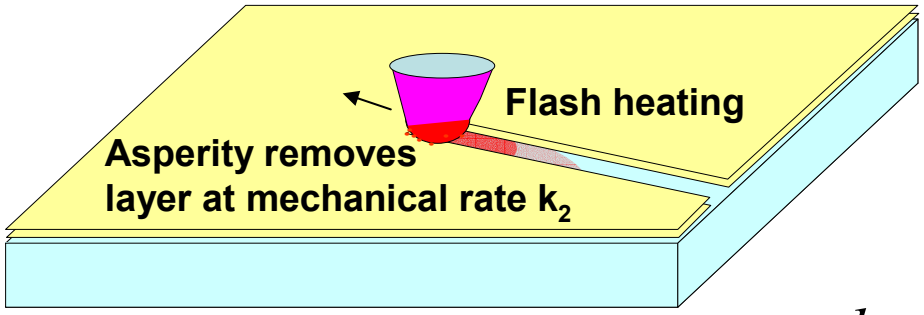
Flash Heating in Chemical-Mechanical Polishing

L. Borucki
Intelligent Planar

J. Sorooshian, Z. Li, Y. Zhuang and A. Philipossian
University of Arizona

Removal Rate Modeling

Silicon dioxide removal rates can be described accurately with a Langmuir-Hinshelwood model plus a model for flash heating by pad asperity tips.



Surface layer grows at chemical rate k_1. Growth is fastest at the flash temperature.

$$k_1 = A \cdot \exp\left(\frac{-E}{kT}\right)$$ Chemical rate

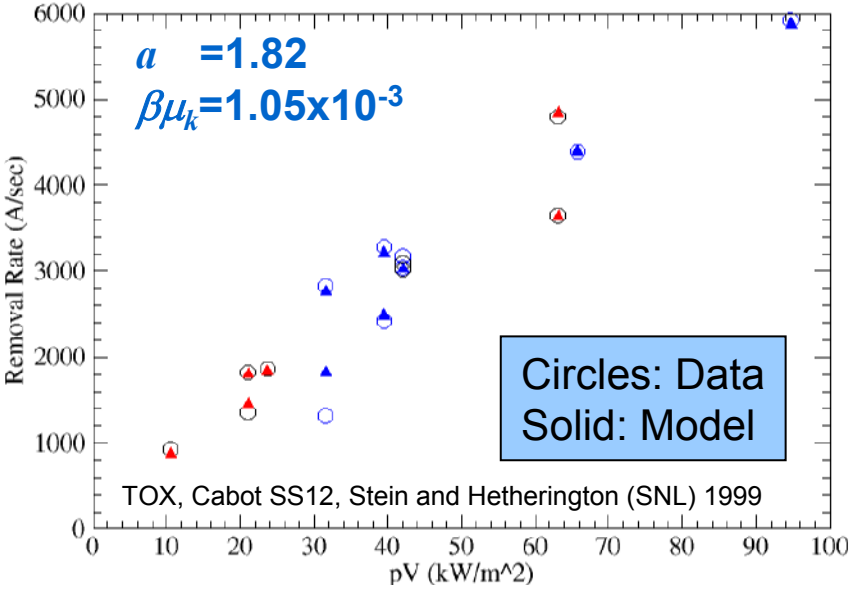
Mechanical rate

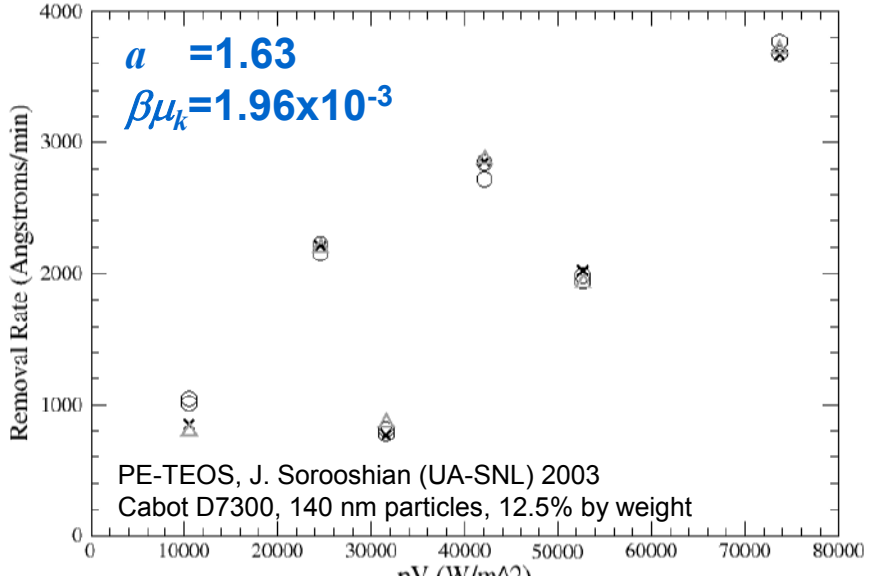
$$k_2 = c_p \mu_k pV$$

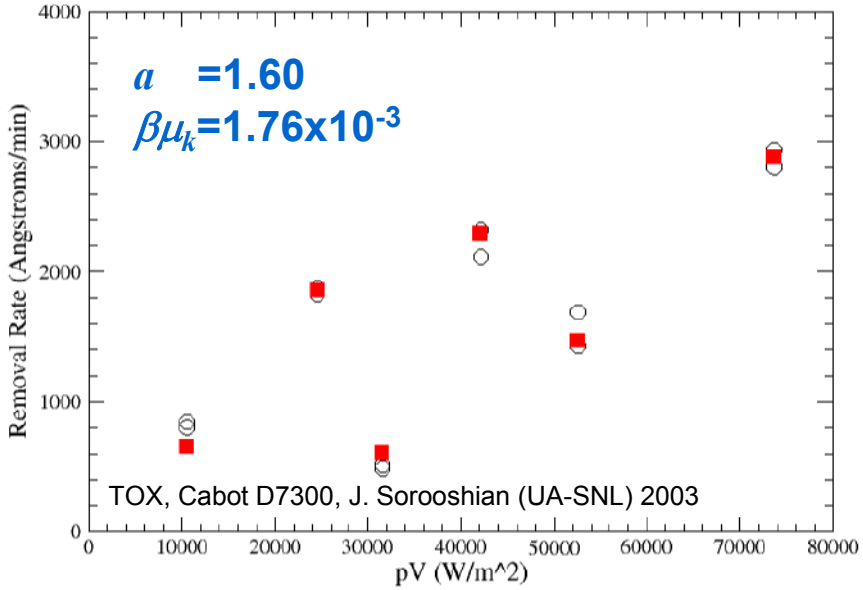
$$T = T_a + \frac{\beta}{V^a} \mu_k pV$$ Flash heating. T is required by every chemistry model.

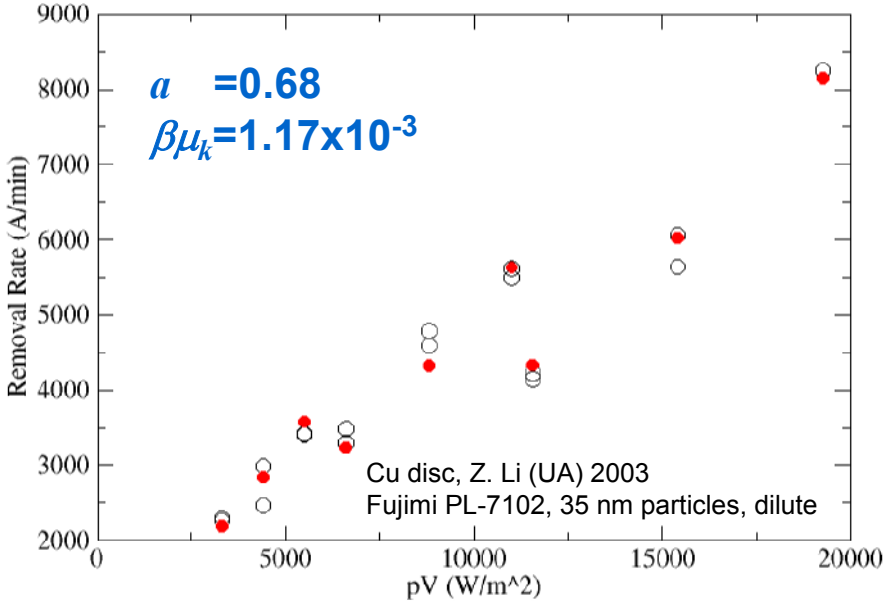
$$RR = \frac{M_w}{\rho} \frac{k_2 k_1}{k_2 + k_1}$$ Langmuir-Hinshelwood Model

Comparisons with Data









Questions

For the flash heating model,

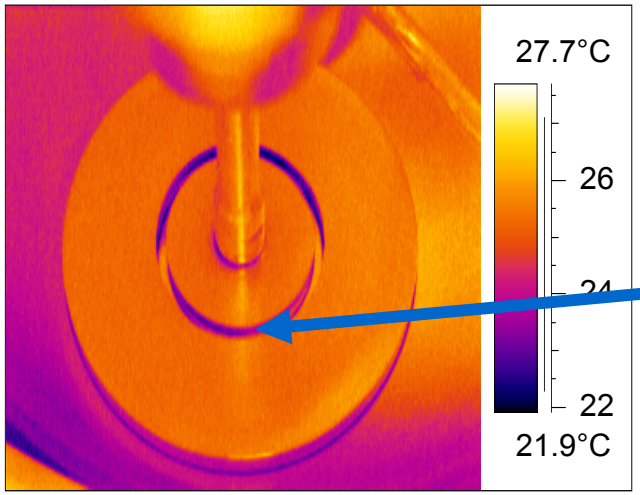
$$T = T_a + \frac{\beta}{V^a}\mu_k pV$$

Why does a power law work so well?

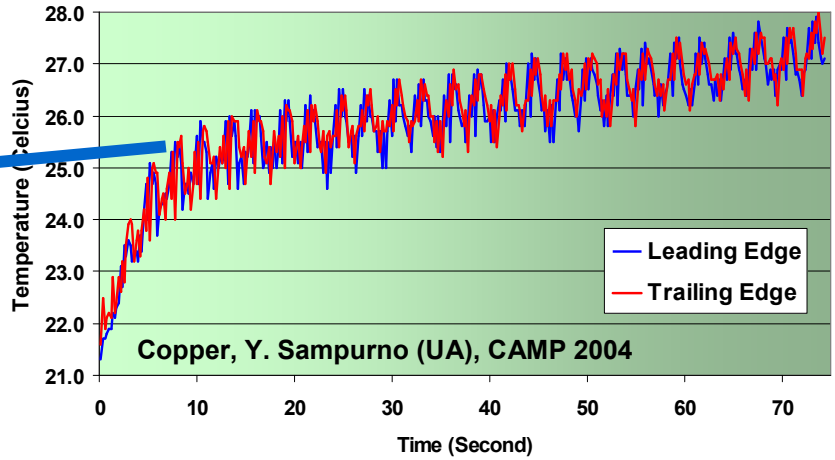
Why do β and a have the observed values?

Why is "a" smaller for copper than for SiO_2?

Body Temperature vs. Flash Temperature

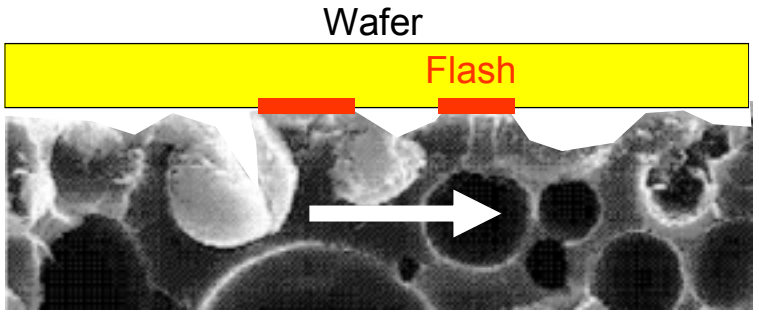


The wafer *body temperature* is the temperature that is measured by an IR camera. Typical temperature increases may be 5-15 °C.

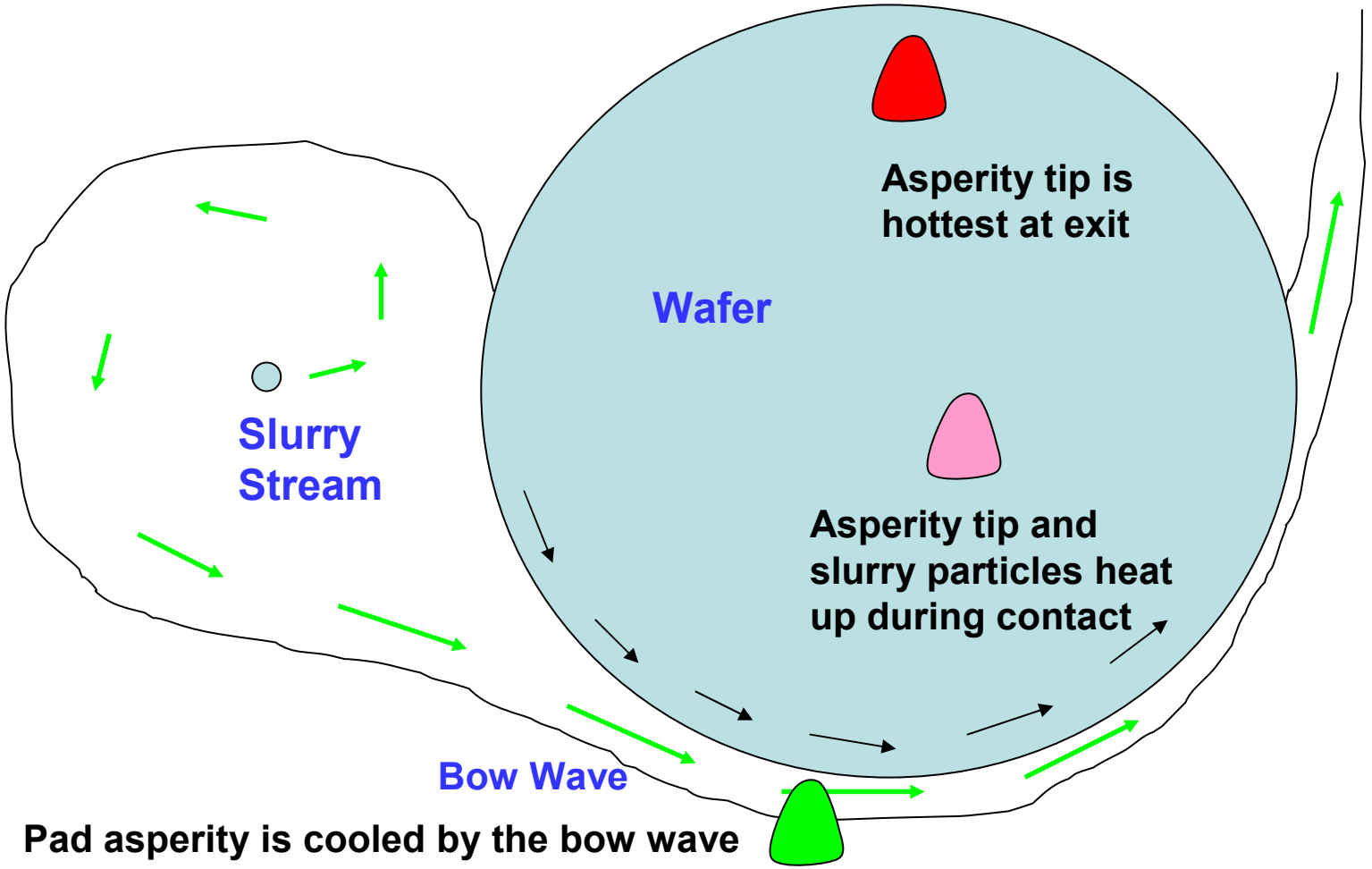


***Flash heating* is very localized, transient heating of the wafer by pad asperities.**

The flash temperature rise can be several tens of degrees C. It is higher than the body temperature increase because the real contact area between the pad and wafer is usually a small fraction of the wafer area.



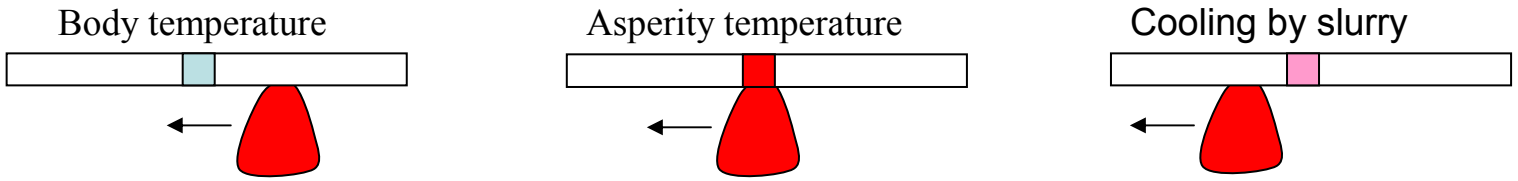
Flash Heating



Pad asperity is cooled by the bow wave within a few C of ambient temperature.

Mathematical Formalization

If there is temperature continuity, the wafer surface temperature transiently matches the asperity temperature during contact, followed by rapid cooling.



The mean reaction temperature taken over the wafer is therefore approximately the mean asperity tip temperature under the wafer.

From a really simple surface heating estimate, the asperity tip temperature rise after contact time τ at sliding speed V and mean real contact pressure p_a is approximately

Fraction of power density transferred to pad (heat partition factor)

COF

$$\theta(\tau) = \frac{2}{\sqrt{\pi}} \frac{\gamma_p \mu_k p_a V}{\sqrt{\kappa \rho C_p}} \tau^{1/2}$$

Pad properties

Mathematical Formalization

The mean asperity tip temperature rise averaged over the wafer surface is

$$\bar{\theta} = \frac{1}{\pi r_w^2} \iint \theta(\tau) dA$$

Since the wafer is tilted, p_a varies with τ. Let's ignore this variation.
Then

$$\bar{\theta} = \zeta(c_w, r_w) \frac{2}{\sqrt{\pi \kappa \rho C_p}} \frac{\gamma_p (p_a / p)}{V^{1/2}} \mu_k p V$$

Geometric factor that depends on the wafer size and location.

The mean reaction temperature then looks like

$$\bar{T} = \bar{T}_b + \frac{2\zeta}{\sqrt{\pi \kappa \rho C_p}} \frac{\gamma_p (p_a / p)}{V^{1/2}} \mu_k p V \quad \overset{?}{\longleftrightarrow} \quad T = T_a + \frac{\beta}{V^a} \mu_k p V$$

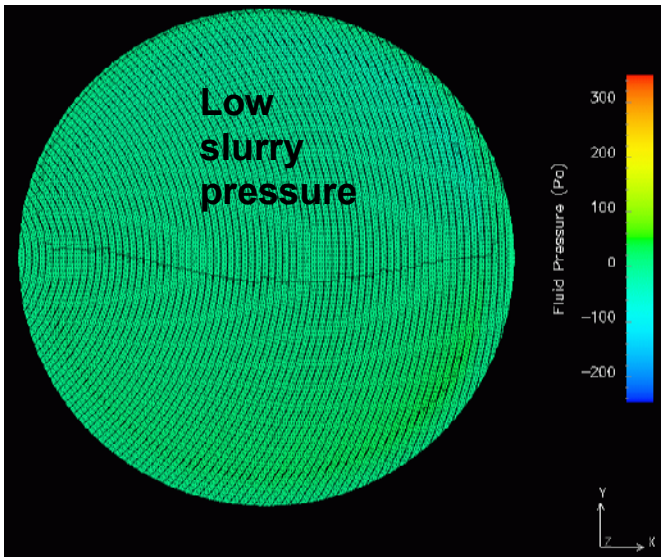
Mean wafer body temperature, which experimentally is a few C above ambient.

Flash Heating: Hydrodynamic Effects

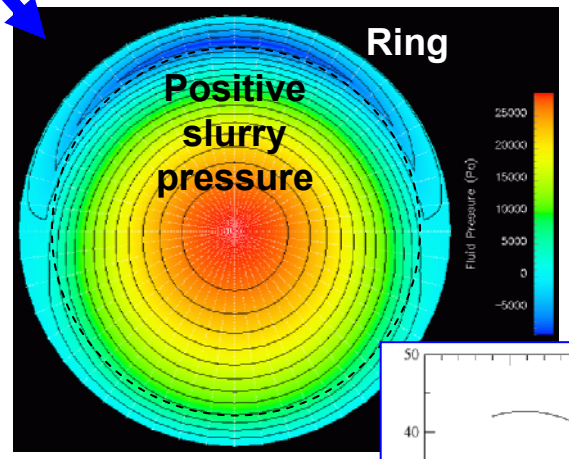
The mean real asperity contact pressure is influenced by hydrodynamic forces that can be related to pad grooving and the polishing head design.

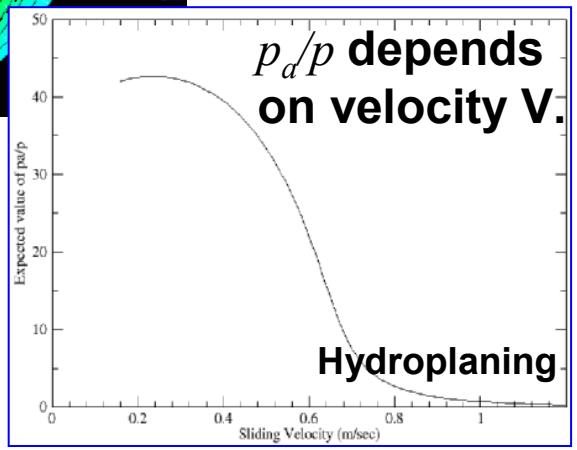
$$\overline{T} = \overline{T}_b + \frac{2\zeta}{\sqrt{\pi\kappa\rho C_p}} \frac{\gamma_p (p_a / p)}{V^{1/2}} \mu_k p V$$

Grooved pad



Plain pad, wafer with retaining ring



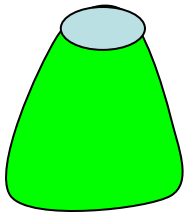


p_d/p **is nearly constant with V.**

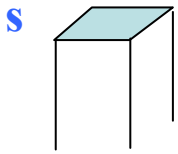
The experiments reported here were all run using k-groove pads.

Finite Element Simulation of Heat Partitioning

Part of the heat partitioning calculation involves estimating the size of the *mean contact area* between asperities and the wafer surface.

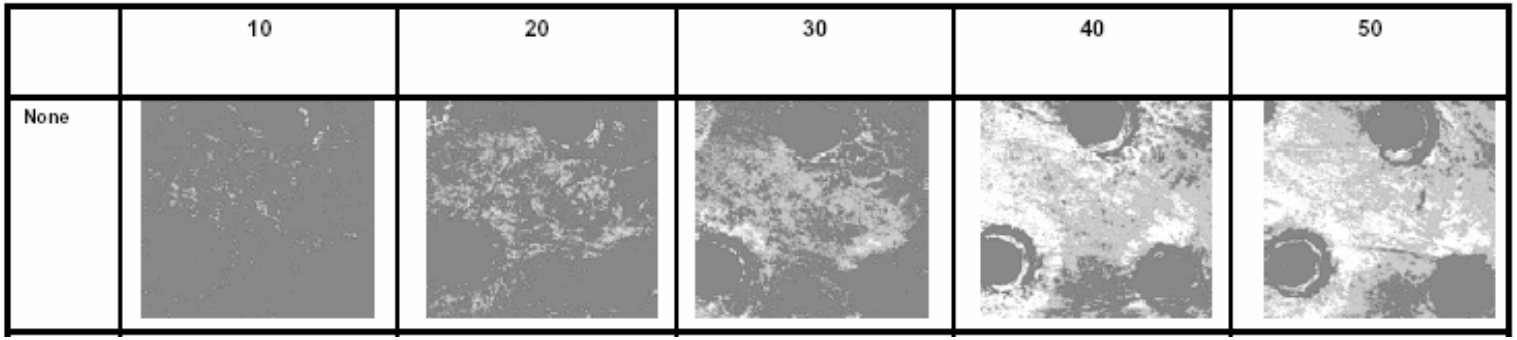


In Greenwood and Williamson theory, an asperity contact is always circular



We approximate it here with a square of the same area for simplicity of geometry construction.

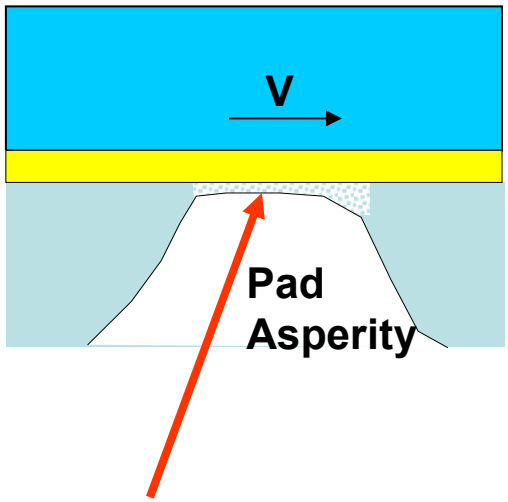
In reality, asperity contacts may be irregular.



Borucki, Lee, Zhuang and Philipossian, AIChE, Nov. 2004

Finite Element Simulation of Heat Partitioning

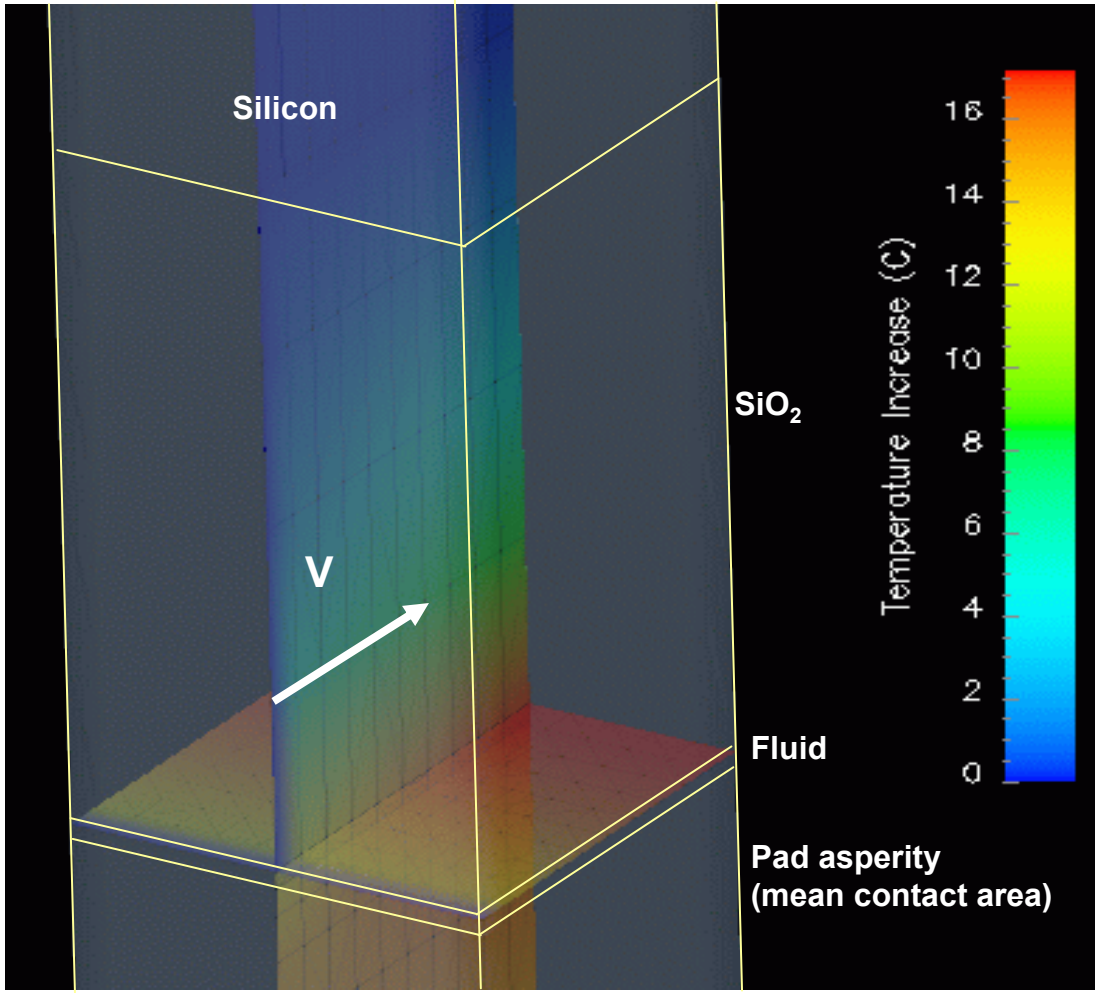
$$\overline{T} = \overline{T}_b + \frac{2\zeta}{\sqrt{\pi\kappa\rho C_p}} \frac{\gamma_p (p_a / p)}{V^{1/2}} \mu_k p V$$



Nanolubrication Layer

Fluid shearing
Active slurry particles (140 nm)
Silica polishing debris (glaze)

Glazing seems to be thermally important.

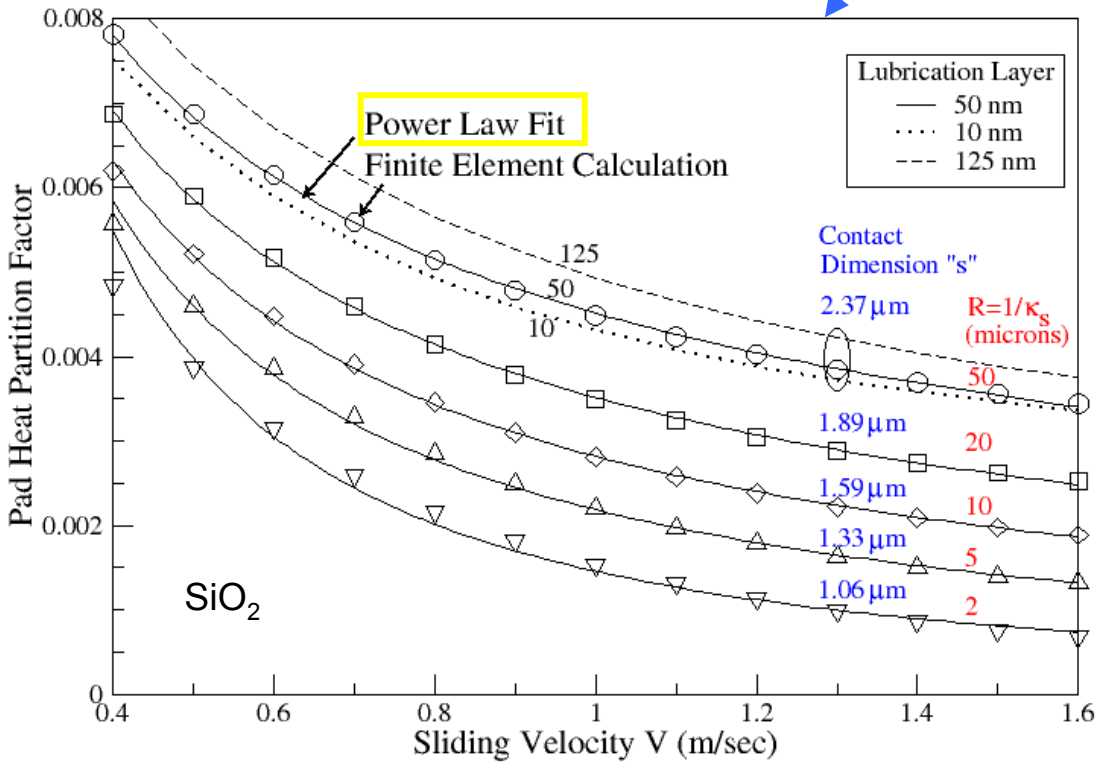


Lubricated Pad Asperity Heating

Finite Element Simulation of Heat Partitioning

The fraction of frictional heat transferred to an asperity depends on velocity and on the thermal properties of the pad, wafer surface and the lubrication layer.

$$\overline{T} = \overline{T}_b + \frac{2\zeta}{\sqrt{\pi\kappa\rho C_p}} \frac{\gamma_p(p_a/p)}{V^{1/2}} \mu_k p V$$



Since $\gamma_p(V,s)$ is a power law

$$\gamma_p = \frac{\gamma_p^1}{V^e}$$

then

$$\overline{T} = \overline{T}_b + \frac{\beta}{V^a} \mu_k p V$$

where

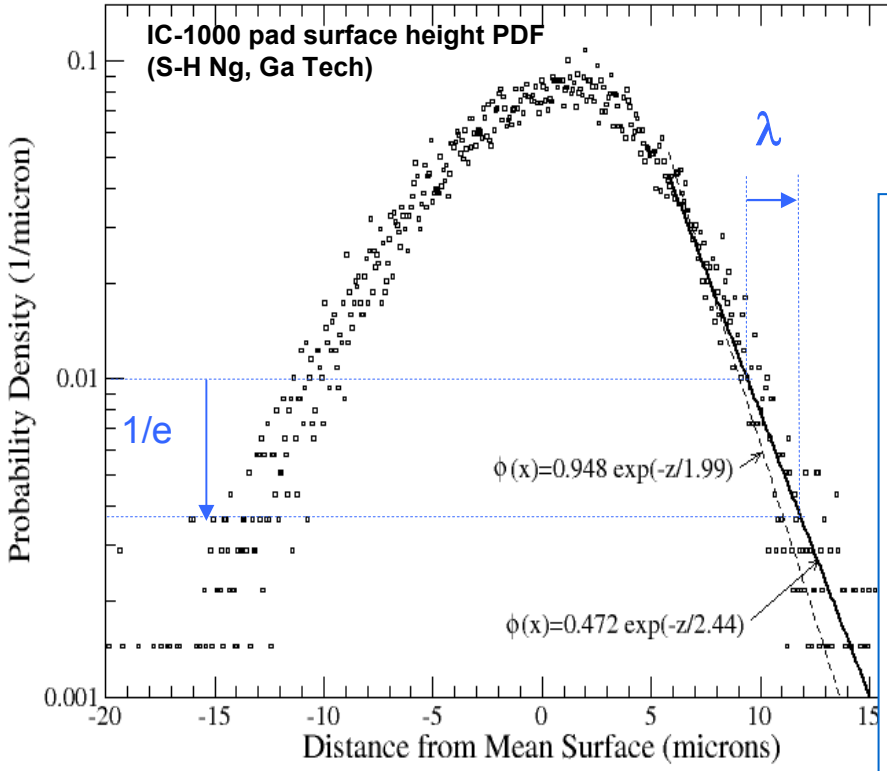
$$a = \frac{1}{2} + e$$

$$\beta = \frac{2\zeta(c_w, r_w)\gamma_p^1(p_a/p)}{\sqrt{\pi\kappa\rho C_p}}$$

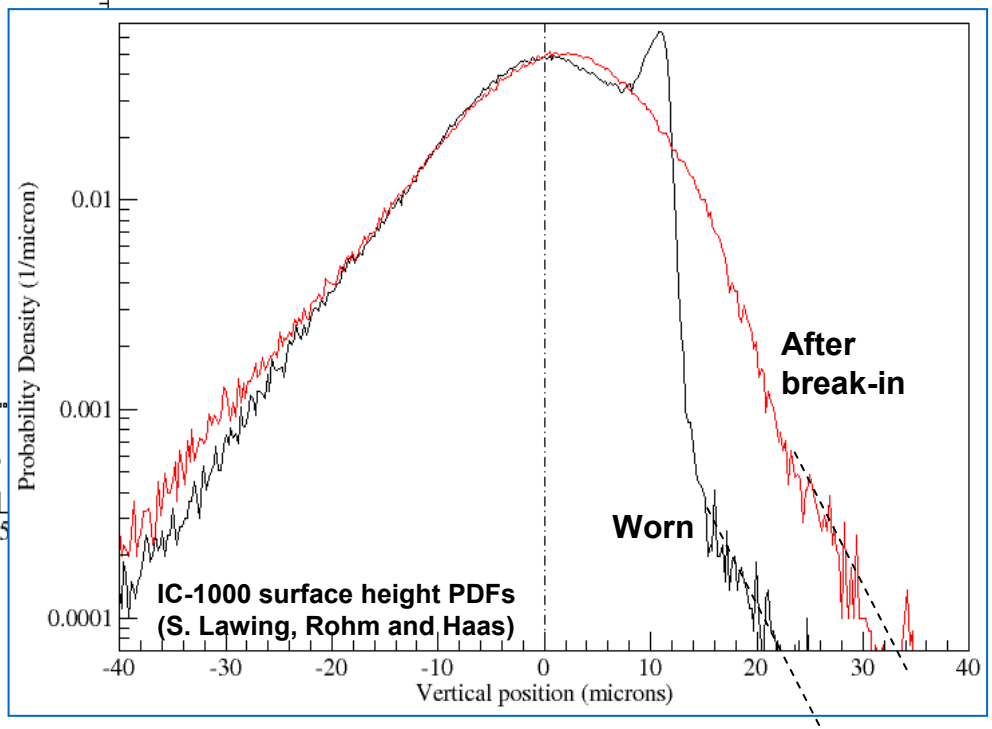
This explains why the power law form of the flash heating model works.

Connection with Pad Properties

The mean contact area depends on the pad summit height distribution. *Surface* height distributions often have an exponential tail, suggesting that *summit* height distributions may be similar. This is a convenient simplifying assumption.



Contacting summit heights can then be described with a single characteristic decay length λ.



Connection with Pad Properties

When the summit heights are exponentially distributed with characteristic decay length λ, it is possible to express the mean contact size s and corresponding scaled pressure p_a/p explicitly in terms of pad material and surface properties:

$$s^2 = \frac{p}{E^*}\frac{\pi^{1/2}}{\eta_s \lambda^{1/2}}\kappa_s^{-1/2}$$

$$\frac{p_a}{p} = \left(\frac{E^*}{p}\right)^{1/2}\frac{4}{3\pi^{5/4}\lambda^{1/4}\eta_s^{1/2}}\kappa_s^{3/4}$$

(for k-groove)

where

κ_s = Mean summit curvature

η_s = Summit area density

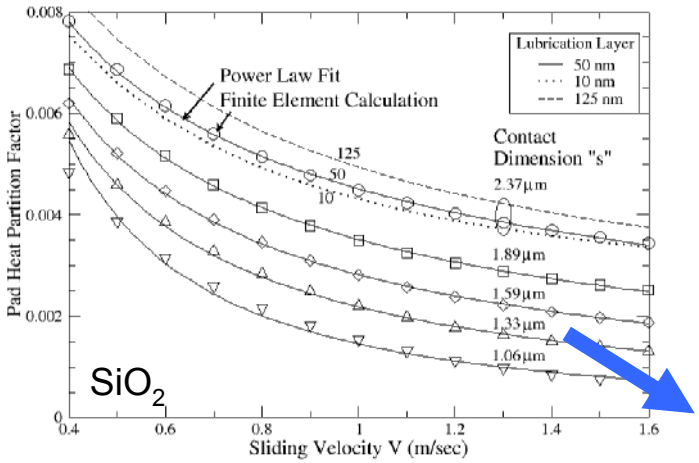
$E^* = \frac{E_Y}{1-\nu^2}$ = Effective modulus

$$\beta = \frac{2\zeta(c_w, r_w)\gamma_p^1(p_a/p)}{\sqrt{\pi\kappa\rho C_p}}$$

$$\overline{T} = \overline{T}_b + \frac{\beta}{V^{1/2+e}}\mu_k pV$$

Connection with Pad Properties

From the calculated pad heat partition factors, the exponent e and coefficient γ_p^1 can be expressed in terms of s.

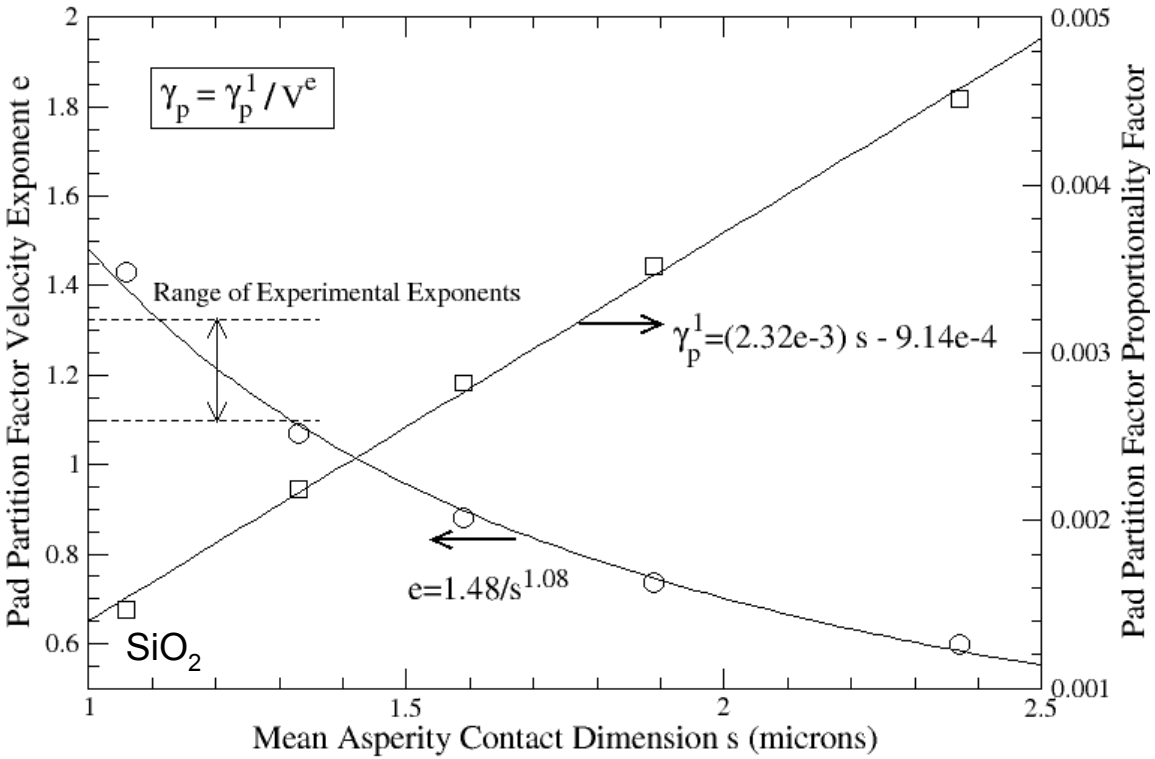


$$\gamma_p = \frac{\gamma_p^1}{V^e} \qquad \gamma_p^1 = 2.32x10^{-3}s - 9.14x10^{-4} \qquad e = \frac{1.48}{s^{1.08}}$$

s in microns

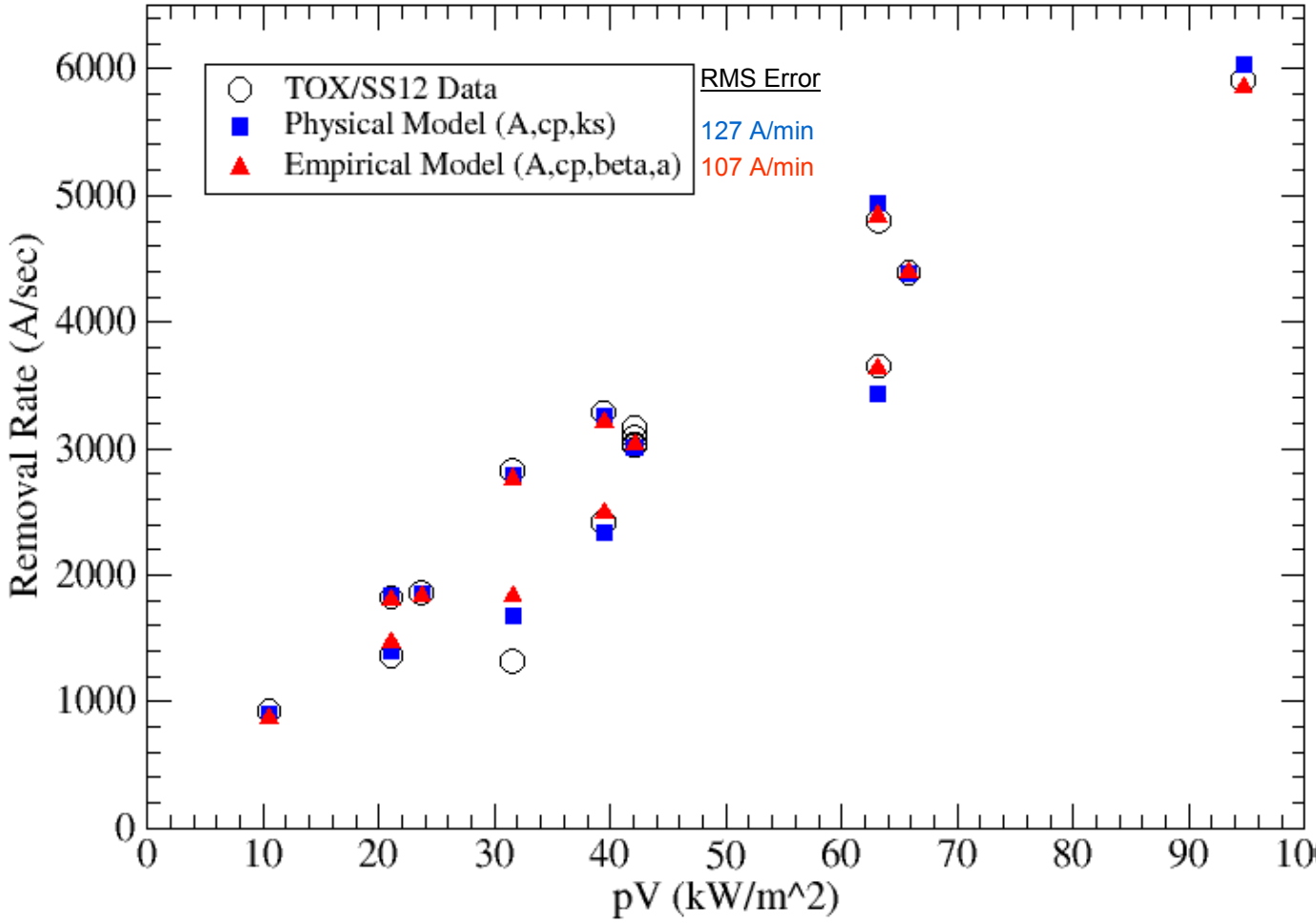
$$s^2 = \frac{p}{E^*}\frac{\pi^{1/2}}{\eta_s \lambda^{1/2}}\kappa_s^{-1/2}$$

Since s is related to load and to material and surface properties, all of the parameters in the flash heating model can be related to measurable quantities.



Comparison of Physical and Empirical Models: SiO_2

Stein and Hetherington 1999 Thermal Oxide Data



Physical Model

E = 0.53 eV

A = 1.87e5 moles/m^2-s

c_p = 5.47e-9 moles/J

E_Y = 285 MPa (Rohm&Haas)

ν = 0.5

η_s = 2e8 / m^2 (Shan)

λ = 2 μm (Ng)

μ_k = 0.4

κ_s = 6.13e5 /m

Empirical model

E = 0.53 eV

A = 7.99e4 moles/m^2-s

c_p = 5.31e-9 moles/J

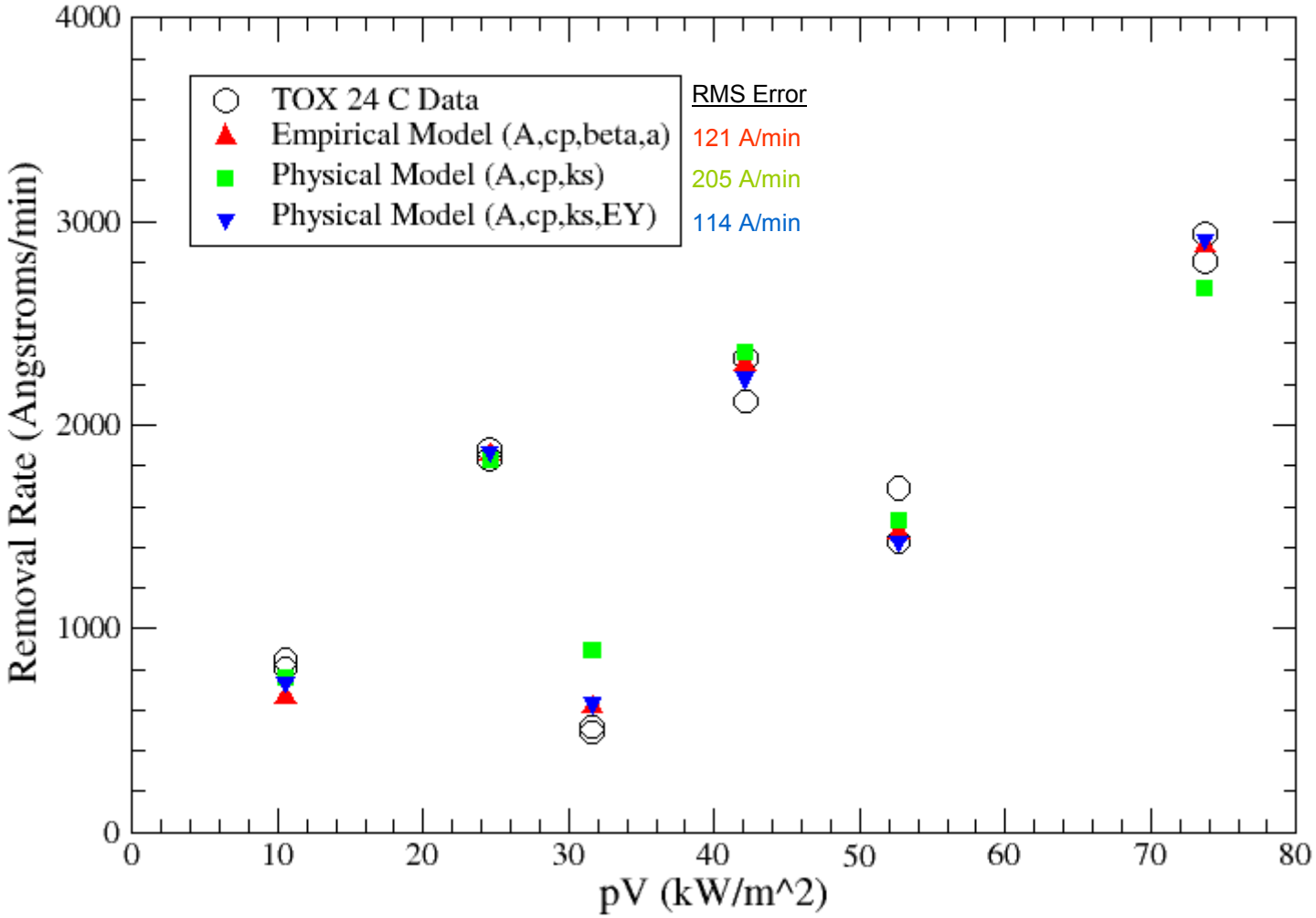
$\beta\mu_k$ = 1.05e-3

a = 1.82

Optimization on blue parameters.

Comparison of Physical and Empirical Models: SiO_2

Sorooshian 2003 Thermal Oxide Data



Empirical model
E = 0.53 eV
A = 7.97e3 moles/m^2-s
c_p = 4.73e-9 moles/J
$\beta\mu_k$ = 1.76e-3
a = 1.60

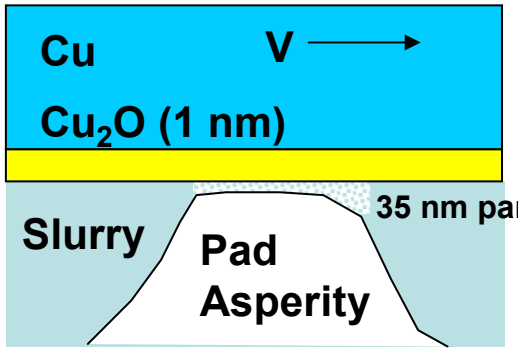
Physical (A,c_p,k_s)
A = 7.32e4 moles/m^2-s
c_p = 4.55e-9 moles/J
κ_s = 2.15e6 /m

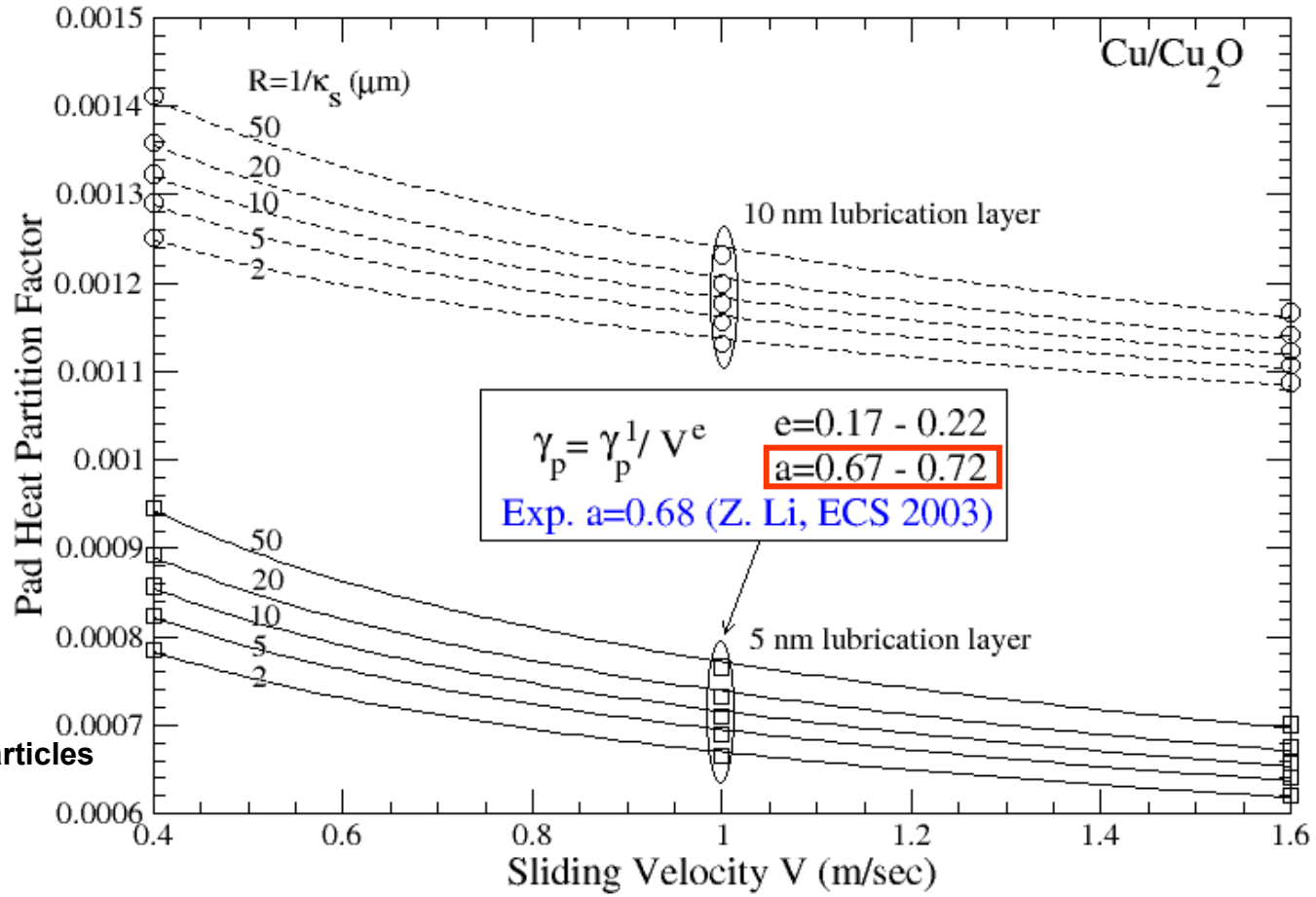
Physical (A,c_p,k_s,E_Y)
A = 2.41e4 moles/m^2-s
c_p = 4.85e-9 moles/J
κ_s = 1.71e6 /m
E_Y = 105 MPa

Other parameters for physical model same as for Stein data.

Heat Partitioning for Copper

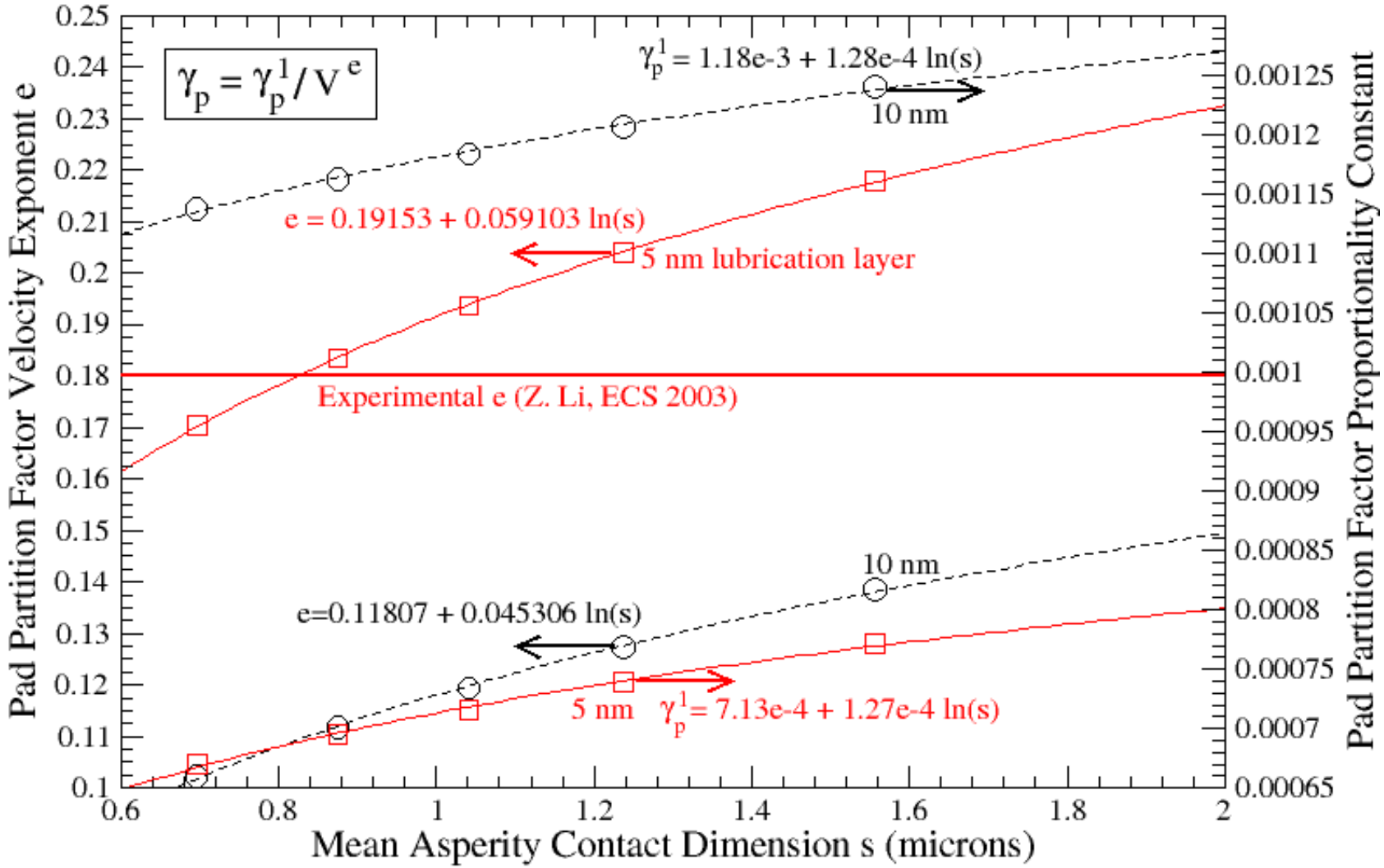
The pad heat partition factor for copper with a thin oxide layer also has a power law form. The exponent e is smaller than for SiO_2. The calculated value of a for a 5 nm lubrication layer is close to experiment.



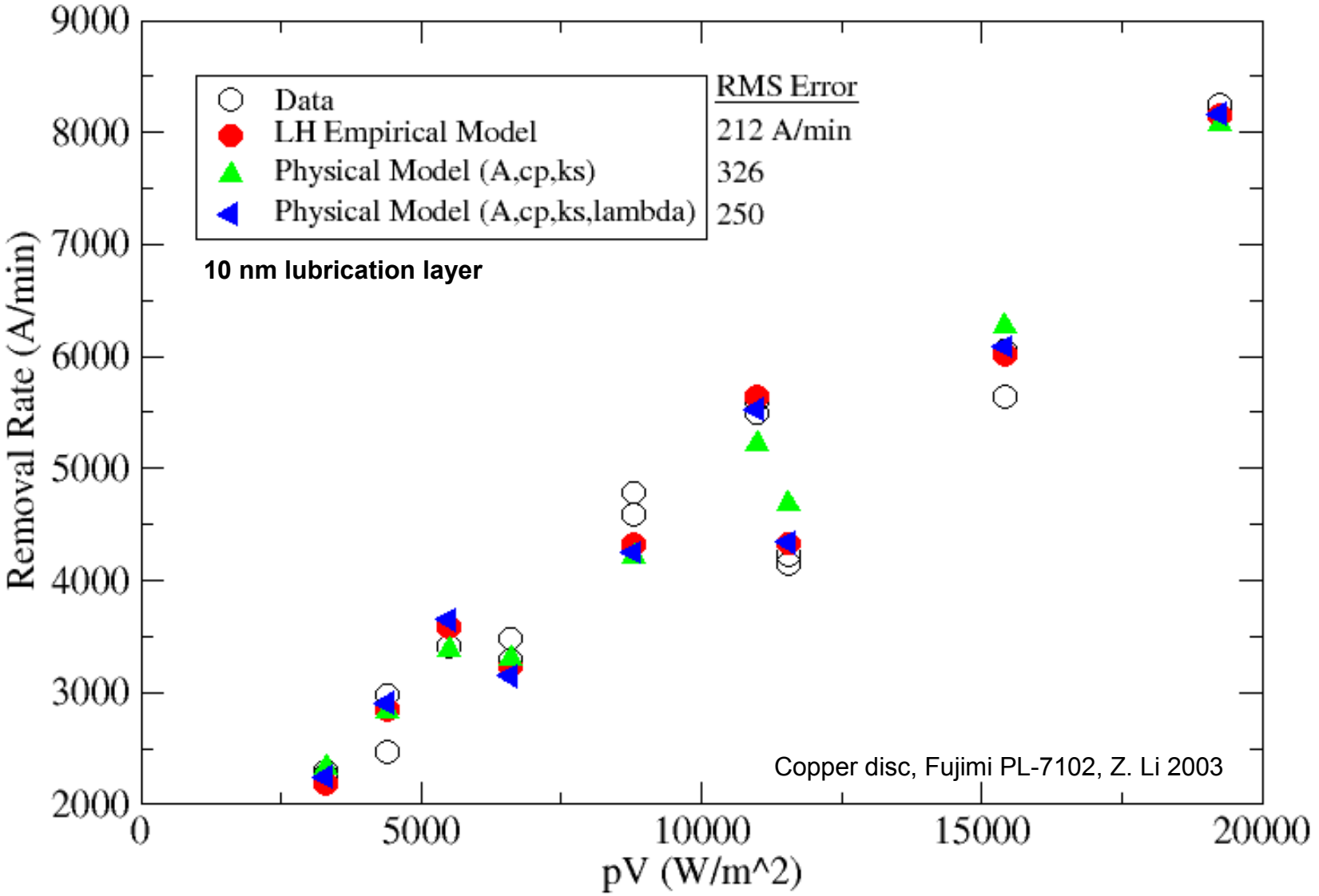


Heat Partitioning for Copper

For copper, *e* increases with contact size – the opposite of SiO_2.

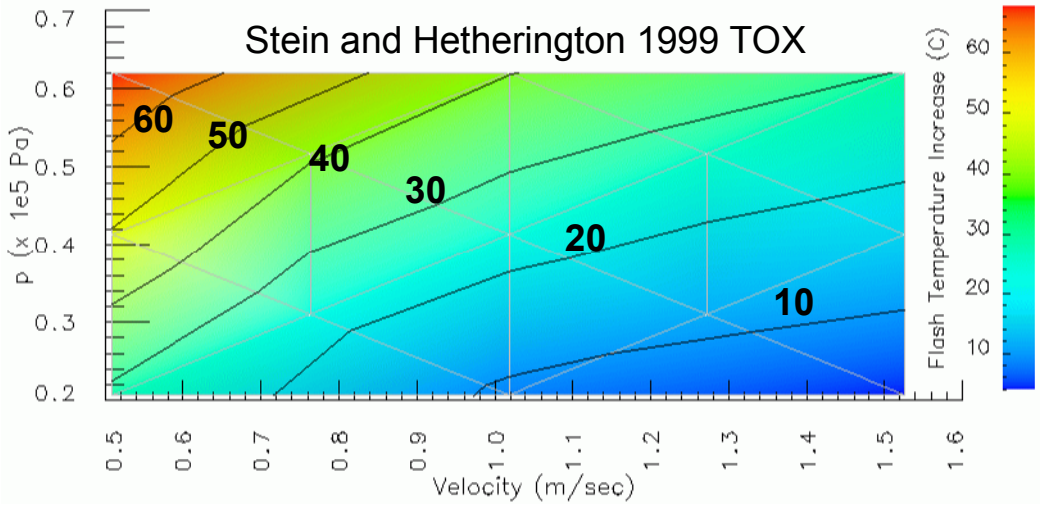


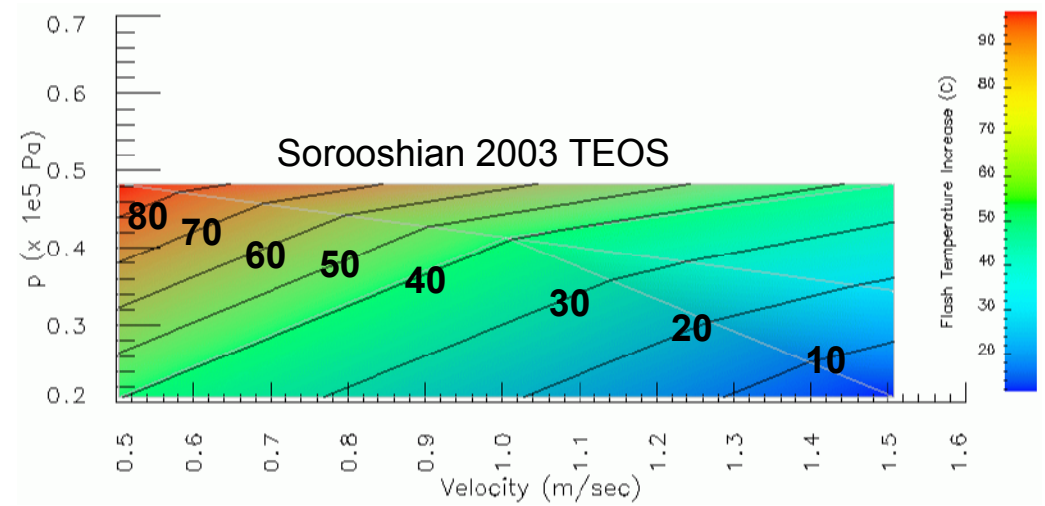
Comparison of Physical and Empirical Models: Cu



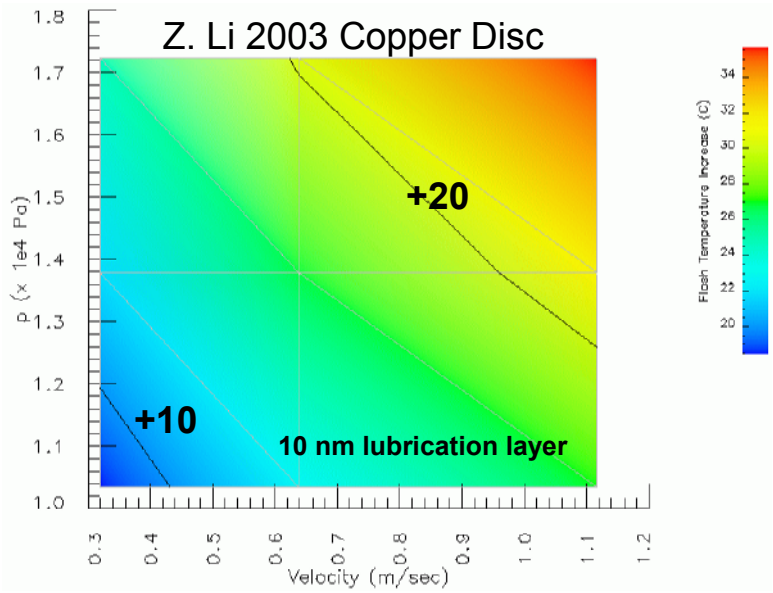
Calculated Flash Temperature Increase

From the physical model.





In the experiments examined here, the predicted flash temperature rise is smaller for copper than for SiO_2. The temperature variation also has a different functional dependence on p and V for the two materials.



Summary

The flash heating temperature rise has a power law form because the pad heat partition factor has this form.

Theoretically calculated values of β and a in the flash heating model are in good agreement with empirically extracted values.

In particular, theory predicts that a should be smaller for Cu than for SiO_2, in agreement with experiment.

β and a can be related to pad mechanical and topographical properties as well as to pad, wafer and slurry thermal properties.

The physical version of the flash heating model fits data about as well as the empirical version, sometimes using one less parameter. Extracted parameters have a testable physical interpretation.