscosity gradually decreases towards zero, the film thickness approaches zero, and the film shape and EHL pressure distribution approach those of the Hertzian dry contact due to vanishing hydrodynamic action, as shown in Fig. 7. Note that, as the zero film thickness cannot be illustrated on a log scale, a cut-off of 1.15 nm is applied to the solution of the lowest speed, $U^* = 3.439 \times 10^{-14}$ (1.15 nm is already smaller than the possible lubricant



Elastohydrodynamic Lubrication (EHL), Fig. 5 EHL solution for a circular contact

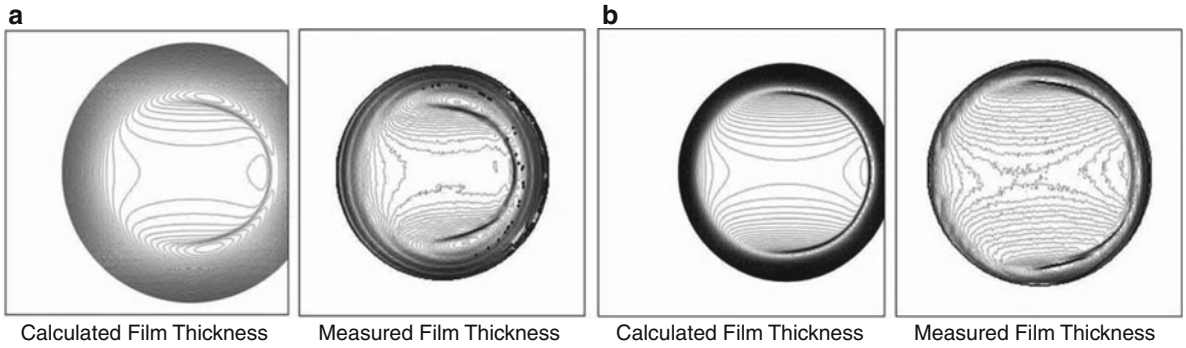
molecular size of most lubricants; it can be considered practically as zero in engineering reality).

When load varies, however, the situation may be more complicated. Figure 8 shows that as the load increases, the central film thickness may first go up and then go down, but, indeed, film thickness variation due to the change of load is not as significant. It is said that the stiffness of EHL film appears to be greater than that of the solid body materials (usually steel). In other words, the increase of load may cause much more elastic deformation of the bodies than the EHL film reduction. The EHL film is often difficult to break down solely due to load increase, unless in some uncommon cases the load becomes extremely heavy, e.g., the case of $W^* = 6.419 \times 10^{-4}$ presented in Fig. 8.

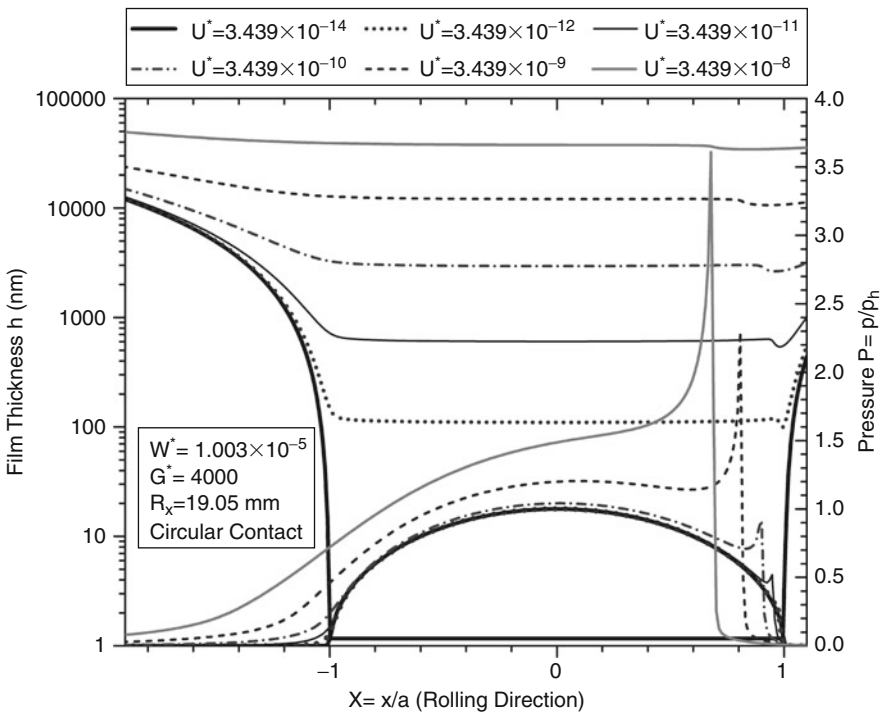
It is also observed that at light loads the surface deformation may be insignificant and the pressure distribution appears to be close to that from the hydrodynamic lubrication theory (no film constriction, no secondary high pressure spike, as in the case of $W^* = 6.269 \times 10^{-9}$ given in Fig. 8). When the load is continuously increased, the solution shows more and more EHL characteristics. If the load is further increased, the solution will gradually approach that of Hertzian dry contact and the film thickness approaches zero eventually, as shown in Fig. 8. The EHL solution for the extreme case of $W^* = 6.419 \times 10^{-4}$ demonstrates a zero central film thickness as well as characteristics nearly the same as those of dry contact.

Effect of Contact Ellipticity

Contact ellipticity, defined as $k = b/a$, may considerably affect EHL performance, especially at high speeds and/or light loads that yield strong hydrodynamic action. Generally, a small contact ellipticity may cause significant lateral flows that have a negative influence on the EHL film formation, while a large ellipticity close to line contact may yield a considerably thicker EHL film. Figure 9 presents a set of EHL solutions with varying contact ellipticity but a constant max. Hertzian pressure, P_h. It can be seen that large ellipticities greater than 4∼6 yield nearly the same film thickness as that of line contact, so these cases can be approximated by line contact in practice. Note that in Fig. 9, dimensionless load parameter W^* has to be greatly changed in order to keep the Hertzian pressure constant.



Elastohydrodynamic Lubrication (EHL), Fig. 6 Comparison of film thickness contours between numerical solutions and experimental results from optical interferometry. 1″ steel ball against sapphire disk, max. Hertzian pressure 1.517 GPa Paraffinic base oil SR600 at 40°C, rolling speed: (**a**) 0.372 m/s; (**b**) 0.133 m/s



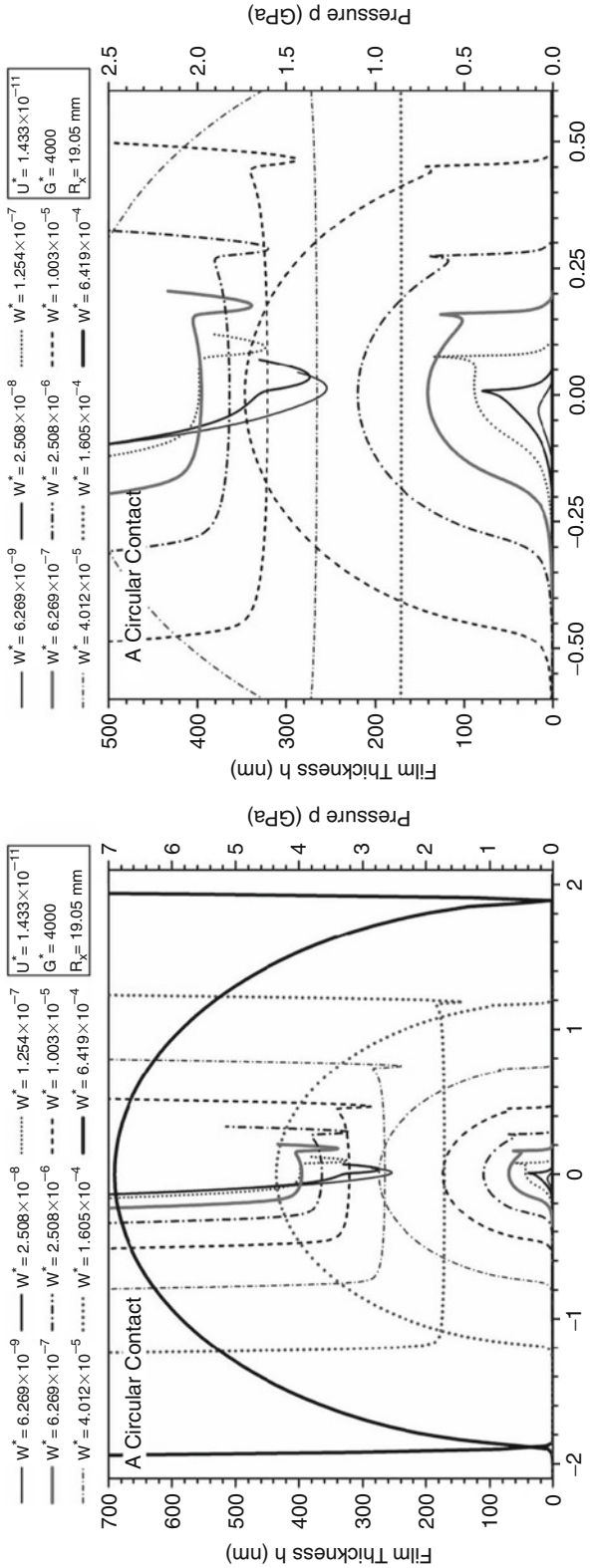
Elastohydrodynamic Lubrication (EHL), Fig. 7 Speed effect on EHL

Effect of Materials Properties

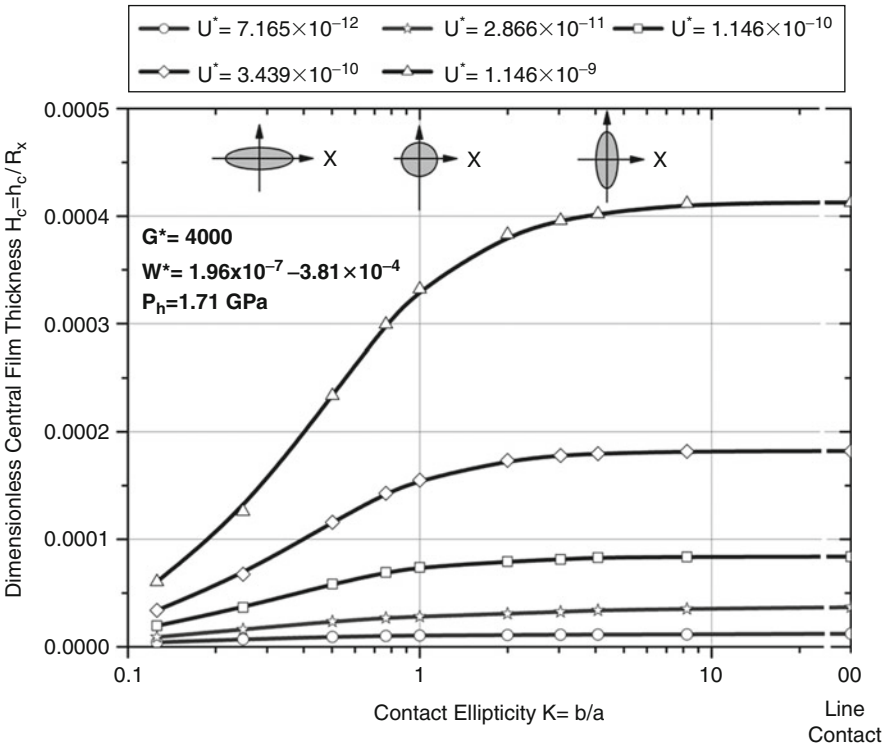
Figure 10 gives EHL solutions demonstrating the effects of material properties on EHL film thickness. The left graph is for the effect of pressure-viscosity coefficient of lubricant, $\boldsymbol{\alpha}$, and the right one for that of effective elastic modulus of the solid bodies, $\boldsymbol{E'}$. It is obvious that the film thickness is quite sensitive to the piezo-viscous behavior of lubricant but insensitive to the elastic properties of solid bodies within the shown parameter ranges commonly found in engineering applications.

Surface Roughness Effect

Preliminary studies were based on the smooth surface assumption, showing basic EHL characteristics as described above. In engineering practice, however, no surface is ideally smooth and roughness is usually of the



Elastohydrodynamic Lubrication (EHL), Fig. 8 Load effect on EHL. The central part of plot on the *left* is enlarged and shown on the *right*

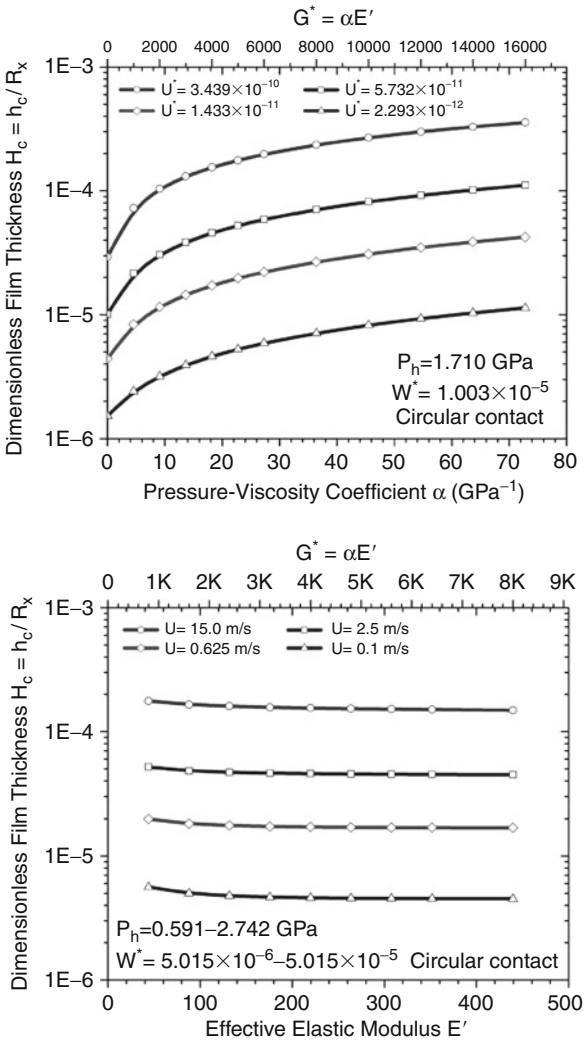


Elastohydrodynamic Lubrication (EHL), Fig. 9 Contact ellipticity effect

same order of magnitude as, or greater than, the film thickness estimated by the smooth surface EHL theory. Effects of surface roughness and topography, therefore, ought to be taken into account in most engineering applications, except in some cases with super-finished surfaces whose roughness is much smaller than the EHL film thickness. Early studies on roughness effect employed mainly stochastic models, e.g., those by Patir and Cheng (1978) for line contacts and Zhu and Cheng (1988) for point contacts. Recently, more attention has been paid to deterministic analyses using digitized real engineering surfaces as input data (e.g., Xu and Sadeghi 1996; Zhu and Ai 1997; and others). The roughness effect is a complicated topic requiring in-depth investigation and lengthy discussions. The present article only gives a snapshot. More details can be found in "► Stochastic Models for Rough Surface EHL," "► Deterministic Models of Rough Surface EHL," and "► Mixed EHL" in the present encyclopedia.

A set of EHL numerical solutions is given in Fig. 11, in which all the cases are under the same operating conditions, but surface roughness heights are adjusted proportionally to obtain different levels of RMS roughness, while keeping the surface topography consistent, in order to study the roughness effect. It is obvious that the roughness can remarkably affect lubrication characteristics as well as subsurface von Mises stress field, which can be closely correlated with surface failures. It is important to note that, for rough surface EHL, the central and minimum film thicknesses defined previously no longer seem to be appropriate, as they are largely dependent upon local surface irregularities, therefore, they may not be representative for overall lubrication effectiveness. The minimum film thickness, for example, is always zero whenever/wherever there is a local asperity contact, and the central film thickness may fluctuate greatly from one moment to another due to moving roughness. New parameters are introduced as follows in order to describe rough surface EHL behavior:

Average Film Thickness $\boldsymbol{h_a}$: For point contacts, $\boldsymbol{h_a}$ is calculated within a certain area around the center of contact, typically within ½ Hertzian radius of the normalized contact zone, so that a sufficiently large area can be included for calculating the average value, but possible edge effect can be avoided. For line contacts, the average value is calculated within ½ Hertzian contact widths on each side along the centerline. When the surfaces are smooth, $\boldsymbol{h_a}$ is almost the same as central film thickness $\boldsymbol{h_c}$ defined previously.



Elastohydrodynamic Lubrication (EHL), Fig. 10 Effects of materials properties

Film Thickness (λ) Ratio: In the last 30+ years the term of λ ratio has been widely used, but defined in different ways, so it should be clarified. In the present article λ ratio is defined as the ratio of average film thickness (defined above) to the composite RMS roughness, $\boldsymbol{\lambda = h_a/\sigma}$. This definition well describes the global lubrication effectiveness in rough surface contacts.

Contact Load Ratio $\boldsymbol{W_c}$ and Contact Area Ratio $\boldsymbol{A_c}$: If the average EHL film thickness is not much greater than the composite RMS roughness, say the λ ratio is smaller than 1.25∼1.50, surface asperity contacts will usually take place, and a certain portion of load is supported by the asperity contacts. The contact area ratio, $\boldsymbol{A_c}$, is defined as the ratio of total asperity contact area to the area of Hertzian contact zone. The contact load ratio, $\boldsymbol{W_c}$, is the load supported by asperity contacts divided by the total load. Note that $\boldsymbol{A_c}$ and $\boldsymbol{W_c}$ are correlated but can be different in value. For full-film lubrication with no significant asperity contact, both $\boldsymbol{A_c}$ and $\boldsymbol{W_c}$ are zero or nearly zero. As the film thickness decreases and the asperity contacts become more and more significant, $\boldsymbol{A_c}$ and $\boldsymbol{W_c}$ increase. For dry contact $\boldsymbol{W_c}$ will reach its maximum value of 1.0, but $\boldsymbol{A_c}$ may be less than 1.0, unless all the asperities are completely flattened.

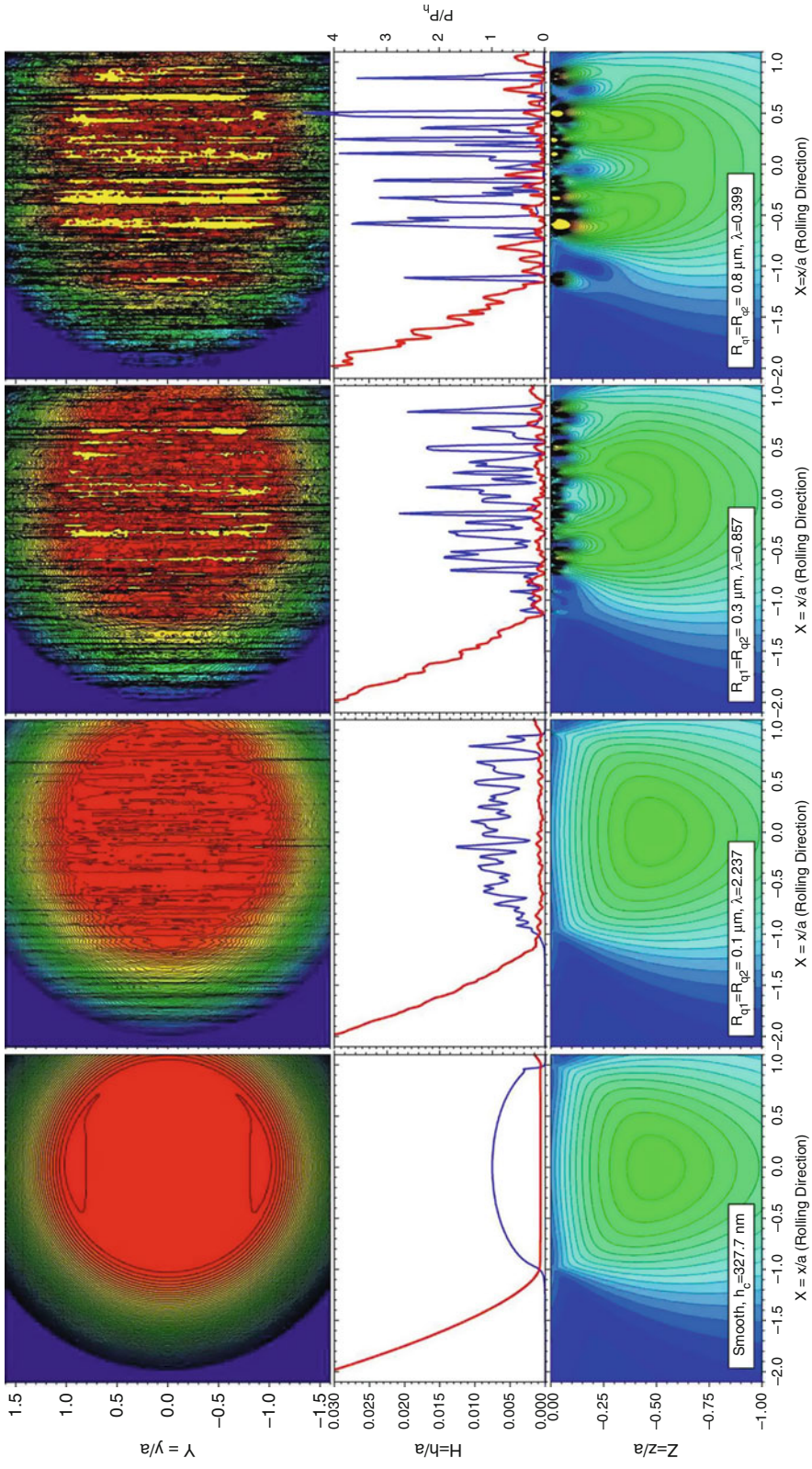
Some typical results showing the roughness effect based on the cases, a few of which are already given in Fig. 11, are summarized in Fig. 12. It can be seen clearly that increasing roughness usually leads to a considerable reduction of λ ratio and significant increases in contact load ratio, friction, and subsurface stresses, which can be directly related to contact severity and surface failures. In general, therefore, better finished surfaces are usually preferred in engineering practice. However, reduction of roughness is often associated with production cost increase. Also, for specific products, surface finishing processes usually need to meet specific requirements, and it is unlikely that there is a universal solution for a large variety of applications. Once again, the effects of surface roughness and topography are complicated topics and are still under research.

Mixed EHL

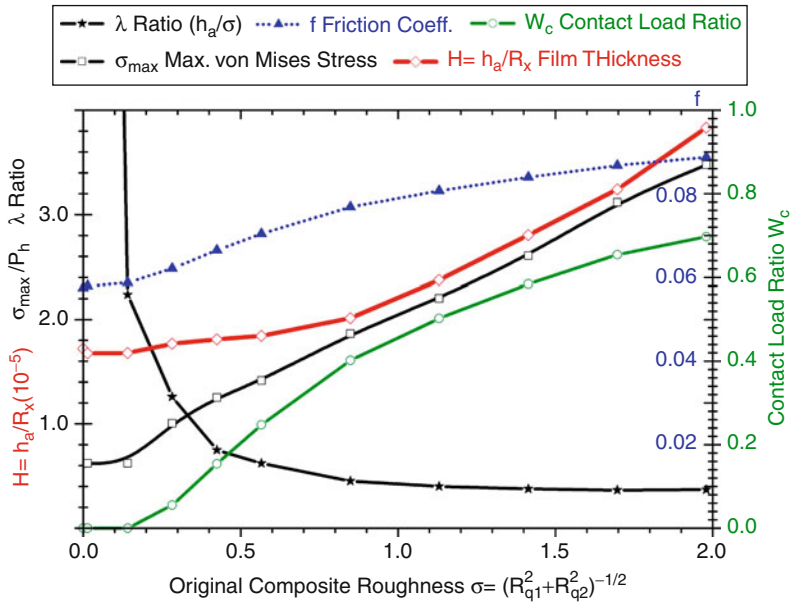
As described above, surface roughness is often of the same order of magnitude as, or greater than, the lubricant film thickness, so that a complete separation between the two rough surfaces is seldom seen in engineering practice. Mixed EHL (also called partial EHL) is the mode in which both EHL films and surface asperity contacts coexist and neither can be ignored. As a consequence, the load is shared between them. Actually, most functional components operate in mixed lubrication regime. Mixed lubrication is of great importance not only because of its wide existence but also because it is a critical transition towards lubrication breakdown and surface failures.

Investigations on mixed EHL have been challenging, mainly for the following reasons:

1. Modeling mixed EHL needs to handle hydrodynamic lubrication and asperity contacts simultaneously. The Reynolds equation can be used for lubrication and a separate dry contact model is often employed for asperity contacts. The boundary conditions in between might not be easy to handle because the borders may generally be irregular in shape and unstable due to moving roughness.



Elastohydrodynamic Lubrication (EHL), Fig. 11 A set of solutions showing roughness effect (circular contact, $U^* = 1.433 \times 10^{-11}$, $G^* = 4{,}000$, $W^* = 1.003 \times 10^{-5}$, $P_h = 1.71$ GPa, two ground surfaces)

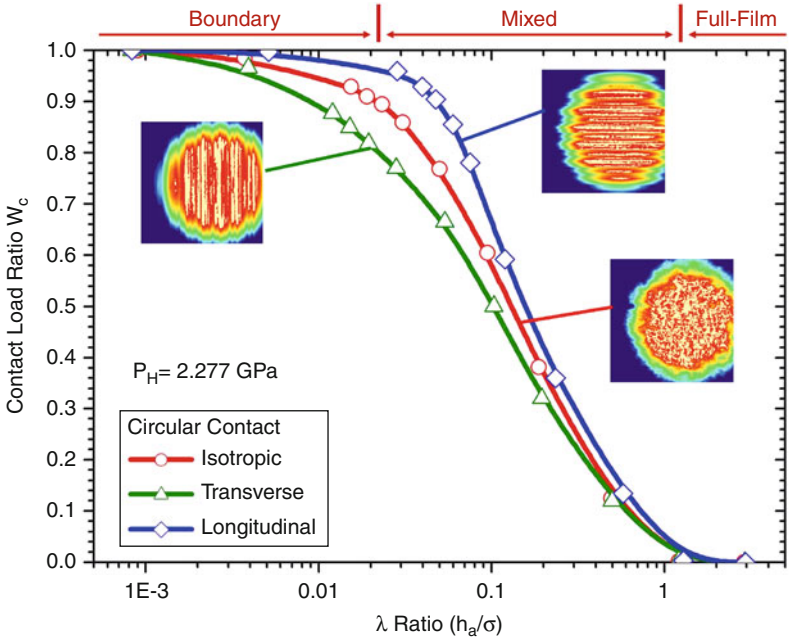


Elastohydrodynamic Lubrication (EHL), Fig. 12 Effect of composite roughness

2. In reality, mixed EHL is often associated with high-frequency machined roughness. Early studies employed stochastic models, which appear to have limitations in handling low λ ratio cases and presenting detailed parameter variations and localized peak values (that could be critical for failure analyses). Deterministic models, however, require advanced computers and improved numerical algorithms when dealing with engineering roughness.
3. Extremely thin lubricant films and rough asperity contacts also pose great challenges to experimental studies. Some techniques previously used, such as those based on measuring capacitance, electric resistance, or x-rays, may no longer be appropriate. Optical interferometry might be the only feasible choice so far, but much improved resolution is required.

Advancements in computer and information technologies have fueled significant breakthroughs in thin-film and mixed EHL research and development since the 1990s. First, optical interferometry has been improved so that measurement of thin films on a nanometer scale with a patterned/textured rough surface is now possible. The contributors include those at Imperial College, London (Johnston et al. 1991; Guantang et al. 2000, and others), Tsinghua University, China (Luo et al. 1996), and Brno University of Technology, Czech Republic (Hartl et al. 1999). Please refer to "► Lubricant Film Thickness Measurement: Colorimetric Interferometry" and "► Lubricant Film Thickness Measurement: Relative Optical Interference Intensity (ROII) Method" for details. Concurrently, deterministic solutions for mixed EHL with real engineering rough surfaces using a unified numerical approach have been obtained by Hu and Zhu (2000) and Holmes et al. (2005). Considering that dry contact is nothing but a special case of lubricated contact under extreme conditions (such as ultra-low speed, ultra-low viscosity, and high pressure concentrated in tiny asperity contact areas), theoretically one should be able to use a unified lubrication equation system to simulate both EHL films and asperity contact simultaneously (see Zhu 2007). Newly developed numerical approaches appear to be capable of simulating the entire transition from full-film and mixed EHL down to dry contact under severe operating conditions using digitized machined rough surfaces. Please refer to "► Mixed EHL" for details.

It is important to note that mixed lubrication is still an active research area and many questions remain unanswered. For example, how to simulate phenomena in ultra-thin films of only several nanometers or less, for which hydrodynamic lubrication theory based on continuum mechanics may become invalid, because the film thickness is of the same order of magnitude as that of molecular sizes of common lubricants. Also, the surfaces may often be covered by boundary films whose formation is dependent largely upon interfacial physics/chemistry,

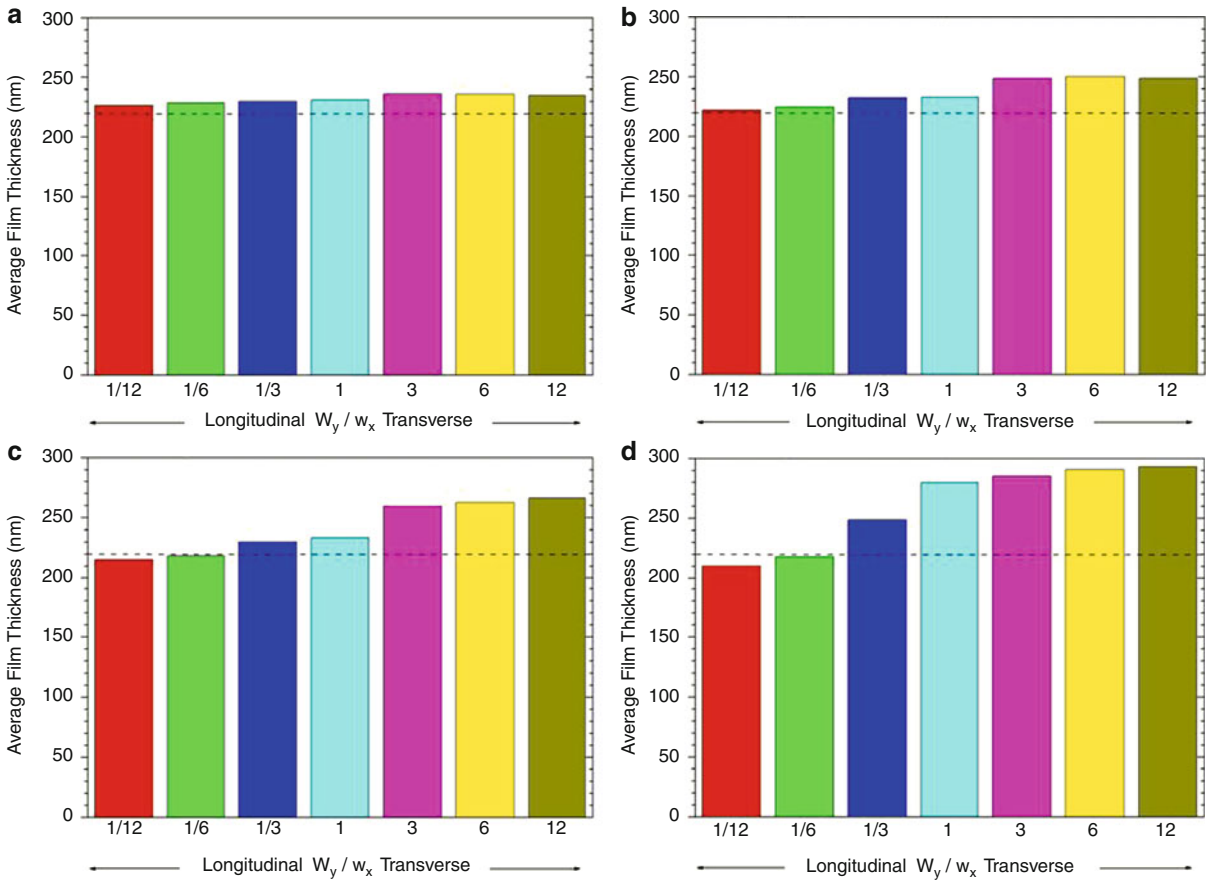


Elastohydrodynamic Lubrication (EHL), Fig. 13 Load sharing as a function of film thickness (λ) ratio

for which modeling is not well developed yet. In addition, experimental techniques need to be further improved when dealing with thin films and rough surface interactions on a nanometer scale. In-depth discussions are beyond the scope of the present article. The following is a snapshot of some fundamentals to date.

Figure 13 illustrates some typical results based on a set of deterministic solutions of rough surface mixed EHL. It can be seen that when the λ ratio is greater than 1.25∼1.50, the contact load becomes zero or negligible, indicating a full-film EHL condition. If the λ ratio is reduced to less than 0.05∼0.1 or so, more than 80% of load is supported by asperity contacts so the situation can be considered as boundary lubrication without significant hydrodynamics. Mixed lubrication is generally considered in a range of λ ratio in between, in which most functional components operate. It can also be observed that the surface topography and orientation have a considerable influence on the lubrication behavior and load sharing in mixed EHL regime. This influence becomes negligible at small λ ratios due to vanishing hydrodynamics in boundary lubrication, or at very large λ ratios when the lubricant film thickness is much greater than the roughness.

A common question is how the roughness and its orientation affect the EHL film thickness if other conditions remain the same. This is of practical importance in engineering design and manufacturing. Unfortunately, it is dependent upon several factors and there does not seem to be a simple answer. For example, in line contacts, roughness transverse to rolling direction may yield a thicker EHL film than longitudinal, but in circular contacts, it might sometimes be opposite due to significant lateral flows that may be negative to lubrication formation but could be enhanced by transverse roughness. In general, however, the roughness and orientation effects on the EHL film thickness, predicted by recent deterministic models, do not seem to be as great as those predicted by early stochastic models (such as those by Patir and Cheng 1978; Zhu and Cheng 1988). Figure 14 presents some results of line contact mixed EHL, obtained from using sinusoidal roughness in order to avoid undesirable influences by randomness of machined surfaces (see Ren et al. 2009, for details). Note that in the figure W_x and W_y are sinusoidal wavelengths in x- and y-directions, respectively. A ratio of W_y/W_x greater than one represents transversely oriented sinusoidal roughness, while that smaller than one means longitudinal. It is observed that for line contacts, transverse roughness may yield thicker EHL film (or greater average gap) than longitudinal under the same operating conditions. Also, in mixed lubrication at small h_{cs}/σ ratios this effect appears to be more significant. This is qualitatively in good agreement with the trend predicted by Patir and Cheng's stochastic model (1978). However, quantitatively the roughness and orientation



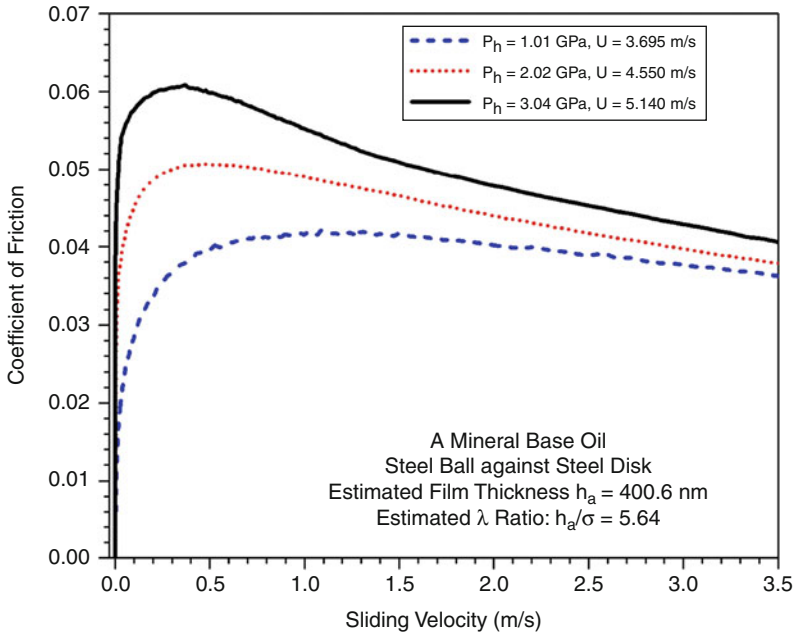
Elastohydrodynamic Lubrication (EHL), Fig. 14 Effects of roughness and orientation on line contact EHL. *Dashed lines* show central film thickness from corresponding smooth EHL solution, h_{cs} (**a**) σ=70 nm, h_{cs}/σ=3.143 (**b**) σ=150 nm, h_{cs}/σ=1.467 (**c**) σ=300 nm, h_{cs}/σ=0.733 (**d**) σ=450 nm, h_{cs}/σ=0.489

effects predicted by the deterministic approach do not seem to be as great in the same range of h_{cs}/σ ratio.

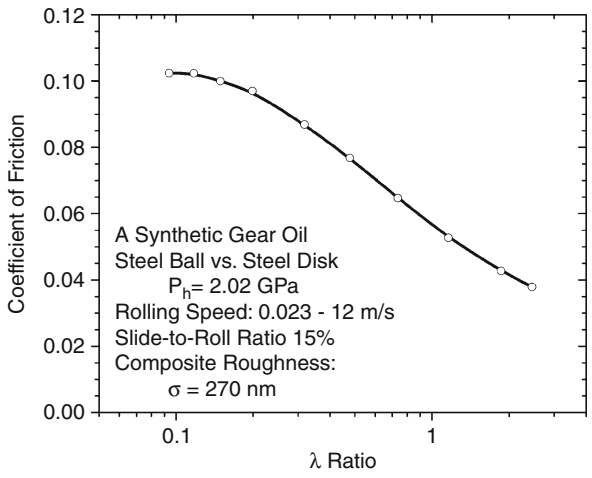
Once again, the rough surface mixed EHL is a complicated subject still under research. All the results presented in the present article show basic characteristics only, which may not be sufficient for every specific case in engineering applications.

Friction in Mixed EHL Contacts

EHL friction, sometimes called traction, is an important subject as it is directly associated with machine efficiency and energy consumption. High friction may also cause excessive heat generation and operating temperature increase that may have a negative impact on lubrication performance due to reduced viscosity. For hydrodynamic lubrication, in which pressure is relatively low and lubricant film is thick so that lubricant shear rate in the film is low, commonly used industrial lubricants can be considered as Newtonian fluids. This means that the shear stress is proportional to shear rate and friction can be reasonably estimated based on viscosity. For the EHL, however, the frictional mechanism becomes more complicated. In the inlet zone of an EHL contact, the pressure is relatively low and the gap is large, so that the shear rate is still low and the Newtonian model may still be acceptable. However, a vast majority of sliding friction is generated in the contact zone, where the pressure becomes high and the film thickness (or gap) very small, resulting in a lubricant shear rate possibly as high as $10^7 \sim 10^8$ 1/s. Under such conditions the lubricant can no longer be considered as Newtonian. Study of lubricant rheology under EHL conditions is a challenging task, as it is difficult to reproduce such high pressure and high shear rate conditions in the lab outside the tiny EHL contact zone. Major



Elastohydrodynamic Lubrication (EHL), Fig. 15 Measured sliding friction in an EHL circular contact. Test conducted by Wedeven Associates, Inc



Elastohydrodynamic Lubrication (EHL), Fig. 16 Friction versus λ ratio. Friction measured by Wedeven Associates, Inc

contributions include non-Newtonian viscous-elastic lubricant models developed by Johnson and Tevaarwerk (1977), and Bair and Winer (1978).

Figure 15 shows typical sliding friction curves under full-film EHL conditions. Basically, as the sliding velocity/lubricant shear rate in the film starts to increase from zero, the friction will increase linearly only within a very narrow range; then it may quickly reach a peak value. This is because the shear stress in the lubricant film cannot go beyond a certain limit, called "limiting shear stress," which is a property of lubricant. Generally, some synthetic lubes, such as advanced gear and engine oils, may have lower limiting shear stress than commonly used naphthenic and paraffinic mineral oils, while traction fluids are purposely formulated to have high limiting shear stress. Also, the limiting shear stress is a function of pressure and temperature, e.g., the higher the pressure, the higher the limiting shear stress, as indicated in Fig. 15. After the friction reaches its peak, if the sliding velocity is further increased, the friction may gradually decrease due to reduction of limiting shear stress caused by temperature increase in the film resulting from heat generation in the sliding contact.

As described above, the limiting shear stress is indeed an important non-Newtonian parameter for EHL friction analyses, in addition to the Newtonian viscosity. Unfortunately, few lubricant suppliers provide systematic limiting shear stress data for their products. Fortunately, EHL friction is relatively easy to measure experimentally, and commercially developed testing devices as well as technical services are available in the market.

In mixed EHL contacts, the friction could be significantly higher than that in full-film EHL, as the total friction now consists of two parts: hydrodynamic friction caused by lubricant shearing in the film (as described above), and boundary friction in the areas of asperity contact. The latter may become dominant as its coefficient is usually higher, typically around 0.07∼0.15. Figure 16 shows a curve of measured mixed EHL friction against calculated λ ratio for a circular contact. It can be seen that as the λ ratio decreases, the friction is increased gradually approaching a limit, which is supposed to be the boundary friction. If the λ ratio is continuously increased, the friction may quickly reduce due to enhanced hydrodynamic action with reduced rough surface asperity contacts and increased lubricant film thickness. This curve is similar to the "Stribeck curve," which was originally developed for showing friction variation in journal bearings. It is obvious that the friction in mixed lubrication can be reduced by the following practical means:

1. Choose advanced lubricants (e.g., synthetic) with low limiting shear stress to reduce hydrodynamic friction;
2. Use advanced additives and/or low friction materials and coatings to reduce boundary friction;
3. Optimize product design/operating conditions, properly engineer surface texture and reduce surface roughness to have λ ratio increased.

Key Applications

EHL is an important branch of lubrication theory, describing lubrication mechanisms in non-conformal contacts, which can be widely found in many mechanical components such as gears, rolling bearings, cams and followers, hydraulic vane pumps, ball screws, traction drives and continuous variable transmissions, and metal rolling tools. These components usually transmit substantial power and motion, and they often carry heavy loads. Due to high stress concentrated in small contact areas, these components are often vulnerable points in machinery. Inadequate lubrication may often cause early surface failures such as excessive wear, scuffing, and pitting due to contact fatigue. A good understanding of the EHL, especially the mixed EHL in which most functional components operate, is vital to product design optimization, performance improvement, efficiency and durability maximization, failure prevention, environment protection, and energy conservation.

Nomenclature

a	Semi-axis of Hertzian ellipse in x-direction, or radius of Hertzian circle
b	Semi-axis? of Hertzian contact ellipse in y-direction
E_1, E_2	Elastic moduli of Body 1 and Body 2, respectively
E'	$2[(1-\nu_1^2)/E_1 + (1-\nu_2^2)/E_2]^{-1}$, effective elastic modulus
G^*	$\alpha E'$, dimensionless material parameter
h	Local film thickness (or gap)
h_a	Average film thickness (or average gap)
h_c	Central film thickness
h_m	Minimum film thickness
l_e	Effective length of line contact
p	Hydrodynamic pressure, or pressure in general
p_h	Maximum Hertzian pressure
R_q	Root mean square (RMS) surface roughness
R_x	$(1/r_{1x} + 1/r_{2x})^{-1}$ for point contact, or $(1/r_1 + 1/r_2)^{-1}$ for line contact, effective radius of curvature in *x-z* plane
R_y	$(1/r_{1y} + 1/r_{2y})^{-1}$, effective radius of curvature in *y-z* plane
S	$(u_2-u_1)/U$, slide-to-roll ratio
U^*	$\eta_o U/(E'R_x)$, dimensionless speed parameter
U	$(u_1 + u_2)/2$, rolling velocity (or entraining velocity)
u_1, u_2	Velocities of Surface 1 and Surface 2, respectively
W^*	$w/(E'R_x^2)$ for point contact, or $w/(E'R_x l_e)$ for line contact dimensionless load parameter
w	Load
W_c	Contact load ratio (load supported by surface contact divided by total load)
x, y	Coordinates (x is chosen to be parallel to rolling direction)
α	Pressure-viscosity exponent used in viscosity equation $\eta = \eta_o EXP(\alpha p)$
η, η_o	Viscosity and viscosity under ambient condition, respectively
λ	h_a/σ, film thickness ratio
ν_1, ν_2	Poisson's ratios of Body 1 and Body 2, respectively
σ	$(R_{q1}^2 + R_{q2}^2)^{0.5}$, composite RMS roughness

Cross-References

- ▶ Deterministic Models of Rough Surface EHL
- ▶ EHL, Full Numerical Solution Methods
- ▶ EHL Governing Equations

- ► Film Thickness Formulas: Line Contacts
- ► Film Thickness Formulas: Point Contacts
- ► Lubricant Film Thickness Measurement: Colorimetric Interferometry
- ► Lubricant Film Thickness Measurement: Relative Optical Interference Intensity (ROII) Method
- ► Mixed EHL
- ► Simplified EHL Solution Methods
- ► Stochastic Models for Rough Surface EHL

References

D. Dowson, G.R. Higginson, A numerical solution to the elastohydrodynamic problem. J. Eng. Sci. **1**, 6–15 (1959)

R. Gohar, A. Cameron, Optical measurement of oil film thickness under elastohydrodynamic lubrication. Nature **200**, 458–459 (1963)

A.N. Grubin, Fundamentals of the hydrodynamic theory of lubrication of heavily loaded cylindrical surfaces (Central Scientific Research Institute for Technology and Mechanical Engineering, Moscow, 1949), Book No.30, (DSIR Translation), pp. 115–166.

G. Guantang, P.M. Cann, A. Olver, H.A. Spikes, Lubricant film thickness in rough surface mixed elastohydrodynamic contact. J. Tribol. **122**, 65–76 (2000)

B.J. Hamrock, D. Dowson, Isothermal elastohydrodynamic lubrication of point contacts, part 1 — theoretical formulation. ASME J. Lubr. Technol **98**, 223–229 (1976a)

B.J. Hamrock, D. Dowson, Isothermal elastohydrodynamic lubrication of point contacts, part 2 — ellipticity parameter results. ASME J. Lubr. Technol **98**, 375–383 (1976b)

M. Hartl, I. Krupka, D. Zhu, EHL film thickness behaviour under high pressure - comparison between numerical and experimental results, in *Proceedings of the IUTAM Symposium on Elastohydrodynamics and Micro-Elastohydrodynamics*, 2005, pp. 217–228

M.J.A. Holmes, H.P. Evans, R.W. Snidle, Analysis of mixed lubrication effects in simulated gear tooth contacts. J. Tribol. **127**, 61–69 (2005)

Y.Z. Hu, D. Zhu, A full numerical solution to the mixed lubrication in point contacts. ASME J. Tribol. **122**, 1–9 (2000)

G.J. Johnston, R. Wayte, H.A. Spikes, The measurement and study of very thin lubricant films in concentrated contacts. Tribol. Trans. **34**, 187–194 (1991)

J.B. Luo, S.Z. Wen, P. Huang, Thin film lubrication.1. Study on the transition between EHL and thin film lubrication using a relative optical interference intensity technique. Wear **194**, 107–115 (1996)

N. Patir, H.S. Cheng, Effect of surface roughness orientation on the central film thickness in EHD contacts, in *Proceedings of the 5th Leeds-Lyon Symposium on Tribology*, Leeds, 1978, pp. 15–21

N. Ren, D. Zhu, W.W. Chen, Y. Liu, Q.J. Wang, A three-dimensional deterministic model for rough surface line contact EHL problems. J. Tribol. **131**(011501), 1–9 (2009)

D. Zhu, On some aspects in numerical solution of thin-film and mixed EHL. Proc IMech, Part J. J. Eng. Tribol. **221**, 561–579 (2007)

D. Zhu, H.S. Cheng, Effect of surface roughness on the point contact EHL. ASME J. Tribol. **110**, 32–37 (1988)

Elastohydrodynamic Lubrication Considering Plastic Deformation

► Plasto-Elastohydrodynamic Lubrication (PEHL)

Elastohydrodynamic Lubrication for Rolling Element Bearings

► Rolling Bearing Lubrication

Elastohydrodynamic Lubrication Friction

► Friction/Traction Behavior of EHL

Elastohydrodynamic Lubrication of Natural Synovial Joints

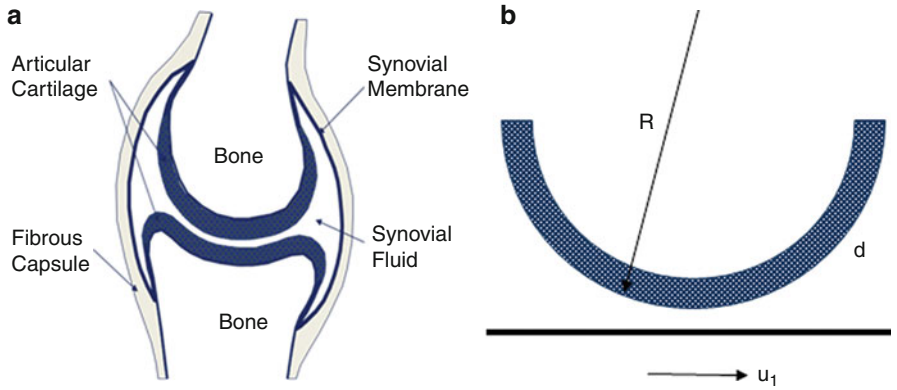
Zhong Min Jin
School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, People's Republic of China
School of Mechanical Engineering, Institute of Medical and Biological Engineering University of Leeds, Leeds, UK

Synonyms

EHL of joints

Definition

Natural synovial joints are remarkable bearings in engineering terms. These man-made devices not only support a large dynamic load, but also allow at the same time a wide range of motions to be accompanied through various activities. Yet, the friction in the natural synovial joint is low and the wear is minimal over the lifetime period. However, joint diseases such as osteoarthritis do occur. Tribology of the bearing surfaces of articular cartilage in the presence of synovial fluid under the load and motion during different activities may play an important



Elastohydrodynamic Lubrication of Natural Synovial Joints, Fig. 1 (**a**) Schematic diagram of synovial joint and (**b**) Simplified lubrication model showing the simplifications of the equivalent geometry (*R*), the combined cartilage thickness (*d*) and the sliding motion (u_1)

role in the understanding of how the healthy joint operates as well as how the joint disease may be developed. One of such important considerations is the lubrication, which is directly related to friction and wear.

Lubrication studies of natural synovial joints date back many decades (Hunter 1743; Reynolds 1886). Many forms of lubrication mechanisms have been proposed to explain the remarkable tribological performance of synovial joints as covered in this *Encyclopedia of Tribology*. It is reasonable to assume that a multi-mode and complex lubrication mechanism exists, covering a continuous spectrum depending on the external factors imposed on the joint. The bearing surface of synovial joints is articular cartilage, which is much softer than the underneath supporting bone and therefore can be expected to deform under load. This, coupled with the geometric characteristics of the bearing surfaces, promoted the introduction of the elastohydrodynamic lubrication (EHL) mechanism for synovial joints (Dowson et al. 1967). Since then, a number of studies have been conducted to understand the elastohydrodynamic lubrication of synovial joints, particularly from a theoretical computational point of view. The common approach is the estimation of the minimum lubricant film thickness, based on the assumption of smooth bearing surfaces, and then the comparison with the surface roughness of articular cartilage to determine the lubrication regime. More and more realistic physiological conditions have been considered as well as the direct incorporation of the surface roughness of articular cartilage.

Scientific Fundamentals

Elastohydrodynamic lubrication modeling of synovial joints requires the cartilage surface geometry, the lubricant of synovial fluid, and the load and motion. Due to the complex three-dimensional geometry of the hip and the knee, a majority of the elastohydrodynamic lubrication analyses have only been carried out for a simplified two-dimensional (plane strain) model of the human ankle joint. Figure 1a shows a schematic diagram of synovial joints. Figure 1b shows the corresponding simplified lubrication model.

Articular cartilage has been reviewed elsewhere in the *Encyclopedia of Tribology*. A ball-in-socket and an ellipsoidal-on-plane geometry can be used to represent the hip and the knee joint, respectively, with appropriate radii of curvature. For the human ankle joint shown in Fig. 1a, b it is possible to use a cylindrical line contact configuration using an equivalent radius (R). It is also possible to just include cartilage in one of the bearing surfaces since the cartilage thickness is much smaller than the bearing radius as well as the contact radius (Medley et al. 1984). The surface roughness of articular cartilage is by no means smooth, even for the healthy specimens with a typical average surface roughness between 2 and 5 μm reported (Dowson and Jin 1986).

The deformation of articular cartilage under loading is time-dependent, as determined from its biphasic composition. Incorporating the time-dependent behavior into the deformation calculation and lubrication modeling has not been easy. Also, under certain conditions, it is possible to neglect the time-dependent deformation. For example, under short- or long-term loading or at physiological walking frequencies, articular cartilage essentially behaves like a simple elastic solid. The main reason for this behavior is the low permeability associated with articular cartilage and that there is little time for the relative

movement between the fluid and the solid phases to occur under rapid loading. Furthermore, due to the small thickness of articular cartilage, it is often possible to simplify the elasticity model, based on a constrained column model where the elastic deformation is directly proportional to the local pressure (Medley et al. 1984). This assumption significantly reduces the computational time and facilitates the EHL solutions of the human ankle joint under transient walking conditions. Furthermore, the porosity of articular cartilage only becomes important when the lubricant thickness is reduced greatly and therefore can be neglected under predominantly EHL conditions (Hlaváček 2010).

The properties of synovial fluid are reviewed in detail separately in the *Encyclopedia of Tribology*. The major consideration of synovial fluid from a lubrication point of view is its rheology. Synovial fluid generally behaves like a non-Newtonian shear thinning fluid, particularly under low shear rates. However, under physiological walking conditions, the shear rate experienced within synovial fluid is quite high, reaching 10^5–10^7 1/s. Under these conditions, the shear stress and shear rate relationship for the non-Newtonian synovial fluid approaches asymptotically; essentially like a Newtonian fluid with a viscosity not much different from that of the base constituent of water. It is therefore only necessary to use a constant viscosity to represent the rheology of synovial lubricant.

Time-dependent loading and motion are generally encountered in synovial joints during walking and other daily activities. Therefore, both the entraining motion and the squeeze-film motion are required in the lubrication modeling of the Reynolds equation. The finite difference method is usually used to discretize the Reynolds equation. The discretized equations are usually solved using a number of methods, such as the Newton–Raphson method.

The predicted minimum film thickness is often used in comparison with the surface roughness to access the lubrication regime using the common engineering approach. It is also possible to directly incorporate the roughness into the EHL analysis (micro-EHL (Dowson and Jin 1986)).

Key Applications

It is useful and informative to estimate the lubricant film thickness, based on different lubrication mechanisms, and the corresponding lubrication regimes. Table 1 summarizes the typical parameters adopted for the EHL analysis of the normal human ankle joint lubrication model shown in Fig. 1b.

Table 2 shows the predicted lambda ratios and the corresponding lubrication regimes.

Elastohydrodynamic Lubrication of Natural Synovial Joints, Table 1 Typical conditions used for the lubrication modeling of a human ankle joint (Medley et al. 1984; Dowson and Jin 1986)

Equivalent radius (R)	0.35 m	Lubricant viscosity	0.01 Pas
Mean total cartilage thickness (d)	2.4 mm	Average entraining velocity	0.0191 m/s
Modulus of elasticity (E)	16 MPa	Average load/unit width	33.7 kN/m
Poisson's ratio	0.4	Average roughness	2–5 μm

Elastohydrodynamic Lubrication of Natural Synovial Joints, Table 2 Predicted lambda ratios and corresponding lubrication regimes based on the common engineering approach

Lubrication mechanisms	Lambda ratio	Lubrication regimes
Hydrodynamic	0–0.005	Boundary
Elastohydrodynamic	0–0.2	Boundary
Micro-elastohydrodynamic	0–10	Boundary to fluid film
Transient elastohydrodynamic	0.1–0.2	Boundary
Transient micro-elastohydrodynamic	1–10	Mixed to fluid film

The effect of geometry contribution to the film formation of lubrication through the hydrodynamic action is small. The predicted lubricant film thickness is much smaller than the surface roughness of articular cartilage and therefore a boundary lubrication regime would be predicted. However, in reality, articular cartilage can readily be deformed due to its relatively low elastic modulus, and this promotes the formation of a substantially increased lubricant film thickness. The predicted minimum film thickness from the elastohydrodynamic lubrication mechanism is increased by at least two orders of magnitude. However, the predicted lubricant film thickness is still substantially smaller than the surface roughness of articular cartilage, also resulting in a boundary lubrication regime.

The consideration of the surface roughness of articular cartilage substantially increases the lambda ratio,

calculated from the deformed roughness. The main reason is that the original surface roughness is largely smoothed out due to the local elastohydrodynamic action. This is also common in engineering in the so-called micro-elastohydrodynamic lubrication mechanism. This action is particularly effective in soft contacts such as articular cartilage. Only a modest pressure perturbation is required to largely smooth out the cartilage roughness. The original roughness may be deformed by two orders of magnitude. As a result, the roughness of the deformed cartilage surface is much smaller than the original one and a significantly increased lambda ratio is predicted (Dowson and Jin 1986).

However, all the lubrication mechanisms discussed above rely on the entraining action through the relative sliding between the two bearing surfaces. When the speed in the joint goes to zero, as at the reverse instant during the walking cycle, the predicted film thickness becomes zero. It is generally known that the squeeze-film action is important under these adverse operating conditions. Through the combined effect of squeeze-film and entraining actions, an almost constant film thickness is generated during steady-state walking cycles (Medley et al. 1984). During the stance phase, the lubricant film thickness mainly developed during the swing phase is largely preserved and maintained. During the swing phase, the lubricant film thickness is largely replenished due to the small load and the large sliding velocity. Nevertheless, the predicted transient lubricant film thickness is still significantly smaller than the original surface roughness of articular cartilage, leading to a predicted boundary lubrication regime. When the transient lubrication mechanism is combined with the micro-elastohydrodynamic action, a substantial increase in the lambda ratio is predicted throughout a walking cycle and a fluid film lubrication regime is likely under steady-state walking conditions.

Despite all the efforts devoted to the study of the elastohydrodynamic lubrication mechanism in natural synovial joints, a complete understanding is still lacking. A wide range of operational conditions in the natural synovial joints means that a multi-mode of lubrication mechanisms is likely, covering from boundary to a full fluid film regime. As yet, such an analysis has not been attempted. It is also important to recognize the use of simplified lubrication models and the limitations (Hlaváček 2010). In the future, the lubrication model for a whole joint, considering other soft tissues as well articular cartilage, is required. In parallel to the refinement of the lubrication modeling of synovial joints, the application of elastohydrodynamic lubrication theories to other biological organs is equally important, such as in the pleural as addressed in other entries in the *Encyclopedia of Tribology*.

Cross-References

► Lubrication Modeling of Artificial Hip Joints

References

D. Dowson, Models of lubrication in human joints, in *Proceeding Symposium on Lubrication and Wear in Living and Artificial Human Joints*, vol. 181(3 J) (Institution of Mechanical Engineers, London, 1967), pp. 45–54

D. Dowson, Z.M. Jin, Micro-elastohydrodynamic lubrication of synovial joints. Eng. Med. **15**, 63–65 (1986)

M. Hlaváček, Lubrication of the human ankle joint in walking. J. Tribol. **132**, 011201-1–011201-8 (2010)

W. Hunter, Of the structure and diseases of articulating cartilages. Philos. Trans. R. **42**(1743), 514 (1743)

J.B. Medley, D. Dowson, V. Wright, Transient elastohydrodynamic lubrication models for the human ankle joint. Eng. Med. **13**(3), 137–151 (1984)

O. Reynolds, On the Theory of Lubrication and its Application to Mr. Beauchamp Tower's Experiments Including an Experimental Determination of the Viscosity of Olive Oil. Philos. Trans. R. Soc. **177**(1), 157–234 (1886)

Elastohydrodynamic Lubrication with a Non-Newtonian Lubricant

► Lubricant Non-Newtonian Effect on EHL

Elastohydrodynamic Lubrication, Line Contacts

► Lubrication Regimes – Line Contacts

Elastohydrodynamics

► Elastohydrodynamic Lubrication (EHL)

Elastomer Composites and Nanocomposites

► Polymeric Elastomers: Material Aspects of Tribology

Elastomers

► Polymeric Elastomers: Material Aspects of Tribology

Electric Contact, Elements, and Systems

Xin Zhou
Innovation Center, Eaton Corporation, Moon Township, PA, USA

Synonyms

Electrical contact

Definition

Electric contact is an electrically conductive material that can conduct currents when a pair of electric contacts joins electrical circuits together.

Scientific Fundamentals

Electric contacts are pairs of conductive materials that join electrical circuits together. Usually these conductive materials are metals and metal alloys. The most commonly used contact materials include Au, Ag, Cu, Al, Ag Oxide, C, AgW, AgC, AgNi, AgPd, AgWC, and CuCr.

The basic elements in an electric contact system are contacts. The contact with positive polarity is called anode, and the contact with negative polarity is called cathode. Current flows from the anode to the cathode. An electric contact system includes a pair of electric contacts, anode and cathode, and a mechanical structure that provides contact force so the electric contacts can join the electrical circuit together with minimum electric contact resistance.

The mechanical structure in an electric contact system varies depending on different applications. For example, in electromechanical switching devices, the mechanical structure includes a contact spring that provides a contact force, a structure that holds and deforms the contact spring to provide the contact force when contacts are joined together, and a mechanical mechanism that operates the mechanical structure to open and close the pair of electric contacts to join the electrical circuit or open the electrical circuit. The mechanical mechanism can be operated manually, pneumatically, or magnetically.

For a contact system in a contactor, as shown in Fig. 1, a force balance equation on the movable contact bridge can be expressed as (Zhou and Theisen 2000):

$$m \cdot \frac{d^2 s}{dt^2} = -F_{spring} - m \cdot g - F_{magnet} + F_{arc} + F_{constriction} \quad (1)$$

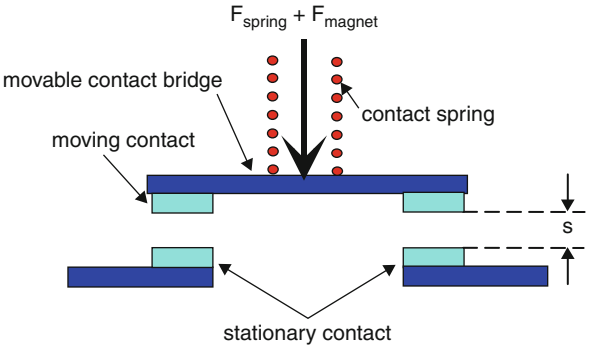
where m is the movable contact bridge mass, g is the gravitational acceleration, s is the contact gap, and t is the time.

The contact spring force F_{spring} can be expressed as:

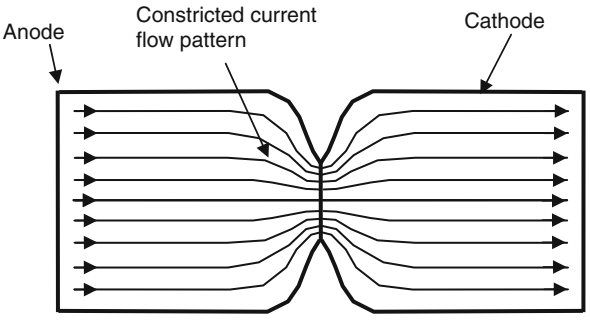
$$F_{spring} = s \cdot \xi + F_{initial} + \left(L_{initial} - L_{final}\right) \cdot \xi \quad (2)$$

where ξ is the spring rate (N/m), $F_{initial}$ is the initial spring force, $L_{initial}$ is the initial spring length, and L_{final} is the final spring length when contacts fully close.

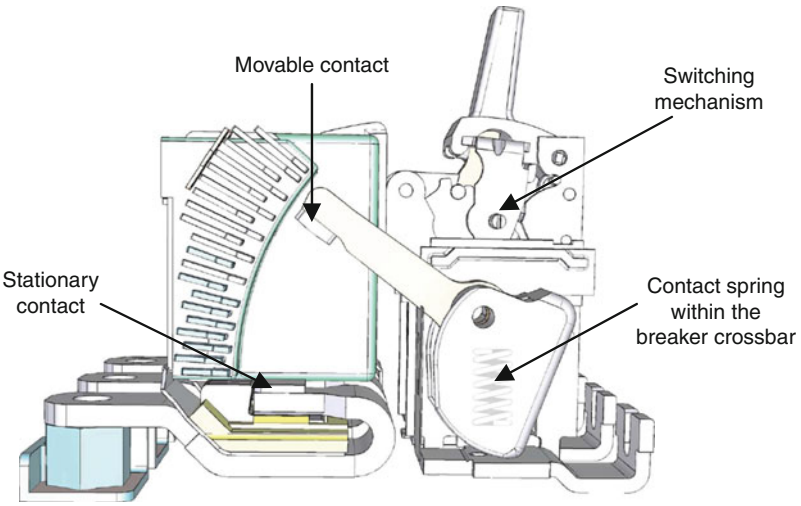
A magnetic force (also called the Larentz force) F_{magnet} is induced by electric current going through the current conducting path. The magnetic force is a function of both current and contact gaps in this case.



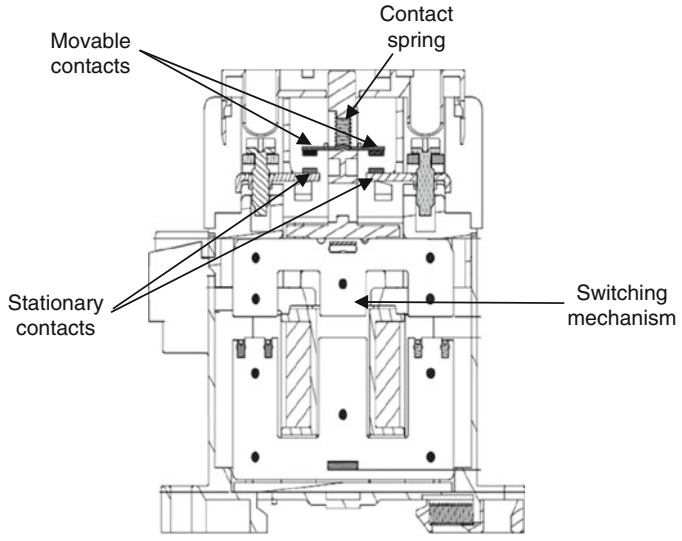
Electric Contact, Elements, and Systems, Fig. 1 A contact system in a contactor (Courtesy of Eaton Corporation)



Electric Contact, Elements, and Systems, Fig. 2 Illustrative schematic of an electric contact and the constricted current flow pattern through the contact spot (Courtesy of Eaton Corporation)



Electric Contact, Elements, and Systems, Fig. 3 A contact system in a circuit breaker (Courtesy of Eaton Corporation)



Electric Contact, Elements, and Systems, Fig. 4 A contact system in a contactor (Courtesy of Eaton Corporation)

Due to the constricted current flow pattern at the contact interface, as shown in Fig. 2, a magnetic force is generated by the constricted current going through the contact interface. This magnetic force is called constriction force or Holm force. The constriction force works against the contact spring force trying to push the contacts open. The constriction force $F_{constriction}$ can be expressed as (Holm and Holm 1967):

$$F_{constriction} = \frac{\mu \cdot I^2}{4 \cdot \pi} \cdot \ln\left(r\sqrt{H \cdot \pi / F_c}\right) \qquad (3)$$

where H is the hardness of the contact material [N/m^2], r is the contact radius [m], F_c is the contact force [N], μ is the permeability [H/m], and I is the current [A]. As soon as the contacts are apart, the constriction force will disappear and an arc will be drawn across the contacts. The arc can

Electric Contact, Elements, and Systems, Table 1 Contact materials for different switching device applications

Applications	Low current (under 10 A)	Medium currents (10–100 A)	Med-high currents (100–1,000 A)	High currents (>1,000 A)
A.C. switches, relays, and contactors	Ag			
	Fine-grain Ag	Ag/Ni		
	AgCu	Ag/CdO	Ag/CdO	
	Ag/Ni	Ag/SnO_2	Ag/SnO_2	
D.C. automotive relays and switches	Ag/Ni	Ag/Ni		
	Ag/Cu	Ag/Cu		
	Ag/SnO_2	Ag/SnO_2		
	Ag/NiO/MgO	Ag/NiO/MgO		
		Pd		
Circuit-breakers USA			WAg	
		WAg	WC/Ag	
		WC/Ag	Ag/SnO_2	
			Ag/W-Ag/CdO	
Circuit-breakers Europe			Ag/CdO	
	Ag/CdO	Ag/CdO	Ag/SnO_2	
	Ag/ZnO	Ag/ZnO	Ag/C-Cu	
	Ag/SnO_2	Ag/SnO_2	Ag/C-Ag/Ni	
Thermostats	Ag			
	Fine-grain Ag	Ag/CdO		
	AgCu	Ag/SnO_2		
	Ag/Ni			
Power interrupters				W/Ag
			W/Cu vacuum	WC/Ag
				W/Cu SF6 or Oil
				Cu/Cr vacuum

Source: Slade 1999

generate a high pressure force F_{arc} $[N]$ on the contact surface trying to push the contacts further apart.

Key Applications

Electric contacts are widely employed in electrical systems ranging from electrical power generation and transmission to distribution and controls. They are there to join electrical cables and busbars together, to connect electrical equipments to electrical power systems, to control electrical power, and to provide protection to electrical systems and personnel when there is an electrical fault. These types of devices normally need to meet various regulations and standards such as UL, IEC, ANSI, KEMA, and others. For electromechanical switching devices, they need to meet code and standard performance requirements such as mechanical endurance, electrical endurance, temperature rise, overload, and short circuit interruption. When a pair of current carrying contacts separate, an arc is drawn across the pair of contacts. The arc has a very high temperature, at 6,000 K or higher. The high arc temperature melts the contact material and causes contact erosion during switching processes. The arc erosion of contacts leads to reduced life of switching devices. Contact erosion and contact weld are the two of the major issues facing these electric contacts in switching devices. Based on their applications, these devices can be categorized into circuit breakers (Fig. 3), contactors (Fig. 4), motor starters, switches, and relays. Different contact materials are used for different types of switching devices based on the application's requirements. Table 1 shows contact materials used in various electromechanical switching devices.

Electric Contact, Elements, and Systems, Table 2 Contact materials for different switching device applications

Type	Contact form	Spring member	Base metal	coating	Typical current (A)	Termination	Common usage
Wire-wire twist	Multiple wire-wire	Coiled spring insert	Cu, Al	none	≤20	None	Household wiring
Wire-screw or lug-screw	Wire-flat or flat-flat	Deformed thread/washer	Cu, Al	none, Sn	≤100	None or crimp	Household wiring/ appliance
IDC	Wire-blate	Cantilevered	Cu, Cu-alloys	none	≤20	None or crimp	Commercial, automotive
Tuning fork	Wire-blate	Cantilevered	Cu, Cu-alloys	none	≤20	None or crimp	Commercial, automotive
Blade-box	Row	Deformed box	Cu-alloys	none, Sn, Ni	≤40	Crimp or solder	Appliance, automotive
Blade-leaf	Beam	Cantilevered	Cu-alloys	Sn, Ag, Au, Ni	≤30	Crimp or solder	Automotive, appliance, commercial
Pin-sleeve	Multiple beam	Cantilevered	Cu-alloys	Sn, Ag, Au	≤10	Crimp, IDC or solder	Commercial, appliance, automotive
Pin-hyperboloid	Wire-pin	Stretched wire	Cu, Cu-alloys	Sn, Ag, Au	≤2	Crimp, IDC or solder	Commercial
Bump-flat	Butt	External	Cu, Cu-alloys	Sn, Pd, Ag, Au	≤2	None or solder	Commercial, automotive

Source: Slade 1999

In electrical connection or electrical joint applications, usually the system includes a pair of electric contacts, and a mechanical structure that holds the electric contacts together with minimum electrical contact resistance. The mechanical structure can be clamps, bolts, lugs, or screw nuts. The mechanical structure can also be the set of electric contacts themselves with spring-loaded force. These types of electric contacts experience no arc erosion, but fretting and corrosion are common problems for these types of contacts. The current range goes from microamps to thousands of amps. Their applications range from high-frequency, low-current communication signals to power transmission, distribution, and controls. Contact materials are also different from those used in switching devices. Thin film coatings such as plating sometimes are used to prevent oxidation and corrosion. Table 2 provides a summary of terminal types, their base metals, and coatings.

Cross-References

- ► Contact Boiling Voltage
- ► Contact Melting Voltage
- ► Contact Temperature of a Moving Solid Surface
- ► Contact Temperature of a Stationary Solid Surface
- ► Contact Temperatures on Coated or Rough Solid Surfaces

References

R. Holm, E. Holm (eds.), *Electric contacts: theory and application* (Springer, New York, 1967)

P. Slade (ed.), *Electrical Contacts: Principles and Applications* (Marcel Dekker, New York, 1999)

X. Zhou, P. Theisen, Investigation of arcing effect on contact blow open process. IEEE Trans. Compon. Packag. Technol. **23**(2), 271–277 (2000)

Electric Discharge Machining (EDM) for Gear Manufacturing

- ► Gear Manufacturing Machines

Electrical Brushes

Koichiro Sawa
Department of System Design Engineering,
Keio University, Kohoku-ku, Yokohama, Japan

Definition

An electrical brush is a conductor to transfer electrical signal or electrical power between a stationary part and a movable part that is usually a rotating commutator, slip ring, or other moving conductors.

Scientific Fundamentals

An electrical brush is mainly a carbon-type brush. In a DC motor it is used to carry out commutation together with a commutator for many years (Shobert 1965; Holm 1967; Slade 1999). Further, it is used to supply exciting current to a generator with a slip ring.

A carbon-type brush is mainly divided into carbon brush and metal-graphite brush.

When many large DC motors for iron mills and electric vehicle traction were manufactured, carbon brushes are predominant. However, presently, such large DC motors are small in numbers in the world, while huge number of small DC motors are produced for office automation apparatus and automotive applications (http:aupac). Metal-graphite brushes are mainly used for such small motors. In addition, precious metal brushes are used for very small DC motors (micro motors) (Slade 1999; http:tanaka).

As electrical brushes supply current from stationary devices to moving ones, low contact resistance and stable sliding motion are requested. Further, sliding motion is accompanied with brush wear, and low wear rate of brush is necessary for a long lifetime (Shobert 1965; Holm 1967; Slade 1999).

Fundamental Characteristics of Carbon Brushes

Contact Voltage Drop

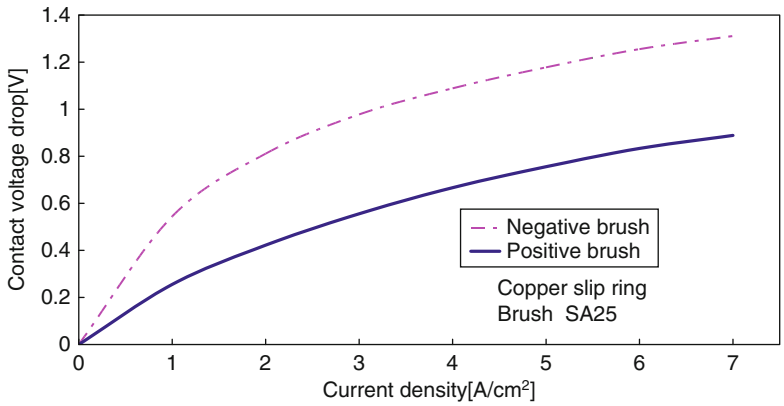
Contact voltage drop is not decided only by the brush but is also affected by mating material of commutator or slip ring.

The brush of DC motor slides on the commutator made of copper. Contact voltage drop between brush and commutator is affected by surface film on the commutator (Shobert 1954; Robert et al. 1995). The V-I relation is generally nonlinear, as shown in Fig. 1. The brush polarity is defined as follows: the brush is called positive or anodic, where the current flow is from the brush to the commutator, and negative or cathodic, where the current flow is from the commutator to the brush (Ichiki 1978).

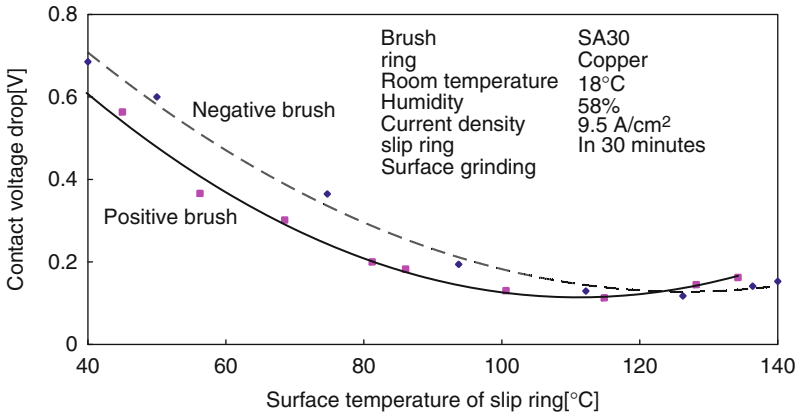
And surface temperature has strong influence on the contact voltage drop (Alley et al. 1991). Figure 2 shows an example. At low temperature, contact voltage drop is high and dependent on polarity of the brush. On the other hand, at high temperature, the contact voltage drop decreases and comes to around 0.1 V at 90–100°C that is independent of brush polarity (Ichiki 1978).

Coefficient of Friction

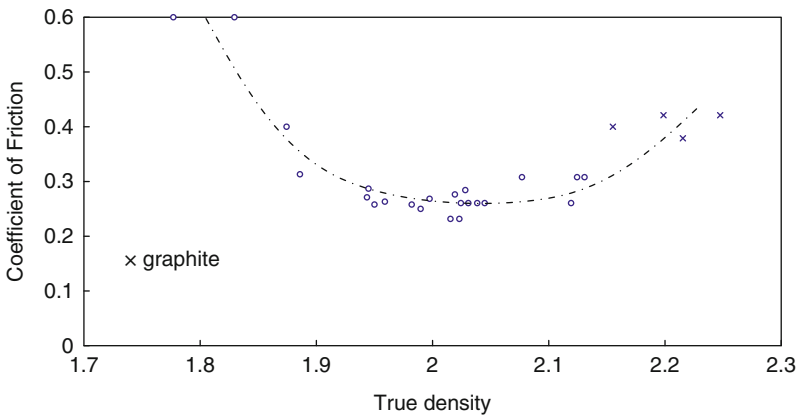
Friction behavior is one of the fundamental characteristics of sliding contact. The friction generates a temperature rise around the brush (brush, holder, commutator, or slip ring) or excites brush vibration.



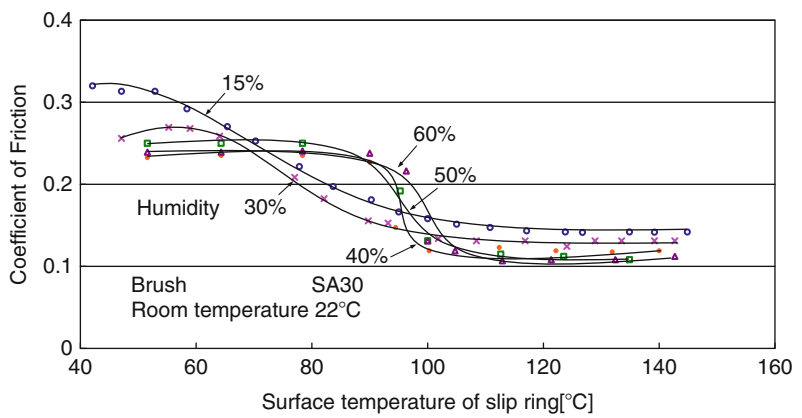
Electrical Brushes, Fig. 1 V-I characteristics of carbon brush



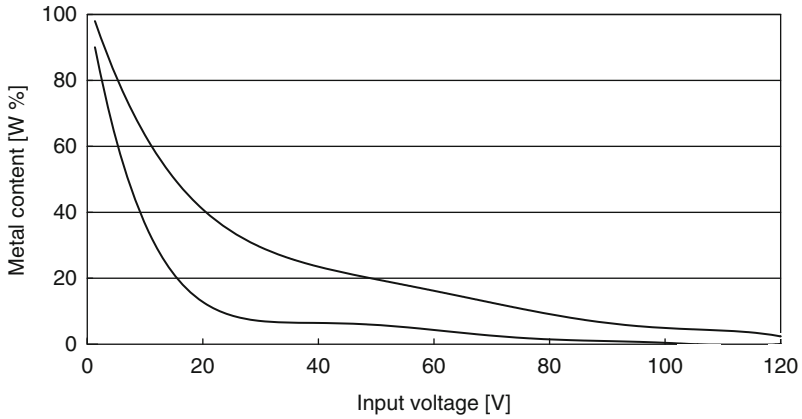
Electrical Brushes, Fig. 2 Contact voltage drop vs. surface temperature



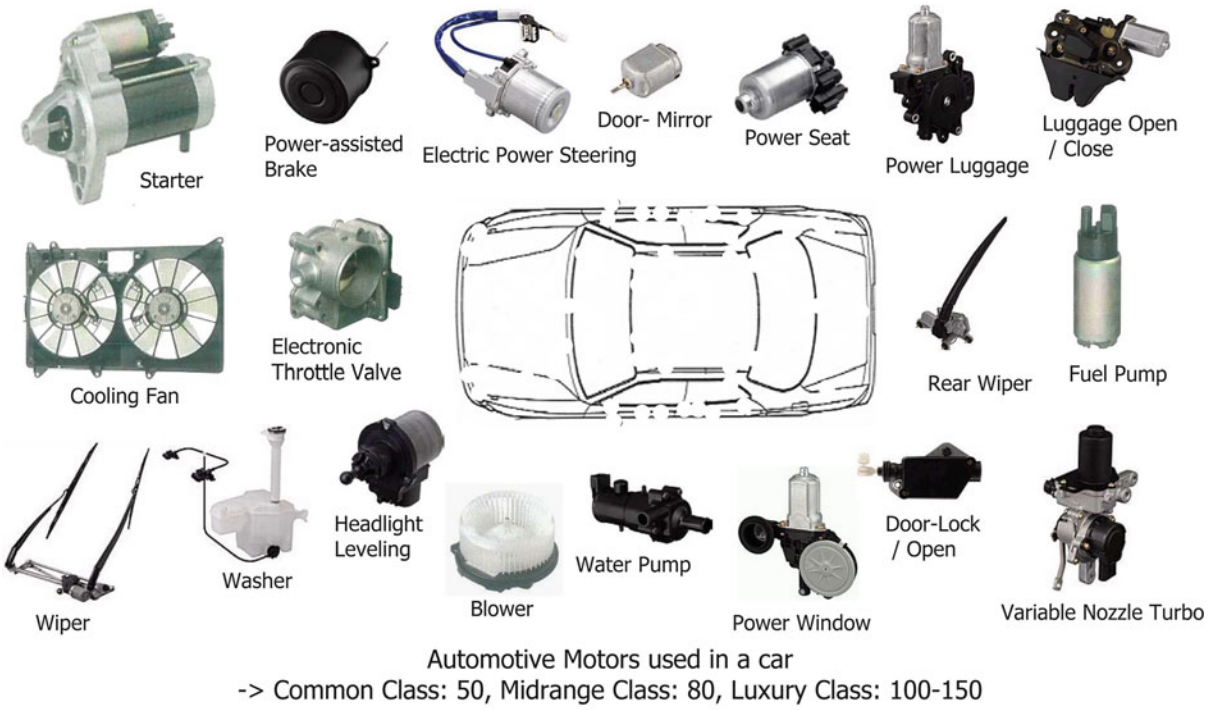
Electrical Brushes, Fig. 3 Coefficient of friction vs. true density of brush



Electrical Brushes, Fig. 4 Coefficient of friction vs. surface temperature



Electrical Brushes, Fig. 5 Metal content vs. input voltage of motor

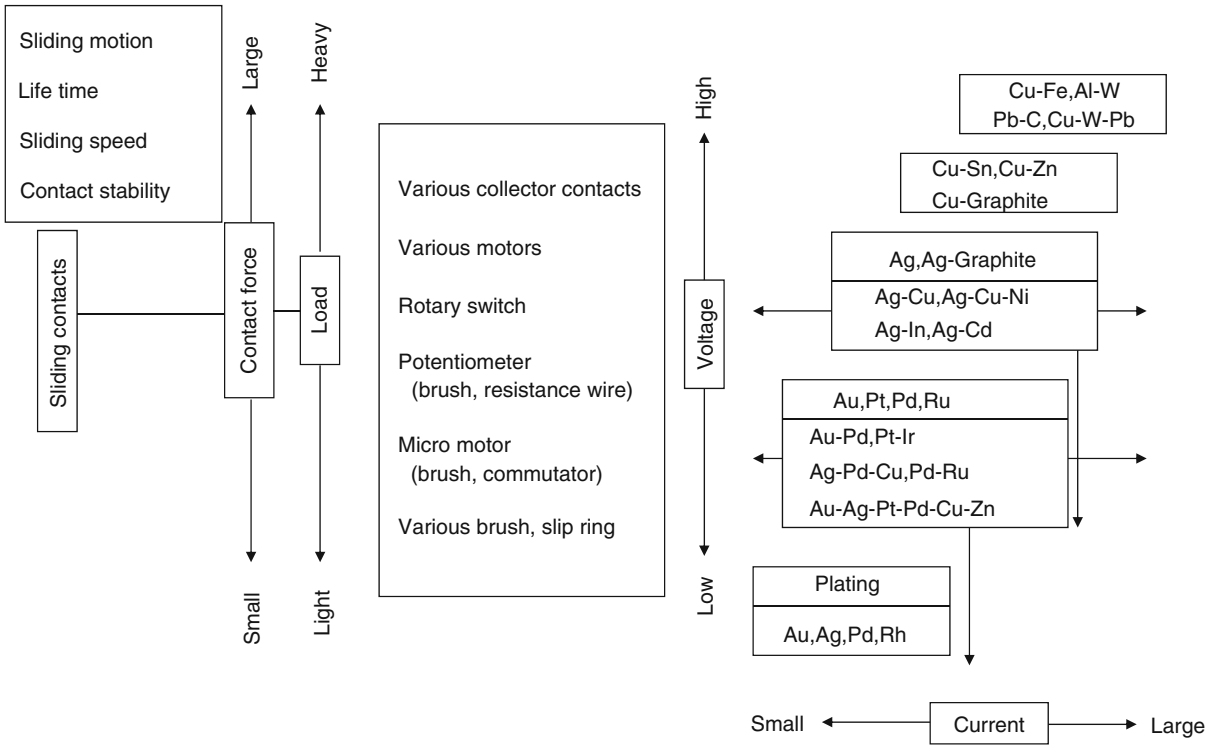


Electrical Brushes, Fig. 6 Various applications to automotive devices

The coefficient of friction for brushes is affected by brush material, surface condition of the commutator, commutator temperature, and atmosphere temperature.

Figure 3 shows a relation between coefficient of friction and true density (Ichiki 1978). The true density or true specific gravity shows graphitization degree. Plotted results are obtained from many kinds of commercially available brushes. The results are dependent only on the true density, independent of other brush specifications. This means that the coefficient of friction is decided by true density, that is, graphitization degree.

In addition, the coefficient of friction increases rapidly below 1.95 of the true density, while it also becomes high around 2.25 of the true density. Generally, the lubricity increases with the graphitization degrees.



Electrical Brushes, Fig. 7 Various applications of precious metals and their alloys

But in case of very high graphitization degree, around 2.25, the surface of a commutator or slip ring is covered with graphite film and then the coefficient of friction becomes high.

Generally the coefficient of friction of suit or pitch cokes grade brush stay almost constant during the operation, while the coefficient of friction for graphite brush increases with the operation time.

Figure 4 shows a relation between coefficient of friction and surface temperature of a slip ring (Ichiki 1978; Yune and Bryant 1988).

The coefficient of friction decreases rapidly with increasing temperature (Dobson 1935). Its behavior is dependent on ambient humidity, but its value becomes almost constant over 100°C. This may be caused by dehydration of cuprous oxide.

Metal-Graphite Brush

The metal content (usually copper) is roughly decided by the input voltage of a motor as shown in Fig. 5. The content is decided as a value between two curves.

For example, a brush of about 50% copper is selected for motors of automobile appliances with 12 V input voltage, and around of 20% for motors of electric power tools with 36 V input voltage (Lu and Bryant 1994; Wingert et al. 1992). Further, carbon brushes are used for universal motors of vacuum cleaners, with the input voltage over 100 V.

Precious Metal Brush

This type of brush is used mainly for micro motors and applications when low contact resistance and high reliability are required for small load current (http:tanaka).

Key Applications

Various grades of brushes are widely used for many applications.

Electro-graphite brushes: large DC motors for steel rolling mills or electrical vehicle traction motors. Currently, they are produced as replacement parts for existing machines.

Carbon brushes: universal motors for vacuum cleaners and electric power tools with commercial power supply of 230 V or 100 V.

Metal graphite brushes: small DC motors for power tools driven by batteries and automobile appliances.

Generally speaking, metal graphite brushes are used where a high current capacity is required and where the contact voltage drop must be kept low.

Resin-bonded brushes: the same application with carbon brushes, especially for cleaner motors with difficult commutation due to high speeds.

Recently, small DC motors are have seen use in electric toothbrushes, hair dryers, and food processors at home, and in outdoor cash dispensers and vending machines.

Among them, automotive small motors have seen a dramatic increase in numbers for more convenient and comfortable cars (http:asmo). Automotive applications can be classified into three groups: auxiliary motors, alternators, and starter motors. There are many auxiliary motors used in automobiles, for example, electric power steering, fuel pumps, door locks, power mirrors, and more. About 120 motors are used for one luxury car. Figure 6 shows a typical example of motor applications in a car.

Precious metal brushes: used for various devices such as potentiometers, trimmers, micro motors, and many kinds of sensors. Various metal materials are provided for many applications as shown in Fig. 7.

Cross-References

► Brush Materials
► Friction Coefficient
► Sliding Electrical Contact Wear
► Sliding Wear

References

D.M. Alley, L.J. Hagen, D.K. Wilsdorf, D.D. Makel, C.G. Moore III, Automated apparatus for long-term testing of electrical brushes. IEEE Trans. Compon. Hybrids Manuf. Technol. **14**, 31–36 (1991)
J.V. Dobson, The effect of humidity on brush operation. Electric J **32**, 527 (1935)
R. Holm, *Electric Contacts* (Springer, New York, 1967)
T. Ichiki, "*Theory and Applications of Electric Brushes*" in Japanese (Corona, Tokyo, 1978)
C.T. Lu, M.D. Bryant, Simulation of a carbon graphite brush with distributed metal particles. IEEE Trans. Compon. Hybrids Manuf. Technol. **17**, 68–77 (1994)
F. Robert, D. Paulmier, H. Zaidi, E. Schouller, Combined influence of an inert gas environment and a mechanical action on a graphite surface. Wear **181–183**, 687–690 (1995)
E.I. Shobert II, *Carbon Brushes* (Chemical Publishing, New York, 1965)
E.I. Shobert II, Electrical resistance of carbon brushes on copper rings. IEEE Trans. Power Appar. Syst. **13**, 788–797 (1954)
P.G. Slade, *Electrical Contacts* (Marcel Dekker, New York, 1999)
B.P.C. Wingert, S.E. Allen, R.C. Bevington, The effects of graphite particle size and processing on the performance of silver-graphite contacts. IEEE Trans. Compon. Hybrids Manuf. Technol. **15**, 154–159 (1992)
Y.G. Yune, M.D. Bryant, Transient nonlinear thermal runaway effects in carbon graphite electrical brushes. IEEE Trans. Compon. Hybrids Manuf. Technol. **11**, 91–100 (1988)
http://www.aupac.co.jp/e/index.html
http://www.tanaka.co.jp/english/
http://www.asmo.co.jp/en/product/index.html

Electrical Contact

► Electric Contact, Elements, and Systems

Electrical Contact Materials

Roland S. Timsit
Timron Advanced Connector Technologies, A Subsidiary of Timron Scientific Consulting Inc., Toronto, ON, Canada

Definition

Electrical contact materials refer to a class of metals that are used in the electrical interfaces of electrical connectors and electrical switches. These materials are characterized by superior electrical conductivity properties and may be used as bulk conductors, such as copper, copper alloys, aluminum, aluminum alloys, silver, and silver alloys. Other conductors such as precious metals, tin, tin alloys, and selected silver alloys are used as coatings on electrical contact surfaces.

Scientific Fundamentals

Criteria for Selecting Electrical Contact Materials

The range of acceptable materials in an electrical connector is determined by the end-use and hence by the service lifetime requirement for the connection. For example, although unalloyed copper is an excellent electrical conductor, its proneness to oxidation and galvanic corrosion generally preclude its use in bare form in an electrical contact. Similarly, copper-base alloys are seldom used in bare form. Thus, electrical connectors fabricated from copper and copper-base alloys are often coated with metals such as silver, tin, or even gold or palladium. The choice of coating material and coating thickness is also determined by requirements on maximum contact

resistance, resistance to mechanical wear, and other factors that limit connector performance.

Gold Deposits

Gold in the form of a thin coating is widely used in high-reliability connectors for computer, telecommunication, and data transmission applications due to its excellent resistance to tarnish and oxidation. Pure gold is soft and has a tendency to cold-weld and seize in a contact and must therefore be used with caution in applications where tribological properties are important. Gold is often deposited in conjunction with small additions (0.1–0.2 wt%) of other metals such as cobalt or nickel to increase hardness and increase resistance to mechanical wear. Gold containing a hardener is identified generically as "hard gold."

Tin Deposits

Tin and tin alloys have been widely used as electroplates on connector contact surfaces. The popularity of tin stems in part from its relatively low cost. The susceptibility of pure tin to forming whiskers often leads to electrical shorts between connector leads. This susceptibility is essentially eliminated through additions of lead. Connectors and other contact components such as lead-frame integrated circuit packages have thus used tin–lead alloys as a surface finish. However, the RoHS legislation (Restrictions of the Use of Certain Hazardous Substances) (RoHS 2012) for the use of lead-free materials has required the removal of lead, including lead from tin alloys used in electronic devices (Warwick 1999). This has necessitated the development of techniques to minimize whisker formation, such as annealing at temperatures slightly below melting, the use of underplate structures, and so on (Osenbach et al. 2005). Because tin and tin-alloy electroplates are soft and prone to mechanical wear and abrasion, they are limited to applications in permanent electrical connections (e.g., bolted connections) and in separable connectors where relatively few connector insertions/withdrawals are expected in their lifetime. For many connectors, tin is used in preference to other materials such as gold and silver because it is relatively resistant to galvanic corrosion and inexpensive, although its electrical resistivity is about seven times that of copper. The high resistivity of tin is commonly accommodated by using a sufficiently thin layer of the material to minimize its deleterious effect on contact resistance. The small Vickers hardness of tin leads to increased area of true contact and often offsets the effects of the large resistivity.

Nickel Deposits

Nickel is often used in separable electrical connections in special applications such as flashlight and automobile batteries. The popularity of nickel stems in part from its relatively low cost and its high resistance to mechanical wear. However, nickel forms a tenacious surface oxide layer that must be fractured to allow passage of electrical current. For this reason, nickel is recommended primarily as an underplate material for surface deposits such as gold, tin, and other coatings on copper-base substrates. A nickel underplate constitutes an effective barrier to diffusion of rapidly-diffusing species, such as gold and tin, into copper-base alloys up to connector service temperatures of about 180 °C (Antler 1999a; Haimovich 1989). Gold and tin coatings are generally deposited on about 2.5-μm-thick nickel layers. The interdiffusion rate between nickel and copper (or copper-base alloys) is negligible, even at temperatures as high as 400 °C (Schwarz et al. 1999b). A nickel underplate mitigates wear in noble metal overlayers (Antler 1999b).

Silver Deposits

Pure silver has the highest electrical and thermal conductivity of all electrical contact materials. The metal does not oxidize in air, but has a tendency to form sulfide and chloride films (Antler 1999b; Chudnovsky 2002). The use of silver in electrical contacts focuses largely on high-power electrical connections (Imrell 1991). Silver is also prone to forming whiskers in the presence of high concentrations of sulfur-containing atmospheric contaminants (Chudnovsky 2003), and to form dendrites in the presence of moisture surface films under the action of a potential difference between contact surfaces, that is, wet electromigration (Krumbein 1994). In recent years, the high price of gold has led to a re-examination of the substitution of gold coatings by silver in electronic connectors.

Palladium and Palladium Deposits

One potential substitute for gold is palladium. Palladium is available not only as a plating but also as a wrought or a cladding material and is appreciably less costly than gold. One major challenge to the use of unalloyed palladium is its susceptibility to the formation of electrically-insulating compounds known as "frictional polymers" in the presence of even small traces of organic contaminants in air [Hermance and Egan 1958]. The tendency of palladium to form frictional polymers in a sliding interface stems from the catalytic action of the metal on organic deposits to form a variety of insulating organic compounds.

One approach to mitigating frictional polymer formation in palladium-palladium contacts is to use a gold flash (thickness of ~ 0.1 μm) or a thin lubricant layer on the palladium surfaces, in order to mitigate access to the palladium surface by polymer-forming organic contaminants.

Promising substitute materials for hard gold as a contact finish include palladium-nickel and palladium-cobalt with nickel and cobalt concentrations ranging from about 10 to 20 wt%. For several years, another promising substitute material for hard gold has been palladium-silver consisting of 60 wt% palladium and 40% silver (Pd60Ag40). This material may be electrodeposited on a contact surface but is also available as a wrought alloy (also known as R156) and may be incorporated into a surface as a clad inlay which is then further fabricated into contact components [Nobel 1984]. One alloy fabricated to offer a gold-rich surface, is a modified version of R156 and has proven to have excellent contact properties when subjected to screening evaluations consisting of corrosion, wear, thermal aging, and reciprocating motion with micron-scale displacement amplitude.

Clad Materials

Clad contact materials represent an attractive alternative to plated layers. The main attributes of cladded layers stem from the diversity and the composition range of cladding materials, which contrasts with the narrow materials choice for electrodeposits. In addition to the advantage of wide alloy range, selected physical properties of claddings such as yield strength and hardness may be controlled by controlling the grain size. Grain size control is achieved relatively easily via metal-working during various stages of the rolling process where the clad is thinned to its final thickness, and by subsequent thermal treatment. Another advantage of claddings is mechanical formability after bonding to the substrate. Clad components may be deformed significantly without tensile fracture of the cladding, in contrast with the generally deleterious effects of deformation on electrocoated surfaces. However, one drawback of claddings is the difficulty of laying them on surfaces that are not either flat or uniformly curved. Another potential drawback is cost.

Cross-References

► Contact Materials in Vacuum Interrupters
► Electrical Contact Materials
► Electrical Contacts: Scientific Fundamentals
► Sliding Electrical Contact Wear
► Tribology

References

M. Antler, Materials, coatings, and platings, in *Electric Contacts: Theory and Applications*, ed. by P.G. Slade (Marcel Dekker, New York, 1999a), pp. 403–432

M. Antler, Tribology of electronic connectors: contact sliding wear, fretting and lubrication, in *Electric Contacts: Theory and Applications*, ed. by P.G. Slade (Marcel Dekker, New York, 1999b), pp. 309–402

B.H. Chudnovsky, Degradation of power contacts in industrial atmosphere: silver corrosion and whiskers, in *Proceedings of the 48th IEEE Holm Conference on Electrical Contacts*, Orlando, pp. 140–147 (2002)

B.H. Chudnovsky, Degradation of power contacts in industrial atmosphere: plating alternatives to silver and tin, in *Proceedings of the 49th IEEE Holm Conference on Electrical Contacts*, Washington, DC, pp. 98–106 (2003)

J. Haimovich, Intermetallic compound growth in tin and tin-lead platings over nickel and its effects on solderability. Welding J., Welding Res. **68**(3), 102–111 (1989)

H.W. Hermance, T.F. Egan, Organic deposits on precious metal contacts. Bell Syst. Tech. J. **37**, 739–766 (1958)

T. Imrell, The importance of the thickness of silver coating in the corrosion behaviour of copper contacts, in *Proceedings of the 37th IEEE Holm Conf. on Electrical Contacts*, Chicago, pp. 237–243 (1991)

S.J. Krumbein, Metallic electromigration phenomena, in *Electromigration and Electronic Device Degradation*, ed. by A. Christou (Wiley, New York, 1994), pp. 139–166

F.I. Nobel, Electroplated palladium-silver (60/40 wt%) alloy as a contact material, in *Proceedings of the 30th Holm Conference on Electrical Contacts*, Chicago, pp. 137–152 (1984)

J.W. Osenbach, R.L. Shook, B.T. Vaccaro, B.D. Potteiger, A.N. Amin, K.N. Hooghan, P. Suratkar, P. Ruengsinsub, Sn whiskers: material, design, processing, and post-plate reflow effects and development of an overall phenomenological theory. IEEE Trans. Electron. Packag. Manuf. **28**, 36–62 (2005)

S.M. Schwarz, B.W. Kempshall, L.A. Effects of diffusion induced recrystallization on volume diffusion in the copper-nickel system. Giannuzzi, Acta Materialia. **51**, 2765–2776 (2003)

RoHS, *Directive of the European Parliament on the Restriction of the Use of Certain Hazardous Substances*. http://ec.europa.eu/environment/waste/rohs_eee/legis_en.htm (2012)

M. Warwick, Implementing lead free soldering-European consortium research. J. Surf. Mt. Technol. **12**, 1–12 (1999)

Electrical Contacts: Scientific Fundamentals

Roland S. Timsit
Timron Advanced Connector Technologies, A Subsidiary of Timron Scientific Consulting Inc., Toronto, ON, Canada

Definition

An electrical contact is a well-defined interface across which an electrical current flows between two conductors.

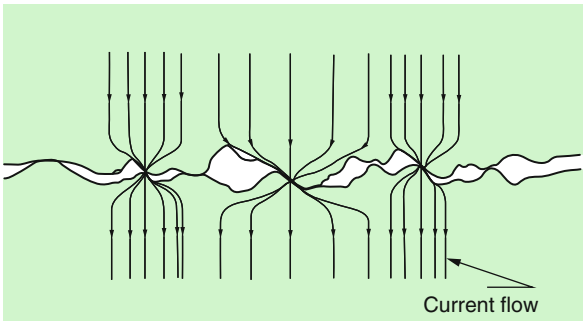
The presence of this interface leads to an electrical "contact resistance." If this resistance is too large, it leads to intolerable Joule heat losses and contact failure.

Scientific Fundamentals

Electrical Contact Resistance

All surfaces are rough on the microscale. This roughness may be characterized in terms of peaks, valleys, plateaus, and so on, in the same way as the landscape of a countryside would be described. An interface between two solids is generated by contact between the protruding surface asperities on each of the contacting bodies. For all solid materials, the area A of true contact is thus a small fraction of the nominal contact area and is given as

$$A = \frac{F}{H} \tag{1}$$



Electrical Contacts: Scientific Fundamentals, Fig. 1
Schematic diagram of contact *a*-spots and current flow in an electrical contact

where F is the contact force and H is the Vickers' microhardness. Since electrical current passes only where the small contact spots (also known as *a*-spots) are electrically conducting (i.e., where electrically insulating films on the contact surfaces are displaced), electrical current is highly constricted as it passes across the interface, as illustrated in Fig. 1. Current constriction gives rise to a *contact resistance* very much like constriction of a water hose increases resistance to water flow. For a circular constriction of radius *a*, the constriction resistance R_C is given as

$$R_C = \frac{\rho}{2a} \tag{2}$$

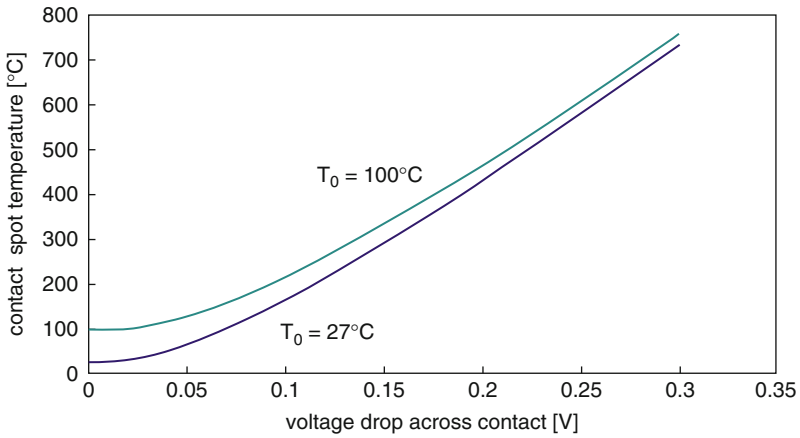
The contact resistance R_C between two conductors of resistivity ρ_1 and ρ_2, held together with a force F, is given as (Holm 1976; Fechant 1996; Timsit 1999)

$$R_C = \frac{(\rho_1 + \rho_2)}{4}\sqrt{\frac{\pi H}{F}} \tag{3}$$

where H is the Vickers' microhardness of the softer of the two materials and F is the contact force.

Contact Spot Temperature

The presence of a finite number of *a*-spots in an interface implies that new contact spots must somehow be formed as others are lost in the presence of wear mechanisms, if the electrical integrity of the interface is to be maintained. Wear mechanisms act to reduce the number of contact spots over time, thus forcing the electrical current through an ever-decreasing number of *a*-spots and leading to a larger contact resistance. In separable and stationary interfaces, Joule



Electrical Contacts: Scientific Fundamentals, Fig. 2 Dependence of *a*-spot temperature on voltage drop across a contact

Electrical Contacts: Scientific Fundamentals, Table 1
Voltage for softening and melting of common electrical conductor materials

Material	Softening voltage (V)	Melting voltage (V)
Al	0.1	0.3
Fe	0.19	0.19
Ni	0.16	0.16
Cu	0.12	0.43
Zn	0.1	0.17
Mo	0.25	0.75
Ag	0.09	0.37
Cd		0.15
In		0.11
Sn	0.07	0.13
Au	0.08	0.43
W	0.4	1.1
Pt	0.25	0.65
Pd		0.57
Pb		0.12
Cu60Zn40		0.2
Cu60Sn40		0.15
stainless steel	0.27	0.55
WC, Co	0.6	

losses stemming from increased contact resistance are largely responsible for the increase in temperature since the contribution of frictional heating is negligible. Significant Joule heating of *a*-spots can have a deleterious effect on the integrity of an electrical contact due to the activation of thermally activated degradation mechanisms such as oxidation and corrosion. High temperatures also reduce the mechanical strength of the mating materials and thus increase mechanical wear due to materials softening.

Because the Joule heat generated in *a*-spots is dissipated largely by thermal conduction to the mating bodies, it turns out that the maximum temperature in *a*-spots is determined by the voltage-drop across the contact interface. Under steady-state conditions, the maximum temperature T_M (in Kelvin) in a contact spot is given as (Holm 1976; Timsit 1999)

$$T_M = \left[T_0^2 + \frac{V^2}{4L}\right]^{1/2} \quad (4)$$

where V is the potential drop across the contact interface, T_0 is the bulk temperature (K) of the contacting bodies far from the interface, and L is the Lorenz constant equal to $2.45 \times 10^{-8}\ V^2/K^2$. As is clear from (4), the evaluation of T_M is independent of the contact material.

Figure 2 shows the relationship of T_M to V for values of the bulk temperature T_0 of 27°C and 100°C. A voltage-drop of several tens of millivolts across a contact leads to *a*-spot temperatures of a few hundred degrees Celsius. The "softening" and "melting" temperatures of various contact materials are listed in Table 1. The temperature-voltage relation is crucial to the understanding of several tribological properties of electrical contacts.

Cross-References

▶ Sliding Electrical Contact Wear
▶ Tribology

References

L. Fechant, *Le Contact Electrique, Phenomenes Physiques et Materiaux* (Hermes, Paris, 1996)

R. Holm, *Electric Contacts, Theory and Applications* (Springer, Berlin, 1976)

R.S. Timsit, Electrical contact resistance: fundamental principles, in *Electric Contacts: Theory and Applications*, ed. by P.G. Slade (Marcel Dekker, New York, 1999), pp. 1–88

Electrical Discharge Machining (EDM)

▶ Gear Manufacturing Machines

Electro- and Electroless Composite Coatings

Sankara Narayanan T.S.N.[1], Seshadri S.K.[2]
[1]National Metallurgical Laboratory, Madras Centre, Chennai, India
[2]Department of Metallurgical and Materials Engineering, Indian Institute of Technology Madras, Chennai, India

Definition

Electro- and electroless composite coatings offer a cost-effective and efficient way to engineer the surface so as to achieve desirable qualities, such as hardness, wear, and abrasion and corrosion resistance (Kerr et al. 2000; Low et al. 2006; Balaraju et al. 2003). These coatings are obtained by codepositing various second phase particles in

electro- or electroless deposited metal or alloy matrix. Almost any particle that can be held in suspension without reacting with the plating bath can be codeposited. Particles used for codeposition with the metal matrix range from hard particles like oxides of Al, Ce, Si, Ti, Th, and Zr, carbides of Si, B, Ti, W, and Cr, nitrides of Si and B and borides of Ti and Zr, and synthetic and natural diamond to soft particles like, graphite, CaF_2, MoS_2, polytetrafluoroethylene (PTFE), WS_2, and advanced materials like carbon nanotube (CNT) and inorganic fullerene.

Scientific Fundamentals

The idea of codepositing various second phase particles in electrodeposited (ED) or electroless deposited (EL) metal or alloy matrix and thereby taking advantage of their desirable qualities, such as hardness, wear and abrasion resistance, corrosion resistance, and so on, has led to the development of composite coatings with a wide range of possible combinations and properties (Kerr et al. 2000; Low et al. 2006; Balaraju et al. 2003). An essential advantage of preparing composite coatings by EL compared with ED is that the former allows accurate reproduction of the base geometry and eliminates the need for subsequent mechanical finishing.

Mechanism of Codeposition of Particles in ED and EL Metal Matrix

The mechanism of co-deposition of particles in ED metal matrix is not yet thoroughly understood. Electrophoresis, convective diffusion, mechanical entrapment, and adsorption are the four commonly discussed mechanisms. Among the various theoretical models put forth to describe the codeposition of particles in ED metal matrix, the models proposed by Guglielmi (1972) and Celis et al. (1987) are significant. Guglielmi's model is based on a two-step adsorption process. The first step involves loose physical adsorption of the particles on the cathode surface with a high degree of coverage but without discharge of the electro-active ions adsorbed on the particles. The fractional coverage follows a Langmuir adsorption isotherm. The second step involves strong electrochemical adsorption of the particles by the applied electrochemical field that is accompanied by the discharge of the electroactive ions. Both steps would take place at the same time all over the surface of the cathode. If a particle is strongly adsorbed on the cathode, then it will be embedded in the growing metal layer. However, Guglielmi's model fails to consider hydrodynamics, particle size, and aging effects, which led to the development of other models. The model proposed by Celis et al. (1987), which takes into account for hydrodynamic conditions, is based on two fundamental postulates:

1. An adsorbed layer of ionic species is created around the inert particles when they are added to the plating bath or during their pretreatment in ionic solutions.
2. The reduction of some of these adsorbed ionic species is required for the incorporation of particles in the metallic matrix.

According to this model, on its way from the bulk of the solution to the site of incorporation at the active cathode surface, the inert particle has to proceed through the following five stages (Fig. 1):

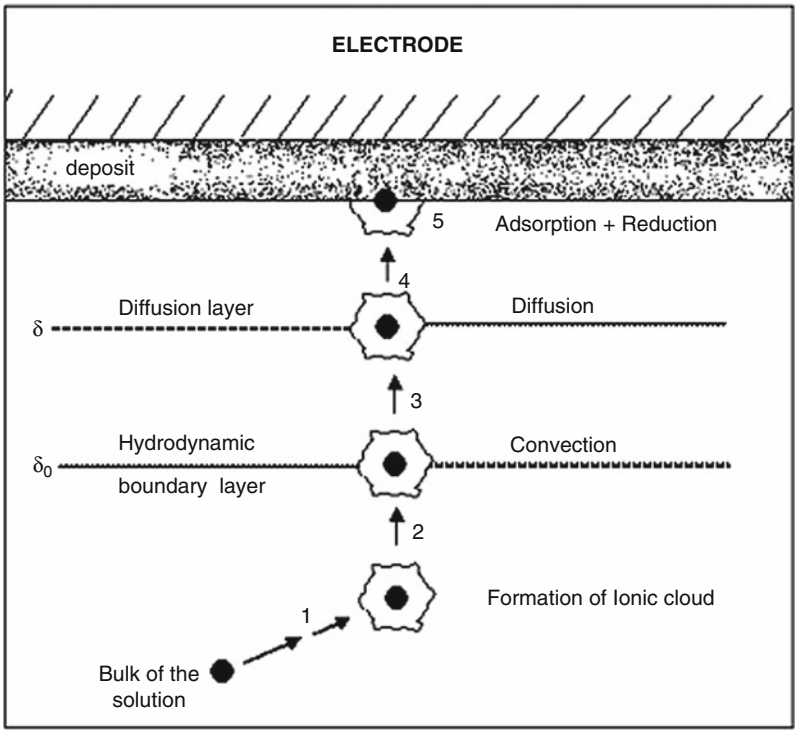
1. Adsorption of ionic species upon the particle surface
2. Movement of the particle by forced convection towards the hydrodynamic boundary layer at the cathode
3. Diffusion of the particle through the diffusion double layer
4. Adsorption of the particle with its adsorbed ionic cloud, at the cathode surface
5. Reduction of some adsorbed ionic species by which the particle becomes irreversibly incorporated in the metal matrix

The various theoretical models used to describe the behavior of metal electrodeposition containing particles along with their characteristics and assumptions are described elsewhere (Low et al. 2006).

EL composite coatings are formed by the impingement of particles on the substrate and subsequent envelopment of these particles by the matrix metal as it is deposited. There is no molecular bonding between particles and metal matrix. The mechanism of particle incorporation in EL metal matrix follows Guglielmi's model. Incorporation of particles in EL metal matrix depends on two factors: (1) impingement and (2) residence time of the particles on the electrode surface. Impingement of particles on the electrode is determined by the flux of particles at the interface, which is a function of their concentration, size, and density. The residence time is a function of mode of agitation and speed, particle shape, and extent of hydrogen evolution (Balaraju et al. 2003).

Factors Influencing Particle Incorporation in Composite Coatings

Incorporation of particles in ED and EL metal matrix is influenced by several factors (Kerr et al. 2000; Low et al. 2006; Balaraju et al. 2003), of which the major ones are discussed below.



Electro- and Electroless Composite Coatings, Fig. 1 Pictorial representation of the five-stage model proposed to explain the incorporation of second-phase particles in electrodeposited metal matrix (Reproduced from Celis et al. (1987) with permission of The Electrochemical Society)

Size and Shape of the Particles

The mass/charge ratio of the particles is the key factor in determining the codeposition of particles in ED metal matrix. Increase in size of the particles beyond 10 μm leads to an increase in the mass/charge ratio, which causes a reduction in adsorption and electrophoretic movement of the particles in the electrical double layer at the cathode, thus reducing the chances for codeposition. In general, the particles must be large and heavy enough to settle in the solution yet not so large as to make the deposit rough or make it difficult for them to be held in suspension. Also, the size of the particles should be selected with reference to the thickness of the ED or EL metal matrix. Though particles in the size range of 2–10 μm are suitable for codeposition in ED and EL metal matrix, those having the size range of 4–7 μm are the easiest to work with. Smaller particles with a narrow size distribution are favored since they yield the maximum level of incorporation and they can be firmly held in the matrix and enable a better integrity between particles and metal matrix. Angular-shaped particles will have a greater tendency to hold on to the surface than the round ones, but results in a rough surface finish.

Composition of the Plating Bath

Codeposition of Al_2O_3, SiC, and so on is less from an acid copper sulfate bath than from alkaline copper cyanide and fluoborate baths, in spite of these baths being operated close to 100% efficiency. The difference is mainly due to the adsorption of cations on the particle surface. Since cation adsorption on particles in acid copper sulfate bath is small, the attractive force for the cathode surface is not strong enough to cause a higher level of incorporation. Similarly, for codeposition of insoluble inorganic particles, Watt's nickel bath (high sulfate bath) is found to be more effective than the chloride and fluoborate baths at the same temperature and pH.

Concentration of Particles in the Bath

Bath loadings of 5–200 g/l are quite common in ED, whereas the concentration of particles used in EL is relatively lower. In ED, incorporation of particles in the metal

matrix varies either linearly or logarithmically. Under steady state condition, the number of codeposited particles equals that approaching the cathode surface, thus exhibiting a linear relationship. However, beyond a critical concentration, collision among the moving particles inhibits the entrapment, which limits the level of incorporation. In EL, particle incorporation in the metal matrix exhibits a logarithmic relationship. In some instances, a slight decrease in the level of incorporation is also observed. The critical concentration at which particles exhibit saturation in incorporation is not very different for various hard and soft particles in EL, whereas in ED, it varies with type of particles. Compared with ED composite coatings, incorporation of particles for a given concentration is considerably higher for EL composite coatings. Moreover, to obtain a particular level of incorporation, a greater amount of particles in the bath is required in the case of ED than EL deposition.

Method Employed for Deposition

Particle incorporation in ED, irrespective of whether it is micron or nano-sized, is higher for pulsed current (PC) than for direct current (DC). In this, the extent of incorporation is higher at a duty cycle of 10% and it decreases with further increase in duty cycle. Hypophosphite-reduced EL nickel plating baths enable a reasonable level of incorporation of particles, whereas the excessive hydrogen evolution limits particle incorporation to less than 2 wt% in borohydride-reduced EL plating baths.

Current Density

The range of current density suitable for codeposition changes from metal to metal. ED Cu, Ni, Fe, and Co composite coatings can be prepared at 0.1–1 A/dm^2 whereas a higher current density of 10–60 A/dm^2 is required for preparing ED Cr composite coatings. The variation in the level of incorporation of particles as a function of current density can be classified in to three types:

- A linear increase in incorporation with current density; e.g., Ni-Al_2O_3, Ni-TiO_2 and Ni-SiC
- A linear decrease in incorporation with current density; e.g., Ni-Cr_3C_2, Ni-flyash, Ni-CeO_2, Ni-graphite, and Ni-Al_2O_3
- An increase followed by decrease with current density, the maximum being at a critical current density; e.g., Ni-γAl_2O_3

Among the three categories, the latter one is followed by most of the systems. At low current densities (below 0.4 A/dm^2), incorporation of particles generally increase with increase in current density for most of the systems. However, after a critical current density, which is specific for individual bath, the extent of incorporation decreases with further increase in current density, irrespective of the temperature and pH of the bath.

Agitation

Various methods of agitation, which include mechanical agitation, circulation by pumping, purging of air, oxygen, nitrogen, ultrasonic agitation, and the plate-pumping technique have been employed to obtain a uniform dispersion of particles in the plating bath. In general, if the agitation is too slow (laminar flow), the particles in the bath may not disperse completely, except when their density is low. On the other hand, if the agitation is too high (turbulent), particles will not have sufficient time to become attached to the surface, and this results in poor particle incorporation. Laminar-turbulent transition region is considered as the most effective agitation condition for achieving maximum incorporation of particles. Ultrasonic vibration assumes significance in the co-deposition of nano-sized particles and nano whiskers. It enhances mass transport, electrode cleaning and emulsification, through the generation of cavitation bubbles and its subsequent destruction, which enables a uniform dispersion of nano-sized particles.

Additives

Organic compounds such as, thiourea, ethylenediamine, and so on, and, cations such as Tl^+ and Cs^+ promote the co-deposition of particles in ED metal matrix. Surfactants are especially important in the incorporation of soft particles like PTFE, graphite, MoS_2, and carbon nanotubes. Addition of surfactants such as sodium lauryl sulfate and Forafac-500 becomes mandatory to avoid agglomeration and to achieve an effective dispersion of the nano-sized particles in the plating bath as well as in the metal matrix. However, beyond a critical concentration, surfactants tend to promote agglomeration of particles, act as a barrier for codeposition, and impart a brittle nature to the metal matrix. When particles are added to electroless plating bath, the surface area loading is increased about 800 times, which warrants addition of stabilizers for successful operation.

Incorporation of micron- Versus Nano-Sized Particles

Both micron- and nano-sized particles could be codeposited with the metal matrix. However, their codeposition mechanism is different. The micron-sized particles are codeposited at the borders and the edges of

the metal crystallites, while the nano-sized particles, besides the borders and edges, are also incorporated inside the nickel crystals. Hence, the embedding mechanism of the nano-sized and micron-sized particles could be characterized as "intra-crystalline" and "inter-crystalline," respectively. The change in embedding mechanism also reflects in the level of incorporation of these particles. The extent of incorporation of micron-sized particles is usually expressed in terms of either weight or volume percentage. In case of sub-micron and nano-sized particles, the number density of the particles in the metal matrix is considered as most appropriate to express the extent of incorporation. The ratio of number density of codeposited particles to the number density of particles in the plating bath is defined as the codeposition efficiency. In general, the codeposition efficiency increases substantially with increase in particle size. Hence, the size and number density of particles in the plating bath are important parameters in the codeposition process (Garcia et al. 2001)

Effective dispersion of particles in the plating bath is a major problem in co-deposition of nano-sized particles because of the strong tendency of these particles towards agglomeration and sedimentation. When the particles are dispersed in the plating bath, the surface charge on the particles is changed from positive to low negative values, which will have a detrimental effect on their dispersion stability. Flocculation and destabilization of particles with increase in bath temperature also becomes a major limitation.

Nucleation, Surface Finish and Morphology, Structure and Texture of Composite Coatings

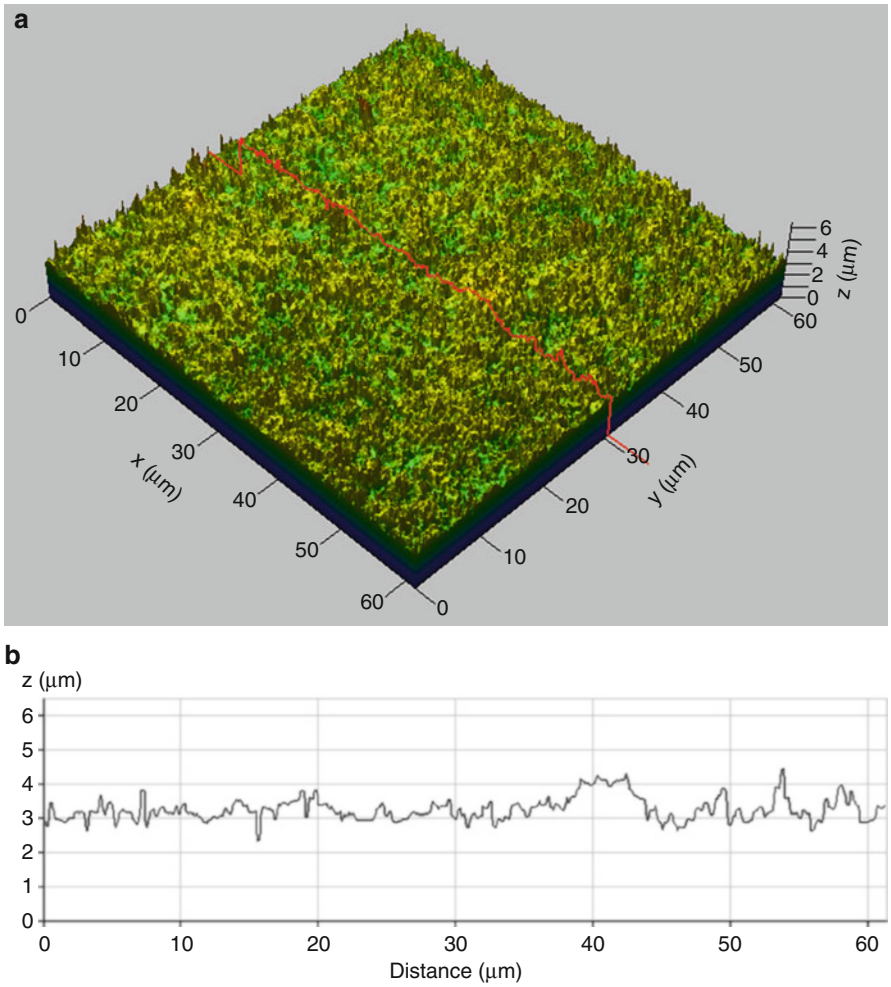
Influence of particles on the nucleation of the metal crystallites during ED assumes significance since adsorption of these particles might either increase the electroactive electrode area or they might cause a blockage of the electrode surface. In general, the dispersed particles do not modify the electrocrystallization mechanism, but they take part in the process by increasing the metal deposition rate. The particles act as a catalyst for the reduction of metal ions, leading to an increase in the number of active nucleation sites (Benea 2009). However, the ability of the particles to act as a catalyst and to promote nucleation depends on the type of particles as well as the plating bath. Particles such as ZrO_2 and CNT promote the nucleation of Ni (Benea 2009); SiC promotes the nucleation of both Zn and Co, whereas Al_2O_3 promote nucleation of Zn but exhibit a negative effect on nucleation of Co (Tulio and Carlos 2009).

Incorporation of particles in ED and EL metal matrix alters the surface finish both in terms of decreasing the luster and increasing the surface roughness. The extent of these changes is a function of the type of particle, particle size, number density of the particles, and thickness of the coating. The surface topography and surface profile of ED Ni-B-Si_3N_4 composite coatings is shown in Fig. 2.

ED Ni coatings have a rather regular surface, whereas composite coatings, in general, develop a rough morphology, constituted by nodular agglomerates or disturbed surface structure. This characteristic of composite coatings could be due to the significant nucleation rate of new crystals on the particle surface before it is covered by the growing Ni layer. The morphological features of composite coatings are also influenced by the type of ED methods (DC, PC, and PRC). Compared with those prepared by DC ED, composite coatings prepared by PC ED possess a smooth and compact surface morphology. In contrast, composite coatings obtained by PRC ED exhibit a coarse surface morphology. The application of a very high instantaneous current density during PC ED enables an increase in nucleation rate that leads to the formation of a finer grain size and more homogenous surface appearance. In PRC ED, though numerous nuclei of nickel could form during the cathodic process, the smaller and less stable crystallites with high energy of formation would dissolve instantaneously during the next anodic process while the larger crystallites continued to grow. As a result, the PRC ED could help to dissolve the smaller crystallites and preserve the larger. Moreover, the adsorption of $Ni(OH)_2$ crystals and/or H_2 molecules on the surface also promotes the formation of a larger micro-structure. All of this results in the coarsened surface morphology of the composite coating obtained by PRC ED (Wang et al. 2005). With the incorporation of second phase particles, the surface of EL Ni-P and Ni-B matrix is changed from bright and smooth to foggy and rough, with nodular protrusions covering the entire surface.

Incorporation of TiC, Si_3N_4, CeO_2, and TiO_2 particles in EL Ni-P matrix and Si_3N_4, particles in EL Ni-B matrix does not alter the structure (Balaraju et al. 2003). However, incorporation of B_4C particles is found to affect the orientation of nickel crystallites without influencing the crystallite dimensions; nickel tends to be less oriented in layers with B_4C particles. Incorporation of SiC particles helps to increase nucleation centers, degree of crystallization, and microstructural stability and prevents grain growth and aggregation of the matrix.

Incorporation of particles such as TiO_2, ZrO_2, SiC, Nb, and WC in ED Ni matrix affect the preferred orientation Ni. ED Ni coatings possess a preferred orientation along the (2 0 0) crystal plane. The preferred orientation of the ED Ni grains in the composite coatings does not depend



Electro- and Electroless Composite Coatings, Fig. 2 Surface topography (**a**) and surface profile (**b**) of Ni-B-Si_3N_4 composite coating electrodeposited at 1 A/dm^2 assessed using a laser scanning microscope (Reprinted from Krishnaveni et al. (2008), with permission from Elsevier)

on cathodic current density, stirring rate, or concentration of particles in the bath. In contrast, the level of incorporation and dispersion (agglomerated or mono-dispersed) of the particles in the ED Ni matrix and method of deposition (DC or PC deposition) seem to influence the preferred orientation to a larger extent. The physico-chemical interaction of particles when they approach the catholytic area is considered responsible for the change in texture of the composite coatings. The H^+ adsorption–desorption phenomena that occurs on the surface of these particles (depending on the pH of the plating bath) inhibits the reactivity of specific chemical species, which impose certain modes of nickel crystal growth. Incorporation of ZrO_2 and SiC in ED Ni matrix exhibits a definite change in preferred orientation. In contrast, Nb and WC do not show any preferential orientation, but rather a random orientation. This is probably due to the nucleation of randomly oriented new Ni crystals on the embedded Nb and WC particles.

Hardness of Composite Coatings

The hardness of ED and EL composite coatings increases with incorporation of hard particles, whereas with soft particles, the hardness tends to decrease. The level of incorporation of particles, distribution of the particles in the metal matrix, the alloying elements of the metal matrix, and heat-treatment determines the hardness of these coatings. The hardness of EL Ni-P composite

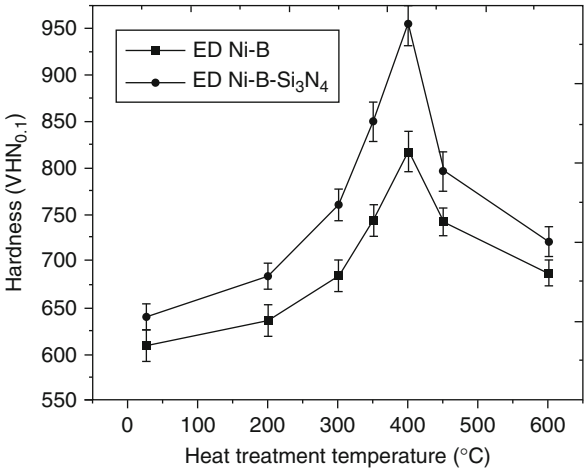
Electro- and Electroless Composite Coatings, Table 1 Hardness of EL Ni-P composite coatings (Data adapted from Balaraju et al. 2003)

Type of EL Ni-P coating	Phosphorus content (wt.%)	Particle content (wt%/vol%)	Hardness ($HV_{0.1}$)	
			As plated	Heat-treated[a]
Ni-P	8.00–9.10	–	410–600	979–1,136
Ni-P-nano diamond	7.60	0.52 wt%	470	939
Ni-P-nano diamond	6.27	2.21 wt%	755	966
Ni-P-SiC (irregular)	8.22	19.60 vol%	705	1,143
Ni-P-Al_2O_3(irregular)	8.22	9.70 vol%	643	1,139
Ni-P-Al_2O_3(spherical)	8.22	28.60 vol%	743	1,248
Ni-P-Al_2O_3(fibres)	8.22	10.70 vol%	640	1,147
Ni-P-Cr_2C_3	7.20	27.00 vol%	645	1195[b]
Ni-P-PTFE	9.5–10.0	25.00 vol%	275[c]	450[c,d]
Ni-P-hexagonal BN	5.5	33.00 vol%	486[e]	753
Ni-P-Si_3N_4	10.10	8.01 wt%	720	1,171
Ni-P-CeO_2	10.18	7.44 wt%	676	1,136
Ni-P-TiO_2	10.40	5.42 wt%	642	1,104
Ni-P-CNT	> 7.00	12.0 vol%	520	1,035

[a]Heat-treated at 400°C/1 h unless otherwise indicated
[b]500°C/12 h
[c]Load 50 g
[d]350°C/2 h
[e]Knoop hardness

coatings codeposited with various hard and soft particles is given in Table 1. The increase in hardness following the incorporation of hard particles such as, Si_3N_4, SiC, and so on could be due to a combination of grain refinement (Hall–Petch strengthening) and dispersion strengthening (Orowan strengthening) mechanisms. The ability of the hard particles to restrain the growth of the grains of the metal matrix and to impede the dislocation movement under loading depends on the inter-particle spacing in the matrix (Garcia et al. 2001). For optimum performance, it is preferable to have an inter-particle spacing between 0.5 and 5 μm. In such instances, the metal matrix carries the load and the particles impede the motion of dislocation, thus enabling a high hardness. The increase in hardness of composite coatings incorporated with micron-sized particles is mainly due to the particle strengthening effect. However, for coatings containing sub-micron particles, where the inter-particle spacing is less than 5 μm, increase in hardness is mainly due to the dispersion strengthening effect. In case of incorporation of nano-sized hard particles, a uniform distribution with very high number density of these particles offers a significant improvement in hardness. The contribution to the increase in hardness of the composite coatings by grain refinement strengthening mechanism is similar for both agglomerated and evenly distributed hard particles in the metal matrix whereas the contribution from dispersion strengthening mechanism is largely due to the uniformly distributed hard particles in the metal matrix. This indicates the importance of achieving a uniform distribution of hard particles to impart a better hardness of the resultant coating. Hence, it is evident that the hardness of composite coatings incorporated with hard particles depends not only on the amount of the particles co-deposited but more importantly on their size and distribution in the metal matrix.

Heat-treatment of ED Ni composite coatings in inert atmosphere or vacuum generally reduces the microhardness. However, the extent of reduction in hardness with annealing temperature for composites is much less when compared with ED pure Ni coatings under similar conditions. ED and EL Ni-P and Ni-B composite coatings incorporated with hard particles exhibit a behavior that is different from pure Ni coatings on heat-treatment. There is a marginal increase in hardness of these coatings when the heat-treatment temperature is increased to 200°C. With further increase in temperature (around 350–450°C depending on the type of



Electro- and Electroless Composite Coatings, Fig. 3 Variation in hardness of ED Ni-B and Ni-B-Si_3N_4 coatings as a function of heat-treatment temperature (With kind permission from Springer Science + Business Media, Krishnaveni et al. (2009), Fig. 4)

matrix), however, the hardness increases rapidly, as the structure of the coating matrix begins to change. The precipitation of hard nickel phosphides and borides, primarily Ni_3P phase in Ni-P coatings and Ni_3B and Ni_2B phases in Ni-B coatings, are considered responsible for the increase in the hardness. When the heat-treatment temperature exceeds beyond a point (~350–450°C depending on the type of matrix), the hardness of these coatings starts to decrease due to the decrease in lattice defects and softening of the matrix following coarsening of the nickel phosphide and boride phases, which reduces the number of hardening sites. The variation in hardness of ED Ni-B and Ni-B-Si_3N_4 composite coatings as a function of heat-treatment temperature is shown in Fig. 3. The volume fraction of the phosphide and boride phases formed during heat-treatment also influences the hardness of the composite coatings. The hardening effect becomes more predominant with an increase in the volume fraction of phosphides and boride phases of the matrix. Thus, the phosphorus and boron content of ED and EL Ni-P and Ni-B composite coatings is a major factor in influencing the hardness of the coating after heat-treatment and the effect of particle is only a supporting factor.

Friction and Wear Behavior of Composite Coatings

ED and EL composite coatings offer better wear resistance compared with their plain counterparts. Wear behavior of these composite coatings is largely a function of the method of deposition, the type and size of particles, the number density of particles in the metal matrix, chemical, microstructural, and textural modifications, and hardness of the metal matrix induced by the particles codeposited in the metal matrix (Garcia et al. 2001). The improvement in wear resistance offered by composite coatings incorporated with hard particles such as SiC, Al_2O_3, Si_3N_4, and so on, when compared with their plain counterparts, is due to (1) the characteristics of these hard particles, such as high hardness and high resistance to plastic deformation; and (2) the dispersion of these hard particles in the metal matrix, which increases the hardness. The uniformly distributed hard particles act as supporting points, thus strengthening the metal matrix and, reduce the direct contact between the metal matrix and the counter disc during sliding which in turn reduce the extent of plastic deformation and adhesive wear, consequently enabling an improvement in wear resistance. To achieve a better wear resistance with composite coatings, it is necessary to ensure that (1) the number density of codeposited particles should be high to realize particle strengthening effect; (2) the particles are uniformly distributed in the matrix without agglomeration so as to attain stronger bonding between the particles and the metal matrix, the benefit arising from dispersion strengthening effect of the particles and a better load bearing capacity; (3) the ratio of coating thickness to the size of the particles is optimized to achieve a better integrity between the particles and the metal matrix so that the particles are not pulled-out of the matrix and induce abrasive wear; and (4) heat-treatment imparts a higher hardness and presents an incompatible surface to the counterface material during wear.

The wear behavior of ED pure Ni, Ni-P, Ni-B and EL Ni-P, Ni-B coatings and their composite coatings have received considerable attention due to their widespread use in engineering applications. ED pure Ni, Ni-P, Ni-B coatings, in their as-plated condition, exhibit cracking, spalling, and severe delamination of the coating along the sliding direction during sliding against a hardened steel disc, which suggests a typical adhesive wear mechanism. The occurrence of severe plastic deformation and delamination indicates that ED Ni coating is rather weak and has a poor load-bearing capacity. The wear behavior of ED Ni, Ni-P, and Ni-B composite coatings is also governed by the adhesive wear mechanism. However, the extent of plastic deformation and adhesive wear is considerably reduced in these coatings.

Adhesive and abrasive wear are the most frequently encountered wear mechanisms in EL Ni-P and Ni-B coatings, in their as-plated condition. Incorporation of

particles in these coatings significantly reduces the extent of adhesive wear. On the other hand, the abrasive wear resistance is largely determined by the hardness of the particle, its level of incorporation and dispersion in the matrix, and its detachment from the matrix during wear. Hard particles such as WC and diamond cause pronounced abrasion of the counterface materials. This is due to the protrusion of these particles, which increases the number of supporting points and causes an increased wearability. Detachment of agglomerated nano-Si_3N_4 particulates from ED Ni-Co matrix enhances the abrasive wear and scuffing.

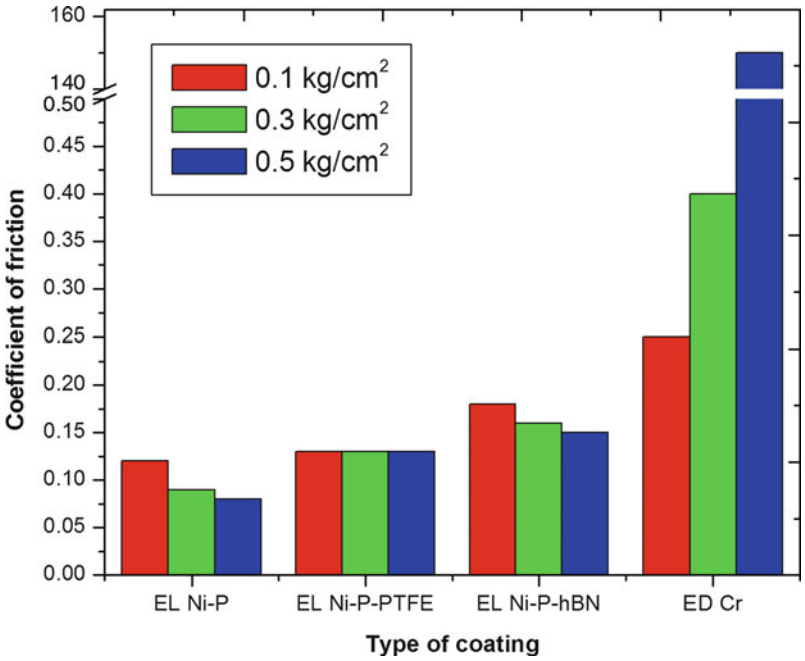
Heat-treatment of ED as well as EL Ni-P composite coatings at 350°C and Ni-B coatings at 300°C and 450°C induces the formation of hard nickel phosphide and boride phases, respectively, which imparts a double strengthening effect: dispersion strengthening of the particles and precipitation strengthening of the metal matrix. Hence, the high hardness of the composite coating after heat-treatment enables a decrease in wear rate. Besides, the formation of hard phosphide and boride phases presents a virtually incompatible surface to the counterface material (hardened steel) since there could be very little solubility between iron of the counterface material and these hard phases that would lead to a better wear resistance. ED and EL Ni-P and Ni-B composite coatings, in their heat-treated conditions, exhibit a bright and smooth finish with fine grooves along the sliding direction with no gross adhesion between the mating surfaces.

Heat-treatment of ED and EL Ni-P and Ni-B composite coatings could, however, have a deleterious effect on the wear resistance under certain conditions. If the heat-treatment temperature exceeds beyond a level, these coatings exhibit a mixed adhesive and fatigue wear mechanism, accompanied by a large plastic deformation and a high coating transfer to the counterface material. This indicates that the tribological response of composite coatings at high temperatures is related to the mechanical properties of the matrix and the ability of the matrix to withstand the applied stresses with increase in temperature. The interaction between certain particles and the metal matrix, for example, the reaction between SiC and Ni-P matrix, which leads to the formation of nickel silicide at 580°C, could also reduce the wear resistance of the composite coatings at elevated temperatures.

The friction coefficient of ED and EL composite coatings incorporated with hard particles, in their as-plated condition, is relatively high compared with their plain counterparts due to their high surface roughness and high mechanical interlocking forces. For a given size of particle, coefficient of friction increases with increase in level of incorporation of particles (Garcia et al. 2001). The coefficient of friction exhibits an abrupt increase during the initial period in case of composite coatings incorporated with hard particles, which is attributed to the high hardness and abrasive properties of the particles. However, it gets stabilized after some sliding distance following the removal of coating and the onset of a friction process between the interacting surfaces, perhaps with some coating particles entrapped in the contact. The formation of a tribochemically assisted layer, for example, the hydroxylated silicon oxide layer formed by the tribochemical reaction between nano-Si_3N_4 and water vapor in the environment (relative humidity of 52–56%), could act as a solid lubricant and decrease the coefficient of friction and wear rate of the mating surfaces. Heat-treatment reduces the coefficient of friction of ED and EL Ni-P and Ni-B composite coatings. Formation of hard phosphide and boride phases as well as poor solubility between iron of the counterface material and these phases presents a virtually incompatible surface during wear. In addition, the build-up of iron oxides at the interface could provide a lubricant film and enables a decrease in the coefficient of friction in these coatings.

For achieving better lubricity, incorporation of PTFE and hexagonal BN particles is considered as the most suitable choice (Ploof 2008). ED and EL composite coatings incorporated with PTFE particles (usually about 19–25 vol% PTFE) is self-lubricating and offers a non-stick and non-galling surface. During dry friction, the PTFE particles act as a solid lubricant and are easily sheared and transferred to the mating surface, thus changing the test system, which contributes to excellent wear resistance and low coefficient of friction. Since PTFE becomes softened at 325°C, ED and EL PTFE composite coatings are best suited for lower-temperature, low-load applications. ED and EL composite coating with 6–8 wt% hexagonal BN particles offers a lower coefficient of friction under higher loads than their PTFE counterparts. Since hexagonal BN can withstand temperatures up to 3,000°C, these coatings can be utilized for high temperature applications (up to the solidus point of the corresponding metal matrix) (Ploof 2008). A comparison of the coefficient of friction of EL Ni-P-PTFE and EL Ni-P-hBN composite coatings with EL Ni-P and ED Cr coatings is shown in Fig. 4.

ED and EL composite coatings containing hard particles offer a superior wear resistance, but their coefficient friction is very high. In contrast, composite coatings with soft particles offer better lubricating properties even though the decrease in hardness of the matrix limits their ability to offer better wear resistance. As a result, hybrid composite coatings containing both hard and soft particles have been developed. EL Ni-P-PTFE-SiC



Electro- and Electroless Composite Coatings, Fig. 4 Comparison of the coefficient of friction of EL Ni-P-PTFE and Ni-B-hBN composite coatings with EL Ni-P and ED Cr coatings (Data adapted from Ploof 2008)

composite coating demonstrates a combination of the advantages of EL Ni-P-SiC (high hardness and wear resistance) and that of EL Ni-P-PTFE coating (a low friction coefficient and lower surface energy) (Huang et al. 2003). ED and EL Ni-P-CNT are yet another interesting class of composite coatings that offer a combination of high wear resistance and low friction coefficient compared with Ni–P–SiC and Ni–P–graphite coatings. The self-lubrication property and the unique topological structure of CNTs enable a significant improvement in tribological property of these coatings. Reinforcement of CNTs in the metal matrix transforms the wear mechanism from fracture to deformation.

Fretting Wear Behavior of Composite Coatings

Incorporation of nano-sized Al_2O_3, ZrO_2, SiO_2, and cubic BN particles in nickel matrix offers better fretting wear resistance compared with their plain counterparts. The improvement in fretting wear resistance of Ni-Al_2O_3, Ni-ZrO_2 and Ni-SiO_2 composite coatings is due to the grain refinement strengthening, dispersion strengthening, and increase in dislocation following incorporation of these particles in the nickel matrix while the self-lubricating property of cubic BN becomes the dominant factor in Ni-cubic BN composite coating. An increase in fretting frequency leads to failure of plain Ni coatings at lower fretting cycles, whereas under similar conditions the composite coatings withstand relatively higher number of fretting cycles. Composite coatings offer better fretting wear resistance up to 400°C due to the tribochemically induced oxide layer formation on the worn surface that provides lubrication effect. However, the performance of composite coatings drastically decreases at 500°C due to the decrease in hardness of the coating and delamination of protective oxide layer. The predominant fretting wear mechanism at temperatures less than 200°C is delamination while a combination of delamination and adhesion is operative when the temperature is increased beyond 200°C.

Tribocorrosion Behavior of Composite Coatings

Tribocorrosion is defined as degradation of materials in fretting, sliding, rolling, or erosion conditions in presence of a corrosive environment by a combined action of chemical–electrochemical–mechanical processes. The mechanism of tribocorrosion is not yet fully understood due to the complexity of the chemical, electrochemical, physical, and mechanical processes involved. The material removal in a tribocorrosion system usually exceeds the sum of mechanical and corrosion contributions measured separately. The tribocorrosion behavior of ED Ni-SiC and

Electro- and Electroless Composite Coatings, Table 2 Applications of electro- and electroless deposited composite coatings in various industries

Type of industry/ Application	Component/Assembly	Problems/Conventionally used coatings/Normal service life	Improvement in performance by adopting ED and EL Ni-based composite coatings
Rubber and plastic industry	Molds	The life of moulds for plastics, rubber etc. usually last for 10,000 moldings	A 50 μm thick ED and EL Ni–P–SiC coating increases the service life by 15 times
	Molds	Accelerated corrosion and abrasion of molds	EL Ni–P–SiC and EL Ni-P-PTFE coatings prevent accelerated corrosion and abrasion of molds and its performance is superior to chrome plating
Foundries	Core boxes	Wear and release of sand cores without breakage	EL Ni–P–SiC coating has been found to reduce wear and helps to release sand cores, without breakage, from core boxes
Oil and gas industry	Butterfly valves	Pick up and galling Increase in the leak rate – unsafe operation of the valves	A Ni–P–PTFE composite coating applied to a butterfly valve prevents pick up and galling and decreases the leak rate and enables safe operation of the valve for cryogenic applications
Textile industry	Thread guides yarn brakes, gears, friction clutches	Wear and friction	EL Ni–P–diamond coatings enable slipless transmission of very high rotational speed, an essential requirement of yarn brakes, infinitely variable gears, and friction clutches
Automobile industry	Carburetor parts, Choke shafts, Piston	Pickup and galling arise during forming and drawing operations	EL Ni–P–SiC coating has been used to overcome problems due to pickup and galling
	Cylinder liners Gears, etc.	Friction and dry lubrication	ED and EL Ni–P–PTFE on carburetor makes it non-stick, provides dry lubrication and low coefficient of friction
		Build-up of gummy deposits	EL Ni–P–PTFE coating minimizes build-up of gummy deposits on the choke shafts
		Wear resistance	ED Ni-SiC composite coating offers enhanced wear resistance
Aerospace industry	Aircraft turbine blades and turbine engine components	Wear resistance	ED Ni-Cubic BN composite coating offers enhanced wear resistance
			ED Ni-SiC, Ni-Al_2O_3 and Ni-Cr_2O_3 composite coating also offers an improved performance
	Inner surface of hydraulic actuating cylinders for the Joint Strike Fighter	Coefficient of friction	ED Co-P-SiC after heat-treatment offers a lower coefficient of friction than hard chrome plating and their wear rates are comparable
			ED Co-P-SiC coating is metallurgically sound and does not contain any of the cracks that are characteristic of hard chrome, which contributes to superior corrosion properties
	Landing gear, helicopter components		ED Co-P-SiC coating does not have any of the fatigue debit that is a major disadvantage for hard chrome plating

Electro- and Electroless Composite Coatings, Table 2 (continued)

Type of industry/ Application	Component/Assembly	Problems/Conventionally used coatings/Normal service life	Improvement in performance by adopting ED and EL Ni-based composite coatings
Cutting tools and grinding wheels	Profiled diamond tools, screw threads	High accuracy micro-finishing – very difficult with conventional electroplating	EL Ni–P–diamond, ED Ni-SiC, Ni-diamond and Ni-garnet composite offer improved performance
	Razor blades	Wear resistance	ED Ni-MoS_2 and Ni-h BN composite coatings offers enhanced wear resistance
Food processing industry	Processing equipment	Microbial adhesion and biofilm formation	Graded EL Ni–P–PTFE coatings could reduce the attachment of thermophilic streptococci by 82–97%
Cooling water systems	Heat exchangers	Scaling	Graded Ni–P/Ni–Cu–P/Ni–Cu–P–PTFE composite coating prevents $CaSO_4$ scale formation on heat transfer surfaces
		Microbial adhesion and biofilm formation	Stainless steel surfaces coated with Ag-PTFE composite coating reduced *E. coli* attachment by 94–98%, when compared with pure Ag coating, stainless steel or Ti
Electronic industry	Integrated circuits such as computer CPUs	Often experience thermal fatigue which produces a lot of heat which reduce their lifetime	ED Cu composite coating incorporated with microcapsules composed of a paraffin core with urethane shell can be used for heat sink applications since it combines the high thermal conductivity of copper with the high heat absorption capacity
		Metals though have a high thermal conductivity, their heat capacity is low	
	Microelectro-mechanical systems (MEMS)	Compatibility of permanent magnets with MEMS	ED Co-Ni-barium ferrite composite coating imparts a hard-magnetic behavior and compatible with MEM. Extent of magnetization can be increased with increase in level of incorporation of barium ferrite
		Performance of microresonant devices at high frequency	Increase in Young's modulus and hardness following incorporation of multi-walled CNT in EL Ni-P matrix enhances the performance of microresonant devices at high frequency
Solid oxide fuel cells	AISI 430 ferritic stainless steel interconnects	Rapidly decreasing electronic conductivity	ED Co/$LaCrO_3$ coatings on AISI 430 stainless steel limit chromium migration and increase high-temperature electronic conductivity
		Chromium volatility	Incorporation of $LaCrO_3$ particles in ED Co matrix not only improves the oxidation resistance but also eliminates scale spallation
		Poisoning of cathode material	
		Poor oxidation resistance	
		Spalling of oxide scales	The area specific resistance of ED Co/$LaCrO_3$-coated AISI 430 stainless steels does not exceed $\sim 0.02\,\Omega.cm^2$ after 900 h at 800°C in air

Ni-ZrO_2 composite coatings reveals that structural modification of the ED Ni matrix, increase in hardness due to dispersion strengthening, ability of these particles to provide a better load bearing capacity, and reduction in ductility of the matrix in the contact region reduced the plastic deformation and wear under tribocorrosion conditions (Benea 2009; Benea et al. 2009). ED Ni-SiC composite coating exhibits no signs of localized corrosion along the wear track area, whereas such an occurrence is clearly observed in Stellite6 hard face coating under tribocorrosion conditions (Benea et al. 2009). The loss due to tribocorrosion is significantly reduced from 54 to 18 g/kg when ZrO_2 particles are codeposited in ED Ni matrix (Benea 2009).

Corrosion Resistance of Composite Coatings

The improvement or impairment of corrosion resistance of ED and EL composite coatings depends on several factors. Decrease in porosity, change in microstructure from columnar to non-columnar structure, reduction in defect size of the metal matrix and chemical stability, and the ability of the particles to screen the metal matrix, to retard oxygen reduction reaction, and to prevent preferential corrosion of grain boundaries and triple junctions are the reasons suggested for improvement in corrosion resistance of composite coatings. Increase in surface roughness, increase in porosity, inhomogeneity of the coatings, and formation of electrochemical cells have been suggested as reasons for the decrease in corrosion resistance of composite coatings.

Key Applications

ED and EL composite coatings have received widespread acceptance to improve wear resistance and to provide a low coefficient of friction for many components in a variety of industries (Wang et al. 2001; Balaraju et al. 2003; Ploof 2008). ED and EL Ni-P PTFE composite coatings have been widely used in molds for rubber and plastic, pumps, ball and butterfly valves, fasteners, aluminum air cylinders, and carburetor choke shafts. EL Ni-P-PTFE coating has been specified for aluminum components in a vane rotary air compressor to minimize friction, noise, and vibration during operation.

ED Ni oil-containing microcapsules (μcaps) composite coatings are an interesting development as they could provide self-sustained lubrication in the interface on demand, especially for surfaces that are not easily accessible. The utility of this coating for electrical connector contacts has shown that they offer self-lubrication and provide cooling action during fretting and transient operation. However, the oil-containing μcaps also develop an interfacial insulating layer that significantly increased the contact resistance (Dervos et al. 1999).

Composite coatings incorporating materials with a phosphorescent nature are another interesting development. The ability of these coatings to emit a distinct, brightly colored light under UV light could serve as an indicator for the extent of wear-off of the coatings that warrant replacement. This type of coating could be valuable in authenticating parts and is especially promising for the identification of genuine OEM parts.

The performance of EL Ni-P-diamond, EL Ni-P-SiC in yarn line abrasive wear test, EL Ni-P-diamond coating in Taber wear test, and EL Ni-P-BN and EL Ni-P-PTFE coatings in Falex test suggests that these composite coatings have considerable advantageous in replacing and surpassing the performance of chrome plating. ED Co-P-SiC coatings, due their superior mechanical properties such as high hardness, wear resistance, and fatigue strength, as well as corrosion resistance, are also considered as an effective replacement for hard chrome plating (John Carpenter et al. 2009).

ED Ni/WC, ED Ni–Co–$LaNi_5$, ED Ni-TiO_2-Ti, and EL Ni-P-TiO_2-supported IrO_2 mixed oxide composite coated electrodes possess a high stability for hydrogen evolution reaction (HER). The activity for HER in these coatings is a function of level of incorporation, size of the particles, increase in real surface area, decrease in the apparent free energy of activation. and formation of oxides and intermetallic phases after heat-treatment.

Incorporation of nano-sized Fe_3O_4 particles imparts a ferromagnetic behavior for ED Cu matrix and the magnetic properties of the composite coatings can be modulated as a function of the level of incorporation of the Fe_3O_4 particles. Some of the applications are presented in Table 2.

Cross-References

- ▶ Electrochemical Deposition
- ▶ Electroplating
- ▶ Sliding Wear
- ▶ Surface Roughness

References

J.N. Balaraju, T.S.N. Sankara Narayanan, S.K. Seshadri, J. Appl. Electrochem. **33**, 807 (2003)

L. Benea, J. Appl. Electrochem. **33**, 1671 (2009)

L. Benea, F. Wenger, P. Ponthiaux, J.P. Celis, Wear **266**, 398 (2009)

J. Carpenter, A. Kertesz, A. Datta, Adv. Mater. Process. **167**(3), 25 (2009)

J.P. Celis, J.R. Roos, C. Buelens, J. Electrochem. Soc. **134**, 1402 (1987)

C.T. Dervos, C. Kollia, S. Psarrou, P. Vassiliou, IEEE Trans. Comp. Packag. Technol. **22**(3), 460 (1999)

I. Garcia, J. Fransaer, J.P. Celis, Surf. Coat.Technol. **148**, 171 (2001)

N. Guglielmi, J. Electrochem. Soc. **119**, 1009 (1972)

Y.S. Huang, X.T. Zeng, I. Annergren, F.M. Liu, Surf. Coat.Technol. **167**, 207 (2003)

C. Kerr, D. Barker, F. Walsh, J. Archer, Trans. IMF **78**(5), 171 (2000)
K. Krishnaveni, T.S.N. Sankara Narayanan, S.K. Seshadri, Electrodeposited Ni-B-Si_3N_4 composite coating: preparation and evaluation of its characteristic properties. J. Alloy. Compd. **466**, 412–420 (2008)
K. Krishnaveni, T.S.N. Sankara Narayanan, S.K. Seshadri, Wear resistance of electrodeposited Ni–B and Ni–B–Si_3N_4 composite coatings. J. Mater. Sci. **44**, 433–440 (2009)
C.T.J. Low, R.G.A. Wills, F.C. Walsh, Surf. Coat.Technol. **201**(1–2), 371 (2006)
L. Ploof, Adv. Mater. Process. **166**(5), 36 (2008)
P.C. Tulio, I.A. Carlos, J. Appl. Electrochem. **39**, 1305 (2009)
W. Wang, F.-Y. Hou, H. Wang, H.-T. Guo, Scripta Mater. **53**, 613 (2005)
Y. Wang, K. Brogan, S.C. Tung, Wear **250**, 706 (2001)

Electroadhesion

► Electrostatic Field Effects on Adhesion

Electrochemical Conversion Coatings

► Chemical Conversion Coatings

Electrochemical Deposition

Cesar Augusto Duarte Rodriguez,
Germano Tremiliosi-Filho
Instituto de Química de São Carlos, Universidade de São Paulo, São Carlos, São Paulo, Brazil

Synonyms

Cathodic deposition; Electrochemical deposition process; Electrodeposition; Electrolytic deposition

Definition

Electrochemical deposition is a process by which a thin and tightly adherent desired coating of metal, oxide, or salt can be deposited onto the surface of a conductor substrate by simple electrolysis of a solution containing the desired metal ion or its chemical complex.

Introduction

A thin, metallic, inorganic or organic coating electrochemically deposited onto a conductor or semiconductor substrate has been used for more than a century to confer surface properties different from those of the base substrate with the aim either of protecting the substrate against corrosive and erosive attack by the surrounding environment or for decorative purposes, providing a particular appearance, such color and luster. Often, coatings are used for a variety of reasons, including creating reflecting surface, conducting paths in printed circuits, magnetic layers, surfaces with specific friction parameters in sliding bearings, and restoration of the surface of worn parts, among others applications. Recently, due to advances in microsystems, microelectronics, and optics, increased use in corrosive environments and in tribological applications has demanded an increase in the quality and properties of coatings. Thus, the electrodeposition of metallic or non-metallic coatings plays an important role in the development of different technologies, including wear resistance.

The electrodeposition process employs different operating parameters, such as temperature, deposition current density and pH, producing different kinds of deposit structures. The production of a nanostructured surface by electrochemical deposition with a controllable roughness is of prime importance in material science dealing with tribology. For example, chromium electrochemically deposited on a steel substrate can be used as an effective surface hardener. Thus, the electrochemical deposition is an alternative and feasible method of preparing surfaces with tribological interest.

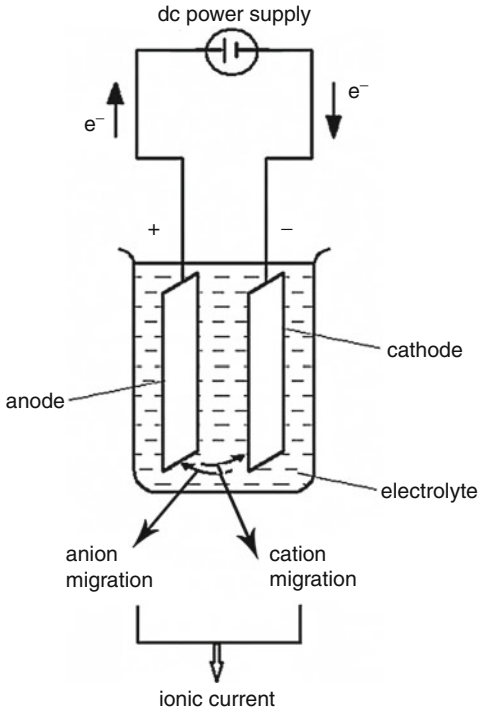
For reasons of economy and/or convenience, the automotive industry and a number of other key industries, have adopted electrochemical deposition even when other methods, such as evaporation, sputtering, and chemical vapor deposition, are options. Electrochemical deposition is a less expensive method compared with those mentioned above. It is also more versatile because can be applied in the simultaneous coating of a large number of samples and can be applied to pieces that are large in size. This discussion will be limited to electrochemical deposition as it relates to preparation of surfaces for tribological applications.

Scientific Fundamentals

Electrochemical Deposition Process (Bagotsky 2006)

To deposit a thin coating (metal, oxide, or salt) by electrochemical deposition it is necessary use an *electrolytic cell*, as shown in Fig. 1.

The *electrolytic cell* is a device consisting of two electronic conductors (*electrodes*) held apart from each other



Electrochemical Deposition, Fig. 1 Schematic representation of an electrolytic cell and its components

and both dipped into an *electrolyte*, usually a dissolved ionic compound. Connection of the electrodes to a source of direct electric current (DC power supply) renders one of them negatively charged and other positively charged. In the electrolytic cell a substance from the electrolyte or the surface of the electrode is chemically transformed and forms an adherent coating on the electrode surface, as required for tribological applications.

A conductor or semiconductor material in contact with an electrolytic solution is called an *electrode*. The electrodes are directly connected to the positive and negative terminals of the external power supply.

Thus, an electrolytic cell is composed by an *electrolyte* (ionic conductor) and two electronic conductors immersed into the *electrolyte* (electrodes). When electrons circulate through the electrode from the electronic external circuit (power supply) in the direction of the electrolyte, it is named *cathode*. *Anode* is the name of the electrode where electrons circulate in the opposite direction.

An *electrolyte* is a solution that conducts electricity and is composed by a solvent and a dissolved ionic compound dissociated in cations and anions, which migrate toward and ordinarily are discharged at the negative and positive electrodes (*cathode* and *anode*, respectively). Positive ions from the electrolyte solution migrate to the cathode and there receive electrons, losing part or all of their positive charge and becoming new lower charge ions or neutral elements or molecules; this process is known as *reduction*. At the same time, negative ions migrate to the anode and transfer electrons to it, also becoming new lower negative charged ions or neutral elements or molecules; this process is known as *oxidation*. These two processes, the transfer of electrons from negative ions to positive ions, are known as an electrochemical reaction.

The electrolyte solution is almost always aqueous, however, some coatings can be electrodeposited in organic electrolytes (e.g., aluminum) and some of the refractory metals (e.g., niobium, molybdenum, and tantalum) can also be deposited from fused salts.

In general, salts, acids, and bases, generically represented by C_xA_y, dissociate into ions when dissolved in water or other solvent according to the following equation:

$$C_xA_y \rightleftarrows xC^{z_+} + yA^{z_-}$$

where $|xz_+| = |yz_-|$, z_+ and z_- are the ion charge numbers. If $[C_xA_y]$ is the original concentration of C_xA_y without dissociation, then, $[C^{z_+}]$ and $[A^{z_-}]$ are the concentrations of the corresponding dissociated cation (C^{z_+}) and anion (C^{z_+}), respectively.

$$[C^{z_+}] = \alpha x[C_xA_y]$$

$$[A^{z_-}] = \alpha y[C_xA_y]$$

where α is the degree of the electrolyte dissociation. A solution formed by a substance that is fully dissociated, $\alpha \to 1$, is called a *strong electrolyte*.

The total concentration of all ionic species in solution, in the limit of $\alpha = 1$ (substance totally dissociated), can be written as $[C^{z_+}] + [A^{z_-}] = (x + y)[C_xA_y]$.

An electrolyte solution contains several types of species; their concentrations are interrelated with one independent concentration, $[C_xA_y]$.

The concentration of each ion, in units of moles L^{-1}, can be ascribed as,

$$[C^{z_+}] = n_{z_+}/V$$

$$[A^{z_-}] = n_{z_-}/V$$

where n_{z_+} and n_{z_-} are the number of moles of C^{z_+} and A^{z_-}, respectively, and V is the unit of the volume.

The electric charge of each ion is

$$Q_{z_+} = z_+Q^0$$

$$Q_{z_-} = z_-Q^0$$

where Q^0 is the elementary charge (1.62×10^{-19} C). The charge of 1 mol of ions is given by z_+F or z_-F, where $F = N_A Q^0 = 96,485\ Coulomb/mol$, which is the Faraday constant, and N_A is the Avogadro's constant ($6.02 \times 10^{23}\ mol^{-1}$).

The electric current (I) in a conductor is measured in Amperes (A). The current density (i) is independent of the conductor cross section area (A), i = I/A. As the electric current in an electrolyte solution is directly related to the motion of ions under the action of an applied electric field, i is proportional to the field strength ($\vec{\in}$),

$$i = \sigma \vec{\in}$$

where σ is the solution conductivity.

In an electrolyte solution the positive (C^{z_+}) and the negative (A^{z_-}) ions will move in opposite directions under the influence of an applied electric field. Thus, the total ionic current is the sum of the partial current due to transport of each ion (positive and negative).

In the electrolytic cell electrons enter through the negatively charged electrode (cathode). Positively charged ions of the electrolyte travel to this electrode, accept these electrons, and are *reduced* or chemically transformed to neutral elements or molecules. In particular, in the case of electrochemical deposition, the surface of the electrode changes with time by the incorporation of a metallic deposit formed on the substrate by the corresponding metallic ion reduction. The negatively charged components of the solution migrate to the other electrode positively charged (anode), release their electrons, and are *oxidized* or chemically transformed into neutral elements or molecules. If the surface of the metal electrode is chemically transformed, the reaction is generally one in which the neutral atoms of the electrode surface dissolve (are oxidized) as metallic cations by giving up electrons. The metallic cations formed simply go to the electrolyte bulk or are immediately precipitated on the electrode surface as an oxide or a salt; this process is known as *anodic dissolution/film precipitation.*

Electrolysis is used extensively for both coating of metals from the corresponding metallic cations on a cathode surface and in the formation of surface oxides or salts on an anode surface. During the course of electrolysis, the power supply applies a DC voltage between the positive and negative electrodes forcing an electric current (electrons) in the external part of the electric circuit that is complemented by an ionic current (transported by ions) into the electrolyte. Thus, during an electrochemical reaction, the electric current is carried by ions (into the solution phase) and electrons (in the external electric circuit). Ionic current in an electrolyte solution is the direct motion of ions under the influence of an applied electric field. Electronic conduction is found in metals, carbon materials (graphite, black carbon, carbon nanotubes, etc.), inorganic materials (some oxides, tungsten carbide, etc.), and a number of organic substances and is the motion of electrons due to an applied voltage.

The potential of an electrode in equilibrium (no net electrochemical deposition occurring) with an electrolyte of its ions of given concentration may be determined by using the Nernst equation, which relates the formal potential of an electrode to the equilibrium potential in the presence of a solution of its ions:

$$\mathrm{E} = \mathrm{E}^{\circ\prime} + (\mathrm{RT/nF})\ln[\mathrm{M}]$$

where E is the equilibrium potential of the electrode, $\mathrm{E}^{\circ\prime}$ is the formal potential, n is the number of electrons needed for the reduction of one metal ion, F is the Faraday constant, R is the gas constant, T is the absolute temperature, and [M] is the molar concentration of metal ions in equilibrium with the electrode. When the electrode potential is made more cathodic than the equilibrium value by imposing an external applied potential on it by a DC power supply, electrochemical deposition occurs until the metal ion concentration is lowered to the value that is in equilibrium with the electrode at the applied potential. By making the electrode sufficiently negative (cathodic), the metal ion remaining in the solution may reduce to a negligible concentration.

Key Applications

The electrodeposition of an adherent coating on an electrode in order to form a surface film with properties or dimensions different from those of the base substrate can be considered as a surface treatment. The thickness of the film growth by electrochemical deposition varies for each specific application. The thickness of a gold film for decorative purposes can be as little as 0.025 μm. For standard nickel-chromium plate on automotive hardware, the thickness is in the range of 25–50 μm. For tribological applications, the thickness can be up to 1 μm to 1 mm and can reach 1 mm or more for *electroforming* (production or reproduction of articles by electrodeposition).

The properties conferred by the surface coating vary with the application; these include improving wear resistance and hardness, appearance, corrosion resistance, frictional characteristics, solderability, specific electrical properties, and many others.

There is the possibility that atoms of a metallic deposit may, in time or at elevated temperatures, diffuse into the substrate structure and form an alloy with the substrate

metal. Such alloys may be brittle or have other undesirable properties. If such effects must be avoided, a barrier layer of another metal, most often nickel, is interposed between the substrate and the deposited coating. Where the application involves exposure to high temperatures, relative coefficients of expansion of the substrate and the deposit also may have to be taken into account.

Before a useful adherent layer can be deposited on a surface, the surface must be prepared to ensure the desired adherence to the substrate (Kenneth Grahm 1971). From a practical point of view, a satisfactory clean surface should not contain foreign material that interferes with the formation of an adherent deposit. The more contaminated the surface, the more cleaning it will require. This implies the removal of gross dirt and soil, heavy oxide or tarnish films, and surface skins of damaged metal produced by prior mechanical operations. A typical cleaning cycle includes (a) pickling to remove gross polishing or buffing compounds; (b) cleaning to remove oils and greases; (c) rinsing; and (d) acid dipping to remove oxide films and rinsing. Some substrates require more specialized cleaning treatment. The operations of electrodeposition include cleaning, rinsing, depositing, and post-electrodeposition treatments. Some agitation of electrolyte solution and temperature control is required. The final surface may require several layers of different metals to provide a wear-resistant surface.

Factors affecting the resulting deposit include pretreatment and cleaning of the substrate metal surface, concentration of metal ions in electrolyte, agitation, current density, temperature, conductance of electrolyte, pH, and addition agents.

Electrochemical deposition operations are conducted with electrical equipment operated by direct current, thus, a rectifier is employed.

Low friction, hard, and wear-resistant coatings can be easily produced by electrochemical deposition. Examples of electrochemical-deposited coatings applied primarily for wear resistance are hard chromium, oxides, nitrides, carbides, and borides. Coating hardness measurement is done by means of nanoindentation. Adherence control can be made by means of a scratch test. Coating wear resistance is checked by means of a pin-on-disc tribometer, and atomic force microscopy (AFM) can be used to evaluate of surface quality.

Increased surface hardness can be extend the lifetime of tools, mechanical components, or wear parts. Many processes have been developed to increase surface hardness, including electrochemical deposition and diffusion techniques. Different techniques have been designed to be applied on specific materials and in specific applications. However, electrochemical deposition is more appropriate to for producing coatings that are extremely hard and corrosion resistant (Dan et al. 1993). Hard chromium coating is described below as an example.

Electrochemical Deposition of Hard Chromium

Hard chromium coating is usually electrodeposited in a thickness ranging from 0.25 to 500 μm. It prolongs the life of metal parts that are subject to friction, abrasion, wear, and corrosion and is regarded as a means of surface hardening. The Vickers hardness is between 900 and 1,100. Most of the deposits are applied on ferrous alloys. The cathodic deposition of chromium commonly uses a sulfate and fluoride containing chromic acid baths. A standard chromium bath contains 150–260 g L^{-1} chromic acid, 1 – 1.7 g L^{-1} of sulfate (sulfate over 120:1 ratio), temperature of 45–65 °C, and current density up to 2 A dm^{-1}. The current efficiency is low, generally in the range of 10–20 %. To prepare the bath, the chromium trioxide or chromium anhydride (CrO_3) has to be dissolved in water to give a solution containing both ions, CrO_4^{2-} and $Cr_2O_7^{2-}$, according to the following reaction:

$$2CrO_3 + 2H_2O \rightarrow 2CrO_4^{2-} + 4H^+ \rightleftarrows Cr_2O_7^{2-} + H_2O + 2H^+$$

The CrO_4^{2-} ion is in equilibrium with $Cr_2O_7^{2-}$.

In the cathodic reaction, the chromic acid ($Cr_2O_7^{2-}$) is reduced to metallic chromium.

$$Cr_2O_7^{2-} + 14H^+ + 12e^- \rightarrow 2Cr + 7H_2O$$

Besides the desired reaction of metallic chromium formation, many undesired reactions occur, such as the hydrogen gas formation in acidic electrolyte.

$$2H^+ + 2e^- \rightarrow H_2$$

Another additional side reaction is the reduction of Cr(VI) to Cr(III) at the cathode surface. Cr(III) cannot be reduced to metallic chromium, according to

$$Cr_2O_7^{2-} + 14H^+ + 6e^- \rightarrow 2Cr^{3+} + 7H_2O$$

Sulfuric acid is the most common sources of sulfate. The sulfate ion in the chromic acid bath acts as a catalyst in the reduction of Cr(VI) to metallic chromium and inhibits hydrogen gas formation. The most used insoluble anodic materials are lead and lead alloy (Pb-Sn, 93–7 %).

Chromium represents a potential environmental problem. Thus, the bath as well the rinsed water have to be effectively treated before being discarded.

Electrochemical Codeposition of a Surface Hardening Composite Coating

Electrochemical codeposition enables the production of a large range of composite materials with specific tribological properties. It is a suitable method for combining the advantages of metallic electrodeposition with those of prepared composite. The main uses of electrodeposited composites are in the wear resistance, corrosion resistance, and lubrication. Thus, inert submicron hard particles held in suspension in an electrolyte can be electrochemically codeposited in a metallic matrix (Hovestad and Janssen 1995). The particles could be of different types, for example, pure metals, ceramics, or organic materials. Incorporation in the metal coating of dispersed ceramics hardened particles like Al_2O_3, WC, TiO_2, SiC, and so on forms composites that have a considerably higher yield strength and hardness than the pure metallic matrix and are protected from abrasion. The resulting apparent surface hardness is usually over 1,300 HV. These hard materials exhibit weak surface adhesion to metal and are too brittle to be used as coating alone. Nevertheless, metal cement is used to hold the particles and promote its good adhesion. The metal cement strongly sticks to the metal base and the occluded particles adhere vigorously to the coating. Particularly, cubic boron nitride (c-BN) is known as the second hardest material after diamond. It has good thermal stability and chemical inertness regarding iron. In contrast, diamond cannot be applied as a hard and wear-resistant material codeposit with iron, where relatively higher temperature is demanded due to the problem of the dissolution of carbon in iron at high temperature. This advantage of c-BN has opened up possibilities for wider use of iron and its alloys in the field of automotive and aerospace industries. Although c-BN can be coated using physical and chemical vapor deposition methods, electrochemical codeposition is less expensive and can be easily codeposited with such metals as nickel. Ni/c-BN composite film may be applied to cutting, drilling, and gridding hard alloys of steel and other materials (Teruyama et al. 2004).

The incorporation of particles, such as Al_2O_3, $BaSO_4$, Si_3N_4, V_2O_5, Cr_2O_3, and so on, in a metal deposit decreases the corrosion rate of the base metal. Furthermore, electrodeposited composite coatings are used to increase the lifetime of metals surface that are in moving contact.

Finally, in addition, polymer particles like polytetrafluoro-ethylene (PTFE) and polyethylene (PE) have also been used to reduce the friction coefficient and to achieve the anti-stick surface of the composite.

Many experimental factors influence the codeposition process that involves a number of steps and usually needs to be performed at high temperature. The main factors that influence the codeposition are (a) the particle concentration in the bath, (b) the particle shape and size, (c) the adsorption of ions on the particle surface, (d) suspension stability, (e) bath constituents, (f) temperature, (g) pH, (h) current density, (i) presence of surfactant, (j) agitation, and (k) surface charge of the particles. Many of these factors are interrelated and influence the codeposition process, however, particle concentration in the bath, current density, and bath agitation seem to be the most important parameters to be controlled.

Attempts to develop mechanistic models for codeposition, which are able to predict the amount of incorporated particles, were only partly successful and offer a promising perspective.

Electrochemical deposition is of prime importance in the preparation of surfaces for tribological applications. It has opened up a wider range of possibilities of applications, improving wear resistance and hardness and corrosion resistance and conferring excellent frictional characteristics to surfaces. Extremely hard and corrosion-resistant coatings can be obtained by electrodeposition of hard chromium. Electrochemical codeposition of inert particles in a metal matrix is also a suitable technique to produce composite coatings for tribological applications.

Cross-References

- ▶ Chromizing
- ▶ Corrosive Wear
- ▶ Duplex Coatings
- ▶ Electrochemical Deposition
- ▶ Electroplating
- ▶ Nanocomposite Coatings
- ▶ Polymer Composites and Nanocomposites

References

V.S. Bagotsky (ed.), *Fundamentals of Electrochemistry*, 2nd edn. (Wiley, New Jersey, 2006)

J.P. Dan, H.J. Boving, H.E. Hintermann, J. Phys. IV **3**, 933 (1993)

A. Hovestad, L.J.J. Janssen, J. Appl. Electrochem. **25**, 519 (1995)

A. Kenneth Grahm (ed.), *Electroplating Engineering Handbook* (Van Nostrand Reinhold Company, New York, 1971)

S. Teruyama, N.K. Shresthia, Y. Ito, M. Iwanaga, T. Saji, J. Mater. Sci. **39**, 2941 (2004)

Electrochemical Deposition for Self-Lubricating Metal Composite Coatings

▶ Self-lubricating Metal Composite Coatings by Electrodeposition or Electroless Deposition

Electrochemical Deposition Process

► Electrochemical Deposition

Electrochemical Plating

► Electroplating

Electrocrystallization

► Electroplating

Electrodeposition

► Electrochemical Deposition

Electrolytic Deposition

► Electrochemical Deposition

Electrolytic Plasma Technology (EPT)

► Ceramic Conversion of Light Alloys

Electromagnetic Interactions and Adhesion

► Basic Concepts in Adhesion Science

Electron Beam Dispersing

► Electron Beam Surface Technologies

Electron Beam Hardening

► Electron Beam Surface Technologies

Electron Beam Surface Alloying

► Electron Beam Surface Technologies

Electron Beam Surface Technologies

ROLF ZENKER
TU Bergakademie Freiberg, Institute of Materials Engineering, Freiberg Zenker Consult, Mittweida, Germany

Synonyms

Electron beam dispersing; Electron beam hardening; Electron beam surface alloying; Electron beam technologies

Definition

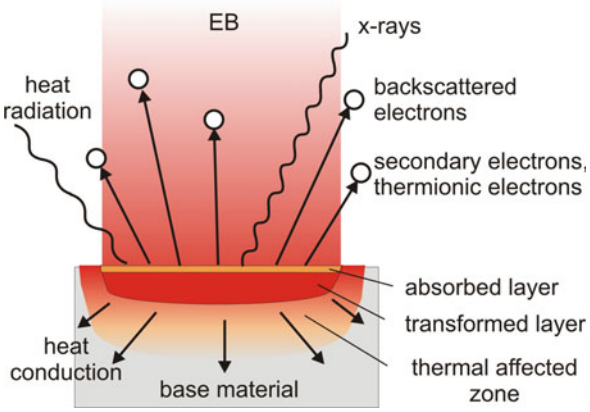
Electron beam surface technologies are thermal high speed heat treatment processes in solid and/or liquid state. Depending on the technological conditions they are carried out as surface annealing, hardening, tempering or surface remelting, alloying, dispersing, or cladding.

Scientific Fundamentals

Principles of Electron Beam Interaction with Materials

The electron beam (EB) is a stream of charge carriers accelerated in an electric field up to a rate of two-thirds the speed of light. A substantial portion of its kinetic energy is converted to heat within a thin layer (some microns, depending on acceleration voltage and material) when it impinges on a metallic surface (Fig. 1). The portion of power converted to heat in the material is about 85% (solid state processes) up to 95% (liquid state processes). The remaining energy is lost through electron backscattering, X-ray, secondary electrons, and heat radiation. The absorption layer is heated and the layer below is warmed up rapidly by heat conduction. After the EB interaction with the material, the material is rapidly

cooled also by heat conduction. The heating and cooling rates are 10^3–10^4 (10^5) Ks^{-1}. Depending on the energy density of the EB (up to 10^5 (10^6) Wcm^{-2}) in the interaction area, different processes take place. If the surface temperature is not higher than the melting point, the material is influenced by solid state processes only. In the case that the surface temperature is higher than the melting point, liquid phase processes take place. The depth of the melting zone depends on whether the beam energy is lower (fusion pool) or higher (keyhole effect).



Electron Beam Surface Technologies, Fig. 1 Effects of EB interaction with material

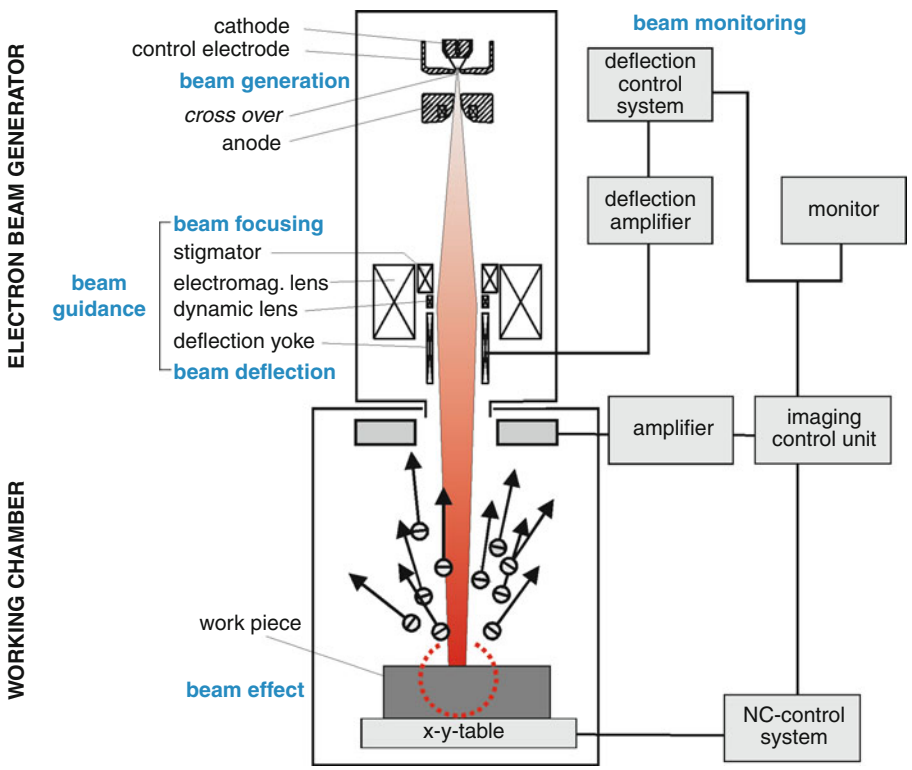
Because of its excellent formability and deflectability with high frequencies (up to 100 kHz), it is possible to generate manifold beam deflection patterns by different beam deflection techniques.

Further advantages of the EB are a good beam profile, large penetration depth, high beam stability, and high efficiency.

EB (surface) technologies mostly are carried out as vacuum processes in order to avoid oxidation and decarburizing (solid state processes), and a good degassing of pores (liquid state processes) is guaranteed.

Electron Beam Facilities

EB facilities generally consist of an EB generator and a working chamber (Fig. 2). Electrons are emitted by thermal emission from a tungsten cathode (electron cloud) and accelerated by an electrostatic field (acceleration voltage for surface technologies 60–150 kV) as



Electron Beam Surface Technologies, Fig. 2 EB facility (principle)

a "crude beam." Beam current (≤200 mA) is controlled by a control electrode. The beam forming and guiding system consists of a stigmator, static and dynamic lenses, and a beam deflection system. For thermal surface treatment, the EB has a diameter of 0.1. . .0.4 mm. The EB generator works under high vacuum conditions (10^{-5}–10^{-6} mbar). Depending on the application, the working chamber with handling systems is evacuated to a soft vacuum (10^{-2}–10^{-3} mbar) or high vacuum (10^{-5}–10^{-6} mbar). For process observation and/or control, EB facilities are equipped with optical and/or secondary electron observation systems.

For different tasks and/or performances, universal or single-purpose EB facilities with different configurations (single-chamber, multi-chamber, shuttle-type, or lock-type facilities) are available.

Beam Deflection Techniques

Usually, the EB is acting in an energy transfer spot. Its size is at least 0.1–0.4 mm (focused beam diameter) up to some millimeters in diameter (defocused EB). When using high frequency beam deflection techniques (≤100 kHz), the beam can act either as an oscillation pattern (circle, ring, trigonometrical functions) or within an energy transfer field of 200 mm × 200 mm maximum. The EB can be deflected so that the energy transfer takes place nearly in situ in every position within this field with lateral homogeneous distribution of energy or locally defined dosed energy distribution.

Related to the EB application, there are two basic techniques. Either the component and/or the EB is/are continuously moved relative to one another when the beam interacts with the material (continuous interacting (CI) technique) or the component is fixed below the beam when the EB is interacting with the material (flash technique) (Fig. 3).

Regarding EB process technology in case of multi-spot deflection technique, the same process is carried out in every spot. Multi-process technologies are characterized by quasi in situ implementation of different processes in different areas treated (Fig. 4).

Thermal EB Surface Technologies

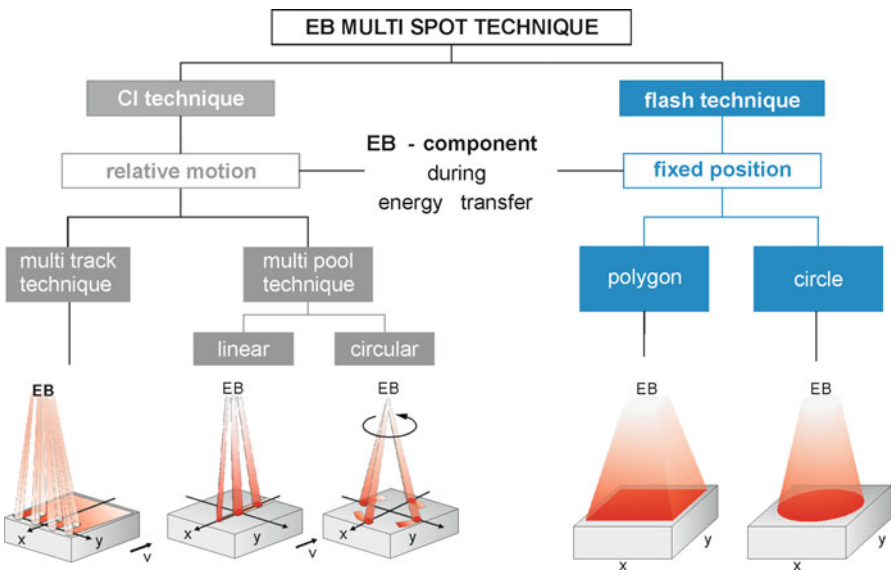
Different thermal EB surface technologies are necessary and available for different applications (Fig. 5).

The selection is mainly determined by the material and the load conditions of the component. Thermal EB technologies are characterized by high productivity, excellent flexibility, large process variety, good process safety, excellent reproducibility, and ecological friendliness.

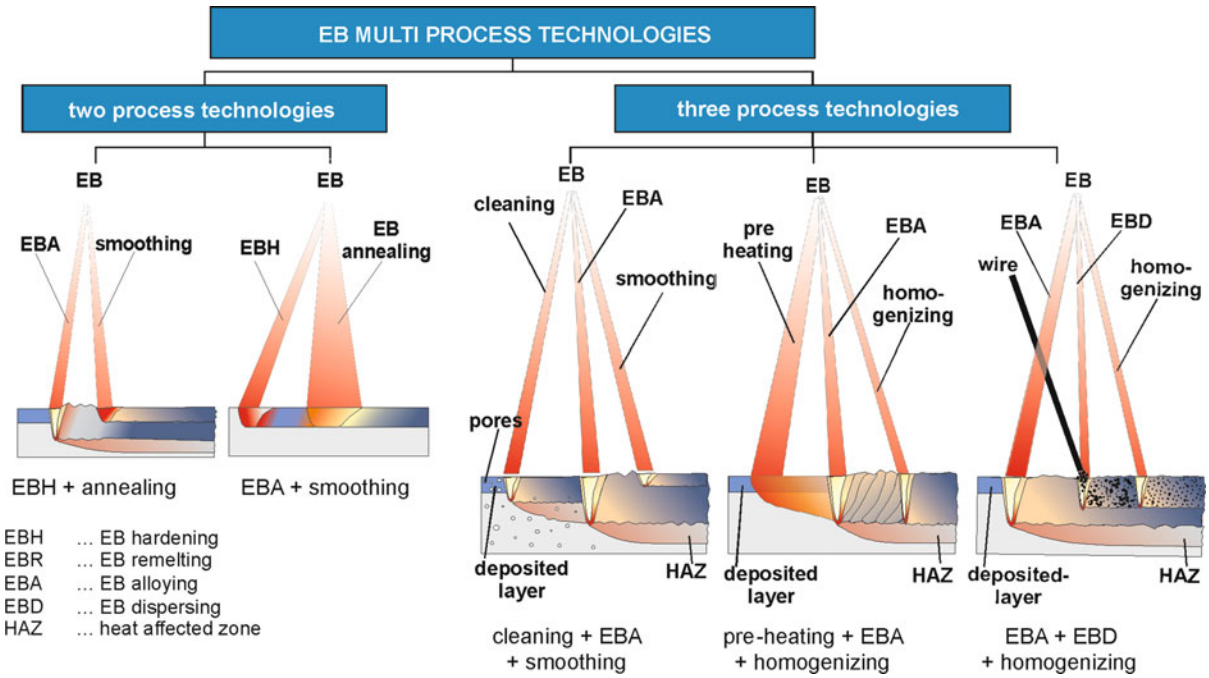
Solid Phase Processes ($T < T_M$)

EB Surface Hardening

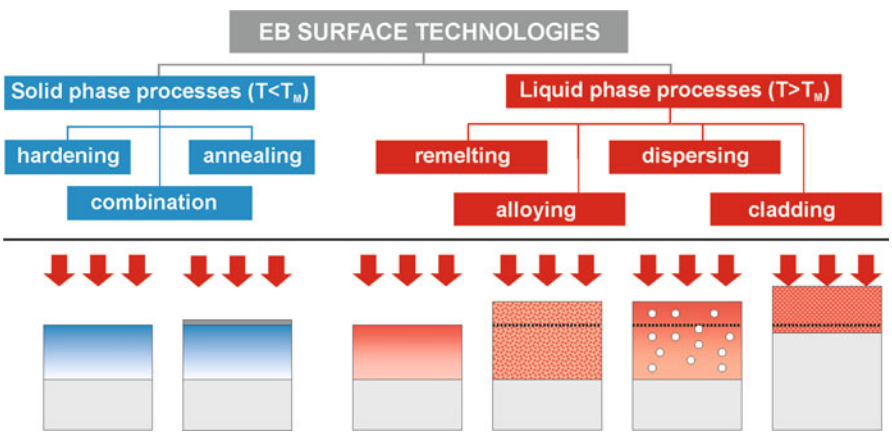
The hardenable material is heated to its austenitizing temperature up to a certain depth (0.1–2.0 mm) and after the effect of the beam has cooled down by self-quenching, martensitic transformation takes place.



Electron Beam Surface Technologies, Fig. 3 Multi-spot beam deflection techniques



Electron Beam Surface Technologies, Fig. 4 Multi-process EB technologies



Electron Beam Surface Technologies, Fig. 5 Thermal EB surface technologies

EB Surface Annealing or Tempering

The material is heated up to a critical temperature in a certain depth and cools down so slowly that martensitic transformation is avoided and recrystallization or precipitation processes are initiated.

Combined EB Surface Technologies

In combination with EB hardening, thermochemical treatment (carburizing, nitriding, nitrocarburizing, boriding) or hard coating is realized before or after EB hardening in order to improve the supporting effect of the base material to the layer without or with deliberately changing the layer's structure and/or composition. These combined surface technologies in principle are two-step processes.

Liquid Phase Processes ($T > T_M$)

These technologies are carried out without or with additions previously deposited or added during EB action.

EB Surface Remelting

The material is heated up to a temperature above T_M up to a certain depth (0.1–3 (8) mm) to obtain a melting bath. After the beam effect, the material solidifies rapidly by self-quenching. Thus, the microstructure but not the nominal composition of the surface layer is changed.

EB Surface Alloying

The surface layer of the material and the additional material will be heated up to a temperature above the melting point T_M of the base material and the material added. During the subsequent self-quenching solidification process, a sequence of metallurgical processes takes place, changing the layer's chemical composition, microstructure, and properties.

EB Surface Dispersing

The surface layer of the material is heated up to a temperature above the melting point T_M of the base material and to a certain depth. The melting point of the additional material is not reached so that the particles of the additional material are dispersed in the molten layer. During the subsequent high-speed self-quenching solidification process, this state becomes "frozen."

EB Cladding

An additional material and a thin layer of the base material are heated up to a temperature above the melting point T_M of both base material and additional material so that the additional material is completely transformed into a liquid state whereas the base material approaches such a state only within a very low depth. The subsequent quick solidification results in a layer of different chemical composition, microstructure, and structural shape that firmly adheres to the base material.

Deposition of Additional Material

The deposition of additional material is carried out either as a two-step process (deposition of additional material before the EB treatment) or as a one-step process (deposition of additional material during interaction of EB with material in the melting bath).

Effects on Microstructure and Properties

Microstructure

The main aim of applying EB surface technology is wear protection. It is, for example, also used to prevent corrosion, to improve plastic behavior, or to reduce internal stresses.

The success of EB surface treatment depends, among other things, on the material to be treated in accordance with the technology chosen in connection with a suitable beam deflection technique related to the geometry of the component, especially its surface contour.

Because of the short local time-temperature cycles, the microstructure is generally very fine-grained and the phases caused are mostly non-equilibrium states after EB treatment.

EB Hardening and Tempering

The solid state processes like EBH and tempering are connected to hardenable materials (steels, cast irons). The martensitic microstructure of hypereutectoid steels has a lath-like morphology with small packets of nearly parallel short laths (Fig. 6a). Hypereutectoid steels are characterized by plate-like martensite with different percentages of retained austenite.

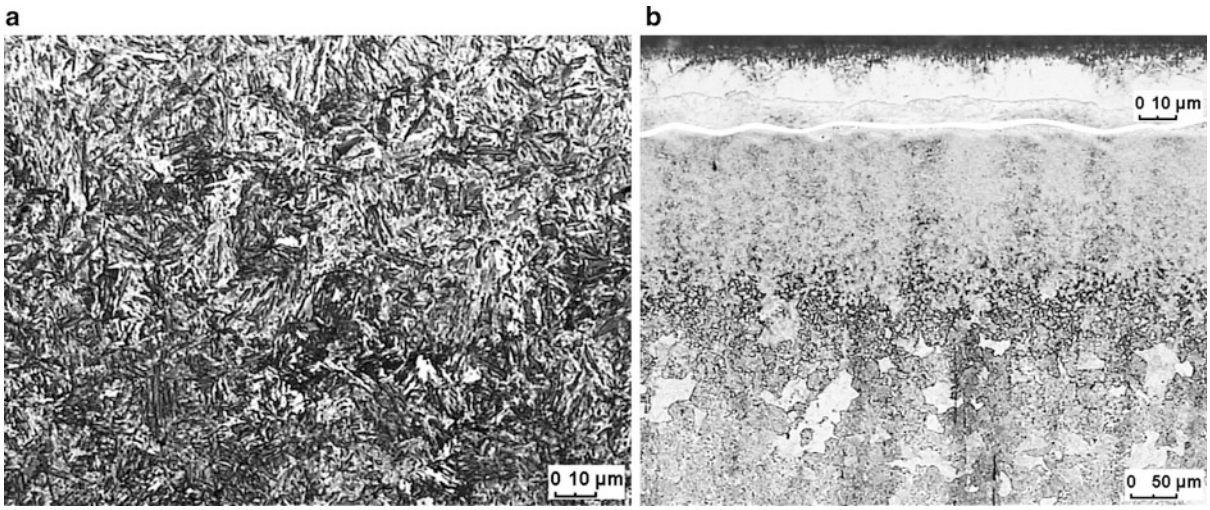
Perlite containing cast irons are also EB hardenable by martensitic transformation of the former perlitic matrix. The graphite is not transformed, but the carbon can diffuse from the graphite into the ferrous matrix at the edges of the graphite particles in spite of the short time-temperature cycle. In case of former ferritic matrix areas, martensite forms along the border between graphite and ferrous matrix after quenching. If a former perlitic matrix is enriched with carbon by diffusion out of the graphite, martensite and retained austenite exist after quenching.

EB Annealing

EB annealing mostly corresponds to local recrystallization processes and/or local relaxation of internal stresses. As a result of this treatment, the deformed grains transform more or less into small polygonal grains in the thermally influenced zone depending on the temperature and how long the EB is in interaction with the material.

Combination of Thermochemical Treatment/Hard Coating with EB Hardening

When EB hardening is applied after thermochemical treatment (TTC), the diffusion layer and the matrix material are martensitically transformed up to a certain depth. The morphology of the martensite depends on the concentration of C/N and alloying elements. The compound layer should not significantly change in its structure and microstructure, apart from the growth of the pore seam (Fig. 6b). This is connected with a reduction of N concentration. This treatment sequence is preferred for alloyed steels with a relatively low tempering stability and the need for both a large depth effect and a wear-resistant nitride surface layer.



Electron Beam Surface Technologies, Fig. 6 Microstructures after EB solid state surface treatments (**a**) Martensite after EB hardening (low alloyed steel) (**b**) Nitride layer (partially transformed compound layer; martensitic diffusion layer) on low alloyed steel after N + EB hardening combination

The combined EBH + N is suitable for steels with high tempering stability (nitriding temperatures >400°C) and is necessary for load conditions like high abrasive wear and corrosion.

In case of combined EBH/hard coating (HC) after EBH, the matrix material is also martensitic up to a certain depth. Depending on the substrate material, other phases (carbides, nitrides, retained austenite) are present. Even though the processing surface temperature during EBH can be higher than 1,000°C, visual appearance, structure, and composition of the thin (2–4 µm) hard-coated surface layer are not influenced significantly if the conditions are optimal and because the EB treatment is realized in a very short temperature-time cycle (milliseconds to seconds) and is carried out in vacuum.

The sequence of combined surface treatment EBH + hard coating is useful only in cases where the layer deposition is carried out at temperatures lower than the tempering temperature of the base materials (e.g., when the hard coating is a PVD or PA-CVD process). The level of the processing temperature in relation to the tempering stability of the bulk material determines the success of this treatment combination. The better the tempering stability of the steel, the smaller the hardness reduction in the earlier produced EBH layer as a result of the hard coating process.

For both combinations, thermochemical treatment and hard coating with EBH, the main task of the martensitic layer is the improvement of the supporting effect for the hard surface layer.

EB Remelting

EB remelting is successfully applicable in the case of cast irons. During remelting, the graphite is completely dissolved. In result of the rapid solidification, the liquid phase is transformed into dendritic ledeburite (Fig. 7a). In the interdendritic areas, martensite is formed. During or after quenching, cracks are initiated because of the high internal stresses. Therefore, the matrix material must be pre-heated up to temperatures >400°C before EB remelting. In this way, very fine perlite forms and crack formation is eliminated.

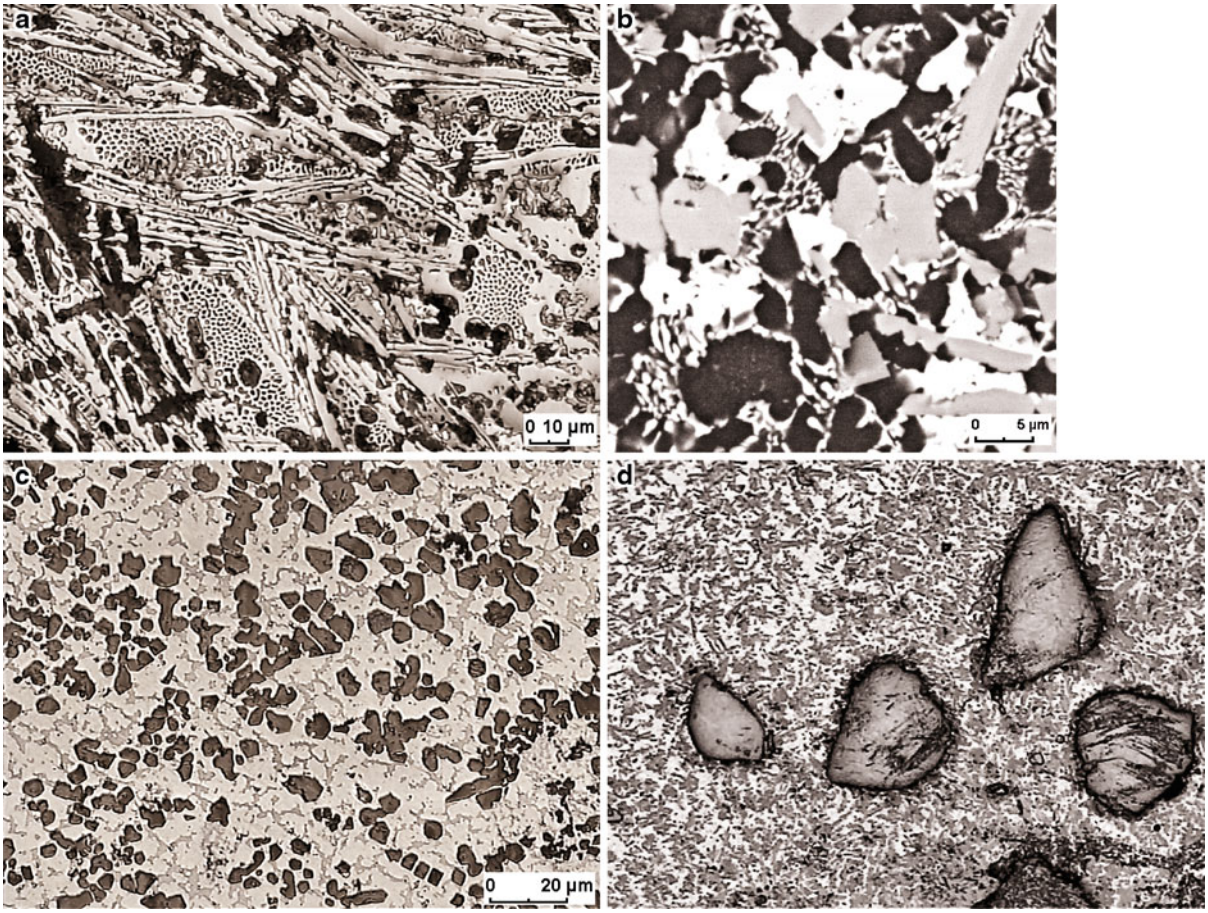
EB remelting of Mg alloys results in a finer microstructure and the precipitation of new phases ($Mg_{17}Al_{12}$) due to rapid self-cooling.

EB densifying – a special variant of remelting – is successfully used for cast Al alloys, porous spraying layers, and sintered materials. This remelting technology results in dense layers with very fine (10 times finer SDAS) microstructures free of pores (and cracks). Because of the high heating and cooling rates, it is possible to get metastable non-equilibrium phases, on the one hand, and to especially avoid undesired dispersion phases, on the other hand. The formation of interdendritic Al-Si eutectic can also be avoided.

This group of EB surface technologies is carried out without changing the nominal composition in the re-melted layer.

EB Alloying

EB liquid state processes, such as alloying, dispersing, or cladding, are suitable for Fe, Al, Ti, and Mg alloys.



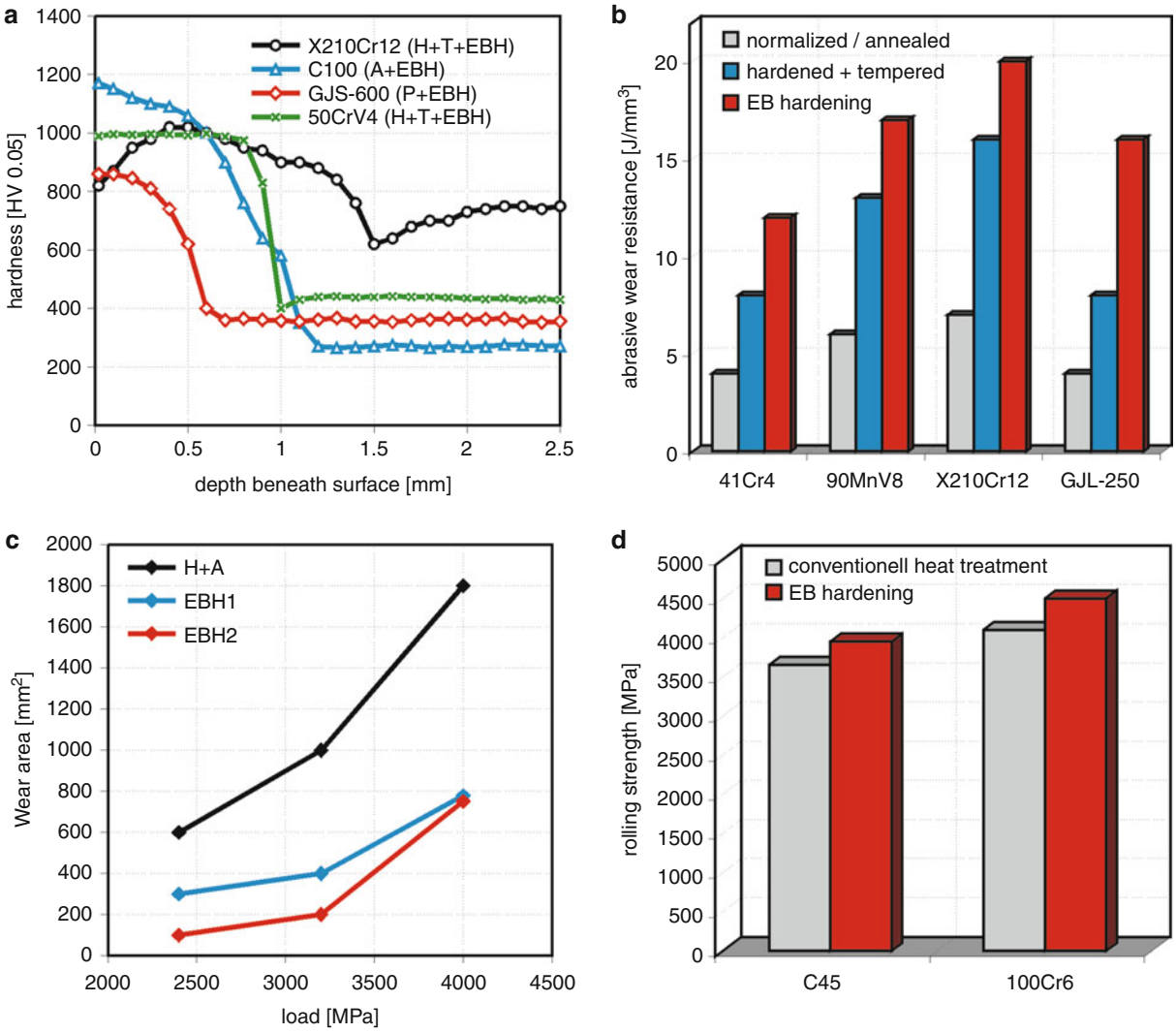
Electron Beam Surface Technologies, Fig. 7 Microstructure after EB liquid phase surface treatments (**a**) Dendritic ledeburite (globular cast iron) after EB remelting. (**b**) Al matrix (AlSi10) and different intermetallic compounds (Cu-Ni18Cr18 addition) after EB alloying. (**c**) Different intermetallic phases in Mg solid solution after EB alloying. (**d**) Embedded WC particles in an Al matrix with intermetallic compounds after EB cladding

The success of these EB technologies depends on the suitable combination of matrix material with additional material, deposition technology, EB technology in connection with beam deflection technique, process parameters, and layer thickness. By goal-directed change of composition, structure, and microstructure, excellent property combinations are caused in the layers.

EB remelting as well as alloying with vanadium (8–10%) results in significant grain refining compared with the base material. Carbides are also refined and precipitations of fine vanadium carbides are formed in addition to the well-known carbides (e.g., M_6C, $M_{23}C_6$).

Enrichment of the matrix solid solution by suitable additional elements (e.g., Ni, Cu, Co, (Fe)) and the formation of intermetallic compounds are initiated by EB alloying of Al cast alloys in the transformed layer (Fig. 7b). It is not only the type of intermetallic compounds but also their morphology that is decisive for the effect on the properties. Regarding layers alloyed with Cu or Ni on Al cast alloys with depths $>$1–3 mm, the criterion of absence of cracks is fulfilled for layer hardness values $<$400HV0.10. In contrast, on spray-formed Al alloys with 12–40% Si and $\leq$10% Ni, Cu, or Fe additions, layer depths $<$0.8 mm with a hardness of up to 600HV0.10 could be produced without any cracks. Some of the limits mentioned for the production of technically usable layers can only be fully utilized through additional measures such as pre-heating and post-heating and/or EB multiple treatment.

The precipitation microstructure resulting from surface alloying of Al (cast) alloys is very fine with regard to all additional material variants examined. The type, amount, size, morphology, and distribution of



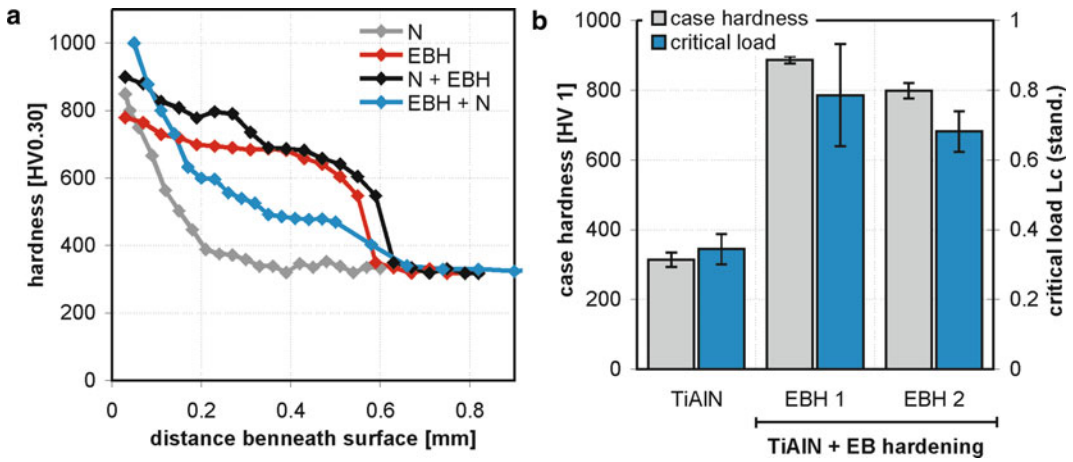
Electron Beam Surface Technologies, Fig. 8 Properties after solid phase EB surface treatment (EB hardening) (**a**) Hardness depth profiles of different steels after EB hardening. (**b**) Abrasive wear resistance of different steels and cast iron after EB hardening in comparison to conventional heat treatment. (**c**) Fretting fatigue after EB hardening of Cr steel in comparison to hardened and tempered state. (**d**) Rolling wear after EB hardening in comparison to conventional hardening and tempering of steels

microstructural components, such as primary silicon and intermetallic phases (e.g., Al_xNi_y, Al_xCu_y, Al_xCr_y, $(Fe, Mn)_xAl_ySi_z$), are controlled and changed in a targeted manner by varying the EB parameters. Due to the proceeding high-speed heating and cooling processes, there is an oversaturation of the Al solid solution and unbalanced phases are also formed.

By means of rapid self-quenching of Mg alloys, the microstructure of Mg alloys consists of very fine α-Mg solid solution, globular and acicular $Mg_{17}Al_{12}$, and the facetted phase Mg_2Si after EB alloying with Al, Si, and/or Ti in case of lower Al concentration. Higher Al concentrations lead to α-Al solid solution, Mg-Al eutectic, and Mg_2Si intermetallic phase with globular morphology (Fig 7c).

EB Dispersing and Dispersion Alloying

In case of EB dispersing, additional insoluble or precipitated finely dispersed particles in the matrix material are responsible for the improvement of the layer properties.



Electron Beam Surface Technologies, Fig. 9 Properties after solid phase EB surface treatment (combined surface treatment) (**a**) Hardness depth profiles after combined EB hardening + nitriding, nitriding + EB hardening in comparison to nitriding and EB hardening. (**b**) Critical load after combined hard coating + EB hardening in comparison to hard coating

EB dispersion alloying aims at the same effects as the combination of EB alloying and EB dispersion. The additional materials must consist of components completely or partially soluble in the matrix material and components that are insoluble in the matrix or can precipitate during cooling or precipitation hardening after EB processing.

In case of cast Al alloys (AlSi10), as a result of dispersion alloying by addition of Ni and WC after the EB treatment, Ni is partially solved in Al solid solution and partially precipitated as intermetallic compound (e.g., Al_3Ni, $AlNi_3$). Additional fine particles of WC are imbedded in the Al matrix (Fig. 7d).

Different very fine intermetallic compounds are imbedded in the Ti matrix of the EB alloyed layer (Al and/or Co additions) on TiAl6V4. A specialty of this $\alpha + \beta$ Ti alloy is that an α' transformation hardened layer exists behind this EB alloyed layer.

EB Cladding

Different technologies exist because of the specific cladding characteristics (the matrix material is influenced only minimally).

The application of cladding for steels or cast iron is known for tools and locally high-loaded components as well as for regeneration of turbine blades, rolls, tools moulds, and dies. The microstructure is very manifold and depends on the special demands in relation to the load conditions.

For Al alloys, successful EB cladding results were achieved by a layer consisting of Al solid solution, fine particles of primary Si particles of intermetallic compounds rich in Ni, for example.

Properties

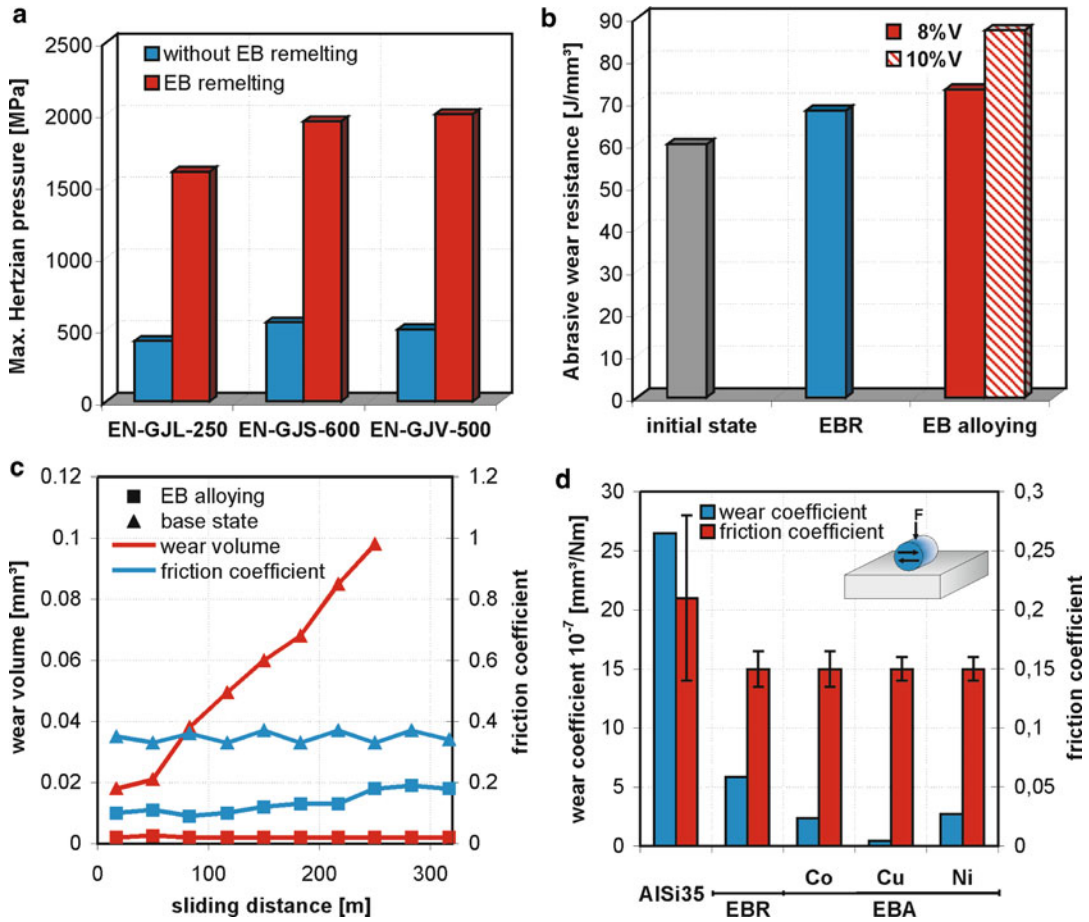
In case of complex load conditions, a functional sharing between surface and core material properties is often required. After a heat treatment of bulk material that increases toughness, a selected EB surface treatment of the highest-loaded areas opens up the possibility to meet these requirements. Depending on the load conditions and the base material, different concepts of EB surface treatment are available for optimum configurations of the layers.

The layer's properties are determined by type, quantity, size, distribution, and morphology of existing phases after transformation. However, EB treatment does not only influence the layer itself, but also the properties gradient.

An additional important factor affecting the properties of the layers is the quality in relation to the existence of pores and/or cracks. Pores mostly are the result of chemical reactions between some components during EB interaction with the material or results from the matrix material. Cracks are the result of internal stresses and are initiated by volume differences, thermal and/or transformation stresses, or differences in thermal expansion coefficients of existing phases.

EB Hardening

The surface hardness is raised by the formation of fine-grained martensite as a result of EB hardening of steels.



Electron Beam Surface Technologies, Fig. 10 Properties after liquid phase EB treatment. (**a**) Hertzian pressure after EB remelting of cast irons in comparison to casting state. (**b**) Abrasive wear resistance of EB remelted and EB alloyed high speed steel in comparison to conventional hardened + tempered state (scratch test). (**c**) Sliding/friction wear volume and friction coefficient (pin-disk wear test) after EB alloying of TiAl6 with NiAl in comparison to base material. (**d**) Wear and friction coefficients (sliding wear test) of AlSi35 after EB remelting and alloying in comparison to base material

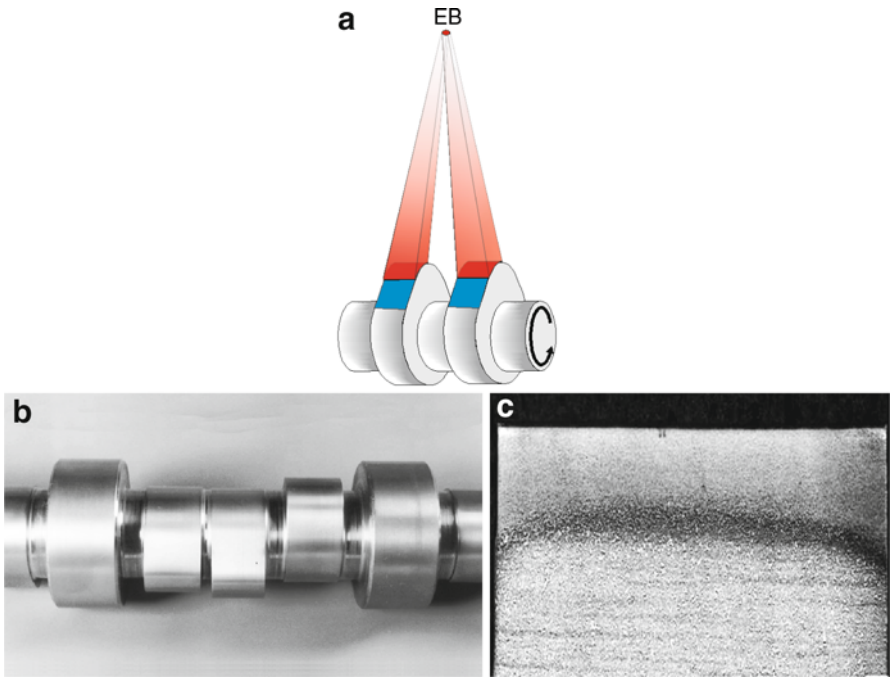
Hardness depth profiles (Fig. 8a) are characterized by transformation products in the EBH layer and both by base material and its pre-heat treatment state. In case of high alloyed steels, the hardness may be reduced in an area close to the surface because retained austenite is formed. The lower the content of carbon and alloying elements of the steel, the steeper the transition from the hardness values of the EBH layer (plateau) to the hardness level of the base material. A hardness minimum at the transition area between EBH layer and base material is typical for hardened and tempered base materials.

Significant improvement of abrasive wear resistance of steels in relation to conventional heat treatment is typical for EBH (Fig. 8b). However, it is not only the effect of the fine-grained martensite in the EBH layer but EB hardened steels also do not have to be tempered in most cases in contrast to conventionally hardened steels. The abrasive wear often corresponds to the surface hardness.

In case of optimum process parameters related to the load conditions, both fretting fatigue (Fig. 8c) and rolling wear resistance (Fig. 8d) can be improved by EB hardening.

EB Annealing

The result of EB annealing is the reduction of strength and improvement of formability of the material and/or reduction of internal stresses. In case of austenitic steels, it is possible to reduce by half the hardness and tensile strength. The effect is not so drastic with regard to non-ferrous materials. Alloys hardened by precipitation lose



Electron Beam Surface Technologies, Fig. 11 EB hardening of cam shaft. (**a**) Two-field technique. (**b**) Cam shaft segment. (**c**) EBH layer

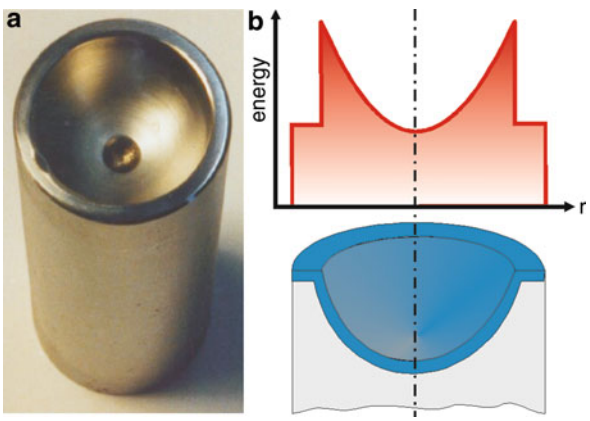
their hardness and strength by thermally influencing ("annealing") the material (e.g., in a zone behind remelting layers) as a side-effect.

Combination of Thermochemical Treatment/Hard Coating with EB Hardening

The combined nitriding + EB hardening and EB hardening + nitriding technologies are used for locally high-loaded materials (abrasive/sliding wear in connection with high local compression and/or corrosion). Every technology leads to a typical hardness depth profile (Fig. 9a). In comparison to the hardness depth profiles after nitriding and EB hardening, the profiles after combined treatment show higher hardness, at least in the area near the surface.

The sequence of the combined surface treatment EB hardening + hard coating is useful only in cases where the layer deposition is carried out at temperatures lower than the tempering temperature of the base materials (e.g., when the hard coating is a PVD or PA-CVD process). Thus, surface hardness as well as critical load (scratch test) are slightly improved.

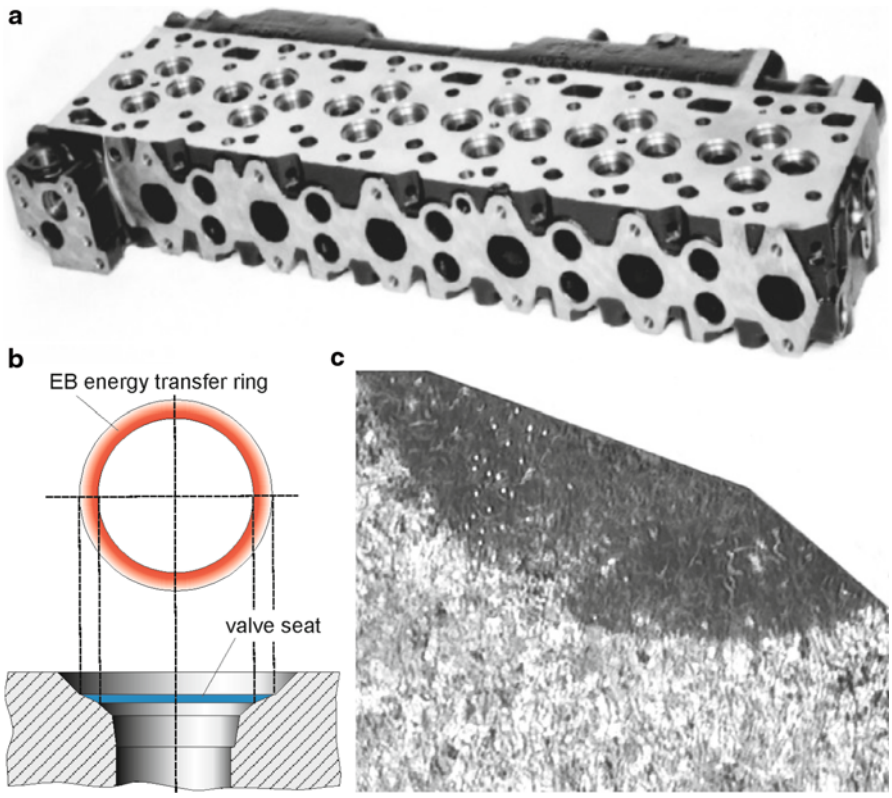
If the sequence of the treatment is applied in the opposite order (HC + EBH), a significant increase of surface hardness and critical load is achieved (Fig. 9b).



Electron Beam Surface Technologies, Fig. 12 EB hardening (flash technique). (**a**) Calotte carrier. (**b**) Energy transfer field and hardened zone

EB Remelting

EB remelting of cast irons also leads to significant increase in hardness, adhesive and abrasive wear, and maximum Hertzian pressure (Fig. 10a) as a result of rapid solidification and transformation to dendritic ledeburite ("white cast iron").



Electron Beam Surface Technologies, Fig. 13 EB hardening of valve seats. (**a**) Cylinder head (cast iron). (**b**) Energy transfer ring. (**c**) EBH zone after finish grinding

Depending on the composition of steels, the effect of EB remelting is differentiated. High-speed steels have a light raised hardness and abrasive wear resistance (Fig. 10b).

Surface remelting of Mg or Al alloys leads to an increase of hardness (two to three-fold). Abrasive wear resistance is mostly raised by ~80%. In the case of Mg alloys, a weak improvement of corrosion behavior is possible. With an AlSi35 alloy, the wear and friction coefficient of EB alloyed layers are drastically reduced in comparison to base material (Fig. 10c).

EB Alloying

The improvement of properties (hardness, wear, and/or corrosion resistance) depends on the kind and fraction of additional alloying elements (i.e., the degree of mixture between base material and additives).

EB alloying noticeably improves surface hardness and abrasive wear resistance of high-speed steels (Fig. 10b). As a result of EB alloying of Ti alloys with NiAl sliding/friction, wear volume and friction coefficient (pin-disk wear test) are increased significantly in comparison to base material (Fig. 10c).

Depending on the additional material (Cu, Ni, Co), wear and friction coefficients (sliding wear test) of AlSi35 after EB remelting and alloying are reduced up to five to six times in comparison to base material (Fig. 10d).

With an optimum combination of EB parameters, additional materials (e.g., Cu, Co, Co, Ni, Fe, also as mixture), and transformed volume (layer thickness) similar results were reached for Al cast alloys.

Layer properties of Mg alloys, especially wear and/or corrosion behavior, can be positively influenced by alloying with additives like Al, Ti, Si, (Ni). This can lead to significantly higher hardness of the surface layer (up to 350HV0.1) and mostly of abrasive wear resistance. In terms of corrosion, the surface layer rich in Al shows corrosion properties close to an Al-Si cast alloy.

EB Dispersing and Dispersion Alloying

In case of cast Al alloys (AlSi10) dispersion alloying improves hardness and adhesive/abrasive wear. Matrix

hardness in the layer is improved by addition of Ni, partially solved in Al solid solution and partially precipitated as intermetallic compound (e.g., Al_3Ni, $AlNi_3$). Additional fine particles of WC, imbedded in the Al matrix, support the effect on the wear resistance.

Dispersing of hard particle materials in combination with Al additives (e.g., AlSi12/SiC) leads to the best (adhesive/abrasive) wear properties of Mg alloys (e.g., AZ31, AZ91). However, optimization is needed to reach a very fine dispersion of the particles. A homogeneous and high level of layer hardness is attainable because of high-volume contents of fine hard particles. There are also noticeable effects with regard to the corrosion behavior in this case.

Key Applications

Solid State Processes

EB Hardening of Camshafts

EB hardening of camshafts is carried out by field technique. Another remarkable aspect is that if neighboring cams have (nearly) the same position and its intermediate distance is not too wide, a two-field technique is applied (Fig. 11a), that is, two cams are treated simultaneously (Fig. 11b), nearly doubling the productivity.

The base materials have to be heat-treatable steels. However, EB hardening is now used for camshafts made from nodular cast iron as well. Because the surface deformation is very small (4– < 10 µm), it is not necessary to apply a grinding process after EB treatment, but a finishing process is carried out in most cases.

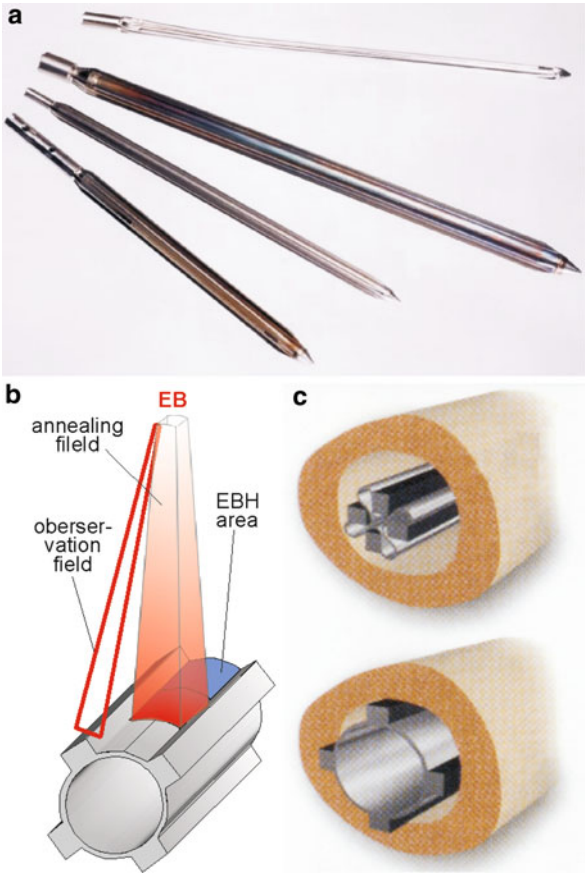
The distinctive feature of the EB process is that the layer thickness may be varied depending on the load around the cam contour. Consequently, the heat transfer into the camshaft can be minimized, which is important with regard to the tempering effects and the deformation of the camshaft. The width of the EBH layer reaches up to the border of the cams in contrast to EB liquid phase processes (Fig. 11c).

EB Hardening of a Calotte Carrier

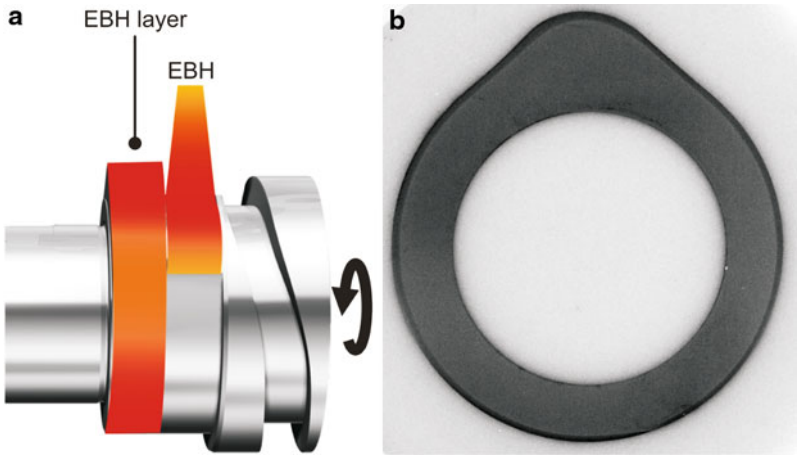
Another technically and economically attractive EB hardening technology is applied in the case of calotte carriers (Fig. 12a). The energy transfer is realized by flash technique. During the interaction of the EB ($\leq$1.0 s), the component is fitted to the beam before crossing the α/γ transformation temperature (processing time $\leq$0.2 s). The surface contour is programmed as a rotation-symmetric energy transfer field with a surface contour congruent to the energy distribution (Fig. 12b). The resulting hardening profile is characterized by an almost constant EBH thickness (DS) along the whole contour that is nearly independent from the incidence angle of the EB.

EB Hardening of Valve Seats

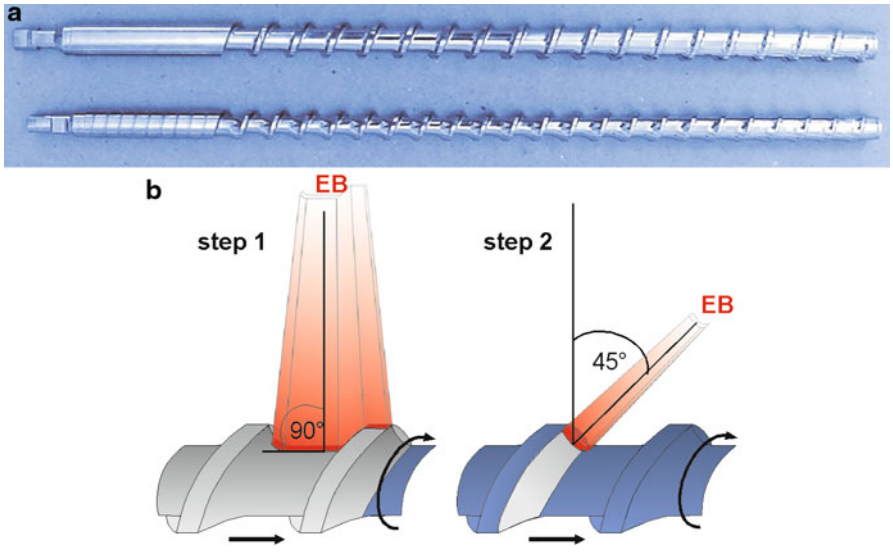
The present heat treatment technology for valve seats in cast iron cylinder heads (Fig. 13a) is induction hardening. The main problem caused by this technology is the strong distortion of the cylinder head, which requires an appropriate rework (grinding). In this application, the EB flash technique is the alternative with the most effective energy transfer method (flash ring, Fig. 13b). It is possible to treat a 4-valve, 6-cylinder head in 60 s. An additional effect is that neither a quenching medium nor subsequent tempering is required, thanks to EB hardening by self-quenching.



Electron Beam Surface Technologies, Fig. 14 EB annealing of implants. (**a**) Medical nails. (**b**) Energy transfer field with positioning line scan. (**c**) Medical nail after folding and after inflation



Electron Beam Surface Technologies, Fig. 15 Combined nitriding and EB hardening of cam segments (**a**) EBH of a nitrided cam segment. (**b**) Hardened cam contour



Electron Beam Surface Technologies, Fig. 16 Combined EB hardening and nitriding of extrusion components. (**a**) Extrusion screw. (**b**) Two-step field technique

The short-time energy transfer with high-energy density is realized with minimum heat transfer into the surrounding material. Expecting an exact positioning of the energy transfer ring, the hardened zone is limited to the contact area of valve/valve seat (Fig. 13c). This EB technology is flexible with regard to different geometrical conditions of the components. An additional important feature is that positioning of the EB to the valve seat can be carried out simultaneously at the beginning (heating) of the heat treatment process.

EB Hardening of Rods

In the case of a special motor concept under unfavorable load conditions, two neighboring steel rods can contact one another, causing abrasive wear at the end face.

The application of EB hardening of the ring-shaped contact zone is one method to minimize the wear. The energy transfer is realized by a rotating energy transfer field along the circular hardening contour.

The hardening depth is $\leq$0.3 mm. Thus, there is no distortion of the rods. The time of treatment is below 1.5 s.

EB Annealing of Medical Implants

A special application of EB annealing is the annealing of medical nails (Fig. 14a) used for implantation in broken bones. The nail, made from austenitic steel for medical applications, has a cross-section consisting of four thin membranes (150–350 μm) and four stabilizing "bridges." The whole cross-section is strengthened by plastic deformation after machining. The energy transfer is only carried out in the area of the thin membranes using an exactly positioned (automatic) positioning control (Fig. 14b) energy transfer field. The softened membranes will be folded after this treatment (Fig. 14c). In this state, they are introduced into the bone and then inflated to the original form (Fig. 14c) in order to stabilize the bone.

Combined Nitriding Plus EB Hardening of Cam Segments

Cam segments are exposed to high dynamic load combined with abrasive/adhesive wear and local high surface pressure, in particular along the cam contours. These demands are satisfied by combined nitriding + EBH. Prior nitriding provides sufficient wear protection in all areas where it is necessary. But the high-loaded cam contours must also be EB hardened (Fig. 15a). The EB parameters are optimized so that the nitriding layer, especially the compound layer, is not (fully) transformed and the good wear resistance remains nearly complete. The hardness depth must be nearly constant over the complete cam contour (Fig. 15b).

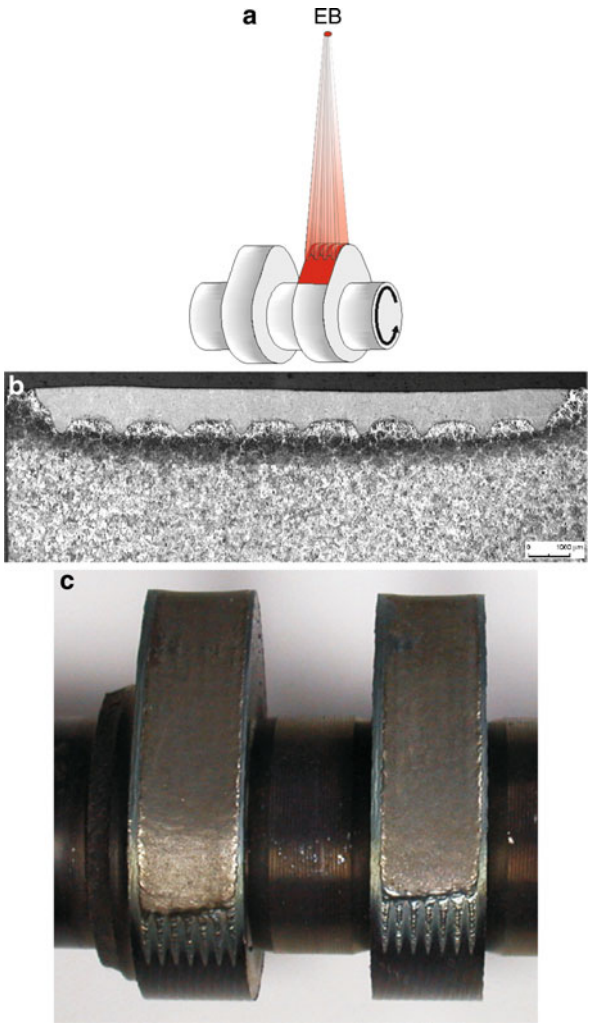
Combined EB Hardening Plus Nitriding of Extrusion Screws

The entire surface contour of extrusion screws (Fig. 16a) is exposed to severe load due to abrasion and corrosion at working temperatures above 400°C. Therefore, a combined treatment is necessary for the entire screw contour. Due to the intricate surface geometry, EB hardening is carried out as a two-step process prior to nitriding (Fig. 16b). The base material (high-alloyed steels) is characterized by good tempering strength so that hardness and strength are not negatively influenced by subsequent nitriding at temperatures above 400°C.

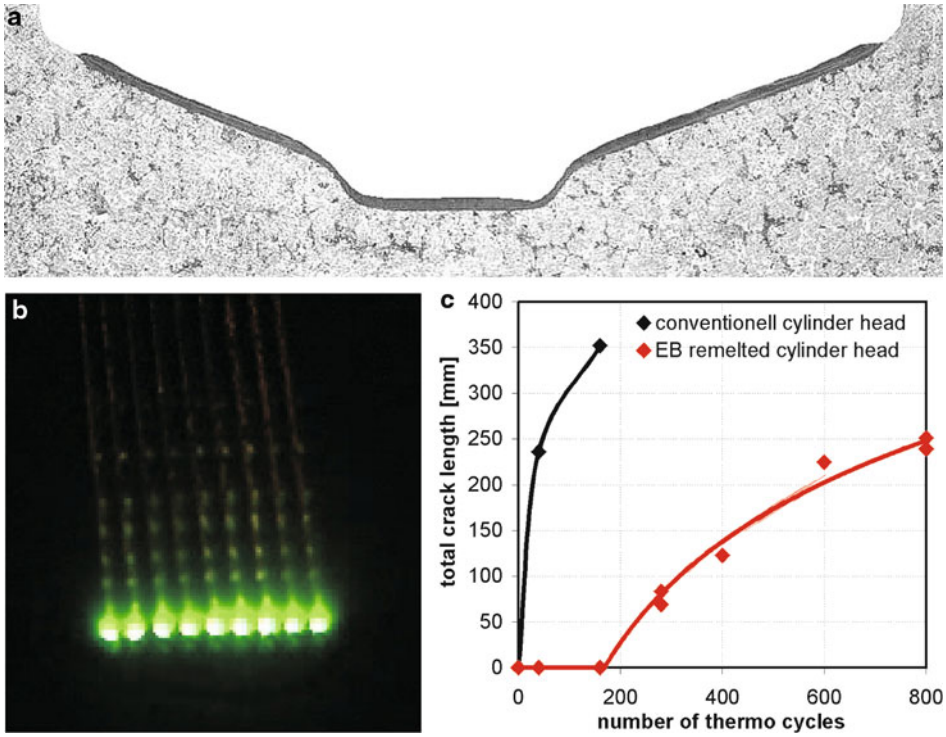
Liquid State Processes

EB Remelting of Camshafts

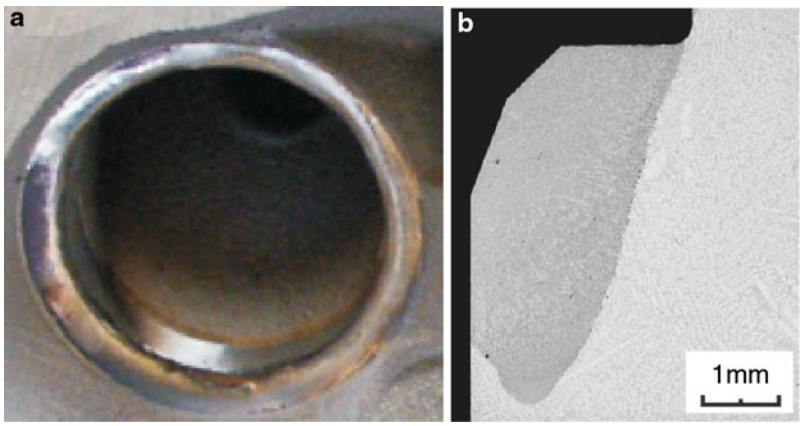
EB remelting of camshafts has been especially developed for ferritic and perlitic cast iron with lamellar and globular graphite using the multi-track technique (Fig. 17a). The surface of the cams is remelted up to a depth of 0.6–0.8 mm (max. 1.00 mm) in a tooth-like manner (Fig. 17b). This beam deflection technique combines the advantages of high processing speed (the whole width of the cam or several cams can be treated simultaneously) with small surface deformation (several small melting pools). The camshafts must be preheated to temperatures ≥400°C. Otherwise, micro cracks that may grow and become macroscopic cracks would arise in the martensite crystals. The camshaft may be destroyed under load. Under these conditions, the cams are sufficiently tough and have a high hardness and wear resistance. Nevertheless, EB remelting (Fig. 17c) must be followed by a grinding process (≤0.3 mm).



Electron Beam Surface Technologies, Fig. 17 EB remelting of cam shaft. (**a**) Multi track technique. (**b**) Remelting layer. (**c**) Remelted cam



Electron Beam Surface Technologies, Fig. 18 EB densifying of cylinder head. (**a**) EB remelting contour. (**b**) Multi-spot EBU of cylinder head. (**c**) Life time thermocyclical load



Electron Beam Surface Technologies, Fig. 19 EB alloying of valve seat (**a**) Valve seat after EB surface alloying with CuNiCr. (**b**) Cross section of alloyed valve seat after finish machining

EB Densifying of Cylinder Heads

A successful method to avoid cracks resulting from heavy thermocycles is to densify the surface layer of the bridge between the valve holes in cylinder heads by EB densifying (Fig. 18a). The very dense, fine-grained microstructure up to a depth of 0.3–8.0 mm is free of pores and very stable against crack initiation and extension. This was achieved by using the EB multi-spot meander technique for remelting the cylinder head around the valve holes (Fig. 18b).

If there is any crack initiation, it takes place at irregularities behind the EB densified layer in the untreated cast structure. The cracks grow very slowly throughout the layer up to the surface. The lifetime of cylinder heads treated this way is six times longer than of those untreated (Fig. 18c).

EB Alloying of Ring Grooves in Pistons

It is a fact that the first ring groove of a piston is heavily loaded both thermally and mechanically. The strength and wear resistance of the piston material (Al alloys) is not high enough to resist these demands so that usually an additional metallic ring is inserted in the mould and after the casting process it is surrounded by piston material.

Another alternative and more effective technology for improving strength and wear resistance of the groove is EB remelting or alloying of the ring groove area. In case of low loaded pistons, EB remelting is effective. With regard to higher loaded pistons, EB alloying is essential. This is done by a double-pool technique. The addition is brought in by a one-step process (e.g., a wire supply). The result of this treatment is a fine-grained zone of some millimeters in dimension with - in case of EB alloying – fine intermetallic precipitations that increase hardness, strength, and wear resistance.

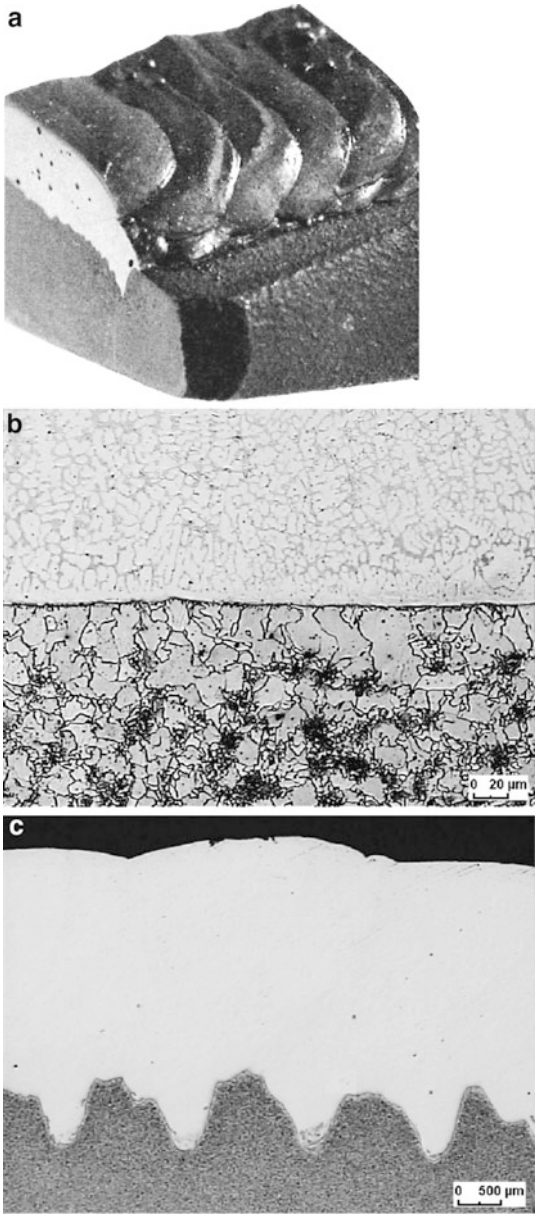
EB Alloying of Valve Seats

The complex loading conditions of valve seats in cylinder heads (thermocyclic load, abrasive wear, surface fatigue, Hertzian pressure) make high demands on the material. As an alternative to the conventional method (inserts of Fe- or Ni-based rings), a solution using EB alloying technology has been developed. In the first step, the additional material is deposited by introducing a sinter ring. In the second step, this area (deposited ring together with some surrounding base material) is remelted by a multi-pool EB technique (Fig. 19a). A homogenous, fine-grained layer with a thickness of $\leq$4 mm without pores or cracks and metallurgical bonding to the base material without gaps can be produced (Fig. 19b). The heat transfer from the valve seat to the cylinder head matrix material was significantly improved, resulting in lower peak temperatures at the valve seat compared with the current production solution (valve seat ring). Under these conditions, the maximum hardness of the EB alloyed layers was $\leq$ 450HV0.10. Wear resistance reaches about the same values.

EB Dispersion Alloying of Bearings

The basic wear condition of friction bearings is the well-balanced ratio between hard, wear-resistant and soft areas as an area with emergency running properties. This is achieved by different methods.

The principle of EB alloying is that some traces of transformed material are generated simultaneously, so that a front of melting pools moving relative to one another can act on the material surface (dynamic multi-track technique).



Electron Beam Surface Technologies, Fig. 20 EB cladding of ploughshare. (**a**) Cladded surface layer. (**b**) Layer: NiCrBSi, base material: steel. (**c**) Tooth-like bonding layer/matrix

The additional material is pre-deposited, partially as a sandwich layer. The result is a lateral remelting pattern of different load-optimized configurations. This technique permits a great variety of different technological possibilities thanks to the flexible, high-frequency deflectability of the EB and provides a high productivity.

EB Cladding of Ploughshare

One cladding application is a technology for ploughshares (Fig. 20a). Because of the special load conditions – high impact load and high abrasive wear and corrosion – the base material's degree of toughness and strength has to be high and the surface must resist wear and corrosion. The demands in relation to the base material are fulfilled by selecting alloyed steel. The properties of the surface are realized by EB multi-spot cladding with additional material based on NiCrBSi (Fig. 20b). In addition to the good layer properties, the beam deflection technique contributes to a good adhesive strength of the layer because of the tooth-like boundary between layer and matrix (Fig. 20c).

Cross-References

- ► Duplex Coatings
- ► Gas Nitriding
- ► Laser Cladding
- ► Laser Surface Alloying
- ► Laser Surface Hardening
- ► Multiplex Coatings

References

A. Buchwalder, *Beitrag zur Flüssigphasen-Randschichtbehandlung von Bauteilen aus Aluminiumwerkstoffen mittels Elektronenstrahl*, Dissertation TU Bergakademie Freiberg, 2007

A. Buchwalder, R. Zenker, Modern thermal electron beam processes. Research results and industrial applications, in *European Conference on Heat Treatment 2008; ECHT'08*: 7–9 May 2008, Verona, 2008, AIM: CD (paper 38)

J. Rödel, *Beitrag zur Modellierung des Elektronenstrahlhärtens von Stahl*, Dissertation TU Bergakademie Freiberg, 1996

G. Sacher, R. Zenker, N. Frenkler, T. Kimme, Kombinierte Randschichtwärmebehandlung – PVD-Hartstoffbeschichtung in Verbindung mit dem Elektronenstrahl- oder Laserstrahlhärten. HTM J. Heat Treat. Mater. (formerly HTM Zeitschrift für Werkstoffe, Wärmebehandlung, Fertigung) **64** (1), 20–27 (2009)

S. Schiller, U. Heisig, S. Panzer, *Electron Beam Technology* (Verlag Technik GmbH, Berlin, 1995)

R. Zenker, *Elektronenstrahl-Randschichtbehandlung – Innovative Technologien Für Höchste Industrielle Ansprüche* (Pro-beam AG & Co, KGaA, Munich, 2003)

R. Zenker, Structure and properties as a result of electron beam surface treatment. Adv. Eng. Mater. **6**(7), 581–588 (2004)

R. Zenker, Elektronenstrahl-Mehrspot-Technik - Neue Möglichkeiten und Perspektiven für die Randschichtbehandlung. Stahl Eisen **2**, 26–28 (2007)

R. Zenker, Modern thermal electron beam processes – research results and industrial application. Metallurgia Italiana **1014**, 55–62 (2009)

R. Zenker, Electron meets nitrogen -combination electron beam hardening and nitriding. Int. Heat Treat. Surf. Eng. **3**(4), 141–146 (2009)

R. Zenker, N. Frenkler, T. Ptaszek, Neuentwicklungen auf dem Gebiet der Elektronenstrahl-Randschichtbehandlung. HTM **54**(3), 143–149 (1999)

R. Zenker, A. Buchwalder, N. Frenkler, S. Thiemer, Moderne Elektronenstrahltechnologien zum Fügen und zur Randschichtbehandlung. Vak. Prax. **17**(2), 66–72 (2005)

R. Zenker, H.-J. Spies, A. Buchwalder, G. Sacher, Combination of high energy beam processing with thermochemical treatment and hard protective coating: state of the art. Int. Heat Treat. Surf. Eng. **4**(12), 152–155 (2007a)

R. Zenker, G. Sacher, A. Buchwalder, J. Liebich, A. Reiter, R. Häßler, Hybrid technology hard coating – electron beam surface hardening. Surf. Coat.Technol. **202**, 804–808 (2007b)

R. Zenker, A. Buchwalder, M. Klemm, Neue Entwicklungen auf dem Gebiet der thermischen Elektronenstrahl-Randschichtbehandlung von Al-Werkstoffen. HTM J. Heat Treat. Mater. (formerly HTM Zeitschrift für Werkstoffe, Wärmebehandlung, Fertigung) **64**(4), 208–214 (2009)

Electron Beam Technologies

► Electron Beam Surface Technologies

Electron Energy Loss Spectroscopy (EELS)

Wanfeng Li[1], Chaoying Ni[2]
[1]Department of Physics & Astronomy, University of Delaware, Newark, DE, USA
[2]Department of Materials Science & Engineering, University of Delaware, Newark, DE, USA

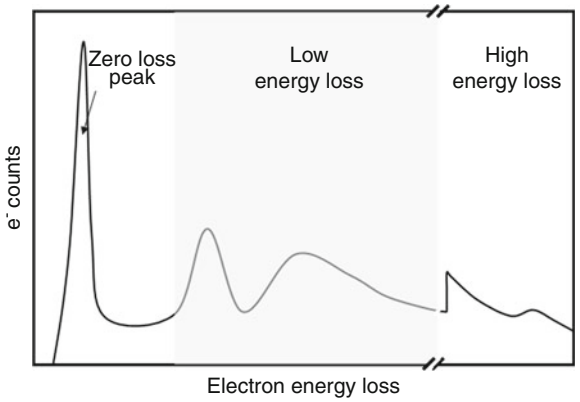
Definition

Electron energy loss spectroscopy (EELS) is a characterization technique to measure kinetic energy change of electrons after inelastic interactions with materials, which provides structural and chemical information of the materials studied.

Scientific Fundamentals

Mechanism

Interactions between the incident electron beams and materials may result in electron energy loss. Because



Electron Energy Loss Spectroscopy (EELS), Fig. 1 Schematic diagram of electron energy loss spectrum showing typical regions of different energy losses

EELS has become a common ancillary technique in transmission electron microscopy (TEM), we will herein mainly focus on TEM-EELS. Although the incident electrons are generally scattered by the total potential of atoms, the scattering process can be further categorized into elastic and inelastic processes depending on whether and how the incident electron responds to the field of the nucleus or to its surrounding electrons. Figure 1 shows an exemplary energy loss spectrum. Generally, the spectrum can be divided into three components according to the origins of the energy loss: zero loss, low energy loss, and high energy loss.

Elastic Scattering

Elastic scattering is a process where an incident electron is scattered by an atomic nucleus. Due to the great mass difference between an electron and a nucleus, the energy exchange between the incident electron and the nucleus is small and usually not measurable in EELS. This process contributes to the zero loss peak of EELS.

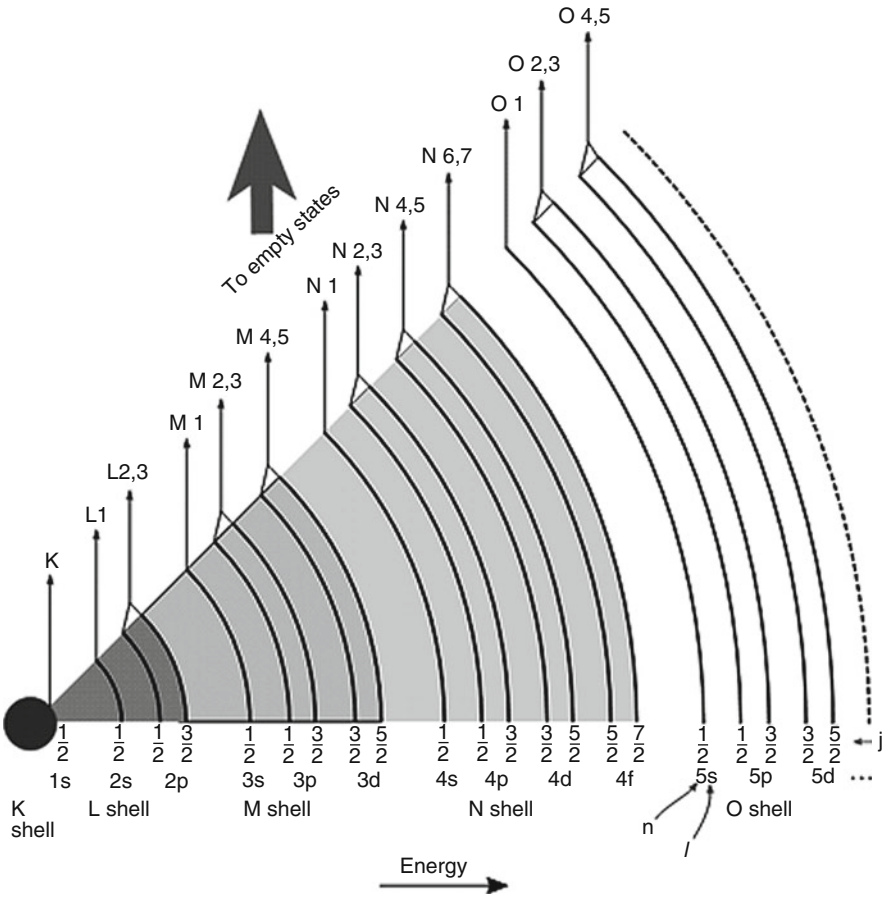
Inelastic Scattering

Inelastic scattering involves the interaction between incident electrons and atomic electrons. Because they have the same mass, the energy exchange is efficient, which means the incident electrons can lose substantial amount of energy during the scattering process. Therefore, the inelastic scattering can be further divided into low energy scattering and high energy scattering or, in other words, outer shell excitation and core shell excitation.

Plasmon excitation is one of the important outer shell inelastic scatterings the incident electrons encounter. It is a collective excitation. The outer shell electrons of the materials are only weakly bound to the atoms, and therefore delocalized. The delocalized electrons are correlated with each other through electrostatic forces. When the incident electrons with sufficient kinetic energy penetrate a solid, the Coulomb interaction between the incident electrons and outer shell electrons will displace the outer shell electrons leaving positively charged holes. The attractive force from these holes will lead to energy loss of the incident electrons. Provided the electron speed exceeds the Fermi velocity, the response of the atomic electrons is oscillatory, resulting in regions of alternating positive and negative space charges (Egerton 2009). Beside the plasmon excitation, there can be single electron excitation in the outer shell excitation. This process usually involves the interband transitions from the valence band to the conduction band. The energy loss due to the outer shell excitation is less than 100 eV.

Compared with the outer shell excitation, the energy loss from the core shell excitation process can be much higher, which is due to the fact that the binding energies of core shell electrons are mostly hundreds or even thousands of electron volts. If sufficient energy from an incident electron is transferred, a core shell electron can be excited from its original core level to an unoccupied state above the Fermi level. Such excitation will appear in EELS as ionization edges, which are superimposed on a background that represents energy loss due to valence electrons. These edges are conveniently classified according to the initial state of the excited electron. It should be pointed out that $L_{2,3}$ edge is in fact a combination of L_2 and L_3. It is called $L_{2,3}$ instead of L_2 and L_3 individually because the energy difference between them is usually too small to be resolved in the spectrum. The same rule applies to $M_{2,3}$, $N_{4,5}$, and so on (Fig. 2). Since core shell electrons are nearly unaffected by the bonding in a solid, and the binding energies of core shell electrons vary in different shells and different elements, such ionization edges can be used to identify the presence of elements in a specimen, which is quite similar to the X-ray energy-dispersive spectroscopy (XEDS).

The core shell electron can receive energy from the incident electron and be scattered out of the initial shell. It can be excited to the vacuum level if getting sufficient energy from the incident electron. However, in many cases, the energy may not be enough to pump the electron to vacuum level. Instead, it may just be scattered to a state in one of a range of possible energy levels above the Fermi level. The final state is determined by the electronic structure of the material. In general, the core-loss intensity $J_c(E)$ is given by an expression known as the Fermi golden rule (Egerton 2009):



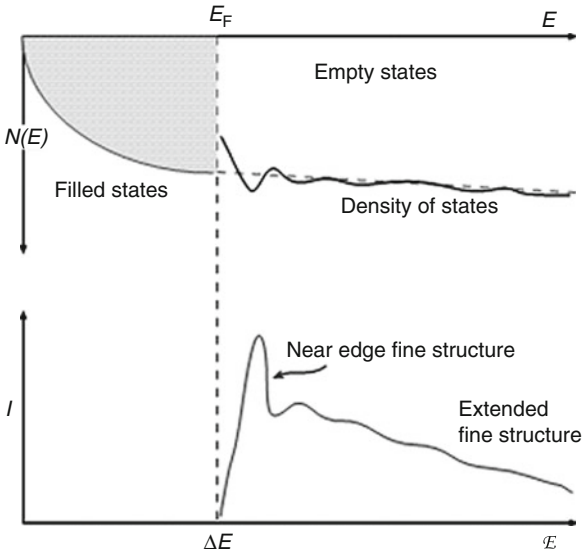
Electron Energy Loss Spectroscopy (EELS), Fig. 2 Full range of possible edges in the electron-loss spectrum due to core shell ionization and the associated nomenclature (Williams and Carter 2009)

$$J_c(E) \propto M(E)^2 N(E) \quad (1)$$

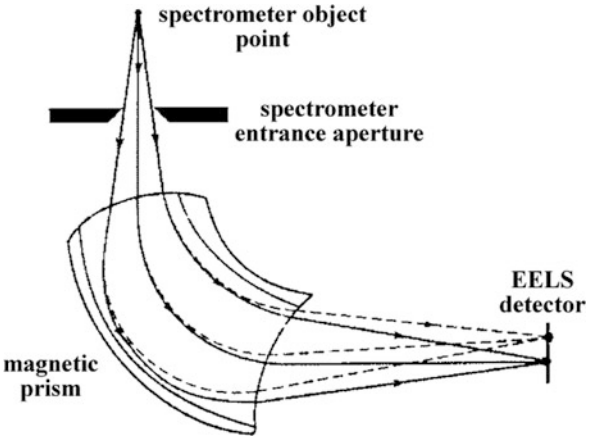
Here, $M(E)$ is an atomic matrix element, and $N(E)$ is the density of final states in the electron transition. Therefore, the density of empty states (DOS) in the conduction band can modulate the core-loss spectrum and lead to variation in the core-loss intensity. This effect can be extended to several tens of eV above the ionization edge onset, which is called energy loss near edge fine structure (ELNES). ELNES reflects the density of states of the materials measured, which is one of the most important applications of EELS in materials science. Figure 3 shows the relationship between the empty DOS and the ELNES. Here, the DOS is atomic position sensitive, which may be called local DOS and is different from the DOS by electrical measurement.

Electron Sources and Detectors

Experimentally, an EELS analysis unit, just like an XEDS analysis unit, is attached to a TEM. The electrons analyzed in EELS are the same as the ones for imaging (i.e., electrons after interaction with the specimen). They are just re-categorized according to their energies. In order to get high energy resolution and signal-to-noise ratio in TEM-EELS analysis, high current density and small kinetic energy distribution are needed for the TEM electron source. The higher current density can insure higher signal-to-noise ratio and the kinetic energy distribution of electrons should be as small as possible to increase the energy resolution. Although LaB_6 and a field emission Schottky gun can offer almost the same beam current, the current density per solid angle from a LaB_6 filament is much lower. As a result, the Schottky gun, which consists of crystallographically oriented tungsten tip coated with ZrO_2 and



Electron Energy Loss Spectroscopy (EELS), Fig. 3 Relationship between the empty DOS and the ELNES intensity in the ionization edge fine structure (Williams and Carter 2009)



Electron Energy Loss Spectroscopy (EELS), Fig. 4 Schematic diagram of EELS spectrometer geometry and the trajectory of electrons when passing through the magnetic prism

exposed to high electric field, suits the EELS application much better. In some advanced analysis, to decrease the energy distribution of the source and hence increase the energy resolution, monochromators can be employed to increase the energy resolution to about 0.2 eV.

One of the important applications of EELS is to analyze composition and electronic structure of a very small area, which is usually done in a dedicated scanning transmission electron microscopy (STEM) or in STEM mode of a TEM. To increase the spatial resolution, which means to acquire a probe size as small as possible, a spherical aberration corrector for the probe forming lens needs be installed. The corrector allows electrons to be focused into a probe with diameter below 0.1 nm, which makes atomic-scale analysis possible.

A spectrometer unit is needed to acquire electron energy loss spectra. The spectrometer is usually placed below the imaging column of a TEM, which is called the post column unit. It can also be positioned inside the TEM imaging column, where the beam is bent into a Ω shape for image stability. The main part of the spectrometer is a magnetic prism, as shown in Fig. 4. Inside the prism, there is a uniform magnetic field generated by an electromagnet. Due to the Lorentz force, electrons passing through the prism are separated and an energy loss spectrum is formed and received by a detector, usually CCD (charge-coupled device), on the energy dispersive plane. A mechanical slit in the energy-dispersive plane can be used for the energy loss selection.

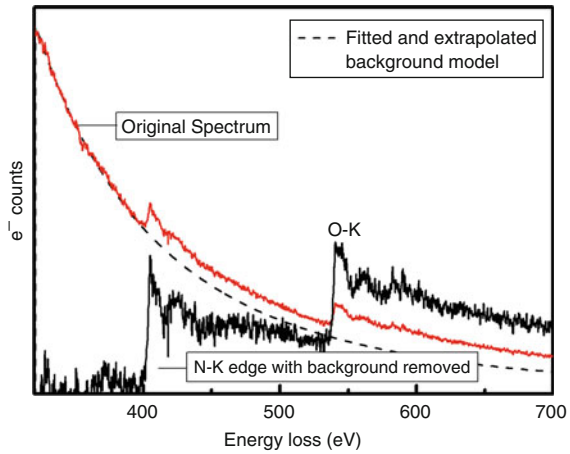
Key Applications

Elemental Identification

As mentioned above, the core loss edges of different elements and different shells of the same element are unique. So the core loss edges can be used for element identification. XEDS and EELS are just different aspects or signals stemming from the same process. The atomic electrons get sufficient energy from the incident electrons and are excited above the Fermi level. The excited atom is not stable, and when it comes back to the ground state, X-ray photons or Auger electrons will be emitted, where the former can be used for XEDS and the latter for Auger spectroscopy, while the transmitted electrons are for EELS. Therefore, both the X-ray photons and the energy loss electrons can be used to identify the atoms excited. In many cases, EELS and XEDS are complementary techniques. Light elements usually have low X-ray yield and the associated low energy X-ray can be easily absorbed even in very thin specimens, which makes the XEDS analysis difficult for light elements. On the other hand, EELS depends on the electrons exiting from the specimen and their energy losses, having no direct correlation with X-ray emission or absorption, making it a suitable technique for light element analysis. Furthermore, due to the isotropic nature of X-ray emission, an XEDS detector can only collect about 1% of the emission. In contrast, due to the forward-scattering nature of the electrons, a collection

efficiency of higher than 50% can be easily reached in the EELS analysis by using a large enough collection angle β. If higher spatial resolution is needed, one can use a smaller collection angle to decrease delocalization effect, with which an atomic-scale resolution can be achieved.

For quantitative analysis, EELS is much more complicated than XEDS. A smaller convergence angle α is usually needed to make sure the spectrum collection angle is larger than the convergence angle ($\alpha < \beta$). Once a spectrum is acquired, the quantification procedure involves background subtraction, deconvolution to remove the plural scattering effect, and edge integration. Some other parameters, such as partial ionization cross section (σ), need to be treated before the quantification process. Partial ionization cross section in EELS analysis is derived from theoretical calculation. The programs of SIGMAK and SIGMAL developed by Egerton are good enough for most applications, even though the model seems oversimplified (Egerton 1996). For details of EELS quantification analysis, one can refer to D.B. Williams and C.B. Carter's book, which explains the analysis in more detail.

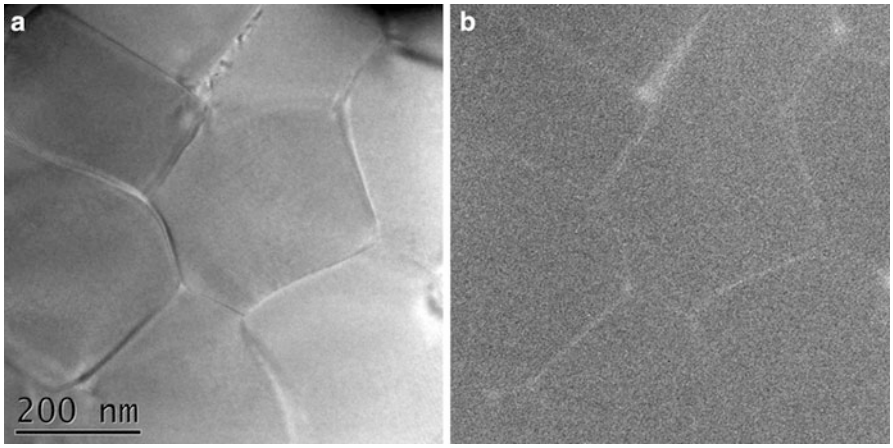


Electron Energy Loss Spectroscopy (EELS), Fig. 5 Electron energy loss spectra presenting nitrogen-N and oxygen-K edges and the background model fitted spectra to reveal more pronounced nitrogen-N and oxygen-K edges (Courtesy of Dr. Y. Li, University of California, Davis)

Example. Figure 5 presents electron energy loss spectra of N and O K-edges taken from an aluminum alloy sample. The signal from N, which is usually difficult to be detected in XEDS analysis, is very obvious, especially after background subtraction. The spectrum with background removed was rescaled to fit the dimension of the figure.

Energy Filtered TEM (EFTEM)

EFTEM is another powerful EELS application. By utilizing the technique, an elemental distribution map can be acquired. Either STEM mode or regular fixed-beam TEM mode can be used for EFTEM operation. In STEM mode, a focused beam scans the region of interest point by point with high energy resolution, but requires long scan time, usually minutes or even hours. In fixed-beam TEM mode, the EFTEM scan takes only seconds. The fixed-beam EFTEM, therefore, is more commonly used in practice, which generally involves two approaches: the two-window method and the three-window method.



Electron Energy Loss Spectroscopy (EELS), Fig. 6 (**a**) Zero loss image of nanocrystalline NdFeB magnet; (**b**) Nd N jump ratio image

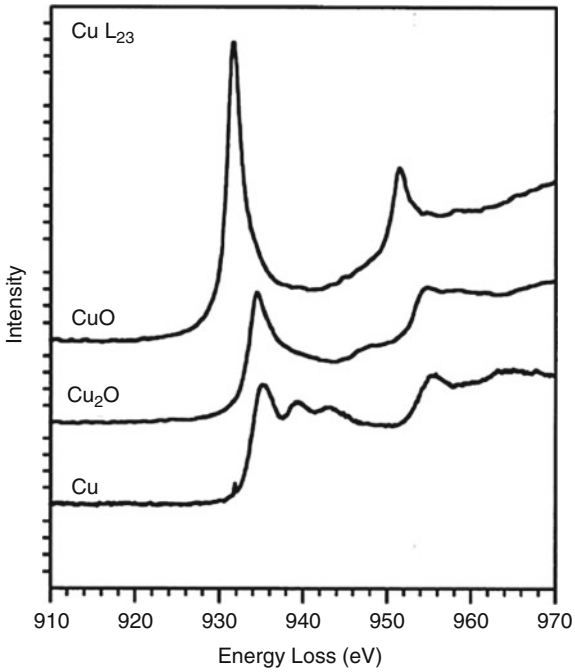
For each method, energy-filtered images need to be taken firstly. The energy-filtered image is achieved by shifting the energy spectrum until the desired energy window passes through the energy selecting slit. The energy shift is achieved by changing the accelerating voltage of the TEM so that electrons of different energies stay on-axis and thus in focus through the spectrometer. In the two-window method, two images, one pre-edge and one post-edge, are taken. An intensity ratio between these two images is then made and displayed. This method is also called the jump ratio method, and the associated analysis is qualitative. For the three-window method, three filtered images are first taken, including two pre-edge and one post-edge images. The pre-edge images are used to calculate the background, usually by employing a power law model, $I = CE^{-s}$, where I is the intensity, E is the energy loss, and C and s are fitting parameters. The fitted background is then subtracted from the post-edge image. Although the three-window method is quantitative, the extrapolated background is usually noisy. The jump ratio method is less susceptible to diffraction effects and can give a less noisy image.

Example. For a NdFeB magnet, if the hard magnetic $Nd_2Fe_{14}B$ grains are separated by a thin layer of Nd – rich phase, the magnetic properties can be increased. The thin layer is about 2 nm thick or less, which makes it difficult to be detected by XEDS. By employing Nd $N_{4,5}$ edge, EFTEM (jump ratio image), however, clearly shows the Nd enrichment in the grain boundary region (Fig. 6b). To avoid artifacts due to thickness difference, the sample was prepared using focused ion beam (FIB).

ELNES

ELNES depends on every detail of the local atomic environment such as coordination, valence state, and type of bonding. Measurement of ELNES gives information of the local electronic structure and is, therefore, helpful in understanding the properties of materials. In the last few years, great progress has been made in the theoretical calculation of unoccupied DOS thanks to the improvement in models of atomic potentials and the computing power, which in turn helps understand or predict features of ELNES.

Example. For transition metals, in the electron energy loss spectra, there are sharp lines corresponding to the transitions to the partially unoccupied d-band. These sharp lines are called white lines, referring the strong signal intensity. For Cu, because its 3d states are all filled, there is no white line in the spectrum as shown in Fig. 7. In CuO, O removes electrons from 3d band of Cu, and the white line appears. Cu_2O has a full 3d band, the same as Cu, and we might not expect a white line. However, the white line appears in Cu_2O, which is due to the hybridization of the Cu and O states that changes the shape of the d band.



Electron Energy Loss Spectroscopy (EELS), Fig. 7 The $L_{2,3}$ edge of Cu, *CuO*, and *Cu_2O*. The spectra are shown offset from one another in the vertical direction for clarity (Reprinted with permission from Keast et al. 2001)

Cross-References

- ► Auger Electron Spectroscopy (AES)
- ► Scanning Electron Microscopy (SEM)
- ► Transmission Electron Microscopy (TEM)
- ► X-Ray Photoelectron Spectroscopy (XPS)

References

R. Egerton, *Electron Energy Loss Spectroscopy in the Electron Microscope*, 2nd edn. (Plenum Press, New York, 1996)

R. Egerton, Electron energy loss spectroscopy in the TEM. Report on Progress in Physics, 016502, 2009

V.J. Keast, A.J. Scott, R. Brydson, D.B. Williams, J. Bruley, Electron energy-loss near-edge structure- a tool for the investigation of electronic structure on the nanometer scale. J. Microsc. **203**, p135 (2001)

D.B. Williams, C.B. Carter, *Transmission Electron Microscopy: A Textbook for Materials Science* (Springer, New York, 2009)

Electroplating

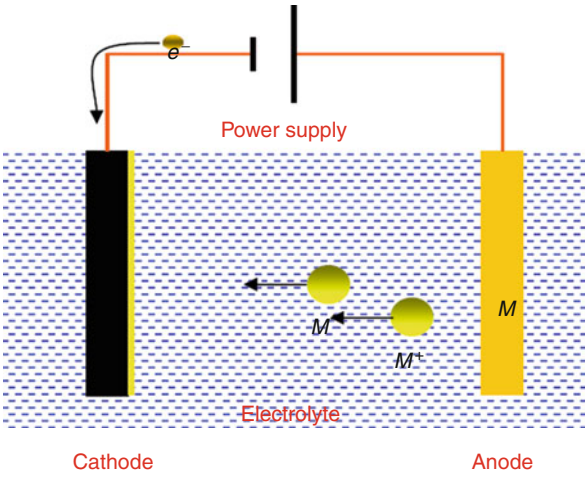
Chang-Dong Gu[1], Tong-Yi Zhang[2]
[1]Department of Materials Science and Engineering, Zhejiang University, Hangzhou, People's Republic of China
[2]Department of Mechanical Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong SAR, People's Republic of China

Synonyms

Electrochemical plating; Electrocrystallization; Electroplating process

Definition

Electroplating is defined as the deposition process of a metallic coating upon an electrically negative charged object by using an external electrical current to reduce metallic ions in electrolyte to metallic atoms, as illustrated in Fig. 1. The object to be coated is called the *cathode*. To complete the electrical circuit, another electrode called the *anode* is connected to the positive terminal of a power supply of direct current (DC), pulsed-current, or DC-predominated current profile. Both cathode and anode are immersed in the electrolyte, which contains the metallic ions to be coated and, usually, other types of dissolved metallic ions and complexing agents. When a DC-predominated current is applied, metallic atoms at the anode, if soluble, are oxidized to metallic ions in the electrolyte. The metallic ions react with the complexing agents in the electrolyte to form complex ions, which may play a very important role in electroplating (Parthasaradhy 1989). At the cathode, the dissolved metallic ions in the electrolyte are reduced, and thus results in metal deposition. If the anode is not soluble into the electrolyte, it is a non-consumable anode such as platinum in most electrolytes. The metallic ions being plated must be replenished periodically in the electrolyte bath as electroplating is going on. In practice, electroless plating, which is a parallel coating technique with electroplating and may be considered the sister coating technique of electroplating, is also generally adopted to produce high corrosion and wear resistant coatings for workpieces.



Electroplating, Fig. 1 Sketch of the electroplating process. Driven by an applied DC-predominated current, metallic ions (M^{Z+}) are dissolved from the anode (M), move in electrolyte to the cathode (object), and are reduced to metallic atoms as a coating on the object

Scientific Fundamentals

Faraday's Law of Electrolysis

In the electroplating process with a soluble metallic anode, metal is deposited on the cathode while the same kind of metal is being dissolved from the anode. The extent of deposition or dissolution is determined by the quantity of electricity passed. The relationship between the two is given by Faraday's law of electrolysis, which is commonly stated as follows:

1. The mass of a substance, m, altered at an electrode during electrolysis is directly proportional to the quantity of electric charge, Q, transferred at the electrode.
2. For a given quantity of electric charge, the mass of an elemental material, m, altered at an electrode is directly proportional to the element's equivalent weight. The equivalent weight of a substance is its molar mass M divided by the valence number, z, of ions of the substance.

Faraday's law of electrolysis is thus given by

$$m = \frac{Q}{F} \cdot \frac{M}{z}, \tag{1}$$

where $F = 96,485\ \mathrm{C\ mol^{-1}}$ is the Faraday constant. Note that M/z is the same as the equivalent weight of the substance altered. Furthermore, the total charge Q is the electric current $I(t)$ integrated over time t. For an electrolysis time t', Faraday's law of electrolysis takes the form

$$m = \frac{\int_0^{t'} I(t)\,dt}{F} \cdot \frac{M}{z}. \tag{2}$$

With Faraday's law of electrolysis, electroplaters are able to work out the weight of deposit, the thickness of deposit, and the duration of electroplating.

Standard Electrode Potential and Electrochemical Series

When a metal is immersed in a solution of its own ions (e.g., copper in copper sulfate solution), which is called a *half-cell* in electrochemistry, two opposite reactions occur at the interface between the solution and the metal. One is the ionization of copper and the other is the deposition of copper, which can be shown as the following equations:

$$Cu \rightarrow Cu^{2+} + 2e^{-}, \tag{3a}$$

$$Cu^{2+} + 2e^{-} \rightarrow Cu, \tag{3b}$$

where e^{-} denotes electron and the valence number is $z = 2$. Equation (3a) states an electrochemical reaction called the *oxidation reaction*, in which a metallic atom loses electrons and becomes a positively charged ion, whereas (3b) states the reduction reaction of Cu, in which the Cu^{2+} ion gains electrons and returns a Cu atom. More general forms of oxidation and reduction reactions can be expressed by

$$M \rightarrow M^{z+} + ze^{-}, \tag{4a}$$

$$M^{z+} + ze^{-} \rightarrow M. \tag{4b}$$

Consider an oxidation reaction. When metallic atoms are oxidized to ions and dissolved into electrolyte, the lost electrons remain in the anode. The positively charged ions, called cations, and negatively charged electrons generate an internal electric field, which is against the oxidation reaction. That is why the reduction reaction occurs simultaneously. Eventually, the two reactions reach the equilibrium state at which the rate of metallic atoms dissolved into the solution as ions equals the rate of metallic ions deposited back as metallic atoms. At equilibrium, a *double layer* is formed at the interface between anode and electrolyte, which may be visualized as a layer with excess positive cations in the electrolyte confronting another layer of an equal number of negatively charged electrons (or anions) in the anode. Thus, there is an electric potential jump crossing the interface. The potential jump is called the *electrode potential*, representing the degree of ease with which a metallic material will be oxidized into ions. If the potential jump is measured at the equilibrium state, it is called the *reversible potential* or *equilibrium potential*. If a half-cell, similar to the one described for Cu, consists of a pure metal electrode immersed in a 1M solution of its ions and at 25°C, it is called a *standard half-cell*. However, an electrochemical cell is needed to measure *electrode potential* with a potentiometer, which ensures no electric current flowing through the cell. In electrochemistry, a reference half-cell is established, to which other cell halves may be compared. This reference half-cell is the *standard hydrogen electrode* (SHE). It consists of an inert platinum electrode in a 1M solution of H^{+} ions, saturated with hydrogen gas that is bubbled through the solution at a pressure of 1 atm. and a temperature of 25°C. The platinum itself does not take part in the electrochemical reaction; it acts only as a surface on which hydrogen atoms may be oxidized or hydrogen ions may be reduced. The *standard electrode potential*, E^{o}, is the measure of the other standard half-cell individual potential by coupling to the standard hydrogen electrode, which is called the electromotive force (EMF). It is very useful to list standard potentials in order for most commonly used metals, and such a list is called the EMF series or electrochemical series. Furthermore, reduction reactions are usually taken in the expression of EMF series. Table 1 is the EMF series, showing that the most negative E^{o} values are placed at the top of the EMF series and the most positive at the bottom. The more negative the EMF is, the easier the metal will be oxidized, or the easier the metal loses electrons. In electrochemistry, it is said that the more negative the EMF is, the more active (or anodic) the

Electroplating, Table 1 Standard electrode potentials

Electrode reaction at 298 K	Standard potential ($E^{o}_{M^{z+}/M}$, volts)
$Li^{+} + e^{-} \rightarrow Li$	−3.045
$K^{+} + e^{-} \rightarrow K$	−2.925
$Ca^{2+} + 2e^{-} \rightarrow Ca$	−2.87
$Na^{+} + e^{-} \rightarrow Na$	−2.714
$Mg^{2+} + 2e^{-} \rightarrow Mg$	−2.73
$Al^{3+} + 3e^{-} \rightarrow Al$	−1.66
$Zn^{2+} + 2e^{-} \rightarrow Zn$	−0.763
$Fe^{2+} + 2e^{-} \rightarrow Fe$	−0.440
$Pb^{2+} + 2e^{-} \rightarrow Pb$	−0.126
$2H^{+} + 2e^{-} \rightarrow H_2$	0
$Cu^{2+} + 2e^{-} \rightarrow Cu$	0.337
$Ag^{+} + e^{-} \rightarrow Ag$	0.7991
$Au^{3+} + 3e^{-} \rightarrow Au$	1.498

metal will be. On the other hand, the more positive the EMF is, the more difficult the metal will be oxidized, or the more difficult the metal loses electrons. In electrochemistry, it is said that the more positive the EMF is, the more inert (or cathodic) the metal will be. When a half-cell is electrically connected with another half-cell to form an electrochemical cell, the metal with relatively lower EMF will be dissolved into electrolyte, whereas the other metal with relatively higher EMF will be deposited. That is why a metal with a more negative value of E^{o} is called a stronger reducing agent, while a metal with a more positive value of E^{o} is called a stronger oxidizing agent.

Nernst Equation and Polarization

The EMF series applies to highly idealized electrochemical cells or standard cells. Altering temperature or/and solution concentration will change the cell potential. Consider the reduction reaction given by (4b). Letting $a_{M^{z+}}$ and a_M denote the activities of M^{z+} and M, respectively, the redox potential ($E_{M^{z+}/M}$) of the half-cell is given by the *Nernst equation*:

$$E_{M^{z+}/M} = E^{o}{}_{M^{z+}/M} + \frac{RT}{zF}\ln\frac{a_{M^{z+}}}{a_M}, \qquad (5a)$$

where $E^{o}{}_{M^{z+}/M}$ is the standard potential of the redox couple, the value of which is listed in Table 1, $R = 8.31$ J/(mol K) is the *gas constant*, and T is the absolute temperature. The activity a_M of pure metal is usually taken to be unity so that (5a) is reduced to

$$E_{M^{z+}/M} = E^{o}{}_{M^{z+}/M} + \frac{RT}{zF}\ln a_{M^{z+}}. \qquad (5b)$$

During electroplating, an electric current goes through the electrochemical cell. In this case, the potentials of two electrodes will not be at the equilibrium values listed in Table 1 because the system is now a non-equilibrium one. The displacement of each electrode potential from its equilibrium value is termed *polarization* and the magnitude of this displacement is called the *overvoltage* or *overpotential*. In electroplating, an applied current induces an overvoltage, which makes the cathode potential more negative and the anode potential more positive compared with the equilibrium values, respectively. Polarization is an unavoidable phenomenon in the electroplating process. Based on the mechanism, polarization can be classified as activation polarization, concentration polarization, and ohmic drop. An electroplating process contains a sequence of steps that occur in series at the interface between the metal electrode and the electrolyte solution. Among the steps there is a bottleneck step (i.e., the step with the lowest rate), which predominantly controls the reaction process. Activation polarization refers the polarization caused by the bottleneck step and the term *activation* is applied here because an activation energy barrier is associated with the slowest rate step. Concentration polarization is caused by diffusion in solution, which limits the reaction rate. Ohmic drop is due to the electric resistance of the electrochemical cell. The three polarization mechanisms are associated with three types of overvoltage, activation overvoltage, concentration overvoltage, and ohmic overvoltage. In electroplating, the activation polarization can be changed by varying electroplating temperature, concentration polarization can be reduced by stirring the electrolyte solution, and ohmic drop can be lowered by changing the electric conductivity of the solution. Electrochemical analysis provides the relations of activation overvoltage, concentration overvoltage, and ohmic overvoltage with electric current density. Controlling the current density is more closely linked with the cathode polarization and is not greatly influenced by large changes in the electrode shape and spacing (Parthasaradhy 1989).

Key Applications

Surface Modification and Coating Technology

Electroplating is one of the most cost effective and simple coating techniques for altering the surface properties of a workpiece in order to improve its corrosion and wear resistance, solderability, electrical conductivity, or decorative appearance. For example, gold electroplating is widely used in electronics for making contacts and connectors. In some cases, the gold plating is used for decorative purposes in jewelry and watch cases. Nickel and nickel-based alloys are widely plated metals in the plating industry, for both decorative and protective applications. Electroplating nickel-based composite coatings, which are structured by fine inert particles, such as WC and SiC, being dispersed in the electroplated nickel, confer the plated workpiece with excellent wear and abrasion resistance. During the electroplating of metal-based composite coatings, small inert particles are dispersed in electrolytes and kept in suspension with forced convection for the metal deposition. Hard chromium plating involves deposition of a thick coating of chromium directly over the substrate, which confers to the plated parts a combination of physical and mechanical properties such as high hardness, abrasion resistance, low coefficient of friction, and good corrosion resistance. Electroplating zinc coating is usually used for protecting steel or iron parts from corrosion due to its

lower standard electrode potential than iron and steels. Therefore, the zinc deposit is also called a *sacrificial coating*, meaning that it corrodes in preference to the coated materials when exposed to a corrosive atmosphere. Electroplated alloy-coatings are also in demand to improve the coating's physical and chemical properties. For example, electrodeposited Ni-Co alloys have been thoroughly investigated as important engineering materials for several decades due to their unique properties, such as high strength, good wear resistance, heat conductivity, electrocatalytic activity, and specific magnetic properties.

Synthesis of Nanocrystalline Metals by Electroplating Method

Nanocrystalline materials are polycrystalline materials with grain size smaller than 100 nm and thus a large volume fraction of grain boundaries. Nanocrystalline materials have experienced rapid development in recent years due to their existing and potential applications in a wide variety of technological areas such as electronics, catalysis, ceramics, magnetic data storage, and structural components. Porosity-free and high-purity nanocrystalline materials are the ideal candidates for exploring the deformation mechanism of low-dimension materials and being the building blocks of microelectromechanical/nanoelectromechanical systems. The electroplating technique is a versatile bottom-up method with a simple operation for manufacturing nanocrystalline materials (Erb 1995). Adjusting electroplating parameters allows one to synthesize different nanocrystalline metals/alloys with specialized microstructures (Erb 1995; Gu et al. 2006, 2007). The electroplating techniques applied to synthesize nanocrystalline structures include conventional DC electroplating and pulsed-current deposition (Erb 1995). Increasing the cathode polarization induces a high crystal nucleation rate during deposition, resulting in finer-grained deposits. Practically, compact, smooth, adherent, and bright deposits are obtained only at relatively moderated values of electrocrystallization overpotentials and by using well-established concentrations of proper addition agents known as levelers and/or brighteners. For example, 2-butene-1,4-diol (also called 1,4-butenediol) is an effective grain refiner (Gu et al. 2007), which is usually adopted in the conventional nickel plating bath as a brightener additive. Interestingly, a layered nanostructured nickel sample, which consists of alternate layers of ultrafine and nano-sized grains, was fabricated by periodically replenishing the grain refiner of 2-butene-1,4-diol during electroplating nickel process (Gu et al. 2007). Nanocrystalline nickel also exhibits enhanced localized corrosion resistance because of the large volume fraction of grain boundaries that provide an increased number of preferential attack sites and therefore disperse the corrosion current (Erb 1995; Gu et al. 2006). Thus, electroplating nanocrystalline nickel coatings on magnesium alloys, which intrinsically are active in chemistry and poor in wear resistance, is an effective method to extend their applications in various industries (Gu et al. 2006).

Electroplating in Advanced Electronic Packaging

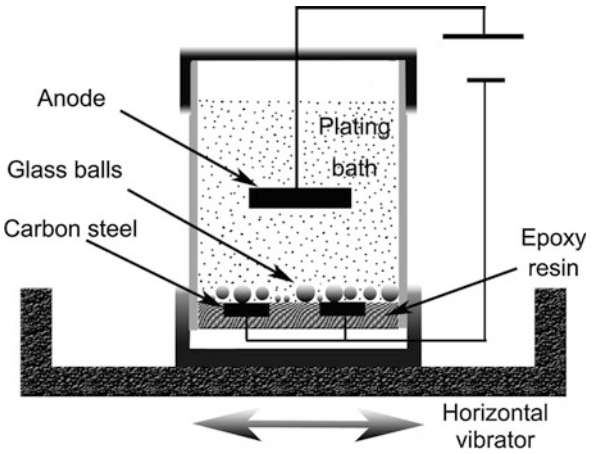
In the electronic industry, the manufacturing of semiconductor devices requires a lead frame as a support for mounting semiconductor chips and as an electrical connection between the chip and the wiring board. The widely used preplated lead frames (PPFs) comprise a copper (Cu) alloy base electroplated with nickel (Ni), palladium (Pd), and gold (Au) sequentially. The Ni layer provides the lead frame with oxidation and corrosion resistance by impeding Cu out-diffusion to the lead frame surface because Cu out-diffusion to and oxidation at the lead frame surface is the most severe failure mechanism of PPFs. The ultrathin Pd and the top Au flash facilitate wire bonding and soldering and provide protection to the underlying Ni layer from oxidation. The technical challenge is how to reduce the thickness of the Au/Pd/Ni layers while maintaining or even enhancing the PPF quality and reliability. An electrodeposition technique has been developed that enables the production of low-defect, continuous, and epitaxy-like Au/Pd layers atop the Ni layer (Liu et al. 2006). The thicknesses of Pd and Au layers are reduced to 10 and 3 nm, respectively (Liu et al. 2006). Introducing a nanometer-thick intermetallic-compound (IMC) layer is an effective means to meet such a requirement. Fu et al. (2006) developed an innovative electroplating technique, by which a nano-Sn layer is introduced in between the Cu base and the Ni layer to form a Cu-Ni-Sn IMC interlayer. The resultant IMC layer is a more effective diffusion barrier than the bare Ni layer; thereby the oxidation and corrosion resistance of the PPFs is boosted (Fu et al. 2006).

Furthermore, three-dimensional (3D) interconnection techniques are being developed to meet the size-reduction challenge of electronic devices, in which the through-wafer interconnect technique is one of the key technologies. Generally, vertical through-silicon holes are fabricated using the deep reactive ion etching (DRIE) technique. Then, the holes are filled with conductive materials using electrodeposition or/and other deposition techniques. Compared with other deposition techniques,

the electroplating technique is cheaper, faster, and easier in operation at lower temperatures. Since local current distribution does not remain uniform in very deep and narrow through-holes, electroplating void-free copper vias with a high aspect ratio of depth to diameter becomes a challenge to the electronic packaging industry. An approach to solve this challenging problem is to control the local current density distribution. Dixit and Miao developed a novel "aspect-ratio-dependent electroplating technique" to fabricate copper vias, where the electroplating parameters, such as forward current, reverse current, and pulse-on time, were continuously varied along with the change in the unfilled via depth (Dixit and Miao 2006). Varying the current profiles was believed to improve the local distribution of current as per the changing depth and help in avoiding void formation (Dixit and Miao 2006). Recently, Gu et al. demonstrated a simple reverse-pulse electroplating copper technique to fill the tapered through-wafer holes having high aspect ratio (Gu et al. 2009). Clearly, the geometry of tapered holes would have much influence on the local current density distributions. The diameter of the hole gradually becomes larger, making the actual current density decrease when the applied current is maintained unchanged. This indicates that the tapered high aspect ratio through-wafer holes could automatically adjust the local current density distribution during the electroplating process. Tapered through-wafer holes have two advantages in the electroplating copper process. Firstly, the taper profile could increase the Cu^{2+} flux and improve the Cu^{2+} diffusion in the holes. Secondly, air bubbles at the surface of cathode, such as hydrogen, could easily escape through the wide mouth of the taper holes. Air bubbles are the main cause for the formation of voids in the electrodeposited copper interconnects.

Mechanically Assisted Electroplating of Ni-P Coatings on Carbon Steel

Numerous mechanical and physical methods have been attempted to enhance electroplating, including using rotating electrodes, laser plating, jet plating, electroplating accompanied by controlled abrasion, and ultrasonic wave-assisted electroplating. A mechanically assisted electroplating technique, which combines the conventional electroplating technique with mechanical ball-rolling, was demonstrated by (Ping et al. 2008). The set-up of mechanically assisted electroplating Ni-P alloys is schematically shown in Fig. 2. As shown in the figure, the plating bath is placed on the top of a vibrator. The vibrator provides a sinusoidal vibration of 1 mm amplitude and 4.5 Hz frequency in the horizontal direction. Two



Electroplating, Fig. 2 Schematic set-up of mechanically assisted electroplating Ni-P alloys (Ping et al. 2008)

carbon steel samples are mounted in epoxy resin as a whole and work as the cathode. After installing the cathode, glass balls with density of 2.2 g/cm^3 and diameters (Φ) of 1, 5, and 7 mm (many balls of Φ1 with a total weight of 12 g, 10 balls of Φ5, and 5 balls of Φ7) are dispersed on the cathode surface. About one third of the surface area of each sample is covered by the glass balls. When the plating bath is vibrated, these glass balls roll horizontally back and forth and provide mechanical ball-rolling to coatings simultaneously during the plating process. The anode is a high-purity nickel ingot and parallel to the cathode surface, as illustrated in Fig. 2. The main ingredients in the electrolyte solution are $NiSO_4$ 250 g/L, $NiCl_2{\cdot}6H_2O$ 15 g/L, NaCl 15 g/L, H_3BO_3 30 g/L, H_3PO_3 6 g/L, and $NaH_2PO_2{\cdot}H_2O$ 20 g/L. The plating temperature is maintained at 70°C. The present electroplating technique with mechanical assistance is different from the electroplating technique with controlled abrasion. The abrasion is conducted out of the electrolyte bath, while the mechanical ball-rolling is simultaneously carried out within the electrolyte bath by vibrating samples being plated with glass balls on it. The mechanically assisted electroplating technique has been applied to fabricate Ni and Ni-P coatings, which possess smooth surfaces, refined grains, increased hardness, and excellent corrosion resistance (Ping et al. 2008). Furthermore, it is found that the mechanically assisted, electroplated, high phosphorus Ni-P coatings are crystalline, while the conventionally electroplated Ni-P coatings with the same phosphorus content are a mixture of amorphous and microcrystalline Ni. The Ni-P coatings electroplated with mechanical

assistance are superior to the conventionally electroplated Ni-P coatings in microhardness and corrosion resistance (Ping et al. 2008).

Cross-References

► Duplex Coatings
► Electrochemical Deposition
► Multiplex Coatings
► Nanocomposite coatings
► Surface Nanocrystallization and Hardening (SNH)

References

P. Dixit, J.M. Miao, Aspect-ratio-dependent copper electrodeposition technique for very high aspect-ratio through-hole plating. J. Electrochem. Soc. **153**(6), G552–G559 (2006)
U. Erb, Electrodeposited nanocrystals: synthesis, properties and industrial applications. Nanostruct. Mater. **6**, 533–538 (1995)
R. Fu, L.L. Liu, D.M. Liu, T.Y. Zhang, In situ formation of Cu-Sn-Ni intermetallic nanolayer as a diffusion barrier in preplated lead frames. Appl. Phys. Lett. **89**(13) (2006)
C.D. Gu, J.S. Lian, J.G. He, Z.H. Jiang, Q. Jiang, High corrosion-resistance nanocrystalline Ni coating on AZ91D magnesium alloy. Surf. Coat. Technol. **200**(18–19), 5413–5418 (2006)
C.D. Gu, J.S. Lian, Q. Jiang, Layered nanostructured Ni with modulated hardness fabricated by surfactant-assistant electrodeposition. Scr. Mater. **57**(3), 233–236 (2007)
C.D. Gu, H. Xu, T.Y. Zhang, Fabrication of high aspect ratio through-wafer copper interconnects by reverse pulse electroplating. J. Micromech. Microeng. **19**(6) (2009)
L.L. Liu, D.M. Liu, R. Fu, Y.F. Kwan, C.H. Yau, T.Y. Zhang, Epitaxy-like protective layers for high-performance low-cost Au/Pd/Ni preplated Cu alloy leadframes. IEEE Trans. Adv. Packag. **29**(4), 683–689 (2006)
N.V. Parthasaradhy, *Practical Electroplating Handbook* (Prentice Hall, Englewood Cliffs, 1989)
Z.X. Ping, Y.D. He, C.D. Gu, T.Y. Zhang, Mechanically assisted electroplating of Ni-P coatings on carbon steel. Surf. Coat. Technol. **202**(24), 6023–6028 (2008)

Electroplating for Self-Lubricating Metal Composite Coatings

► Self-lubricating Metal Composite Coatings by Electrodeposition or Electroless Deposition

Electroplating Process

► Electroplating

Electrostatic Condition Monitoring

► Bearing Wear Monitoring Using Electrostatic Charge

Electrostatic Field Effects on Adhesion

Yonggang Meng
State Key Laboratory of Tribology, Tsinghua University, Beijing, People's Republic of China

Synonyms

Adhesion and electrostatic field; Electroadhesion; Sticking

Definition

Electrostatic forces in electrical double layers involved in adhesion of solids.

Scientific Fundamentals

Adhesion Phenomena

Adhesion is a phenomenon where a force or energy is required to separate two dissimilar bodies that are already in contact, in the normal direction of the contact surface. The maximum force needed in the separation process is referred to as adhesion force, and the energy required during the separation process is regarded as the work of adhesion. As a matter of fact, adhesion exists not only in the separation or detaching process of a contact, but in the formation of a contact as well. Owing to the tremendous difficulty in measurement of interaction forces between two attached bodies, when we talk about experimental adhesion force, in general, it means the value measured in separating a solid bond, which is probably not the same as that in contact or in the process of forming a bond.

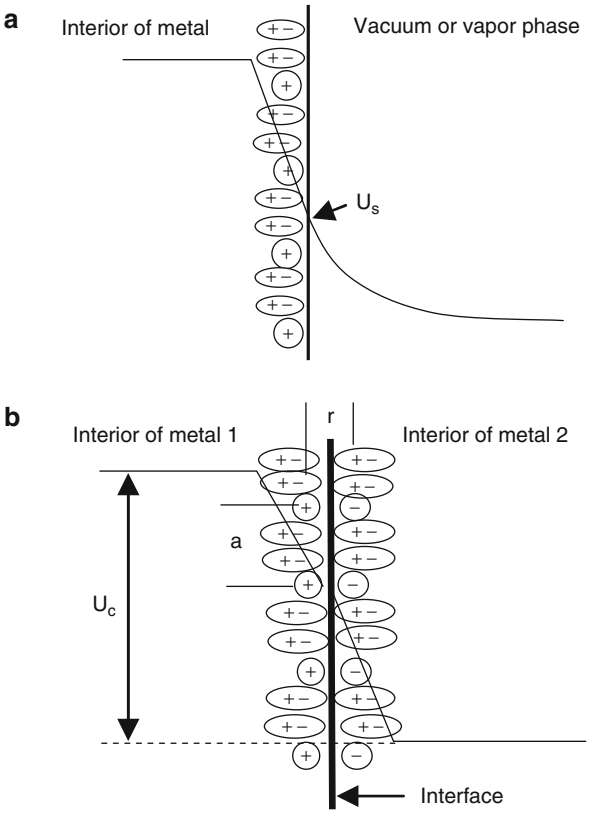
There are four types of basic interactions contributing to adhesive force: chemical bonds, intermolecular bonds, electrostatic force, and meniscus force. Chemical bonds, including ionic, covalent, and metallic bonds, have the strongest bond energy and the shortest interaction range (about 0.15–0.3 nm) among the basic interactions. The hydrogen bond is one of the chemical bonds; its bond energy is relatively small. Intermolecular bonds are also known as the van der Waals force or Lifshitz–van der Waals force, which includes three components: dispersion

(London effect), dipole–dipole (Keesom effect), and dipole-induced dipole (Debye effect). Intermolecular forces have an interaction range as far as 10 nm or greater (with retardation effect) and exist everywhere, whether chemical interactions appear or not. When two contacting surfaces have the same or opposite net charges, there is an electrostatic force acting between them, which contributes to adhesive force. It is unclear which mechanism, electrostatic force or intermolecular force, is the predominant one in adhesion. Derjaguin and colleagues argued that the electrostatic effect plays an extremely important and general role in adhesion phenomena (Derjaguin et al. 1978). This viewpoint is still the subject of debate (Kinloch 1980). For instance, the movement of geckos on walls and ceilings is explained and modeled fully by the van der Waals force (Tian et al. 2006). Nevertheless, there is no doubt that electrostatic forces make a significant contribution to the work of adhesion due to their long-range characteristics. The meniscus force presents when the contacting surfaces of solids are fully or partially separated by an interlayer of liquid and surrounded by the vapor of the liquid. The meniscus force may be attractive or repulsive depending on the contact angles of the liquid on the solids. Detailed descriptions of the mechanisms of adhesion are available from a variety of reference materials, including references (Comyn 1997; Lee 1991).

Electrical Double Layer and Electrostatic Forces

In most of the cases of adhesion of solids, electrostatic forces are generated from the electrostatic field within the electrical double layer, which is formed at an interface due to charge exchange and redistribution during the attachment of two bodies. The formation of electrical double layers depends on the electrical properties of the mating surfaces.

For metal/metal contacts, charge exchange and electrical double layer formation can be described with the mechanism of electron tunneling from one metal to the other. Metals (excluding amorphous metals) have crystalline structures and common electron gas. At 0 K temperature, free electrons occupy energy levels up to the Fermi energy, E_f. When a free electron with Fermi energy removes from inside the metal to infinity place, where the potential of the electron is assumed to be zero, an energy equal to the work function ϕ_m, which is a characteristic property for the individual crystalline plane of each metal, is required. If the metal surfaces trap some amount of net charges, decay electrostatic fields exist outside the surfaces. Let U_s denote the surface potential of the electrostatic fields (see Fig. 1a). Thus, the total energy



Electrostatic Field Effects on Adhesion, Fig. 1 Potential distributions around (**a**) an isolated metal surface and (**b**) a metal/metal interface

of an electron at the Fermi energy relative to infinity is equal to $\varphi_m = -\phi_m - eU_s$, which is referred to as the electrochemical potential of an electron inside metal. When two unlike solids, metal 1 and metal 2, are brought near each other, the electrons in the side with the higher electrochemical potentials have the tendency to flow into the other side once the separation between the two sides is less than the tunneling distance. As a result, the surface losing the electrons is in excess of positive charge, and the surface accepting the electrons is negatively charged. The charge exchange does not cease until the electrochemical potentials of electrons at the interface reach the same level, i.e.,

$$-\phi_{m1} - eU_{s1} = -\phi_{m2} - eU_{s2} \quad (1)$$

Finally, an electrical double layer is formed by the positive charges on one side and the negative charges on the opposite side, and a contact potential, $U_c = U_{s2} - U_{s1}$, is generated, which is given by

$$U_c = \frac{\phi_{m1} - \phi_{m2}}{e} \quad (2)$$

where e is the electronic charge (1.6×10^{-19}C). As shown in Fig. 1, besides the electrical double layer across the interface, there exists a molecular thick layer, across which electric potential changes abruptly, bound to metal surfaces. This skin layer is due to electron wave functions extending beyond the ion cores at utmost surfaces. It can be viewed as a layer of dipoles, or molecular capacitor, and is also called the electrical double layer in some textbooks. However, since the bound double layers locate on each side of the interface, not across the interface, they have little contribution to adhesion forces.

A significant feature of the electrical double layers of metal/metal contacts is that the exchanged charges close around the interface on both sides of the contact, and there are no diffusion layers, as in the cases of semiconductors, dielectrics, and electrolyte solutions. Therefore, the electrostatic interactions between the electrical double layers of metal/metal contacts are often treated as those of planar capacitors. Presuming that the opposite charges are uniformly distributed on the two sides of the interface with charge density σ, and that the separation between the charge surfaces, r, as shown in Fig. 1b, is uniform, too, from Gauss's law, the electric field in the intervening medium is

$$E = \frac{\sigma}{\varepsilon\varepsilon_0} = \frac{U_c}{r} \tag{3}$$

where ε_0 is the vacuum dielectric constant, 8.85×10^{-12}C^2/Nm^2, and ε the relative dielectric constant of the intervening medium. From the electrostatics of planar capacitors (Feynman et al. 1989), the electrostatic energy, U, stored in the electrical double layer with a unit area can be written as

$$U = \frac{1}{2} E\sigma r = \frac{1}{2} U_c \sigma \tag{4}$$

and attractive electrostatic pressure, P_A, is given by

$$P_A = \frac{1}{2} E\sigma = \frac{\sigma^2}{2\varepsilon\varepsilon_0} = \frac{U_c\sigma}{2\varepsilon\varepsilon_0 r} \tag{5}$$

From (5), we can find that the electrostatic component of adhesion force is independent of the separation r. Because the values of work function for a variety of metals range from 3 eV to 5 eV, the magnitude of contact potential U_s should be less than 2 V from (2). Assuming the separation $r = 0.5$ nm, and the relative dielectric constant $\varepsilon = 1$, a rough estimation of the maximum charge density from (3) would be 35.4 mC/m^2, and the maximum electrostatic pressure from (5) would be 70 MPa, which is a value comparable to the van der Waals pressure for metals (about 100 MPa for the same separation). However, the charge density on the metal surfaces is limited by the breakdown of air, not exceeding 30 μC/m^2. This corresponds to an electrostatic pressure 5×10^{-5} MPa, far less than the van der Waals pressure.

However, the planar capacitor model is valid only when the separation r is much greater than the intervals, a, between discrete charges (see Fig. 1b). When the separation is comparable to the average interval, the assumption of uniformity of charge distribution on surfaces no longer holds. A detailed theoretical analysis of the effect of discrete structure of charges of the electrical double layer on the electrostatic component of adhesion can be found in Appendix D of reference (Derjaguin et al. 1978). The analysis result shows that the electrostatic component of adhesive pressure increases with the decrease in the average interval for a fixed separation. When $r = 0.5$ nm, the electrostatic pressure is 3 MPa for a = 10 nm, and 580 MPa for a = 1 nm. When $a \gg r$, the electrostatic pressure can be expressed as

$$P_A = \frac{Ne^2}{4\pi\varepsilon\varepsilon_0 r^2} \tag{6}$$

where N denotes the number of cation–anion pairs per unit area. In practical situations, the uniform separation condition in the model is difficult to satisfy because of the roughness of solid surfaces.

For metal/semiconductor contacts, the charge carriers (electrons in n-type or holes in p-type) in the zone of semiconductor near the interface transfer to the facing metal side under the potential difference between the Fermi energy levels of metal and semiconductor. At thermodynamic equilibrium, the energy band structure of the semiconductor side is distorted near the interface so that the Fermi energy level at the semiconductor side equals that of the metal side, and an electrical double layer is formed across the interface. Because the charge density (electrons or holes) in doped or undoped semiconductors is considerably less than in metals, the charge exchange density for metal/semiconductor contacts is less than that for metal/metal contacts. Meanwhile, the charge penetration depth into the semiconductor, which can be modeled by Poisson's equation, is much larger than that in metals, on the order of micrometers. A theoretical prediction of the electrostatic component of adhesive force for metal/semiconductor contacts has been given by Derjaguin and colleagues (1978).

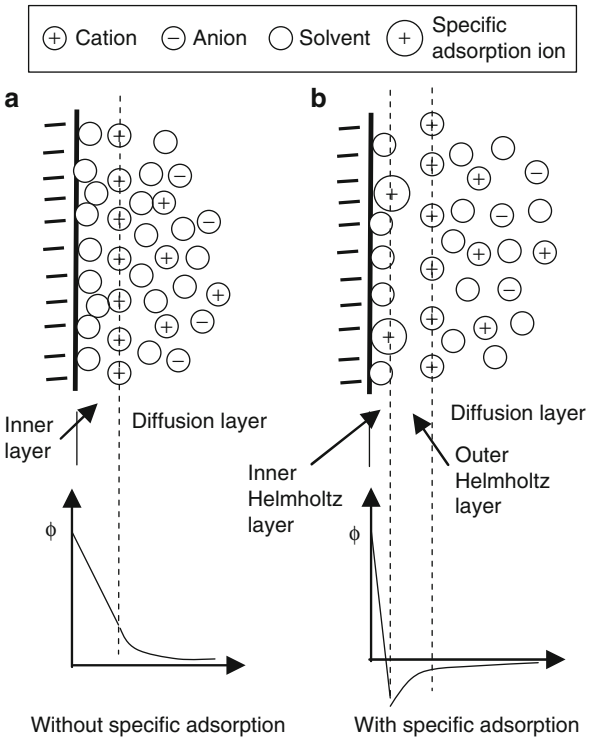
For metal/polymer contacts, donation and acceptance of electrons are considered to take place between the Fermi energy level of the metal and the surface states, which are located in between the highest occupied molecular orbitals

(HOMO) and the lowest occupied molecular orbitals (LOMO), of the polymer. The adhesion behaviors between metals and polymers are explained by the principle of the hard-soft acid-base (HSAB) instead of the electrostatic effect (Lee 1991), although a theory of electrical double layer by the donor-acceptor pairs of general solid contacts is available (Derjaguin et al. 1978).

Besides the dry solid/solid contacts, adhesion and interactions between solids separated by an intervening liquid phase have been extensively studied in the context of colloid science and electrochemistry. In this case, at each of the interfaces of solid and liquid, there exists an electrical double layer. When the distance between two solid/liquid interfaces is close enough, the electric fields of the opposing electrical double layers overlap, resulting a repulsive or attractive electrostatic force between the two surfaces. The formation of electrical double layer at a solid/liquid interface is considered due to the dissociation of ions from solid surface to liquid phase and/or the binding of ions from liquid phase to solid surface at natural conditions. The charge density on one or both of the solid surfaces can also be changed to some extent (limited by the electrochemical reactions) by applying an external electric voltage, especially when the solids are conductors or semiconductors. Structures of the electrical double layers formed at solid/liquid interfaces are not as compacted as those of metal/metal contacts due to the lower charge density in liquid phase and the Brownian motion of ions. Figure 2a schematically shows charge and potential distributions of the electrical double layer of an isolated solid/liquid interface in the absence of specific adsorption. The charges in the liquid phase distribute in two layers, inner layer (or Helmholtz layer) and diffusion layer (or Gouy-Chapman layer). If the liquid phase contains some amount of specific adsorption ions, such as I^-, Br^-, SN^-, and surfactant ions, the inner layer can be further divided into an inner Helmholtz layer and an outer Helmholtz layer, as shown in Fig. 2b.

The potential distribution in the diffusion layer obeys the Poisson-Boltzmann equation, and analytical and numerical solutions for single and overlapped electrical double layers have been obtained (Israelachvili 1985; Usui 1984). Based on the electric potential distribution, electrostatic force F and energy W per unit area between two solid surfaces can be found. For two planar surfaces with the same charge density, the repulsive electric pressure can be approximately expressed by

$$P = \frac{2\sigma^2 e^{-\kappa D}}{\varepsilon\varepsilon_0} \tag{7}$$



Electrostatic Field Effects on Adhesion, Fig. 2 Electrical double layer at solid/liquid interface. (**a**) Without specific adsorption. (**b**) With specific adsorption

where σ is the charge density on the solid surface, κ the Debye length of the electrical double layer, D the separation between the two solid surfaces, and ε the relative dielectric constant of the liquid phase, if the surface potential of solids is lower than 25 mV. For two identical spheres of radius R, the electric repulsive force between them is given approximately at the same surface potential conditions.

$$F = \frac{2\pi R\sigma^2 e^{-\kappa D}}{\kappa\varepsilon\varepsilon_0} \tag{8}$$

By incorporating the double layer interaction with van der Waals force between particles, the so-called DLVO theory (Derjaguin, Landau, Verwey, and Overbeek) of colloidal stability has been established.

Experimental Studies of Electrostatic Effects on Adhesion of Solids

There are numerous reports of experimental work on adhesion of solids. Among them, some work emphasizes the role of electrostatic effects in adhesion, and makes efforts to clarify the contribution of electrostatic

interactions to the total adhesion force. However, it is almost impossible to separate the pure electrostatic interactions from the other components, including chemical and van der Waals forces, in the experimental study of adhesion. In addition, back-flow, tunneling, and air discharging during the detachment of two charged surfaces make it difficult to accurately measure electrostatic charge and force.

A variety of test methods have been used to measure adhesion of solids, including the particle detachment test with centrifugal or acceleration force, the peeling test, surface force apparatus (SFA), atomic force microscope (AFM), and so on. SFA was originally designed for measurement of surface forces between two atomically smooth, curved mica surfaces with or without an intervening medium. The range of materials has been extended to thin metallic and polymeric films. AFM can be used to measure surface forces between a standard tip (usually silicon, Si_3N_4, or diamond) with a tip radius of dozens of nanometers and any flat surface. The tip can be replaced by a microsize sphere to change the material and force range. Both SFA and AFM utilize cantilever beams for sensing the surface forces. Adhesion force is identified as the "jump-out" or "pull-off" point on the force curve acquired in an attachment-detachment process.

Electrical double layer forces and van der Waals forces between mica surfaces have been measured on SFA in water, dilute KNO_3, and $Ca(NO_3)_2$ solutions and other various electrolytes as well as in nonaqueous polar liquids in the surface separation ranging from several hundreds of nanometers down to a few nanometers. The measured force vs. separation curves are in good agreement with the DLVO theory.

Charges can be injected into some polymer materials (e.g., PE, PTFE, PS, and PVCH) by various methods and not be neutralized in air for a relatively long time. Comparison of the adhesion force of charged polymers with that of noncharged polymers in dry contact conditions has been done on SFA. It has been reported that charged polymers, in spite of the pairs of negatively charged to negatively charged or negatively charged to positively charged, present a larger adhesive force than that of noncharged polymers. The higher the surface charge density, the larger the adhesive force. The reason for the increase of adhesion force is considered due to the increase of surface energy of polymers by charging. The higher surface energy results in a large contact area according to JKR theory, thus leading to a higher adhesion force in detaching.

Applying an external electric field on a solid surface can also change its surface potential. An AFM test of a Langmuir-Blodgett (LB) monolayer deposited on silicon substrate under DC and AC external electric fields has shown that DC voltage application can increase the attractive force between the LB film and the AFM tip in the approach stage, but causes little change in the "pull-off" force compared with the case of no voltage application.

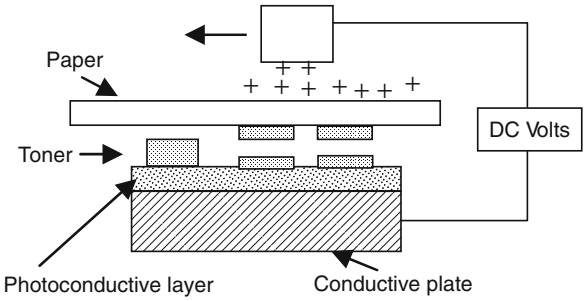
It is well known that adhesion of solids depends not only on surface properties and structures of solids but also on bulk properties and structure. Stiff structures made of harder materials (i.e., greater Young's modulus) can store greater elastic energy than flexible structures made of softer materials in the loading process, and thus in the subsequent unloading process the restoring elastic energy of stiff structures is larger than that of flexible structures relative to adhesion energy. As a result, stiff structures remain a smaller real contact area and adhesion force after unloading than flexible structures under the same loading conditions. Therefore, in macromachines adhesion is usually not a serious problem. However, in micro/nanosystems, adhesion or stiction becomes a significant issue because of the decrease in stiffness of micro/nanostructures due to scaling down. For example, in RF-MEMS switches, repeated contact will cause charge accumulations in the dielectric medium and and increase in electrostatic adhesion force. Once the elastic restoring force of beams or membranes cannot overcome the adhesion force, the device will fail. Hence, solving adhesion problems is an important subject in micro/nanomanufacturing.

Key Applications

Electrophotography

Electrophotography is the term coined to describe the core technology on which the modern photocopying industry is based. A synonym for electrophotography is xerography, which means the formation of electrostatic latent images on a photosensitive layer or a photoconductive plate. An electrophotography process involves the following five basic steps: (a) sensitizing the photoconductive layer by corona discharging; (b) exposing the photoconductive layer to form an electrostatic latent image; (c) developing the latent image with fine particles; (d) transferring the developed image to paper; and (e) fixing the image by fusing. Among them, steps (a), (c) and (d) are closely related to the electrostatic effect. Figure 3 provides an illustration of the transfer step in the process.

The photoconductive layer is usually a film of amorphous selenium or an organic photoconductive coating. The toner can take the form of either dry powders or colloidal suspensions. The toner pigments mainly consist



Electrostatic Field Effects on Adhesion, Fig. 3 A schematic of the transfer step in the electrophotography process

of carbon black or magnesium oxide with a diameter of a few micrometers, mixing with some amount of charge control agents. Controlling the charge density of carbon black particles is key to achieving high quality of photocopying. For a review of the history, principles, and applications of electrophotography, the reader is referred to reference (Schaffert 1965) and other publications.

Anodic Bonding

Anodic bonding is a process widely used in the manufacture of a variety of sensors, microfluidic systems, and MEMS devices (Maluf 2000). The purpose of the process is to join a bare silicon wafer with a sodium-containing glass substrate (for example, Corning Pyrex 7740 and 7070) together without any adhesive. It can also be used to bond two silicon substrates. In this case, a thin glass film containing sodium must be deposited on one of the mating surfaces before anodic bonding.

Anodic bonding is also known as electrostatic bonding. A large DC voltage (500–1,500 V) is applied across the two substrates, with the glass held at the negative potential. The application of the electric field causes the cations (Na^+) to migrate away from the silicon/glass interface towards the cathode, leaving behind fixed negative charges. When all mobile cations have reached cathode, which can be monitored by measurement of current during the bonding process, the bonding is finished. The intimate contact is maintained by the electrostatic force between the fixed negative charges in the glass and the positive charges in the silicon. Meanwhile, the electrostatic force enhances the formation of chemical bonding between silicon and glass.

Anodic bonding is usually conducted at the temperature range of 200–500 C in a vacuum, air, or an inert gas environment. The silicon oxide film on silicon wafer surface affects the bonding strength greatly. Thermal stress is an important concern for the success of the anodic bonding. Selection of the glasses with a coefficient of thermal expansion comparable with that of monocrystalline silicon ($2.6 \times 10^{-6}/°C$) is necessary.

Cross-References

- ▶ Basic Concepts in Adhesion Science
- ▶ Contacts Considering Adhesion
- ▶ Electrostatic Field Effects on Friction

References

J. Comyn, *Adhesion Science* (Royal Society of Chemistry Paperbacks, Cambridge, 1997)

B.V. Derjaguin, N.A. Krotova, V.P. Smilga, *Adhesion of Solids* (Consultants Bureau, New York, 1978)

R.P. Feynman, R.B. Leighton, M. Sands, *The Feynman Lectures on Physics* (Addison-Wesley, Reading, 1989)

J.N. Israelachvili, *Intermolecular and Surface Forces* (Academic Press., London, 1985)

A.J. Kinloch, Review: the science of adhesion, part 1, surface and interfacial aspects. J. Mater. Sci. **15**, 2141–2166 (1980)

L.-H. Lee, The chemistry and physics of solid adhesion, in *Fundamentals of Adhesion*, ed. by L.-H. Lee (Plenum Press, New York, 1991)

N. Maluf, *An Introduction to Microelectromechanical Systems Engineering* (Artech House, Boston, 2000)

R.M. Schaffert, *Electrophotography* (Focal Press, London, 1965)

Y. Tian, N. Pesika, H. Zeng, K. Rosenberg, B. Zhao, P. McGuiggan, K. Autumn, J. Israelachvili, Adhesion and friction in gecko toe attachment and detachment. Proc Natl Acad Sci USA **103**(51), 19320–19325 (2006)

S. Usui, Electrical double layer, in *Electrical Phenomena at Interfaces: Fundamentals, Measurements and Applications*, ed. by A. Kitahara, A. Watanabe (Marcel Dekker, New York, 1984)

Electrostatic Field Effects on Friction

Yonggang Meng

State Key Laboratory of Tribology, Tsinghua University, Beijing, People's Republic of China

Synonyms

Friction and electrostatic fields; Potential-controlled friction; Voltage-controlled friction

Definition

Electrostatic fields at solid-solid and solid–liquid interfaces are intrinsic factors involved in friction and lubrication and can be externally controlled in certain cases.

Scientific Fundamentals

Interaction Between Electrostatic Field and Friction

After being rubbed with a piece of silk, a glass bar will be charged positively, and an electrostatic field is formed around it. This electrical phenomenon is referred to as triboelectrification, which has been identified in most friction pairs. On the other hand, if an external electric field is imposed onto a friction pair, its coefficient of friction may increase, decrease, or change little, depending on the constitutions of the friction pair, lubrication condition, and the polarity and magnitude of the applied electric field. The relationship between friction and electrostatic field is complicated.

Friction of Dry Contacts

When two dissimilar solid bodies are in contact, an electrical double layer (see "► Electrostatic Field Effects on Adhesion") will form across the interface of contact. The electrical double layer generates an intrinsic electrostatic potential difference and an attractive electrostatic force between the contacting bodies. It is often postulated that the stronger the intrinsic electrostatic potential difference is, the higher the sliding friction will be. Furthermore, if an external electric field can simply cancel out the intrinsic electrostatic potential difference, then a reduced friction is expected. Based on the acknowledgment of the role of electrostatic field during friction, a number of experimental investigations have been done by dozens of individual research groups on the correlation between contact potential and coefficient of friction and on effects of external electric field application on friction for a variety of friction pairs, such as metal/metal dry contacts, metal/polymer dry contacts, metal/ceramic dry contacts, and so on. Most of the experimental results reported a reduction of friction coefficient, in the range of 10–40%, when the self-generated contact potential during friction was counteracted by properly manipulating the external electric field. However, it seems that there exists no simple relationship between friction and electrostatic potential at interfaces. One of the reasons for the complexity is attributed to the fact that, unlike the case of adhesion, the real electrostatic contact potential during friction changes considerably with time due to triboemission and surface damage, which strongly depend on load and speed conditions of running. In addition, the reduction in friction coefficient under external electric field applications cannot be interpreted soley by the possible change in attractive electrostatic interaction. Other mechanisms, such as oxide film formation on metallic rubbing surfaces, may be responsible for the decrease of friction, as pointed out by some researchers.

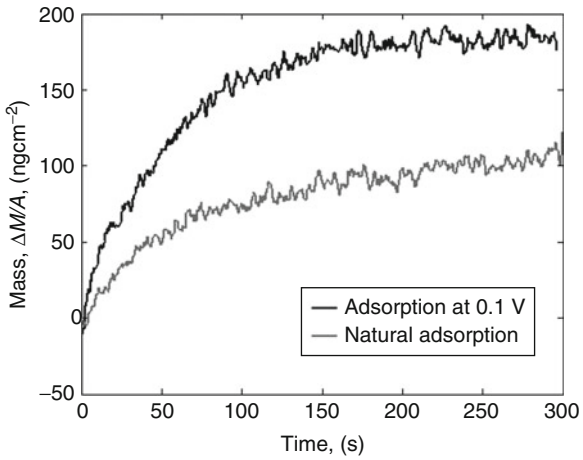
Imposing an external electric field would diminish the attractive electrostatic force between contacting bodies, but cannot change the force from attractive to repulsive. A repulsive electrostatic force can be generated between a pair of contacting bodies made of electrets, both of which carry same kind of excessive charges. With the purpose of realization of low friction, various efforts have been made to dope negative or positive charges onto surface layers of friction pairs. However, some experimental work has shown that charge doping cannot result in a substantial reduction in friction. On the other hand, electrostatic levitation can be realized by deliberate design and control of electrostatic field around a charged object in a vacuum. Based on the principle of electrostatic levitation, frictionless gyroscopes and bearings have been invented and used in practice.

Friction in Boundary Lubrication

When a solid contact is immersed in liquid, the micro and nano gaps separated by real contacting asperity junctions (Bowdern and Tabor 1950) are filled with the liquid, provided that the solid surfaces are fully wetted by the liquid. Thus, a solid–liquid interface is formed on both sides of the contact. Generally, an electrical double layer will form at each of the solid–liquid interfaces, and the interaction force between the two neighboring solid–liquid electrical double layers is repulsive. This repulsive force may counteract the surface force due to the van der Waals effect and plays an important role in particle aggregation behaviors as well as in friction. If the liquid phase is aqueous electrolyte, the strength of the electrical double layer interaction can be modulated to some extent by changing the pH value of the electrolyte and/or the surface potential of the contact with a potentiostat. According to predictions made by Bockris and Sen (1972) and later by Kesall et al. (1993), a peak value of coefficient of friction appears at the condition of zero surface potential, while both the negative and positive surface potentials lead to a lower friction. This theory was applied to explain the experiment results of friction of platinum/platinum contact in pure dilute sulfuric acid, which were described and interpreted using a different mechanism, hydrogen gas deposition under the negative surface potentials and oxygen gas deposition under the positive surface potentials, in reference (Bowdern and Tabor 1950). The experimental work on iron-on-iron sliding friction in a simple salt solution and theoretical analysis reported in reference (Zhu et al. 1994) also support the mechanism of electrostatic repulsion effect on friction. It is worth noting that the effect of electrostatic

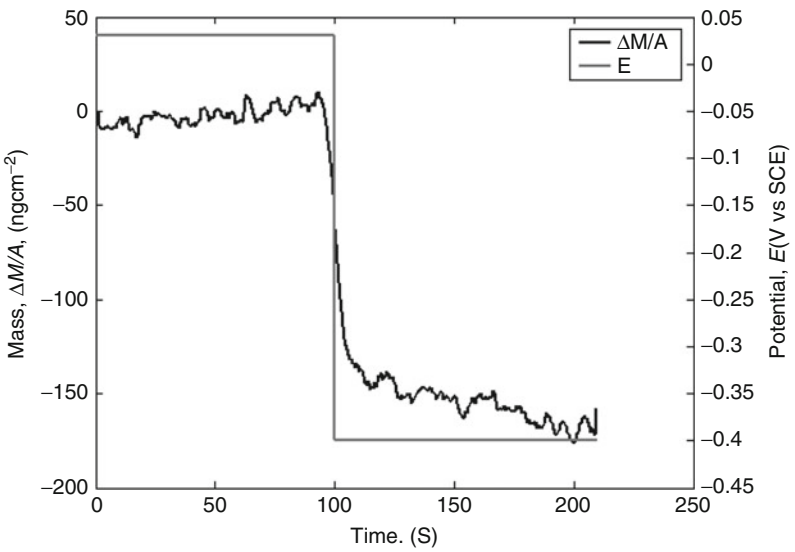
repulsion on friction becomes less remarkable with the increase in normal load because the electrostatic repulsion force per unit area is limited, typically less than a few MPa in most of practical situations.

A substantial effect of surface potential on friction of metals has been found in the presence of lubricious additives, such as fatty acid, stearate acid, and other surfactants, in aqueous solutions, even at heavy load conditions. Modern surface analysis tools, including atomic force microscope (AFM) and quartz crystal microbalance (QCM), have been used to identify the adsorption, desorption, and morphology of surfactant molecules on metal surfaces under different surface potentials, combined with electrochemistry measurement. For example, the adsorption processes of dodecyl sulfate anions on a stainless steel surface at natural surface potential (or OCP, open circuit potential, around +0.03 V with respect to a saturated calomel reference electrode, or vs. SCE, in the test case) and a controlled positive potential of +0.1 V verses SCE are shown in Fig. 1, based on measurements on an electrochemistry QCM (EC-QCM). The liquid phase used in the measurements was a solution of sodium dodecyl sulfate with a concentration of 0.5 mM. It is clear that the positive surface potential can enhance adsorption of the anionic surfactant molecules on a metal surface, as shown by the change in the mass of adsorbate in the figure. In contrast, desorption of the anionic surfactant molecules occurs quickly when the surface potential drops suddenly from the OCP to −0.4 V verses SCE, as shown in Fig. 2



Electrostatic Field Effects on Friction, Fig. 1 Adsorption of dodecyl sulfate anions verses time curves under the natural and positive surface potentials on a stainless steel surface

The adsorption and desorption phenomena under the control of surface potential is reversible as long as the magnitude of potential is not so high that severe electrochemical reactions occur on the surface. For cationic surfactants, the effect of surface potential on adsorption/desorption is inverse, i.e., a negative potential enhances their adsorption from bulk solution onto metal

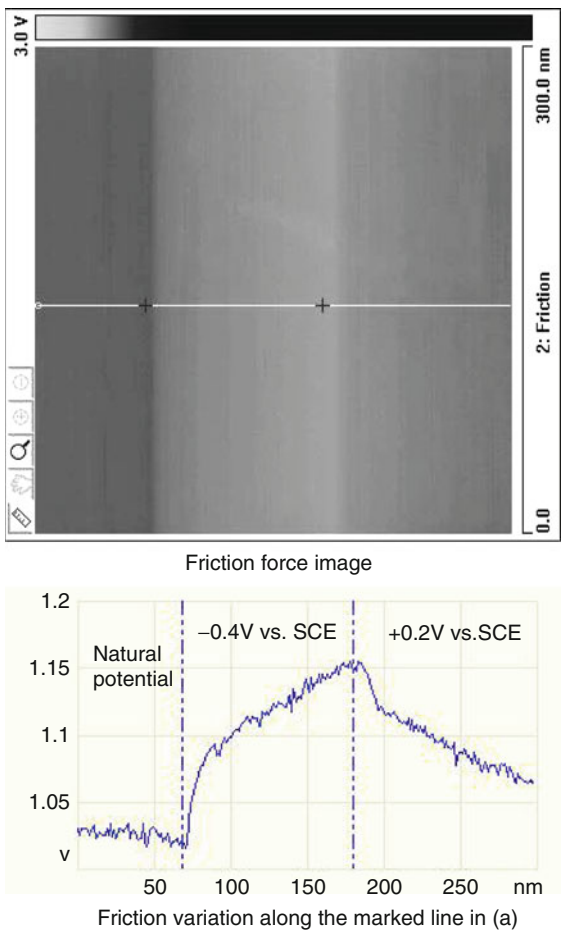


Electrostatic Field Effects on Friction, Fig. 2 Desorption of dodecyl sulfate anions from a stainless steel surface due to a sudden drop of surface potential from natural potential to −0.4 V verses SCE at time=100 s

surfaces, while a positive potential inhibits adsorption and leads to desorption. The morphology of the surfactant molecules adsorbed on metal surfaces is also influenced by the surface potential. For example, at small positive potentials, dodecyl sulfate anions form hemimicellar aggregates. When the positive charge on the metal surface either becomes equal to or exceeds the charge of adsorbed surfactant, the hemimicellar aggregates melt to form a condensed monolayer. The transition between the hemimicellar and condensed states of the surfactant film is reversible (Burgess et al. 1999).

Because the surfactant layers adsorbed or deposited on metal surfaces play a role in boundary lubrication, tuning of the adsorption/desorption of surfactants by changing surface potential implies that friction in boundary lubrication can be controlled in the range from a well-lubricated condition (i.e., the metal surface is covered by a boundary lubrication film of surfactants) to the extremely poorly lubricated condition (i.e., no protecting surfactant film on metal surface). Both nanoscale friction tests with lateral force microscope and macroscale friction tests with ball-on-disk machine have verified the dramatic transition of the boundary lubrication states caused by the change in surface potential, as shown in Figs. 3 and 4, respectively. Figure 3 displays the variation of the output signal (proportional to frictional force) when a Si_3N_4 tip scans across a polished stainless steel surface immersed in 0.5 mM sodium dodecyl sulfate solution. During the scanning process, the surface potential of stainless steel was switched from the natural potential to −0.4 V verses SCE and finally to +0.2 V verses SCE with a potentiostat. Consequently, friction was relatively lower at the first phase, then increased to a higher value quickly in the second phase, and finally declined gradually in the third phase.

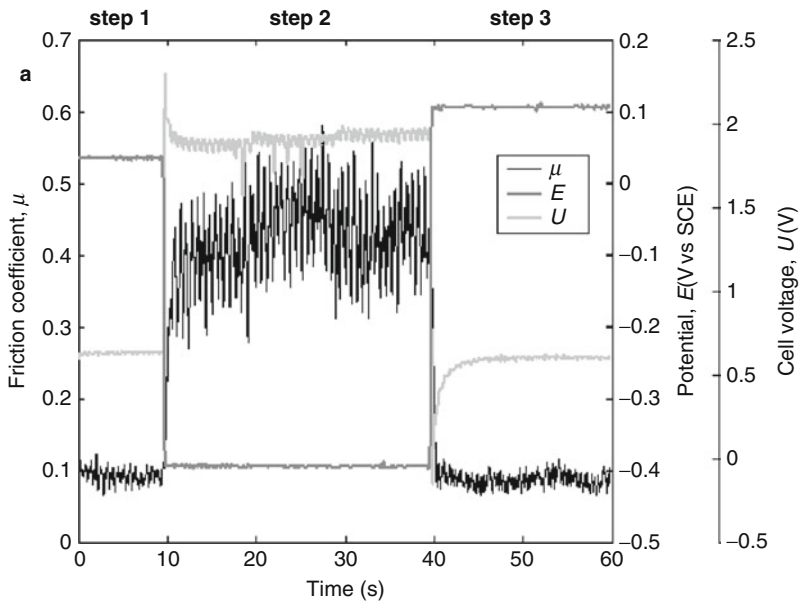
Figure 4 shows a typical experimental result of friction coefficient of ZrO_2 ball-on-stainless disk in 0.5 mM/L sodium dodecyl sulfate solution. The mean contact pressure in the test was 100 MPa, and the linear sliding velocity was a constant of 57 mm/s. During the sliding friction test, only the surface potential of the stainless steel sample was intentionally changed in three stages. In the first 10 s of running, a minor natural potential (about +0.03 V vs. SCE) existed as in conventional situations, and the average friction coefficient over the stage was 0.1. In the next 30 s of running, the surface potential of the metal disk was shifted to the value of −0.4 V verses SCE, and the average friction coefficient increased to about 0.43, which is the same magnitude as in the test with pure water. In the last 10 s of test, the surface potential was elevated to the value of +0.1 V verses SCE, and the average friction coefficient dropped to 0.09. The



Electrostatic Field Effects on Friction, Fig. 3 Variation of friction force during the three different potential phases. (**a**) Friction force image (**b**) Friction variation along the marked line in (**a**)

friction coefficient at the medium negative potential is about four times higher than that at the natural or a minor positive potential.

It should be noted that neither hydrogen nor oxygen gas production was observed on the metal surface in the above experiments, because the surface potentials used in the experiments were not great enough to cause hydrolysis of water. Usually, to prevent surfaces from electrochemical oxidation, the highest positive surface potential allowed for metallic parts is limited to a small magnitude, although a higher potential is more favorable to the formation of an anionic surfactant film on metal surfaces. On the other hand, excessive anodic potentials should also be avoided because, with the shifting of surface potential from neutral toward the negative direction, friction coefficient saturated



Electrostatic Field Effects on Friction, Fig. 4 Changes of the applied external voltage U, surface potential of metal disk E and friction coefficient μ during a ball-on-disk friction test

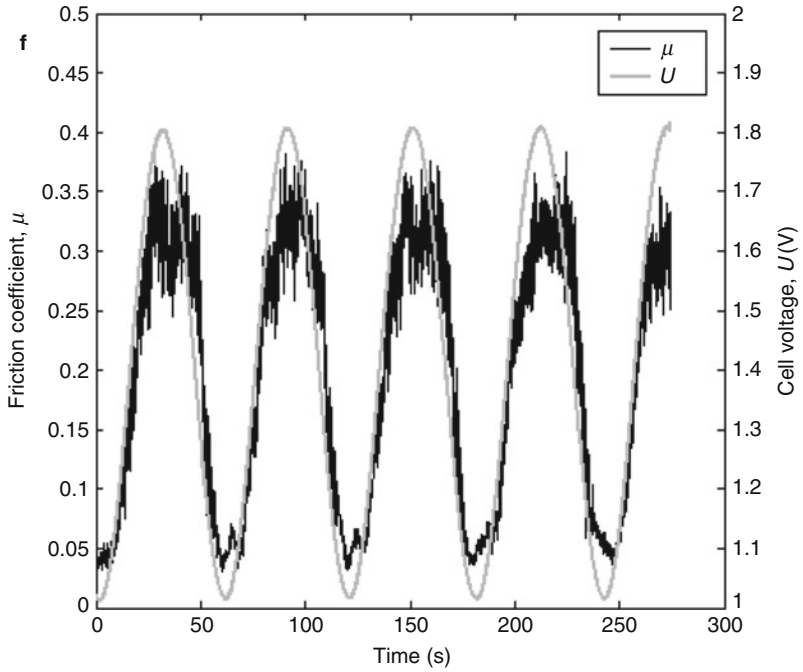
at a medium anodic potential to the magnitude corresponding to pure water, while electric current increases exponentially with the overpotential, resulting in high electric power loss and severe damage to metal surfaces.

With oil lubrication, it becomes difficult to measure and control surface potential of contact surfaces because of the high impedance of oil phase. Nevertheless, electrostatic field can influence friction in boundary lubrication, especially when polar lubricious additives are contained in lubricants. For example, it has been reported that the effect of a friction reduction by up to 35% can be achieved by utilizing an in-situ electro-charging technique in the friction test of steel ball on steel plate with the lubricant of base mineral oil blended with 2 mass percent zinc organo dithiophosphates (ZDPs) (Tung and Wang 1991). The improved performance is attributed to an increase in the electrochemical reactivity of the ZDP-mineral oil blends and the formation of phosphate or sulfate films on the metal surfaces. When liquid crystals (LCs), the molecule orientation and viscosity of which are sensitive to electric fields, are used as lubricant, application of a d.c. or low-frequency a.c. voltage across a steel-to-steel contact can also result in a remarkable reduction of friction coefficient under boundary lubrication condition (Kimura et al. 1994). It is suggested that the application of electric fields may enhance the lateral attraction between the LC molecules, thus either strengthening monomolecular films by compensating for the possible absence of chemical bonds to metal surfaces, or maintaining thick multimolecular films.

Friction in Fluid Lubrication

In fluid lubrication conditions, the mating surfaces of a friction pair are completely separated by a flowing fluid film, the load capacity and friction characteristics of which are dominated by the hydrodynamic behaviors of lubricant. Electrostatic fields are involved in hydrodynamic lubrication in the following two circumstances.

One circumstance is encountered when fluid lubricant contains movable ions, such as aqueous salt solutions. Near solid–liquid interfaces, the ions in the liquid phase distributed in the forms of a compact layer and a diffusion layer, obeying the Poisson-Boltzmann equation, according to the theory of electrical double layer. When the liquid flows through the narrow gap of lubrication region under pressure, the ions in the diffusion layer of an electrical double layer are carried downstream, generating a streaming current in the direction of the liquid flow. In steady state, a conduction current of ions flows in the opposite direction of the streaming current, which gives rise to a reduction of flow rate in the direction of pressure gradient. This dragging effect of electrical double layer on lateral liquid flow is often regarded as an electroviscous effect. Theoretical modeling and numerical analysis have shown that the electroviscous effect can enhance the load-carrying capacity



Electrostatic Field Effects on Friction, Fig. 5 Simultaneous variations of external voltage and friction coefficient with time

of fluid lubrication, especially when the lubricant film thickness is in the same order of the characteristic thickness of electrical double layer (Bai et al. 2006).

The other circumstance is in fluid lubrication with electrorheological fluids (ERFs). ERFs are suspensions of dielectric particles in insulating oils. In the absence of an external electric field, ERFs are Newtonian fluids. When a strong electric field, usually in the order of MV/m, is applied, ERFs become solid-like materials with a shear yield stress as high as tens of kPa. The transition is due to the polarization and aggregation of the suspended particles under electric fields and is reversible. ERFs have a tunable capability to sustain normal, shear, and tensile loads (Tian et al. 2002) and have potential applications in lubrication of journal bearings, clutches, and other machine components.

Key Applications

Wire Drawing Enhanced by Electric Field Application

Wire drawing is a metalforming process in which a metal wire is pulled through a tapered die hole at high speed to become thinner and longer. In industry, an oil-in-water (O/W) emulsion containing surfactant and fatty acid is often used as a lubricant to reduce friction between the deforming wire and die. To enhance lubrication performance of O/W emulsion in copper wire drawing, a technique of electric field application was used in an industrial wire drawing machine (Su 1997). In front of the drawing die, a tubular auxiliary electrode was fitted around the wire, and the emulsion lubricant was pumped through the tube. A d.c. voltage was applied over the copper wire and the auxiliary electrode. It was reported (Su 1997) that as high as 35% reduction in drawing force was achieved by using the electric field application technique. In addition, the surface quality of drawn wires was improved remarkably, and die life was prolonged significantly, compared with no field wire drawing.

Active Control of Sliding Friction

As described in the preceding section, sliding friction coefficient of metals in sodium dodecyl sulfate aqueous solution reaches the minimum value μ_{min} at a small positive surface potential E_p, and the maximum value μ_{max} at a medium negative surface potential E_n. Assuming that friction coefficient changes linearly and promptly in the range [μ_{min}, μ_{max}] when surface potential is changed in the range [E_p, E_n], then the surface potential required for any prescribed value of friction coefficient in the range [μ_{min}, μ_{max}] can be estimated. This implies that sliding friction can be controlled actively in a sequential or feedback way. Figure 5 shows such a demonstration of friction control. Here, the external voltage applied between the metal surface and a counter electrode, which is related to the surface

potential proportionally, is modulated in the range of 1–1.8 V sinusoidally with a period of 60s. The friction coefficient follows the change in the external voltage simultaneously and shows a sinusoidal variation in the range from 0.05 to about 0.35. This opens a unique way to realize smart transmission, braking, and damping by using friction.

Cross-References

► Electrostatic Field Effects on Adhesion
► Friction (Concepts)

References

S. Bai, P. Huang, Y. Meng, S. Wen, Modeling and analysis of interfacial electro-kinetic effects on thin film lubrication. Tribol. Intern. **39**(11), 1405–1412 (2006)

J.O.M. Bockris, R.K. Sen, Variation of the coefficient of friction with potential for a solid-solution contact: a revised calculation. Surf. Sci. **20**, 237–241 (1972)

F.P. Bowdern, D. Tabor, *The Friction and Lubrication of Solids* (Clarendon, Oxford, 1950)

I. Burgess, C.A. Jeffrey, X. Cai, G. Szymanski, Z. Galus, J. Lipkowski, Direct visualization of the potential-controlled transformation of hemimicellar aggregates of dodecyl sulfate into a condensed monolayer at the Au(111) electrode surface. Langmuir **15**, 2607–2616 (1999)

G.H. Kelsall, H.A. Spikes, Y.Y. Zhu, Electrochemical effects on friction between metal oxide surfaces in aqueous solutions. J. Chem. Soc. Faraday Trans. **89**, 267–272 (1993)

Y. Kimura, K. Nakano, Kato T, Control of friction coefficient by applying electric fields across liquid crystal boundary films. Wear **175**, 143–149 (1994)

Y.-Y. Su, Enhanced boundary lubrication by potential control during copper wire drawing. Wear **210**, 165–170 (1997)

Y. Tian, Y. Meng, S. Wen, Electrorheological fluid under elongation, compression, and shearing. Phys. Rev. E **65**(3), 031507 (2002)

S.C. Tung, S.S. Wang, In-situ electro-charging for friction reduction and wear resistant film formation. Tribol. Trans. **34**(4), 479–488 (1991)

Y.Y. Zhu, G.H. Kelsall, H.A. Spikes, The influence of electrochemical potentials on the friction and wear of iron and iron oxides in aqueous systems. Tribol. Trans. **37**(4), 811–819 (1994)

Ellipsometry

Jinmin Zhao
Spintronics, Media and Interface Division,
Agency for Science, Technology and Research (A*star),
Data Storage Institute, Singapore, Singapore

Definition

Ellipsometry is an optical technique that uses polarized light to characterize thin films, surfaces, and material microstructures; it is commonly used to determine the thickness and optical constants of both layered and bulk materials. It measures the change in the polarization state of light reflected or transmitted across the boundary where the change in material properties occurs. The polarization change is represented as the amplitude ratio Ψ and the phase difference Δ, which makes the measurement highly precise and sensitive. Ellipsometry can also be applied to characterize composition, crystallinity, and doping concentration.

Many types of ellipsometry are available in the market. Single-wavelength ellipsometry normally employs 632.8 nm wavelength. Spectroscopic ellipsometry covers the infrared (IR), near infrared (NIR), visible (VIS), or ultraviolet (UV) spectral range. Imaging ellipsometry combines laser ellipsometry with a charge-coupled device (CCD) camera. In situ ellipsometry is designed for real-time monitoring in a process control. The choice of ellipsometry is usually determined by the best wavelength range of light to meet a given application.

Scientific Fundamentals

Polarized Light

Ellipsometry uses polarized light for the measurement. There are three forms of polarized light: linear, circular, and elliptical. If two mutually perpendicular waves P and S are in-phase, the resulting light will be linearly polarized, as shown in Fig. 1a. The relative amplitudes determine the resulting orientation. If these two waves are 90° out-of-phase but equal in amplitude, the resulting light is circularly polarized, as shown in Fig. 1b. If these two waves are of arbitrary amplitudes and phases, the resulting light is elliptically polarized, as shown in Fig. 1c (J.A. Woollam Co. 2000).

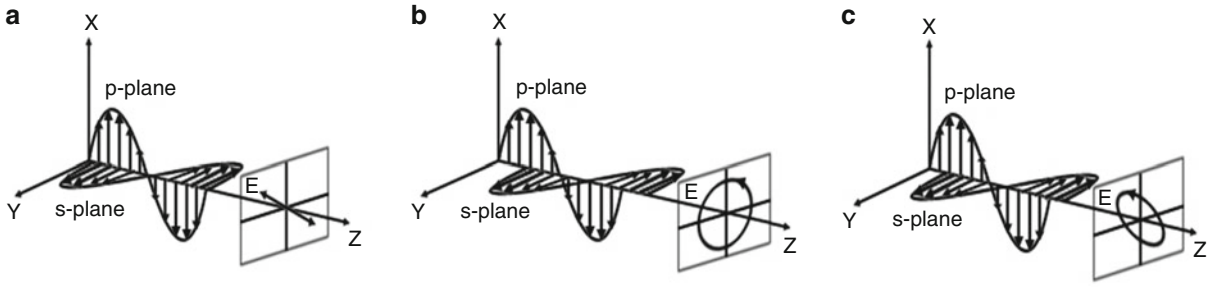
Ellipsometry Measurement

Ellipsometry measures the change in the polarization state as light reflects from or transmits through a material structure. The incident light is linear or circular with both p- and s- components. The reflected or transmitted light becomes elliptical and contains the information about the material after undergoing the amplitude and phase changes of both p- and s- components. These changes are the ellipsometry measurement, commonly written as:

$$\tan(\Psi)e^{j\Delta} = \frac{\tilde{R}_p}{\tilde{R}_s} \quad (1)$$

where

$\tilde{R}_p$, $\tilde{R}_s$ – Fresnel reflection coefficients of the p- and s- components separately



Ellipsometry, Fig. 1 Two linearly polarized light beams combined to demonstrate polarization: (**a**) linear; (**b**) circular; (**c**) elliptical

$\tan(\Psi)$ – Magnitude of complex ratio of $\tilde{R}_p$ and $\tilde{R}_s$
Δ – Phase difference between $\tilde{R}_p$ and $\tilde{R}_s$

Because ellipsometry measures the ratio of two values, it has high tolerance of the common fluctuation in the light source intensity so that it can be highly accurate and reproducible. Since the ratio is a complex number, it also contains the "phase" information (Δ), which makes the measurement highly sensitive to tiny changes in material microstructures.

For a single film on an optically thick substrate, as shown in Fig. 2, ellipsometric data Ψ and Δ can be calculated from

$$\tan(\Psi)e^{j\Delta} = \frac{\tilde{r}_{p1}+\tilde{r}_{p2}e^{-j\Gamma}}{1+\tilde{r}_{p1}\tilde{r}_{p2}e^{-j\Gamma}} * \frac{1+\tilde{r}_{s1}\tilde{r}_{s2}e^{-j\Gamma}}{\tilde{r}_{s1}+\tilde{r}_{s2}e^{-j\Gamma}} \qquad (2)$$

where r_{p1}, r_{p2} and r_{s1}, r_{s2} are the Fresnel reflection coefficients at two different interfaces for p- and s- components, respectively. The subscript 1 denotes the ambient-film interface while the subscript 2 denotes the film-substrate interface. Γ is the phase difference between the reflected components of light. They can be expressed by Fresnel equations:

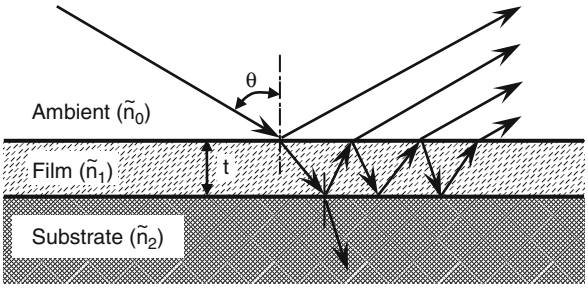
$$\tilde{r}_{p1} = \frac{\tilde{n}_1^2\cos\theta - \tilde{n}_0(\tilde{n}_1^2 - \tilde{n}_0^2\sin^2\theta)^{1/2}}{\tilde{n}_1^2\cos\theta + \tilde{n}_0(\tilde{n}_1^2 - \tilde{n}_0^2\sin^2\theta)^{1/2}} \qquad (3)$$

$$\tilde{r}_{p2} = \frac{\tilde{n}_2^2(\tilde{n}_1^2 - \tilde{n}_0^2\sin^2\theta)^{1/2} - \tilde{n}_1^2(\tilde{n}_2^2 - \tilde{n}_0^2\sin^2\theta)^{1/2}}{\tilde{n}_2^2(\tilde{n}_1^2 - \tilde{n}_0^2\sin^2\theta)^{1/2} + \tilde{n}_1^2(\tilde{n}_2^2 - \tilde{n}_0^2\sin^2\theta)^{1/2}} \qquad (4)$$

$$\tilde{r}_{s1} = \frac{\tilde{n}_0^2\cos\theta - (\tilde{n}_1^2 - \tilde{n}_0^2\sin^2\theta)^{1/2}}{\tilde{n}_0^2\cos\theta + (\tilde{n}_1^2 - \tilde{n}_0^2\sin^2\theta)^{1/2}} \qquad (5)$$

$$\tilde{r}_{s2} = \frac{(\tilde{n}_1^2 - \tilde{n}_0^2\sin^2\theta)^{1/2} - (\tilde{n}_2^2 - \tilde{n}_0^2\sin^2\theta)^{1/2}}{(\tilde{n}_1^2 - \tilde{n}_0^2\sin^2\theta)^{1/2} + (\tilde{n}_2^2 - \tilde{n}_0^2\sin^2\theta)^{1/2}} \qquad (6)$$

$$\Gamma = (4\pi t/\lambda)(\tilde{n}_1^2 - \tilde{n}_0^2\sin^2\theta)^{1/2} \qquad (7)$$



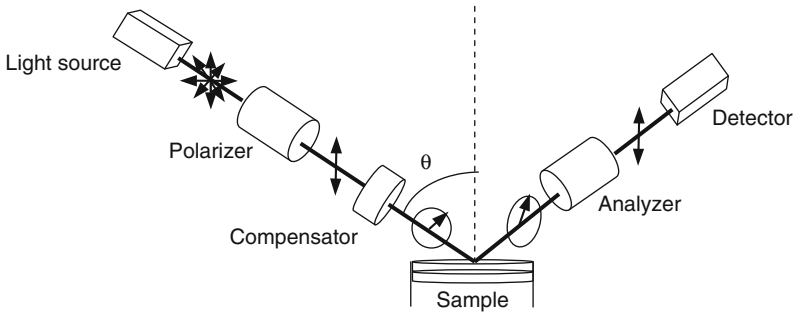
Ellipsometry, Fig. 2 Schematic of a single film on an optically thick substrate

Here, $\tilde{n}_0$, $\tilde{n}_1$, $\tilde{n}_2$ are the refractive indices of the ambient, film, and substrate, respectively; θ is the angle of incidence, λ is the optical wavelength, and t is the film thickness.

Ellipsometry Instrument Configuration

The ellipsometry instrument generally consists of a light source, a linear polarizer, a sample, a polarization analyzer, and a detector. Among various ellipsometry systems, the primary difference lies in the sections that polarize the incoming light beam and resolve the polarization state of the reflected or transmitted light beam. There are three main types of ellipsometry configurations: null ellipsometry, polarization modulation ellipsometry, and rotating element ellipsometry.

The configuration of a null ellipsometry is shown in Fig. 3. The light source produces a randomly polarized light, which is then sent through the polarizer, allowing the light of a preferred electric field orientation to pass and thus yielding a linearly polarized light that in turn transmits through the compensator. The compensator converts the incoming light to a circularly polarized light. The latter strikes the sample surface, reflects, becomes an elliptically polarized light, and then passes through the polarization analyzer. The amount of light allowed to pass depends on



Ellipsometry, Fig. 3 Schematic diagram of a null ellipsometry instrument

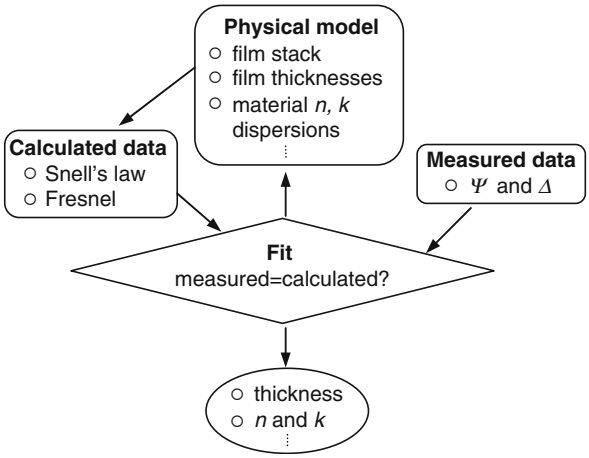
the analyzer orientation relative to the electric field "ellipse" coming from the sample. The detector catches the light and converts it to an electronic signal to determine the reflected polarization. This information is compared with the known input polarization to determine the polarization change caused by the sample reflection. The output results are the measured ellipsometric data Ψ and Δ.

Data Analysis Scheme

Ellipsometry is an indirect measurement technique. It is impossible to measure the sample parameters of interest (thickness, optical constants, etc.) directly with light. Instead, one measures the ellipsometric data Ψ and Δ. The measured data are then modeled in order to determine the parameters of interest. For a multi-layer material structure, the constructed model is a layered description that consists of the parameters of interest of each layer. The final model may give best fits to the measured data. The data analysis procedure is elaborated in Fig. 4.

First, a sample is measured, and then a model is constructed to describe the sample. The model is used to calculate the values of sample parameters from Fresnel's equations (see Eqs. 3–7), which describe each material with its thickness and optical constants. The calculated data are compared with the measured data. Any unknown material properties can then be varied to improve the agreement between the measured and calculated data. Finding the best match is typically achieved through the regression analysis. An estimator, mean-squared error (MSE), is used to quantify the difference between the measured and calculated data. The unknown parameters are allowed to vary until the minimum MSE is reached. A best-fit set of parameters corresponds to the lowest MSE if several local minima exist.

The accuracy of ellipsometry measurement depends mainly on how close the constructed model is to reality



Ellipsometry, Fig. 4 Flowchart of ellipsometric data analysis procedure

and whether the model structure is unique. In other words, it depends on the ways in which analysts deal with the data. Therefore, the final model should be examined for accuracy, sensitivity, and uniqueness. A good MSE is necessary but not sufficient to substantiate the correction of the final model. Other indicators, including the figure of merit (or 90% confidence), two-parameter correlation, and uniqueness test, also need be considered. Most importantly, the accuracy of the final model still requires alternate measurement techniques for verification.

Key Applications

Ellipsometry applicability depends mainly on three factors of a material structure: a) roughness at the surface of the sample or interfaces between layers of the sample; b) thin film thickness; and c) thin film uniformity within a measured spot.

Ellipsometry works best for smooth surfaces or interfaces; the surface or interfacial roughness should be less than about 10% of the wavelength of light used in the measurement. Additionally, ellipsometry works well for thin film characterization when the film thickness is within the preferred wavelength range of light. For example, the thin films from about 5 to 1,000 nm can be characterized with the light of about 500 nm wavelength. Finally, for ellipsometry analysis to be valid, it is best if the variation in thin film thickness is less than about 10% over the width of the measurement spot on the surface.

Determination of Substrate Optical Constants

Ellipsometry can be used to obtain the optical constants of both opaque and transparent substrates in a layered material structure. The accurate determination of substrate optical constants is critically important for further accurate characterization of thin films deposited on the substrates. The best technique for this type of analysis is to use the variable angle spectroscopic ellipsometry because it provides the ability to acquire ellipsometric data at more than one angle of incidence so that the measurement error can be averaged over the entire data set (Tompkins and McGahan 1999).

Opaque Substrate

Ellipsometry allows a unique determination of the optical constants of an unknown opaque substrate by directly inverting the measured ellipsometric data Ψ and Δ values at each measurement wavelength. However, the accuracy of the obtained optical constants may be affected by the presence of the substrate surface overlayer (generally due to roughness and/or native oxidation). In order to consider the effect of such an overlayer on the measurement of the substrate optical constants, one must estimate the roughness and/or the thickness of the oxidation layer. This estimated overlayer will then be included in the model used to accurately determine the optical constants of the substrate.

Transparent Substrate

The transparent substrate can well assist the characterization of metallic thin film because it allows one to acquire intensity transmission data in addition to ellipsometric data. However, a major problem in measuring a transparent substrate is the back-side reflections, which occur in the transparent substrate if both sides of it are smooth and the substrate is so thick that the reflections from the back side are incoherent with the desired reflection from the front side when entering into the detector. These unwanted reflections must be accounted for in the fitting model or suppressed by roughening the back side. One can also focus the beam small enough to separate the front and back reflections or suppress the back-side reflections via the index matching techniques applied to the back side opposite the measurement side (Synowicki 2008).

In the analysis, ellipsometric data are acquired at multiple angles of incidence, but only the Ψ data are fit using the Cauchy model to determine the refractive index of the substrate. If the extinction coefficient of the substrate needs to be determined, it is necessary to acquire transmission data at the normal incidence at the same spot on the substrate where the ellipsometry data are measured, and to fit the acquired data in the model in the UV spectral range.

Characterization of Very Thin Films

The characterization of a very thin film presents a challenge because its optical constants are often strongly correlated with its thickness. The degree of such correlation depends on the characterization method and the sample structure. Several methods can be used to reduce the coupling and ensure the unique solution for both thickness and optical constants of the very thin film by reducing the number of the unknown properties in the model and/or increasing the measurement information.

Common examples of such types of thin films include native oxides, gate oxides on microelectronic devices, organic films and a diamond-like carbon layer on the surface of magnetic recording medium, a metallic layer in a magnetic random access memory device, self-assembly monolayers and absorbing layers, and many others. The ellipsometry technique can be applied to study these thin films in the following areas.

Thickness of a Very Thin Transparent Film

When the optical constants of a very thin film are available, ellipsometry can be used to determine its thicknesses (less than 5 nm or so). One can use the published optical constants from the literature, or the corresponding bulk optical constants for the thin film, so that only the film thickness is unknown. In short, the thickness measurement of a very thin film can be performed if its optical constants are accurately known in advance. In addition, the phase information from Δ is sensitive to thin film thickness down to the sub-nanometer level, making the thickness of a monolayer film measurable. This feature is useful for the in situ process control of dielectric thin film deposition where the thickness change of the thin film in the sub-nanometer scale needs to be detected.

Existence of Multiple Films

Ellipsometry can be used to distinguish one thin film material from another when their optical constants are significantly different. It is easy to identify a thin metallic film from a thin dielectric film because of the different shapes of the optical constant dispersions; however, it is difficult to distinguish a very thin dielectric film from another such film due to similar shapes of their optical constant dispersions. Similarly, it is easy to determine the thickness of each layer if the stacked thin films have different optical constant dispersions, for example, a thin aluminum oxide film deposited on a thin layer of silicon nitride a silicon wafer. On the other hand, it is difficult to determine the individual thicknesses if the stacked thin films have similar dispersions, for instance, a thin aluminum oxide film deposited on a thin aluminum on a silicon wafer.

Thickness and Optical Constants of a Thin Metallic Film

Except for the strong correlation between the thickness and optical constants, most thin metallic films have rather complicated dispersion of optical constants, further complicating the characterization of thin metallic films. When properly implemented, ellipsometry measurements can simultaneously and uniquely determine the thickness and optical constants of thin metallic films. Several effective methods used for the unique and sensitive measurements of thin metallic films are suggested as below (Tompkins and McGahan 1999; Hilfiker et al. 2008).

Optical Constant Parameterization

This method uses a mathematic dispersion equation to describe the optical constants of thin metallic films. This helps to significantly reduce the total number of unknown properties in the model and ensure the resulting optical constants to maintain Kramers-Kronig consistency with a physically reasonable shape. Usually, the optical constants of thin metallic films are described by a summation of "oscillator" terms (Lorentz, Gaussian, Tauc-Lorentz, etc.) due to their multiple absorption features. This makes this method less effective in reducing the correlation. Therefore, optical constant parameterization is often best utilized in combination with other methods mentioned below.

Interference Enhancement

This method measures a special type of sample structure where a thick dielectric layer is between a thin metallic film and a silicon wafer. The thick dielectric layer significantly enhances the interaction between light and the thin metallic film at multiple incident angles to provide new information. When data from multiple angles are fit simultaneously, a unique solution for the thickness and optical constants can be obtained. The effectiveness of this method depends on the thicknesses of both thin metallic layer and underlying thick dielectric layer, and the latter thickness must be sufficient to provide interference in the former optical dispersion.

Combined Ellipsometry and Transmission

This method combines ellipsometry with intensity-based optical measurements from the same thin metallic film. The extra transmission measurement helps break the correlation between its thickness and optical constants. When two types of data are fit simultaneously, unique results are generated. The main limitation of this method is that it requires a transparent substrate and accurate transmission measurement. In addition, the substrate optical constants must be carefully characterized in advance because any small absorption in the substrate may affect the measurement of the optical constants of the thin metallic film.

Multiple Sample Analysis

This method measures multiple samples of the same metallic film but different thickness. The data from multiple samples are fit simultaneously using a common set of the optical constants. Such simultaneous analysis can reduce the correlation and thus uniquely yield the optical constants and thicknesses of multiple samples at one time. However, this technique requires the optical constants of each sample to be nearly identical and a large number of samples; thus it is also time consuming.

The abovementioned methods work well on a variety of thin metallic films, but have their own limitations. In practice, they are best used in combination.

Determination of Effective Optical Constants of Mixed Materials

Ellipsometry can be used in combination with an effective medium approximation (EMA) model to determine the effective optical constants of a mixed material. There are three types of EMAs: the Lorentz-Lorenz, Maxwell-Garnett, and Bruggeman models. These EMA models can be used for any combination of transparent and metallic materials. The most commonly used EMA model is the Bruggeman EMA model (BEMA). It is a mathematical equation describing the relationship between the effective optical constants of a mixed material and the optical constants of its constituent materials by

assuming a) each material retains bulk-like properties and b) materials are sufficiently mixed over the macroscopic area, but grain sizes are smaller than 10% of the light wavelength to be considered homogenous.

In the BEMA model, the effective dielectric function of a mixed material is obtained by solving the following equation:

$$f_1 \frac{\tilde{\varepsilon}_1 - \tilde{\varepsilon}}{\tilde{\varepsilon}_1 + 2\tilde{\varepsilon}} + f_2 \frac{\tilde{\varepsilon}_2 - \tilde{\varepsilon}}{\tilde{\varepsilon}_2 + 2\tilde{\varepsilon}} + f_3 \frac{\tilde{\varepsilon}_3 - \tilde{\varepsilon}}{\tilde{\varepsilon}_3 + 2\tilde{\varepsilon}} = 0 \qquad (8)$$

where $\tilde{\varepsilon}$ is the effective dielectric function of the mixed material, $\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, \tilde{\varepsilon}_3$ are the well-defined dielectric functions of each material, and f_1, f_2, f_3 are the volume fractions of each material, respectively. This equation is valid for a mixture with three materials, or with two materials by fixing f_3 at zero (Tompkins and McGahan 1999).

For some mixed materials, such as polysilicon, PECVD nitrides, oxides, and silicon oxynitrides, their effective optical constants can be adequately determined using the BEMA model.

Evaluation of Surface or Interfacial Roughness

This application does not aim to characterize the surface or interfacial roughness itself but rather models the surface or interfacial effects properly in order to obtain the accurate precise measurements of films underneath the surface or above and below the interface. When the surface or interfacial roughness is treated as a discrete thin film that consists of a mixture of two materials above and below the surface or interface, a BEMA model is usually used to model this layer by mixing the optical constants of the material with "void" and allowing its thickness to be adjustable. For convenience, the roughness in the BEMA model is usually equivalent to a mixture of 50% material and 50% void; therefore, the thickness of the roughness layer can be determined individually.

It is particularly important to model the surface roughness in this manner when determining the optical constants of the substrate with a microscopic surface roughness. In this case, the thickness of the surface roughness layer should be less than about 10% of the light wavelength.

Cross-References

► Diamond-Like Carbon Coatings
► Nanolubricants
► Polymer Nanolayers
► Surface Characterization and Description

References

J.N. Hilfiker, N. Singh, T. Tiwald, D. Convey, S.M. Smith, J.H. Baker, H.G. Tompkins, Survey of methods to characterize thin absorbing films with Spectroscopic Ellipsometry. Thin Solid Films **516**, 7979–7989 (2008)

R.A. Synowicki, Suppression of backside reflections from transparent substrate. Phys. Stat. Sol. (c) **5**(5), 1085–1088 (2008)

H.G. Tompkins, W.A. McGahan, *Spectroscopic Ellipsometry and Reflectometry: A User's Guide* (Wiley, New York, 1999)

J.A. Woollam Co., Inc, *Guide to Using WVASE32®* (Wextech Systems, New York, 2000)

Elliptical Contact of Surfaces

► Hertz Theory: Contact of Ellipsoidal Surfaces

Empirical Fractals

► Fractal Nature of Surfaces

End Face Seal

► Mechanical Seals

Endurance Limit

► Fatigue Limit

Energetic Condensation

► PVD: Ion Plating

Energy Description of Sliding

► Thermodynamic Modeling of Wear

Energy Dissipation and Temperature Determination in Lubricated Contacts

Francis E. Kennedy
Thayer School of Engineering, Dartmouth College, Hanover, NH, USA

Definition

Frictional energy dissipation is the process by which mechanical energy is dissipated into internal energy, in the form of heat, in a sliding or rolling/sliding contact. In this entry it is assumed that the solid surfaces in the contact are separated completely by a fluid film (hydrodynamic or elastohydrodynamic lubrication) and that all energy dissipation takes place within the fluid film.

Scientific Fundamentals

Introduction – Technological Importance of Heating in Lubricated Contacts

Whenever a fluid lubricant passes between two solid surfaces in relative motion, there is some resistance to the fluid motion. That resistance is caused by shear stresses within the moving fluid, and it results in a dissipation of kinetic (mechanical) energy or work. Nearly all of the dissipated energy is transformed into internal energy in the fluid film in the form of heat, causing an increase in the temperature of the lubricant, and frequently a temperature increase in the solid components that are separated by the fluid film. The increased temperature can have many effects on the behavior of the tribological contact, including:

- Changes in viscous properties of liquid lubricants
- Significant reduction in lubricant film thickness at higher lubricant temperatures
- Degradation, e.g., oxidation, of the lubricant
- Deterioration or removal of boundary lubricant films
- Thermal deformation of contacting solid materials
- Scuffing or scoring of contacting components, resulting from lubricant deterioration and softening of contacting materials

In order to more accurately predict the behavior and performance of a lubricated contact and to avoid the problems caused by thermal effects, it is important to be able to determine the temperature changes resulting from frictional energy dissipation.

Energy Dissipation in Fluid Lubricated Contacts

If a full film of fluid separates the contacting solids, the frictional energy dissipation takes place within the fluid film. This is true whether the lubrication condition is hydrodynamic or elastohydrodynamic. Friction in this case is dictated by the fluid shear stress and its interaction with the surfaces of the contacting solids. The shear stress is determined by the viscosity of the fluid and the velocity gradient within the fluid film near the surface. Thus,

$$\tau = \eta \frac{\partial v}{\partial y} \tag{1}$$

This shear stress, and the accompanying fluid velocity, gives rise to work or energy dissipation in the fluid. That energy dissipation is governed by the energy equation, which is a mathematical statement of the conservation of energy within a volume element of a flowing fluid. For the case of an incompressible Newtonian fluid with constant thermal properties, the energy equation is as follows (Szeri 1998):

$$\rho C_p \left(\frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y} + w\frac{\partial T}{\partial z} \right) = K \left[\frac{\nabla^2 T}{\nabla x^2} + \frac{\nabla^2 T}{\nabla y^2} + \frac{\nabla^2 T}{\nabla z^2} \right] + \Phi \tag{2a}$$

where

$$\Phi = 2\eta \left[\left(\frac{\partial^2 u}{\partial x^2} \right) + \left(\frac{\partial^2 v}{\partial y^2} \right) + \left(\frac{\partial^2 w}{\partial z^2} \right) \right] + \eta \left[\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right)^2 + \left[\left(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y} \right)^2 \right] + \left[\left(\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z} \right)^2 \right] \right] \tag{2b}$$

Φ is the heat of dissipation or the dissipation function; it gives the heat generated per unit time and unit volume owing to internal friction in the fluid element.

Note, the energy equation (2a) may be written as

$$\rho C_p \left(\frac{\partial T}{\partial t} + V \cdot \nabla T \right) = K \nabla^2 T + \Phi \tag{3}$$

so the energy equation reduces to Fourier's law for heat conduction if there is no energy dissipated within the fluid (e.g., if the fluid is at rest).

Different formulations for the energy equation and the dissipation function are available for compressible fluids or for turbulent flow conditions (e.g., Szeri 1998).

The determination of the temperature field within the fluid requires the solution of (2a), using knowledge of the velocity field in the fluid, along with appropriate thermal boundary equations. The velocity field requires solution of

the fluid momentum equations, usually Reynolds equation. This important equation is presented in another entry in this encyclopedia (see entry "► Reynolds Equation").

Thermal Boundary Conditions

For a more complete discussion of thermal boundary conditions in lubricated contacts, see the entry "► Thermal EHL Theory."

In general, thermal boundary conditions are required at each of the four bounding surfaces of the lubricating film:

- Fluid Inlet
  In some cases the temperature of the lubricant is known at the inlet to the bearing, but in many other cases a cool supply fluid is mixed with hotter lubricant that is being carried into the bearing by the runner or shaft. This mixing condition may be complicated by such factors as reverse flow or starvation, so the analysis of the inlet condition may be as complicated as the thermal analysis of the fluid within the bearing itself (Pinkus 1990).
- Fluid Outlet
  It can usually be safely assumed that there is no temperature gradient in the flow direction at the outlet of a bearing. However, cavitation in the outlet region may require special consideration (Pinkus 1990).
- Upper Solid Interface (at $y = h$) and Lower Solid Interface (at $y = 0$)
  In some cases the temperatures of the solid surfaces could be known and prescribed, giving isothermal bearing surfaces. This is convenient for computation, but in practice it is difficult to know the constant temperature value a priori. In other cases it could be assumed that there is no heat flow into the bearing and shaft (or runner) surfaces. This is the adiabatic condition, and it requires that the temperature gradient in the y-direction go to zero at the upper and lower surfaces; i.e., $\left(\frac{\partial T}{\partial y}\right)_{y=0} = \left(\frac{\partial T}{\partial y}\right)_{y=h} = 0$. This simplifying assumption has been found to be acceptable in many cases, but a better condition for most cases is to couple the thermal analysis of the fluid with that of the solid at the upper and lower boundaries and to prescribe that the temperature and heat flux (or temperature gradient) be continuous across the two interfaces. This requires that the heat flux into the solid surface be related to the temperature gradient in the fluid by a simplified version of Fourier's law, as follows:

$$q_{y=0} = K\left[\frac{\partial T}{\partial y}\right]_{y=0} = q_1 \quad \text{and} \quad q_{y=h} = -K\left(\frac{\partial T}{\partial y}\right)_{y=h} = q_2 \tag{4}$$

where the fluid film thickness is h and q_1 and q_2 are the heat fluxes into the upper and lower solid surfaces, respectively. The two heat fluxes, which may be a function of position, are the same as those used in the solution for the temperature field in the solids; this is discussed in the accompanying entry "► Contact Temperature of a Stationary Solid Surface" or "► Contact Temperature of a Moving Solid Surface," depending on the kinematics of the solid surfaces relative to the point of contact. Therefore, complete determination of the temperature field in a fluid lubricated contact usually requires a coupled solution of the energy equation, Reynolds equation, and the equations for solid surface temperatures. This is generally accomplished by a numerical analysis.

Key Applications

There are numerous lubricated tribological components in which frictional energy dissipation and the resulting frictional heating are important. Among the most important effects of frictional heating on tribological phenomena are the following:

- It has long been known that the viscosity of lubricants in journal or thrust bearings is significantly affected by the lubricant temperature (Kennedy et al. 1997). Heating of the lubricant within the lubricated conjunction plays a major role in determining the effective temperature of the lubricant, and therefore its viscosity. For this reason, accurate evaluations of the fluid film thickness and pressure distribution within a hydrodynamic bearing, particularly for heavily loaded or high speed bearings, frequently require a thermohydrodynamic analysis (Khonsari 1997). A complete thermohydrodynamic analysis includes an analysis of the temperature distribution within the fluid film (by solving (2a) above), as well as a solution of Reynolds equation or equivalent for the pressure and film thickness distributions (Szeri 1998). Frequently, it is also necessary to couple the thermal analysis of the fluid to a solution for the surface temperatures on the solid surfaces, as discussed above.
- Heavily loaded concentrated contacts, such as those found in rolling element bearings, cams and tappets, gears, or traction drives, are lubricated by an elastohydrodynamic fluid film. Shearing of the lubricant film in these contacts causes heating that affects the thickness of the lubricant film, traction characteristics of

the contact, and sometimes the deformation of the solid components. Accurate evaluation of the tribological characteristics of many such contacts requires a thermal elastohydrodynamic analysis, which involves the coupled solution of the Reynolds, energy and elasticity equations, along with expressions for the temperature- and pressure-dependence of the lubricant's viscous properties, and usually an analysis of surface temperatures in the solid components (Cheng and Sternlicht 1965; Kim et al. 2001). Thermal effects in concentrated contacts can cause a significant reduction in lubricant film thickness, especially in concentrated contacts that involve a significant amount of sliding.

- The effectiveness of boundary lubrication is strongly dependent on the temperature seen by the boundary lubricant within the contact. Friction and wear with lubricants containing physically or chemically adsorbed additives deteriorate rapidly when the contact temperature reaches a critical value at which the additive molecules desorb (Fein et al. 1959). The role of heating in the lubricant transition has been well established (Ettles et al. 1994). Lubricants containing extreme pressure (EP) additives rely on high contact temperatures to promote the formation of protective films on the contact surfaces (Spikes and Cameron 1974).
- When the fluid film thickness in a lubricated conjunction is too small to prevent contact between asperities of the solid surfaces, a condition of mixed or partial lubrication prevails. This occurs particularly when the applied load is very high or when the relative velocity between the solid components is quite small, and/or when the solid surfaces are rough. As solid-to-solid contact occurs, some of the normal load is carried by these contacts and frictional energy dissipation within those contacts becomes larger than fluid frictional losses. In such cases, a major determinant of contact temperatures is frictional heating of the rough solid surfaces. This is discussed in the entry entitled, "► Contact Temperatures on Coated or Rough Solid Surfaces." Analyses of flash temperature rises in cases of mixed lubrication have been carried out by Zhu and Hu (2001) and others.
- Scuffing or scoring is an important failure mode for lubricated sliding or rolling/sliding components such as gears or cams. Most models for scuffing failure are based on a critical contact temperature that causes scuffing by either weakening the lubricant film so it is unable to support the load (Dyson 1975) or increasing the shear stress in the film to a limiting value (Jacobson 1990).

Cross-References

- ► Contact Temperature of a Moving Solid Surface
- ► Contact Temperature of a Stationary Solid Surface
- ► Contact Temperatures on Coated or Rough Solid Surfaces
- ► Reynolds Equation

References

H.S. Cheng, B. Sternlicht, A numerical solution for the pressure, temperature, and film thickness between two infinitely long, lubricated rolling and sliding cylinders. Trans. ASME J. Basic Eng. **5**, 695–707 (1965)

A. Dyson, Scuffing – a review. Tribol. Int'l **8**, 77–87 (1975)

C.M.M. Ettles, O.S. Dinc, S.J. Calabrese, The effect of frictionally generated heat on lubricant transition. Tribol. Trans. **37**, 420–424 (1994)

R.S. Fein, C.N. Rowe, K.L. Kreuze, Transition temperatures in sliding systems. ASLE Trans. **2**, 50–57 (1959)

B. Jacobson, Mixed lubrication. Wear **136**, 99–116 (1990)

F.E. Kennedy, E.R. Booser, D.F. Wilcock, Tribology, lubrication and bearing design, in *CRC Handbook of Mechanical Engineering*, ed. by F. Krieth (CRC Press, Boca Raton, 1997), pp. 3–128–3–169

M. Khonsari, A review of thermal effects in hydrodynamic bearings. ASLE Trans. **30**, 19–33 (1997)

H.J. Kim, P. Ehret, D. Dowson, C.M. Taylor, Thermal elastohydrodynamic analysis of circular contacts. Proc. Instn. Mech. Eng. **215 J**, 339–352 (2001)

O. Pinkus, *Thermal Aspects of Fluid Film Lubrication* (ASME Press, New York, 1990)

H.A. Spikes, A. Cameron, Additive interference in dibenzyl disulphide extreme lubrication. ASLE Trans. **16**, 283–289 (1974)

A.Z. Szeri, *Fluid Film Lubrication* (Cambridge University Press, Cambridge, UK, 1998)

D. Zhu, Y.Z. Hu, A computer program package for the prediction of EHL and mixed lubrication characteristics, friction, subsurface stresses and flash temperatures based on measured 3-D surface roughness. Tribol. Trans. **44**, 383–390 (2001)

Energy Dissipation in Ultraprecision Machining

► Tribological Aspects of Ultraprecision and Nanometric Cutting

Engine Condition Monitoring Based on Oil Analysis

Bernardo Tormos
CMT-Motores Térmicos, Universitat Politècnica de València, Valencia, Spain

Synonyms

PdM – predictive maintenance based on oil analysis

Definition

There are four primary methods of monitoring equipment's condition: vibration, thermography, acoustic emission, and oil analysis. The first two present several limitations for application in reciprocating internal combustion engines; the last one is the most commonly applied technique in these cases. Thus, oil analysis is one of the most important techniques applied for condition monitoring and must be understood as a diagnostic maintenance tool. Used oil testing provides interesting information about the condition of the oil, the equipment in which it is being used, and oil suitability for further use. Used oil analysis is comparable to a medical analysis with a blood test. Like blood, lubricating oil contains a good deal of information about the lubricated system in which it circulates.

Scientific Fundamentals

Engine Oil Analysis Objectives

Data from used oil analysis can be used to assist in identifying incipient mechanical failures or in determining the quality and useful life of the oil. Thus, potential engine component failure and premature lubricant failure may be detected prior to a major engine failure and subsequent expensive repairs and equipment unavailability (probably even more expensive) may be avoided. Oil analysis may also be used to identify improper maintenance procedures and unsatisfactory equipment parts or components.

There are three main areas that must be covered by oil analysis:

- An assessment of the oil condition revealing when the lubricating oil is ready to be changed, or if it is fit for further service. Important savings can be obtained by maximizing oil drain periods with the confidence that condition monitoring provides.
- Detection of potential contaminants in lubricating oil that can be indicative of engine subsystem malfunctions.
- Detection of abnormal wear.

Applying a proper and well-defined oil analysis program, multiple benefits can be obtained (Macián et al. 2003):

- Avoidance of catastrophic failures and consequent related costs
- Reduction in unscheduled downtime
- Effective maintenance scheduling
- Improved engine reliability
- Reduction in maintenance costs
- Maximization of oil drain periods
- Reduction in fuel consumption

The most critical factor in oil analysis is, perhaps, the interpretation of oil analysis results. A proper interpretation of the oil test should be related with three areas mentioned before: oil condition, oil contamination, and wear. Diagnosticians should draw from different sources to obtain a proper interpretation of oil analysis results, including Original Equipment Manufacturers' (OEM) recommendations, oil baselines and typical industry literature, and their experience and mechanical expertise, to correctly interpret test results into a correct picture of equipment condition. Usually, a oil analysis report includes a recommendation outlining any necessary corrective maintenance actions.

Engine Lubrication

Internal-combustion engines can be considered as among the most difficult lubricated systems. Loads and temperature supported by engine oil can be on the same level that can be found in other lubricated systems but, as an additional effect, the presence of combustion products and residues that may contaminate the lubricant becomes a much more difficult process for them.

Type of combustion (spark-ignition or compression-ignition), type of cycle (two strokes/four strokes), and, related with it, fuel type and quality will affect lubrication requirements and specific problems (Caines and Haycock 1996).

The majority of combustion products are exhausted to the atmosphere, but a significant proportion leaks past the piston rings (an effect known as blow-by), contaminating the engine oil and causing adverse effects. Combustion products appearing on blow-by gases are carbon dioxide (CO_2), unburned hydrocarbons (UHC), nitrogen oxides (NO_x), water, and acidic components.

The water vapor that passes to the relatively cooler crankcase via blow-by can condense to liquid form and settle on different engine mechanisms, causing rust, or mix with the lubricant, forming sludge (formerly quite common in cooler engine parts as rocker cover). Dispersing additives are used on engine lubricants to avoid these effects. The effect of high temperatures on hotter parts of the engine helps to flash off the water into oil and exhaust it via positive crankcase ventilation.

Acidic components, usually related to the impurities contained in the fuel, the most important one being sulfur, are produced as a consequence of the combustion process and conditions related in the combustion chamber. These acids can produce corrosion and corrosive wear in engine

parts, but also act as a catalyst both for the degradation of the oil and also for gum and varnish formation. Today, sulfur content in fuels is quite reduced for heavy-duty applications (for instance, European standard EN 590 limits sulfur content to 10 ppm) but remains an important problem in large engines fuelled with residual fuels or gases. Alkaline additives are used in the oil to neutralize the acidic components and avoid the harmful effects on engine and lubricant.

An additional contaminant, mainly in diesel engines, is soot (carbon particles resulting from the incomplete fuel combustion), which also appears in the blow-by. This product mixed with water can produce gray/black sludge in the crankcase and oil passages, resulting in lubricating problems. Soot can also adhere to varnish deposits and accelerate the buildup of carbon if the varnish-forming tendency is not controlled. Two additional negative effects can be related to soot in oil: viscosity increases and a contribution to an increase in engine wear related with soot's abrasive character. The employment of exhaust gas recirculation (EGR) in modern diesel engines has lead to a clear increase in the soot load to be supported by engine oil and consequently an increase in oil requirements for this application. Dispersant additives are used to keep the soot particles finely divided and off of engine surfaces.

Test for Evaluation of Oil Condition

Next, the most typical tests applied for used engine oil condition assessment are presented.

Viscosity

Viscosity is undoubtedly one of oil's most important physical properties. Thus, it is one of the first parameters measured to determine oil condition in many applications. For engines, viscosity, or more precisely SAE viscosity grade, is used by engine manufacturers to define the engine fresh oil viscosity to be used in their engines.

The viscosity of used engine oil, usually measured at 100°C or 40°C (ASTM D445), can present significant changes, increasing or decreasing over the original value of fresh oil. Reasons for these changes are diverse and are presented in Table 1.

Taking into account multiple situations that can be found, and even situations where antagonist effects can be compensated for, viscosity cannot be used as unique condition estimator.

To define a normal trend in engine oil viscosity within an oil use period, which type of oil is used must be taken into account. Thus, for multigrade oil (containing VI improvers), the normal trend is a decrease in viscosity in the oil use period. For single-grade oils, the normal trend is an increase in viscosity in the oil use period (Fig. 1).

Engine Condition Monitoring Based on Oil Analysis, Table 1 Typical causes for oil viscosity changes

	Viscosity decreases	Viscosity increases
Lubricant structure changes	– Shearing of the viscosity index improver – Breakdown of lube molecules	– Oxidation – Polymerization – Evaporation losses
Contamination	– Fuel dilution – Oil mixing (lower viscosity)	– Build-up of suspended insoluble matter (mainly soot) – Water – Oil mixing (higher viscosity)

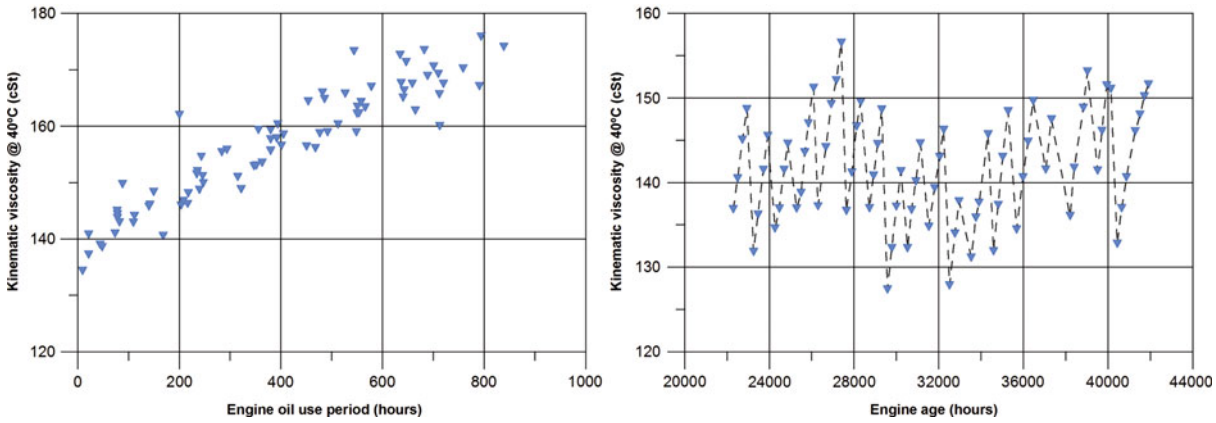
Limits applied for viscosity variations in used oil analysis are usually recommended by OEMs, but, as a general reference, the change to a different SAE grade or ±30% variation from fresh oil value can be used.

Base Number (BN)

As has been mentioned, the acidic environment supported by engine lubricating oil tends to negatively affect the lubrication process and must be neutralized in order to reduce corrosion. Basic chemical additives are used for this aim. The alkalinity of oil is measured by titration through an acid and expressed in mg KOH/g. There are two commonly accepted measurement methods: ASTM D2896, often used for fresh oils, and ASTM D4739, which is often applied to measure the reserve alkalinity of used oils (differences between them are mainly due to the acid and solvent in which the oil is dissolved and used to run the test).

The comparison between the alkalinity reserve of the fresh oil and that of the used oil allows the determination to be made of whether the used oil is still capable of neutralizing acid residues. Too low a BN of a used oil can be due to diverse causes: heavy oxidation of the oil; oil that has been in service for too long; insufficient oil level; a defective cooling system producing overheating; excessive blow-by ratio; use of a fuel containing a high sulfur content; use of an inappropriate lubricant; or contamination of the oil by fuel or water.

Today, looking at the mandatory levels for sulfur content in diesel fuels (in heavy-duty applications) or



Engine Condition Monitoring Based on Oil Analysis, Fig. 1 Viscosity evolution in a monograde oil versus oil use period (*left*) and versus engine use period (*right*) including oil changes

unleaded fuel for gasoline engines, it may seem that BN levels for fresh oils are too high. This is a result of the additives (detergents, dispersants, or some antioxidants), the main target of which is not acid neutralization, which contribute to the alkalinity reserve.

Limits applied for BN variations in used oil analysis are usually recommended by OEMs. A drop of 50% from fresh oil value can be used as a general reference.

Acid Number (AN)

The AN test is a method used to estimate lubricant degradation, usually as a consequence of effects such as oxidation, contamination, and additive depletion. The acidity of the oil is usually measured by titration through a base and is expressed in mg KOH/g (ASTM D664).

Engines that support higher combustion temperatures or those fuelled with residuals fuels containing high levels of impurities (sulfur, vanadium, water, etc.), and consequently higher oil oxidation rate, can be monitored using this parameter.

Examples of BN and AN evolution in used engine oils can be seen in Fig. 2.

As a general rule of thumb, when AN is equal or higher than BN for a sample of used oil, this can be considered an indication for oil change.

Oxidation/Nitration Products

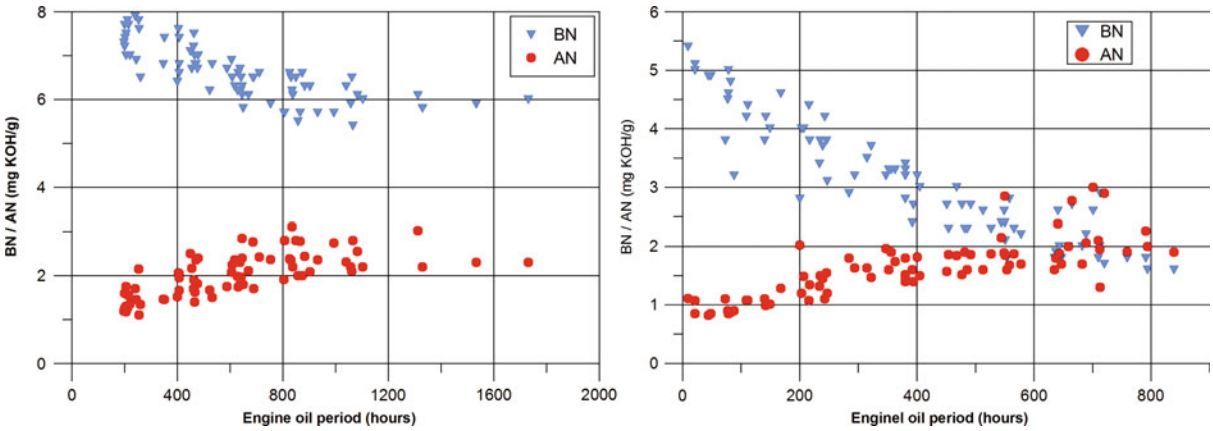
Oxidation is the primary mechanism of lubricant degradation. Oxidation of lube oils occurs when the hydrocarbon constituents of lube oil combine chemically with oxygen to form a wide variety of oxidation products such as aldehydes, ketones, hydroperoxides, and carboxylic acids. The rate at which oil molecules react with oxygen depends on different factors, the most critical being temperature.

Carboxylic acids contribute to the acidity of the engine oil and deplete its basic reserve as neutralization takes place. Sludge and varnish formation is other potential problem, as a consequence of oxygenated reaction by-products such as hydroperoxides and carboxylic acids combining to form larger molecular species (polymerization). Thus, the effect of prolonged oxidation is that chemically the oil becomes acidic, causing corrosion and a physical increase in viscosity.

In addition to oxidation products, nitration products are also formed when organic compounds are exposed to high temperatures and pressures in the presence of nitrogen and oxygen. These are generally in the form of nitrogen oxides such as NO and NO_2. In addition to causing oil thickening and some of these products being acidic, nitration products are the major cause of the buildup of varnish.

An increase in the nitration level of engine oil can indicate an abnormal fuel/air ratio, improper spark timing, or excessive blow-by ratio. It can also reflect severe operating conditions, such as high loads or low operating temperature.

As has been mentioned, different effects appear when oil is oxidized/nitrated, thus, using parameters previously presented such as viscosity, BN, or AN, an estimate of oxidation level can be obtained. In used engine oil analysis today, the most common technique to quantify oxidation/nitration is Fourier transform infrared (FTIR) spectrometry.



Engine Condition Monitoring Based on Oil Analysis, Fig. 2 BN and AN evolution versus oil period for a type of engine working with different oils and in different types of services

OEMs usually offer maximum levels for these parameters in engine used oil ranging between 20 and 30 Abs/cm, but trend analysis of oxidation/nitration levels could be more interesting for detection of potential problems.

Additive Depletion

Oil additives have a limited lifetime because they are consumed as oil ages. As it has been mentioned, alkaline additives are used up by neutralizing corrosive acids produced by the combustion process, and a measure of oil reserve can be made using the BN test. Antioxidants, which are sacrificial additives in nature, deplete with time before the base oil begins to oxidize. It is accepted that when 70–80% of the antioxidant reserve is depleted, physical changes within the oil begin to occur. An estimation of antioxidant reserve can be obtained using different tests. One of the most commonly used in engine oil analysis is the comparison of fresh and used oil sample peak areas' voltammograms (measurement of an electric current with respect to varying an applied voltage), also known as "remaining useful life." Dispersant additives, the main function of which is to suspend contaminants (soot, water, etc.) and prevent them from agglomerating and depositing on the hotter parts of the engine, become "loaded" as oil ages. A blotter spot test can be used as an estimation of lubricant dispersancy performance.

Test for Evaluation of Oil Contamination

The causes of engine oil contamination are diverse and can be classified according to external and internal sources. As external contamination, ingested silicon (dust) can be considered the most important. Of internal origin, soot, fuel, water, and anti-freeze are considered.

Fuel Dilution

Fuel dilution of the oil is unfavorable for the engine, since it involves a viscosity reduction and, consequently, resistance of the oil film (Fig. 3). The principal causes of dilution are a defective fuel injection system; a defective air inlet (obstructed air filter), thus affecting fuel/air ratio and consequently combustion; and incomplete combustion due to too low a working temperature or insufficient compression (as a consequence of excessive ring-liner wear or unregulated valves).

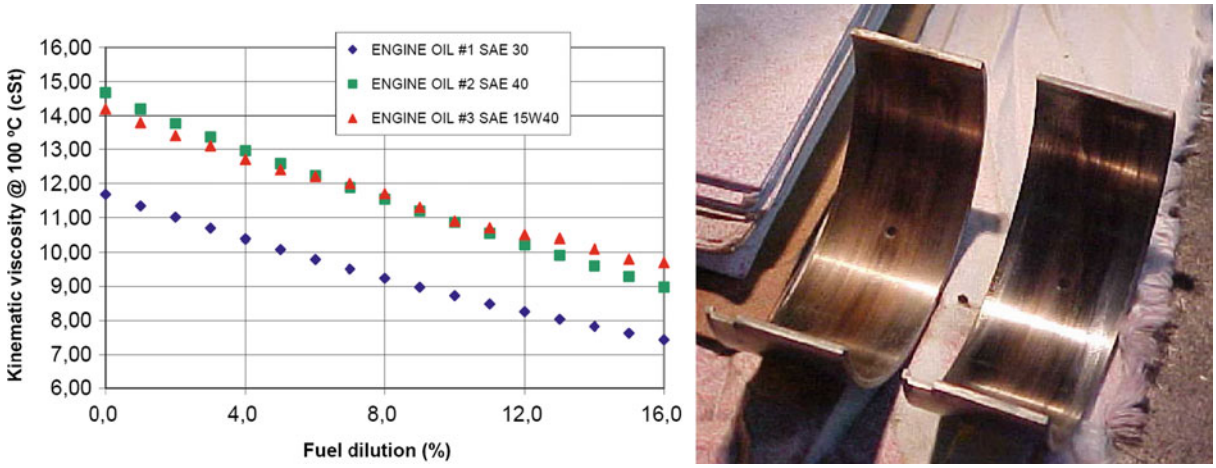
Dilution of used engine oil can be measured precisely by gas chromatography (GC) or by Fourier transform infrared spectroscopy (FTIR). Others methods used, with lower accuracy, are related to the reduction of oil flash point or viscosity correlations.

General limits for fuel dilution in engine oil range from 3% to 5%, depending on the reliability considered. Over this limit, total or partial oil change is recommended.

Water/Anti-freeze

Water contamination problems can stem from two different sources:

- As an external contamination caused by condensation due to too low a working temperature, "stop and go" in-service usage (as in urban transport service), or defective crankcase ventilation.
- Coolant leakages, which are considered among the major sources of catastrophic engines failures in service. There are many different ways coolant can leak into engine oil: deteriorated seals, a cracked block, cavitation erosion of cylinder liners, so on.



Engine Condition Monitoring Based on Oil Analysis, Fig. 3 Effects on viscosity measurements as a consequence of fuel dilution (*left*) and wear generated on engine main bearings as a consequence of oil dilution (*right*) in a real case

Negative effects appear when oil is contaminated by water, such as with additive hydrolysis, viscosity variations, etc. Additionally, water that is evaporated and exhausted is more difficult to detect. In these cases, other fingerprints or clues must be found. Typical compounds or additives used for anti-freeze should be considered (such as glycol, sodium, boron, etc.), but it must be taken into account that additive treatments used in anti-freeze formulations vary considerably between aftermarket suppliers and OEMs

Water content of oil is usually measured by the Karl Fisher apparatus for accurate quantification, although sometimes screening tests, such as the crackle test, are performed on a previous stage. General limits for water contamination in engine oil range from 0.2% to 0.3% depending on engine type and service. If signs of coolant leakages are detected, quick and proper action must be taken in order to avoid subsequent failures.

Soot

Soot, which is approximately 98% of carbon by weight, is formed during the combustion process and exhausted to the atmosphere in engine exhaust gases. Soot is considered a negative emission and is strongly regulated all over the world. Some of the soot formed enters the crankcase with combustion gas via blow-by. Soot particles present a reduced size but tend to agglomerate to form larger particles. Consequences for lubrication related to soot load are viscosity increase, dispersancy reduction, deposits, and wear.

Measurement of soot can be accomplished using different methods, including thermogravimetric analysis (TGA), which estimates the concentration of soot as a percentage by weight, this technique is accurate but quite expensive; FTIR analysis (which has a lower cost and good correlation with TGA measurements); and insolubles tests (using pentane or toluene). The most commonly used method currently is probably FTIR analysis.

Limits associated with this parameter are related to engine type and characteristics (diesel vs. gasoline, turbocharged, EGR, etc.) and type of lubricant. As a general reference, soot load greater than 3% requires serious evaluation.

Silicon (Dust)

Silicon is the most important indicator of dust entry into an engine. There have been several studies of the causes of premature wear in components, and results vary from study to study, but one thing is clear: external contamination of lube oil by silicon is a major cause of accelerated wear.

Particles of airborne dust vary in size, shape, and abrasive properties, and in an engine the ingress of atmospheric dust takes place primarily through the air intake. Dust particles not retained by filters (due to low filter efficiency, filter damage, or air intake damage) pass to the lubricating oil via blow-by and remain suspended in it. Particles similar in size to the oil film clearance in the main lubricated parts of the engine do the most damage. Once a dust particle enters the oil film, it forms a direct link between the two surfaces, nullifying the effect of the oil film. Thus, the immediate consequence is a "scratching" of the surface as the particle is dragged and rolled across the surfaces. A second consequence is

that the particle introduced between the two surfaces changes the loading of the surface from an even distribution to a load concentrated on the particle with a huge increase in pressure at this point. The increase in pressure causes a deflection of the surface, which will eventually result in metal fatigue and the surface breaking up.

As soon as a dust entry problem occurs there is an increase in the silicon concentration in the oil and an acceleration of the wear pattern (iron, tin, lead, etc.). If oil samples are taken periodically, dust entry can be detected at a very early stage and an effective corrective action can be taken, assuring the life of the component and avoiding extra maintenance costs.

In summary, the proper evaluation of silicon should take place using trend analysis rather than by establishing a static alarm level. Most engine manufacturers recommend trend analysis but also offer a maximum alarm level. In order to obtain a proper diagnosis, engine (or vehicle) environment and maintenance procedures should also be taken into consideration.

Engine Wear Evaluation

Wear metals are generated by friction between moving metallic surfaces in mechanical systems. Despite lubrication, wear-metal generation occurs to some degree in all oil wetted systems, and the lubricant serves as a repository for the wear metals. Wear metals may also be generated by corrosive action, cavitation, and other processes. Thus, information related directly to the condition of the assembly exists in the circulating lubricating fluid.

Advanced warning of abnormal wear in high value critical assets can provide important options otherwise unavailable to decision-makers. Secondary damage may be avoidable by identifying and removing worn parts, thereby reducing subsequent maintenance costs that would result if a catastrophic failure were to occur. In addition, a better understanding of the nature of the problem can be obtained, reducing uncertainty about maintenance procedures. The quantitative monitoring of wear condition is a complex and difficult problem. A large variety of methods have been developed to quantify the presence of pollutant elements in the oil caused by engine wear (Hunt 1996), spectrometry being the most frequently used method. One important aspect of oil analysis that requires bearing in mind is that wear debris quantification does not always correlate with the real wear that one intends to measure. Measurements are affected by different factors that should be compensated or accounted for if a proper wear analysis is to be achieved. Also, for practical operating equipment, there are many factors that affect the wear of parts, such as engine age, type of service, environmental conditions of work, engine metallurgy, and so on. These are difficult to evaluate.

Taking into account the limitations and difficulties described above, several approaches to improve wear condition determination have been presented based on various methodologies, ranging from statistical approaches to correction of particular operating condition effects or normalization methods (Huo et al. 1997).

There are some factors conditioning wear quantification. This problem can be divided into three parts: firstly, the problems related to the limitations of the measurement techniques; secondly, the effects of operating conditions on the wear measurements; and finally, the particular engine characteristics (manufacturer, engine age, environmental conditions, type of service, etc.). These are described in more detail below:

- Today, all techniques used to quantify wear debris in oil analysis present certain limitations. The most widely used analytical technique for wear quantification in oil analysis programs is probably spectrometry. This type of technique involves a particle size limitation. For an inductively coupled plasma (ICP) spectrometer (the most common type of spectrometer technique used in oil analysis laboratories), a typical maximum size is 5 μm for approximately 100% recovery. If a particle is larger than the size stated, then it will not give off a signal at all and the user will be under the impression that there is no debris. Furthermore, a key concept is that machines wearing in an abnormal mode will produce unusually large amounts of particles and a particle distribution with proportionally more large particles. The different regimes of wear, from mild to severe, are characterized by different sized particles, the most severe being associated with particles larger than 1 μm. Thus, the detection of severe wear is restricted by the size limitation associated with spectrometry techniques. Although it is known that there are important limitations when using spectrometer measurements for wear condition monitoring, it is also important to note that at present it is probably the best available and most commonly used option. It is important to mention that, on one hand, researchers are working on different areas to reduce the size effect limitations and, on the other hand, for more critical applications, additional methods such as particle counting and ferrous density analysis are also used to detect abnormal levels of wear debris.
- Oil consumption rate is an important consideration, as used oil should be periodically topped up. An inevitable consequence of oil consumption is that the prevailing

contaminant concentrations will increase even though the real wear rate may have remained constant. On the other hand, if fresh oil is added to maintain the optimum level in the crankcase, dilution reduces the contaminant particle concentration of the system. The equilibrium value obtained will depend on the sump size, oil consumption rate, volatile component, and the real rate of wear. Similar arguments can be applied to the concentration and depletion of metallic oil additives.

- It is possible to find a metallic element that originates from engine wear and is also present as an additive in the fresh oil. Obtaining a compensated concentration that does not reflect this initial value is considered convenient, so that the parameter obtained gives a more accurate idea of the contamination effect and engine wear. In this way, samples from different oils (with different additive levels) can be compared.

Wear limit tables sometimes offered by engine manufacturers are based on extensive research and testing, and these tables reflect mainly average situations and can be used as a guideline only. As has been mentioned, aspects such as type of service, engine age, environment, oil consumption, and operator skill will have significant relevance that will affect wear rate measurements. Because of these factors, each engine must be treated on its own merits. It is far more beneficial to assess the wellbeing of an engine or lubricant on the basis of a trend analysis.

When an abnormal level or rate of production of wear metals is detected, the chemical identity of the abnormally produced particles will provide clues concerning the identity of the parts being worn. Some metallic elements will specifically identify an impending problem, while others provide only general information that abnormal wear is occurring. A good knowledge of engine metallurgy can help detect damaged parts.

Key Applications

Two main pathway applications can be considered for engine condition monitoring based on oil analysis, and these two applications should not be considered as exclusive but as complementary.

On one hand, on-site analysis is performed mainly as a screening test to detect abnormal situations on engine lubrication oil. Accuracy in this type of analysis is obviously reduced, but a quick and easy answer is obtained. For this application, simple and effective devices are used: comparative viscometers, crackle test, constant dielectric measurement, remaining useful life, and so on. The rapid answer and the intimate knowledge that the staff performing these tests have about the equipment and its service condition can contribute to a good diagnosis and rapid action in order to avoid some incipient fault.

On the other hand, an external lab providing oil analysis offers accuracy, much greater oil analysis capability, knowledge of equipment from different customers (different environments, different situations), and so on. A deeper analysis and diagnosis can be obtained from data from an external lab.

It is important to bear in mind that the success and effectiveness of an oil analysis program is dependent upon reliable samples. A reliable sample is one that is truly representative of the circulating lube in the equipment being evaluated. Diagnostic processes will be meaningless if the oil sample fails to effectively represent the real condition of the oil in service.

Definition of effective oil sampling procedures is highly recommended in order to assure consistency of an oil analysis program (Troyer and Fitch 2001). These sampling procedures must take into account general goals that need to be reached: maximizing information obtained, minimizing possible data disturbances and proper sampling frequency. Basic sampling procedure should include at least sampling location, sampling frequency, sampling method, and material requirements.

Cross-References

► Used Oil Analysis

References

A. Caines, R. Haycock, *Automotive Lubricants Reference Book* (SAE, Warrendale, 1996). ISBN 1-56091-525-0

T.M. Hunt, Oil debris monitoring, in *Handbook of Condition Monitoring*, ed. by B.K.N. Rao (Elsevier Science, Oxford, UK, 1996)

Y. Huo, D. Chen, S. Wen, Monitoring of the wear condition and research on the wear process for running equipment. Tribol. Trans. **40**, 87–90 (1997)

V. Macián, B. Tormos, P. Olmeda, L. Montoro, E. Anubla, Results and benefits of an oil analysis programme for railway locomotive diesel engines. Insight **45**, 402–406 (2003)

D. Troyer, J. Fitch, *Oil Analysis Basics* (Noria Corporation, Tulsa, 2001). ISBN 0967596417

Engine Lubricants

Charles A. Passut
Afton Chemical Company, Richmond, VA, USA

Synonyms

Motor oil

Definition

Engine lubricants are oils of lubricating viscosity used to lubricate the internal parts of engines used in the generation of power.

Fundamentals

Engine Lubricants cover a broad spectrum of lubricating oils which are characterized by the types of engines, the fuel used, the combustion process, and the engine service. This discussion is limited to lubricants used in internal combustion engines. The lubricants used in engines consist of mixtures of base lubricating oils, which are distinguished by their viscosities, and performance additives. The base lubricating oils can be petroleum based (mineral oils), synthetically produced, or natural plant oils. The performance additives cover a range of types including, pour point depressants, viscosity modifiers, anti-oxidants, wear inhibitors, dispersants, detergents, basic additives, friction modifiers, extreme pressure agents, emulsifiers, rust inhibitors, corrosion inhibitors, and foam inhibitors.

Reciprocating Engines

The most common internal combustion engines are the reciprocating or piston engines. These come in two basic types: those that operate with the four-stroke cycle and those operating with the two-stroke cycle. The two-stroke cycle engines provide a power stroke in each cylinder with every revolution while the four-stroke cycle has a power stroke every two revolutions. The means of combustion can be ether by spark ignition or by compression ignition (diesel cycle). Spark ignition engines can run at higher speeds than compression ignition engines. The fuels can range from natural gas, LPG, alcohols, diesel oils, and bio-diesel to heavy residuals. These engines can be water cooled or air cooled. Service can include aircraft, automotive, stationary, railroad, and marine. All these variations will have different lubricant requirements and special lubricant formulations may be blended for each.

Engine lubricants for reciprocating engines will perform a variety of functions. The lubricant is used to provide hydrodynamic lubrication to the engine bearings. These bearings are generally plain bearings that require specific oil viscosities to maintain an oil film between the bearing surfaces. This oil film will be determined by the oil viscosity, which will be a function of temperature, pressure, and shear rate in the bearing. A viscosity classification system is provided by the Society of Automotive Engineers (SAE) in the engineering standard J300. These SAE viscosity grade are usually specified by the equipment manufacturer for specific ambient temperature operating conditions. Petroleum-based fluids provide high viscosities at elevated temperatures and pressures. The SAE viscosity classification system describes two basic types of lubricants, monograde or single grade oils and multigrade oils. The polymer type and content of the lubricant will determine the oils' viscosity shear rate relationship. The oil may be used as a hydraulic fluid to operate valve lash adjusters or as power to unit injectors in diesel engines. The viscous oil provides some gas sealing around the piston rings in the engine. Foam inhibitors are used to treat the engine lubricants to reduce foam in the crankcase, which can lead to entrained air entering the oil pump inlet and producing a compressible air in oil emulsion.

Those same viscous oil properties that promote oil film thickness and prevent metal contact also provide viscous drag on the engine. This is a parasitic loss of energy that can be reduced by using the lightest (lowest viscosity) oil that will still maintain engine service life. High fuel costs will increase the emphasis on reducing engine friction both in the hydrodynamic and the boundary friction areas of the engine. In the boundary lubrication regime the frictional properties are determined more by the surface chemistry of the lubricant, not by the physical properties of the oil. Boundary layer friction can be reduced with the addition of surface active friction modifier additives that form low-friction films on the engine parts.

Engine lubricants provide wear protection in metal-to-metal contact or in the boundary lubrication regime. Wear inhibitors, extreme pressure agents, and friction modifiers provide oil films that can be thick enough to prevent metal-to-metal contact and prevent surface welding and seizure. Some anti- wear additives provide sacrificial films that prevent metal adhesion.

Engines lubricants are also used to provide cooling for internal engine parts. The lubricant is often sprayed on the under-crowns of the engine pistons to reduce the piston temperatures. Oil may be circulated through oil coolers to help control engine temperatures.

A characteristic of internal combustion reciprocating engines is the contamination of the engine lubricant with combustion gases that pass by the piston rings and through the piston ring gaps (blow-by). These gases contain incomplete combustion by products. These contaminants are acidic and promote oxidation. Diesel engines and direct injection gasoline engines may produce soot in the combustion process, which will contaminate the engine lubricant. These solid and insoluble materials are dispersed in the oil by using ashless dispersants. Insoluble varnish type deposits that can collect on high-temperature surfaces such as pistons, piston ring grooves, and piston rings are prevented with basic detergent additives. At low

operating temperatures the blow-by may contaminate the oil with water and acidic oxidation products. The engine lubricants contain rust inhibitors and basic compounds to neutralize the acids and prevent rust.

Depending on the amount of fuel being consumed and the specific power output of the engine the bulk oil temperature can approach 135C. These temperatures will promote oil oxidation and the corrosion of non-ferrous parts. Engine lubricants are fortified with a variety of oxidation and corrosion inhibitors including basic materials.

Two-stroke cycle engines can have special lubricant requirements. The four-stroke cycle engines typically maintain an oil system separate from the fuel. This may be in the engine crankcase or in a separate dry sump system. Oil is delivered to the engine by an oil pump or in simple systems by just splash lubrication from the crankcase. Two-stroke cycle diesel engines may have similar oil pumping and delivery systems. Some two cycle gasoline engines with the Reid valve system use the crankcase to compress the intake charge and mix the oil with the fuel. These engine lubricants typically contain light napthenic or aromatic solvents to promote easy mixing with the fuel. They also contain high molecular weight petroleum fractions or synthetic oils that will provide anti-scuffing performance for the pistons and cylinders and lubrication for the bearings. The bearings in these engines are typically roller or tapered (anti-friction) bearings that operate in the elasto-hydrodynamic lubrication regime. Some two-stroke cycle spark ignition engines may deliver the fuel without oil or by direct fuel injection into the cylinders. The lubricant in this case is delivered to the critical engine parts by an oil pump but the oil is still burned in the combustion process.

Rotary Engines

Several rotary engines have been developed from the original Wankel design. These engines have been used in a variety of mobile applications. The rotary engine is actually a variation of the reciprocating engine that uses an eccentric drive and rotor to replace the piston and connecting rod. This crankcase application has requirements similar to the four-stroke cycle reciprocating engine. The engine lubricant is circulated through the rotors for cooling and provides lubrication to the bearings. The lubricant is not exposed to as high an exhaust blow-by rate as in the reciprocating engine. Some lubricant may be injected into the trochoid housing to provide lubrication to the rotor apex seals. This injection is similar to the injection in some two-stroke cycle engines. Crankcase temperatures can be high and anti-oxidants are an important part of the lubricant formulation.

Gas Turbine Engines

The engine lubricant for the gas turbine engine has similar requirements to the reciprocating engine but it is not exposed to the blow-by contaminants. The combustion process is continuous and occurs in stages in the engine. The primary lubrication requirements are for appropriate viscosity oils to be used in the ball or roller bearings. These oils need to be of relatively low viscosity to minimize viscous energy losses and to prevent sliding of the roller bearings during acceleration. These lubricants may be exposed to high temperatures and antioxidants are important lubricant components.

Key Applications

In considering the choice of engine lubricants the engine design, fuel type and the type of service must all be taken into consideration.

Passenger Car Engines

One of the largest segments of the engine lubricant industry is the passenger car engine oil. Often referred to as "motor oil" these lubricants have sophisticated requirements which require complex additive formulations. Extensive testing is required to meet the certification processes for engine and vehicle manufacturers (OEMs) and international standards organizations. Two such organizations are the American Petroleum Institute (API) and the Technical Association of the European Lubricants Industry (ATIEL). This testing is to simulate the variety of situations to which these engines will be subjected. Some will be started at very low temperatures and will require low viscosity oils at low temperatures. This requirement will be covered by the SAE viscosity grade described in the Society of Automotive Engineers standard J300. The viscosity of the oil will be noted at both the low temperature or "W" grade and the viscosity at high temperature, such as 5W-30. This oil would then provide starting viscosity at low temperature like a 5W and at high temperatures would run as an SAE 30 grade oil. The ability to meet different viscosity requirements at widely different temperatures requires lubricating oil with a high viscosity index. Viscosity index is an arbitrary scale which indicates the relative rate of change of viscosity with temperature. A high viscosity index lubricant has a lower rate of viscosity decrease with increasing temperature. Engine lubricant viscosity index is often increased with the addition of polymer thickeners called viscosity modifiers of viscosity index improvers.

Engine operation at low temperatures and short trip driving can produce a significant amount of condensed water and fuel in the engine crankcase. These

contaminants will promote rust and corrosion in the engine. Engine Lubricants usually contain materials with base number to neutralize the acids produced in these situations. Surface active materials are also added to provide a protective coating on metal parts. These rust and corrosion problems can be aggravated when fuels containing alcohols are used. In diesel powered vehicles the use of bio-fuels can result in significant accumulation of high boiling point fatty acid methyl esters (FAME) in the crankcase. These materials will affect the oxidation stability of the lubricant and require additional antioxidants in the lubricant.

Engine operation at low speeds and long idle periods can result in significant fuel dilution of the lubricant in gasoline and diesel powered vehicles. In addition diesel vehicles may encounter high soot levels in the lubricant. These contaminants can produce oxidation products that result in high levels of insoluble materials. Lubricating oils require dispersants and detergents to suspend these materials and prevent sludge deposits in the engine or the formation of hard varnish deposits on internal engine surfaces.

Passenger vehicles operating at high loads and in high ambient temperatures may experience increased oil temperatures. Lubricating oils must be equipped with a significant level of antioxidants to prevent oxidation of the oil, which leads to oil thickening.

Emission control systems in passenger vehicles are sophisticated and have required many changes in lubricant formulations. Since some of the lubricating oil will be burned in the combustion process, the oil and oil additives will be consumed, and the oxidation products will reach the emission control systems catalyst and traps. Additives that can poison these systems such as sulfur and phosphorus are limited in the oil formulation. In the case of diesel particulate traps the total amount of ash containing material in the lubricant may be restricted.

Commercial Vehicles

Commercial vehicles include gasoline- and diesel-powered highway vehicles and also those used in non-highway applications such as farming and construction applications. These vehicles typically operate at higher load factors than passenger cars and can produce increased thermal stresses on the engine lubricant. At high cylinder pressures blow-by gas rates and contaminants are increased. In diesel engines frequent accelerations may result in high soot contamination rates. The soot particles in diesel engines must be highly dispersed by the engine lubricant to limit engine wear, filter plugging and increased oil viscosity. Commercial vehicles powered by natural gas and LPG can experience high operating temperatures and high levels of nitrogen oxides in the blow-by similar to diesel engines. Most commercial engine applications have specific OEM requirements and may require extensive proof of performance testing. The API and ATIEL organizations also provide performance classifications for commercial diesel engine oil applications.

Railroad diesel engines represent a special case for commercial lubricants. These engines run at lower speeds than the smaller highway diesels and are defined as medium speed diesels. Lubricants for these engines have specific formulation requirements which prohibit certain additive metals and require sufficient oil base number to operate with available diesel fuels. The oil formulations are blended to meet special bearing metallurgy requirements. These oils require extensive field testing to obtain OEM approvals.

Emission control systems for diesel engines often utilize cooled EGR to reduce nitrogen oxide emissions. Cooled EGR in diesel engines introduces increased contaminants into the lubricant, such as soot, water, and combustion acids. The particulate traps on diesel engines require low ash oils to reduce fouling and nitrogen oxide traps require low sulfur oils to prevent catalyst poisoning. The OEMs specify chemical concentration limits for lubricant ash, phosphorus, and sulfur for emission controlled engines.

Stationary Engines

Engines involved in electrical power generation, oil and gas field pumping and agricultural irrigation are examples of stationary power plants. This service is often at constant engine speeds and loads. These conditions can permit a buildup of combustion chamber and piston and piston ring groove deposits which will not be dislodged by frequent accelerations. These engines often run for long periods of unattended operation with extended drain intervals. Oils formulated for this service require extended oxidation stability and high detergency to prevent piston and piston ring deposits. Specific OEM approvals are required for these heavy duty operations.

Marine Applications

Marine service varies from recreational outboard engines to heavy-duty diesel engines for large ships and inland marine operations. Oils for recreational marine use may be formulated for two- or four-stroke cycle gasoline engines. The oils are often certified by the National Marine Manufacturers Association (NMMA). Testing includes requirements for viscosity, corrosion, filter plugging, foaming, aeration, and engine performance tests.

The two-stroke cycle oils are also evaluated for ring sticking and combustion chamber deposits that can cause pre-ignition and detonation. These engines are often run at high loads but are maintained at low temperatures because of low-temperature ambient cooling water.

Diesel marine applications require oils with high detergency to prevent corrosion in the cylinders caused by the combustion of diesel fuel sulfur. Sulfur levels will vary widely depending on the fuel boiling range and the refining source.

Air-Cooled Engines

Air-cooled engines are available in most of the applications cited previously except for marine applications. Several types of engine applications are predominately air-cooled such as motor cycles and reciprocating engine aircraft.

Lubricants for air- cooled engines are often the same as those for water cooled engines in the same service. This would include passenger cars, commercial vehicles, and stationary power plants. These engines often run at higher temperatures and require additional oil antioxidants and more stable viscosities at high temperatures.

Air-cooled aircraft piston engines are heat sensitive and primarily require engine lubricants with the correct viscosity. Most of these aircraft operate on SAE 50 grade oils but multigrade oils are becoming more popular. These lubricants are formulated with antioxidants, dispersants, and detergents. Combustion chamber deposits can be critical since these engine applications are also subject to pre-ignition or detonation.

Motorcycles and snow mobiles are usually powered by small gasoline powered air-cooled engines. These high rpm engines can be two- or four-stroke cycle with specific OEM requirements for the lubricants. Four-stroke cycle engines are sensitive to viscosity and require specific high-temperature viscosity to prevent wear and scuffing. Some motorcycles require heavy SAE 50 oils. The two-stroke cycle oils may be mixed with the fuel or injected in the critical engine areas to be burned during combustion. These two-stroke cycle oils have detergents and dispersants to limit combustion chamber and power cylinder deposits. Detonation can be a problem because of high operating temperatures and ash levels in the oil.

Many small engines used to power lawn and garden equipment are air-cooled and have modest detergent and dispersant requirements. These engines have very often been two-stroke cycle engines but tighter emission controls have forced some to become four-stroke cycle engines. These engines are viscosity sensitive and the OEMs specify specific viscosity grades at a variety of ambient temperatures.

Cross-References

- ▶ Aircraft Engine Lubricants
- ▶ Detergents
- ▶ Engine Oil Test Equipment
- ▶ Friction Modifiers
- ▶ Fuel Economy: Lubricant Factors
- ▶ Hydrodynamic Lubrication
- ▶ Locomotive Engine Oils
- ▶ Lubricant Viscosity
- ▶ Marine Lubricants
- ▶ Used Oil Analysis
- ▶ Viscosity Index Additives

References

E.R. Braithwaite, *Lubrication and Lubricants* (Elsevier, New York, 1967)

W.J. Bartz, *Engine Oils and Automotive Lubrication* (Marcel Dekker, New York, 1992)

A. Schilling, *Automobile Engine Lubrication* (Scientific Publications (GB), Broseley, 1972)

Society of Automotive Engineers Inc, *SAE Handbook* (Society of Automotive Engineers, Warrendale, 2009). Standard J300

Engine Oil Development Tools

- ▶ Engine Oil Test Equipment

Engine Oil Test Equipment

David L. Glaenzer

Afton Chemical Corporation, Mechanicsville, VA, USA

Synonyms

Engine oil development tools; Motor oil test equipment; Research engine test facility

Definition

It is a testing facility where engine oils are tested and evaluated to meet the performance requirements of original equipment manufacturers (OEMs). It is also an R&D facility to assist in the development of suitable engine oils for applications automotive lubrication.

Scientific Fundamentals

Engine oil research and development activities require sophisticated engine test equipment. Automotive manufacturers are usually interested in developing engines that

exhibit high fuel efficiency, are compliant with current and future emission requirements, and have good drivability and durability. Typically, tests are completed to measure engine performance, fuel efficiency, durability, and exhaust emissions. Engine oil research and development activities need to determine the oil's effect on the engine's ability to stay compliant with its performance parameters. A research engine test facility allows for development of engines and engine oils through the measurement, recording, and control of many operating parameters.

Information gathered from sensors during engine operation is processed and logged by data acquisition systems. Actuators controlling engine speed are used in combination with dynamometers to arrive at the desired torque and power output level. Commanded by the test cell computer, the dynamometer creates a load on the engine to simulate actual vehicle operation or some other desired test conditions.

Engine test cell is a term used to describe a room in which a fired test engine is operated. The test cell can have many forms and sizes depending on the application. It may be a permanent part of a structure or may be a stand-alone installation. Whatever shape or form an engine test cell takes, there are certain aspects that must be considered. An overview of those aspects follows.

General Purpose Test Cell

General purpose engine test cells generally have the capability of handling 300 kW engine output power. These test cells are the most prevalent and are used for basic engine research, checking emissions, endurance testing, tuning engines for race applications, and engine oil testing. The engine and dynamometer are usually bolted to a rigid mounting plate that may be isolated from the rest of the building with a seismic mounting block and elastomeric pads to minimize vibration in the test facility.

The test cell is usually adaptable to take a wide variety engines within a given speed and power range. Adjustable mounting plates allow changes to be made with relative ease, although most stands are used to run the same engine for many months.

Engines are changed by lifting the engine into place and using adaptors to fasten to the engine's designed mounting locations. Once mounted on the test stand, connections are made to the various supply lines for fuel, coolant temperature control, and oil temperature control. A transducer cabinet and temperature thermocouple cabinet is located near the engine and provides connection of various pressure and temperature sensors to the data acquisition system. The cabinets are sometimes mounted on a moveable boom to provide accesses to the engine for ease of installation and removal.

Exhaust from the engine is generally directed upward to remove as much heat from the test cell as soon as possible. Various types of insulation are used depending on the application and temperatures. Pressure and sample taps are located as required. In most applications a butterfly valve is used to control exhaust back pressure.

Coolant and temperature control vary with the type of application, but generally include pumps, heat exchangers, and control valves for flow and temperature and may include both heating and cooling elements. Proper sizing of heat exchangers and valves will result in more accurate control at the desired test conditions.

Combustion air is supplied to the test cell from an outside source. This combustion air may be humidity controlled and is usually temperature controlled. Combustion air pressure is sometimes controlled at the inlet to the engine using a butterfly or other type of valve to control inlet restriction.

Speed and dynamometer torque are controlled by remote throttle and dynamometer excitation and are measured by various means. Typically, a strain gauge load cell is attached to the dynamometer moment arm.

Fuel flow may be measured. Mass flow meters are typically used. For consistency, the fuel temperature may be controlled prior to the flow meter and/or the engine inlet.

Dynamometer

A dynamometer or "dyno" for short is a machine used to measure torque and rotational speed from which power produced by an engine, motor, or other rotating primary mover can be calculated. A dynamometer that is designed to be driven is called an absorption dynamometer. A dynamometer that is designed drive or absorb is called an active or motoring dynamometer.

An absorption dynamometer usually includes a means of measuring torque or rotational speed. An absorption unit consists of some type of rotor inside a housing. The rotor is coupled to the equipment under test or an engine with a driveshaft, all of which must be capable of rotating safely at the desired test speed. A dynamometer is usually equipped with some means of measuring torque. This is accomplished by measuring the braking torque between the rotor and the rotor housing. By mounting the rotor housing such that it is free to rotate, except that it is restrained by a torque arm, it is possible to measure torque. Torque arms are usually connected to scales or load cells that are used to provide an electrical signal that is proportional to torque. Another way to measure

torque is to fix the dynamometer housing and use a torque-sensitive couple between the power source and the dynamometer. The torque measurement system is usually calibrated with weights that are traceable to some standard.

Control Console

Generally located outside of the test cell is the control console. The control console houses the data acquisition system as well as various displays to monitor-controlled processes. A work area as well as a chair is usually provided. The control console is located near a window so the operator can observe the inside of the test cell.

HVAC

An engine running in a test cell can generate a tremendous amount of heat. Air handling should be sized to provide the proper environment for operation of the test stand as well as employee safety. Multi-level controls may allow a test cell to have different temperature and air flow characteristics when the engine is operating than those conditions when the engine is at rest, allowing for more comfortable working conditions for the test engineers and technicians. Ambient air is usually filtered to prevent contamination from foreign material.

Doors and Windows

Doors for moving engines in and out of the test cell should be of sufficient size to safely maneuver an engine on a build-up stand into the test cell. There should be at least two doors for personnel egress if a situation warrants they leave the test cell immediately. All doors should contain some sort of sound insulation and must open outward for egress. Doors should be adequately marked with Exit signs. A window is usually provided near the control console and offers a view into the test cell. Windows are usually double pane insulated glass.

Walls

Wall material should be adequate to insulate sound from the rest of the building as well as provide a surface for the mounting of various pieces of support hardware needed to operate and monitor the test engine. The walls or overhead building steel will usually support an overhead crane system, so they must be of adequate strength. On occasion, a wall will be built as a blow-out wall. In the event of an explosion inside the test cell, this wall will provide a means for the dissipation of the energy from the explosion and help prevent damage to the rest of the building. Walls should also provide sufficient strength to contain all rotating and reciprocating parts in the case of failure at high speeds and loads.

Lighting

Adequate lighting is a must. On occasion, a test cell may become clouded with fumes from the engine. Sufficient lighting must be available to see what is going on in the test cell from the control console window. Care should be taken when locating the lighting to avoid shadows created by fire sprinklers, exhaust piping, or other services. Guards should be used to cover light bulbs if there is a condition that may lead to bulb breakage.

Engine Handling

Engines are usually rolled to the test stand on an engine cart and then moved with an overhead chain hoist system to the mounting location. Many test stands have a single beam type chain hoist located in the center of the test cell that is used for engine as well as dynamometer replacement. A bridge crane system has the utility of moving heavy material to most locations in the test cell. All components of the hoisting system must be sized properly for the loads being transferred. Periodic inspection of hoisting systems is warranted.

Guards

Guards are used to prevent personnel from coming into accidental contact with rotating parts or hot surfaces. Guards are also used to contain a driveshaft in the event of a catastrophic failure, preventing the driveshaft from damaging the test stand equipment or personnel. Care should be taken when designing guards to assure they will be sufficient for their intended use.

E Stops

Emergency stop buttons should be located both inside and outside the test cell. Such stop buttons turn off all power except for HVAC and lighting and are in prominent locations and adequately marked. Typically, emergency stop buttons are located near exit doors.

Fire Control

Several different methods are used for fire control. Handheld fire extinguishers are usually available inside and outside the test cell. Fire extinguishers are usually dry powder or CO_2. On occasion, a test cell may be connected to a central CO_2 distribution system that will engulf the test cell in CO_2 when activated. Fire sprinkler systems are almost always present in test cells. On occasion, a foam-type system is used.

Data Logging

Much has changed over the years in the area of recording, processing, and storage of data. The old method of hand logging data on paper may be outdated; however, it may still have value in certain circumstances. Traditionally, observations were made of measurement devices and logged on paper, and conclusions were based on the logged data. Chart recorders were employed to monitor both stability and trends. Today, there are many different systems to log data by computer. Such data logging has the ability to record significant observations in a short period. Data is tabulated and linearized within the computer and displayed in tabulated form. Data is stored for future retrieval and post-processing analysis.

Health and Safety

Because a test cell houses an engine and dynamometer with high-speed reciprocating and rotating parts, it is important to consider the safety of personnel as well as protecting the site location. Attention must be given to sound control using measures such as insulation, double glazed windows, and sealed doors to provide suitable working conditions outside the test cell. Guarding and containment of rotating shafts is important. There may be many hot surfaces such as exhaust pipes, coolant and oil lines that require insulation for protection. Consideration should be given to providing suitable alarms for carbon monoxide and combustible gases that may be emitted into the test cell. Walkways must have adequate clearance for people as well as the movement of materials. A test cell should have sufficient means of egress should the situation warrant evacuation. Most importantly, fire protection and detection should be considered. Flame, heat, and smoke detectors can be used to shut off fuel and power to the engine in addition to triggering fire extinguishing systems. In many instances there are redundant systems for fire protection such as sprinklers and CO_2.

Key Applications

Engine Protection

Modern engine oil must perform a number of functions to keep the engine running properly. Engine sludge is formed during low temperature operation when oil does not have the ability to prevent contaminants, such as fuel and water from the combustion process, from forming. Sludge can lead to restrictions in oil flow through orifices and can impede oil pickup through the oil screen by the engine oil pump causing lubricant starvation and wear. Varnish and carbon deposits on piston rings can prevent the rings from properly sealing the combustion gases in the engine cylinder. This can lead to high blowby gas flow and additional contamination of the engine oil. Oil consumption can increase, which can lead to inefficient catalytic converter operation and increased exhaust gas emissions. Oil may oxidize and become viscous leading to premature engine failure. Modern engine oil is also an integral part of fuel economy, containing additives to increase fuel economy or being formulated in such a way to gain fuel economy through reduced viscosity.

The Engine as a Tool

Testing of engine oils requires precise, repeatable engine oil tests. When engine oil is evaluated on a test stand, the intent is to provide a result in days or hours that would take weeks, months, or years to see in real-world applications. Factors that contribute to the degradation of engine oil are magnified; temperatures may run higher or lower than normal, oil charge levels may be lower, operating speeds and torque output levels may be higher. All of these factors must be precisely controlled to yield a repeatable test. The test engine itself is usually built more precisely than an over-the-road engine. Care is taken to measure and control tolerances to a fine level. Cleanliness is of utmost importance. Following operation, the engine is disassembled and then various parts are measured and rated for wear and/or deposits using standardized procedures.

Engine Oil Evolution

Engine oil has changed over time to satisfy the needs of the modern engine. The American Petroleum Institute (API) sets minimum performance standards for lubricants. Motor oil may be composed of a lubricant base stock only and is referred to as non-detergent motor oil. Fully formulated motor oil contains additives to improve the oil's detergency, extreme pressure performance, and ability to inhibit corrosion of engine parts. The API has two general service classifications, the "S" class for spark ignition engines and the "C" class for compression (diesel) ignition engines. The International Lubricant Standardization and Approval Committee (ILSAC) also has standards for motor oil. Their latest standard for spark ignition engines, GF-4 was approved in 2004. The ACEA (*Association des Constructeurs Europeens d'Automobiles)* performance/quality standards are used in Europe. Their classifications are arguably more stringent than the API and ILSAC standards. The Japanese Automotive Standards Organization (JASO) has come up with their own set of performance and quality standards for engines of Japanese origin. Several automobile manufactures have also developed standards that they require for specific applications.

ILSAC is in the midst of specifying their latest performance category, GF-5. This category will satisfy all previous categories and add a few enhancements. Fuel economy of the engine oil will be evaluated with an updated test, the Sequence VID. Deposit tests such as the Sequence VG will have stricter pass/fail limits. The inclusion of catalytic converter phosphorus poisoning protection will be addressed.

Engine oil will continue to evolve as we see a tightening of fuel economy standards, a change of fuels available for automotive use, and an increase in use of hybrid vehicles. The engine test cell will continue to be a tool for use for the foreseeable future.

Cross-References

► Additive Chemistry Testing Methods
► Engine Lubricants

References

M. Plint, A. Martyr, *Engine Testing Theory and Practice* (Butterworth-Heinemann, Oxford, UK, 1995)

Engine Oils and Nanoparticles

► Nanoparticles in Automotive Applications

Engine Tests

► Additive Chemistry Testing Methods

Engineering Surfaces

► Topography of Engineering Surfaces

Entraining Velocity

► Gear Sliding

Environmental Effects on Vapor Phase Lubrication

► Vapor Phase Lubrication for Micro-Machines

Environmentally Assisted Cracking Under Repeated Loading

► Corrosion Fatigue of Metallic Alloys

Environmentally Assisted Fatigue

► Corrosion Fatigue of Metallic Alloys

Environmentally Friendly Lubrication Issues

Richard E. Kuhlman
Afton Chemical Corporation, Southfield, MI, USA

Synonyms

Benefits of Eco-friendly Lubricants; Lubricant Formulation for Reducing Greenhouse Gas; Use of Biodegradable Lubricants

Definition

Environmentally friendly lubricants are defined as those that are readily biodegradable in nature. This paper will discuss the issues surrounding the many types of these fluids and greases that exist in world commerce. It will also note the various criteria that are used to measure biodegradability. The degree of biodegradability is largely dependent upon the base oils used, but can also be affected by performance additives. Base oils include, but are not limited to, naturally occurring esters such as soybean oil, sunflower oil, rapeseed oil, and castor oil, along with synthetic esters such as polyol esters and di-isotridecyl adipate. Naturally occurring esters can also be enhanced through the introduction of genetic changes in the plants, and/or through structural changes in their chemistry following their harvest and extraction.

Scientific Fundamentals

The earliest forms of lubrication were based on biodegradable substances. These were largely comprised of animal fats. Accordingly, one can imagine that their presence in commerce had no long-term effect on the environment.

Biodegradable lubricants were used almost exclusively until the latter part of the eighteenth century when petroleum products began to appear. Even so, animal fats have continued to be used, although usually in a chemically modified state, right up to the modern era. Some of the more common in the early years, such as sperm whale oil, are no longer used for conservation reasons, and have been replaced by synthetic versions of their structure. In fact, it was the over-harvesting of whales in the mid-nineteenth century, coinciding with Drake's discovery of oil in Pennsylvania, which triggered a long-term downtrend in the use of animal fats. Interestingly, in general, none of the earlier biodegradable lubricant components were chosen for use with that attribute in mind. They were used because they were cost effective and available. Today it is a different story. Researchers have brought a great many chemicals to the market, pairing them with readily biodegradable base fluids, in an effort to change the way equipment is lubricated, especially in environmentally sensitive areas. As one would expect, governments in some areas of the world are more aggressive than others in embracing this new technology. The use of these products is expected to become more widespread as the twenty-first century progresses and regulations that encourage their use are strengthened. This effort will be accompanied by an ever-increasing list of building block materials that the lubricant formulators will have at their disposal. This chapter offers the reader a review of the existing environmentally friendly products in use today, and compares their relative benefits and trade-offs with conventional petroleum-based lubricants.

The discovery of petroleum created access to inexpensive sources of fuels and lubricants, thus ending dependence on naturally occurring products. Only in times of economic stress, such as the World Wars, did commerce return to the use of these earlier technologies. The oil embargos of the early 1970s also heightened the awareness of alternative sourcing. Each time, however, the supply/demand curve returned to a level that discouraged further development of products, limiting their use to total loss applications, that is to say, applications in which virtually none of the lubricants are recovered from the environment.

Early work was carried out on oils that were extracted from naturally occurring agricultural products. These oils had many positive attributes, but also had performance and appearance issues that limited their utility. Progress has been made on at least two fronts. Firstly, researchers have been successful in modifying the natural products to eliminate, or at least mitigate, shortcomings in performance. Such efforts include, but are not limited to, improvements in low temperature operability, as well as increased hydrolytic and oxidative stability. The second method used to create improvements has been to introduce hybridized seeds that result in agricultural products that are more adaptable for use as lubricants. Of course, chemical modification of the hybrid crops can also be utilized. A good example of this approach can be seen in a 2008 publication from the United Soybean Board (Soy Lubricants). It highlights four general opportunities to increase the utility of soy:

- Biotechnology to produce more stable oil from seed.
- Non-transgenic modification to produce more stable oil.
- Modification of the oil through chemical or mechanical processing to increase stability while maintaining good oil properties.
- Chemical additives that improve stability, offering the most rapid and cost-effective route to commercialization.

Other products, such as castor oil, have also been explored in a like manner. One such project involved the modification of natural castor oil by an isomerization reaction that extended the carbon chain, thus lowering the pour point and improving the viscosity index (Tao et al. 2004).

Vegetable oils have certain issues in terms of their performance. On the plus side, they can have superior lubricity. They also carry a high viscosity index (VI). Soy can range in VI up to 223 versus 90–100 for API Group I petroleum products. Flash point is another attribute. Soybean oil has a flash of 610°F compared with 390°F for a typical, common petroleum base. Toxicity is also generally lower for naturally occurring products. Lastly, they have an advantage in evaporation (NOACK) of up to 20% less than mineral oils (Schneider et al. 2006).

On the negative side, vegetable oils in their natural form exhibit a lower level of oxidative stability than petroleum-based products. However, chemical modification can be used to enhance this property. One method is to increase the level of oleic acid in the product. Canola oil, rapeseed oil, sunflower oil, and soybean oil have all been modified to contain high oleic acid content. Another negative is the relatively high pour point of vegetable oils. This property can be modified through both genetic engineering of the seeds themselves, and/or through the

addition of chemical pour point depressant additives (Honary et al. 2001).

Lastly, there are certain cost penalties involved in the use of environmentally friendly fluids. The base fluid itself can be, at a minimum, several times as expensive as mineral oil. One must add to that the prospects of more frequent service intervals and the potential need to derate equipment operating parameters in order to accommodate their use.

Environmental Considerations

Firstly, one must recognize that most synthetic and mineral oils are inherently biodegradable, meaning that they do not persist in the environment over the long term. Some applications, however, require a greater degree of biodegradability. Fluids for these applications are referred to as readily biodegradable. The end user can be guided in the choice of environmentally acceptable products by referring to results of testing for this attribute. One common method is to measure the amount of oil that is converted to CO_2 when added to an aqueous solution containing common soil and sewage bacteria (Pearson and Spagnoli 2000). Hydraulic fluids and greases are considered as readily biodegradable if more than 60% of the test material carbon is converted to CO_2 within 28 days. The test method is known as the OECD 301B. Vegetable oils and some esters meet these criteria, while mineral oils do not. Polyglycol-based materials also fail, with only 6–38% conversion to CO_2 in the 28-day period (Pearson and Spagnoli 2000; Schneider et al. 2006).

In addition to the speed of degradation, one must take into account the issue of toxicity. The relative toxicity of a fluid is determined by the chemistry of the additive components as well as the base stocks. The level of toxicity can be determined by conducting acute toxicity studies with rainbow trout, using OECD 203. Toxicity is expressed as the concentration of material in ppm that results in a 50% fish kill after 96 h (LC_{50}). The generally accepted threshold for toxicity of lubricants is $>$1,000 ppm (Levi Pearson and Spagnoli. 2000; Willing et al. 2001).

There exists a series of state-sponsored symbols, designed to assist the end user in identifying lubricants that meet the readily biodegradable criteria (Schneider et al. 2006). The earliest of these symbols is the Blue Angel (Der Blauer Engel), which was established to cover a variety of environmentally friendly products in Germany (Bartz et al. 1988). The aim, according to the Blue Angel criteria, is to promote the substitution of conventional mineral oil-based lubricants by eco-friendly lubricating oils and greases. In 1990, a line of eco-label, readily biodegradable lubricants and forming oils was launched. Following this action, and throughout the 1990s, about 5% of the hydraulic fluids in the German marketplace qualified for Blue Angel status. By 2005, this number had increased to 9% for stationary applications (primarily hydroelectric power plants, locks, dams, etc.) and 19% for mobile applications (primarily agricultural equipment, construction machinery, and forestry equipment).

The general order of biodegradability for common lubricants is (in decreasing order): vegetable oils, synthetic esters, mineral oils and alkybenzenes, PAGs, and finally polyalpha olefins (PAO) (Padavich and Honary 1995). It should be noted that 2cSt PAOs do meet the minimum standard for biodegradability. They are based on essentially a dimer of decene without much chain branching. No effort in this chapter will be made to quantify the usage of any of these types, as their consumption is variable and based on many factors, including economic considerations.

Generally speaking, however, central Europe has been more aggressive in the encouragement of the use of environmentally friendly lubricants than other world areas, including the United States. Their action has had a direct effect on the type of base oils that are used and the choice of additive components. Owing to the relative availability of agricultural products, rapeseed-based lubricants are popular in Europe, along with synthetic esters, while soybean-derived fluids are more popular in North America. In addition to rapeseed and soybean by-products, safflower, sunflower, crambe, and corn oil have been used. As with any agriculturally derived product, there can be important differences in the quality and performance of these naturally derived materials from year to year, largely dependent on climatic conditions. This is an important issue when planning the introduction of eco-friendly technologies.

Along with the positives, in terms of their biodegradability in our environment, there are other issues that the fluid formulator must contend with when blending environmentally friendly lubricants. As previously stated, vegetable oils have relatively poor oxidation control and poor hydrolytic stability when compared with highly refined petroleum-based oils, although their thermal stability and corrosion protection properties are comparable. The oxidative stability of synthetic esters, on the other hand, varies according to their type. The monounsaturated fatty esters have better oxidation stability than vegetable oil but poorer than mineral oils, whereas thermal stability is somewhat better than that of mineral oils. Saturated monoesters, polyol esters, and dibasic acid esters have better oxidation and thermal stability than mineral oils, but the hydrolytic stability of all ester-based

lubricants is lower than that of mineral oils. Finally, the lubricity characteristics of vegetable oil, fatty esters, and synthetic esters are much superior to those of mineral oils (Wagner et al. 2001).

Ecologically friendly lubricants can play a particularly successful role in cases in which the lubricant is in a total loss application. Examples of this include bar and chain oil for forestry, rail grease for railroad track usage, two-stroke oils, and mold release fluids. Essentially 100% of the bar and chain oil used in logging is introduced to the environment. None of it is recaptured for later use. Bar and chain oils are needed to lubricate the cutting chain, while protecting the bar on which it rides from wear. Likewise, rail greases are dispensed at the point of use and are totally expended. Specialized lubricants are available for both of these applications. However, it should be noted that industrial fluids, which represent a large portion of the total fluids used each year, are typically captured, removed from the site, reworked, and brought back for future use. Techniques can range from simple filtering and dehydration to a rigorous re-refining of the used oil. Since this type of treatment lessens the impact of disposal of used fluids, and makes additional use of post service virgin product, the practice can rightfully be referred to as being environmentally friendly. It should also be noted that not all used fluids are reworked in such a manner. Some are burned for fuel value.

Let us take a closer look at rail lubrication. This is not a new tribological application, but rather has been in place for quite some time. Rails, especially on curves, have been lubricated in order to reduce friction, thus increasing efficiency and reducing rail and wheel wear. The lubricant can be applied at rail side or by on-board dispensers on the locomotive. As noted above, all of the lubricant that is dispensed in this operation stays in the environment. In the 1990s, rail lubrication began to use biodegradable products. These could be in the form of a fluid or light grease. In addition to the obvious advantages to the environment, these products are less apt to create a build-up at the point of usage over time. This is not simply a cosmetic advantage, but rather can serve to reduce plugging of dispensing mechanisms (Sudhir Kumar 1999; Anand and Chhibber 2006).

Two-stroke oils and concrete mold release lubricants are also good examples of total-loss applications. Two stroke oils are meant to be consumed. However, the ones that are used in marine applications have an opportunity to leak into the water. This leakage can be rendered less critical, although still important, if the oils in use are readily biodegradable. Concrete mold release lubricants pose a similar but different problem. They stay behind on the concrete once the molds are removed, and can subsequently be washed off by the elements. Once removed from the concrete in this manner, the lubricants can pose a threat to the immediate environment, as well as causing contamination for neighboring bodies of water (Anand and Chhibber 2006).

There are also environmental advantages for fluids that are not designed to be used in a total loss situation. Hydraulic fluids used in forestry or snow grooming are two good examples. It is common for transfer hoses to rupture in the field. Hose failure under high pressure can lead to a significant amount of fluid being injected into the environment. Thus, in sensitive areas, especially near watersheds, the use of biodegradable products can lessen the long-term effect on the surroundings.

Tables 1 and 2 contain the typical physical properties of representative environmentally friendly lubricants.

The reason that naturally occurring vegetable oils can be used successfully in total loss applications, such as mold release oils and for chain saw bar lubrication, is that there are no high temperature excursions. For most applications, the oils have to be genetically and/or chemically modified. In Europe, predominately rapeseed oil and sunflower oil are used. Chemically, these are esters of glycerin and long-chain fatty acids (triglycerides). The alcohol component (glycerin) is the same in all vegetable oils, while the fatty acid components are plant-specific and therefore variable. Natural triglycerides are very rapidly biodegradable and are highly effective as lubricants. However, their thermal, oxidative, and hydrolytic stability are limited (Bartz et al. 1988; Schneider et al. 2006).

The reason for the thermal and oxidative instability of vegetable oils is the double-bond portion of the fatty acid part and the beta-CH group of the alcoholic components. Double bonds in alkenyl chains are especially reactive with the oxygen in air. The beta-hydrogen atom is easily eliminated from the molecular structure. This leads to the cleavage of the esters into acid and an olefin. Hydrolysis is also a complicating factor. This problem can be mitigated through chemical modification. Vegetable oils are split into their oleochemical components such as fatty acids or fatty acid methylesters and glycerin before they are modified (Bartz et al. 1988; Schneider et al. 2006).

Reaction strategies beyond this step are too numerous to discuss here. Chemical modification can be aided with improved starting materials. Therefore, much attention has been given to genetic modification of the seed itself. For instance, a high oleic version of sunflower seeds was developed in the 1970s.The oil thus formed was extremely oxidatively stable when compared with products using seeds from a non-modified source. Future work will

Environmentally Friendly Lubrication Issues, Table 1 Typical physical properties of vegetable oils (refined and stabilized) and fatty acid esters (Anand and Chhibber 2006)

Properties	Rapeseed oil	Acetylated castor oil	Partially hydrogenated rice bran oil	Nonyl ester of acetylated hydrogenated castor oil fatty acids	Nonyl ester of partially hydrogentated rice bran oil fatty acids
Viscosity at 40°C	45.92	117.61	24.13	48.06	20.78
Viscosity at 100°C	8.97	15.75	6.26	8.33	5.58
Viscosity index	213	142	230	149	232
Pour point (°C)	−30	−30	−6	−25	−30
Flash Point (°C)	>300	280	276	>300	289
Noack evaporation loss (%)	1.3	1.6	1.9	1.4	1.6
Cu strip corrosion 3 h @ 100°C	1a	<1a	<1a	<1a	<1a
4 Ball wear (40 kg/h)	0.35	0.31	0.37	0.36	0.39
Friction coefficient	0.090	0.801			
Weld load (kg)	201	195			
Air release @ 50°C (min)	4	3	4		
Foaming characteristics					
ASTM D 892, 25°C	Nil/nil	Nil/nil	Nil/nil	Nil/nil	Nil/nil
Acid no. increase	1.57	0.56	0.3	0.13	0.15
Viscosity increase at 40°C	(64.29)	(141.01)	(30.26) 25.4%	(51.98) 6.1%	(24.08%)
Sludge value. . .insolubles	0.14	0.08	0.05	Nil/nil	0/02
Biodegradability	>90	>90	>90	>90	>90

need to be centered on improving the overall performance of modified vegetable oils while lowering the finished product cost. It is only through this route that these products will earn an increased portion of the overall lubricants market (Honary et al. 2001).

Life Cycle Analysis

No review of environmentally friendly lubricants would be complete without mention of the discipline referred to as "life cycle analysis." As with most forms of human endeavor, not all outcomes are positive. For example, no discussion of naturally occurring lubricants would be complete without mentioning the energy that is used in soil preparation, planting, fertilizing, and harvesting. These processes require considerable amounts of fossil fuel to be burned by the tractors that do the work. Fossil fuel is also typically used in the drying process prior to market. Trucks that transport the seeds to the processing plants further the use of diesel fuel and gasoline. Each of these events releases a significant amount of carbon to the atmosphere, a process that is trying to be mitigated through the use of environmentally friendly fluids in the first place.

A detailed life cycle analysis can be found in a comprehensive publication by Miller et al. in the journal *Environmental Science & Technology* (Miller et al. 2007). This work centered around the life cycle of soybean products when compared with comparable petroleum-based lubricants. The environmentally friendly lubricants, although representing a small portion of the total lubricants in use today, are increasing in popularity due to a combination of positive technical properties and a desire to reduce overall impact on the environment by lubrication practices. Rigorous Monte Carlo analyses were carried out in order to quantify the pluses and minuses of lubricant type. The focus was on the effect that the life cycle had on the elements carbon and nitrogen. The study concluded that, on balance, there are tangible benefits to be gained through the use of soybean-based lubricant products, although these benefits are not as great, or as clear-cut as one might originally envision. However, there may be benefits to certain populations that rely greatly on imported fossil fuel and it precursors. Agricultural-based materials offer the opportunity for renewable products that may benefit a net importer in addition to moderate atmospheric carbon and nitrogen reduction.

Environmentally Friendly Lubrication Issues, Table 2 Properties of complex esters and dibasic acid ester (Anand and Chhibber 2006)

Characteristics	TMP vegetable oil trioleate fatty acid ester	Polyol Diester	Polyol complex esters	Di-isotridecyl adipate	Di-isotridecyl sebacate	Unsaturated fatty acid ployester
Density 20°C	0.915	0.935	0.948	0.912	0.923	0.938
Viscosity at 40°C	46	32.8	57.9	26.6	36.7	47.2
Viscosity at 100°C	9.00	5.82	8.02	5.30	6.70	9.30
Viscosity index	190	121	124	136	141	189
Pour point (°C)	−39	−54	−49	−40	−52	−40
Acid value	1.2	0.02	0.92	0.3	0.1	0.6
Flash point (°C)	290	290	300	245	259	300
Corrosion Cu strip 3 h @100°C	1a	<1a	<1a	<1a	<1a	<1a
4 Ball wear (40 kg, 1 h) WSD (mm)	0.35	0.42	0.4	0.44	0.46	0.39
Volume of Foam (ML)						
Sequence 1 25°C	10/0	12/0	7/0	5/0	Nil/nil	7/0
Sequence 2 95°C	8/0	10/0	5/0	3/0	Nil/nil	5/0
Sequence 3 25°C	10/0	12/0	7/0	5/0	Nil/nil	7/0
Air release @ 50°C	4	5	5	5	5	5
Oxidation stability 95°C for 120 h						
Acid no. increase	1.1	0.5	0.65	0.1	0.05	0.81
Viscosity increase (%)	+22.5	+6	+13	−1.1	−0.5	+26
Sludge value/100 ml	2.3	0	3.8	0	0	4.5
Biodegradability	90	90	80	85	85	90
Noack evaporation loss (%)	2.3	1.8	1.3	6.5	5.3	0.9

Summary

Because of the high cost associated with the use of environmentally friendly lubricants, today these lubricants are mainly used in a total loss lubrication situation. The list includes bar and chain oils for logging, rail greases, two-stroke oils for marine application, and mold release fluids.

Although vegetable oils have particularly good tribological properties, combining good boundary friction lubricity and general wear protection with stable viscosity/temperature behavior and low volatility, the growth of environmentally friendly fluids has been rather slow during the latter part of the twentieth century and the first part of the twenty-first. Fluids and greases that are formulated to be more compatible with the environment, both in service and as used products, are intrinsically more expense to formulate and, in most cases, represent some sort of compromise in terms of the performance that they provide and their productive time in the equipment. Many innovations have been made that bring these products more in line with the performance that end users have grown to expect from more conventional lubricants. This progress will continue, making the emerging fluids and greases even more cost competitive. Governments will also play a significant role in the success of environmentally friendly lubricant products. Regulations requiring their use will be helpful in spurring interest. Finally, life cycle analysis must continue to be part of the equation when establishing a compelling case to use environmentally friendly lubricants.

Cross-References

► Natural Oils as Lubricants
► Oil Life

References

O.N. Anand, V.K. Chhibber, Vegetable oil derivatives: environment-friendly lubricants and fuels. J. Synth. Lubr. **23**, 91–107 (2006). doi:10.1002/jsl.14. Published online in Wiley InterScience (www.interscience.wiley.com)

W.J. Bartz, Technische Akadeie Esslingen, Ostfildern, Germany, Lubricants and the environment, Nov 22 (1988)

T. Di-Hua, Y. Bin, Modification of the chemical structure of an environmentally-friendly castor oil lubricant. J. Synth. Lubr. **21**, 59–64 (2004). doi:10.1002/jsl.3000210106

L.A.T. Honary, Biodegradable/biobased lubricants and grease. Machinery Lubrication Magazine, Sept (2001)

S. Kumar, Top-of-rail lubrication system for energy reduction in freight transport by rail, SAE Technical Paper Series #1999-01-2236 (1999)

S. Levi Pearson, J.E. Spagnoli, Environmental lubricants –an overview of onsite applications and experience. Lubrication Engineering Magazine, April (2000)

S.A. Miller, A.E. Landis, T.L. Theis, R.A. Reich, A comparative life cycle assessment of petroleum and soybean-based lubricants. Environ. Sci. Technol. **41**(11), 4143–4149 (2007)

R.A. Padavich, L.A.T. Honary, A market research and analysis report on vegetable-based industrial lubricants, SAE Technical Paper Series # 952077 (1995)

M.P. Schneider, Plant-oil-based lubricants and hydraulic fluids. J. Sci. Food Agric. **86**, 1769–1780 (2006). Copyright © 2006 Society of Chemical Industry

Soy Lubricants, Higher viscosity, lower evaporation loss and great potential, USB (United Soybean Board) www.soynewuses.org/Lubricants

H. Wagner, R. Luther, T. Mang, Lubricant base fluids based on renewable raw materials their catalytic manufacture and modification. Appl. Catal. A Gen. **221**, 429–442 (2001)

A. Willing, Lubricants based on renewable resources – an environmentally compatible alternative to mineral oil products. Chemospere **43**, 89–98 (2001)

EP Oil or Lubricant

► Gear Lubricants

Epicyclic Gear Trains

Robert F. Handschuh
NASA Glenn Research Center, Cleveland, OH, USA

Synonyms

Planetary gear systems; Planetary gear trains; Solar planetary; Star planetary

Definition

An epicyclic gear train is a coaxial speed reducer or increaser stage comprised of a sun gear, planet gear(s), and a ring gear (Townsend 1992; Coy et al. 1985). The ratio attained from the gear train depends on the component that has its rotational motion constrained or controlled. The gears can be spur, helical, or double helical in these gear trains. An epicyclic gear train is an excellent load-sharing device. The input torque is split between the planet gears, thus increasing the load capacity of the gear train.

Scientific Fundamentals

Manufacturing Methods

The gears of an epicyclic gear train are manufactured with the same tooling as would be required for any other external or internal gears. The gear surface geometry is typically involute, but other tooth forms can be used in this type of gear train.

Epicyclic Gear Train Types

There are three basic forms of an epicyclic gear train. Figure 1 shows the three types as (a) planetary (ring gear fixed), (b) star (carrier fixed), or (c) solar (sun gear fixed). High reduction ratio is typically accomplished using a planetary gear train while smaller reduction ratios are designed with the solar arrangement. A description of each of the arrangements and the ratio attained is provided in Table 1 (Townsend 1992). If the epicyclic gear train is driven as a speed increaser, the ratio is the reciprocal of the values calculated in Table 1.

Epicyclic Gear Train Assembly

In order to have planets with equal circumferential spacing (the angle between each planet is identical), the following equation must be satisfied:

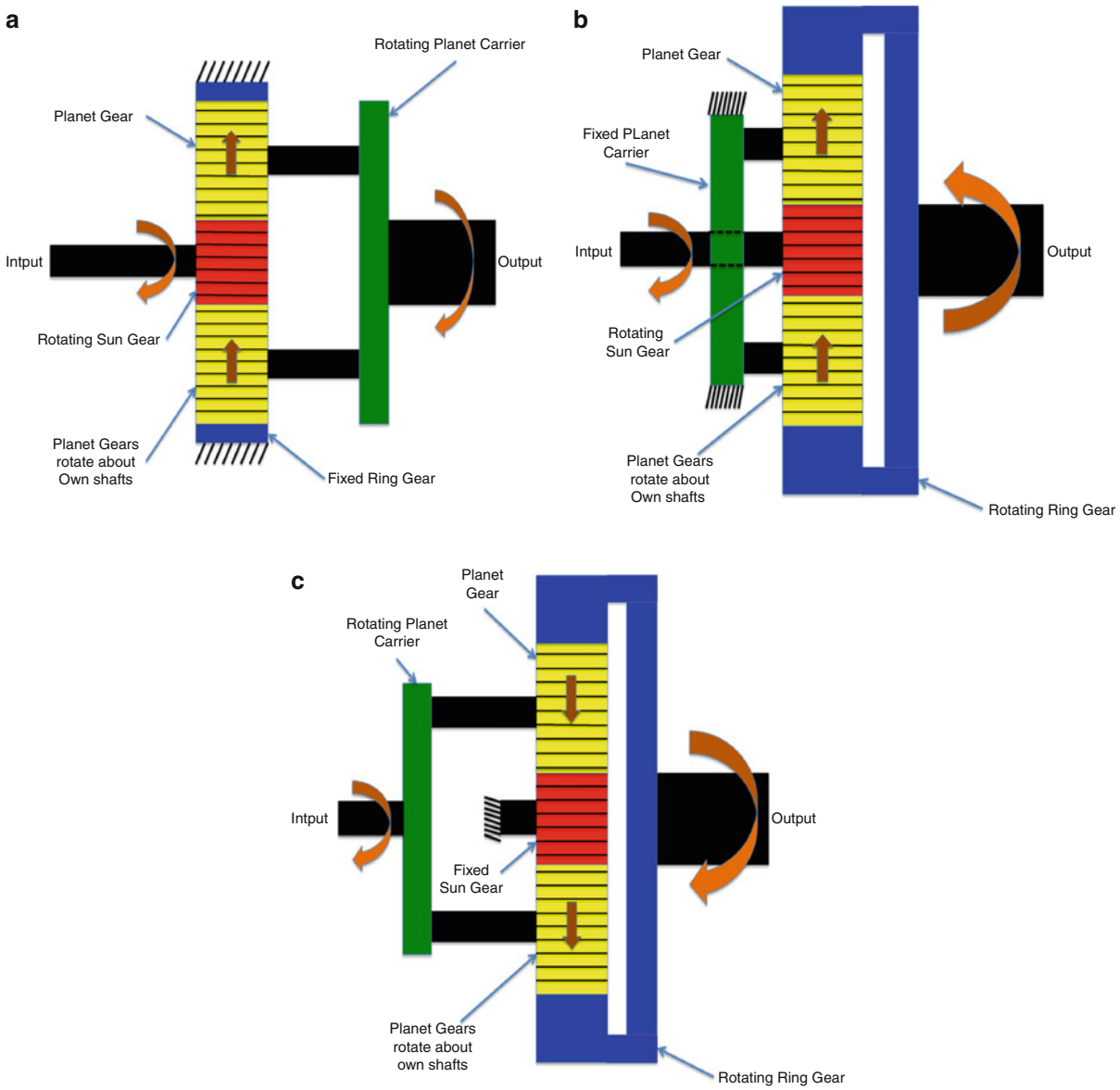
$$I = (N_3 + N_2)/m \tag{1}$$

where I is an integer, m is the number of planets, N_3 is the number of teeth on the ring gear, and N_2 is the number of teeth on the sun gear.

As an example, assume: $N_1 = 18$, $N_2 = 27$, and, $N_3 = 63$. From these choices of tooth count and assuming that it is a planetary gear system, the ratio would be equal 3.3333:1. For this case $N_3 + N_2$ is equal to 90. This is divisible (integer, by the number of planets) by 3, 5, and 6. More planets would not fit because they would interfere with one another.

Compound Epicyclic Gear Trains

A planetary with two planets on a common shaft is called a compound planetary. An example of how this epicyclic



Epicyclic Gear Trains, Fig. 1 (**a**) Planetary gear train. (**b**) Star planetary. (**c**) Solar planetary

Epicyclic Gear Trains, Table 1 Epicyclic gear train configuration data

Epicyclic type	Fixed member	Input member	Output member	Overall ratio	Typical range of ratios
Planetery	Ring	Sun	Carrier	$(N_3/N_2) + 1$	3:1–12:1
Star	Carrier	Sun	Ring	(N_3/N_2)	2:1–11:1
Solar	Sun	Ring	Carrier	$(N_2/N_3) + 1$	1.2:1–1.7:1

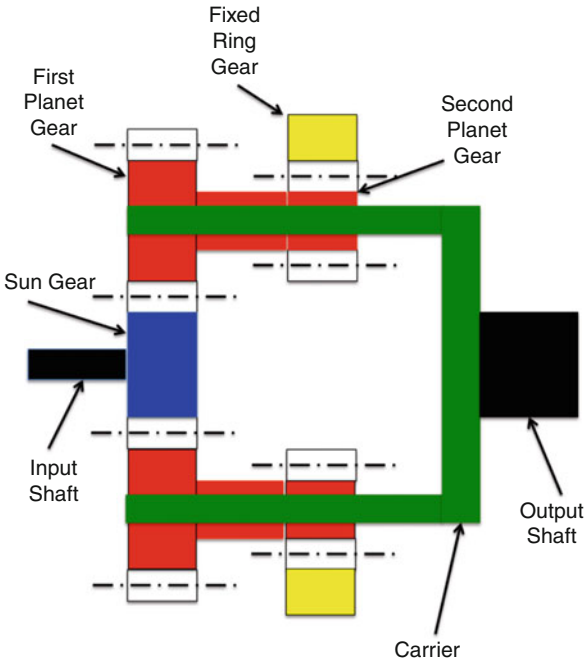
Note:
N_1 – number of teeth on planet gear
N_2 – number of teeth on sun gear
N_3 – number of teeth on ring dear

gear train is arranged is shown in Fig. 2. To calculate the ratio of the various configurations, the necessary equations are shown in Table 2. As was mentioned in the previous section, there are requirements that need to be met for this type of gear train to be assembled. The following equations must be satisfied to assemble or have a certain number of planets:

$$d_{p3} = d_{p2} + d_{p11} + d_{p12} \tag{2}$$

where:

d_{p3} = ring gear pitch diameter
d_{p2} = sun gear pitch diameter



Epicyclic Gear Trains, Fig. 2 Compound epicyclic gear train

d_{p11} = first planet pitch diameter
d_{p12} = second planet pitch diameter

Note that the number of teeth on each of the gears could be substituted for the pitch diameters.

and

$$I = \frac{(N_{11}N_3 + N_{12}N_2)}{m} \tag{3}$$

where m is the number of planets and I must be an integer.

Coupled Epicyclic Gear Trains

Combining various elements of simple planetary gear trains with combinations of multiple sun, planet carrier and ring gears can form coupled epicyclic gear trains. From Reference (Townsend 1992), some examples of these combinations are given in Fig. 3. For these more complicated gear trains the requirements for assembly and number of planets can be treated as a simple planetary for each stage, but the overall ratios will need to be computed using the tabular method. This method can be found in textbooks on the subject (Townsend 1992; Mabie and Ocvirk 1975).

Key Applications

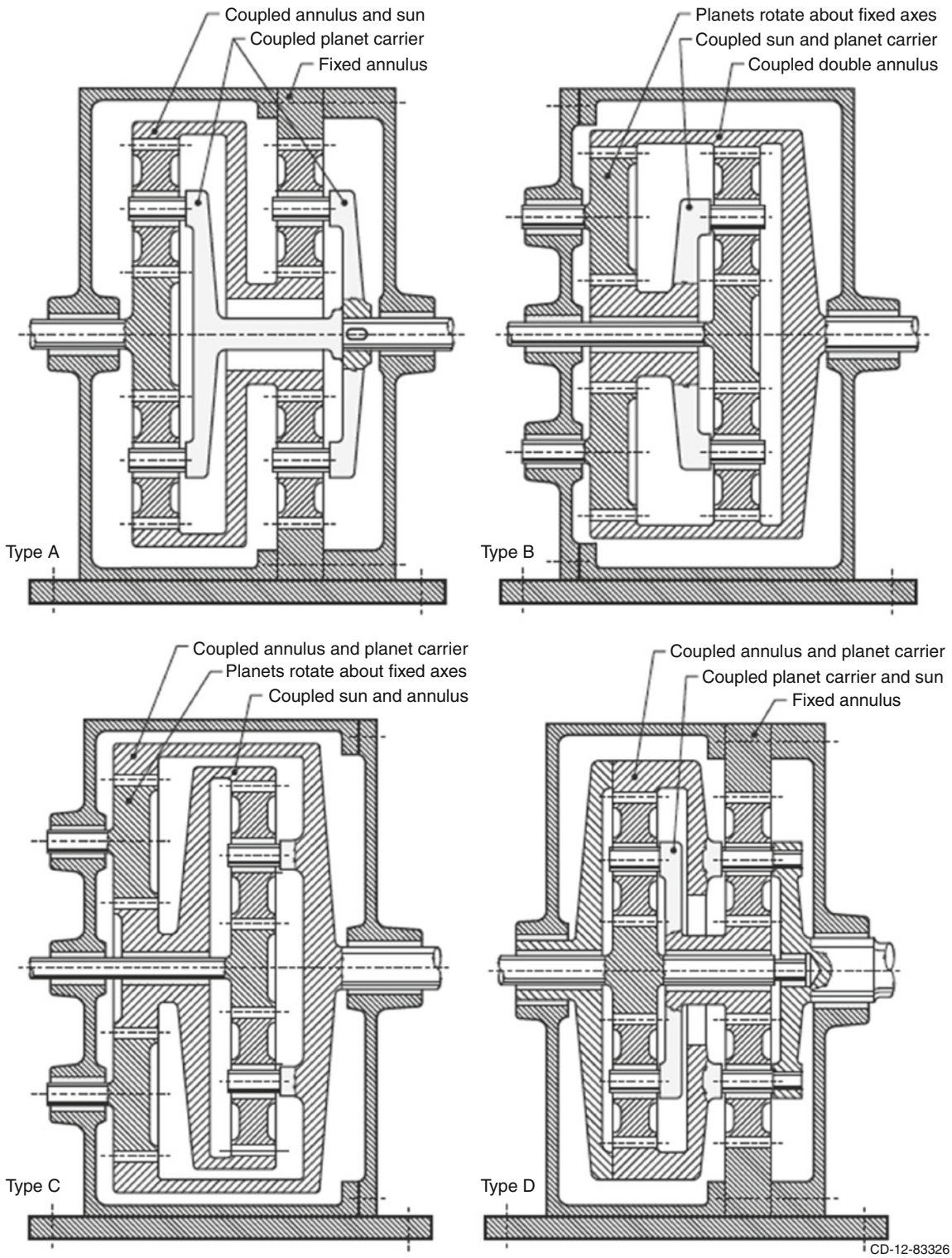
Application of Epicyclic Gear Trains

The application of this type of gear train is very common in automobile automatic transmissions, helicopter transmissions (Fig. 4a, b), and gas turbine engine propulsor applications (Fig. 5, McCune et al. 1993). In these applications, the co-linearity of the gear train offers many advantages for packaging. In automobiles, the planetary system components are locked or released to provide the transmission with multiple ratios. In the gas turbine application, the epicyclic gear train is used to match the turbine and fan performance – providing a highly efficient overall

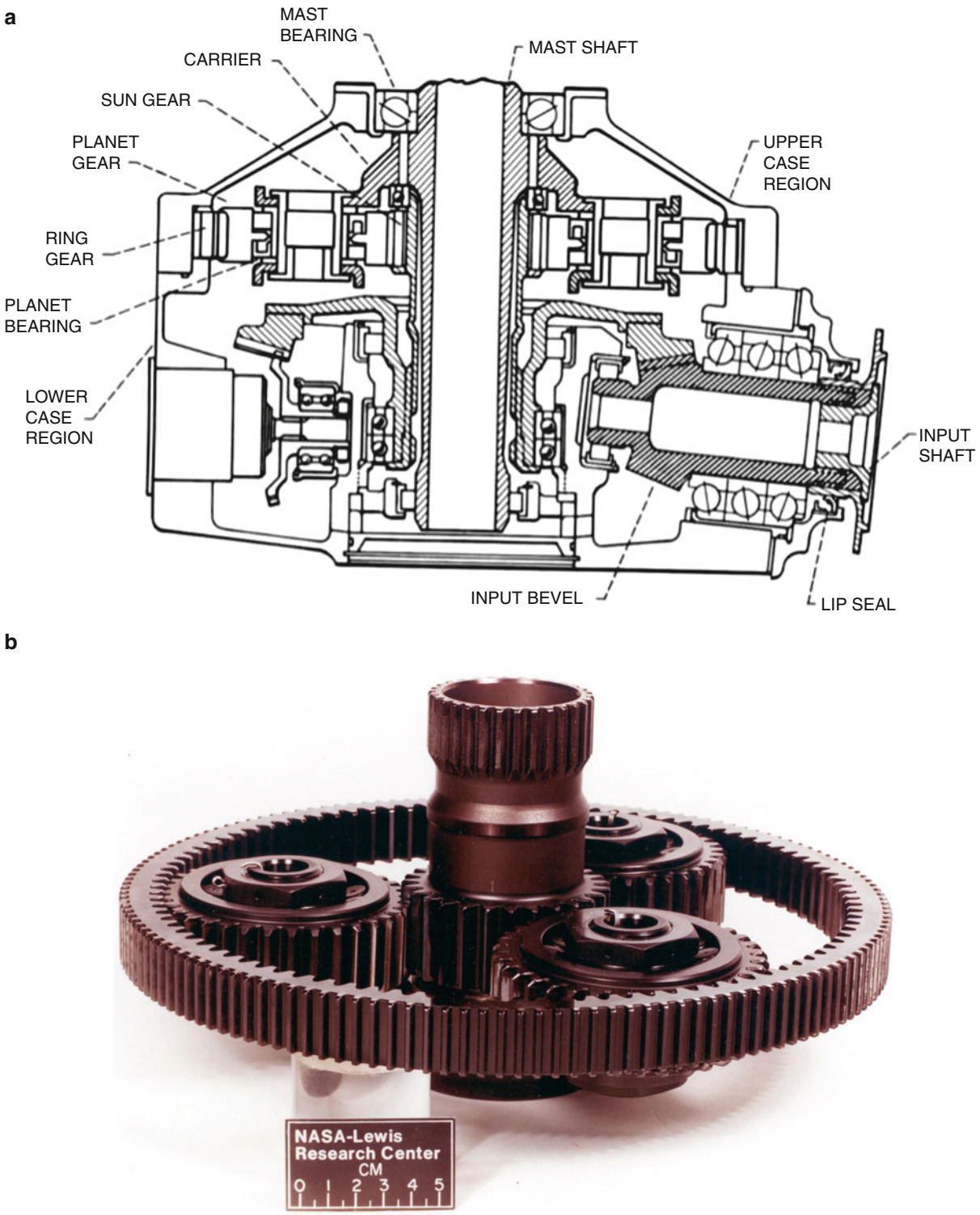
Epicyclic Gear Trains, Table 2 Compound epicyclic gear train ratios

Epicyclic type	Fixed member	Input member	Output member	Overall ratio	Typical range of ratios
Compound planetary	Ring	Sun	Carrier	$(N_3N_{11})/(N_2N_{12}) + 1$	6:1–25:1
Compound star	Carrier	Sun	Ring	$(N_{11}N_3)/(N_2N_{12}) + 1$	5:1–24:1
Compound solar	Sun	Ring	Carrier	$(N_2N_{12})/(N_3N_{11}) + 1$	1.05:1–2.2:1

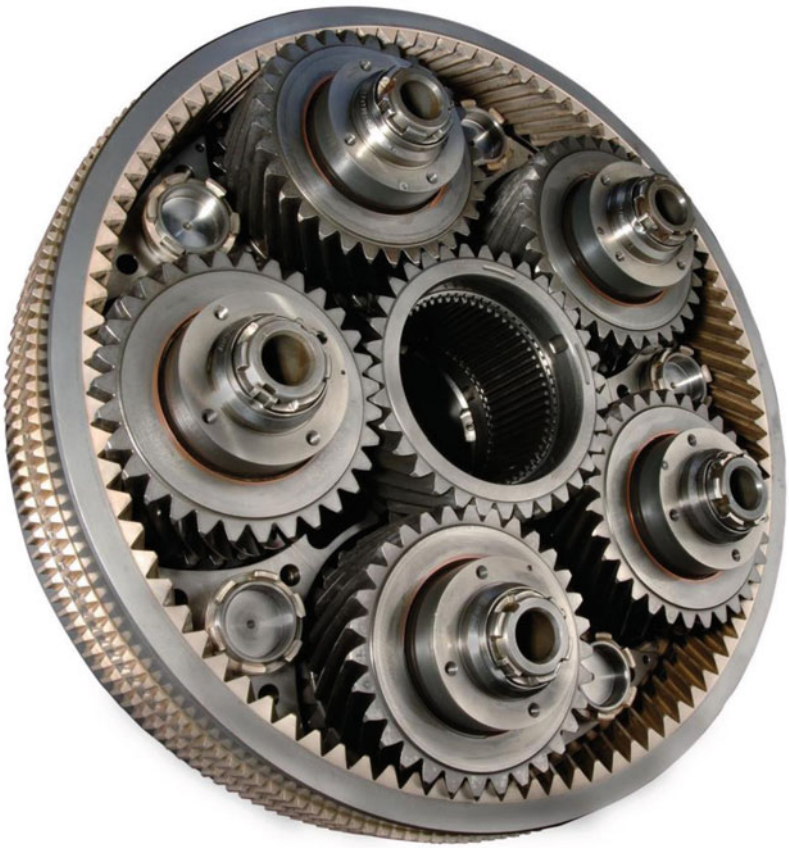
Note:
N_{11} – number of teeth on planet gear first reduction stage
N_{12} – number of teeth on planet gear second reduction stage
N_2 – number of teeth on sun gear
N_3 – number of teeth on ring gear



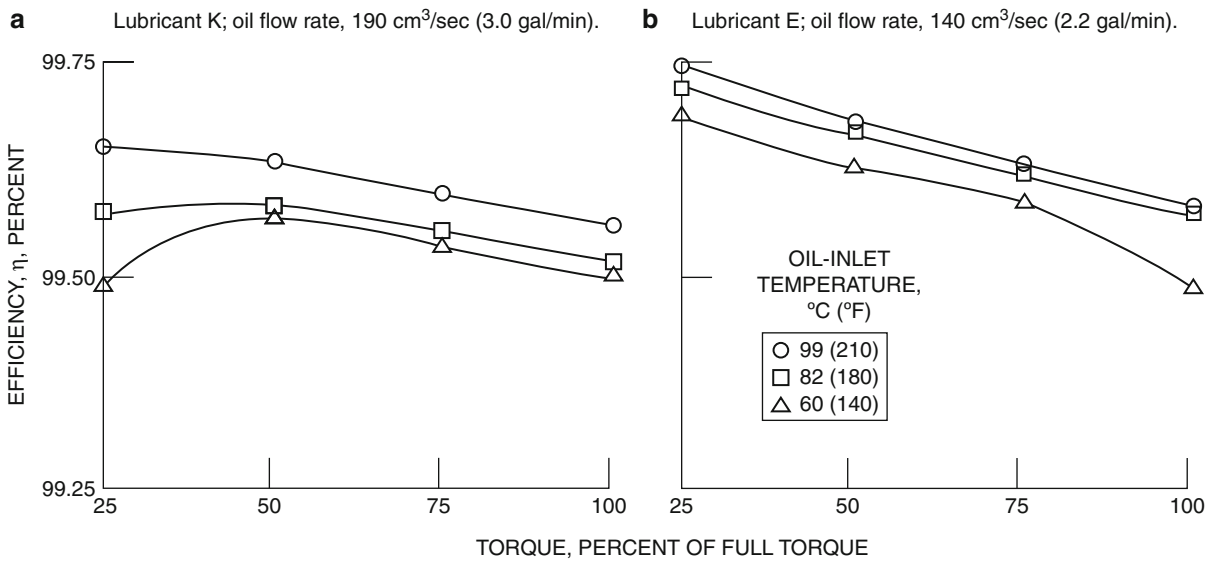
Epicyclic Gear Trains, Fig. 3 Examples of coupled epicyclic gear trains



Epicyclic Gear Trains, Fig. 4 (**a**) Cross-section of a two-stage (bevel then planetary) helicopter transmission (Coy et al. 1985). (**b**) Photograph of a three-planet epicyclic gear train (Coy et al. 1985)



Epicyclic Gear Trains, Fig. 5 Geared turbofan engine star planetary (McCune et al. 1993)



Epicyclic Gear Trains, Fig. 6 Example of aerospace planetary gearing efficiency as a function of lubricant type and temperature (Handschuh et al. 1988)

propulsion package. In the helicopter drive-system, the epicyclic gear train(s) are used as the final stage(s) before outputting the large torque – low speed to the main rotor. The load sharing capabilities and large reduction available make this arrangement very advantageous.

Performance of Epicyclic Gear Trains

The performance of epicyclic gear trains has been studied in several papers in work conducted at NASA (Anderson et al. 1980; Handschuh et al. 1988). An example of the efficiency is shown in Fig. 6. In this figure, the efficiency of a four-planet gear stage from a helicopter (313 kW) as a function of applied torque is shown for two different operational conditions. For this aerospace arrangement (high-accuracy components), the efficiency ranged from 99.4% to 99.7%. This high efficiency can be expected from aerospace manufactured components. A typical less accurate epicyclic gear stage should produce at least a 99% efficiency.

Special Considerations for Epicyclic Gear Trains

In epicyclic gear trains, several design considerations need to be included to attain long life. Planet spacing, number of planets, and choices for tooth numbers can affect the fatigue life of the gear and bearing components. In aerospace planetary gear stages, usually the sun gear life is of the most concern due to the high cycle count. The planet gear has to be carefully designed due to the reversed bending stress every revolution. Also important is the flexibility of the ring gear. In aerospace systems the ring gear attachment to the gear housing is usually designed such that the ring is able to flex and therefore improve load sharing between the planet gears.

Cross-References

► Gear Efficiency
► Helical Gears
► Spur Gears

References

N. Anderson, S. Lowenthal, *Spur-Gear Efficiency at Part and Full Load*, NASA TP-1622, 1980

J. Coy, D. Townsend, E. Zaretsky, E. *Gearing*, NASA RP-1152, 1985

R. Handschuh, D. Rohn, *Efficiency Testing of a Helicopter Planetary Reduction Stage*, NASA TP-2795, February 1988

H. Mabie, F. Ocvirk, *Mechanisms and Dynamics of Machinery*, 3rd edn. (Wiley, New York, 1975)

M. McCune, *Initial test results of 40,000 horsepower fan drive gear system for advanced ducted propulsion systems*. AIAA, SAE, ASME, and ASEE, Joint Propulsion Conference and Exhibit, 29th, Monterey, CA, 28–30 June 1993

D. Townsend, *Dudley's Gear Handbook*, 2nd edn. (McGraw-Hill, New York, 1992)

ε-N Curve

► Strain-Life Theories

Equation of State

► Temperature and Pressure Dependence of Density and Thermal Conductivity of Liquids

Equations for EHL

► EHL Governing Equations

Equations of Newtonian Fluid Flow

► Navier-Stokes Equation and Applications in Lubrication

Equivalent Discrete Spherical Convolution (EDSC)

► Elasticity Theory for Spherical Bearings

Equivalent Discrete Spherical Convolution (EDSC) for Contact Analysis

► Contact Mechanics for Spherical/Aspheric Bearing

Equivalent Spherical Convolution (ESC)

► Elasticity Theory for Spherical Bearings

ESC-Equivalent Spherical Convolution for Spherical-Bearing Lubrication Analysis

► Lubrication Theory for Spherical Bearings

ESC-Equivalent Spherical Convolution Considering Spherical/Aspheric Geometry

► Geometry of Spherical/Aspheric Bearings

ESC-Equivalent Spherical Convolution for Contact Analysis

► Contact Mechanics for Spherical/Aspheric Bearing

ESC-Equivalent Spherical Convolution for Joint Simulation

► Biotribological Joint Simulation System

ESC-Equivalent Spherical Convolution for Spherical-Bearing Friction Prediction

► Friction Prediction for Spherical Bearings

ESC-Equivalent Spherical Convolution for Spherical-Bearing Wear Modeling

► Wear Modeling of Spherical Bearings

Eshelby's Equivalent Inclusion

► Inhomogeneous Inclusion in Materials

Eshelby's Inclusion

► Inclusions Subjected to Eigenstrain

Evaluation of the Fractal Dimension of Rough Surfaces

► Fractal Characterization of Surfaces

Exo-electron Emission and Tribochemistry

► Triboemission, Triboplasma Generation, and Tribochemistry

Expanded Austenite

► Low Temperature Carburization

Experimental Biotribology of Joint Replacements

► Friction in Joint Replacements

Externally Pressurized Bearings

► Hydrostatic Thrust Bearings

Externally Pressurized Bearings Under Hydrodynamic Conditions

► Hybrid Hydrostatic/Hydrodynamic Bearings

Externally Pressurized Journal Bearings

► Hydrostatic Journal Bearings

Extreme Pressure Additives

► Ashless Phosphate Esters

Extreme-Temperature Lubrication

► Tribochemistry of Extreme-Pressure Additives