Heatable cantilevers

The cantilevers used in this study have been described previously [1]. As shown in Fig. S1, the cantilever consists of four cantilever beams and two electrical heaters. In this work, only the heater directly behind the tip was used.

\[ V_{tot} \]

\[ I \text{(mA)} \]

\[ V_{lev} \]

\[ R_{el} \text{(k}\Omega) \]

\[ P \text{(mW)} \]

\[ T_{\text{max}} = 550 \text{ } ^\circ \text{C} \]

Temperature calibration

The heater temperature was calibrated using the measurement of the electrical resistance of the heater using the circuitry shown in Fig. S2a. During the experiments a constant bias, \( V_{tot} \), is applied. The electrical resistance is measured from the voltage drop across the heater. The resulting current–voltage diagram is shown in Fig. S2b. The electrical resistance changes as a function of the power dissipated in the heater due to the temperature dependence of the electrical resistivity of the silicon.

To determine the electrical resistance \( R_{el} \) of the heater as a function of temperature, we exploit the fact that the maximum electrical resistance occurs at a temperature determined by the doping level in the heater. For the cantilevers used here this occurs at 550 °C, see Fig. S2c. For other temperatures, we assume that the dissipated electrical power is proportional to the temperature rise in the heater. The temperature calculation using this method has been shown to have an uncertainty of less than 30% and is thereby comparable to the uncertainty of other methods used to measure temperatures in microscale structures. The calibration process and the underlying assumptions were described in detail by Menges et al. [2].

Quantification of the thermal resistance

Figure S3 explains the extraction of thermal resistances from electrical measurements in detail. During an experimental loading and unloading curve, we apply a constant bias \( V_{tot} \) and measure the voltage drop across the cantilever heater, \( V_{lev} \), as shown in Fig. S3a. From this, the electrical resistance is determined, and, using the temperature calibration described above, the heater temperature is calculated (Fig. S3b). From this, together with the electrical power dissipated in the heater (Fig. S3c), the total thermal resistance is extracted (Fig. S3d). Finally, the thermal resistance of the tip-surface contact is obtained after subtracting the contribution from the cantilever alone (Fig. S3e). The last step is explained using Fig. S4.
Figure S3: Measured and calculated signals of heated tip approaching towards and retracting from the taC surface. a: Measured voltage drop across cantilever (see Fig.S2a). During the measurement the voltage $V_{tot}$ was set to 3.046 V. b: The heater temperature $T_{heater}$ calculated from a, using the temperature calibration described above. c: The heater power $P$ calculated from A, using the temperature calibration described above. Using the circuitry shown in Fig. S2a, $P$ is approximately constant during the experiment. d: The total thermal resistance $R_{th\text{-tot}}$ calculated from b and c. The thermal resistance is given by the cantilever only, when the tip is out of contact, and the tip-surface contact in parallel with the cantilever when the tip is in contact with the surface. e: The thermal resistance of the tip-surface contact only, $R_{th\text{-contact}}$, as calculated from d.

Tip-polishing procedure

From continuum theory contact models, one expects large stress distributions for a flat punch pressed onto a flat surface. The stress peaks at the periphery of the contact. In contrast, a spherical tip pressed onto a flat surface has a pressure distribution with a maximum at the center. The polishing procedure used here has been discussed previously in detail [3–6]. It is argued that the polishing progresses in response to the local stress in the contact zone. This implies that a worn truncated cone has a terminating surface that is not perfectly plane. Instead the shape of the surface interpolates between the limiting cases (spherical tip versus flat punch) in such a way that the stress should be distributed almost equally across the contact zone. A similar observation has been made for macroscopic wear [7], where the dependence on the local wear progression as a function of local pressure should be even stronger. In either case, the deviation from a perfectly plane boundary needed to level out the stress is very small and not apparent from the visual inspection of the scanning electron micrographs. We note, that in light of the findings in this manuscript, the notion of local stress and wear progression needs to be studied in more detail.

Load forces of 5, 25 or 50 nN were applied during the tip-polishing. This matches roughly the averaging region used to determine the load-dependence of heat transport of 5-30 nN.

Application of a continuum-theory model of roughness

In the main text we argue that the increase of contact area with applied force is well described using an atomic scale roughness picture as proposed in atomistic simulations [8]. An important cross-check is to compare known contact models based on nominally flat rough surfaces. A particularly successful description on roughness on arbitrary length scale has been given by Persson [9]. In his work the two contacting bodies are treated using continuum theory of elasticity. An analytical equation for the dependence of ”real” contact area on applied load can be given for a material with roughness described with a single Hurst exponent.
We studied the roughness of both contacting surfaces, the taC surface and the worn silicon tip after transfer to an atomic force microscope in ambient. The roughness of the taC surface was measured using a commercial atomic force microscope operating under ambient (Bruker, Dimension V) with commercial cantilevers using tapping mode. Fig. S5a shows the surface topography which exhibits roughness on several length scales. From images like this, 2D-power spectral densities of the surface topography can be extracted, as shown in Fig. S5b. The taC surface could be measured down to ≈ 3 nm before finite tip size effects and electrical noise started to dominate. The spectrum can be well approximated with a single power law in the range between approximately 3 and 300 nm. From the slope of the curve (the exponent of the power law) we calculate a Hurst exponent[10] of $H = 0.84$. The rms-roughness of the surface depends on the length scale chosen. On a 24 nm-length scale (the average of the contact diameters of the tips used for heat transport) the rms roughness is 0.09 nm.

It is more challenging to measure the surface roughness of the blunted tips. We ruled out tip shape reconstruction from imaging of sharp surface objects (like an array of inverted tips) because the imaging process using contact mode is likely to lead to severe wear and will alter the surface of interest. Instead, we glued the cantilevers onto a substrate and used the more gentle tapping mode to image the tips using commercial AFM cantilevers. (Note, that the cantilevers used in this study are too compliant to use with tapping mode.) With this method, we can determine an upper limit of the rms roughness of about 0.6 nm. We also note that these measurements were performed sufficiently long after the thermal transport measurements for a full native oxide layer to form which may increase the roughness relative to the state during the transport study. For more quantitative estimates of the roughness we used transmission electron microscopy (TEM) images of tips blunted using the same blunting procedure. Peak-to-peak roughness values of 0.5, 0.7, 0.7, and 1.1 nm were obtained on four such tips. The latter value was limited by the image resolution. Note, that an expected remaining small curvature of the surface was not subtracted for the computation of these values. We note, that our measurements are consistent with other published TEM images of silicon and silicon nitride tips worn on ultra-nanocrystalline diamond surfaces[11]. Further, these data are consistent with the surface roughness observed on native silicon oxide covering silicon wafers, varying typically between 0.1 and 1.0 nm rms on a 1 micron scale depending on sample preparation[12, 13]. It is likely that the tip roughness exceeds that of taC which is only composed of a single atom species. In conclusion, we estimate an rms roughness of $h = 0.22 \pm 0.04$ nm for the tips and use this value in following calculation.

We are not able to fully characterize the spectral components of the roughness of the polished tip apices due to the limited resolution of our SEM and TEM images and due to the difficulty of imaging the full area of the flattened tips using AFM techniques. We note however that the diameter of flattened tips places an upper limit on the possible spectral components of the roughness. In addition, TEM images indicate that the tip polishing procedure appears to reduce the amplitude of the larger wavelength components of the roughness and as discussed above, we expect a very small residual curvature due to the leveling out of the stress distribution during polishing. In summary, we expect that the tips are nominally flat on the length scale of the tip diameter, meaning that the roughness spectrum is flat up to a certain smaller length scale below which the amplitudes of the spectral components decay with decreasing length scale, following a corresponding Hurst exponent. The Hurst exponent $H$ of silicon oxide surfaces is found in the literature to be significantly smaller than $H = 1$; reports suggest $H = 0.1$ may be more realistic for our situation[14–17].

To proceed, we also need the elastic properties of the materials, described by the Poisson ratio $\nu$, and the elastic modulus $E$. The modulus is estimated to be dominated by the more compliant of the two surfaces, i.e. the silicon oxide covered silicon tip. The modulus of SiO$_2$ is $E = 72$ GPa, and the Poisson ratio is taken as $\nu = 0.17$.  

---

**Figure S5:** Roughness of the taC surface. a Atomic force microscopy (AFM) image of surface topography of the tetraedral amorphous carbon (taC) sample used in this study. b Two dimensional power spectral density of surface height AFM images on taC. The solid line is a fit using a single exponent power law.
The final parameters needed to apply Persson’s model are the maximum and minimum length scales of the problem, \( L \) and \( \lambda \), respectively. \( L \) is given by the diameter of the contact (24 nm in our case), and \( \lambda \) defines the cutoff at lower spatial wavelength. We can then calculate the change of contact area with applied load from Eq. (6b) of [9], valid for \( L \gg \lambda \):

\[
\frac{\partial A}{\partial F} = \frac{2}{\pi} \frac{L}{h} \frac{1}{E} \left( \frac{1 - H}{1 + H} \right) \left( \frac{\lambda}{L} \right)^{1-H} \frac{1}{\nu^2} \left( 1 - \frac{1}{\nu^2} \right)
\]

As apparent from the equation, the result depends only weakly on \( \nu \).

For the model put forward in this paper we estimate \( \lambda \approx 1/\sqrt{\rho} \), where \( \rho = 3.8 \times 10^{18} \text{m}^{-2} \) is the surface atom density of silicon oxide (the less dense of the two facing surfaces) calculated from its mass density. Altogether we have \( L = 24 \text{ nm} \), \( \nu = 0.18 \), \( h = 0.22 \pm 0.04 \text{ nm} \), \( E = 72 \text{ GPa} \), \( H = 0.84 \), and \( \lambda = 0.51 \text{ nm} \). (We expect that our system actually has a lower \( L \) and a lower \( H \), because the silicon oxide terminated silicon tip will dominate the contact morphology. Nevertheless, we use the Hurst exponent of the taC and the full diameter of the tip for \( L \) to give a more conservative estimation of \( \partial A/\partial F \).

With this we obtain \( \frac{\partial A}{\partial F} = 1.2 \times 10^{-10} \text{ Pa}^{-1} \). Using \( \rho \) from above this corresponds to a force per contacting atom of: \( 1.8 \pm 0.9 \text{ nN} \) (error from uncertainty in parameters). The correspondence with the value extracted from atomistic models is striking \((0.5 \pm 0.15 \text{ nN})\). We note that Mo et al. found that Persson’s model is consistent with the pressure distribution function obtained in atomistic simulations. Here, we find, that the absolute value is also reasonably well reproduced, assuming a cutoff length scale equivalent to the atomic distance length scale.

To test the model predictions for roughness limited to larger length scales, we also compute the value for a cutoff \( \lambda = 1.5 \text{ nm} \), and obtain \( \partial A/\partial F = 1.4 \times 10^{-10} \text{ Pa}^{-1} \).

Combination with a continuum-theory model of thermal boundary conductance

We can combine the results of the application of Eq. 1 with thermal transport models not based on quantized conductance. For the sample system taC–Silicon experimental data exist [18] for the thermal boundary conductance per unit area: \( G_{\text{int}} \approx 1.6 \times 10^8 \text{ W/K/m}^2 \). Assuming this value is largely scale independent, we can calculate the load or pressure dependence of thermal boundary conductance:

\[
m = \frac{\partial G}{\partial F} = G_{\text{int}} \frac{\partial A}{\partial F}
\]

Using \( \frac{\partial A}{\partial F} = 1.7 \times 10^{-10} \text{ m}^2/\text{F} \) (calculated for roughness down to the atomic scale) we obtain \( m = 0.009 \text{ W/(KN)} \).

If, for comparison, we use a larger cut-off length scale and \( \partial A/\partial F = 1.5 \times 10^{-11} \text{ Pa}^{-1} \), we find \( m = 0.011 \text{ W/(KN)} \).

We note that both these values are several orders of magnitude below the experimentally determined value. We attribute this to the effect that, through transport quantization, the conductance per unit area stops decreasing with contact asperity size when the length scale reaches nanometric to atomic dimensions (at room temperature). Therefore, while continuum theories may be applied to understanding the interaction mechanics of rough surfaces, the thermal transport appears to be governed by finite size effects.

For completeness we calculate the pressure dependence of thermal boundary resistance assuming a roughness cut-off length of 1.5 nm in combination with quantized thermal transport. Up to \( \approx 1.5 \text{ nm} \), the value of the quantized conductance remains constant at one conductance channel. The conductance per channel of diameter of 1.5 nm \((A_0 = \pi(1.5 \text{ nm})^2/4)\) is therefore also \( G_0 = G_Q N\tau = 1.2 \times 10^{-9} \text{ W/K} \) (see main text). Then the pressure dependence is \( m = \frac{G_0}{A_0} \frac{\partial A}{\partial F} = 0.02 \text{ W/(KN)} \).

Limits of the quantized transport picture

From the discussion above and the widely applicable models of the mechanics of contacts of rough surfaces, it appears that the model of roughness-induced quantized thermal transport is applicable to many systems.

However, the model is valid only valid in the low nominal pressure regime. At high pressures, the notion of independent nanoscale contact points must eventually break down. The higher the pressure, the larger the ”real”contact area becomes, and at some point, for our case at around 500 MPa, one cannot talk about independent contact points any longer. According to Eq. 1, the amount of pressure necessary to reach this point also depends on the magnitude of roughness and weakly on the absolute length-scale involved.

Fig. S6 visualizes the limits of independent contact point formation. A simple atomistic calculation was used to predict the number of independent contact islands as a function of applied nominal pressure. We consider an array of 10'000 independent atoms that are, for simplicity of visualization, positioned on a hexagonal grid with the same density as expected for our amorphous interface (the surface atom density is limited by the less dense silicon oxide and estimated to be \( 3.8 \times 10^{18} \text{ m}^{-2} \).) In consecutive steps, an additional random atom position is chosen that makes contact until the entire array of 10000 sites forms a continuous contact. (This estimation corresponds to a flat roughness spectrum \((H \approx 0)\) and no long-range elasticity. This is an over simplification of the actual situation, but is still useful to visualize how the model breaks down at high pressure and to provide an upper bound for where this break down is expected.
theory[8, 9] has a similar value, namely 770 MPa.

Limits of roughness

One may ask also, if surfaces can be too smooth to be described by this picture. Let us consider the minimum corrugation needed to fulfill the requirements of this model. For this we evaluate the force-interaction range of facing atoms. A Lennard-Jones potential using typical magnitudes (characteristic distance $= 0.34 \text{ nm}$, binding energy $1.42 \text{ meV}$ from [12] with a force per atom of $0.5 \text{ nN}$) yields about $3 \text{ pm}$ spread from atom to atom to allow for a gradual accumulation of individual point contacts with increasing load. The average distribution in surface height of the most protruding atoms that form the contact must therefore be on average larger than $3 \text{ pm}$.

To estimate the roughness of our system we start with the smoother of our two surfaces: the taC sample measured on a length-scale of a few tens of nanometers (characteristic for the nominal contact area of the tips in this paper) of about $0.05 \text{ nm}$ (about $0.1 \text{ nm}$ on a 1 micron scale). TaC is known to be exceptionally smooth. The opposing surface, consisting of the silicon oxide covered tip, is certainly rougher. Typical roughness values for native oxides are in the range of $0.2 \text{ nm}$ on a 1 micron scale. Our measured value lies at $0.22 \text{ nm}$ with some uncertainty. Overall, we can safely assume an rms-distribution of distance between opposing atoms of at least $75 \text{ pm}$ as a lower bound. We note that the surface roughness used in the atomistic simulations by Mo and Slufarska [8] is of similar magnitude ($80 \text{ pm}$).

Next, we calculate the fraction of atoms in contact according to the simplification put forward by Mo and Slufarska [8]. The nominal pressures in our experiment are around $p_{exp} = 320 \text{ MPa}$. The pressure corresponding to all atoms in contact is $p_{max} = 0.5 \text{ nN}/(0.36 \text{ nm})^2$, using surface the density of the silicon atoms. The ratio of the two is the expected fraction of atoms in contact and amounts to $8\%$. If we further use the common approximation of a Gaussian surface height distribution (neglecting surface elasticity), we find that the mean position of the 10% contacting atoms is $130 \text{ pm}$ above the average of the remaining 90%. We conclude that in our system we are not near a situation of the two surfaces being two flat for quantized transport. In contrast, we find that it may be difficult to find surfaces that are too smooth other than single crystals. The same conclusion can be reached considering continuum theory based on Eq. 1.

The opposite case, a surface too rough to observe quantized transport, is not predicted by the models discussed here. If with increasing load existing contact sites grow in size rather than the formation of new contact sites, then the model has to be modified. In particular surfaces that are not described with a Hurst exponent, such
as artificially made surfaces, will behave differently. In addition, a surface with relatively small long-range elasticity in combination with a large Hurst exponent may reduce the validity of the model to smaller nominal pressures. Furthermore, an extremely rough surface may involve extreme values of load according to Eq. 1. In this case, plastic deformation may occur, leading to an increase of existing contacting islands instead of the formation of new ones. Apart from this, we also expect that the model breaks down for contacts for which the adhesion force is so large that even without applied load the limit of independent heat transport channels (discussed using Fig. S6) is reached.

Both limits, too much and too little roughness, are actually not only set by the roughness and the Hurst exponent alone. Instead the combination with the elastic properties has to be considered. In the limit of a stiff (rigid) material, only few contact points will form. In the other limit of very compliant material, "full" contact may already be reached in response to adhesive forces alone.

**Kinematic Effects**

We also note that in AFM based indentation experiments, kinematic effects result in a change in the angle of the cantilever and tip as the sample is displaced. This angle change can effect the pressure distribution within the contact zone and for large sample displacements can result in a peeling or the contact rather than an abrupt decohesion event. For the experiments described here, care has been taken to minimize these effects and we expect a negligible impact on the pressure distribution within the region of the indentation experiment used to analyze the thermal transport. For a typical experiment we calculate the magnitude of the angle variation to be 3.4 mrad resulting in a variation of the position of the edges of a 20 nm-wide tip apex of ±0.035 nm. This is 10 times smaller than the estimated peak-to-peak roughness at the contacting surfaces. The long-range elasticity of the contacting materials is expected to accommodate the resulting minor pressure and shear stress variations in the contact zone.