Nano Research

Open Access

Research Article

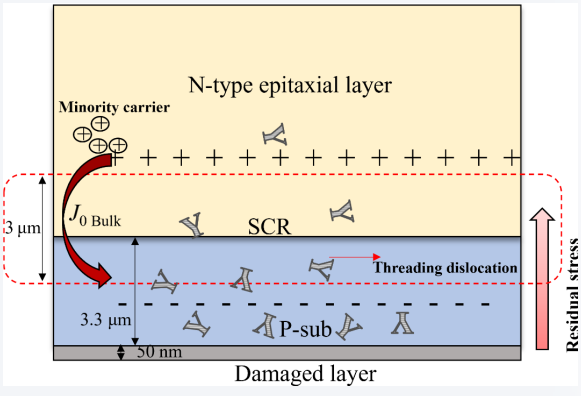
# Wafer-scale flexible silicon transistor: The role of thinning induced stress and defects on device performance

Neng Gao[1], Zhaoxian Wang[4], Xin Song[4], Dongliang Liu[4], Ying Chang[6], Wei Jian[2,3], Taisong Pan[1], Wei Qiu[5], Ying Chen[4] ✉, Yuan Lin[1] ✉, and Xue Feng[2,3] ✉

[1] *School of Materials & Energy, University of Electronic Science & Technology of China, Chengdu 610054, China*
[2] *State Key Laboratory of Flexible Electronics Technology, Tsinghua University, Beijing 100084, China*
[3] *AML, Department of Engineering Mechanics, Tsinghua University, Beijing 100084, China*
[4] *Institute of Flexible Electronics Technology of Tsinghua University Zhejiang, Jiaxing 314000, China*
[5] *Tianjin Key Laboratory of Modern Engineering Mechanics, Department of Mechanics, Tianjin University, Tianjin 300072, China*
[6] *School of Mechanical Engineering, Tianjin University of Commerce, Tianjin 300134, China*

**Cite this article:** *Nano Research*, **2026**, *19*, 94908642. https://doi.org/10.26599/NR.2026.94908642

**ABSTRACT:** Ultra-thin and flexible silicon chips are pivotal for high-performance conformal electronics. Backside grinding reduces chip thickness below 50 μm, enabling mass production of ultra-thin flexible chips in a low-cost way, yet the resulting defects and residual stress on the backside may propagate to the frontside circuits degrading the electric performance, especially for the critical-thickness devices. Here, commercial pulse-width modulation (PWM) chip wafers (involving tens of thousands of dies) based on bipolar transistors have been thinned to the thickness of ~ 20 μm by mechanical grinding with different parameters. Experimental results reveal a thickness-dependent bifurcated failure mechanism: Short-circuit current decays progressively with thickness reduction, while leakage current exhibits a catastrophic surge below the critical thickness (~ 18 μm). This work reveals a dual degradation mechanism in ultrathin ICs: Mechanical grinding not only amplifies substrate parasitic coupling via geometric thinning but also generates stress fields that induces dislocation rearrangement-aggregation cascades, ultimately dictating electrical failure modes. Chemical mechanical polishing (CMP) and reactive ion etching (RIE) have been deployed to inhibit leakage current surges by removing grinding-induced damaged layers and relieving interfacial residual stress, which collectively validate the stress-defect interaction as the governing mechanism of electrical failure in ultra-thin chips (UTCs). Hopefully, this study can throw light on the impact of mechanical grinding thinning on the electrical performance of ultra-thin chips paving the way to the wide applications of high-performance flexible electronics in the future.




**KEYWORDS:** ultra-thin chips, flexible silicon transistor, pulse-width modulation (PWM), leakage current, residual stress, mechanical grinding thinning

## 1 Introduction

In recent years, with the rapid development of the Internet of Things, artificial intelligence, and virtual reality technologies [1, 2], traditional rigid electronic devices have struggled to meet the demands for seamless integration and imperceptible wearability. As a result, flexible electronic devices have attracted considerable attention due to their deformable and reconfigurable properties, specifically in the fields like smart sensing, energy, and healthcare [3, 4]. These devices are gradually forming the foundation for emerging applications such as wearable systems, mobile healthcare, and smart cities [5–11]. Achieving system-level flexibility and fully integrated function necessitates the development of flexible integrated circuits, also known as the ultra-thin chips (UTCs). The UTCs exhibit a dramatically reduced stiffness compared to a traditional thick chip, and its superior form factor confers an advantage for integration on flexible substrates.

**Received:** January 15, 2026; **Revised:** March 11, 2026
**Accepted:** March 16, 2026

✉Address correspondence to Ying Chen, chenying@ifet-tsinghua.org; Yuan Lin, linyuan@uestc.edu.cn; Xue Feng, fengxue@tsinghua.edu.cn

As a viable fabrication route, backside mechanical grinding employs precision grinding wheels to thin the chips down to 5–50 μm producing the UTCs [12], whose rotating wheels are embedded with abrasive micro-particles, such as diamond. The backside of the wafer is pressed against this rotating abrasive surface to remove the silicon substrate material, whereby the wafer must withstand the pressure of the grinding wheel. The surface substrate undergoes the physical damage, fracture, and removal, resulting in a certain thickness of damage layer on the processed surface. This damage layer inevitably introduces residual stress, dislocation structures, interface layers, and other additional defects [13]. Previous studies have preliminary revealed the impact of the defects on the mechanical degradation of ultra-thin chips induced by grinding thinning processes. Yoriko Mizushima's study reports that coarse grinding (#320 grit grinding) and fine grinding (#2000 grit grinding) generate damage layers with thicknesses of less than 5 μm and 200 nm, respectively. These layers include amorphous, polysilicon, and crystalline defects such as stacking faults and dislocations [14]. Zhang et al. have shown that the subsurface damage (SSD) layer in silicon after grinding exhibits a degradation in mechanical properties such as the Young's modulus, ultimate tensile strength (UTS), and fracture strain compared to the ideal silicon [15]. Jian et al. theoretically and experimentally reveal that the SSD depth induced by different grinding parameters determines the bending strength of the UTC [16].

As chip thickness decreases, the stress-defect coupling caused by the backside damage layers may profoundly influence the electrical characteristics. Residual stress generated during grinding modifies the carrier transport pathways through band structure modulation [17], while dislocation-related defect clusters potentially exacerbate charge leakage by forming trapping centers [18–20]. Such defect-related mechanisms have been preliminarily explored through strain engineering with deliberate introduction of controlled stress fields. Yet the residual stress introduced during the backside grinding process typically exhibits a non-uniform distribution, with higher stress concentrations near the surface and a gradual decrease in magnitude with increasing depth into the silicon substrate. This stress is primarily compressive in nature, especially in the near-surface region, due to the plastic deformation and lattice distortion caused by mechanical abrasion and surface damage. Besides, the damaged layer also contains localized microstructural defects. To address the coupled effects of high near-surface stress gradients and localized defect clustering in UTCs, the existing stress-defect interaction models based on uniform distribution assumptions are inadequate. To overcome this limitation, large-scale wafer data analysis is required to reveal the asymmetric clustering of stress and defects in spatial domains and localized extreme variations in defect density. By leveraging statistical insights, customized process can be developed to suppress harmful coupling effects, thereby enhancing the electromechanical reliability of UTCs.

This study systematically investigates the performance degradation mechanisms of wafer-scale flexible silicon transistors (total thickness < 50 μm) induced by mechanical back-grinding processes. Wafer-scale chip probing (CP) testing reveals two typical electrical failure behavior: a gradual attenuation of short-circuit output current with decreasing thicknesses, and abrupt leakage current fluctuations below a critical thickness threshold. Through characterization combining transmission electron microscopy (TEM) and Raman spectroscopy, the synergistic effects of residual stress distribution and dislocation density gradients have been systematically investigated. A qualitative correlation model has been proposed linking circuit-level anomalies to atomic-scale lattice distortions, encompassing parasitic substrate coupling induced by thickness loss and stress-driven dislocation rearrangement-triggered leakage current anomalies. Process optimizations demonstrated that chemical mechanical polishing (CMP)/reactive ion etching (RIE) effectively suppressed leakage current fluctuations, thereby enhancing manufacturing yield for sub-20 μm ultra-thin chips. This work provides insights into the electrical failure mechanisms of the UTC, paving the way for the development and wide application of the high performance flexible electronics in the coming future.

## 2 Results

### 2.1 Chip thinning and characterization

As illustrated in Fig. 1(a), the backside mechanical grinding is systematically applied to pristine 525 μm-thick silicon wafers with the size of 6 inch through a sequential thinning process achieving the target thicknesses of 50, 35, 30, 25, and 18 μm. Wafer thinning process is typically subdivided into coarse grinding and fine grinding. Coarse grinding, defined as the primary material removal stage employing abrasive wheels, achieves rapid substrate thinning by prioritizing efficiency over surface integrity. Characterized by high-efficiency material removal rates, this process removes ~ 80% of the target thickness using 300-mesh diamond grinding wheels operating at wheel speed of 3500–4000 rpm and feed rate of 5–10 μm/s. Fine grinding is designed to mitigate residual stress distribution and optimize surface integrity post-coarse grinding. Employing ultra-fine abrasives (2000–8000 mesh) at wheel speed of 3000–3500 rpm and feed rate of 0.5–1 μm/s, this stage reduces brittle fracture propagation while achieving thickness uniformity. After grinding, further processes such as CMP, wet etching, or dry etching are selectively used to remove the damaged layer. CMP, in particular, smooths the silicon wafer using chemical corrosion and mechanical force. During the polishing process, a polishing slurry is added, which chemically reacts with the wafer surface material to soften it. The softened material is then removed through micro-mechanical friction from the abrasive particles on the polishing pad. After the thinned chips are diced and packaged, they are divided into four groups: (i) 20 μm via 3000# grinding; (ii) 18 μm via 8000# grinding; (iii) 18 μm via 8000# grinding + CMP; (iv) 18 μm via 8000# grinding + RIE.

Figure 1(b) shows the cross-section of the thinned PWM chip as observed by the scanning electron microscope (SEM). The image reveals a 17 μm-thick active layer of the vertical NPN transistor comprising three functionally engineered layers: a 3 μm metal layer serving as the electrical interconnects, an 8 μm n-type epitaxial layer integrating heavily doped collector regions and carrier transport junctions, and a 6 μm buried layer enabling parasitic current suppression through doping gradient optimization.

The morphology of the back surface of the chip is further characterized by TEM, as shown in Fig. 1(c). At $H$ = 0 μm (where $H$ denotes the depth measured from the back surface), TEM characterization reveals the presence of damage layers with varying thicknesses. TEM characterization reveals the presence of damage layers with varying thicknesses. The thickness of the damage layer for the 3000# wheel is about 200 nm, while the thickness of the damage layer for the 8000# wheel is approximately 50–120 nm. Notably, the damage layer observed in the CMP- and RIE-treated

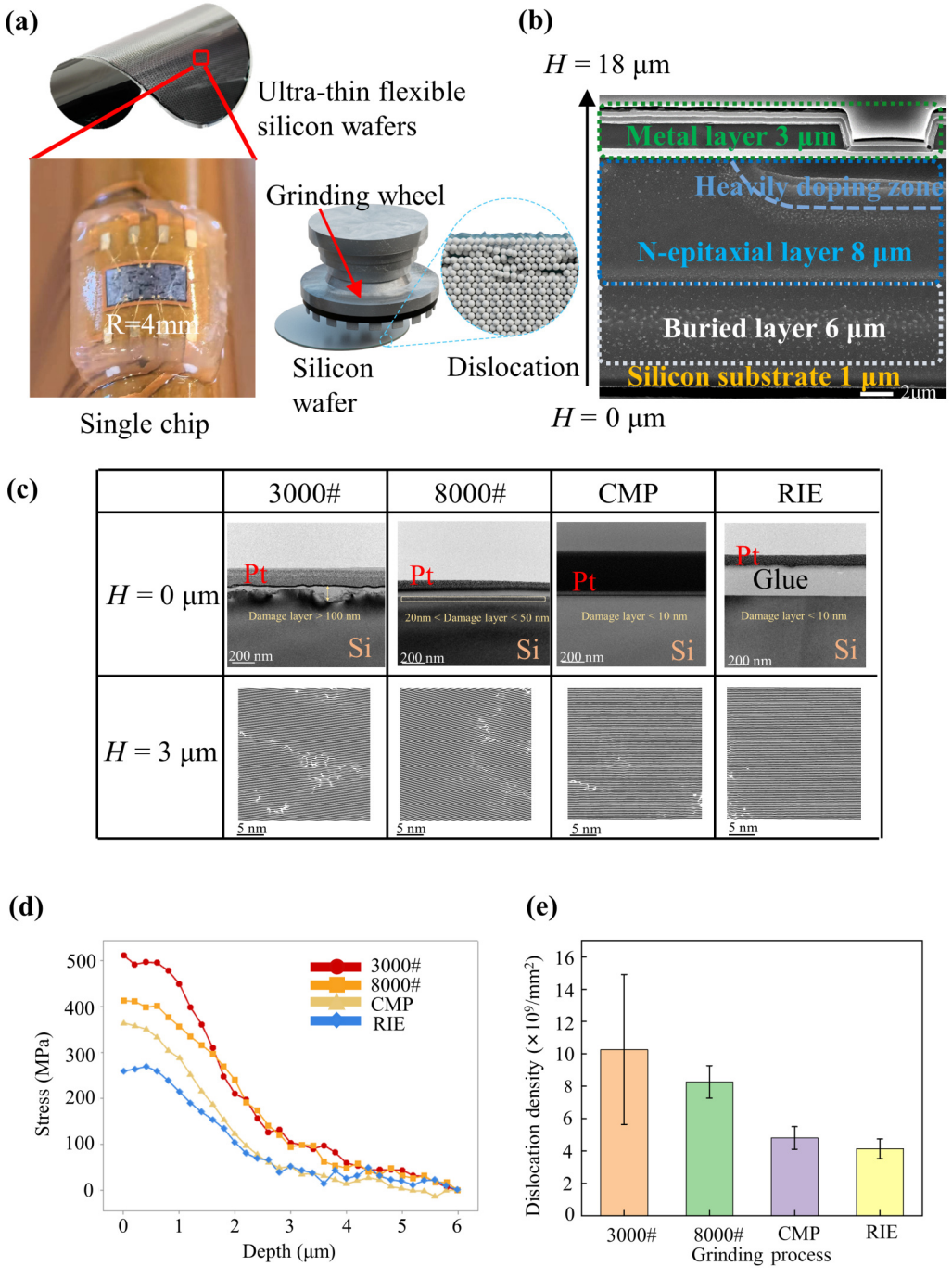



**Figure 1** (a) Grinding, packaging and bending of ultra-thin chips. (b) A cross-sectional SEM image of ultra-thin bipolar PWM chip. (c) TEM images of the bottom of the chip with different grinding processes. (d) Comparison of residual stress distribution along depth direction under different grinding process. (e) Dislocation density statistics for different grinding process at 3 μm from the back surface of the chip.

chips is found to be less than 10 nm. To systematically evaluate the residual stress distribution induced by backside damage layers and its depth-dependent propagation characteristics, the Micro Raman spectroscopy is adopted. Under the illumination of a laser with a specific wavelength, the position of the Raman peak is sensitive to changes in stress/strain, thereby allowing for the precise localization of stress variations within microregions. Accordingly, three independent micro-regions were defined on each sample's backside, each arranged in a 3 × 3 array of detection sites with an individual area of 2 μm × 2 μm, forming a dense network of 27 measurement points to provide high-resolution data for modeling depth-dependent stress propagation. Dual-wavelength Raman spectroscopy is employed with depth-optimized configurations: 532 nm excitation for shallow subsurface screening (0–2 μm) and 633 nm excitation for full-depth profiling (0–6 μm), both acquired at 0.2 μm axial intervals. Data processing involved two-stage filtering. Firstly, the shallow-layer (0–2 μm) spectral anomalies are identified as quality control benchmarks to reject outliers in the full-

depth (0–6 μm) dataset through cross-wavelength consistency checks. Intensity-threshold filtering is applied prior to weighted averaging of valid spectra per sampling site. Residual stress values are ultimately derived using material-specific Raman shift-stress factors correlating Raman shifts to lattice strain, as shown in Fig. 1(d). Specifically, for equi-biaxial stress in single-crystal silicon, the conversion is $\sigma = -435(\omega - \omega_0)$ (in MPa, with $\omega$ in $cm^{-1}$ and $\omega_0 = 520\ cm^{-1}$) [21]. The derived stresses are primarily compressive, with their distribution exhibiting a pronounced correlation to both grinding parameters and post-thinning treatments. For the 3000# wheel-processed sample (red curve), the stress peaks at ~ 450 MPa at 0.2 μm depth, consistent with its thicker damage layer (~ 200 nm), followed by an exponential decay to < 50 MPa beyond 2 μm depth. In contrast, the 8000# wheel (orange curve) demonstrates a lower stress of ~ 300 MPa at the surface, which rapidly decreases to ~ 80 MPa within 1.5 μm distance. The CMP-treated and RIE-treated samples (blue and green curves, respectively) display near-surface stress levels below 100 MPa, corroborating their ultrathin damage layers (< 10 nm). Specifically, CMP achieves a uniform stress gradient, decreasing from ~ 90 MPa at the surface to < 20 MPa at 3 μm depth. RIE treatment further suppresses residual stress, maintaining values below 50 MPa across the entire 0–6 μm depth range, likely due to anisotropic etching that eliminates lattice distortions introduced during mechanical grinding.

To quantitatively evaluate the correlation between residual stress distribution and crystalline defects near the back surface, cross-sectional specimens are prepared via focused ion beam (FIB) milling at $H$ = 3 μm depth among four group samples. TEM characterization was conducted using a 300 kV electron beam with probe-corrected imaging mode to resolve atomic-scale lattice distortions. Acquired micrographs undergo spatial frequency decomposition via two-dimensional Fourier transform (2D-FT), isolating periodic defect signatures from background noise. Inverse Fourier transform (IFT) reconstruction with frequency-domain filtering generated dislocation density maps, where grayscale intensity gradients quantitatively correlated with Burgers vector distributions. Post-processing statistical analysis quantifies the dislocation line densities within the 3 μm depth region, as visualized in Fig. 1(e). Mechanical grinding with coarser wheels (3000#) induced the highest dislocation density (attributed to plastically deformed zones from abrasive fracture), while finer grinding (8000#) reduced defect populations through partial stress relaxation. Advanced treatments demonstrated defect suppression mechanisms: CMP achieved stress homogenization via synergetic chemical etching and mechanical planarization, whereas RIE preferentially removed strained lattice regions through directionally selective etching. This phenomenon corresponds to the magnitude of the residual stress at 3 μm for each group in Fig. 1(d), indicating that there is an obvious correlation between residual stress and dislocation density.

## 2.2 Electrical performance testing

CP testing, a pre-packaging screening method, enables full-array electrical characterization across tens of thousands of dies, providing statistical significance for defect identification and yield evaluation. To maintain flatness and suppress mechanical distortion-induced failures during electrical testing of ultra-thin wafers, temporary bonding to rigid carriers is essential for minimizing measurement errors. The ultra-thin wafer with framed protective film is aligned and bonded to the cleaned temporary glass carrier by an ultraviolet (UV)-release tape in vacuum for 20 s at 80–100 °C, followed by constant-pressure curing for 60 s. The bonding process is further evaluated by measuring the total thickness variation (TTV) using a three-dimensional (3D) Line Confocal Scanner. It turns out that the TTV values are below 40 μm, as shown in Fig. S1 in the Electronic Supplementary Material (ESM), which is smaller than the probe overdrive (OD) range (50–150 μm) in CP testing, ensuring stable electrical contact during full-array probing.

During CP testing, each die is evaluated against over 40 electrical parameters ranging from functional verification to operational stability criteria. Dies passing all tests are mapped in green, failure dies in red, and untested edge dies for process boundary exclusion in yellow. As shown in Fig. 2(a), statistical analysis of chip probing results for thinned wafers (18–50 μm) indicates that the yield degradation caused by thinning is relatively minor. However, a notable exception occurs in the 8000# +18 μm samples, where the post-thinning yield drops drastically. To investigate this anomaly, failure rates of key electrical parameters are analyzed across different thicknesses and thinning processes (Fig. 2(b)). The data reveals that the yield collapse in the 8000# +18 μm sample is primarily attributed to catastrophic failure of the start-up current ($I_{st}$), with a failure rate approaching 90%. In contrast, parameters such as overflow status, voltage reference, discharge current, frequency of oscillator, and voltage feedback remain largely unaffected by thinning, exhibiting failure rates below 10%. This stark contrast highlights the unique sensitivity of $I_{st}$ to extreme

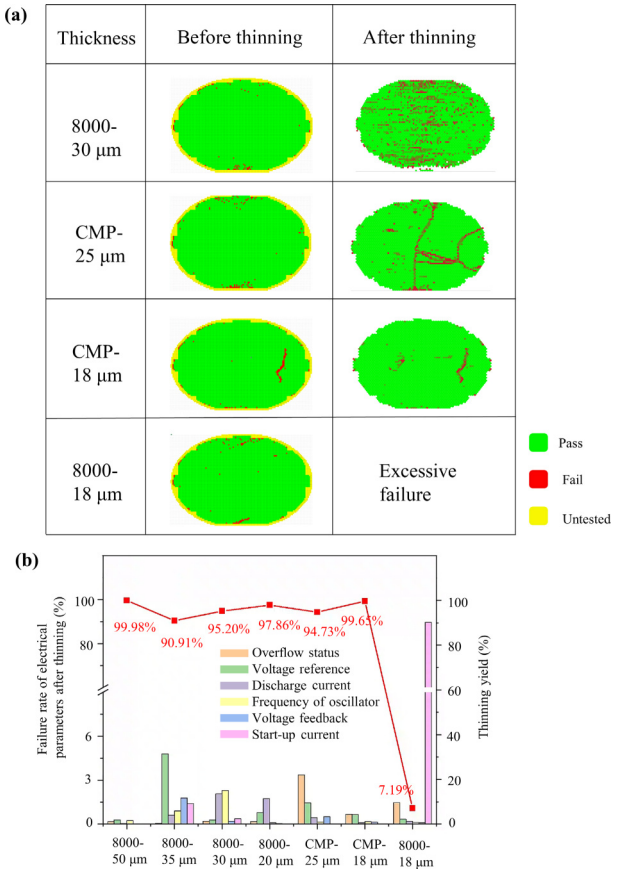



**Figure 2** (a) Ultra-thin chip yield statistics before and after thinning. (b) Statistical analysis of post-thinning electrical parameter failures (left $Y$-axis) and chip thinning yield (right $Y$-axis).

thinning (18 μm), while other parameters demonstrate robust stability under the same process conditions.

Further analysis of failure probability distribution and startup current characteristics (Figs. 3(a) and 3(b)) let us notice that the leakage current bursts in the start-up current ($I_{st}$) strictly depend on two concurrent conditions: ultrathin geometry (≤ 18 μm) and absence of post-thinning CMP. As shown in Fig. 3(a), the cumulative probability curves of the start-up current ($I_{st}$) for wafers ≥ 25 μm (30, 35, 50, and 525 μm) exhibit statistical consistency in large-scale datasets, with over 99% of $I_{st}$ values concentrated below 50 μA. In stark contrast, the 18 μm wafer shows a sharp rightward shift in $I_{st}$ distribution, where 96% of devices above 200 μA, indicating a dominant leakage failure mode. Furthermore, to examine the potential influence of wafer edge effects, we performed additional statistical analysis by selecting approximately 100 chips each from the central region, the top edge, and the left edge of the thinned wafer. The specific locations of these sampled chips are illustrated in Fig. S2(a) in the ESM, which shows the spatial distribution of the selected points across the wafer. As shown in Fig. S2(b) in the ESM, the cumulative probability distributions of $I_{st}$ from these different regions exhibit no significant differences and closely overlap with the overall distribution, indicating that the chip location on the thinned wafer has negligible correlation with the start-up current behavior. Critically, results in Fig. 3(b) reveal that

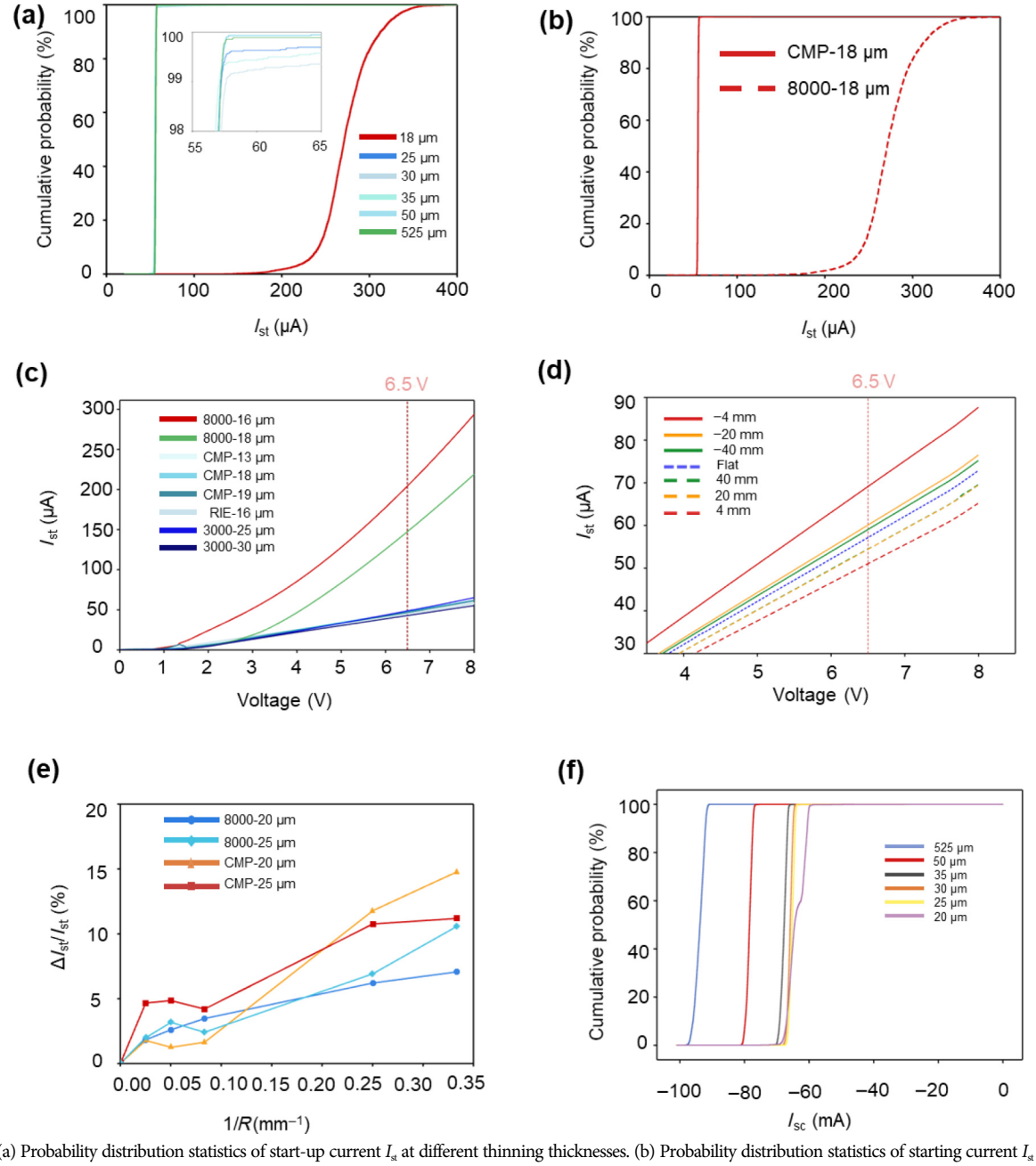



**Figure 3** (a) Probability distribution statistics of start-up current $I_{st}$ at different thinning thicknesses. (b) Probability distribution statistics of starting current $I_{st}$ at different grinding processes. (c) The start-up current $I_{st}$ plotted as a function of voltage for different grinding processes and different thickness (6.5 V as reference voltage for CP tests). (d) The start-up current $I_{st}$ of the 25 μm-CMP group plotted as a function of voltage for different bending radius. (e) The start-up current $I_{st}$ rate of change plotted as a function of curvature for different thinning thicknesses and grinding processes. (f) Probability distribution statistics of short-circuit output current $I_{sc}$ with chip thinning thicknesses.

this anomaly is fully suppressed by CMP: the CMP-treated 18 μm wafer ("CMP-18 μm") regains a cumulative probability of > 99% at $I_{st} \leq 50$ μA, aligning with the stable behavior of thicker wafers. Furthermore, the short-circuit output current, $I_{sc}$, exhibits a notable degree of variability during the CP test, which is represented in Fig. 3(f). Detailed statistical data, including mean values, medians, standard deviations, and value ranges, is summarized in Table S1 in the ESM. The data demonstrates that the short-circuit output current declines in proportion to the reduction in wafer thickness. The short-circuit output current of the 20 μm wafer is observed to decrease by approximately 30% in comparison to the unthinned wafer.

To ascertain the impact of stress and leakage on the ultrathin devices, the 4200 Semiconductor Analyzer is employed for systematic electrical characterization of the packaged thin chips. Figure 3(c) shows the correlation between the chip startup current ($I_{st}$) and voltage under varying processing conditions. The results indicate that an increase in leakage occurs when the chip thickness is less than 18 μm, exhibiting a three- to five-time increment compared to the baseline. However, the leakage surge dissipates following the CMP process. In order to further refine the effect of stress on leakage, we proceeded to perform bending tests on ultra-thin chips where leakage did not surge. Figure 3(d) shows the correlation between the chip startup current ($I_{st}$) and voltage under different bending conditions for a 25 μm-thick chip processed by CMP. And the negative bending radius implies reverse bending. Figure 3(e) illustrates the correlation between startup current $I_{st}$ and the bending curvature for distinct processing conditions at a voltage of 6.5 V. The findings indicate that the alteration in leakage resulting from bending stress is not substantial, and the variation in leakage of the chip does not exceed 15% when the maximum bending curvature that the chip can withstand is reached. This suggests that the leakage surge in ultra-thin chips with a thickness of less than 18 μm may be the result of a combination of factors rather than stress alone.

### 2.3 Mechanism analyses

The first effect of chip thinning is thickness loss, which means an increase in substrate resistance and substrate voltage, leading to the formation of parasitic transistors, and an increase in the width of the depletion region, which further reduces the width of the base region or effective channel length, affecting the current–voltage characteristics of the chip.

Figure 4(a) shows the basic structure of a typical NPN transistor. The substrate resistance $R_S = R_1 + R_2$ through which the substrate current $I_s$ flows, $R_1$ represents the transverse resistance passing through the substrate, and $R_2$ represents the longitudinal resistance under the substrate contact. The relative size of $R_1$ and $R_2$ depends on the doping, epitaxy, and isolation processes, $R_1$ is usually much bigger than $R_2$ under the standard bipolar process. The resistance of $R_1$ depends on different geometric shape factors, including the cross-sectional area of the contact between the injection source and the substrate and the distance between them. As an ultrathin chip becomes thinner, the cross-sectional area of its substrate decreases, leading to an increase in $R_1$ and a consequent increase in substrate resistance $R_S$. An increase in the substrate resistance $R_S$ further causes the substrate bias voltage to increase, which may form a parasitic PNP transistor. This parasitic PNP transistor consists of the collector region, base region and substrate of the NPN transistor.

Further consideration is given to study the effect of parasitic transistors on the circuit. For an isolated PN junction, the relationship between the current flowing through the junction and the voltage across it conforms to Shockley's law, which can be expressed as

$$I = I_{SO}\left[\exp\left(U/U_T\right) - 1\right] \tag{1}$$

The current flowing through collector, emitter and substrate is expressed as $I_1$, $I_2$, and $I_3$, respectively

$$\begin{pmatrix} I_1 \\ I_2 \\ I_3 \end{pmatrix} = \begin{pmatrix} 1 & -\alpha_R & 0 \\ -\alpha_F & 1 & -\alpha_{SR} \\ 0 & -\alpha_{SF} & 1 \end{pmatrix} \begin{pmatrix} I_{ES}\left[\exp\left(U_{BE}/U_T\right) - 1\right] \\ I_{CS}\left[\exp\left(U_{CE}/U_T\right) - 1\right] \\ I_{SS}\left[\exp\left(U_{SC}/U_T\right) - 1\right] \end{pmatrix} \tag{2}$$

where $U_T = kT/q \approx 0.026\text{V}$, $T$ = 300 K, $I_{ES}$, $I_{CS}$, and $I_{SS}$ represent the reverse saturation currents of the emitter-substrate, collector-substrate, and substrate junctions, respectively, when each junction is considered independently. Each $\alpha$ is the current gain of these junctions under the specific connection method. $I_E$, $I_B$, $I_C$ and $I_S$ respectively represent the terminal currents flowing through the emitter, base, collector and substrate (The circuit equivalent diagram is shown in Fig. S4 in the ESM). According to Kirchhoff's law

$$\begin{pmatrix} I_E \\ I_B \\ I_C \\ I_S \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & -1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} I_1 \\ I_2 \\ I_3 \end{pmatrix} \tag{3}$$

The current $I_C$ flowing through the collector and the current $I_S$ flowing through the substrate can be expressed as

$$I_c = \alpha_F I_{ES} e^{U_{BE}/U_T} - (1-\alpha_{SF}) I_{CS} e^{U_{BC}/U_T} - (1-\alpha_{SR}) I_{SS} e^{U_{SC}/U_T} \tag{4}$$

$$I_S = -\alpha_{SF} I_{CS} e^{U_{BC}/U_T} + I_{SS} e^{U_{SC}/U_T} \tag{5}$$

As the substrate bias voltage increases, a positive substrate voltage causes the base region to be depleted, thereby reducing the barrier for electron injection from the emitter to the base region, which leads to an increase in the collector current $I_C$ with the increase of the substrate voltage

$$I_C R_C + U_{SC} = I_S R_S \tag{6}$$

Set the above equations together

$$e^{U_{BE}/U_T} = \frac{\alpha_{SF} I_C - (1-\alpha_{SF}) \dfrac{I_C R_C + U_{SC}}{R_S} - (1-\alpha_{SF}\alpha_{SR}) I_{ss} e^{U_{SC}/U_T}}{\alpha_{SF}\alpha_F I_{ES}} \tag{7}$$

using $I_{ES} e^{U_{BE}/U_T}$, $I_{CS} e^{U_{BC}/U_T}$ much bigger than $I_{SS} e^{U_{SC}/U_T}$, we can simplify Eq. (4)/Eq. (5)

$$\frac{I_C}{I_S} = \frac{\lambda}{I_{CS}\alpha_{SF}} \tag{8}$$

$$\lambda = (1-\alpha_{SF}) I_{CS} - \alpha_F I_{ES} e^{\frac{U_{BE}-U_{BC}}{U_T}} \tag{9}$$

Since the NPN transistor works in saturation region ($U_{BE} \approx U_{BC}$), $\lambda$ can be regarded as a constant coefficient, and the equation can be rearranged as

$$I_C = \frac{\lambda U_{SC}}{I_{CS}\alpha_{SF} R_S - \lambda R_C} \tag{10}$$

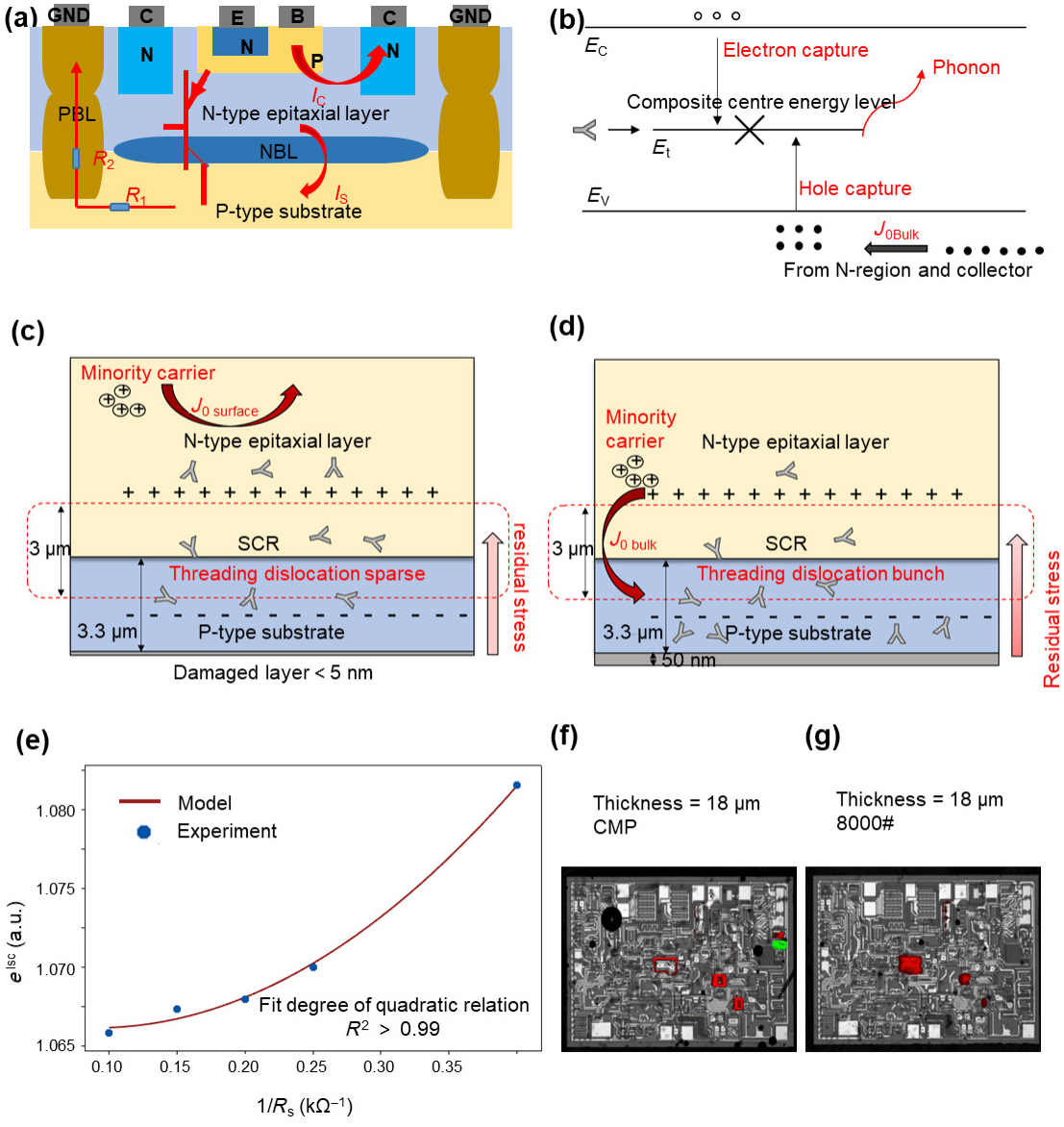



**Figure 4** (a) NPN device structure profile. (b) Carrier trapping at dislocation composite centers. (c) Surface leakage current dominated leakage current paths. (d) Bulk leakage current dominated leakage current paths. (e) The short-circuit output current $I_{sc}$ plotted as a function of the substrate resistance $R_S$. Obirch images of CMP (f) and 8000# (g) chips at 18 μm thickness.

Substitute Eq. (10) into Eq. (7)

$$e^{\frac{U_{BE}}{U_T}} \propto \frac{1}{R_S^2} \tag{11}$$

From Eq. (11), it can be seen that when the UTC becomes thinner, the substrate resistance $R_s$ increases, which causes the collector current $I_c$ and substrate current $I_s$ to increase, and further affects the emitter junction voltage $U_{BE}$, which is quadratically inversely proportional to the size of the substrate resistance. As an important parameter of the transistor, changes in the emitter junction voltage will inevitably cause changes in other circuit parameters, which will have an impact on the electrical performance of the chip.

Residual stresses introduced during the grinding process may become another important factor affecting the chip circuitry. For a PN junction, the leakage current can be described using the following equation

$$J_{0\,\mathrm{Surface}}L + J_{0\,\mathrm{Bulk}}S = I_{0\,\mathrm{Total}} \tag{12}$$

where $J_{0\,\mathrm{Surface}}$ is the surface leakage current density, $J_{0\,\mathrm{Bulk}}$ is the body leakage current density, $L$ is the PN junction perimeter, $S$ is the junction cross-sectional area and $I_{0\,\mathrm{Total}}$ is the total leakage current. When silicon is subjected to a certain amount of stress, the six-degree merging energy valley $\Delta_6$ near the bottom of the conduction band splits into two groups of discrete valleys $\Delta_2$ and $\Delta_4$. One of the splitting valleys must have a smaller value of energy compared to the original $\Delta_6$ valleys, which becomes the new conduction band edge. This often results in a reduction of the forbidden band gap width of the silicon material under the influence of the stress. Even though the study by Song et al. [22] indicates that the 111-grain orientation of silicon does not split the energy valley under stress, it still reduces the forbidden bandwidth under compressive stress. In bandgap theory, the smaller the forbidden bandwidth is, the easier it is for the electrons to obtain enough energy to jump from the valence band to the conduction band, and even in the absence of an

external electric field, the conduction band will have more free carriers, which leads to an increase in the body-leakage current. This phenomenon persists even in the absence of an external electric field, resulting in an increase in the body-leakage current due to the presence of a greater number of free carriers in the conduction band. Using $n_i$ and $E_g$ to represent the intrinsic carrier concentration and bandgap width of the material, respectively, the bulk leakage current is governed by the following relationship

$$J_{0\,\mathrm{Bulk}} \sim n_i \sim \exp\left(-\frac{E_g}{2kT}\right) \tag{13}$$

As shown in Fig. 3(c), leakage current exhibits a sudden increase of 3–5 times when chip thickness drops below 18 μm, but this abrupt change disappears in the CMP and RIE process groups. Further investigation of bending stress effects reveals that bending-induced leakage variations remain below 15%, even under maximum bending curvature where the simulated stress at the chip center exceeds 500 MPa (significantly higher than residual stress). These results indicate that the leakage anomaly in ultra-thin chips (< 18 μm) is not solely stress-dependent, but may involve synergistic effects of dislocation density accumulation. As shown in Figs. 4(c) and 4(d), in addition to directly affecting the bandgap width of the material, the residual stress may have an indirect effect on the chip by affecting the dislocation distribution inside the chip. Under the influence of stress, dislocations tend to migrate to positions of lower energy. When dislocations bunch, their strain fields partially cancel each other out, thereby reducing the total energy of the system. This energy-reducing mechanism of dislocation bunching occurs more frequently under high stress conditions. Concurrently, molecular dynamics simulations show that diffusion processes with small activation volumes, such as atomic rearrangement-assisted dislocation nucleation, are suppressed under high stress conditions [23]. Such atomic rearrangement may play a significant role in the process of dislocation annihilation. This implies that in regions of high stress, dislocations are more likely to bunch and less likely to annihilate and vice versa in low-stress regions. The effect brought about by dislocations on ultrathin chips is shown in Fig. 4(b), which can act as effective carrier trapping sites or complexation centers, allowing the distribution of defect energy levels with different energies within the band gap, affecting the carrier trapping and complexation processes in terms of both rates and activation energies, and consequently leading to the enhancement of the bulk leakage current. The leakage current generated by carriers in the dislocation-associated defect states can be calculated by the following formula

$$J_{0\,\mathrm{Bulk}} = q n_i W \sigma_n V_{th} N_D N_{TD} \tag{14}$$

In the formula, $q$ represents the elementary charge of an electron; $W$ is the depletion width, $\sigma_n$ is the capture cross-section of the dislocations, $N_D$ is the dislocation density, $N_{TD}$ is the number of traps per unit length of dislocation, and $V_{th}$ is the thermal velocity of the carriers. The equation indicates that the bulk leakage current is directly related to the density of threading dislocations at the transistor junction. In conclusion, the residual stresses generated by mechanical grinding may exert a considerable influence on chip leakage, both directly through their impact on the bandgap of the silicon material and indirectly through their effect on the rearrangement of dislocation densities.

The theoretical predictions for both $I_{sc}$ and $I_{st}$ were subsequently validated through experimental characterization. In the device current test, the short-circuit output current ($I_{sc}$), exhibited significant variations in the CP tests, as shown in Fig. 3(f). The data shows that the short-circuit output current decreases proportionally with the decrease in wafer thickness. The short-circuit output current is reduced by about 30% for the 20 μm wafer compared to the unthinned wafer. In fact, the $I_{sc}$ is closely related to the junction voltage $U_{BE}$

$$U_{BE} = I_{SC} R_{SC} \tag{15}$$

Eq. (11) can be written

$$e^{\frac{I_{SC} R_{SC}}{U_T}} \propto \frac{1}{R_S^2} \tag{16}$$

Figure 4(e) shows the relationship between $e^{I_{sc}}$ and $\frac{1}{R_s}$ according to experimental data. After fitting, $e^{I_{sc}}$ and $\frac{1}{R_s}$ show accurate quadratic relationship, which proves the theoretical prediction in Eq. (16).

As shown in Figs. 4(c) and 4(d), the phenomenon of sudden change in chip leakage current is actually undergoing a process of change from being dominated by surface leakage current to being dominated by bulk leakage current. This process is accompanied by a change in the leakage current path, with the surface leakage current flowing from the collector to the emitter and the bulk leakage current flowing directly from the collector to ground. To characterize this change process, the Optical Beam Induced Resistance Change (OBIRCH) technique is used, which is a failure analysis technique for semiconductor devices and integrated circuits that allows fast and accurate location of leakage paths in integrated circuits using laser scanning and thermal effects. Figures 4(f) and 4(g) show the leakage current paths determined by Obirch characterization under two different leakage conditions, normal and abnormal. It can be observed that the leakage current path undergoes an abrupt change, where the current changes from flowing from the collector to the emitter under normal conditions to flowing directly from the collector to ground. This indicates that the main part of the leakage current undergoes a transition from surface leakage current to bulk leakage current. Since all the samples have the same surface treatment process, it can be assumed that they also have the same surface leakage current per unit circumference. Therefore, the sudden change in leakage current must be the result of a significant increase in body leakage current. In the theoretical prediction, we have predicted that the joint effect of stress and defects will lead to the increase of leakage current through the bandgap theory and the composite theory, and in fact, the thickness of the chip is also a necessary factor for the increase of leakage current. Figure 3(a) shows that no leakage mutation occurs in the chips with a thickness of more than 20 μm. Through Fig. 1(d), it is not difficult to find out that the stress influence of the damage layer on the chip is mainly concentrated in the first 3 μm, so when the chip is thick enough, the stress influence of the damage layer on the depletion region is small, which is difficult to cause the change of the dislocation density and the change of the bandgap. However, as the chip becomes thinner, the stress effect gradually covers the depletion region, which in turn affects the bandgap width. Based on the stress characterization results in Fig. 1(d), the variation of bandgap width in the depletion region of the 18 μm ultrathin chip is less than 0.01 eV, consistent with the hydrostatic

deformation-potential model for silicon [24], which means that the difference in leakage current between the thick chip and the thin chip with similar impurity levels should be less than 20%. In Fig. 3(b), the total leakage current of the thin chip reaches 5 times of that of the thick chip. Considering that the surface current does not change, the actual increase in body leakage current should be more than 10 times. This indicates that the increase of leakage current in thin chips is not only caused by the bandgap difference, but the change of dislocation density caused by stress plays a more important role.

## 3 Conclusion

This study demonstrates that the electrical degradation of UTCs arises from parasitic effects induced by thickness loss and stress-defect interactions within grinding-induced damaged layers. In large-scale wafer-level CP testing, parasitic coupling caused short-circuit output current fluctuations within CP test specification limits, while stress-driven dislocation rearrangement near the critical thickness (18 μm) provoked abrupt leakage surges, causing wafer yield to plummet to 7.19%. Crucially, the integration of post-thinning CMP/RIE processes effectively suppressed these leakage anomalies by decreasing and redistributing residual stress, restoring yield to > 94% even at 18 μm. Furthermore, Raman spectroscopy and TEM characterization detected residual stress gradients and high-density dislocation clusters in subsurface regions beneath the damage layer, while bending experiments established defect dominance in stress-defect interactions governing electrical degradation. Importantly, similar electrical degradation phenomena with varying degrees of severity were also observed in other types of thinned bipolar chips (such as IDO, DCDC, and especially high-gain operational amplifier chips), where multiple key parameters deteriorated with decreasing thicknesses. This indicates that the underlying stress-defect coupling and dislocation-related failure mechanisms identified in this work are broadly relevant to bipolar integrated circuits beyond the PWM devices studied here. This work investigates the interplay between residual stress and dislocation rearrangement, along with their stress-dislocation coupling from backside to active regions, establishing a theoretical foundation for reliability enhancement in backside thinning processes of ultrathin integrated circuits. Looking forward, the stress-relief approach and critical thickness threshold identified here provide a practical pathway for scaling ultrathin integrated circuits to sub-15 μm regimes, with relevance extending to other bipolar and advanced device architectures. The underlying stress-defect mechanisms offer a transferable framework for backside thinning of diverse materials and structures, while the combined characterization methods can support future process diagnostics. Further integration of *in-situ* monitoring and adaptive post-processing holds promise for enabling reliable ultrathin devices in 3D integration, flexible electronics, and high-performance applications.

## 4 Methods

### 4.1 Materials

This study utilized a commercial silicon-based (111) bipolar PWM chip (KA3845, Yangzhou Yungu Microelectronics Co., Ltd., China) as the starting material. The initial wafer had a thickness of 525 μm and a diameter of 150 mm. The fundamental circuit structure was shown in Fig. S4 in the ESM.

### 4.2 Mechanical grinding

Mechanical grinding experiments on single-crystal silicon wafers were performed using a single-axis precision surface grinder (DAG-810, Disco Corp., Japan), while CMP was subsequently carried out on a tri-axis grinder (PG200RM, ACCRETECH., Japan). Resin-bond diamond grinding wheels with mesh sizes of #3000 and #8000 were employed for mechanical grinding. Subsequently, two separate post-thinning processes were applied to different sample groups: One group underwent CMP using a colloidal silica-based slurry (pH 11–13, containing 35 wt.%–40 wt.% $SiO_2$ and 1 wt.%–5 wt.% tetramethylammonium hydroxide (TMAH)) with a platen rotation speed of 300 rpm, polishing time of 60 s, slurry flow rate of approximately 500 mL/min, and resulting surface roughness (Ra) of ~ 1.5 nm; another group underwent RIE using $SF_6$ as the etching gas with a flow rate of 90 sccm, RF power of 100 W, and etching time of 300 s.

### 4.3 Characterization

The SEM (SU-8230, Hitachi, Japan) was employed to characterize the cross-sectional microstructure of the chips. The focused ion beam system (Helios G5, Thermo Fisher Scientific, USA) was utilized for preparing of cross-sectional TEM specimens. The transmission electron microscope (Talos F200X, Thermo Fisher Scientific, USA) was applied to analyze subsurface damage layers and deep dislocations within the thinned wafers. The subsurface stress distribution along the depth direction of the chip backside was analyzed using a confocal Raman spectrometer (inVia, Renishaw, UK) with switchable 532 and 633 nm lasers.

### 4.4 Electrical performance testing

CP testing was performed in an integrated system interfacing an automated wafer prober (UF200R, ACCRETECH., Japan) with a semiconductor test system (V93000 SOC, Advantest., Japan), utilizing low-stiffness probes (1 gf/mil spring constant) with planar circular tips (Ø25 μm) to minimize the mechanical stress during multi-site contact. Wafer dicing employing diamond-impregnated blades produces individual dies. Following dicing, the ultrathin dies are mounted onto polyimide (PI) substrates and wire bonding establishes electrical interconnects between device pads and flexible circuits. Subsequent encapsulation involves silicone elastomer dispensing over the active surface, thermally cured at 120–150 °C to achieve conformal protection. The Semiconductor Analyzer (Keithley 4200A-SCS, Tektronix, USA) was employed for systematic electrical characterization of the packaged chips.

**Electronic Supplementary Material**: Supplementary material (including wafer thickness variation data, chip electrical performance statistics, and circuit diagrams) is available in the online version of this article at https://doi.org/10.26599/NR.2026.94908642.

## Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (No. 2023YFF0616800) and

Zhejiang Province Key Research and Development Project (No. 2021C05007-04).

## Declaration of competing interest

The authors declare no competing interests.

## Author contribution statement

All authors contributed to the preparation of this manuscript. N. G.: Methodology, conceptualization, writing, data curation, design, and preparation. Z. X. W. and D. L. L.: Methodology, conceptualization, theoretical guidance. X. S.: Process assistance. Y. C.: Process assistance. W. J.: Process assistance, methodology. T. S. P.: Visualization, methodology, conceptualization. W. Q.: Process assistance. Y. C.: Project administration, visualization, methodology, conceptualization. Y. L.: Visualization, methodology, funding acquisition, project administration, resources. X. F.: Visualization, methodology, funding acquisition, project administration, resources.

## Use of AI statement

None.

## References

[1] Bollam, P.; Gothwal, E.; GBCS Tejaswi, V.; Kumar, S.; Deb, S. Mobile collaborative augmented reality with real-time AR/VR switching. In *Proceedings of the ACM SIGGRAPH 2015 Posters*, Los Angeles, 2015, pp 25.

[2] Ji, H.; Wang, J.; Gao, J.; Liu, X. H. Research on telemedicine technology and implement based on Virtual Reality. In *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, Xi'an, China, 2016, pp 1581–1586.

[3] Yao, G.; Kang, L.; Li, J.; Long, Y.; Wei, H.; Ferreira, C. A.; Jeffery, J. J.; Lin, Y.; Cai, W. B.; Wang, X. D. Effective weight control via an implanted self-powered vagus nerve stimulation device. *Nat. Commun* **2018**, *9*, 5349.

[4] Yao, G.; Jiang, D. W.; Li, J.; Kang, L.; Chen, S. H.; Long, Y.; Wang, Y. Z.; Huang, P.; Lin, Y.; Cai, W. B. et al. Self-activated electrical stimulation for effective hair regeneration via a wearable omnidirectional pulse generator. *ACS Nano* **2019**, *13*, 12345–12356.

[5] Kim, J.; Son, D.; Lee, M.; Song, C. Y. O. N.; Song, J. K.; Koo, J. H.; Lee, D. J.; Shim, H. J.; Kim, J. H.; Lee, M. et al. A wearable multiplexed silicon nonvolatile memory array using nanocrystal charge confinement. *Sci. Adv.* **2016**, *2*, e1501101.

[6] Long, Y.; Wei, H.; Li, J.; Yao, G.; Yu, B.; Ni, D. L.; Gibson, A. L.; Lan, X. L.; Jiang, Y. D.; Cai, W. B. et al. Effective wound healing enabled by discrete alternative electric fields from wearable nanogenerators. *ACS Nano* **2018**, *12*, 12533–12540.

[7] Cheng, T.; Zhang, Y. Z.; Lai, W. Y.; Huang, W. et al. Stretchable thin-film electrodes for flexible electronics with high deformability and stretchability. *Adv. Mater.* **2015**, *27*, 3349–3376.

[8] Bao, Z. N.; Chen, X. D. Flexible and stretchable devices. *Adv. Mater.* **2016**, *28*, 4177–4179.

[9] Kim, D. H.; Rogers, J. A. Stretchable electronics: Materials strategies and devices. *Adv. Mater.* **2008**, *20*, 4887–4892.

[10] Novoselov, K. S.; Fal'ko, V. I.; Colombo, L.; Gellert, P. R.; Schwab, M. G.; Kim, K. A roadmap for graphene. *Nature* **2012**, *490*, 192–200.

[11] Zhu, P. Y.; Li, Z. H.; Pang, J. B.; He, P.; Zhang, S. Y. Latest developments and trends in electronic skin devices. *Soft Sci.* **2024**, *4*, 17.

[12] Zheng, K. W.; Cai, S. S.; Chen, Y.; Ma, Y. J.; Feng, X. Manufacture and characterization of ultrathin flexible chips. In *2021 5th IEEE Electron Devices Technology & Manufacturing Conference (EDTM)*, Chengdu, China, 2021, pp 1–3.

[13] Tao, H. F.; Liu, Y. H.; Zhao, D. W.; Lu, X. C. Ductile deformation and subsurface damage evolution mechanism of silicon wafer induced by ultra-precision grinding process. *Tribol. Int.* **2023**, *189*, 108879.

[14] Mizushima, Y.; Kim, Y.; Nakamura, T.; Sugie, R.; Hashimoto, H.; Uedono, A.; Ohba, T. Impact of back-grinding-induced damage on Si wafer thinning for three-dimensional integration. *Jpn. J. Appl. Phys.* **2014**, *53*, 05GE04.

[15] Zhang, Z. W.; Chen, P.; Qin, F.; An, T.; Yu, H. P. Mechanical properties of silicon in subsurface damage layer from nano-grinding studied by atomistic simulation. *AIP Adv.* **2018**, *8*, 055223.

[16] Jian, W.; Wang, Z. X.; Jin, P.; Zhu, L. J.; Chen, Y.; Feng, X. Subsurface damage and bending strength analysis for ultra-thin and flexible silicon chips. *Sci. China Tech. Sci.* **2023**, *66*, 215–222.

[17] Smeys, P.; Griffin, P. B.; Saraswat, K. C. Influence of post-oxidation cooling rate on residual stress and pn-junction leakage current in LOCOS isolated structures. *IEEE Trans. Electron Dev.* **1996**, *43*, 1989–1993.

[18] Grillot, P. N.; Ringel, S. A.; Fitzgerald, E. A.; Watson, G. P.; Xie, Y. H. Electron trapping kinetics at dislocations in relaxed $Ge_{0.3}Si_{0.7}$/Si heterostructures. *J. Appl. Phys.* **1995**, *77*, 3248–3256.

[19] Giovane, L. M.; Luan, H. C.; Agarwal, A. M.; Kimerling, L. C. Correlation between leakage current density and threading dislocation density in SiGe p-*i*-*n* diodes grown on relaxed graded buffer layers. *Appl. Phys. Lett.* **2001**, *78*, 541–543.

[20] Eneman, G.; Simoen, E.; Delhougne, R.; Verheyen, P.; Simons, V.; Loo, R.; Caymax, M.; Claeys, C.; Vandervorst, W.; De Meyer, K. Analysis of the leakage current origin in thin strain relaxed buffer substrates. *J. Electrochem. Soc.* **2006**, *153*, G379.

[21] Qiu, W.; Ma, L. L.; Li, Q.; Xing, H. D.; Cheng, C. L.; Huang, G. Y. A general metrology of stress on crystalline silicon with random crystal plane by using micro-Raman spectroscopy. *Acta Mech. Sin.* **2018**, *34*, 1095–1107.

[22] Song, J. J.; Zhang, H. M.; Hu, H. Y.; Dai, X. Y.; Xuan, R. X. Valence band structure of strained Si/(111)$Si_{1-x}$ $Ge_x$. *Sci. China Phys. Mech. Astron.* **2010**, *53*, 454–457.

[23] Barchuk, M.; Röder, C.; Shashev, Y.; Lukin, G.; Motylenko, M.; Kortus, J.; Pätzold, O.; Rafaja, D. Correlation between the residual stress and the density of threading dislocations in GaN layers grown by hydride vapor phase epitaxy. *J. Cryst. Growth* **2014**, *386*, 1–8.

[24] Van De Walle, C. G. Band lineups and deformation potentials in the model-solid theory. *Phys. Rev. B* **1989**, *39*, 1871–1883.