UNIVERSITY OF CALIFORNIA
RIVERSIDE

Design-for-Reliability on Thermal Management and ESD Protection of Integrated Circuits

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

by

Cheng Li

March 2022

Dissertation Committee:
Dr. Albert Wang, Chairperson
Dr. Ming Liu
Dr. Sheldon Tan

The Dissertation of Cheng Li is approved:

____

____

____
Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

I would like to take this opportunity to sincerely express my gratitude to my research advisor, Dr. Albert Wang, heartfeltly, during my Ph.D. studies at the University of California, Riverside (UCR), for continuous support, encouragement, supervision, and academic mentorship. His conscientious spirit towards work and rich experience in life illuminate me a lot and formulate my own way of critical thinking and time arrangement skills, especially communication skills. I have been inspired by his motivation, broader knowledge, and affluent ideas through my research periods and I hope to absorb them to guide my future work career and thrive in my rest life. It’s been a great fortune for me to have him as my advisor.

I would like to thank the Department of Electrical and Computer Engineering at the University of California, Riverside, especially the members of my dissertation committee, Dr. Ming Liu and Dr. Sheldon Tan, for their supports of my dissertation research.

I would thank all my laboratory mates with whom I have conducted researches together, including Dr. Fei Lu, Dr. Chen Qi, Dr. Chenkun Wang, Dr. Feilong Zhang, Mr. Mengfu Di, and Ms. Zijin Pan. Through discussion on the research work and help from life, we make a unified group which is beneficial to my study. Also, I want to thank my senior laboratory mates Dr. Zongyu Dong and Dr. Bo Qin, their professional guidance on my work and concerns for my life.

I appreciate the opportunity to do the design internship on Fairchild-Onsemi and thanks to Dr. Bin Zhao and Mr. Burl Fermandes. Also, I want to express my gratitude to the group of the School of Integrated Circuits, Tsinghua University for my research support

such as Dr. Huaqiang Wu, Dr. Junxu, Dr. Rongren Liang, Dr. Weijun Cheng, Mr. Tao Zhong, and Ms. Li Zong.

Last but not least, I would like to express my deepest gratitude to my lovely parents Mr. Jinchuan Li and Ms. Li Liang, also my passionately devoted fiancée Ms. Shuai Xu. Thanks for their generous and endless love to me. They are always with me when I go through tough situations and provide encouragement and support to me.

ABSTRACT OF THE DISSERTATION

Design-for-Reliability on Thermal Management and ESD Protection of Integrated Circuits

by

Cheng Li

Doctor of Philosophy, Graduate Program in Electrical Engineering
University of California, Riverside, March 2022
Dr. Albert Wang, Chairperson

Transistor self-heating is a grand challenge in modern integrated circuit (IC) chip designs. Chip-scale thermal management is required to ensure IC design reliability. As IC technologies rapidly advance into few-nm nodes, while performance, complexity and size of chips continue to increase, heat dissipation becomes a technical bottleneck to advanced chips. For example, data servers and mobile electronics increasingly rely on high clock frequency, multiple-core CPU and GPU, which are extremely power-hungry and are heavy heat generators. However, portable devices, such as smartphones, have little room to accommodate traditional heat dissipation means. Further, advanced IC technologies, such as silicon-on-insulator (SOI) and FINFET, are inherently in thermal conduction, hence, making transistor self-heating even severer at the chip level. In addition, advanced 3D packaging makes it much harder to dissipate heat from IC dies. Together, advanced high-performance, complex chips made in advanced IC technologies are essentially big heat generators, unfortunately, poor thermal conduction makes IC chips increasingly suffering from heat-induced performance degradation and reliability problems. Existing on-chip thermal sensors used to monitor temperature typically use thermocouple, thermistor and PN diode devices, which are placed on a chip in a coplanar manor, i.e., laterally side-by-side, with the circuit blocks to be monitored. There are several major technical disadvantages of making accurate full-chip thermal management impractical using the existing thermal-sensing techniques. First, a thermal sensor is laterally far away from a transistor to be monitored, making it impossible to accurately detect any real hot spots, which are often at the corners of a conduction channel of a transistor. Second, these thermal sensors are bulky, hence, impossible to construct a very large thermal sensor network on a chip to realize full-chip thermal mapping. Consequently, existing on-chip thermal sensing may only achieve circuit block level thermal-sensing resolution, which makes practical thermal management impractical. In order to fundamentally address the transistor self-heating problem and enable effective full-chip thermal management, novel thermal sensing techniques are needed to accurately detect the transient hot spots at the single transistor level, which hence requires transistor-level thermal sensing resolution. Further, to enable accurate run-time full-chip thermal management, a large on-chip thermal sensor mash network is required to achieve transistor-level thermal sensing resolution, which in turn, make full-chip thermal management practical. In this dissertation, I proposed and prototyped a novel under-FET thermal sensor device structure, which utilizes a vertical PN junction. This PN junction thermal sensor is made inside a through-silicon via (TSV) type vertical hole, which is placed directly under a MOSFET, hence being able to detect the transient hot spot in a MOSFET channel. Since the TSV-like under -FET thermal sensor

does not take any extra “lateral” Si die area and is placed right underneath the heating source, i.e., the MOSFET channel, it allows constructing a large thermal sensor mash network on a chip to realize transistor-level thermal sensing resolution without taking extra IC die area. Using the new under-FET thermal sensing technique, a machine-learning (ML) algorithm is proposed to enable run-time full-chip thermal management, which can fundamentally resolve the transistor self-heating-induced chip performance degradation and thermal reliability problems.

Electrostatic discharge (ESD) protection is another major IC reliability challenge, particularly for complex chips implemented in advanced IC technologies. For the past six decades, substantial R&D efforts have been devoted to developing various on-chip ESD protection solutions. Yet, as IC technologies continue to shrink, while IC performance and complexity continuously increase, on-chip ESD protection for advanced ICs becomes extremely challenging. In general, any ESD protection structures inevitably induce the ESD-design overhead problem, including parasitic capacitance, noises and leakages, as well as ESD device size and layout problem. The semiconductor industry urgently needs novel on-chip ESD protection solutions to overcome the ESD-design overhead problems. In this desertion, I proposed and demonstrated three novel on-chip ESD protection structures aiming to address the ESD-design overhead problem. The first new ESD protection structure proposed is a vertical TSV-like diode, which is a truly vertical PN-diode ESD protection device residing inside a TSV-like vertical hole under a bonding pad. Unlike any traditional PN diode ESD protection devices, which always require lateral discussion elements for electrical interconnections, the new TSV-like ESD diode can conduct ESD pulses and ESD-induced heat vertically, hence, significantly improving ESD protection while minimizing Si die area consumed by large ESD protection structures. The second new ESD protection structure is a cell-based Sudoku-type diode-trigger silicon-controlled rectifier (DTSCR) low-triggering ESD protection sub-circuit structure. The third novel ESD protection structure is a single-crystalline graphene-based nano-electromechanical system (gNEMS) ESD protection switch device. This gNEMS ESD protection structure is novel in that it is a mechanical switch made in the back-end-of-line (BEOL) in CMOS, which can be turned on/off extremely fast to provide both human body models (HBM) and charged device model (CDM) ESD protection. Both theoretical and experimental studies were conducted to validate the three novel ESD protection structures

Chapter 1 introduces the transistor self-heating problem, which is the root cause of the thermal reliability problem of modern IC chips. Chapter 2 discusses the new TSV-type under-FET thermal sensor device and the ML-based full-chip thermal management method. Chapter 3 gives an induction of ESD protection fundamentals and graphene material background. Chapter 4 presents the novel vertical TSV-like ESD protection diode structure of the TCAD mixed-mode ESD protection design calibration flow. Chapter 5 discusses the new Sudoku-type ESD protection array and theoretical design analysis. Chapter 6 discusses the novel single-crystallin graphene-based gNEMS ESD switch structure. Chapter 7 summarizes my Ph.D. research achievements.

Contents

List of Figures

Chapter 1 Introduction to heat dissipation and thermal management

1.1 Heat dissipation background

1.1.1 Heat dissipation system for different sizes of applications

Considering the computer heat dissipation, we can find that there are different methods to release the heat generated from the chip to its surroundings. The only unchanged thing is that as long as any chip starts to run, it will generate heat. Therefore, under the situation that we do not change the amount of the heat generation, we need to develop the methods for heat dissipation, which dynamically balance the generated heat and dissipated heat. Seen from figure 1.1, the different application covers from the largest data center to the smallest wearable device like a smart watch. For the data center, there are lots of computer servers running all the time which generate massive heat

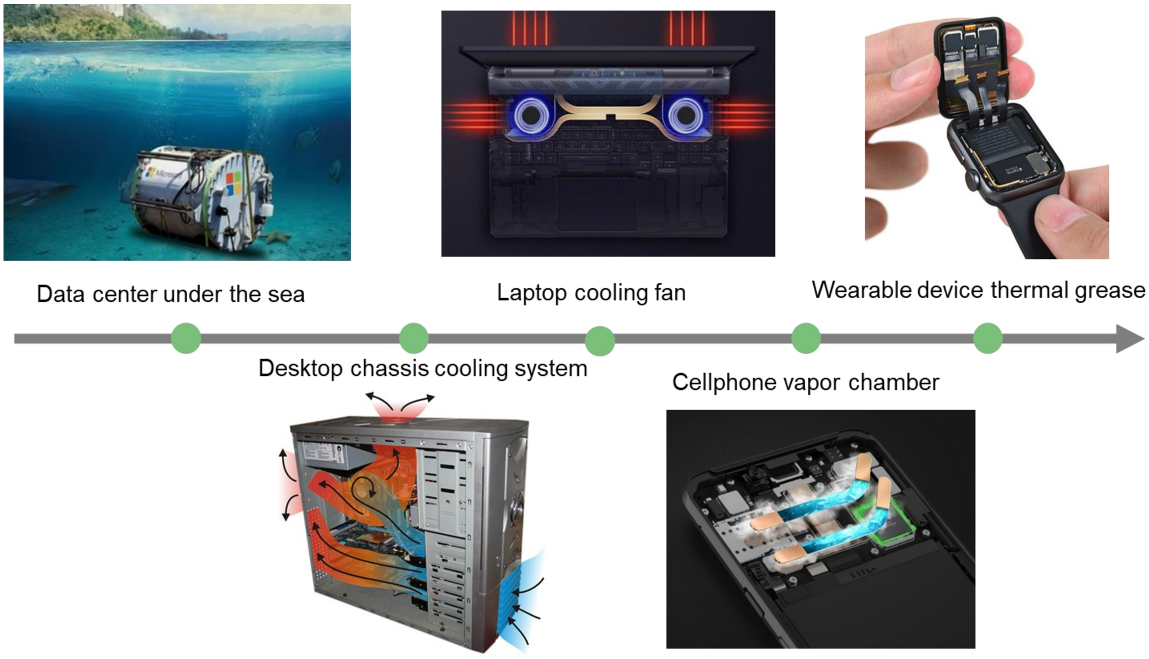
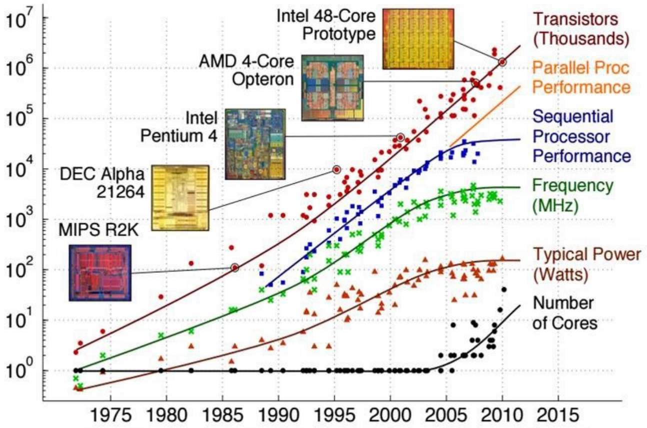

Data center under the sea
Laptop cooling fan
Wearable device thermal grease
Desktop chassis cooling system
Cellphone vapor chamber

Figure 1.1 Different heat dissipation system for different applications

simultaneously. Now the best method nowadays is to place the data center under the sea to utilize the natural and abundant sea water to flow around the data center to take away plentiful heat. For the size scaling to the desktop, the main heat dissipation tool is the chassis cooling system. Normally, many fans are placed above the massive heat generation source like the computer processing unit (CPU), graphic processing unit (GPU), and power unit. The heat generated by these modules is delivered by their fans to the whole chassis then via the fans installed on the surface of the chassis to form a good air flow channel to exhale the hot air and inhale the environment cold air. For some severely loaded desktop system, they even use the water-cooling system which use the higher heat capacity fluid to take the heat more efficiently than air. Then the next scaled device is the laptop which can only use one or two small sizes of cooling fans plus the conductive copper tube. Compared with the desktop which directly installed the cooling fan above the heat generation source combining the heat sink to increase the heat conductivity, the laptop can only use the copper tube to pass the heat to the small cooling fan to indirectly dissipate the heat due to the limited volume of the laptop, which inevitable eliminate the efficiency of heat dissipation. Then moving to the smaller volume device, cellphone, which can only have the vapor chamber similar to the copper tube but without the cooling fan. After the heat passes through the vapor chamber to transfer the heat from the chip to another location of the cellphone in order to average the temperature. The last device is a wearable device like a smart watch, which has almost no heat dissipation solution but for the little effective thermal grease. Of course, there is no tool to facilitate heat dissipation and we have only counted on less heat generation. The law reveals the phenomenon that with the volume of the device decreasing, there are fewer methods for heat dissipation. Since today's devices are continuing to scale, the heat dissipation is more and more useless and we should pay more attention to temperature monitoring and controlling.

1.1.2 Heat generation with the chip development

After the first integrated circuits were developed, the Moore law always guides the IC development like scale down and the transistor density doubles a while year. As seen from figure 1.2, the transistor density arises from 102 orders to 106 orders within these 30 years from the 1980s to the 2010s. A high density of transistors will definitely make the space between each transistor is limited which is unbeneficial for heat dissipation. Of course, more numbers of transistors mean more probability of heat generation source. Not only to add the number of transistors, but the modern CPU also boosts the clock frequency of the transistor in order to increase its computing ability in turn of adding the heat generation. Also, considering today's needs, the chip integrates more and more functions including fast charging, 5G communication, and individual GPU modules, which are all the main



Data partially collected by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond

Figure 1.2 The historical trend and CPU scaling development

heat generation sources. Not only from the hardware requirements but also the software submits the high requirement for the CPU and GPU computing ability. Today work typically needs to operate much software simultaneously and interact within them, which needs the heavy load of CPU supporting for the multitasks running. Besides, better and better vision quality needs the complexity of the high GPU computing abilities. These requirements are shown in figure 1.3 also will cause high-power consumption during running, which in turn will generate lots of heat. It is noted that these requirements normally happen in laptops and cellphones which have limited tools for heat dissipation. Therefore, it will intensify the temperature issue.

Figure 1.3 The illustration for the multitasks running and high-quality of vision during the game.

1.1.3 CMOS Technology influence on the heat dissipation

With the technology node development, the complementary MOSFET (CMOS) structure has developed with both the front end of the line (FEOL) and the back end of the line (BEOL). For the FEOL, the structure has changed a lot from the plenary CMOS to the 3D structure of FinFET and even the gate all around (GAA). As shown in figure 1.4, we can find the difference between the plenary CMOS structure and the FinFET structure. For the plenary CMOS structure, the core heat source, the channel, is adjacent with the substrate, the source, and the drain with a large area connected which is beneficial for the heat conduction from the channel to the surrounding part for the heat dissipation. However, for the FinFET, since the channel arches up with the three-side covered by the gate oxide compared with the channel of the plenary CMOS with one top side covered, has only one bottom side connecting to the substrate for the heat conduction. Also, FinFET is an advanced technology node with a very short length resulting in the small er area connected to the substrate, which is a huge obstacle for heat dissipation [1].

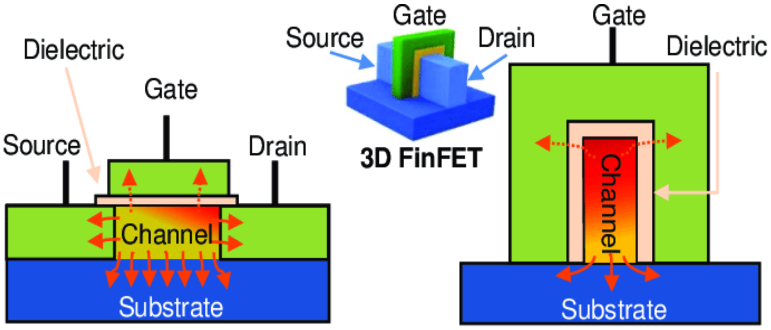


Figure 1.4 Comparison of heat flow between the plenary CMOS and FinFET.

Another structure different from the plenary CMOS is the silicon on the insulator (SOI) structures mainly for the radio-frequency (RF) application to decrease the noise coupling seen in figure 1.5. SOI structure has one layer of buried silicon dioxide underneath the active layer called the BOX. Since SiO_2 has bad thermal conduction, the heat accumulation inside the channel cannot flow past the substrate. Therefore, the heat has to accumulate among the channel, source, and drain parts so as to increase the temperature. Nowadays, many circuits design uses these two technologies to have the main drawback of self-heating. In the meanwhile, BEOL has more and more stacks for the advanced technology node. Because each metal layer has the insulation layer as the isolation mainly comprising of the SiO_2 and Si_3N_4 which are both poor thermal conductors, the heat can also transfer via the

BEOL to the surrounding environment, in turn of adding the difficulties for the heat dissipation.

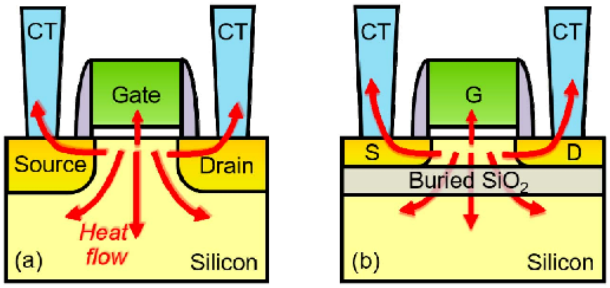


Figure 1.5 Comparison of heat flow between the plenary CMOS and SOI.

The next issue is the packaging technology after the die fabrication. Previously, the packaging is 2D. However, in order to satisfy the scaling down the law, the industry begins to use the 2.5 and 3D packaging which makes many dies stacked together to save the plenary area for the chip. Therefore, for one chip packaging, there are maybe many dies stacked vertically which mean more heat generation source accumulate inside a small volume. Therefore, there are less effective air flow or heat sink for the 3D packaging chip and its temperature easily increase during the chip running [2]. Figure 1.6 shows the illustration for the 3D stacked chip packaging and we can find that the dies far away from the heat sink normally suffers more self-heating issue. According to the right figure example, the peak temperature of the 3D packaging chip is 15% higher than the 2D packaging chip.

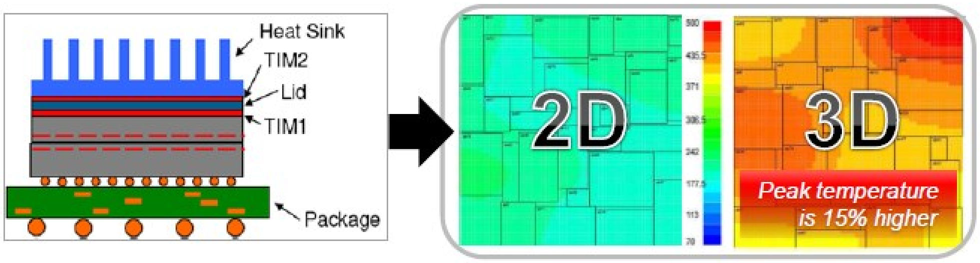


Figure 1.6 Illustration of the 3D packaging chip impact on the temperature

1.2 Thermal management background

1.2.1 Temperature monitor necessity

Since the easy heat generation and hard heat dissipation become a hot potato with the chip development, it is necessary to develop a method to monitor the chip temperature in order to keep it at an acceptable level. Normally the circuit performance will dramatically degrade with the increasing temperature. Even in some cases, the function will be impaired. Because of the intrinsic MOSFET properties, with the increasing temperature, the designed goals of the circuits cannot be fulfilled. Besides, even not on that high temperature, the relatively high temperature will still create many potential problems like largely decreasing the lifetime of the circuits and may even burning down some MOSFETs. In this situation, we must control the temperature of the die to guarantee its good performance and function. Therefore, the important or large scaled circuits like the CPU will have several thermal sensors monitoring the chip temperature based on several hot spots like figure 1.7 [3].

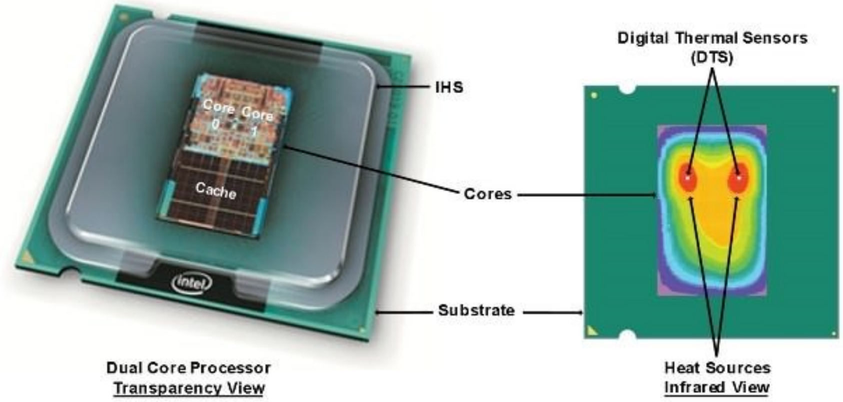


Figure 1.7 Illustration of thermal sensors on the chip

1.2.2 Thermal sensors category

A thermal sensor on-chip normally has several types like thermocouple, thermistor, and PN junction. The thermocouple is an electrical device consisting of two different conductive materials forming the electrical junctions at different temperatures due to the Seebeck effect. The Seebeck effect is the electromotive force that develops across two points of an electrically conducting material when there is a temperature difference between them. Figure 1.8 shows the mechanism of the thermocouple circuit utilizing the Seebeck effect. Thermocouple utilizes this effect to detect one end of the temperature with the aid of the other end in the reference temperature environment via the self-induced voltage changing. The temperature changes responses quickly and it can get a wide temperature range. However, thermocouple normally is not sensitive to minor temperature changing and it always needs the reference temperature atmosphere where hardly coexists with the testing atmosphere within the same environment of the small chip. The thermistor is a type of resistor whose resistance is sensitive to the atmosphere temperature. It has better sensitivity compared with the thermocouple and easy structure just a resistor and its testing range is relatively wide. Nevertheless, thermistors are usually composed of ceramic or polymer material which is not compatible with CMOS fabrication. Also, the worst thing is that the thermistor has very bad linearity because of its self-heating which makes it hardly construct an equation to accurately describe the relationship between the temperature and resistance [4]. The PN junction is the only potential candidate that has good sensitivity, is CMOS fabrication compatible, and has a relatively simple structure. Also, it has good linearity to enable us to build the relationship between temperature and the current or voltage. Though the PN junction has a limited temperature range, the range is enough for IC temperature changing. Also, not a relatively fast response on temperature change is not a big issue for temperature monitoring on IC because the thermal map does not need to refresh at that high speed. In total, the PN junction of the thermal sensor is suitable for IC temperature monitoring [5] [6].

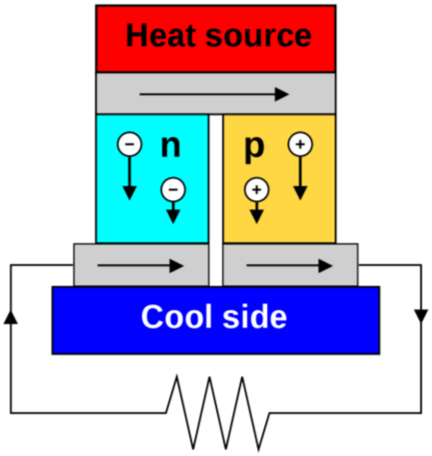


Figure 1.8 The Seebeck effect residing in the thermocouple.

1.2.3 Run-time thermal management flow

For the full thermal management flow, there are five key procedures. The first is the fabricate the under-FET thermal sensor and locate it close to the target transistor. The second step is using the PN junction thermal-electrical property to sense the temperature change and convert it to the current change according to equation 1.

$$I = I_0(e^{\frac{qV}{nkT}} - 1) \qquad (1)$$

In this equation, the I_0 represents the saturation current, q is the charge, V is the voltage, n is the ideal factor (ideal diode is 1 and the non-ideal diode is normally 2~3), k is the Boltzmann factor and the T is the temperature. After getting the current signal, using the analog and digital circuits output the regulated data. Then combined the software to convert the acquired data to the real-time thermal map. Finally, using the feedback control circuits to decrease the power consumption of the over-heated transistor to maintain the system stability. In terms of the controlling methods, we can decrease the clock frequency, decrease the power supply via regulating the LDO for voltage or current mirror for current, and so on [7] [8].

Chapter 2 Under-FET thermal management for design reliabilityThe motivation of Under-FET thermal management

2.1.1 The conventional thermal monitor system

Thermal management is a grand challenge and reliable issue for the ICs stated as previously and due to the more severe inherent self-heating along with the more advanced technology process, thermal management is a necessity for a modern chip. However, the existing on-chip thermal sensing has many limitations. The reported thermal sensors including the thermocouple, thermistor, diode, BJT as well as CMOS circuits block like the bandgap or oscillator circuitry are all based on the block level [9] [10] [11]. Therefore, they cannot support this ultimate full-chip thermal management goal because of several technical problems. First, existing sensors are co-planar in nature which means that they are laterally placed next to or near the IC cores. Thus, the side-by-side proximity sensing cannot achieve the accurate temperature data of the target transistor because sensors are located too far away from the heating source to identify the hot spot like the drain or channel of the victim MOSFET. Second, the existing on-chip thermal is too large compared with the target MOSFET and many thermal sensors are not compatible with the CMOS process. Third, in order to get a full-chip thermal mapping with fine spatial resolution, we need a large in-die sensing mesh network that is not compatible with the existing co-planar sensors. Since these sensors are equivalent to the core circuit in size, they might take the unbearable area overhead on a chip. Facing the practical thermal management, these co-planar sensors are practically either under-functional that fails to detect overheating or too conservative that often leads to an overreaction in controlling IC operations, such as setting too much headroom for a power transistor in power chips. As seen in figure 2.1, the chip is divided into five blocks which are named from u1 to U5. Each block will have one thermal sensor which is responsible to monitor the temperature of the whole block [12] [13] [14]. However, for the real heat running test, the heat will flow to the surrounding areas and the temperature of its self-block will be influenced by the neighboring blocks which makes the monitored temperature according to the pre-divided blocks useless. Therefore, the full chip with the fine spatial resolution is an urgent need.

2.1.2 The conventional thermal monitor system

Ideally, precision on-chip thermal sensors will be embedded in a die to realize full-chip temperature mapping that allows dynamically controlling IC operations, hence realizing full-chip thermal management at circuit block or transistor level. This general on-chip thermal sensing and management methodology requires a chip-scale thermal sensing network, comprising thermal sensors, readout and signal processing circuitry, and global

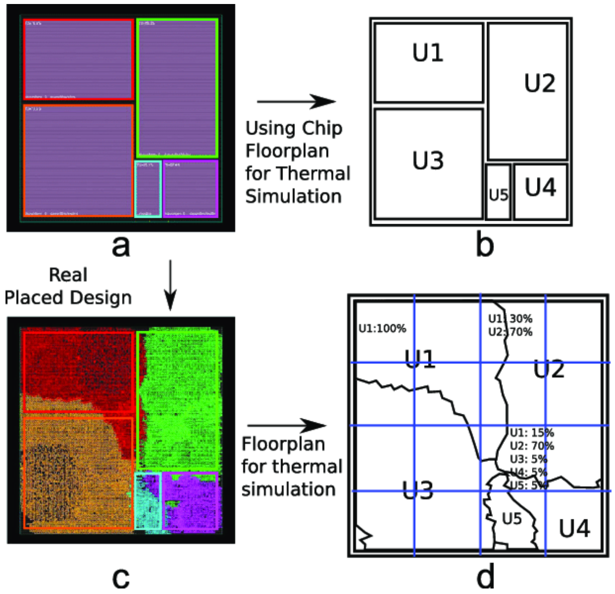


Figure 2.1 Illustration of the block level thermal sensors floorplan on the chip

thermal control circuitry, to achieve real-time thermal sensing with high spatial resolution, ideally down to transistor level. As such, an over-heated transistor may be tuned to cool down temporarily by reducing biasing and/or clock speed, per the thermal mapping and feedback controlling, to ensure lifetime and performance. Instead of using the on-existing thermal sensors on a block level, the high-resolution thermal map generation requires large numbers of the thermal sensors to get enough specific temperature data, close to the real heat source like the drain or channel of the target MOSFET, and not taking the design area. Figure 2.2 depicts the conceptual under-FET thermal sensor implemented by the in-hole diode [15]. The diode consists of the vertical PN junction inside the TSV-like deep shaft located in the substrate close to the MOSFET channel. The deep shaft is a high-aspect-ratioed hole created from the backside of a Si substrate analogous to a non-through-silicon TSV hole. Careful design will be excised to maintain a safety margin between the channel bottom and the sensor hole in order to avoid any adverse impact on MOSFET performance. This novelty structure has several advantages: First, since the in-hole diode is embedded in the substrate underneath the transistor, it will not consume any extra area, which is suitable to be placed in a large number to form the thermal mesh network covering the whole chip. Second, since the sensor is placed right underneath the channel of MOSFET to sense the self-heating, it can precisely identify the self-heating source and ensure high sensing accuracy due to the extremely close to the hot spot of the victim device. Third, the thermal sensor can be readily made upon the industry CMOS process. Fourth, the diode sensor can be naturally interconnected to the readout and control circuitry on the die to realize full-chip thermal mapping and management.

2.2 In-hole diode thermal sensor fabrication

The fabrication procedure starts with a 4-inch heavily doped P-type Si wafer (a resistivity < 0.0015Ω•cm via Boron doped). Ideally, a standard TSV-like process module would be used to fabricate the new in-hole diode sensors, i.e., to etch a high-aspect-ratioed TSV hole that will stop at the bottom of the active Si layer under N-/P-wells that host a MOSFET, followed by a PN diode sensor formed in the deep hole. Unfortunately, due to the lack of advanced TSV etching tool in our cleanroom facility, we have to develop an alternative process flow to fabricate the in-hole diode sensor on this work for demonstration, which can be readily modified and integrated into standard CMOS processes in the future. In the meanwhile, 2D TCAD simulation was used to optimize and guide the design and fabrication of in-hole diodes, critical to selecting the process recipes, in this work. Since this in-hole diode structure is high symmetric or centroid to the center line like a cylinder, the 3D TCAD simulation is not much improved compared to the 2D. What's more, 3D TCAD simulation is computing hungry and here the 2D TCAD is just a guideline for the fabrication instead of the accurate simulation.

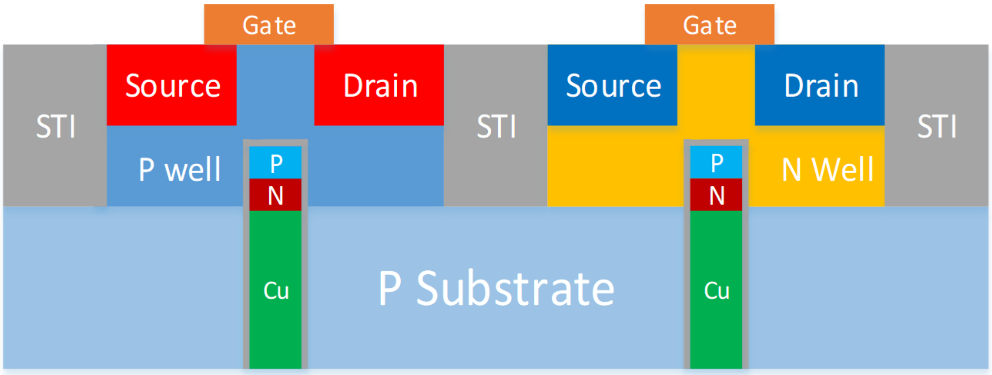


Figure 2.2 Illustration of the new in-hole diode thermal sensor concept in CMOS where the sensor is embedded in a deep hole in the Si substrate under individual MOSFET for transistor-level sensing resolution.

2.2.1 The deep silicon etching procedure

Because of the limitation in our cleanroom, if a deep TSV-like shaft would be created, we would not be able to etch off the SiO_2 layer at the bottom without etching off the SiO_2 layer on the side wall. Otherwise, we cannot form the in-hole PN diode. While, if the SiO2 on the side wall is etched off, it would cause the metal contact problem. To address this problem, we developed a deep ring hole approach as depicted in figure 2.3. Figure 2.3a shows a mask used to define the deep ring hole pattern and AZ4620 photoresistor (PR) is used for deep reactive ion etching (DRIE) protection. Figure 2.3b shows a zoom-in view for the etched deep ring hole by Bosch DRIE etcher. The etched deep ring hole is around 90µm to 100µm, similar to a typical TSV hole. A SiO_2 layer is grown via thermal oxidation to cover the inner side wall of the ring hole to isolate the in-hole diode from the surrounding Si as depicted in Figure 2.3b. Figure 2.3d gives a top view of the device area by optical microscope showing the deep ring hole and the in-hole diode area. In this figure, we can see the deep ring around the center silicon pole. Figure 2.3c is the SEM cross-section view of the side wall of the etched deep ring hole following the cut-line shown in Figure 2.3d, where the unique etching pattern on the side wall came from the Bosch DRIE processes of alternatively using SF_6 plasm bottom etching and C_4F_8 plasma side protection flow. A group of in-hole devices with different dimensions were designed with the outer ring diameters being 400µm, 600µm, 800µm, and 1000µm, and the inner ring diameters of 250µm, 400µm, 500µm, and 600µm, respectively. The etched deep ring hole has a width of 75µm to 200µm corresponding to the different dimensions like the diameters.

The next step is to create a TSV-like deep hole to host the in-hole PN diode. Since a thin SiO_2 layer was grown to cover all surfaces of the etched deep ring hole structure, grown in an oxidation furnace, we need to etch the SiO_2 above the center silicon pole to open the window for the followed long-time high-power DRIE follows to form the desired deep hole to host the in-hole PN diode. Because of the previous ring structure, the SiO_2 etching will be easy due to its location not being on the bottom of the TSV shaft. Since the top edge of the silicon sidewall is vertical, the photoresist is hardly hung along with the etch to provide enough thickness in order to protect the SiO_2 from etching off during the

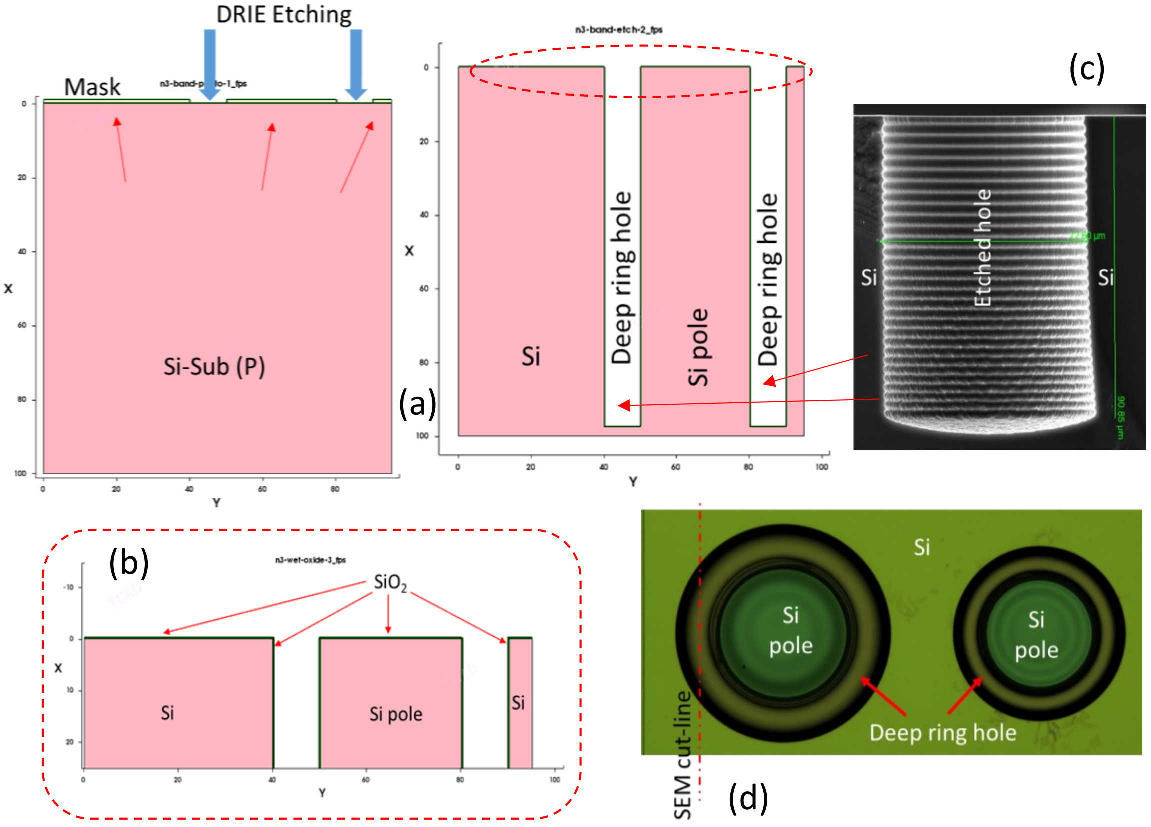


Figure 2.3 (a) Si substrate covered by a mask for DRIE etching to create a deep ring hole, (b) a zoom-in view for the deep ring hole etched by DRIE, covered by the required SiO_2 isolation layer, (c) SEM X-section view of the etched side wall of the deep ring hole, and (d) a top view of in-hole diodes by optical microscope.

long-time high power DRIE. Therefore, during this step, a combined PR/Al layer stack is used where the patterned Al layer serves as a hard mask to keep the shape of the top edge of the deep shaft because it has very high etching selective during the DRIE. Figure 2.4a is the TCAD simulation that shows the deep ring structure is patterned by the PR/Al mask. The cyan stands for aluminum and the light-yellow stands for the photoresist.

Figure 2.4b shows the optical image with the Al hard mask covering the outside of the deep ring hole including the corner with the center-left open for DRIE etching. A very thick PR (positive photoresist AZ4601) layer, filling the deep ring hole, serves to protect the SiO_2 layer that covers the inner wall surface of the deep hole during DRIE etching, which is fairly difficult to handle because of requiring careful craftsman's work for PR baking, exposure and developing to the ensure the fine patterns. A mask is used to define the opening of the deep shaft area, which will be etched to the bottom by DRIE etching to form the required high-aspect-ratioed deep shaft shown in figure 2.4c. Then the mask will

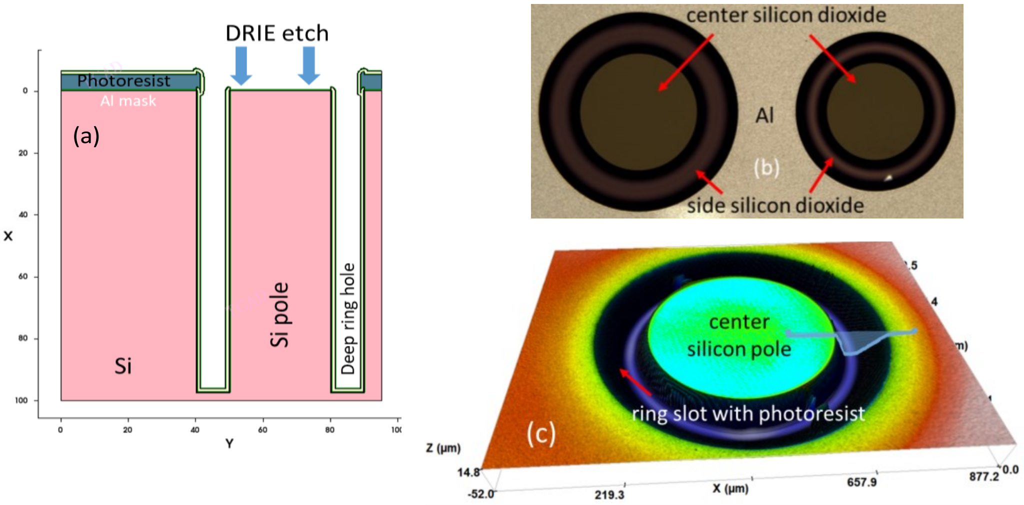


Figure 2.4 (a) TCAD simulation shows the deep hole using a hard Al-mask to protect the top edges of the hole, (b) confocal microscope 3D image of ring structure with photoresist covering the SiO_2 side wall (c) using Al as the hard mask to protect the corner from being etched.

be etched to the bottom by first the normally SiO_2 above the silicon pole etching followed by DRIE etching to form the required high-aspect-ratioed deep shaft as shown in figure 2.5a. After the etching is finished, we will strip the photoresist and then wet etch off the aluminum hard mask to show the bottom silicon exposed and with the side wall still covered by the SiO_2 insulator. Figure 2.5b shows the whole image after the second DRIE etching where the center part looks gray, which is the silicon opening not covered by SiO_2. To this step, we can get the desired TSV structure with half P part of the PN junction exposed at the bottom of the deep shaft.

2.2.2 PN junction formation

The next step is to form the PN junction at the bottom of the deep hole. An intrinsic poly-Si layer of about 0.3µm was deposited onto the bottom of the deep shaft as illustrated in Figure 2.6a. From figure 2.6a, we can see the polysilicon above the bottom silicon pole, attached to the sidewall, and above the top plan outside the deep shaft. It is noted that we only need the center part of the polysilicon to be used from the PN junction, therefore we need to etch the redundant polysilicon attached to the sidewall and above the top plane.

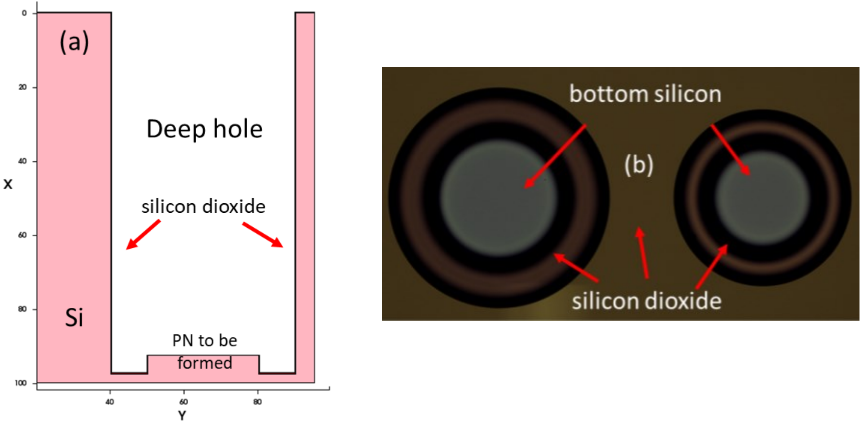


Figure 2.5 (a) using thick PR to define the deep hole pattern for etching by TCAD simulation (b) after DRIE etching, the center bottom silicon is exposed without SiO_2 covering.

Similarly, we need to use the thick PR (negative photoresist NR1000PY) layer to cover the center bottom polysilicon to protect it from being etched in the next step. Figure 2.6b shows the thick PR above the center bottom polysilicon. Then after the next isotropic over-etching, the polysilicon will be fully etched till stopping at the SiO_2 layer. Then only the center bottom of the polysilicon will remain.

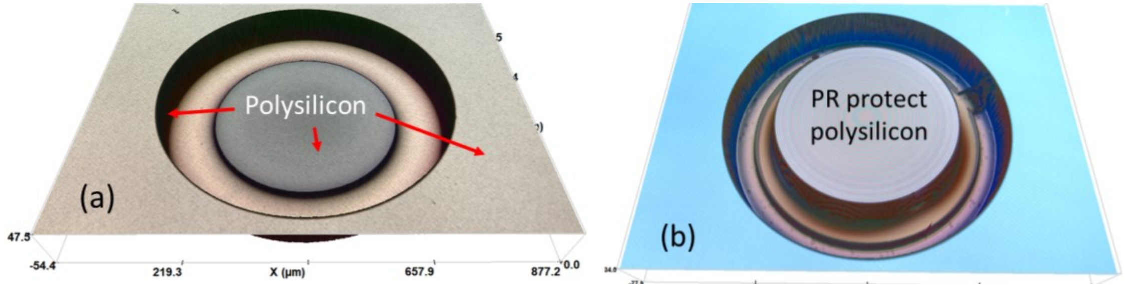


Figure 2.6 (a) image for polysilicon deposition around the bottom, side wall and the top plane, (b) negative photoresist protects the polysilicon above center bottom pole silicon from being etched.

After this procedure, N-doping of phosphorous is implanted into the polysilicon convex region. Of course, the PR still needs to be used to open the implantation window. Careful TCAD simulation was conducted for implantation and annealing to optimize the poly-Si/Si PN diode. Since the PN junction is the core part of this structure, we should delicately modify the implanting dosage and energy. It is known that the energy of the doping will influence the ion accumulation depth and the dosage will influence the junction line shapes. Therefore, according to Figures 2.7a and 2.7b, we can find that the optimized implantation recipe was 1x10^{13} cm^{-3} for dosage and 30ekV for energy. Under this condition, it can form a good PN junction. As seen from figure 2.8c, when we change the dosage un or down 10keV like 20keV and 40keV, the ratio of the P or N part close to the PN junction will change. The same thing for the dosage, if we change it up or down 10 times, the junction line shape will become either too large or too small, which is not a good junction.

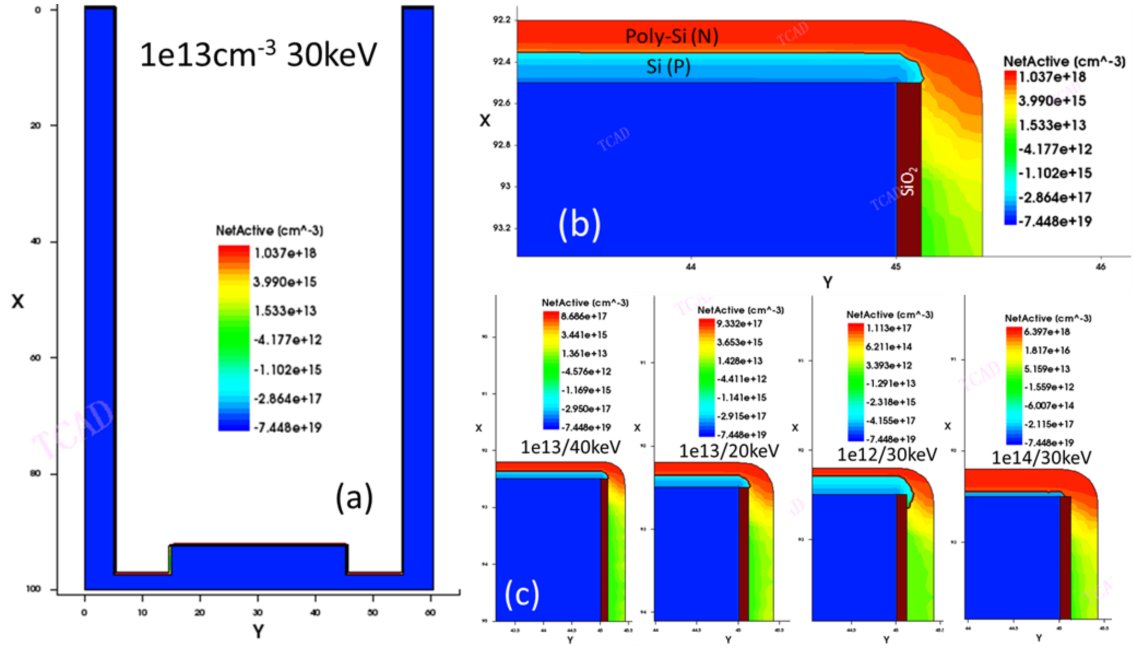


Figure 2.7 TCAD process simulation was used to optimize the deep DRIE etching process: (a) the high-aspect-ratioed deep shaft to be formed, (b) a poly-Si/Si PN diode to be formed at the bottom, and (c) different doping recipes to optimize the PN junctions.

After implanting, we need to lift off the PR and send it into the furnace with the nitrogen gas filling in for annealing. The annealing condition can still be simulated in the TCAD to maximize activate the implanted ions and repair the impaired lattice. Only after this, the PN junction can be formed and right in the center bottom polysilicon-silicon. Figure 2.8a shows the 3D image of the in-hole diode with the polysilicon-single silicon PN junction. We can find that the depth of the TSV-like hole is about 100μm with the bottom gray material which is the phosphorus-doped polysilicon. Also, we make an SEM image to see the cross-section view of the in-hole diode with the zoom-in PN junction shown in Figures 2.8b and 2.8c respectively. From the SEM image, we can see the vertical sidewall of the in-hole diode and the depth of this shaft. What's more, zoom in on the bottom part, we can see the silicon and polysilicon interface which identify the PN junction.

2.3 In-hole diode characterization

2.3.1 Thermal property validated by TCAD simulation

The last step is the interconnection part. After the PN junction formation, we deposit the seed layer like TSV fabrication to form a thin layer of metal used for later electroplating. Here we use the isotropic physical vapor deposition (PVD) to deposit the TiW/Cu along the sidewall and the bottom. Then using electroplating fills copper into the deep shaft to form the interconnect from the in-hole diode to pad. Figure 2.9a shows the in-hole structure after the Cu is filled. We will set the top of the copper pillar as one of the pads and the bottom silicon as the other pad. Therefore, the simulated current flow lines via TCAD from

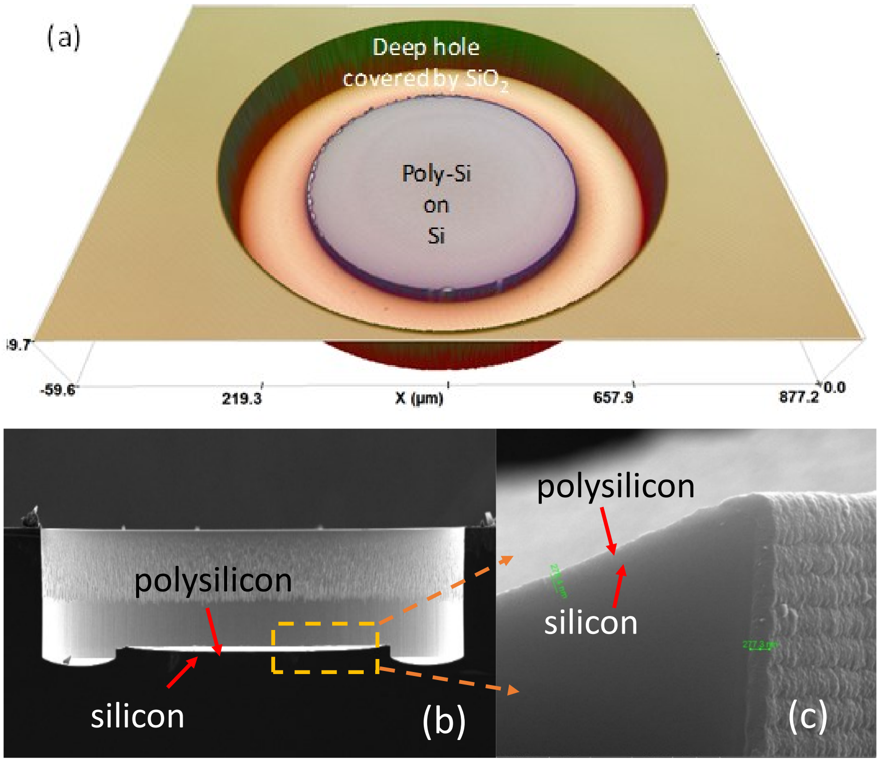


Figure 2.8 Images for the poly-Si/Si PN diode sensor formed at the bottom of the deep shaft under a MOSFET in the Si substrate: (a) a 3D image by confocal microscope, (b) a cross-section view of the PN junction by SEM, and (c) zoom-in of the poly-Si/Si junction corresponding to the yellow box region in (b).

the anode to the cathode in the PN diode formed, which is depicted in figure 2.9b. We can find the hot spots accumulate around the PN junction which validates our design. The current flow line is vertical to the PN junction and from the top pad to the bottom one.

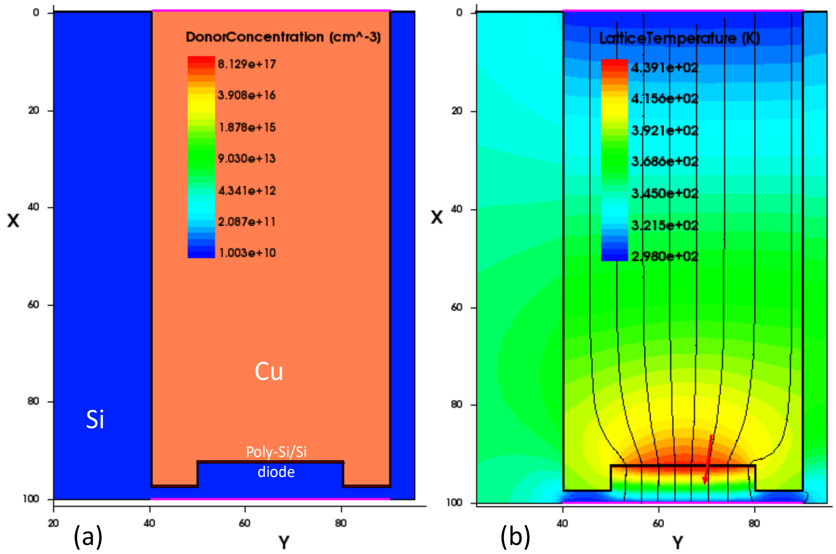


Figure 2.9 TCAD simulation shows (a) the vertical poly-Si/Si PN diode at the bottom of the deep shaft, and (b) the I-V curves of the diode sensor.

Figure 2.10 shows the simulated I-V-T curves for an in-hole diode with a diameter of 30μm. It clearly shows that the I-V curves are very sensitive to temperature variation across the 10°C to 120°C range, indicating a good thermal sensing feature for the in-hole diode proposed, with a temperature coefficient (TC) of about 1.25mV/°C across the 0.6V to 0.9V biasing range. When the temperature increases, we can find that with the fixed voltage, the current will increase either and with the fixed current, the voltage will decrease.

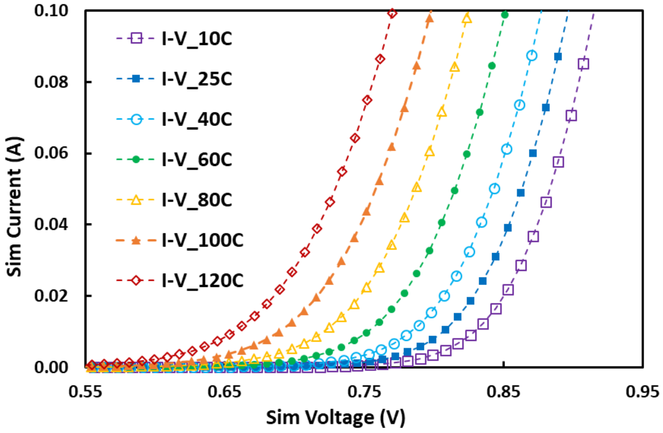


Figure 2.10 TCAD simulation shows desired I-V-T behaviors for the new in-hole diode thermal sensor across a wide temperature range.

2.3.2 Thermal property validated by I-V-T testing

With the aid of the instrument of Keysight precision semiconductor parameter analyzer and ATT thermal chuck system, comprehensive DC I-V sweeping tests were then conducted using a testing set-up comprising a Cascade probe station with a thermal chamber. In case of burning down the thermal sensor, we set the current compliant of 1μA by sweeping the voltage from 0 to 10V. First, on the ascending temperature from 10°C to 120°C, and then on the descending temperature from 120°C to 10°C, we can find the current

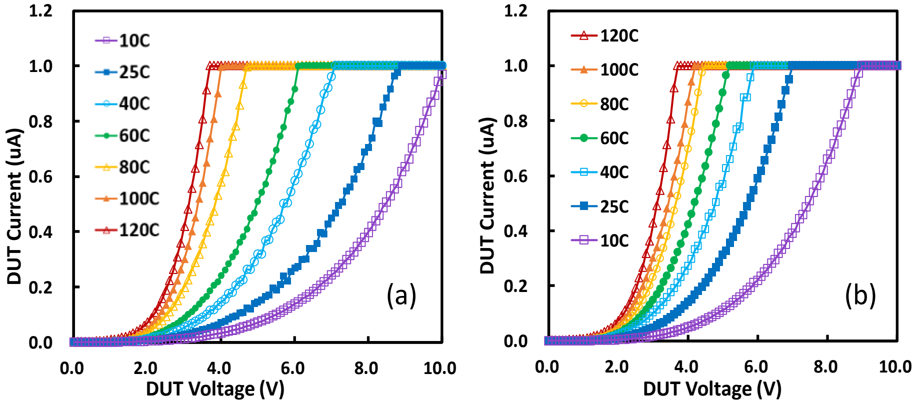


Figure 2.11 Measured I-V-T curves for a sample in-hole diode sensor with a diameter of 250μm: (a) I-V curves with ascending temperature sweeping, and (b) I-V curves with descending temperature sweeping.

changes conforming to the law. With the fixed current, the voltage decreases with the temperature increasing. In turn, with the fixed voltage, the current increases with temperature increasing, which is the desired monotonous I-V-T behavior. Figure 2.11 presents the measured I-V-T curves for a sample in-hole diode with a diameter of 250μm. are readily observed.

A repeated DC sweeping test about 11 times was done for a sample of in-hole diode with a diameter of 400μm illustrated in figure 2.12a. We can find that the I-V curve is very stable during the 11 times DC test under the compliant current of 1μA from 0V to 8V under the room temperature. Figure 2.12b depicts the extracted I-T (at a given voltage bias) and V-T (at a given current level) curves for a sample in-hole diode with a dimension of 250μm. Because the desired monotonous I-V-T behaviors are readily observed, the fabricated in-hole diodes can be confirmed to function well as a thermal sensor. We have done an amount of measurement of samples to validate the novel concept of the proposed under-FET thermal sensor. However, we can find some unperfect issues residing in the testing data

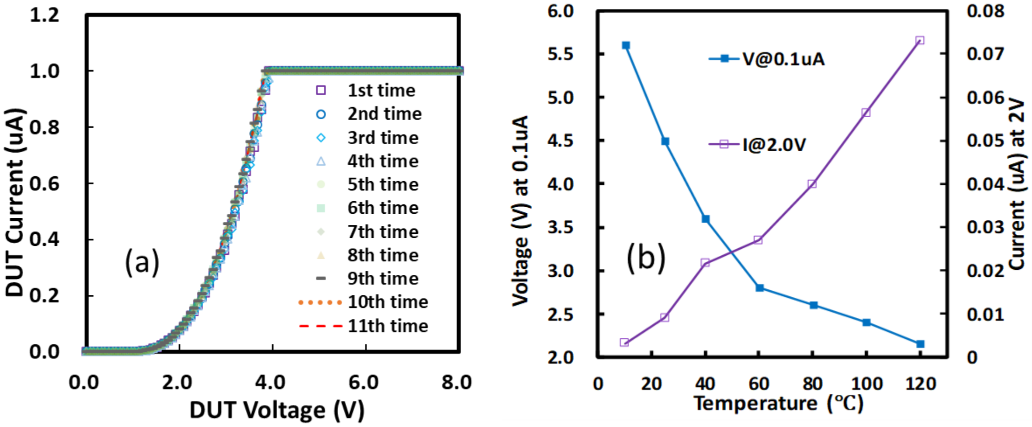


Figure 2.12 (a) Measurement shows that a sample in-hole diode sensor with a diameter of 400μm is stable after repeated DC sweeping tests, (b) Measured V-T and I-T curves show monotonous trend against temperature that is required for a diode thermal sensor.

like the asymmetry in sensor characteristics for the prototype devices, which can be attributed to several factors such as polysilicon quality, the thermal residual effect due to the thermal stressing routines and device asymmetry, etc. Improving the performance of the under-FET thermal sensor is limited by the facilities of the cleanroom.

2.4 Run-time thermal monitoring system

2.4.1 Thermal management topology

The novel under-FET thermal sensor is a critical enabler for realizing run-time full-chip smart thermal management of complex chips made in advanced IC technology nodes. It is necessary to have both accurate and comprehensive time-dependent electrothermal modeling for the under-FET sensors and victim devices for achieving the IC design goal. Through systematic and statistical device characterization, the compactor behavior models can be developed to accurately map the device thermal failure characteristics (e.g., temperature variation, degradation due to thermal aging, lifetime reduction, thermal runaway, etc.) with device functional parameters (e.g., current, voltage, power, frequency, etc.) in the time domain. Through these models, we can monitor and control the ICs for real-time thermal management. One of the applications is to build the sensors along with the time-variant electrothermal models integrated into the cell libraries to enable accurate full-chip thermal validation in the post-simulation phase to pursue a more reliable design goal. Another application is to realize the full-chip thermal mapping focusing on each transistor to assist thermal failure analysis. Among these applications, the main one is to embed the sensors underneath the die to realize full-chip thermal management with spatial thermal monitoring map down to transistor level.

In a nutshell, the in-die thermal management principle is straightforward. Trough the sensor underneath the victim devices to acquire the accurate temperature sensing data neat the heat generation source during the operations. The run-time sensing data will be compared with a pre-developed F-I-V-T-t model for the sensor-victim pair (including mutual coupling effects with neighboring devices on a chip, where "F" denotes the complex thermal failure characteristics (e.g., aging, lifetime, performance degradation, and thermal runaway, etc.). In terms of the closed-loop feedback control circuitry, the thermal sensor will initiatively trigger a control action to governor the risky victim whenever the temperature rises beyond the threshold per observation from the under-FET sensor [16]. As shown in figure 2.13, the simple thermal management action may include temporarily reducing the biasing voltage from the low voltage regulator (LDO) or scaling down the driving current from the current mirror. What's more, using the VCO to change the divider so as to modify the output frequency or changing the load capacitance to decrease the power is also the method to decrease the power consumption to control the over-heated victim devices.

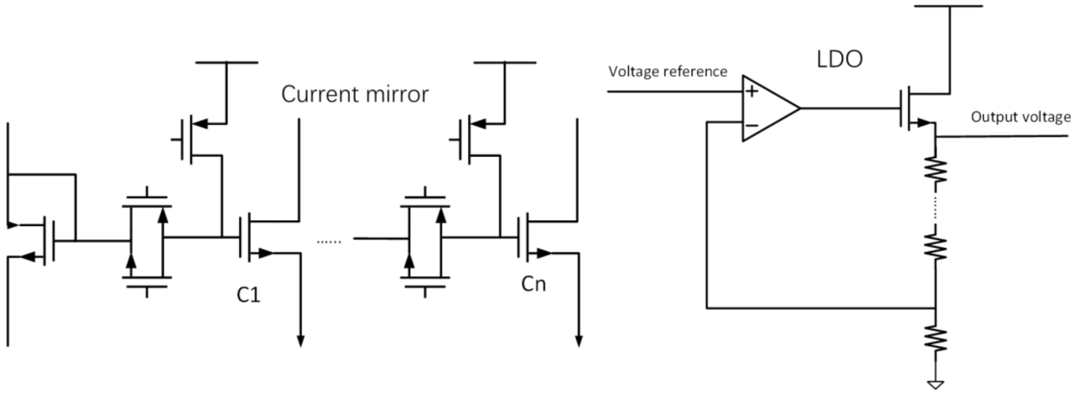


Figure 2.13 Schematic of LDO adjusting the bias voltage and current mirror scaling the driving current

Figure. 2.14 depicts an exemplar functional diagram for a thermal management chip with an embedded large in-die thermal sensing mesh network comprising the new under-FET sensors. Since a large number of in-die thermal sensors are used, practically, a MUX (multiplexer) is needed to stream in the parallel incoming thermal sensing data. A transimpedance amplifier (TIA) will amplify and convert the weak current signals generated by the under-FET sensors into voltage signals for control. If complex thermal management is needed for sophisticated large chips, single-end to double-end (S2D) circuits and high-speed ADC will be needed. Embedded memories and digital signal processor (DSP) blocks may be used to facilitate machine learning (ML) to realize smart thermal management to enable predictive thermal management actions instead of conventional reactive thermal controlling practices. Fine-resolution (transistor level) real-time thermal mapping can be obtained using a large sensor mesh, which is entirely different from the existing block-level thermal mapping method using traditional co-planar sensors.

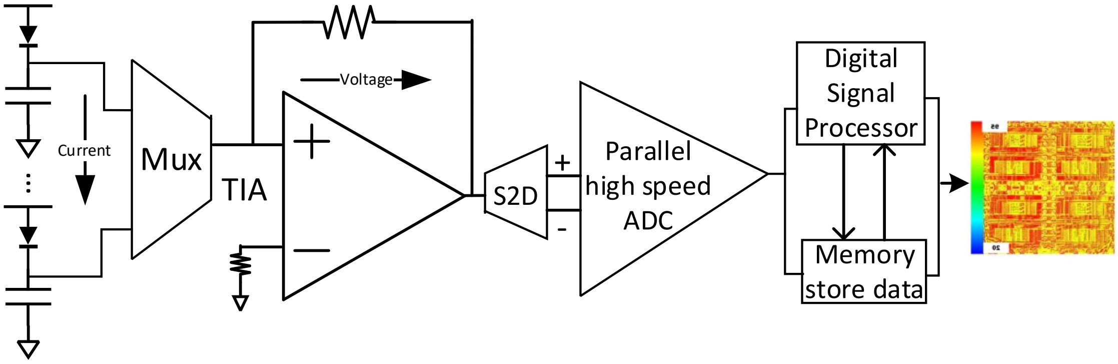


Figure 2.14 A block diagram for thermal management using under-FET thermal sensor mesh network.

2.4.2 Exemplar thermal monitoring blocks for thermal management

From figure 2.15, we can find the schematic of the thermal sensor monitoring and controlling circuits as an example for thermal management. The core circuit contains the TIA plus the temperature sensing diode, comparator circuits as a simple ADC, and a power domain transferring buffer. TIA is used to convert the current into the voltage signals which is beneficial to the later threshold decision. Then the voltage signal transfers into the comparator which is used as a simple ADC here for threshold judgment. If the sensing voltage is smaller than the reference voltage, the comparator will produce a signal to control the victim device releasing from the over-heated situation. As we can know that in order to increase the sensitivity of the in-hole diode, we build the sensing circuits within the 3.3V power domain. However, for the many victims operating blocks, the normal power supply voltage is 1.1V. So, we need a level shifter to transfer the voltage from the 3.3V domain into the 1.1V domain and used it as a buffer to drive the control signal of other operating blocks.

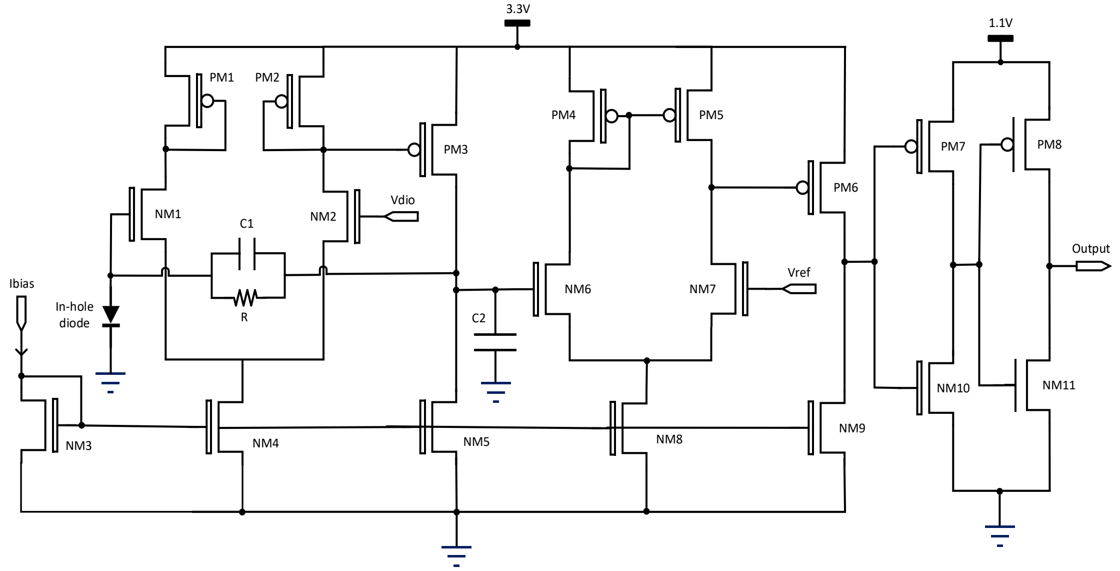


Figure 2.15 Brief schematic of the temperature sensing and feedback control circuits with in-hole diode.

The signal produced by the under-FET thermal sensors converts the temperature signal into the current signal with the fixed voltage bias. Using the analog-A behavior model, we can convert the previous testing data into the simulation needed format and build a temperature sensor model in the library. The model is still based on the I-V curve by sweeping the voltage from 0V to 10V with the compliant current of 1µA. Besides, it contains a set of I-V curves based on the temperature ranging from 10°C to 120°C. We can see that the Spectre simulation curve matches the testing curve very well except the current reaching 1µA limit because of the not enough sampling near the current limit especially for the large current with the same fixed voltage.

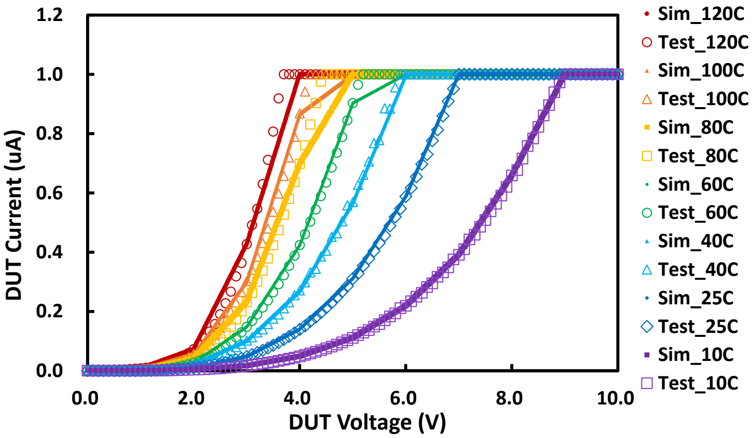


Figure 2.16 In-hole sensor behavior model developed using the test data is validated by Spectre simulation.

Here, we use the TSMC 40nm ultra-low power process to build this circuit. First, the circuit needs a current bias from the bandgap circuit. NM3, NM4, and NM5 constitute the current mirror to provide the tail current for the first stage and the second stage of the TIA amplifier. NM1 and NM2 are the differential input and PM1 and PM2 are the diode-connected active load to increase the gain of the first stage. Normally, the first stage of the gain is not enough, we need the second stage of the PM3 to further increase the gain to a relatively large level about at least 30dB above. We add the C1 and C2 capacitance as the compensation to keep the stability over the wideband. Figure 2.17a depicts the TIA stability is good until about 4MHz with a phase margin of about 90 degrees. The gain curve and phase curve are both under the post-simulation over all the PVT corners. Figure 2.17b displays the post-simulation results for the TIA block showing the required power supply reject ratio (PSRR) of about -50dB below 1MHz and a low noise level of less than -90dBV/$\sqrt{Hz}$ overall PVT corners across a temperature range from -40°C and 125°C. The low PSRR of ~3mV and noise level of ~0.1mV obtained in linear scale are much lower than the desired 30mV output voltage resolution, confirming that the TIA designed is

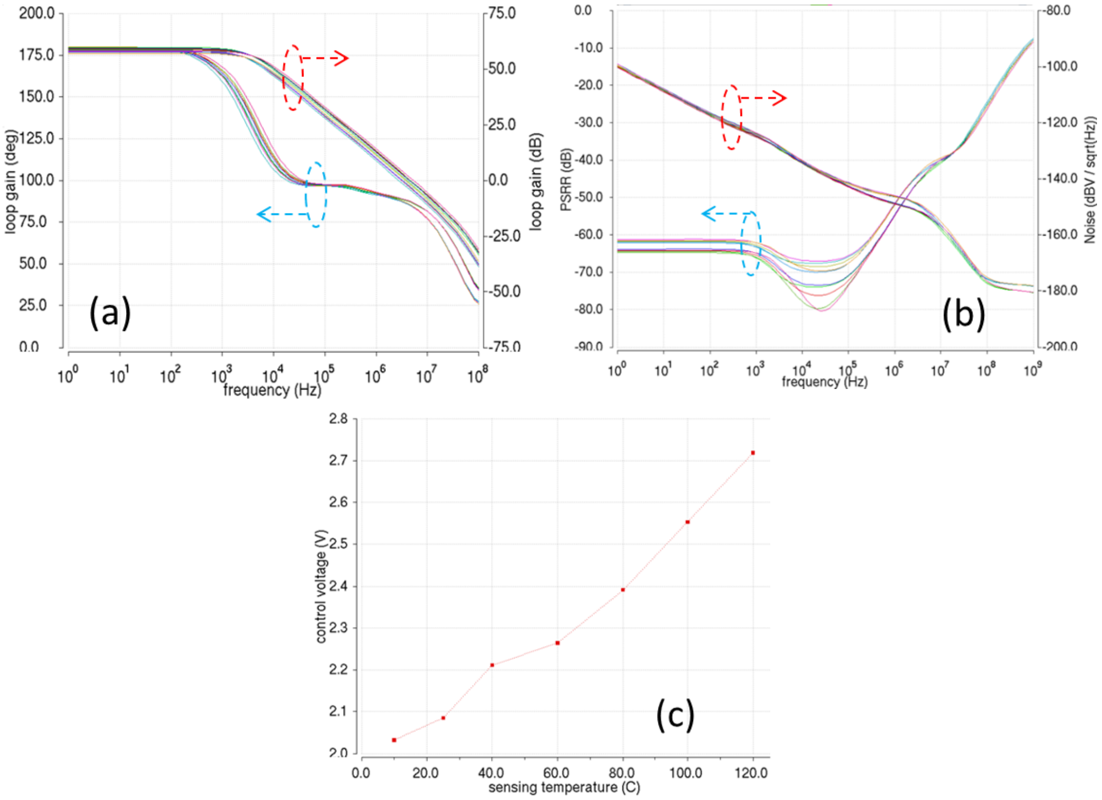


Figure 2.17 Post simulation for TIA shows: (a) good stability across the frequency from 1Hz to 100MHz over all corners, (b) low PSRR and noises across frequency from 1Hz to 1GHz over all corners (-40°C to-125°C), and (c) desired monotonous temperature sensing and voltage controlling for thermal management.

noise-immune for the in-hole thermal sensors. Figure 2.17c shows the desired monotonous relationship between the input sensing temperature and output control voltage from the TIA. We set the reference voltage with 2V and set the different ambient temperatures from 10℃ to 120℃ like the simulation split. We can find that the output voltage increases with the temperature increasing just like what we designed before from 2.03V to 2.73V.

In terms of a comparator, the NM3, NM8, and NM9 constitute the current mirror-like TIA to provide the tail current. PM4 and PM5 are the current mirrors as the active load to provide a very large gain. Combining the second stage PM6, these circuits can provide at least 50dB above gain. Remember we need to choose the long length MOSFET to decrease the offset. We will use several comparator circuits with different Vref voltages from 2.0V to 2.8V as a simple ADC to get the digital output. Remember that, control signal should be within the power domain of 1.1V, therefore we use PM7 and NM10 as the high voltage inverter combined with PM8 and NM11 as the low voltage inverter to constitute the buffer to output the control signal, which realize transferring the 3.3V power domain into 1.1V power domain. Figure 2.18a gives the post-simulation results for the comparator block confirming its functions. Fig. 16a shows PSRR of -40dB and low noise of -90dBV/$\sqrt{Hz}$ overall PVT corners across a temperature range from -40℃ and 125℃. The low PSRR of ~10mV and noise level of ~0.1mV obtained in linear scale are much lower than the desired 30mV output voltage resolution, confirming the comparator block designed is noise-immune for the in-hole sensors. Figure 2.18b confirms the desired flipping output control signals within 90ns driven by the in-hole sensing data overall PVT corners across a temperature range from -40°C and 125°C. With these two blocks, we can get a quick response and a larger driving ability.

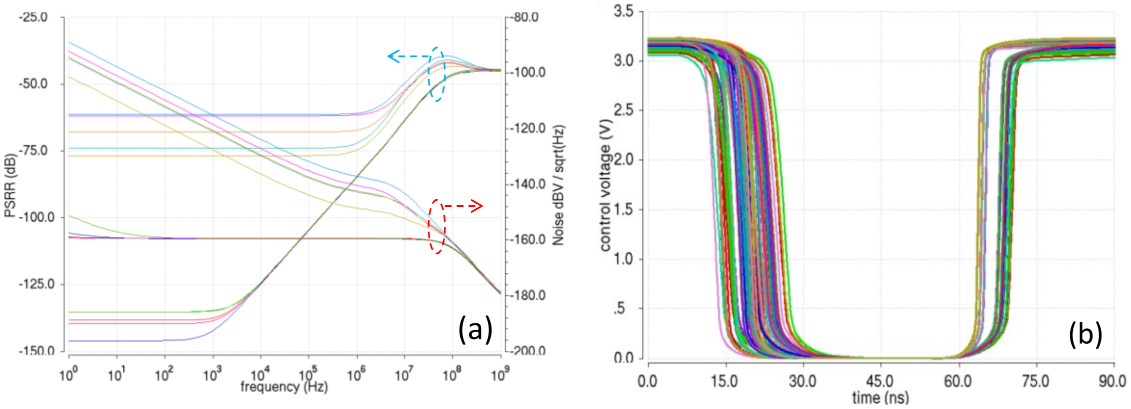


Figure 2.18 Post simulation for comparator block shows: (a) low PSRR and noises from 1Hz to 1GHz over all corners across -40°C and 125°C temperature range, and (b) required flipping output control signals (-40°C and 125°C) driven by the in-hole sensing data for PA operation scaling against self-heating.

2.4.3 Exemplar victim modules and machine learning strategy

The under-FET based thermal management method is validated using a multi-PA wireless module, comprising one Bluetooth PA and two Wi-Fi PAs, designed in TSMC 40nm ultra-low power CMOS where PA transistors are livelily monitored by the new under-FET in-hole diode sensors. Each under-FET in-hole sensor delivers rum-time current signals generated by temperature variation of the overtop victim PA MOSFET. Circuit simulation of the PA chip was conducted using the Verilog-A behavior models for the in-hole sensors from measurement to validate the thermal management functions. Figure 2.19a is the layout of the monitoring circuit with one TIA and three comparators. We put the capacitor above the resistor to save the area and add a dummy for each MOSFET and resistor to decrease the mismatch. Many layout tips like symmetric placement, a signal cross line can get the lowest offset to increase the sensing accuracy.

Figure 2.19b is the layout combining the monitoring circuits and the target PA circuits showing a small area budget for the sensing circuitry compared to the core circuit under monitoring. Also, one monitoring circuit can control many target blocks, so it will not take much area from the die.

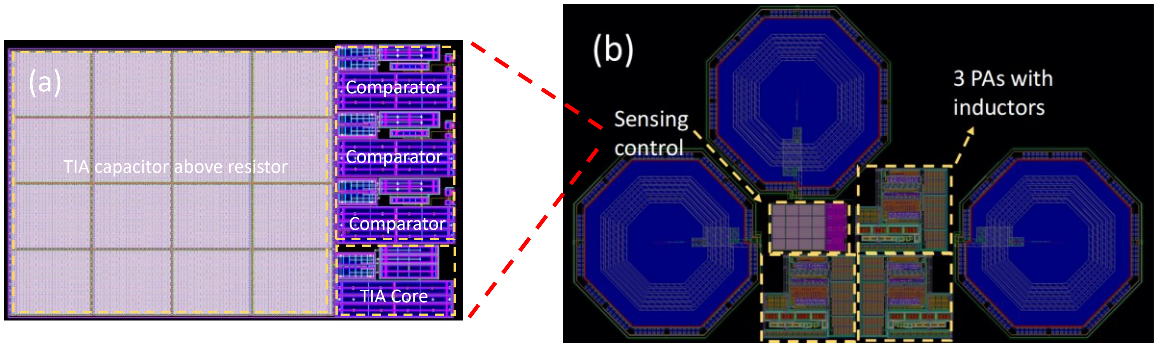


Figure 2.19 (a) Sensing control layout of monitoring circuit layout including TIA and 3 comparators zoomed in view from PA module layout (b) Layout for the PA module with embedded under-FET in-hole thermal sensors designed in 40nm CMOS.

Figure 2.20 shows the class E PA performance managed by the control voltage signals in response to the in-hole temperature sensor underneath the PA. It is readily observed that the PA self-heating can be dynamically monitored and controlled. A pre-set and/or field-programmable overheating threshold triggers the PA operation-scaling routines in run-time through temperature-sensitive feedback (from normal 650mV downscaling to reduced 350mV in this demo). The precise and efficient PA operation scaling is enabled by accurate temperature sensing by the under-PA in-hole thermal sensor.

This prototype demo circuit, though simple, validates that the novel under-FET in-hole temperature sensor can be used to enable full-chip run-time smart thermal management with self-learning, which opens a door for next-generation ML-enabled chip-scale thermal management. Machine learning can optimize the run-time thermal management to get the

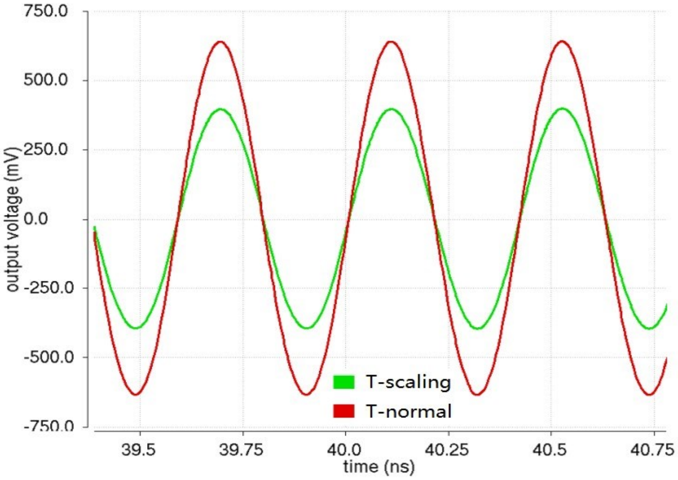


Figure 2.20 Simulation for the PA module confirms the thermal management function through temperature-based PA performance scaling controlled by varying voltage biasing tuned by the under-PA in-hole temperature sensors, hence mitigating thermal impacts in real time. T-normal: PA operation under normal temperature; T-scaling: PA operation scaled down triggered by overheating sensed.

smart full-chip thermal map. Thermal sensing data for the whole chip in real-time is a heavy load for the DSP and memories which will consume a lot of power for the computing and a large amount of storage space. However, carefully think, not every block or transistor has the quickly changed temperature simultaneously or very quickly. In order to save the memory and decrease the processor load, we can utilize machine learning to dynamically adjust the refresh rate for each thermal sensor according to its location and task situation based on the previous data training. For example, some area does not produce much heat during some operating conditions, thus we can decrease the refresh data. In turn, we can increase the refresh rate when facing a boosting situation. Meanwhile, we can know that some blocks are always under the low power consumption situation and thus we can always set the refresh rate very low to save the computing resource, which we call it smart partition. Another application for the ML is to predict the temperature trend for controlling. Use ML to timely predict temperature trends including changing speed, temperature maximum, and spot location which assist to find the hot area and give control signal to decrease the power consumption ahead. Give the control signal ahead to decrease the hot FET to eliminate the high temperature via decreasing its power consumption. Also, a delicate feedback control method based on ML can minimize functional impairment. This novelty under-FET thermal management opens a door for ML-enabled real-time full-chip smart thermal management for future ICs.

Chapter 3 Introduction to the ESD protection and graphene

3.1 ESD protection background

ESD failure is the main issue on the design reliability for the semiconductor integrated circuits (ICs) field. The induced ESD damage accounts for 35%-50% of all the ICs failures resulting in lots of billions of dollars lost each year and becoming the thorny problem for the IC industry [17]. As with the development of the technology node shrinking down, the thin gate oxide and low drain break down voltage makes the protected circuits more vulnerable to the surged ESD. Since the MOSFET structure has moved into the SOI and FinFET stage, due to the severe self-heating effect, the ESD current handleability faces a big challenge. Also, the BEOL of the advanced technology makes the lower metal resistance large which needs careful consideration during the ESD design. In the meantime, the high frequency circuits like 5G communication circuits designed at multi-gigahertz and high-speed circuits like serializer/deserializer (SerDes) at tens gigabits per second are very sensitive to the ESD induced capacitance which may dramatically degrade the performance of the circuit [18] [19] [20].

3.1.1 ESD characterization and standard models

ESD is an extremely fast electrostatic discharging phenomenon between two closed objects each with the opposite polarity charges. The instant surging pulse may have as high as 40mA current and 20kV voltage according to the different ESD standards. Large current will normally cause the device thermal breakdown and large voltage will result in the dielectric being punched through. High protection level of ESD device normally means the large area which will consume the treasured design area and large parasitic capacitance impacting the high frequency and high-speed circuits performance. This paradox needs careful design and validation via silicon testing data and then optimizing the design. These procedures may be repeated several times which makes the ESD design time increase a lot shown in figure 3.1. The figure also shows the different percentages of ESD failure during the different scenarios such as device failure, BEOL failure, ESD network failure, and cross-domain failure. In the latter chapters, we mainly propose the new structure of the ESD device aimed to improve the ESD performance to overcome the device failure. For the rest three factors, we need to focus on the ESD floor plan and be skilled in the combination of different ESD devices.

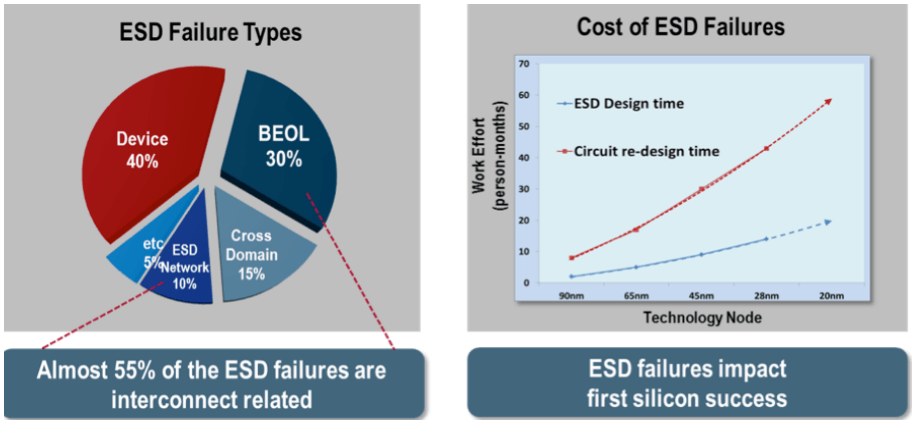


Figure 3.1 ESD failure statics and ESD design time

According to the discharging happening situation, we normally divide into three models named human body model (HBM), machine model (MM), and charged device model (CDM) for on-chip ESD protection. As shown in figure 3.2, HBM stands for the die touched by a human being during the process or handling which has a large instant voltage and is induced from outside. Normally, HBM equivalent circuits mainly consist of a 1.5kΩ resistor in series with a 100pF capacitor with other parasitic components and its surging pulse duration consists of 17-22ns for rising time and 150ns for the decay time [21]. MM

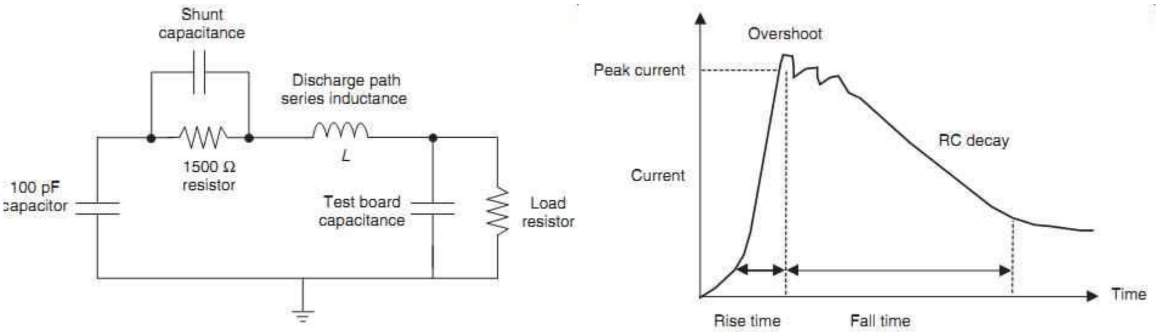


Figure 3.2 HBM equivalent circuit model with parasitic and pulse waveform

stands for the die touched by a machine during process or handling during packaging. Normally MM equivalent circuits consist of a 200pF capacitor with the parasitic components and its waveform normally faster than HBM pus the oscillation switching from the positive to the negative. It is noted that MM pulse has lower instant voltage but higher instant current than the HBM waveform shown in figure 3.3. CDM stands for die's capacitance charging and discharging during contact with the environment. Normally CDM equivalent circuits is a little complicated because its event occurs from inside. In the equivalent circuits, the primary capacitance forms between the device under test (DUT) and the field charging plate. The rest capacitances are the capacitance between DUT and ground plane as well as the field charging plate and ground plane. During the discharging process, the discharge is initiated by the pogo pin and flows through other parasitic inductance and resistance shown in figure 3.4 [22]. It is noted that CDM waveform has the

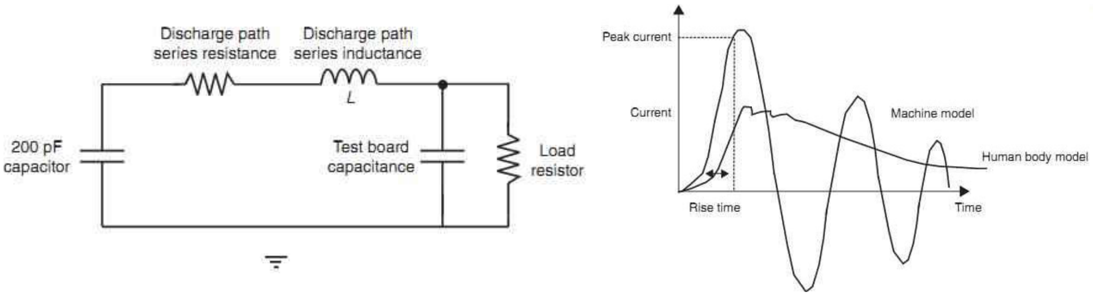


Figure 3.3 MM equivalent circuit model with parasitic and pulse waveform

fastest response with rising time 250ps and oscillation fast damping to zero which has the highest instant current compared with the HBM and MM.

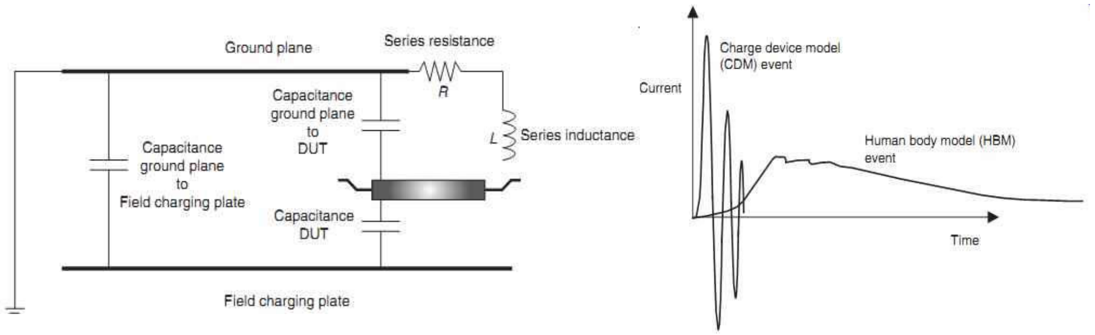


Figure 3.4 CDM equivalent circuit model with parasitic and pulse

3.1.2 ESD testing methods and design window

To characterize the ESD device performance, we normally use the transmission line pulse (TLP) instrument to characterize the properties of the ESD device under HBM

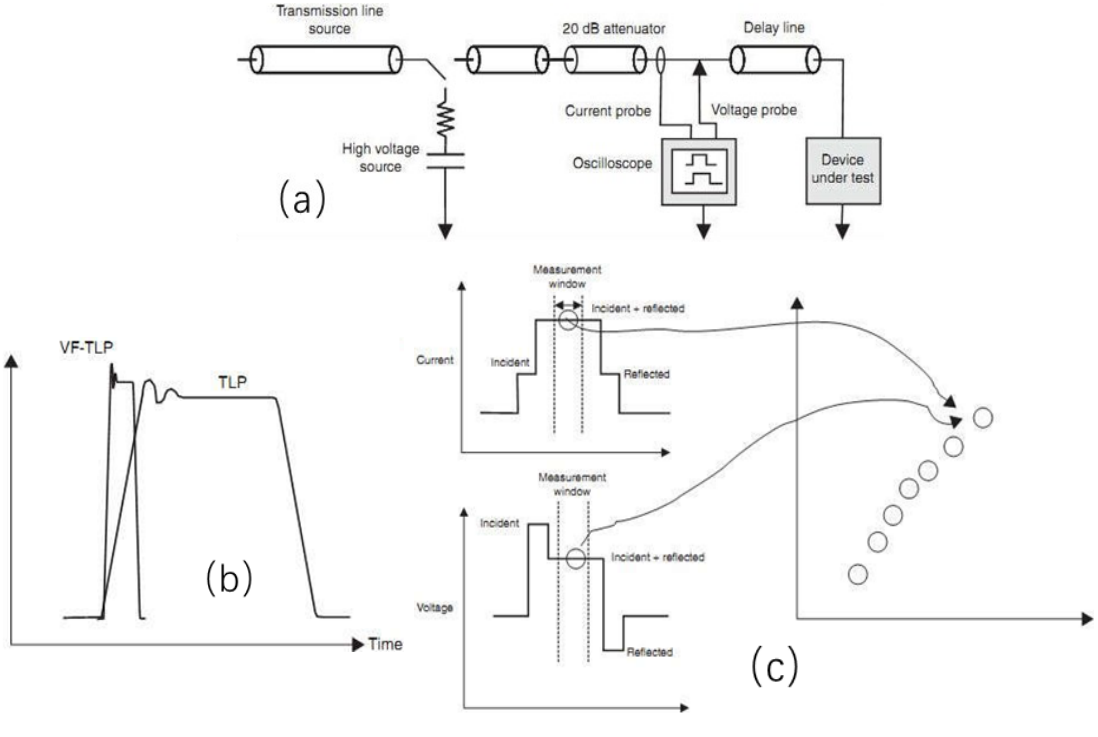


Figure 3.5 (a) TLP/VFTLP time domain reflectometry configuration (b) comparison between TLP and VFTLP waveform (c) TDR results for current and voltage

zapping [23]. The purpose of the TLP method is to build a test that can evaluate the repeatability and reproducibility of the devices via a defined pulse in order to characterize the ESD critical parameters. TLP uses the transmission line to charge and discharge to the DUT and analyze the incident and reflected waveform. Under the quasi-static integration between the 20%-70% of the added waveform from the incident and reflection, it can give the voltage and current spot under one charging and discharging event. After a series of higher and higher charging and discharging events, it can depict the I-V curve. It is noted that in order to mimic the HBM waveform, the square TLP waveform is formed with a duration of 100ns and a rise time of 10ns. Very fast transmission line pulse (VFTLP) is another similar instrument to mimic the CDM waveform [24]. Known from the name, the biggest difference between the TLP and VFTLP is that the VFTLP has the faster square pulse with a duration of 1ns and rise time of 100ps to mimic the CDM zapping. Figure 3.5a depicts the TLP/VFTLP time domain reflectometry configuration, figure 3.5b shows the TLP and VFTLP square-like pulse, and figure 3.5c displays the measurement methods for current and voltage. Last but not least is that TLP and VFTLP are non-destructive testing methods and can reveal more details than the HBM or CDM zapping such as the triggering voltage (V_{t1}), holding voltage (V_h), on-resistance (R_{on}), the breakdown current (I_{t2}) and the breakdown voltage (V_{t2}).

ESD design window is the restricted critical parameters range commonly defined by ESD rules and circuits design rules displayed in figure 3.6. There are mainly two types of ESD devices: one is the non-snackback device like a diode, diode string, and RC clamp; the other is the snackback device like gate ground NMOS (ggNMOS), silicon control

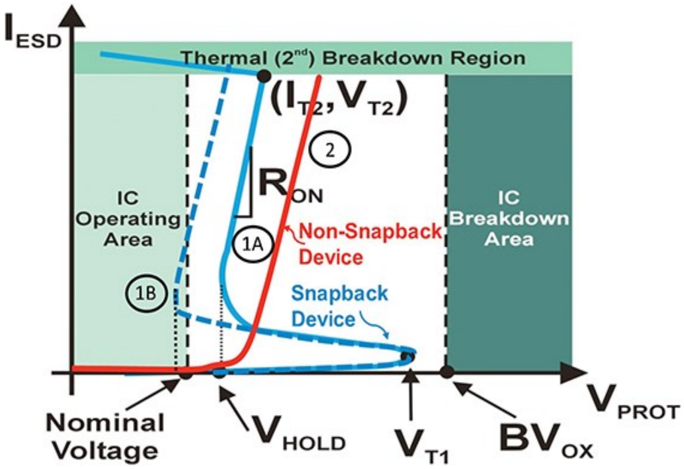


Figure 3.6 ESD design window

rectifier (SCR), and diode-trigger silicon control rectifier (DTSCR). The ESD design window must be confined between the IC operating voltage and IC breakdown voltage because the ESD does not join in the work during the normal operating condition and below the breakdown area to prevent the ESD not damaged. For both types of devices, we must make sure the V_{t1} is larger than the nominal voltage in case of ESD being turned on during normal operating conditions. The I_{t2} and V_{t2} are smaller than the breakdown voltage in order to protect ESD devices from breakdown. What's more, the on-resistance of both types of ESD devices should be as small as possible so as to easily discharge the ESD current. For the snackback device, we must also make the Vh is larger than the nominal voltage and smaller than the V_{t1}. If V_h is inside the IC operating area, it will easily cause the latch-up effect which may be triggered during the operating area.

3.1.3 ESD common structures and protection strategy

The common ESD devices have the below three types: diode, ggNMOS, and SCR whose cross-section structures are shown in figure 3.7. The diode is called P plus N well diode because the good type is n doped. The positive ESD current injects into the anode of the diode from the P plus part, then goes through the N well, and finally leaves from the N plus to the cathode. Since the diode is forward bias, it has a low on-resistance after about 0.7V triggering voltage. Cathode connecting the gate, N plus source, and P plus pick up forms the ggNMOS. The injecting current goes from the anode to the N plus drain area and induces the avalanche effect to turn on the parasitic bipolar junction transistor (BJT) which forms among the source, drain, and P well connecting to the P plus. Then large current goes from the N plus source to the cathode to discharge the ESD pulse [25]. In terms of the SCR, half of the part connects to the cathode and the other is connected to the anode. There are two parasitic BJT embedded inside the SCR structure: one is formed among the N plus, P well, and N well; the other is formed among the P well, N well, and N plus. Same as the ggNMOS, the injecting current causes the avalanche effect to turn on both parasitic BJT and finally flow out of the cathode.

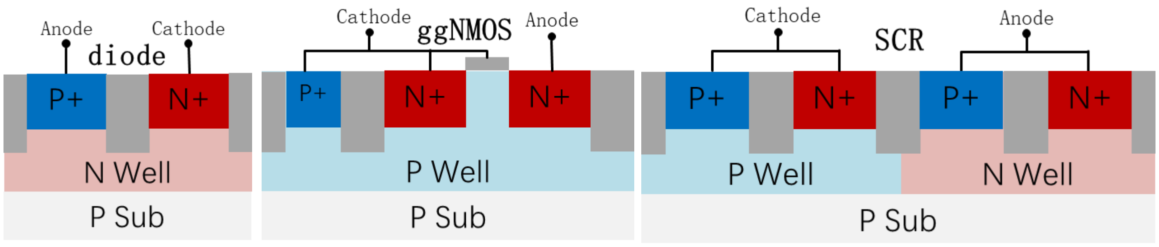


Figure 3.7 ESD cross-section structure of diode, ggNMOS and SCR

ESD protection strategy contains many schemes and one of the widely used is the double diode protection for the input and output (IO) pin. One of the diodes connects to the supply bus and the other connects to the ground bus. As seen from figure 3.8, there is also the RC clamp located between the power supply bus and the ground bus. It is noted that during the normal operation status, the diode and RC clamps are turned off and be turned on during an ESD event to provide a very low resistance discharging path. For a positive pulse with respect to Vdd, the current passes through the upper diode to the supply pin. For a negative pulse with respect to Vdd, the current goes into the supply pin, flows through the power clamp, and then passes the lower diode to the ground. For a positive pulse with respect to Gnd, the current passes through the upper diode, along the supply rail to the power clamp, and out to the ground pin. For a negative pulse with respect to Gnd, the current passes through the lower diode and out to the IO pin. It is worth to be noticed that the discharge path must not only be able to handle the current, but it must also be able to do so without allowing large voltages to develop across the parallel circuitry due to I-R effects. For some

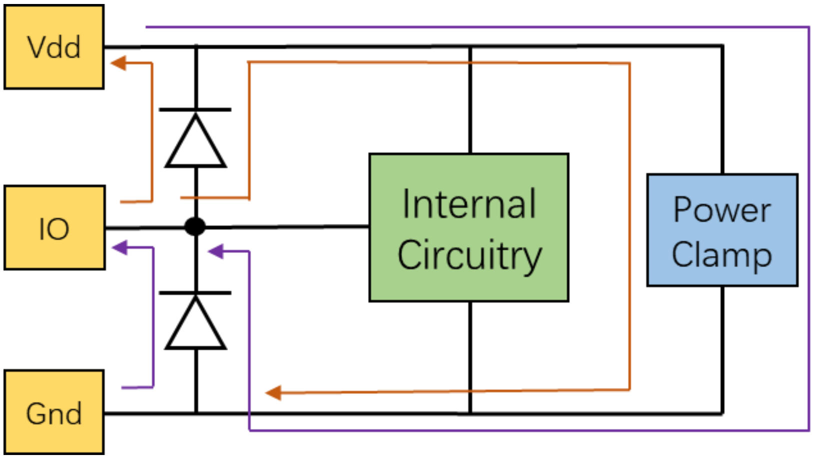


Figure 3.8 Double diode ESD network and discharging current path

ESD protection circuits, the lower diode can be replaced by the ggNMOS or SCR and the upper diode can be replaced by the diode string.

3.2 Graphene material background

3.2.1 Graphene advantageous properties

Graphene is an extremely thin layer of an innovative material that is widely called 2D material. Since graphene was first discovered by Andre Geim and Konstantin Novoselov from the University of Manchester exfoliating the graphene via the scotch tape, it has been populated studied, and explored many outstanding properties covering many areas from chemical, material, physics, and electrical engineering [26]. The 2D material is composed of honeycomb hexagonal lattice and can be extended into a very large plane but with very few layers of thickness even to one atom thickness. As seen from figure 3.9, graphene is an allotrope of carbon in the form of a plane with sp2-bonded atoms of a molecular bond length of 0.142nm. Thru stacking layers of graphene, it can form the graphite with van der Waals forces attracting each other between the layers of 0.335nm space. Graphene is the thinnest known material in this world which can only have the thickness of one atom. Although it is the lightest material of 0.77mg/m^2, it is very tough with a tensile strength of about 130GPa and Young's modulus of 1TPa which is much stronger than the steel. What's more, it has extraordinary electron mobility over 2×10^5cm^2/V•s, which makes it the best electrical conductor and behaves outstanding in heat conduction [27] [28].

3.2.2 Graphene material synthesis

Normally, there are a few processes to synthesis the graphene material such as micro-exfoliation and chemical vapor deposition (CVD). For micro-exfoliation, previously stated the scotch tape method, is the first developed method to discover graphene. It can be done at room temperature, low cost and have a good quality, which is normally used in academic research. However, the scale of the graphene is very small, its shape is unregular and not the uniform quality control. Therefore, the graphene synthesized by the micro-exfoliation cannot be used for the product, especially for the industry. The other method is CVD whose main advantage is the large scale of the graphene, uniformly producing control and the regular shapes. Besides, with more and more studies on graphene synthesis, the different quality of graphene can be fabricated, and also the fabricated condition can be eliminated a lot [29] [30]. Figure 3.10 shows the facility where the mainly CVD method synthesizes the graphene. The inlet of the gas contains hydrocarbon as the main source of graphene and the Ar/H_2 as the auxiliary gas. These gases flow into the quartz tube surrounded by the high-temperature tube furnace to participate in the deposition of the carbon atoms. The copper foil is normally used as the catalyst and substrate for graphene growth. Carefully

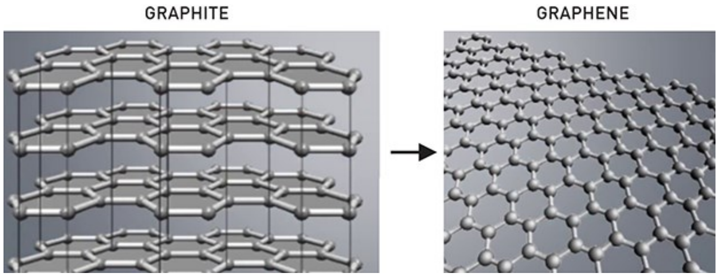


Figure 3.9 A sketch of graphite and graphene lattice structures.

controlling the growing conditions such as the percentage of the carbon content gas and the catalyst as well as the growth temperature is critical for the graphene quality.

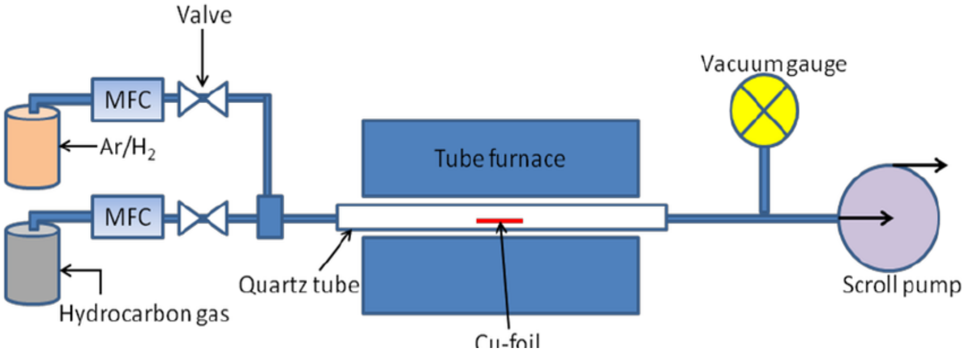


Figure 3.10 A sketch diagram of the graphene synthesis facility and needed sources.

Chapter 4 TSV-based ESD protection and TCAD flow for calibration

4.1 Vertical TSV-like diode ESD protection design

4.1.1 Vertical TSV-like diode for distribution ESD protection

In general, higher ESD protection means a larger layout size of the same ESD device types. For high pin count chips in more advanced technologies, ESD device sizes and layout planning become a major headache to IC designers. In addition, CDM ESD protection is a new design challenge to advanced ICs. However, the traditional large-size side-by-side planar ESD protection structures are not suitable for the new internally distributed CDM ESD protection scheme [31]. Figure 4.1b shows the concept of the new embedded TSV diode ESD device which is the same structure as the under-FET thermal sensor, which is a type of 3D heterogeneous structure for the whole ICs [32]. Because of the core in-hole vertical PN junction, it can be used as the ESD diode for ESD protection. Figure 4.1a illustrates the distributed ESD protection based on the vertical TSV-like diode. Since ESD charges exist in many domains and if the die area is large, the lump ESD devices are normally too far away from the ESD charge accumulation area which will result in ineffective ESD discharging especially for the CDM due to the internal discharging

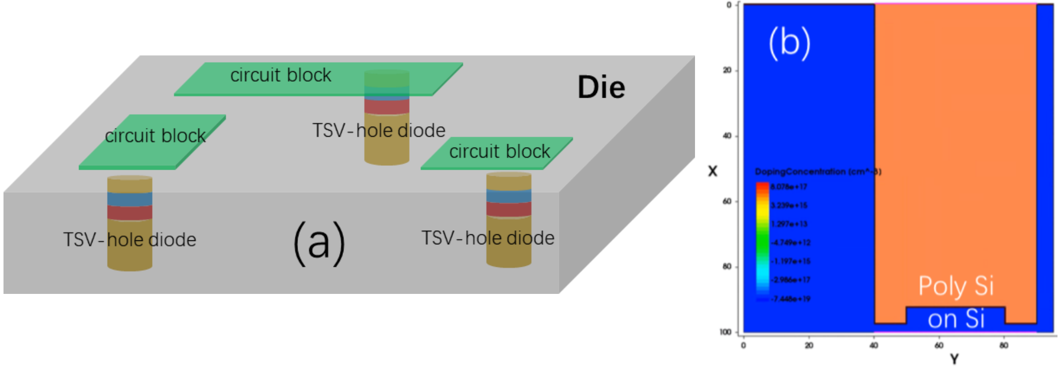


Figure 4.1 Concept of the new in-TSV ESD diode for distributed ESD protection (a) and its X-sectional view by TCAD (b).

mechanism. Therefore, the distributed ESD protection will be proposed in case of some domain areas are damaged by the local ESD discharging. It is found that, unlike conventional in-plane planar ESD protection structures that require careful lateral electrical connection (metals or diffusion regions), an in-TSV ESD protection diode has one terminal connected to an overhead pad and the other electrode vertically and directly connected to a local GND to the backside of the substrate. It has several advantages compared to the traditional side-by-side planar ESD protection structures. First, it largely decreases the consumption area taken by the planar ESD and its diffusion interconnects due to the TSV embedded diode readily placed underneath the circuits block. Second, it is very layout-friendly. Third, the underneath location can allow building the distributed ESD protection mesh on an IC die to improve the protected efficiency. Fourth, it can reduce the series ESD discharge resistance of grounding wires and dissipate ESD heat easily through the TSV Cu

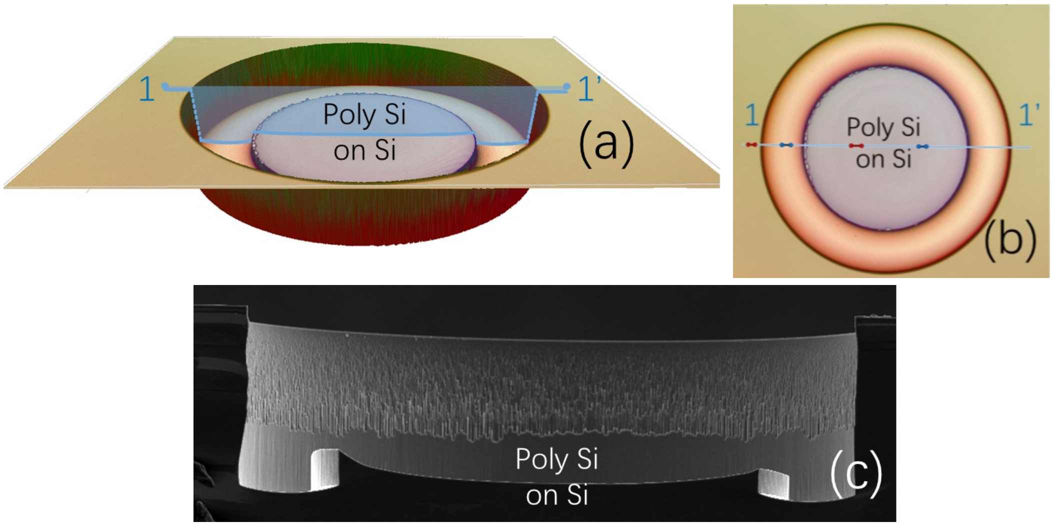


Figure 4.2 Images for prototype in-TSV poly-Si/Si PN diode ESD protection device fabricated inside a 100μm deep TSV hole: (a) a 3D image by confocal microscope, (b) a X-section view along 1-1’ cutline of the PN diode by SEM, and (c) a top view of in-hole ESD diodes by optical microscope along 1-1’ cutline.

pillar vertically [33]. Almost the same fabrication process as the under-FET thermal senor, the vertical TSV-like ESD diode is fabricated in our cleanroom with the complicated procedures due to the facilities limitation. Figure 4.2 shows the final structure with an effective diameter of 400μm inside the bottom of a 100μm deep hole via different microscopes. Figure 4.2a is the confocal 3D image that shows the 3D structure with the 1-1' cutline displayed in figure 4.2b top view. The PN junction line is shown in the SEM image of figure 4.2c.

4.1.2 TCAD comprehensive simulation on vertical TSV-like diode

Extensive TCAD ESD simulation was conducted to validate the new TSV-like ESD diode and guide its designs. Figure 4.3 shows the TCAD simulated vertical TSV-like diode with a diameter of 30um and 100um depth. It is noticed that we follow the real process almost the same as the procedure in the cleanroom to achieve the real device. The TCAD cannot only guide the fabrication process but also can predict the ESD behavior just like the HBM and CDM zapping. Figures 4.3a and 4.3b are the HBM zapping procedure with the HBM pulse of about 1μs defined by the JDEC standard. It shows the maximum I_{t2} current of about 28.3A equalizing the 42.5kV and also the small zoom-in diagram shows the V_{t1} is about 0.76V. From figure 4.3b, we can find that the maximum lattice temperature achieves 1678°C which is close to the silicon melting temperature. Similarly, figure 4.3c shows the CDM zapping with a CDM pulse of about 1ns defined by the JDEC standard. We can find the maximum I_{t2} is about 199A which refers to 17.5kV combined with the maximum lattice temperature of about 1561°C approaching the silicon melting temperature. Also, the zoom-in diagram from figure 4.3c shows the V_{t1} is also about 0.7V. from figure

4.3, we can find the TSV-like hole diode behaves very well with a current handleability of about 60V/μm^2 for HBM and 25V/μm^2 for CDM.

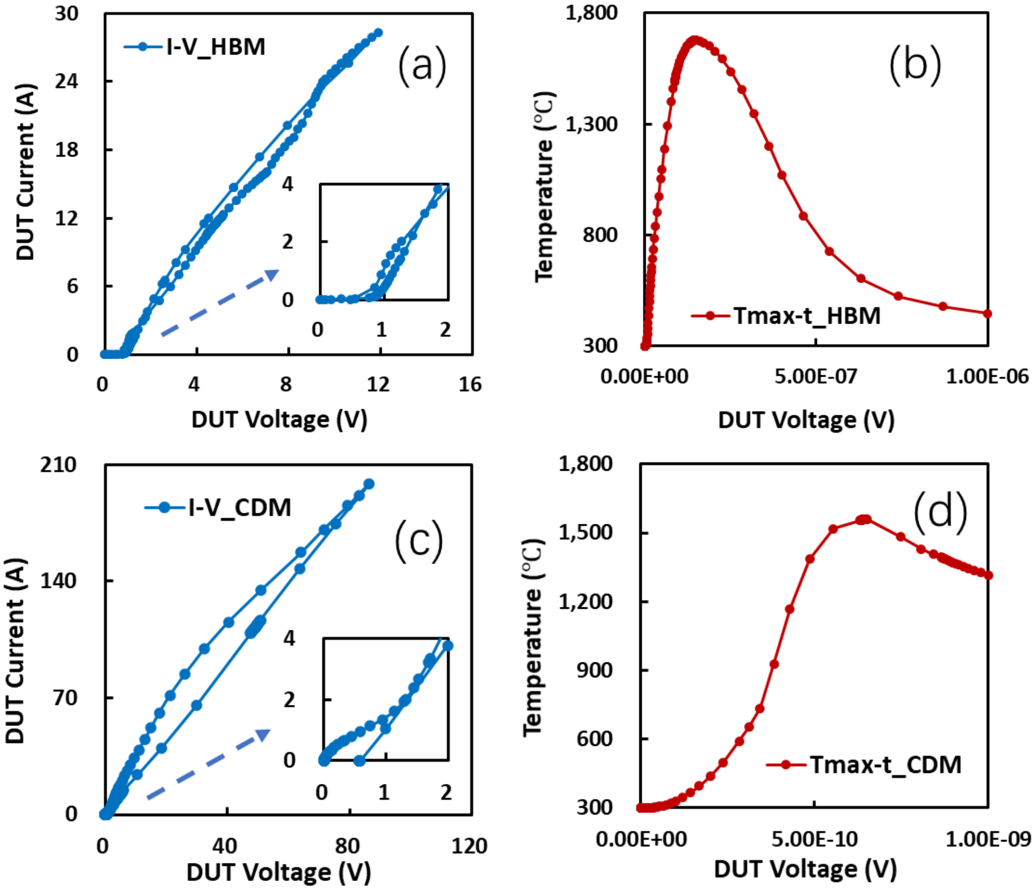


Figure 4.3 TCAD transient ESD simulation on vertical TSV-like diode: (a) I-V curve zapped by HBM waveform, (b) T_{max}-t curve zapped by HBM waveform, (c) I-V curve zapped by CDM waveform, (d) T_{max}-t curve zapped CDM waveform.

Figure 4.4 shows the TLP waveform for (a) and the VFTLP waveform for (b). These are still transient waveforms but we just integrated multiple pulses into one curve according to the TLP and VFTLP definitions. Figure 4.4a shows the TLP pulse about 100ns pulse width and 10ns rise time simulating results with I_{t2} ~20A in the maximum lattice temperature 1663°C. Also, the zoom-in diagram shows the V_{t1} is about 0.77V. Same as before, we can get the current handleability for TLP about 28mA/μm^2. Same as before, figure 4.4b shows

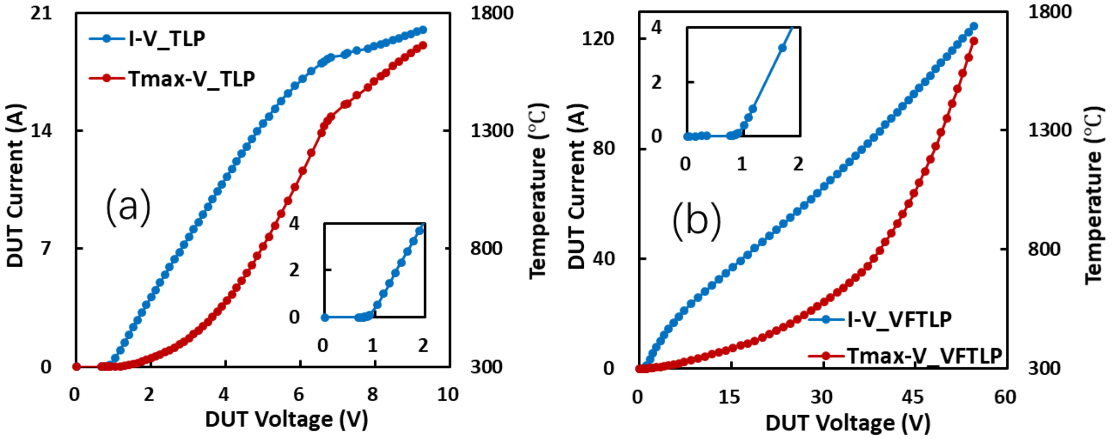


Figure 4.4 TCAD transient ESD simulation on TSV-like hole diode: (a) I-V and T-V curves for TLP waveform, (b) I-V and T-V curves for VFTLP waveform.

the VFTLP waveform with 1ns pulse width and 100ps rise time pulse. The I_{t2} is about125A under lattice temperature 1676°C combined with the V_{t1} 0.7V from the zoom-in diagram from figure 4.5. We can also calculate the VFTLP current handleability of about 177mA/µm^2.

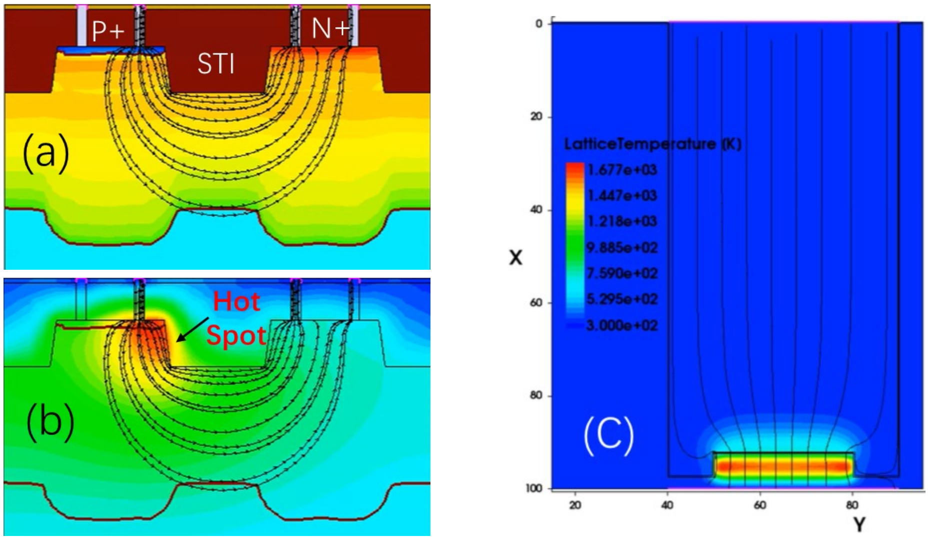


Figure 4.5 TCAD transient ESD simulation reveals: (a) the ESD discharge current flowlines are curved around the STI plug, causing (b) severe local overheating at the STI corner, i.e., a hot spot for a traditional planar STI ESD diode, (c) the vertical evenly distributed ESD discharge current flowlines and uniform thermal map under HBM zapping for a sample TSV-like ESD diode.

In terms of the traditional ESD diode device shown in figure 4.5a with P plus N well structure and the middle shallow trench isolation (STI) part. When an HBM pulse zaps the device from the anode of P plus entering to the cathode of N plus leaving, the current flow lines are displayed in figure 4.5b. We can find the obvious uneven distributed phenomenon where the current crowds with the inner contact area and accumulates along with the middle STI, which will result in the local hot spot surrounding this area. However, it is found that the vertical PN junction diode has a straight current path instead of the bent current path inside the traditional diode structure, which is shown in figure 4.5c. We can find that because of the bent current flow, the STI corner will have the heaviest current density, where is normally the hot spot limiting the ESD device performance. For our device, the active part is straightly stacked up and down, which can evenly distribute the current density in order to maximize the ESD current discharge ability.

4.1.3 TLP validation on vertical TSV-like diode

The prototype vertical TSV-like ESD didoes fabricated were characterized by TLP testing (Barth 4002 TLP tester). As shown the figure 4.6a, displays the reversed function.

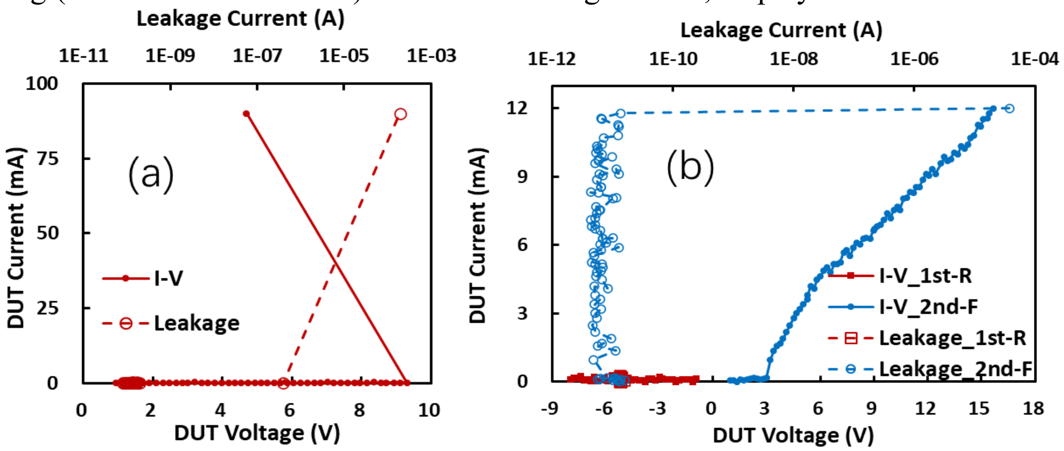


Figure 4.6 TLP testing on TSV-like diode: (a) reverse TLP zapping shows the current blocking until broken-down, (b) first reverse zapping and second forward zapping shows the TSV-like diode function.

We give the TLP to find its punch through voltage about 9.31V, which is satisfied IC design needs. The leakage current suddenly enlarges meaning the PN junction has broken down. Figure 4.6b is the sequential TLP testing mimicking the real operating situation to test its reliability during different ESD zapping situations. Firstly, we give the reverse TLP zapping until 7.98V below the reverse breakdown voltage, which shows the normal working status. Then we zap the forward TLP pulse until the forward breaks down. We can find it still successfully working with V_{t1} 1.91V and I_{t2} 12.0mA. It behaves very well though it is lower than the simulation results due to several factors. First, the TCAD simulation was for an ideal vertical ESD diode made of single-crystal Si PN junction, while the prototype device is a poly-Si/Si PN junction structure. Second, the prototype TSV-like ESD diode fabricated in our cleanroom was not yet optimized in terms of the device structure, PN junction formation, impurity doping, and metal contact formation.

Figure 4.7 presents the TLP-measured transient ESD discharge I-V curves for five TSV-like ESD diode samples, which shows desired ESD discharge I-V behaviors and low leakages. The multiple-sample testing also confirms the uniformity and stability of the new vertical in-TSV ESD didoes. The TLP testing results prove that the new vertical in-TSV ESD diode structure works for ESD protection. The extracted ESD triggering voltage is about $V_{t1} \sim 2.0V$ and the ESD thermal breakdown current is about $I_{t2} \sim 13.0mA$, which validates that the whole fabrication proves to be feasible. The foundry can also largely decrease the diode size by shrinking its diameter with advanced lithography to fulfill the advanced technology node. In total, the new in-TSV ESD structure has the potential to transform the full-chip ESD protection design practices for next-generation ICs.

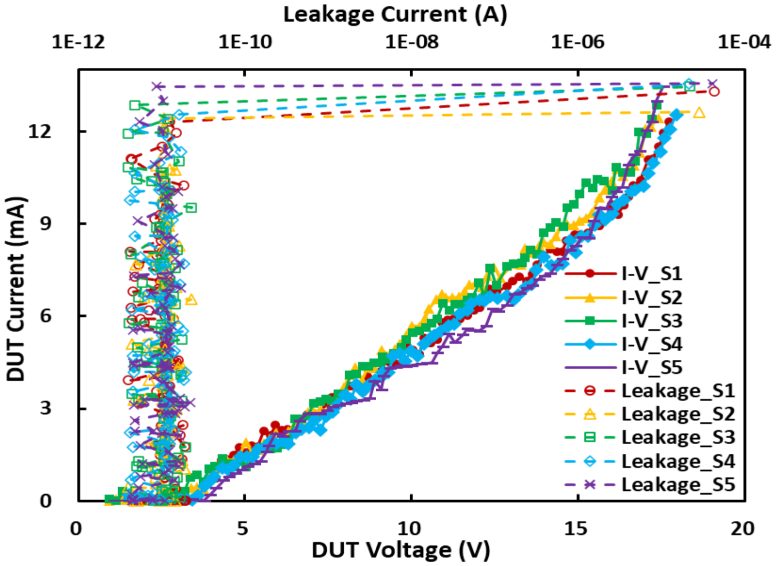


Figure 4.8 TLP-measured transient ESD discharge I-V curves for prototype TSV-like ESD diodes confirm the ESD protection function and low ESD-induced leakage.

Figure 4.8a shows the schematic of distributed CDM ESD protection network based on dummy circuits of a 3-stage oscillator designed in the 45nm SOI [34]. It is found that the vertical TSV-like diodes are integrated into this illustrated schematic which is applicable for the CDM distributed ESD protection. Instead of CDM current flowing through the transistor PM2 and NM8, the TSV-in diode can be placed locally to discharge the CDM accumulated charges ahead in order to protect the transistors. It is displayed in figure 4.8b that exemplary transient voltage analysis for V_{GS} of PM2 and NM8 of the ICs using TSV-

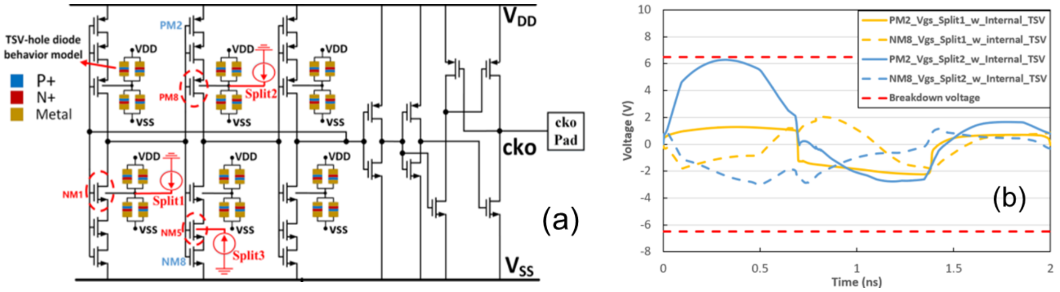

TSV-hole diode behavior model
P+
N+
Metal
PM2
VDD
VSS
PM8
Split2
NM1
Split1
NM5
Split3
NM8
V_{DD}
cko
cko Pad
V_{SS}
(a)
10
8
6
4
2
0
-2
-4
-6
-8
Voltage (V)
0
0.5
1
1.5
2
Time (ns)
PM2_Vgs_Split1_w_Internal_TSV
NM8_Vgs_Split1_w_internal_TSV
PM2_Vgs_Split2_w_Internal_TSV
NM8_Vgs_Split2_w_Internal_TSV
Breakdown voltage
(b)

Figure 4.7 (a) A 3-stage oscillator with TSV-based internally distributed CDM ESD network, (b) transient voltage analysis for V_{GS} of PM2 and NM8 using TSV-based can pass 350V CDM zapping.

based internal-distributed ESD protection network can pass 350V CDM ESD zapping without reaching the breakdown voltage for the oxide (BV_{OX} < 6.5V).

4.2 TCAD Mixed-mode ESD flow for Calibration

4.2.1 TCAD ESD simulation flow overview

Technology Computer-Aided Design (TCAD) refers to using computer simulations to develop and optimize semiconductor processing technologies and devices. The TCAD is used to perform a process simulation using Sentaurus Process, to set up meshing strategies for devices, and to simulate the electrical, thermal, and optical characteristics of various semiconductor devices. In Sprocess simulation, processing steps such as etching, deposition, ion implantation, thermal annealing, and oxidation are simulated based on physical equations, which govern the respective processing steps. After the structure is created, use the “Sdevice” tool to define the HBM and TLP pulses and simulate the I-V behavior and thermal behavior of our device under ESD events. Sentaurus Device produces output files containing electrode names and resulting voltages, currents, charges, times, temperatures, and so on. Svisual and inspect are the view tools in Sentaurus Workbench. Svisual is the tool to view TDR files which contains the device structure, temperature distribution, current density, and other simulation information. Inspect is the tool to view PLT files which shows the simulation results by plotting the current versus voltage curve and temperature versus time curve. The flow is displayed in figure 4.9. To obtain a set of parameters that allow an accurate predictive simulation of ESD behavior, a good calibration methodology is very essential. In ESD simulation, physics models such as Recombination, Auger Avalanche, Mobility, Electro-Thermal are used to simulate the electrical and thermal behavior of the device when encountered with a large HBM or TLP pulse. By adjusting the parameters in these models, we can calibrate the critical parameters in ESD analysis such as holding voltage V_h, trigger voltage V_{t1}, turn-on resistance R_{on} and failure current I_{t2}.

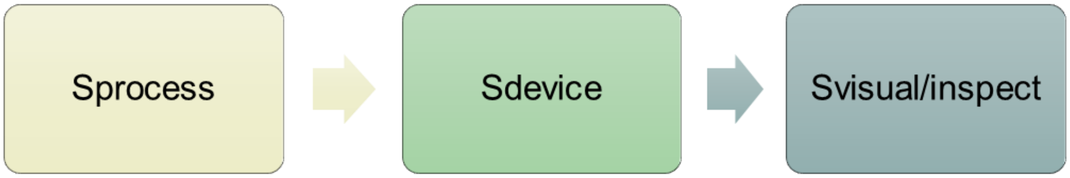


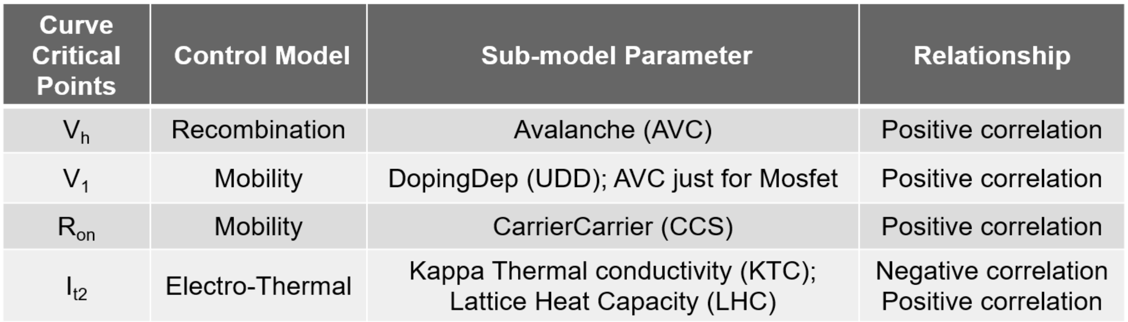
Figure 4.9 TCAD simulation process flow.

4.2.2 TCAD ESD calibration methodology

It is important to understand that, the proper TCAD ESD simulation calibration should follow the sequences: Step-1: to design simple ESD test patterns (must be single-finger ESD structures) for TCAD calibration; Step-2: to test the Si devices fabricated using TLP; Step-3: using the TLP testing data to calibration TCAD ESD simulation curves. After calibration, TCAD ESD simulation flow can be used to design, optimize and predict ESD protection designs for the given process technology. Because the models used in TCAD are only theoretical models, there might be a large difference between the simulation results and experiment results. Hence, a good calibration methodology is essential to obtain a set of parameters that allow an accurate predictive simulation of ESD behavior. Since TLP is a set of ascending pulses, we need to use the inspect tool to integrate all the transient I-V curves before breakdown. ESD-related modeling is addressed by: Impact ionization, Mobility, Recombination, Bandgap narrowing, electrothermal effects, and contact resistance. Table 1 shows the main physical models used to calibrate the TLP I-V curve. UDD which refers to doping dependence sub-model belonging to the mobility model is normally used to calibrate the triggering voltage V_{t1}. AVC which refers to the avalanche sub-model belonging to the recombination model is normally used for calibrating the triggering voltage V_{t1} and holding voltage V_h. CCS which refers to carriercarrier scattering sub-model belonging to the mobility model is normally used to calibrate the on-resistance R_{on}. LHC which refers to lattice thermal conductivity sub-model and KTC which refers to kappa thermal conductivity sub-model both belonging to the electro-thermal model are normally used to calibrate the breakdown I_{t2}.

4.2.3 TCAD ESD calibration results

When preparing to calibrate the device like diode or MOSFET, it is better to use the calibrated kit structure like the single finger diode and MOSFET which will improve the calibration accuracy. It is noted that the TCAD fabrication process must comply with the real silicon fabrication procedures and be validated by the operation validation between the simulation and testing. After this, we can come to the ESD performance calibration. There are mainly four steps for calibration: Firstly, calibrate the V_{t1} to make the curve coarsely match which is to modify the AVC and UDD parameters. When you change the UDD or AVC larger, the V_{t1} will become larger, vice versa, it will get smaller. After adjusting the AVC and UDD the R_{on} and I_{t2} may also change to a certain extend. Secondly calibrate the



Curve Critical Points	Control Model	Sub-model Parameter	Relationship
V_h	Recombination	Avalanche (AVC)	Positive correlation
V_1	Mobility	DopingDep (UDD); AVC just for Mosfet	Positive correlation
R_{on}	Mobility	CarrierCarrier (CCS)	Positive correlation
I_{t2}	Electro-Thermal	Kappa Thermal conductivity (KTC); Lattice Heat Capacity (LHC)	Negative correlation Positive correlation

Figure 4.10 Main physical models related to TLP calibration

R_{on} to make the curve coarsely match, which is to modify the CSS parameters. When you change the CSS larger, the Ron will become larger, vice versa, it will get smaller After adjusting the CSS the V_{t1} may change a little and I_{t2} may change to a certain extend. Thirdly calibrate the I_{t2} to make the curve coarsely match, which is to modify the LHC and KTC parameters. When you change the LHC larger and KTC smaller, the I_{t2} will become larger, vice versa, it will get smaller. After adjusting the LHC and KTC the V_{t1} and Ron may change a little. Last repeatedly calibrate the V_{t1}, $R_{on,}$ and I_{t2} to make the curve almost match. After slightly adjusting the above parameters to get an almost matching curve.

Here we will take some examples to demonstrate the calibration results with the 55nm CMOS process. Figure 4.11 is the structure of the very low triggering N well diode (LLVNW) with the N plus (NL) equaling to P plus (PL) and STI (SL) of 1μm. Figure 4.11a depicts the diode structure with N well and two contacts according to the real device structure and doping distribution. Figure 4.11b shows the heat distribution map which reveals the hotspot locates in the interface between middle STI and silicon as well as no

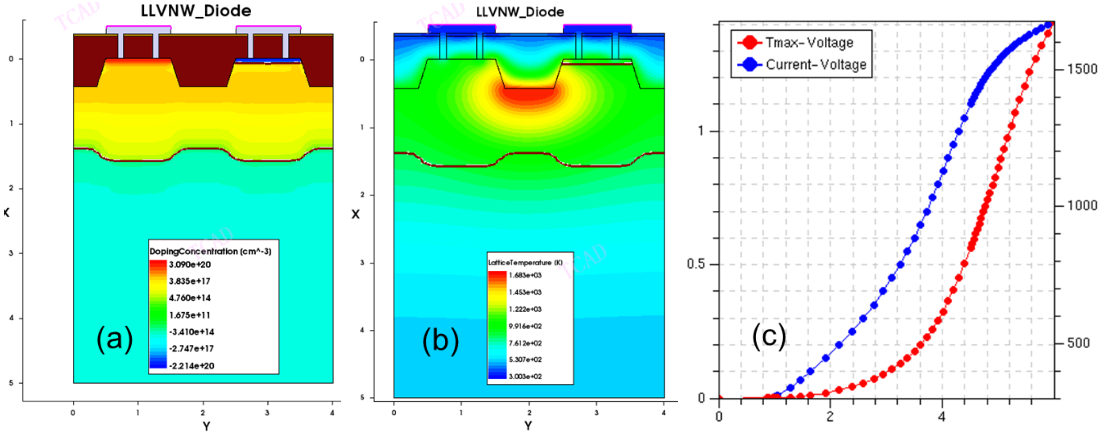


Figure 4.11 TCAD ESD simulation of LLVNW diode (NL=PL=SL=1) under TLP zapping: (a) cross-section of TCAD simulated structure, (b) heat distribution map showing the hotspot location, (c) I-V and T_{max}-V curve until the silicon melting temperature.

shifting because of equal active area length which has the same current density. Figure 4.11c shows the I-V curve and T_{max}-V curve under the TLP TCAD simulation where the simulation will stop at the silicon lattice melting temperature of about 1683℃. In this diagram, we can read the triggering voltage V_{t1}, on-resistance $R_{on,}$ and breakdown current I_{t2}. Figure 4.12 shows the comparison data (N plus, Plus as well as STI ~1μm) between the TLP testing about pulse width of 100ns and rise time of 10ns and the TCAD TLP simulation. It is found that the testing I_{t2} is determined by the leakage current when the I_{leak} suddenly enlarges meaning the device faces thermal breakdown. Therefore, the testing I_{t2} is 1.31A, and simulation I_{t2} is 1.41A with a relative error of about 7.6% totally satisfying the goal. Meanwhile, we can find that the on-resistance R_{on} and triggering voltage V_{t1} is almost matched for the testing data and simulation data. Figure 4.12b is the zoom-in view of figure 4.12a.

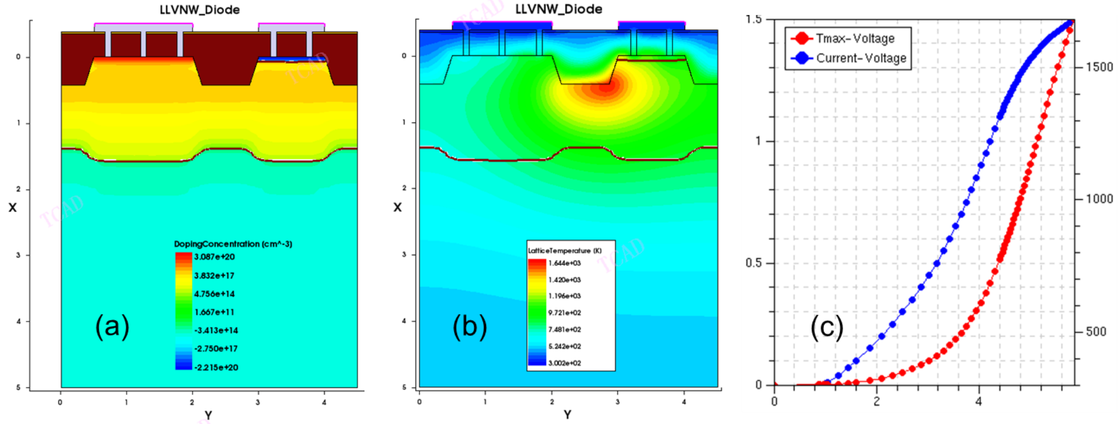


Figure 4.12 TCAD ESD simulation of LLVNW diode (NL=1.5 PL=SL=1) under TLP zapping: (a) cross-section of TCAD simulated structure, (b) heat distribution map showing the hotspot location, (c) I-V and Tmax-V curve until the silicon melting temperature.

It is noted that for the calibration we must have the same adjusted parameters for a batch of calibrated devices which all have the acceptable error difference. If the calibrated parameters are only good for one device but not for the other one, the calibration has no meaning. Now let's adjust the dimension of the diode to validate our calibration results. Take figure 4.13 as an example whose structure is the very low triggering N well diode (LLVNW) with the N plus (NL) equaling to 1.5μm while P plus (PL) and STI (SL) equaling to 1μm. Same as before, figure 4.13a, figure 4.13b, and figure 4.13c represent the structure, heat distribution map, and the I-V curve as well as T_{max}-V curve. Figure 4.13b reveals the hotspot locates in the interface between middle STI and silicon but right shifting because of larger N active area which makes the current crowding around the P plus area. Figure 4.14 shows the comparison data of the LLVNW diode (N plus ~1.5μm and P plus as well as STI ~1μm) between the TLP testing about pulse width of 100ns and rise time of 10ns and the TCAD TLP simulation. It is found that the testing I_{t2} is 1.58A and simulation I_{t2} is 1.50A with the relative error of about 5.1% totally satisfying the goal. Meanwhile, we can

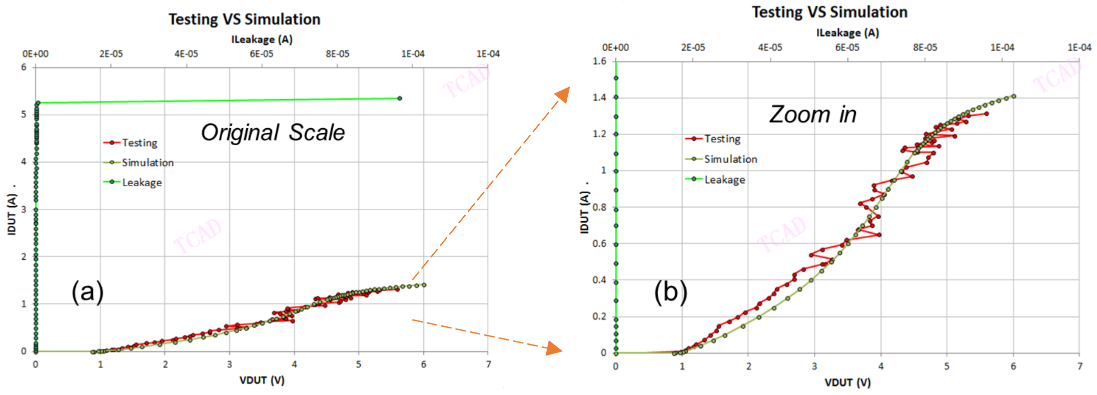


Figure 4.13 Comparison of LLVNW diode (NL=PL=SL=1) under TLP zapping between the testing data and simulation data shows a good matching mainly on the V_{t1}, R_{on} and I_{t2}.

find that the on-resistance R_{on} and triggering voltage V_{t1} is almost matched for the testing data and simulation data. Figure 4.14b is the zoom-in view of figure 4.14a.

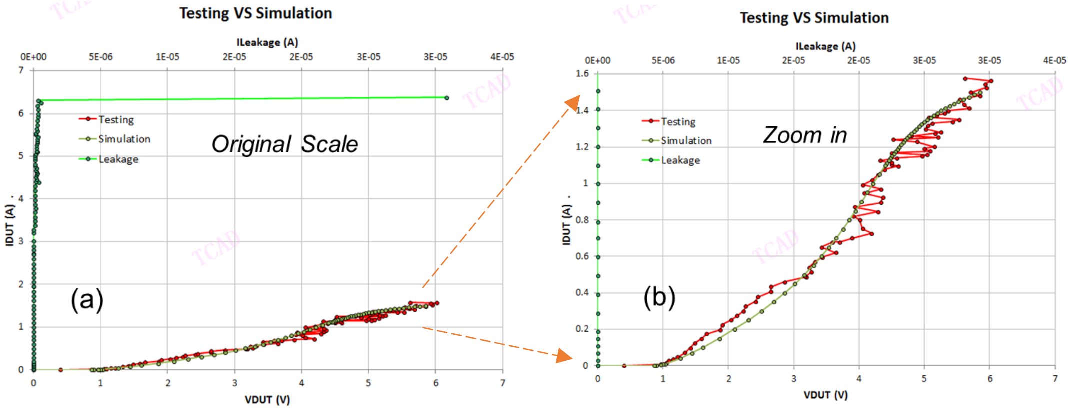


Figure 4.14 Comparison of LLVNW diode (NL=1.5, PL=SL=1) under TLP zapping between the testing data and simulation data shows a good matching mainly on the V_{t1}, R_{on} and I_{t2}.

Next, let's change the good type of the diode from N well to P well to validate our calibration results. Take figure 4.15 as an example whose structure is the very low triggering P well diode (LLVPW) with the N plus (NL) equaling to P plus (PL) and STI (SL) of 1µm. Same as before, figure 4.15a, figure 4.15b, and figure 4.15c represent the structure, heat distribution map, and the I-V curve as well as T_{max}-V curve. Figure 4.15b

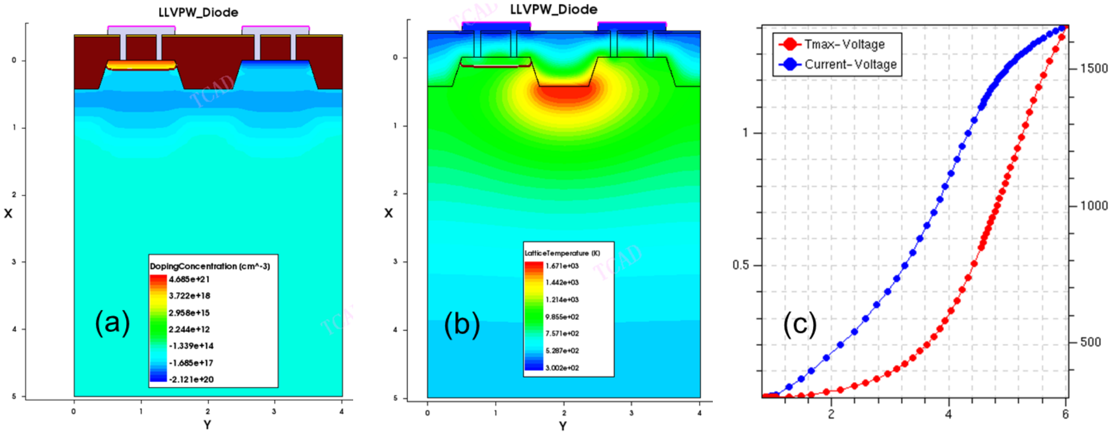


Figure 4.15 TCAD ESD simulation of LLVPW diode (NL=PL=SL=1) under TLP zapping: (a) cross-section of TCAD simulated structure, (b) heat distribution map showing the hotspot location, (c) I-V and T_{max}-V curve until the silicon melting temperature.

shows the heat distribution map which reveals the hotspot locates in the interface between middle STI and silicon as well as no shifting because of equal active area length which has the same current density. Figure 4.16show the comparison data of LLVPW diode (N plus, P plus as well as STI ~1μm) between the TLP testing about pulse width of 100ns and rise time of 10ns and the TCAD TLP simulation. It is found that the testing I_{t2} is 1.45A and simulation I_{t2} is 1.41A with the relative error of about 2.8% totally satisfying the goal. Meanwhile, we can find that the on-resistance R_{on} and triggering voltage V_{t1} is almost matched for the testing data and simulation data. Figure 4.16b is the zoom-in view of figure 4.16a.

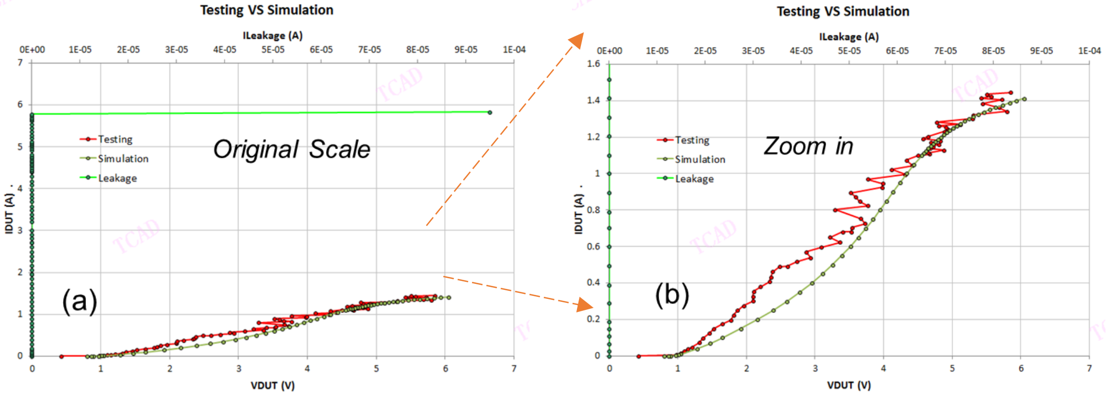


Figure 4.16 Comparison of LLVPW diode (NL=PL=SL=1) under TLP zapping between the testing data and simulation data shows a good matching mainly on the V_{t1}, R_{on} and I_{t2}.

Last, let's change the doping of the diode from very low triggering to normal low triggering to validate our calibration results. Take figure 4.17 as an example whose structure is the low triggering N well diode (LVNW) with the N plus (NL) equaling to P plus (PL) and STI (SL) of 1μm. Same as before, figure 4.17a, figure 4.17b, and figure 4.17c represent the structure, heat distribution map, and the I-V curve as well as T_{max}-V curve. Figure 4.17b shows the heat distribution map which reveals the hotspot locates in the

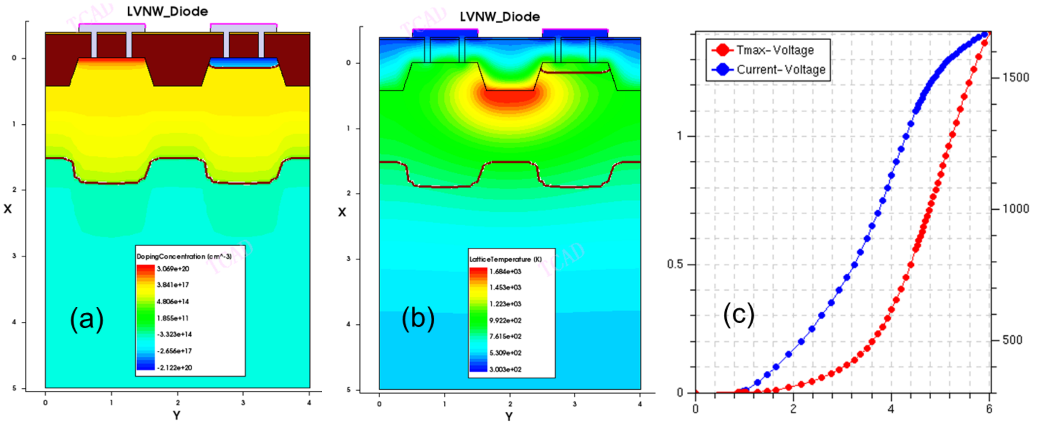


Figure 4.17 TCAD ESD simulation of LVNW diode (NL=PL=SL=1) under TLP zapping: (a) cross-section of TCAD simulated structure, (b) heat distribution map showing the hotspot location, (c) I-V and T_{max}-V curve until the silicon melting temperature.

interface between middle STI and silicon as well as no shifting because of equal active area length which has the same current density. Figure 4.18 shows the comparison data of LLVPW diode (N plus, Plus as well as STI ~1µm) between the TLP testing about pulse width of 100ns and rise time of 10ns and the TCAD TLP simulation. It is found that the testing I_{t2} is 1.45A and simulation I_{t2} is 1.41A with the relative error of about 2.8% totally satisfying the goal. Meanwhile, we find that the on-resistance R_{on} and triggering voltage V_{t1} is almost matched with testing data. Figure 4.18b is the zoom-in view of figure 4.18a.

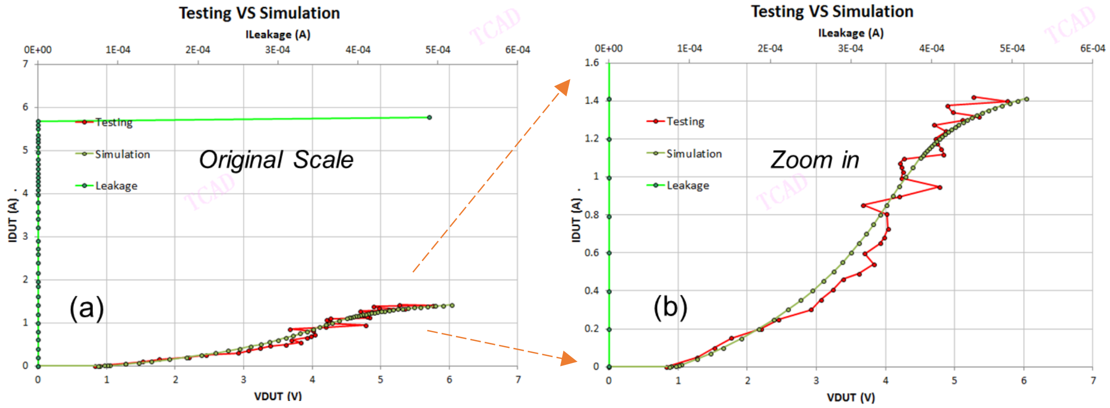


Figure 4.18 Comparison of LLVPW diode (NL=PL=SL=1) under TLP zapping between the testing data and simulation data shows a good matching mainly on the V_{t1}, R_{on} and I_{t2}.

We also study the generation process of the hotspot in the exemplar LLVNW diode. From a set of diagrams within the time domain shown in figure 4.18, the diode is zapped under HBM pulse and this set of diagrams shows the hotspot distribution. It is found that the hotspot mainly arises near the interface of silicon and oxide because of the poor thermal conductivity of silicon dioxide. Since the silicon and tungsten are the materials where the current main flows, the hotspot is located around here. At first, the hotspot arises neat the tungsten contact since the current crowing here. Then, the hotspot transfer underneath the middle STI due to the poor thermal conductivity of oxide. Also, the hotspot near the tungsten is relatively higher than other locations but in fact, isn't absolutely high. Therefore, the real hotspot is still located underneath the middle STI.

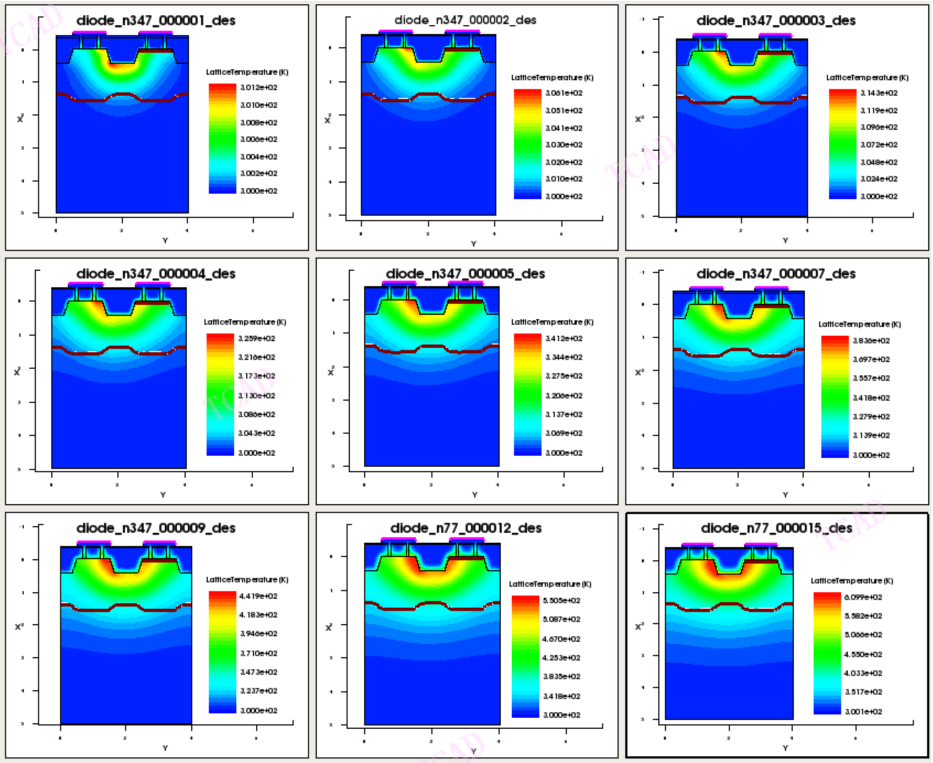


Figure 4.19 A set of hotspots distributed diagrams on the time domain under the TCAD HBM simulation showing the hotspot movement.

Chapter 5 3D TCAD ESD simulation and sudoku DTSCR design

5.1 2D vs 3D TCAD ESD simulation

5.1.1 Diode structure simulated by the TCAD for 2D and 3D

Due to the complex multi-level coupling effects involving transient, electrical, thermal, process, device, and layout parameters, physics-based ESD device modeling is still inadequate for circuit level simulation. Hence, TCAD ESD simulation has been widely used for ESD protection design prediction. Since there exists ununiform and inefficient thermal conduction under ESD stressing, ESD thermal failure is very sensitive to physical geometries of ESD protection device structures, where the edge or corner current and thermal crowding effects become serious design concerns. Consequently, ESD layout design becomes critical to the success of practical ESD protection designs, which cannot be addressed by a circuit or 2D TCAD simulation. Though over these years in the industry, widely used pseudo-3D also called 2.5D TCAD ESD simulation techniques are not useful in understanding ESD layout effects and achieving ESD layout optimization.

To guarantee the accuracy of the TCAD simulation, the ESD protection device used in this work is built via the real silicon fabrication process based 55nm memory CMOS process [35]. Figure 5.1 displays the 3D P plus N well diode structure created by the true 3D TCAD process simulation via the Synopsys Sentarus Process software. During the TCAD process simulation, it covers the whole complete and realistic fabrication process recipes including thermal and deposition oxidation, two steps of STI deposition, multiple steps of implanting and annealing on a different part of the device, multiple steps etching including isotropic and anisotropic. Illustrated in figure 1, the ESD diode generated by 3D

TCAD has three dimensions: X-dimension for device thickness, Y-dimension for device width, and Z-dimension for device length (typically referred to as ESD device finger length for multiple-finger ESD protection structures). For 3D TCAD simulation efficiency, and considering that an ESD event is extremely fast and the ESD-induced heating quickly reaches to thermal equilibrium around tiny hot spots, the 3D ESD diode dimension was truncated to a meaningful size of X=5µm, Y=4µm, Z=11µm especially for the substrate depth (X-axis). Only SiO_2 layer, Si_3N_4 layer, and tungsten contact layer were included in the 3D TCAD simulation since the focus is on ESD heating inside Si instead of the back end of the line. The P+ anode (A) and N+ cathode (K) regions are separated by STI plugs. Following the real product of the layout designs, the contact lines are evenly distributed into 12 metal contacts shown in the top Y-Z plane of figure 5.1a. From figure 5.1b which is the transparent view, it is found that the P plus and N plus area underneath the top dielectric layers. The 2D cross-section of the X-Y cutline view shown in figure 5.3c displays the specific shape of the plug STI and the doping distribution of the P plus, N plus, and N well which will influence the electrical properties of the diode under the HBM discharging event. Since without the later BEOL interconnects and the pads, we define the pad as the end of the distributed metal contacts. Figure 5.1d shows the X-Z cross-section along the contacts paralleling line above the N plus area.

For comparison studies, the same ESD diode structures were created using the same memory process recipes by 2D and pseudo-3D TCAD simulation, which has been commonly used by the industry, as shown in Figure 5.2. In a typical 2D TCAD simulation, the 2D ESD diode is created as shown in Figure 5.2a, which is equivalent to the X-Y cross-section of the 3D ESD diode given in Figure 5.1c. It has the same N plus, P plus, and plug STI part as well as the N well. Since true 3D TCAD simulation is very computing hungry, which means consuming lots of time and common issues like un-convergency, 2D TCAD simulation has been widely used in TCAD simulation, which is however very inaccurate in predicting ESD device behaviors, particularly for TCAD ESD simulation. Alternatively, the industry has been using the pseudo-3D TCAD ESD simulation over years where an ESD device created by 2D TCAD is simply extended in the Z-dimension (hence, also

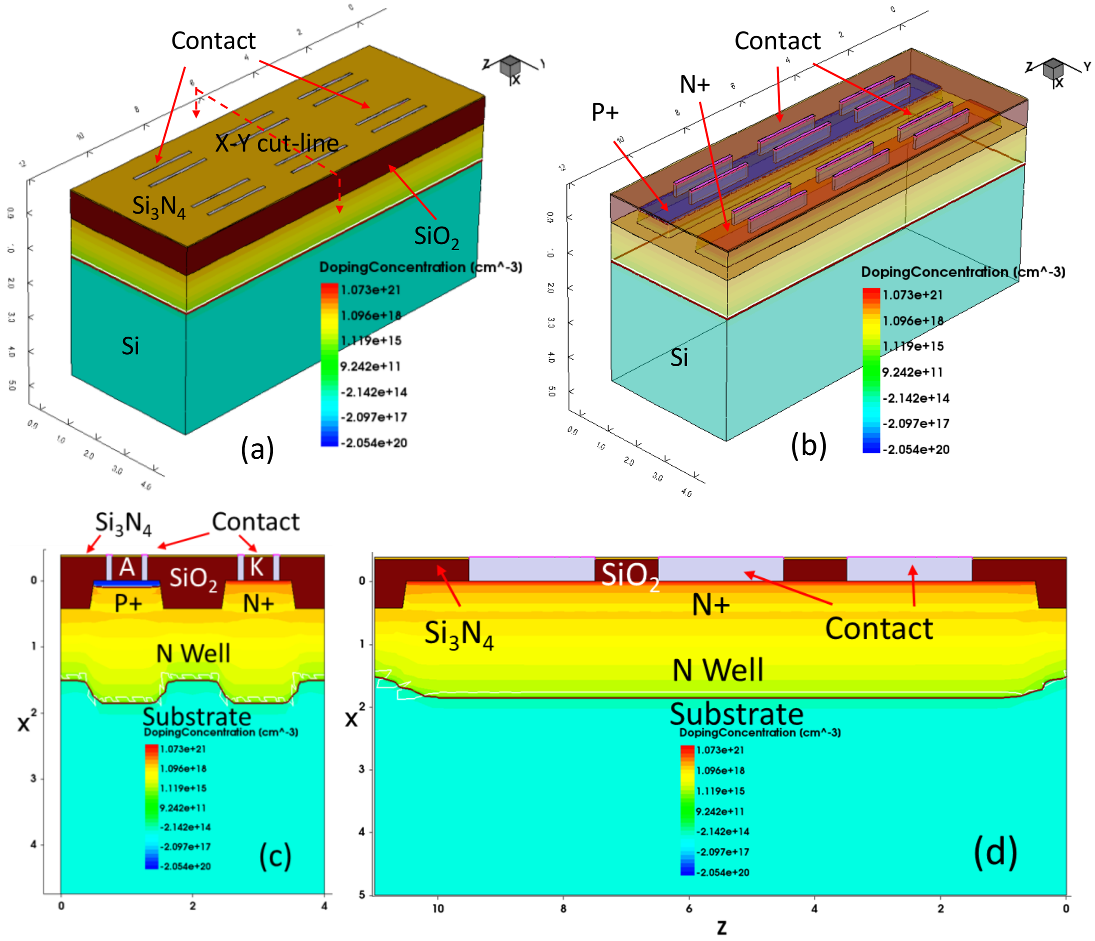


Figure 5.1 A 3D ESD diode generated by true 3D TCAD process simulation that contains 12 evenly distributed ESD metal contact lines per ESD Design Rules. (a) 3D ESD diode created, (b) 3D ESD diode view in transparent mode, (c) a cross-section view of the 3D ESD diode along the X-Y cut-line, (d) a cross-section view of the 3D ESD diode along the X-Z cut-line.

referred to as 2.5D). In this way, it can be easily fabricated in the TCAD process simulation and save a lot of time. What's more, it can be simulated either in the forms of the 2D or extending into 3D under the HBM ESD zapping to see the effects of the electrical properties. Also, the 2.5D structure is easily extended into the 3D which will omit lots of 3D internal unsymmetric domains and thus it will eliminate lots of un-convergency issues and improve the computing speed. Figure 5.2a shows the 2.5D diode structures with the dielectric layers above the active silicon layer with the uniform Z-axis extension like the whole straight contact and STI. From figure 5.2c which is the transparent view, it is found that the P plus and N plus area underneath the top dielectric layers.

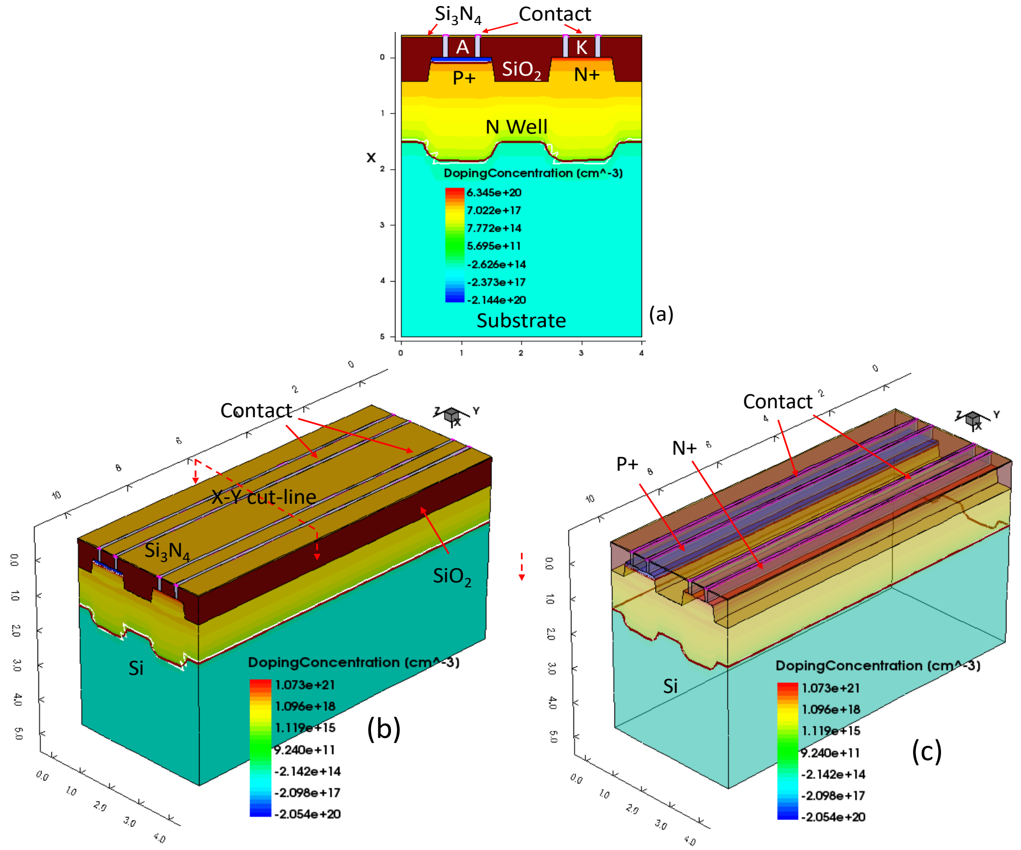


Figure 5.2 The same ESD diode structures created by (a) 2D TCAD (equivalent to Fig. 1c), and (b) psedo-3D TCAD with (c) being its transparent view.

5.1.2 Mesh density influence on the TCAD simulation

Before conducting HBM ESD stress mode using Synopsys Sentarus device tool which results from the HBM equivalent waveform, the mesh generation is a necessity for the 3D TCAD simulation. In principle, a denser device mesh structure gives more accurate simulation results. However, a balance check should be done in device mesh generation because increasing in device mesh density will dramatically increase TCAD simulation time, and beyond a certain point, the mesh density may not meaningfully further improve ESD simulation accuracy. In this work, a careful balance check leads to three mesh splits Fine mesh, Finer mesh, and Finest mesh. For a fair comparison, the physics model and algorithm selected in TCAD simulation must be the same for the ESD diodes created by 2D, pseudo-3D, and true 3D TCAD ESD simulation.

Figure 5.3 presents a set of transient ESD I-V curves (a, c & e) and T_{max}-t curves (b, d & f) for the same ESD diodes created by 2D and 3D TCAD process simulation, each for the Fine, Finer, and Finest mesh cases, respectively. It is found that the different mesh densities will influence the TCAD simulation results and with the density increases, the simulation will be sensitive. Therefore, to keep the results meaningful, we need to check the mesh density balance until insensitive. It is expected that the transient I-V-T behaviors for the ESD diodes generated by 2D and 3D TCAD simulations are very different. Specifically, the critical ESD discharging resistance (R_{on}) is very different with R_{on} being higher for a 3D ESD diode due to several factors: First, the fast transient ESD discharging behaviors are very sensitive to the 3D ESD device structure because of the edge/corner effect leads to current crowding and hence seriously localized heating, i.e., ESD hot spots. Therefore, the 3D structure simulation normally has an easier thermal failure. Second, the ESD discharging routing is determined by the ESD discharging paths, which are dictated by the physical layout designs, e.g., the layout of the ESD contacts, which in turn will lead to ESD discharging current crowding and local ESD hot spots. Compared with the 3D structure, the 2D will not consider the complex discharging path or corner effect which will totally influence the efficiency of the current flow. That's the reason that the on-resistance of 2D TCAD simulation is lower than that of 3D TCAD simulation on the P plus N well diode under the HBM zapping. What's more, it is obvious that with the mesh density increasing, the influence of the obstacles preventing current flow in the 3D structure will be more severe, which will enlarge the difference of on-resistance between the diode structure under 2D simulation and 3D simulation. Figure 5.3a shows the I-V curve under the fine mesh condition and figure 5.3b shows the T_{max}-t curve under the fine mesh condition. We can see the on-resistance of 3D simulation is a little higher than 2D and the lattice temperature of 3D is higher than 2D, which means the earlier silicon failure under HBM ESD zapping. Then with the density of mesh increasing from fine to finer until to the finest which is insensitive to the results, the difference on the Ron is aggravating shown in figure 5.3c and figure 5.3e. In the meantime, the T_{max}-t curves in figure 5.3d and figure 5.3f reveal the more severe self-heating in the 3D structure due to the local hotspot around the unevenly distributed discharging current. Thus, the highest lattice temperature is increasing.

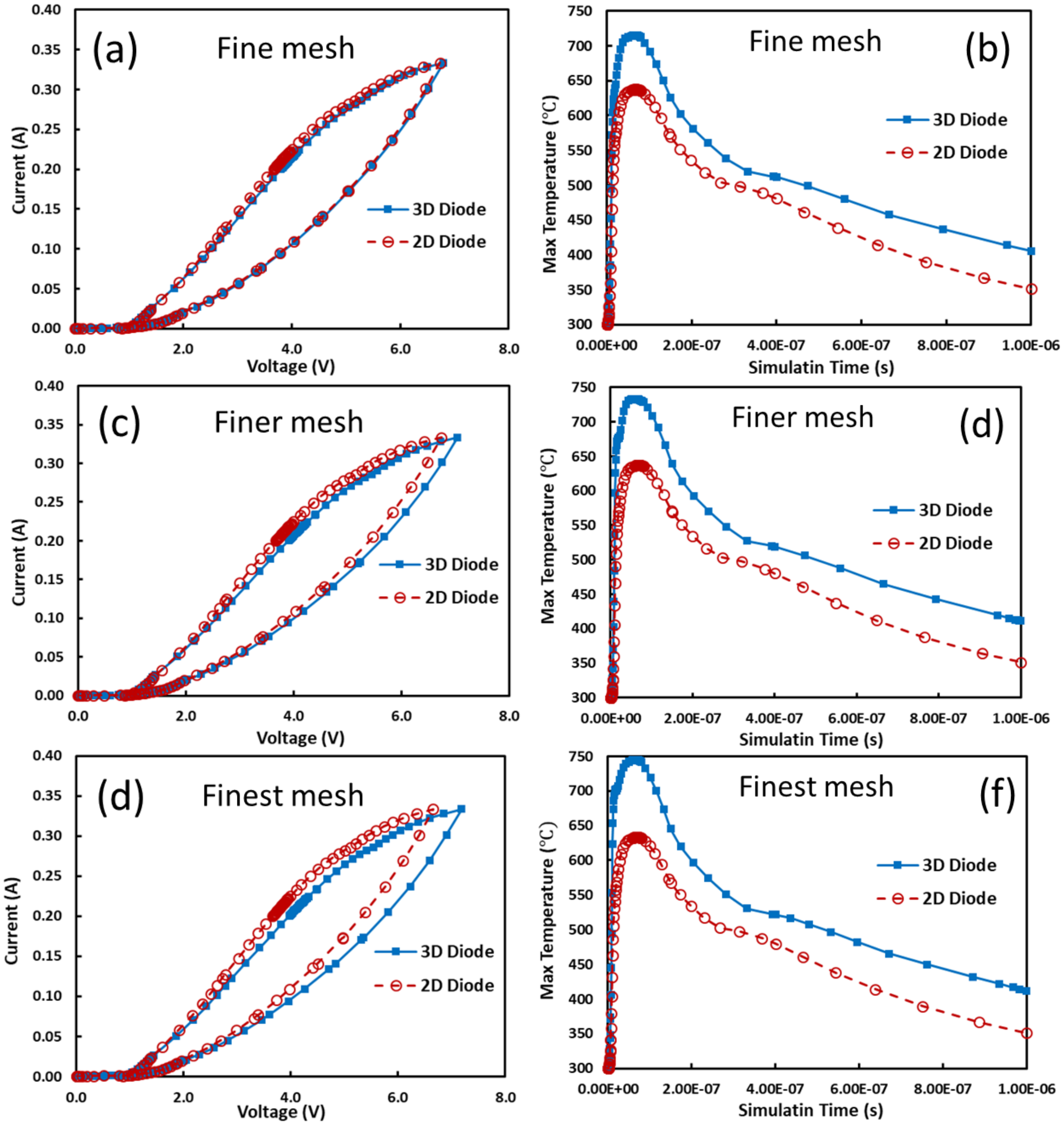


Figure 5.3 Transient ESD I-V-T curve comparison by TCAD ESD simulation for the 2D and 3D ESD diodes with varying mesh densities: (a) I-V curves for Fine mesh, (b) T_{max}-t curves for Fine mesh (c) I-V curves for Finer mesh, (d) T_{max}-t curves for Finer mesh and (e) I-V curves for Finest mesh, and (f) T_{max}-t curves for Finest mesh.

5.1.3 I-V-T behaviors on the TCAD ESD simulation

For figure 5.4, we compare the transient ESD TCAD simulated I-V and T_{max}-t curves for different diode structures created under the 2D, pseudo-3D, and true 3D TCAD process simulations. From figure 5.4a, it is readily observed that the triggering voltage (V_{t1}) is almost the same for these three structures but the on-resistance (R_{on}) is slightly different for these three structures. For the 2D and pseudo-3D, the Ron is almost the same while lower than the true 3D structure due to the assumption of uniform conduction in Z-dimension, which explains why the commonly used pseudo-3D TCAD ESD simulation is as useless as that 2D ESD simulation. However, for the true 3D structures, the current

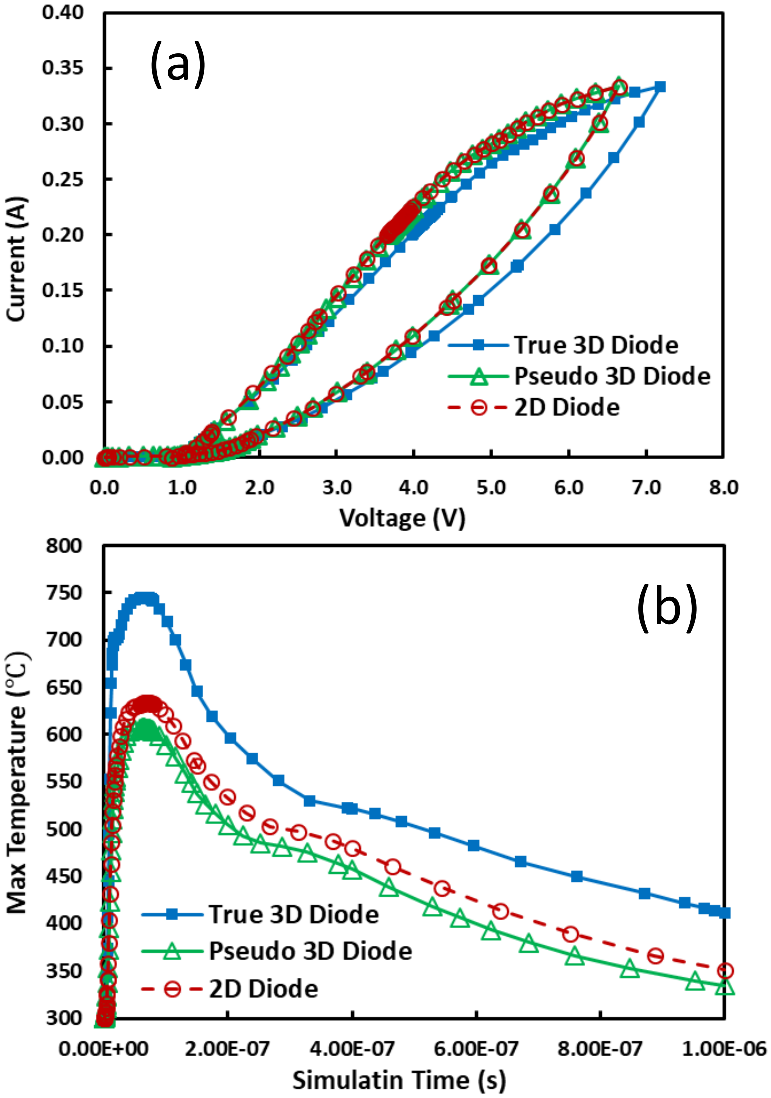


Figure 5.4 Transient ESD I-V-T comparison for 2D, pseudo 3D and true 3D ESD diodes: (a) I-V curves, and (b) T_{max}-t curves.

crowding will be reflected much more to enlarge the current path which results in a little larger on-resistance. In terms of figure 5.4b, true 3D TCAD ESD simulation clearly shows that 2D and pseudo-3D ESD simulation grossly overestimates ESD performance in practical designs. Specifically, T_{max} ~ 745 °C was observed in the true 3D ESD diode as compared to T_{max}~ 630 °C in the 2D ESD diode under the same HBM stress. The T_{max} difference is significant, which may lead to either early ESD failure or grossly over-design of ESD structures if 2D or pseudo-3D TCAD ESD simulation is used to predict ESD performance in Si.

5.1.4 ESD thermal map and design guidelines

The impact of true 3D ESD structure on ESD performance can be analyzed in detail through thermal mapping by TCAD ESD simulation. Figure 5.5 shows the ESD thermal map under transient HBM ESD stressing, including the hot spots, for the 3D ESD diode created by true 3D TCAD process simulation. Figure 5.5a shows the 3D transparent view of the heat distribution map and we can find that the hotspot is not evenly distributed along the Z-axis and of course for the 2D view of the X-Y plane. Fig. 5.5c gives a top view of the ESD thermal map that shows uneven thermal distribution within the ESD diode where ESD hot spots are localized in the center for two possible reasons: first, thermal dissipation is always weak in the center, and second, the assumed even ESD discharging across the ESD diode due to evenly arranged ESD metal contact lines per common ESD DRs make the uneven heat dissipation even worse. Figure 5.5d clearly shows the uneven ESD heating effect along the Z-dimension of the 3D ESD diode where the hotspots are still underneath the plug STI shown in the 2D view of figure 5.5b. However, for the multiple plug STIs, the heat still accumulates around the middle two plugs STIs and especially forms the hotspots on the apace between two plugs STIs in the Z and Y-axis.

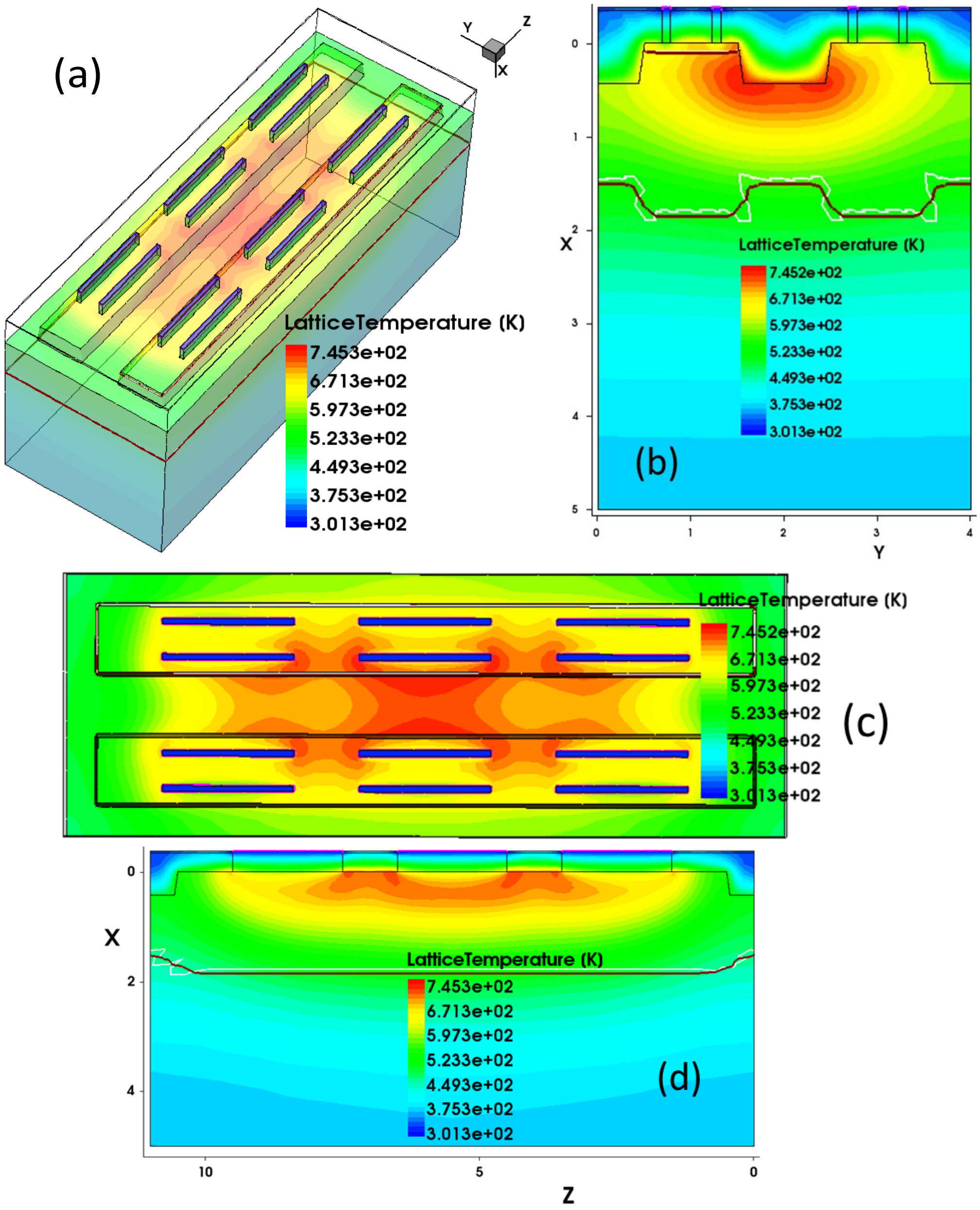


Figure 5.5 3D ESD thermal map by 3D TCAD simulation shows major impact of ESD physical layout design: (a) 3D transparent view, (b) X-Y cross-section view, (c) top view (Y-Z cross-section), and (d) X-Z cross-section view.

In terms of the 2D and pseudo-3D structure, this uneven ESD heating, and thermal distribution effect, mainly along the Z-dimension, simply cannot be revealed in the TCAD

ESD simulation. Figure 5.6a shows ESD thermal map for the 2D ESD diode, equivalent to Figure 5.6c. The hotspots are underneath the plug STI. Figure 5.6b displays the pseudo-3D transparent heat distribution map and we can find the heat evenly distributed along the Z-axis but not evenly on the X-Y plane. Furthermore, figure 6d and figure 6e show an evenly distributed ESD heating and dissipation behavior for the pseudo-3D ESD diode on the Y-

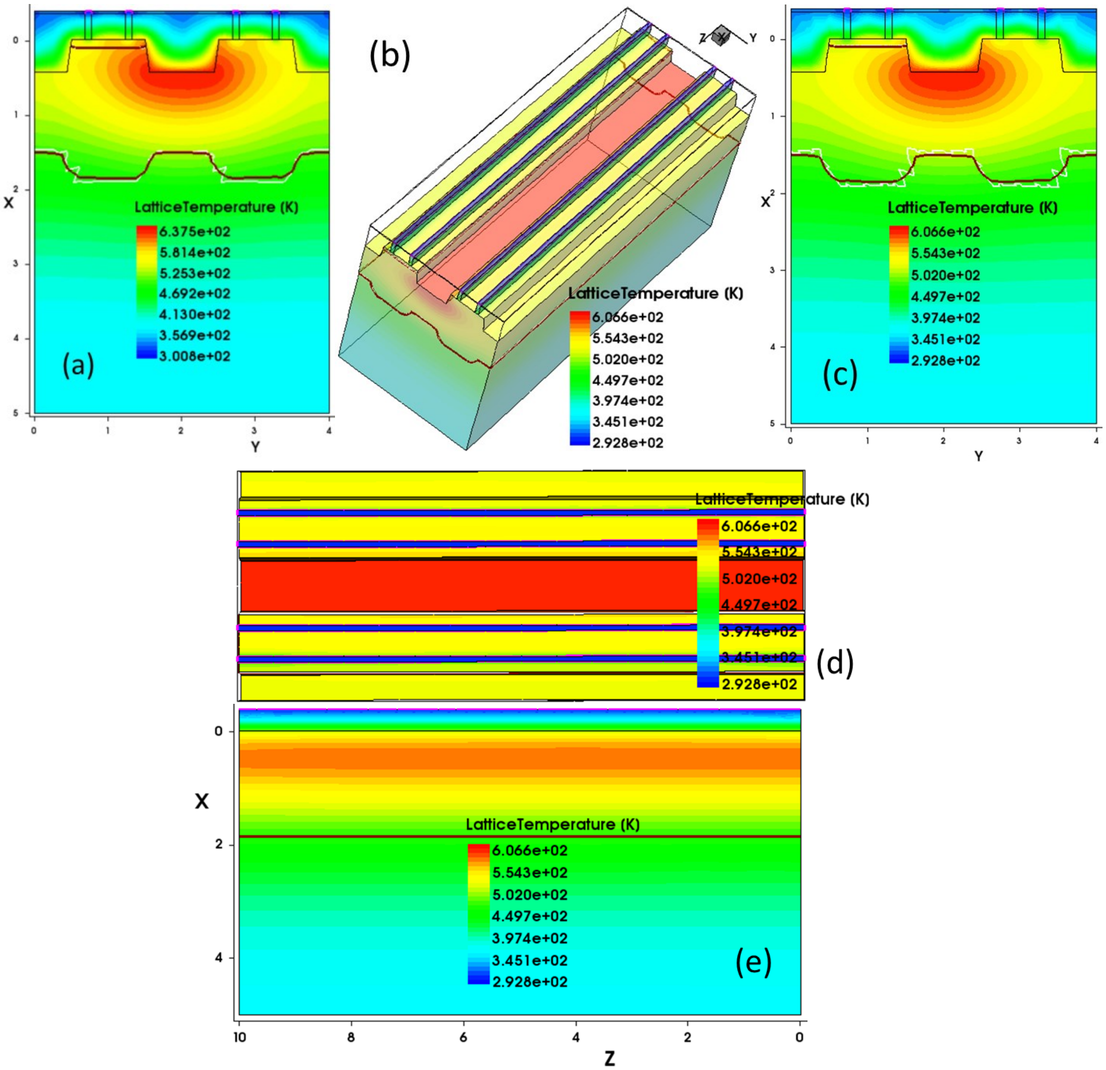


Figure 5.6 ESD thermal map for the ESD diode by 2D/2.5D TCAD ESD simulation: (a) 2D ESD diode thermal map, (b) pseudo 3D view of the 2.5D ESD diode, (c) X-Y cross-section view of the 2.5D ESD diode, (d) Y-Z cross-section view of 2.5D ESD diode thermal map, and (e) X-Z cross-section view of 2.5D ESD diode thermal map.

Z and X-Z plane which is entirely misleading. ESD heating and thermal dissipation are 3D in nature and are seriously affected by the physical layout design and edge/corner effects, which can only be examined by true 3D TCAD ESD simulation that will provide useful ESD design guidelines for practical ESD protection designs. When facing the usual even ESD finger and metal interconnects layout strategy commonly given in the industrial ESD, the design rules based on the 2D or pseudo-3D can cause early ESD failures or grossly over-design of ESD device sizes. In terms of the design guideline, especially for a well-thought-out and unevenly distributed ESD layout design approach, including uneven ESD metal interconnects and ESD finger structures true 3D TCAD ESD simulation, should be adopted in practical ESD protection designs [36].

5.2 Overview of sudoku and finger structures for both SCR and DTSCR

5.2.1 Conventional finger-type SCR and DTSCR structure

Generally, ICs require higher ESD protection robustness, meaning larger ESD device size and more IC performance degradation due to the inherent ESD-induced parasitic effects and IC physical design problems. However, there always exists a paradox for the more robust ESD protection level and less taken area. Accordingly, SCR-type ESD structures are the highest area efficiency ESD devices that have been used for improved ESD robustness [37]. While an inevitable issue about the SCR is the relatively large triggering voltage which makes it not suitable for some circuits. Therefore, the DTSCR ESD structures become attractive to low-voltage (LV) ICs because of the reduced ESD triggering voltage (V_{t1}) offered by low diode triggering voltage though a little lower area efficiency than the SCR but still very high compared with other ESD devices [38].

Nevertheless, conventional finger-type ESD protection device is inherently layout-unfriendly and area-inefficient. In this work, the novel scalable sudoku SCR type ESD devices and finger type SCR devices were designed and fabricated in a foundry 22nm FDSOI CMOS [39] [40]. The FDSOI technology used is a hybrid SOI process allowing bulk Si area for special devices and the SCR type ESD structures in this work are built in such Si area where the bottom oxide is removed. Figure 5.7 shows the conventional finger SCR structure under the TCAD HBM simulation. Figure 5.7a is a 3D view of the only

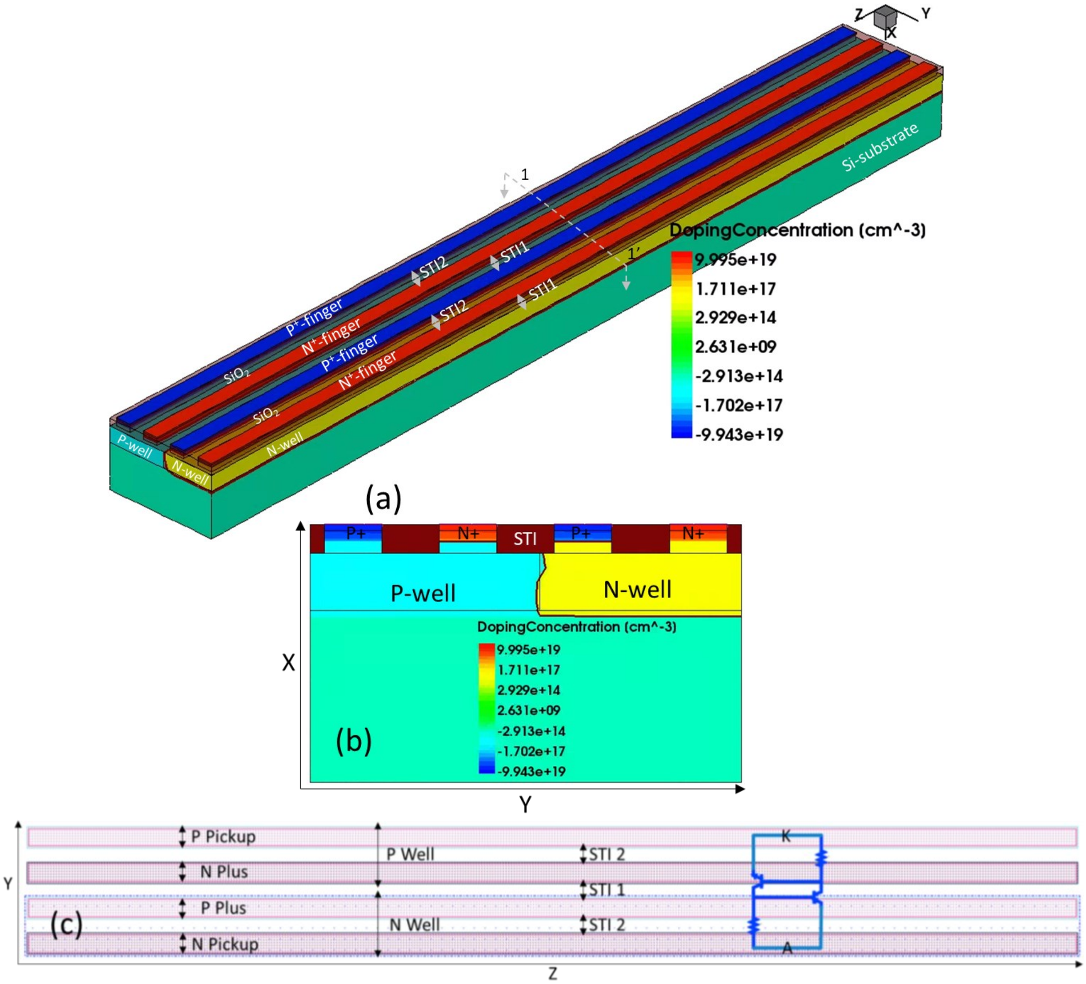


Figure 5.7 A conventional Finger-SCR ESD core array by 3D TCAD where SCR ESD devices are formed across the cell boundaries: (a) 3D view, (b) cross-section view on X-Y plane along the 1-1' cutline, and (c) actual Y-Z plane layout view with equivalent circuit.

silicon part with N plus finger and P plus finger above the N well and P well separately, which has the length of 1µm and width of 60µm. There are three STIs made of SiO_2 separating the four-plus fingers with the same length and width. Considering the N well and P well blocks, the total area of the finger SCR is 420µm^2. Figure 5.7b shows the cross-section view of the 1-1’ cutline in the 3D view from figure 5.7a. It is found that half of N plus and P plus are embedded in the P well and the other in the N well forming the parasitic BJTs which are shown in figure 5.7c. Figure 5.7c displays the real layout with the top Y-Z plane and also marks the equivalent circuits based on the 22nm FDSOI.

Similarly, the finger DTSCR is also built in the 3D TCAD simulation shown in figure 5.8. Figure 5.8a is a 3D view of the only silicon part with N plus finger and P plus finger above the N well and P well separately, which has the length of 1µm and width of 60µm. What’s more, we put the triggering diode having N plus and P plus under the N well to form the P plus N well diode separated by the STI. The width of the diode is still 60µm but the length of the diode including N plus, Plus and STI inside the diode is 0.5µm. There are four STIs made of SiO_2 separating the four plus fingers and diode with the same length and width. Considering the N well and P well blocks, the total area of the finger DTSCR is 570µm^2. Figure 5.8b shows the cross-section view of the 1-1’ cutline in the 3D view from figure 5.8a. It is found that half of N plus and P plus are embedded in the P well and the other in the N well forming the parasitic BJTs connecting with the triggering N well diode which is shown in figure 5.8c. Figure 5.8c displays the real layout with the top Y-Z plane and also marks the equivalent circuits based on the 22nm FDSOI.

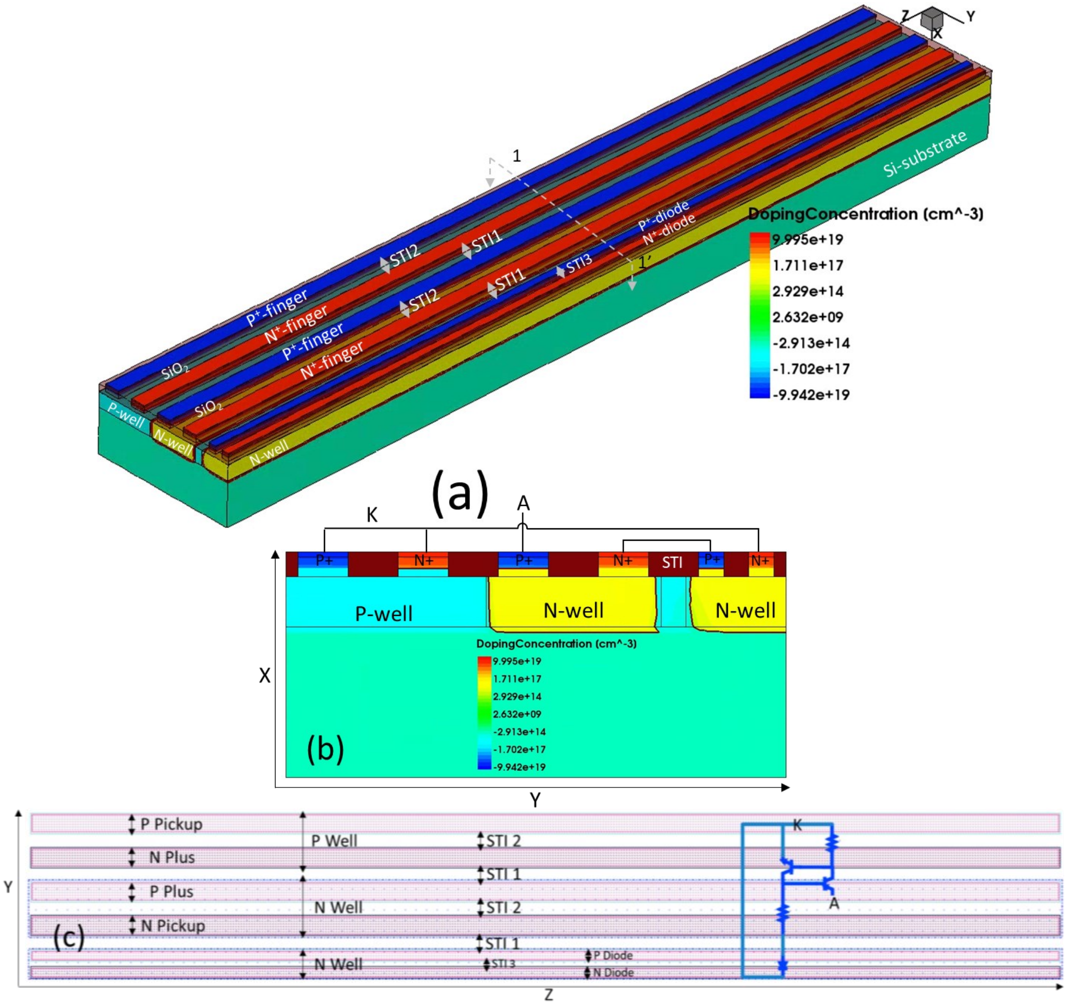


Figure 5.8 A conventional Finger-DTSCR ESD core array by 3D TCAD where SCR ESD devices are formed across the cell boundaries: (a) 3D view, (b) cross-section view on X-Y plane along the 1-1’ cutline, and (c) actual Y-Z plane layout view with equivalent circuit.

5.2.2 Sudoku-type SCR and DTSCR core structure

The design splits include sudoku SCR core and LV DTSCR ESD arrays with varying dimensions for comparison. To optimize and evaluate the novel Sudoku devices, true 3D TCAD ESD simulation is required because the area efficiency cannot be meaningfully studied without carefully considering the edge/corner effects of the array structures. Figure. 5.9 depicts an exemplar sudoku SCR ESD core array with 3X3 cells. There are two types of SCR ESD cells, each contains a central P+ diffusion pickup surrounded by an N+ diffusion ring residing in a P-well and a central N+ diffusion pickup surrounded by a P+ diffusion ring inside an N-well, respectively shown in the figure 5.9a which is a 3D view of the sudoku SCR. As depicted in the cross-section view shown in figure 5.9b, a working SCR ESD structure is formed across the boundary of two adjacent SCR cells of different types, with two electrodes of the anode (A) and cathode (K). It is found in the figure 5.9c that the dimensions for the central P+/N+ diffusions and the surrounding N+/P+ diffusion rings are set to 1μm and the isolation between the central P+ and its surrounding N^+-ring (same for the central N+ and its surrounding P^+-ring) is defined as the inner-cell isolation (STI2) and the inter-cell isolation (STI1) with the same width of 1μm. For a large Sudoku SCR ESD array, each inner ESD cell has four across-boundary working SCR ESD devices, while the edge ESD cells have three and the corner cells have only two across-boundary working SCR ESD devices. Therefore, there is a total of 12 effective discharging edges with a total width of 60um equalling the finger SCR. We can find the equivalent circuit in figure 5.9c revealing the parasitic BJTs in the sudoku SCR. Compared to conventional finger-type SCR ESD devices, the sudoku SCR ESD array has higher area efficiency in terms of ESD discharging. Which has a dimension of 17μm on each side and a total array area of 289μm^2.

Based on the sudoku SCR ESD core structure, new sudoku DTSCR ESD arrays were designed by 3D TCAD ESD simulation. Figure 5.10 depicts the 3D sudoku DTSCR ESD array created by the 3D TCAD process simulation. As seen from figure 5.20a, the 3X3 sudoku DTSCR ESD protection array is similar to a sudoku-SCR core array except that

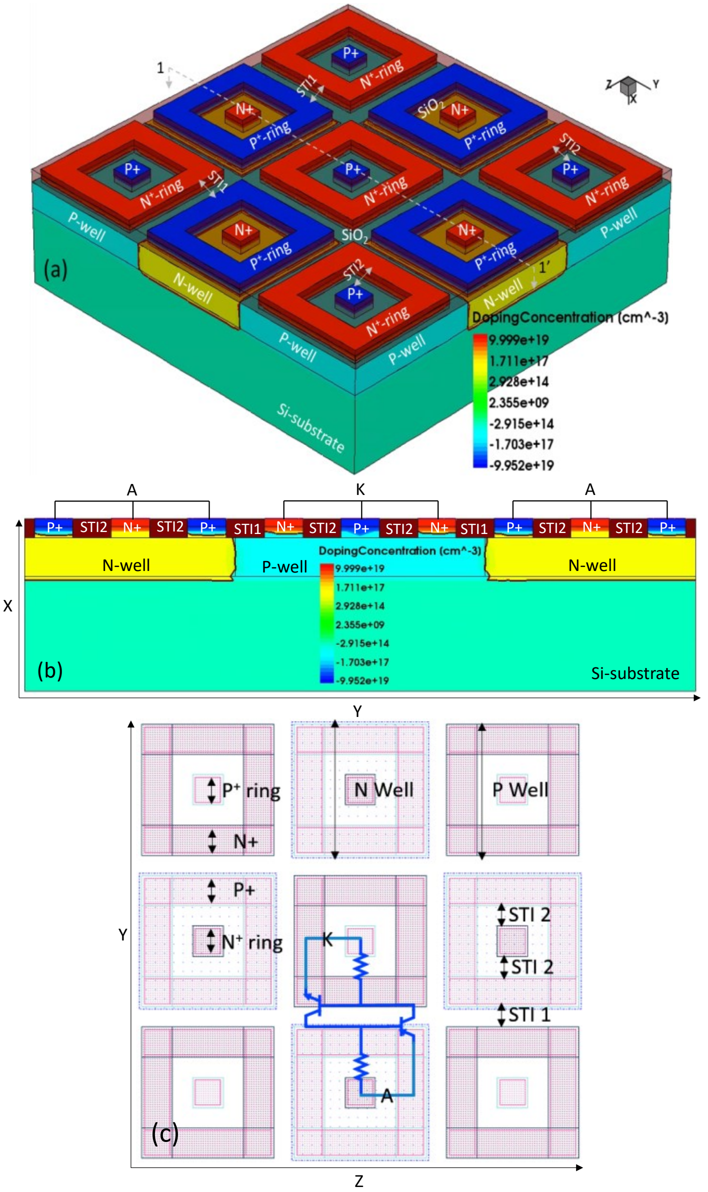


Figure 5.9 A 3X3 sudoku SCR ESD core array by 3D TCAD where SCR ESD devices are formed across the cell boundaries: (a) 3D view, (b) cross-section view on X-Y plane along the 1-1' cutline, and (c) actual Y-Z plane layout view with equivalent circuit.

ESD triggering diodes are integrated into the four sides of the array to reduce the ESD V_{t1} for LV ICs. In addition, in order to connect the triggering diode into the core cell part, we make the edge and corner cells feature an "open" layout for better interconnection. It is noted that we still keep the same effective dimension of the discharging path same as the sudoku SCR array. The sudoku DTSCR ESD array still features 12 across-boundary ESD discharging paths, which has an effective width of 5μm to compose a total of 60μm shown

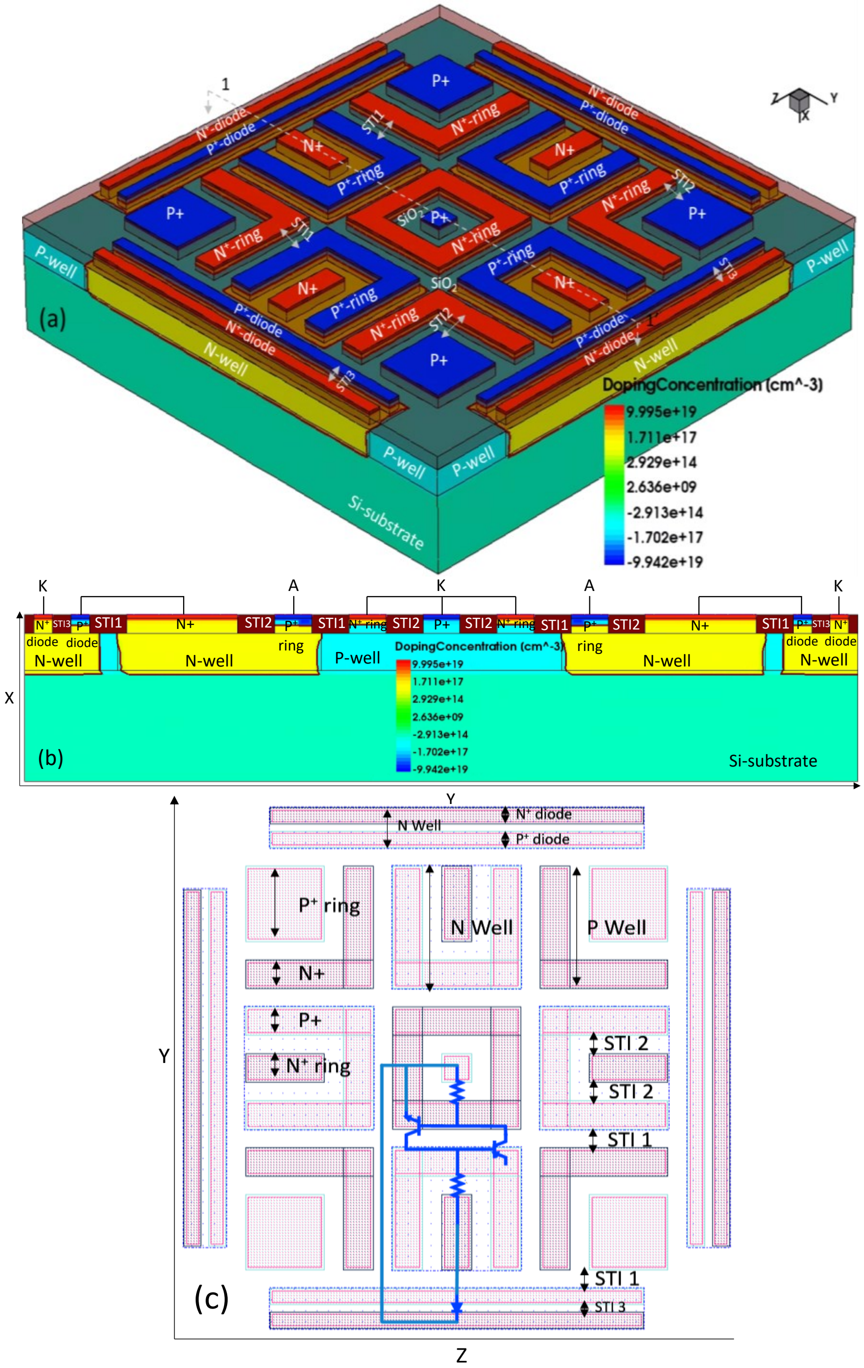


Figure 5.10 A 3X3 sudoku DTSCR ESD array by 3D TCAD where DTSCR ESD devices are formed across the cell boundaries: (a) 3D view, (b) cross-section view on X-Y plane along the 1-1’ cutline, and (c) actual Y-Z plane layout view with equivalent circuit.

in figure 5.10c, which equals to the finger type of DTSCR. Figure 5.10b shows the cross-section view of the 1-1' cutline from figure 5.10a, which reveals the triggering N well diode isolated by the P well guarding. It is found that the dimensions for the central P+/N+ diffusions and the surrounding N+/P+ diffusion rings are set to 1µm and the isolation between the central P+ and its surrounding N^+-ring (same for the central N+ and its surrounding P^+-ring) is defined as the inner-cell isolation (STI2) and the inter-cell isolation (STI1) with the same width of 1µm like before. Furthermore, we put the triggering diode having N plus and P plus under the N well to form the P plus N well diode separated by the STI. The width of each diode is 15µm but the length of the diode including N plus, Plus and STI inside the diode is 0.5µm. Therefore, the total with of the diode is 60µm the same as the finger type DTSCR for the rational comparison later. Also, the equivalent circuit including the parasitic BJTs is depicted in figure 5.10c and the sudoku DTSCR ESD array has a total area of about 484µm^2.

5.3 3D TCAD transient HBM ESD simulation analysis

5.3.1 HBM simulation comparison of sudoku array and finger SCR

Transient 3D ESD simulation was conducted by stressing the sudoku SCR ESD core array structure and conventional finger type SCR structure with a real HBM waveform of

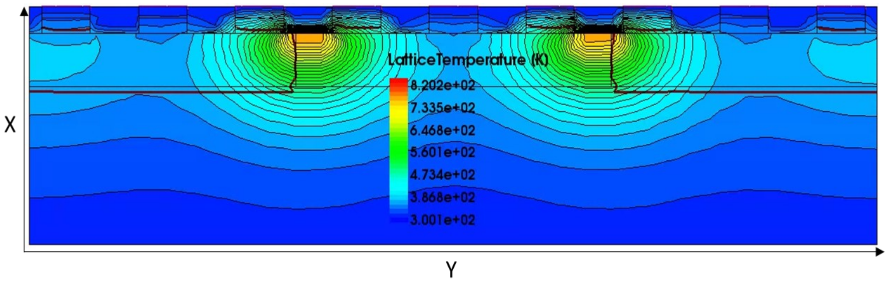


Figure 5.11 Cross-section view of transient ESD heating for the 3X3 Sudoku-SCR ESD array under 5kV HBM ESD stressing reveals transient ESD discharging behaviors.

5KV. Figure 5.11 depicts the cross-section view of the transient ESD discharge heating map across two adjacent SCR cells. We can see that the main hotspots are located in the interface of the N well and P well underneath the STI. Since the parasitic BJTs are induced in this area which discharges lots of ESD current when the BJTs are turned on under HBM zapping, a large amount of heat is generated from here shown in figure 5.11 which depicts two heat accumulated areas around the boundary of two adjacent cells.

Figure 5.12a and figure 5.12b present the transient 3D ESD discharging current density and lattice temperature contours, showing the critical 3D ESD discharging behaviors across the ESD cell boundaries. The high current density will stand where the main discharging

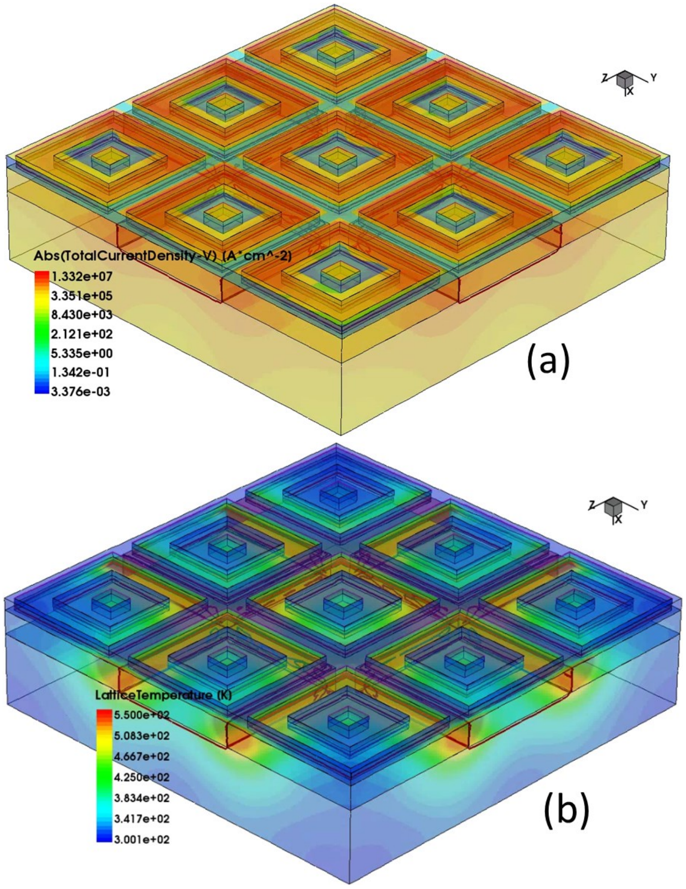


Figure 5.12 3D ESD TCAD simulation under 5kV HBM stressing for 3X3 sudoku SCR ESD core array: (a) 3D ESD discharging I-density map, and (b) 3D lattice temperature T-map.

current flow through and the high temperature will show the potential current handleability where it determines due to the local hot spots. Clearly, for the inner ESD cells, both transient ESD discharging current and heating peak across the cell boundaries. However, the outer edges of the edge/corner cells do not contribute to ESD discharging, hence, showing lower transient ESD heating. Specifically, there exist 12 ESD discharging channels in this 3X3 Sudoku-SCR ESD core array. It is obvious that while the Sudoku-SCR ESD core array is generally area-efficient in ESD discharging, the outer edges of the edge/corner cells reduce the total area efficiency, which will be discussed later. Figure 5.13 shows the comparison of conventional finger type SCR transient simulation under 5kV

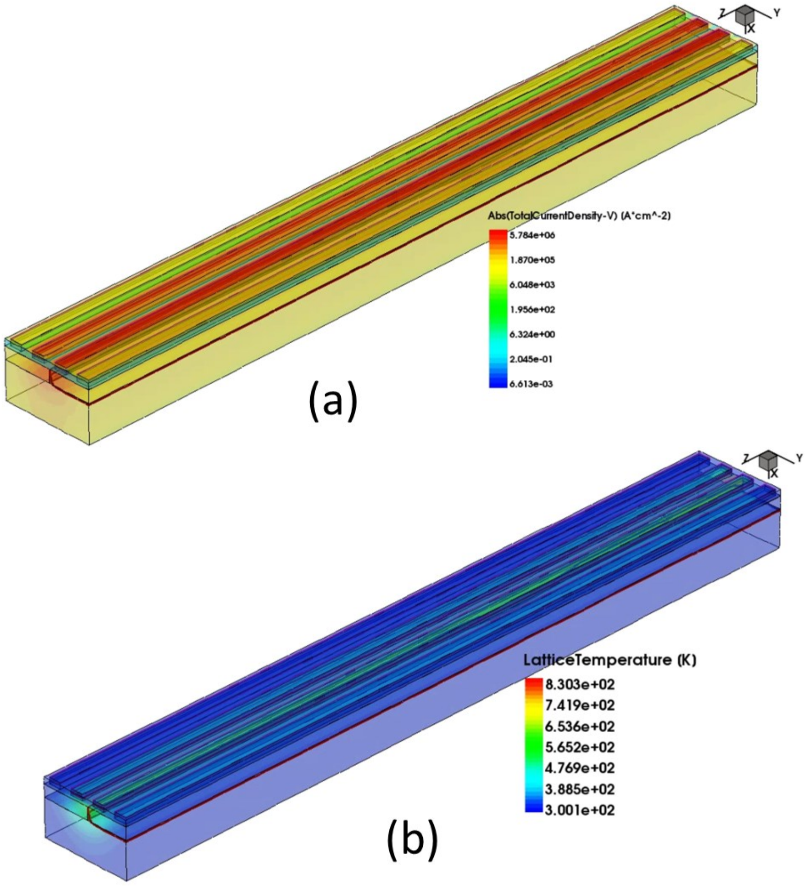


Figure 5.13 3D ESD TCAD simulation under 5kV HBM stressing for finger SCR ESD device: (a) 3D ESD discharging I-density map, and (b) 3D lattice temperature T-map.

HBM stressing. It is found in the current density map that the main current accumulates in the long boundary of N well and P well still underneath the inter STI due to the parasitic BJTs. Meanwhile, the highest temperature arises in the same place due to the poor thermal conductivity of silicon dioxide. The very long edge side does not contribute to the current discharging which results in the lower area efficiency compared to the sudoku SCR array.

Figure 5.14a and figure 5.14b depict the transient ESD discharging I-V and T_{max}-t characteristics for the 3X3 sudoku SCR ESD core array structure and its finger-SCR core device counterpart, respectively. The HBM simulated I-V curves show that Sudoku-SCR has similar ESD triggering voltage (V_{t1}), holding voltage (V_h), and ESD discharging resistance (R_{on}) as that for the finger-SCR structure. However, it is readily observed that the peak lattice temperature for the Sudoku-SCR structure (664℃) is much lower than that for its finger-SCR counterpart device (872℃), indicating a much higher ESD current handling capability for the Sudoku-SCR ESD core, hence, higher ESD protection robustness.

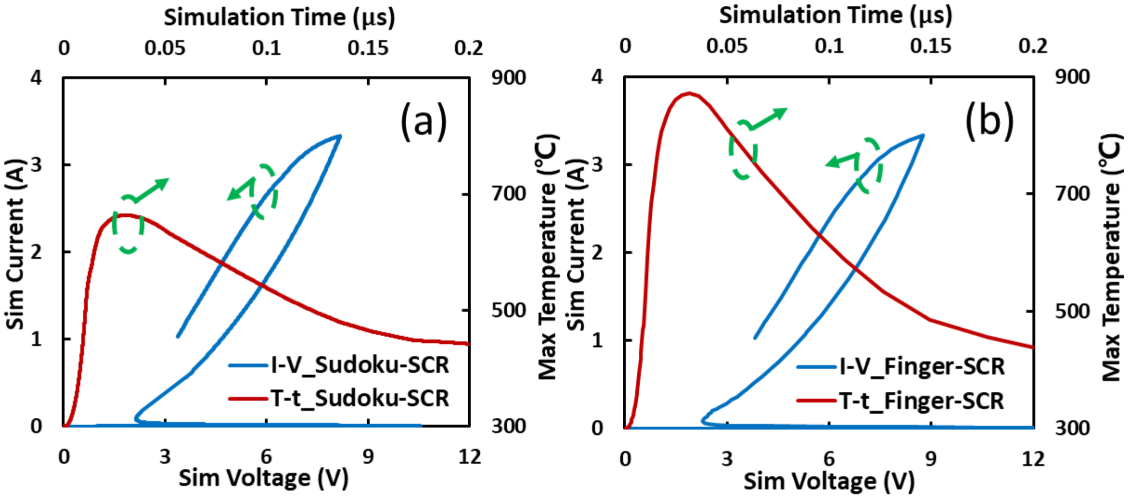


Figure 5.14 HBM-simulated I-V and T_{max}-t curves based on transient 3D TCAD ESD simulation for (a) 3X3 sudoku SCR ESD core array structure, and (b) finger SCR ESD core device.

5.3.2 HBM simulation comparison of sudoku array and finger DTSCR

In terms of the sudoku DTSCR core array, a 3D transient ESD simulation was conducted for the Sudoku-DTSCR ESD array structure using an HBM waveform of 5KV. Figure 5.15 depicts the cross-section view of the transient ESD discharge heating map across two adjacent DTSCR cells. We can see that the main hotspots are still located in the interface of N well and P well underneath the STI instead of neat the triggering diode area because the diode just conducts a little current to turn on the parasitic BJTs at the beginning. Seen from figure 5.15, depicts two heat accumulated areas around the boundary of two adjacent cells.

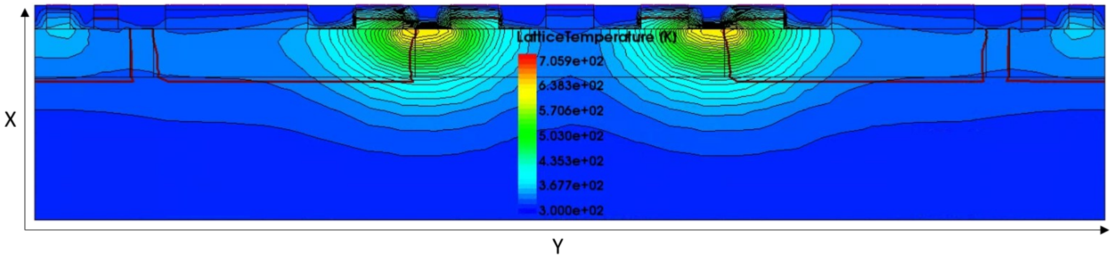


Figure 5.15 Cross-section view of transient ESD heating for the 3X3 sudoku DTSCR ESD array under 5kV HBM ESD stressing reveals transient ESD discharging behaviors.

Figure 5.16a and Figure 5.16b present the transient 3D ESD discharging current density and lattice temperature contours. The detailed 3D ESD discharging behaviors are readily observed across the cell boundaries where transient ESD heating also peaks. There are 12 ESD discharging channels in the 3X3 sudoku DTSCR ESD array. Similar to a Sudoku-SCR core array, the main contribution for ESD discharging comes from the inner cells because the edge/corner cells do not have all four cell boundaries conducting ESD pulses. The high current density will stand where the main discharging, current flow through, and the high temperature will show the potential current handleability where it determines due to the local hot spots. Clearly, for the inner ESD cells, both transient ESD discharging current and heating peak across the cell boundaries. However, the outer edges of the edge/corner cells do not contribute to ESD discharging, hence, showing lower transient ESD heating. What's more, the triggering diode does not contribute to the large current discharging either which will decrease the efficiency of this 3X3 device. However, according to the scalability discussed later, the effect of non-contribution can be eliminated a lot. Figure 5.17 shows the counterpart of conventional finger type SCR transient simulation under 5kV HBM stressing. It is found in the current density map that the main

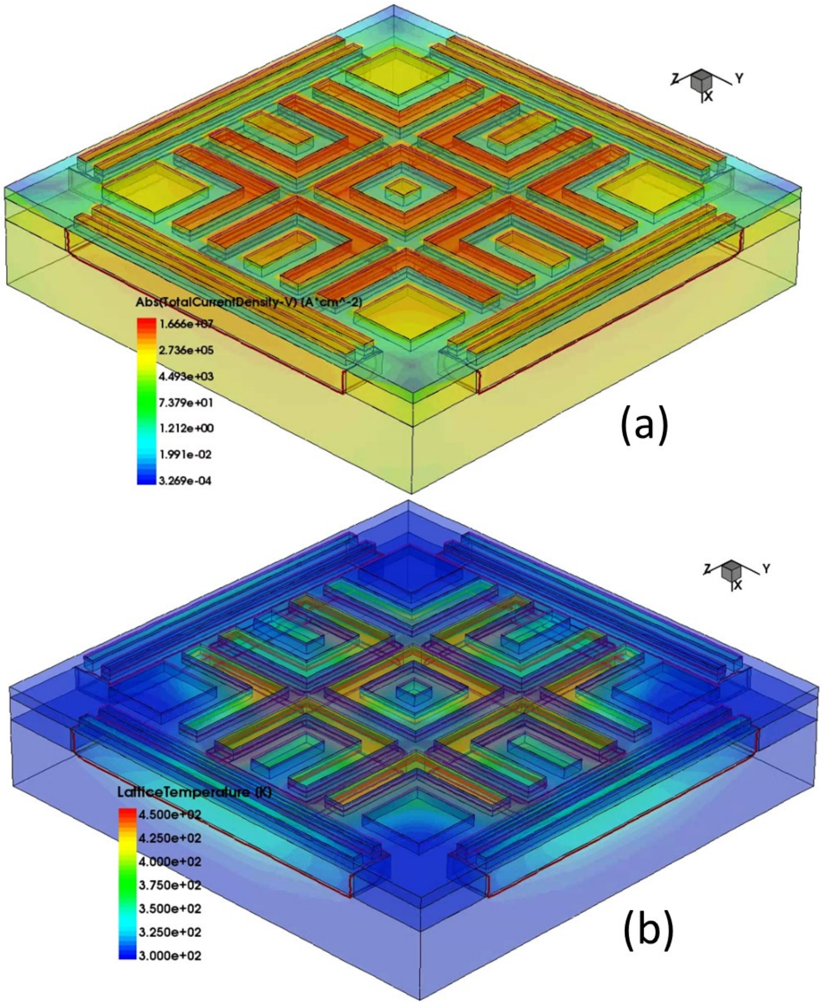


Figure 5.16 3D ESD TCAD simulation under 5kV HBM stressing for 3X3 sudoku DTSCR ESD core array: (a) 3D ESD discharging I-density map, and (b) 3D lattice temperature T-map.

current accumulates in the long boundary of N well and P well still underneath the inter STI due to the parasitic BJTs. Meanwhile, the highest temperature arises in the same place due to the poor thermal conductivity of silicon dioxide. The very long edge side including the triggering diode part does not contribute to the current discharging which results in the lower area efficiency compared to the sudoku DTSCR array.

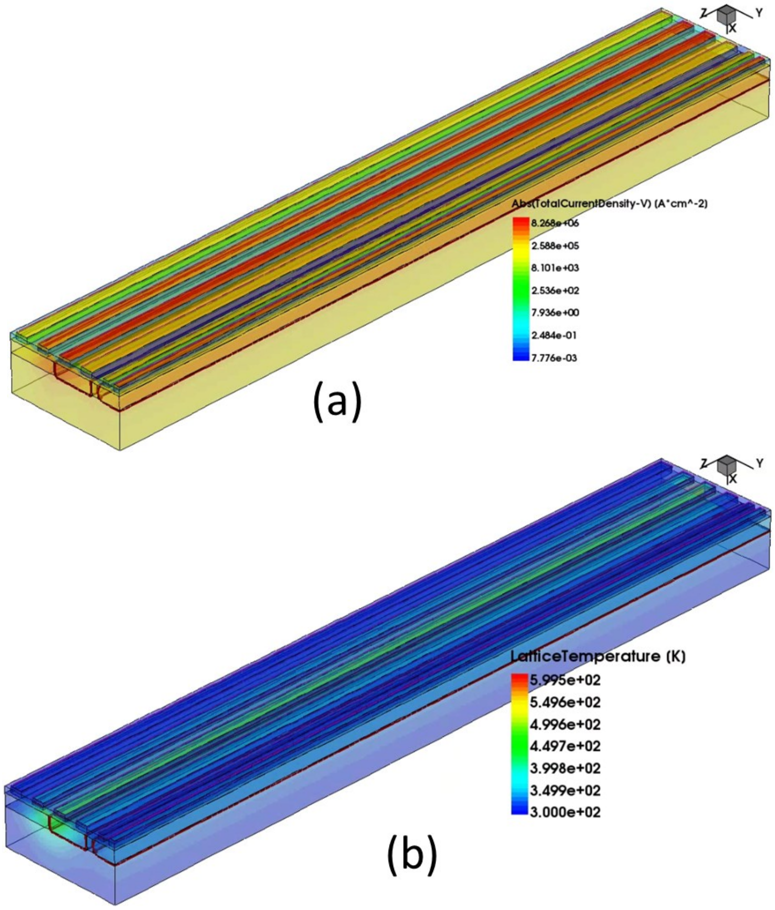


Figure 5.17 3D ESD TCAD simulation under 5kV HBM stressing for finger DTSCR ESD device: (a) 3D ESD discharging I-density map, and (b) 3D lattice temperature T-map.

Figure 5.18a and figure 5.18b present the transient ESD discharging I-V and T_{max}-t curves for the 3X3 sudoku DTSCR ESD array structure in comparison with its finger-DTSCR device counterpart, respectively. The HBM simulated I-V curves show that

Sudoku-DTSCR has similar ESD V_{t1}, V_{h}, and R_{on} as that for the finger-DTSCR structure. However, it is clearly observed that the peak lattice temperature is 610°C for the Sudoku-DTSCR array that is much lower than 718°C for its finger-DTSCR counterpart, hence, the Sudoku-DTSCR ESD array is much more robust for ESD protection. Meanwhile, combining figure 5.14, we can see the DTSCR has a much lower triggering voltage Vt1 than that of the SCR no matter on the sudoku or finger structures, which validates the effectiveness of triggering diode integrated into sudoku SCR core array to forming sudoku DTSCR device.

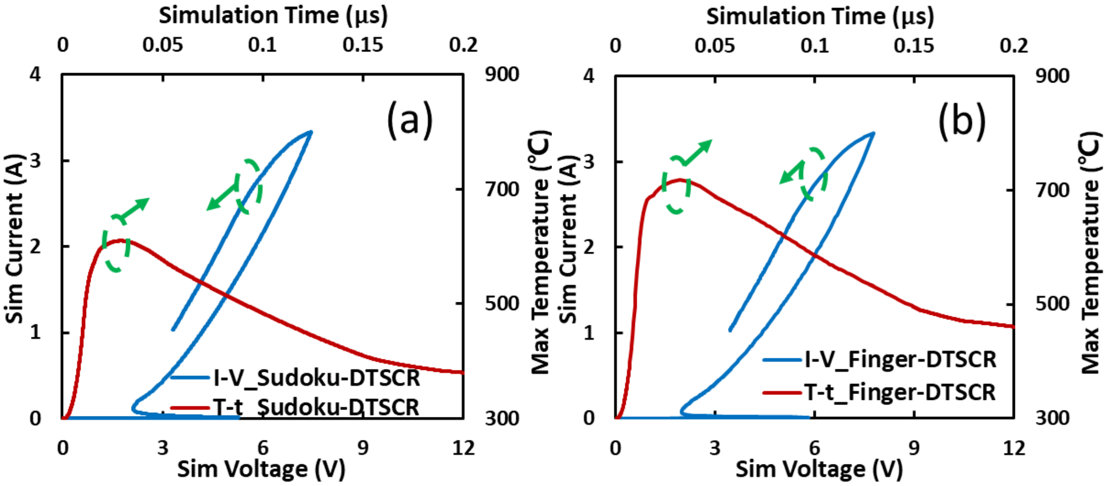


Figure 5.18 HBM simulated I-V and T_{max}-t curves based on transient 3D TCAD ESD simulation for (a) 3X3 sudoku DTSCR ESD core array structure, and (b) finger DTSCR ESD core device.

5.4 3D TCAD transient CDM ESD simulation analysis

5.4.1 CDM simulation comparison of sudoku array and finger SCR

Similarly, a 500V CDM transient ESD simulation was conducted for the Sudoku-SCR core array by 3D TCAD. The CDM has a much faster surging pulse than HBM and can be used to demonstrate the sudoku array structure performance under such fast transient stressing. Figure 5.19a and figure 5.19b depict the transient 3D ESD discharging current density and lattice temperature contours. It clearly shows that the inner ESD cells discharge ESD transient across the cell boundaries where both discharging current and heating peak. The outer edges of the edge/corner cells do not contribute to ESD discharging, hence, showing no transient ESD heating. The high current density will stand where the main discharging current flow through and the high temperature will show the potential current handleability where it determines due to the local hot spots. Clearly, for the inner ESD cells, both transient ESD discharging current and heating peak across the cell boundaries. However, the outer edges of the edge/corner cells do not contribute to ESD discharging, hence, showing lower transient ESD heating. For the counterpart of finger SCR devices,

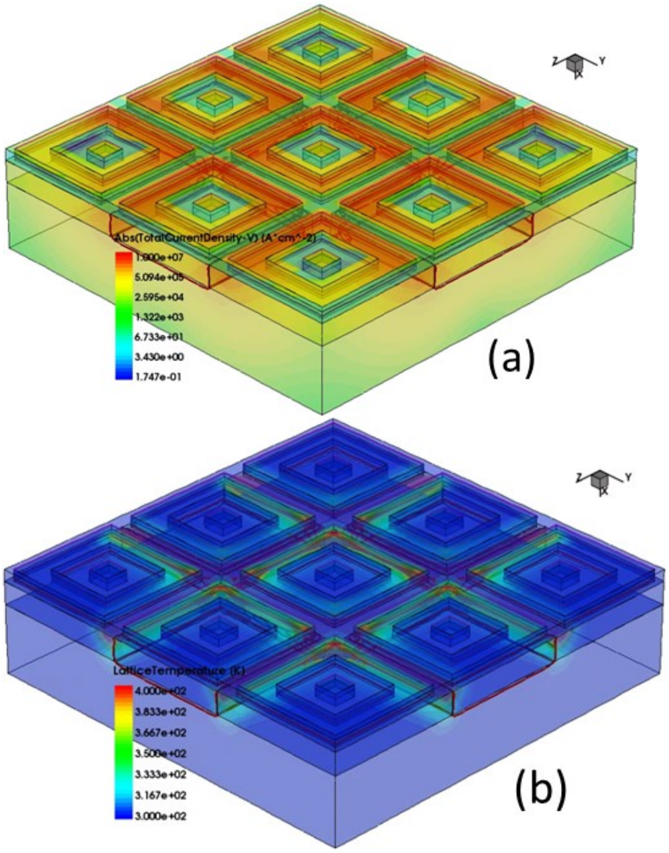


Figure 5.19 3D ESD TCAD simulation under 500V CDM stressing for 3X3 sudoku SCR ESD core array: (a) 3D ESD discharging I-density map, and (b) 3D lattice temperature T-map.

under the 500V CDM stressing, figure 5.20 shows the current density and maximum temperature distribution. Similar to the HBM situation, it is found in the current density map that the main current accumulates in the long boundary of N well and P well still underneath the inter STI due to the parasitic BJTs. Meanwhile, the highest temperature arises in the same place due to the poor thermal conductivity of silicon dioxide. It is noted that compared with figure 5.12 and figure 5.13 for both finger and sudoku SCR devices, the temperature under CDM stressing is not as high as that under HBM stressing due to less time duration for the heat accumulation from the much faster CDM surging pulse.

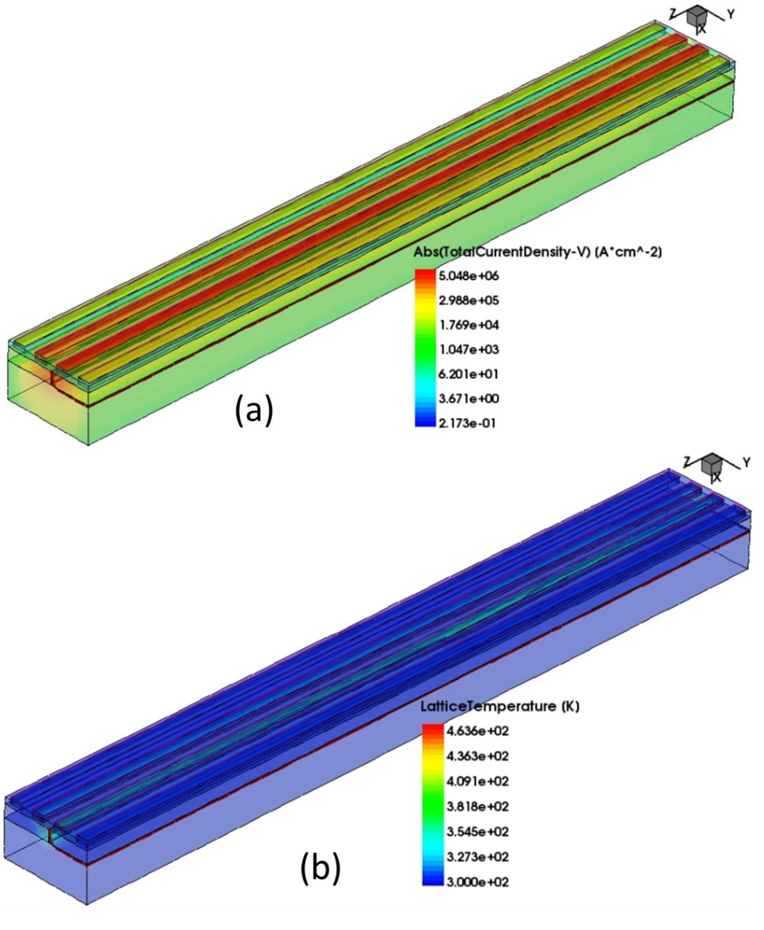


Figure 5.20 3D ESD TCAD simulation under 500V CDM stressing for finger SCR ESD device: (a) 3D ESD discharging I-density map, and (b) 3D lattice temperature T-map.

Figure 5.21a and figure 5.21b present the transient ESD discharging I-V and T_{max}-t characteristics for the 3X3 sudoku SCR ESD core array and its finger-SCR counterpart, respectively. It shows similar R_{on}, V_{t1}, and V_h for sudoku SCR and fingers SCR. However, the peak T_{max} of ~841°C for sudoku SCR is much lower than T_{max} of ~1120°C for the finger-SCR device, confirming that sudoku SCR is more area-efficient than its finger SCR counterpart for CDM ESD protection. This indicates a much higher ESD current handling capability for the Sudoku-SCR ESD core, hence, higher ESD protection robustness.

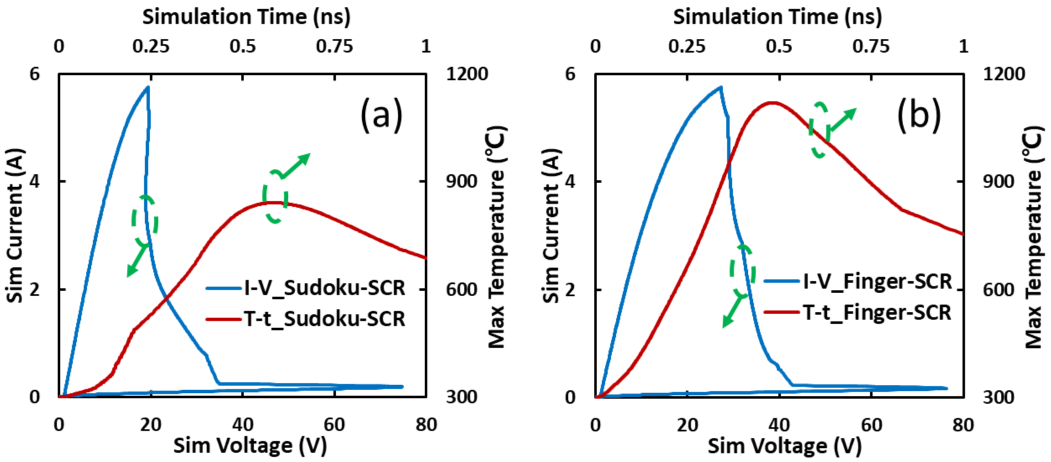


Figure 5.21 CDM simulated I-V and T_{max}-t curves based on transient 3D TCAD ESD simulation for (a) 3X3 sudoku SCR ESD core array structure, and (b) finger SCR ESD core device.

5.4.2 CDM simulation comparison of sudoku array and finger DTSCR

Likewise, a 500V CDM transient ESD simulation was conducted for the Sudoku-DTSCR ESD array structure. Figure 5.22a and Figure 5.22b present the transient 3D ESD discharging current density and lattice temperature maps. Clearly, CDM ESD discharge occurs across four cell boundaries for the inner ESD cells where both discharging current and heating peak simultaneously, while the outer edges of the edge/corner cells do not discharge ESD currents. The high current density will stand where the main discharging, current flow through, and the high temperature will show the potential current handleability where it determines due to the local hot spots. Clearly, for the inner ESD cells, both transient ESD discharging current and heating peak across the cell boundaries. However, the outer edges of the edge/corner cells do not contribute to ESD discharging, hence, showing lower transient ESD heating. What's more, the triggering diode does not contribute to the large current discharging either which will decrease the efficiency of this 3X3 device. However, according to the scalability discussed later, the effect of non-contribution can be eliminated a lot. Figure 5.23 shows the counterpart of conventional

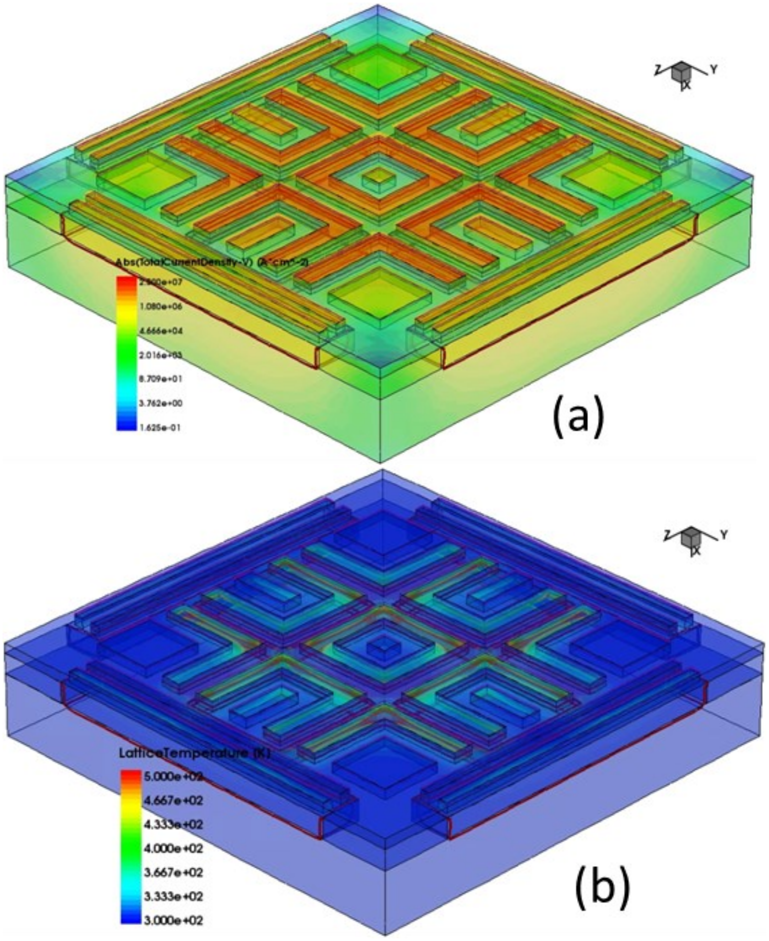


Figure 5.22 3D ESD TCAD simulation under 500V CDM stressing for 3X3 sudoku DTSCR ESD core array: (a) 3D ESD discharging I-density map, and (b) 3D lattice temperature T-map.

finger type SCR transient simulation under 500V CDM stressing. It is found in the current density map that the main current accumulates in the long boundary of N well and P well still underneath the inter STI due to the parasitic BJTs. Meanwhile, the highest temperature arises in the same place due to the poor thermal conductivity of silicon dioxide. The very long edge side including the triggering diode part does not contribute to the current discharging which results in the lower area efficiency compared to the sudoku DTSCR array.

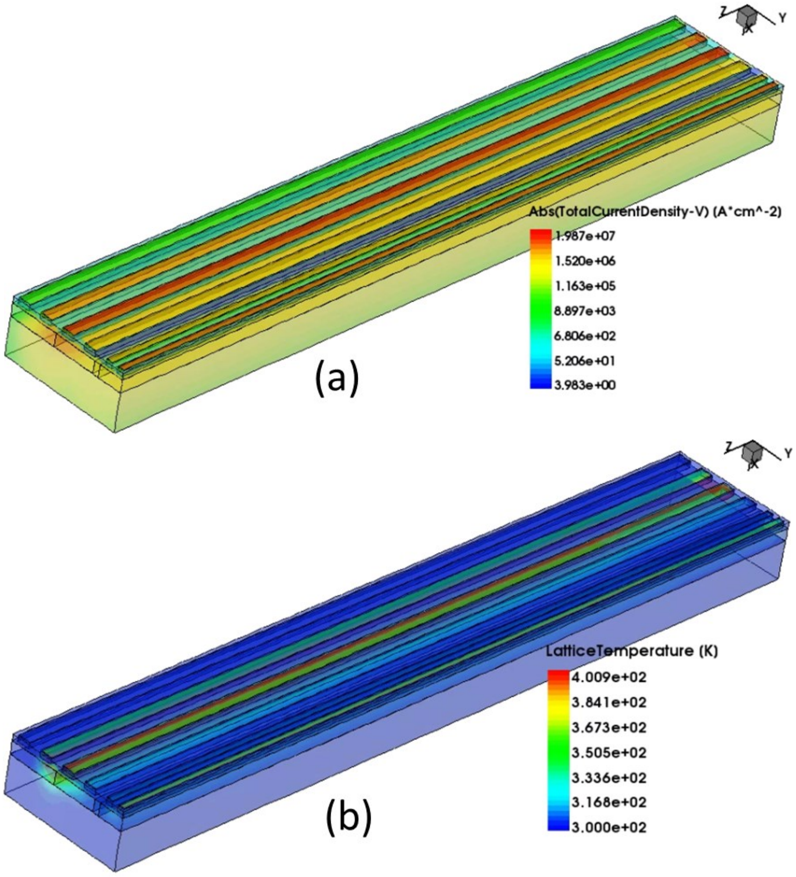


Figure 5.23 3D ESD TCAD simulation under 500V CDM stressing for finger DTSCR ESD device: (a) 3D ESD discharging I-density map, and (b) 3D lattice temperature T-map.

Figure. 5.24a and Figure 5.24b give the transient ESD discharging I-V and T_{max}-t curves for the 3X3 sudoku DTSCR ESD array and its finger DTSCR counterpart, respectively. It shows similar R_{on}, V_{t1}, and V_{h}. for both devices. However, Sudoku-DTSCR has a lower peak T_{max} of ~752℃ than T_{max} of ~1102℃ for its finger-DTSCR counterpart, confirming higher CDM ESD area efficiency for the Sudoku-DTSCR array. It is worth noting that, comprehensive TCAD ESD simulation calibration for accuracy requires process details and Si iterations, which unfortunately were unavailable in this work. Therefore, readers are advised to consider the trends of ESD characteristics suggested by TCAD simulation, not any accurate values.

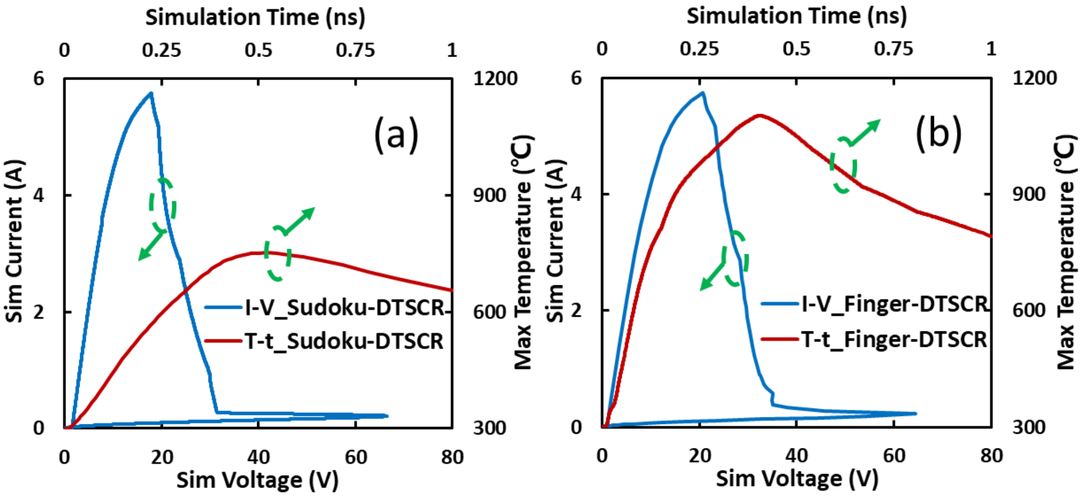


Figure 5.24 CDM simulated I-V and T_{max}-t curves based on transient 3D TCAD ESD simulation for (a) 3X3 sudoku DTSCR ESD core array structure, and (b) finger DTSCR ESD core device.

5.5 Silicon validation on different ESD devices

5.5.1 TLP measurement and analysis

We set the rise time of 10ns and pulse width of 100ns TLP configuration to conduct for HBM protection evaluation. Figure. 5.25a compares measured ESD discharging I-V curves by TLP testing for the Sudoku-SCR ESD core array structure and its finger-SCR ESD device counterpart. It is found in the figure 5.25a that the Sudoku-SCR ESD array has significantly higher ESD current handling capability (I_{t2}) and ESD protection area efficiency (J_{t2}) over the finger-SCR ESD device counterpart, i.e., I_{t2}~3.50A (J_{t2}~12.1mA/µm^2) for Sudoku-SCR array and I_{t2}~2.48A (J_{t2}~5.9mA/µm^2) for the finger-SCR device. Similarly, Figure 5.25b depicts the TLP-measured ESD discharging I-V characteristics for the Sudoku-DTSCR ESD array structure and its finger-DTSCR ESD device counterpart. It shows I_{t2}~3.13A (J_{t2}~6.47mA/µm^2) for Sudoku-DTSCR array and I_{t2}~2.24A (J_{t2}~3.93mA/µm^2) for its finger-DTSCR counterpart, again confirming much higher ESD discharging area efficiency for the novel Sudoku-DTSCR array structure. TLP testing shows that the Sudoku-SCR array dramatically improves the ESD area efficiency by ~105% over its finger-SCR device counterpart, while Sudoku-DTSCR ESD arrays increase the ESD area efficiency by ~64.6% over traditional finger-DTSCR devices, respectively. Furthermore, we can find that the triggering voltage of the DTSCR is much lower than the SCR for both the sudoku and finger devices. This testing validates under the

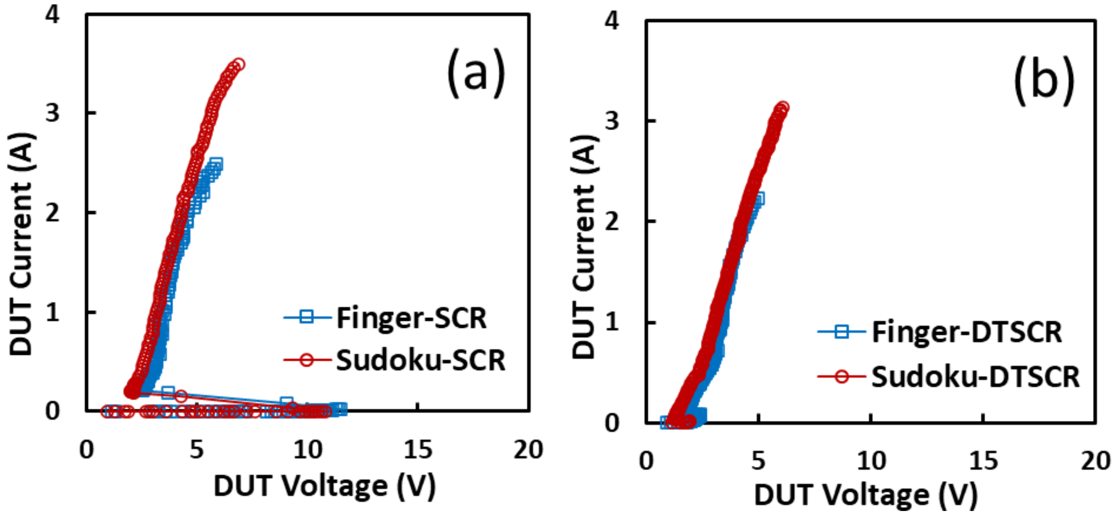


Figure 5.25 Comparison of TLP measured ESD discharging I-V curves for: (a) sudoku SCR ESD array and its finger SCR device counterpart, (b) sudoku DTSCR ESD array and its finger DTSCR device counterpart.

TLP stressing the much-improved high efficiency of sudoku devices upon the finger devices no matter on SCR or DTSCR.

5.5.2 VFTLP measurement and analysis

We set the rise time of 100ps and pulse width of 1ns VFTLP configuration to conduct CDM protection evaluation. Figure 5.26a compares measured ESD discharging I-V curves by VFTLP testing for the Sudoku-SCR ESD core array structure and its finger-SCR ESD device counterpart. It is readily observed that the Sudoku-SCR ESD array has higher ESD current handling capability and area efficiency over its finger-SCR counterpart. Figure 5.26b, clearly shows that the Sudoku-DTSCR ESD array structure is much more ESD robust and has much higher ESD discharging area efficiency, i.e., I_{t2}~13.1A (J_{t2}~27.1mA/μm^2) for Sudoku-DTSCR array and I_{t2}~9.92A (J_{t2}~17.4mA/μm^2) for finger-DTSCR device counterpart. VFTLP testing confirms that Sudoku-SCR array improves the ESD area efficiency by ~84.8% over its finger-SCR device counterpart, while Sudoku-

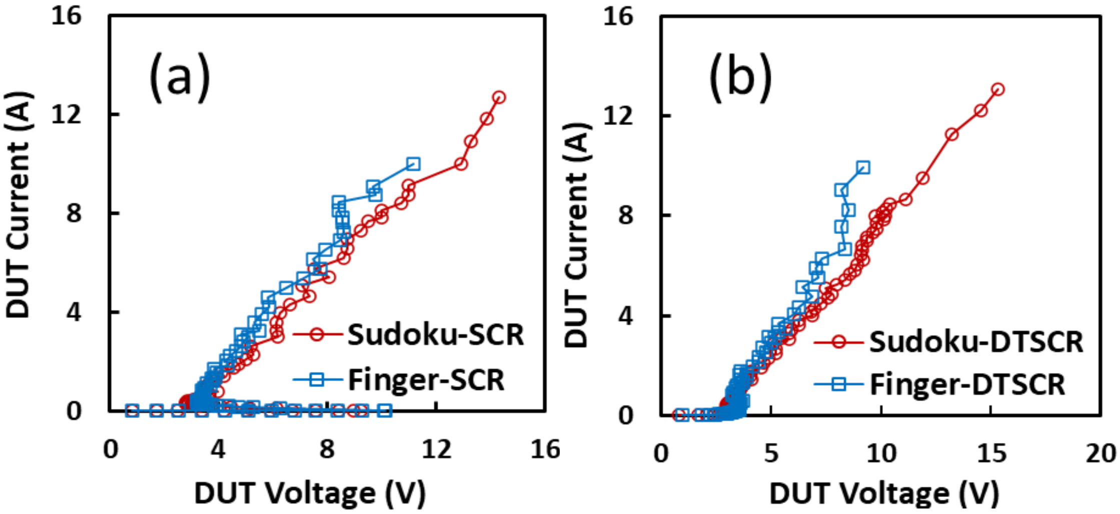


Figure 5.26 Comparison of VFTLP measured ESD discharging I-V curves for: (a) sudoku SCR ESD array and finger SCR device counterpart, (b) sudoku DTSCR ESD array and finger DTSCR device counterpart.

DTSCR ESD arrays increase the ESD area efficiency by ~55.5% over traditional finger-DTSCR devices, respectively, in addition to reducing ESD triggering voltage V_{t1}. Furthermore, we can find that the triggering voltage of the DTSCR is much lower than the SCR for both the sudoku and finger devices. This testing validates under the VFTLP the much-improved high efficiency of sudoku devices upon the finger devices no matter on SCR or DTSCR.

5.6 Sudoku ESD array design scalability

Clearly, sudoku SCR/DTSCR ESD arrays have advantages of higher ESD area efficiency, design scalability, and being layout-friendly. However, scalability design is rather involving and requires careful design trade-offs to maximize ESD area efficiency. How to design the sudoku array is a complexity. The sudoku array is generally area-efficient due to the maximum utilizing the across-boundary all-perimeter ESD conduction to exaggerate the ESD discharging path around each cell while the edge and corner boundary is wasted during the ESD discharging, which will eliminate the area-efficiency, especially for the small sudoku arrays. Nevertheless, smaller cells tend to improve ESD discharging uniformity in a large Sudoku array and may suffer from more corner/edge current/heat crowding at cell levels. Therefore, it comes to the problem: how to select the cell dimension including the line width of P+/N+ rings may affect the equivalent cell ESD

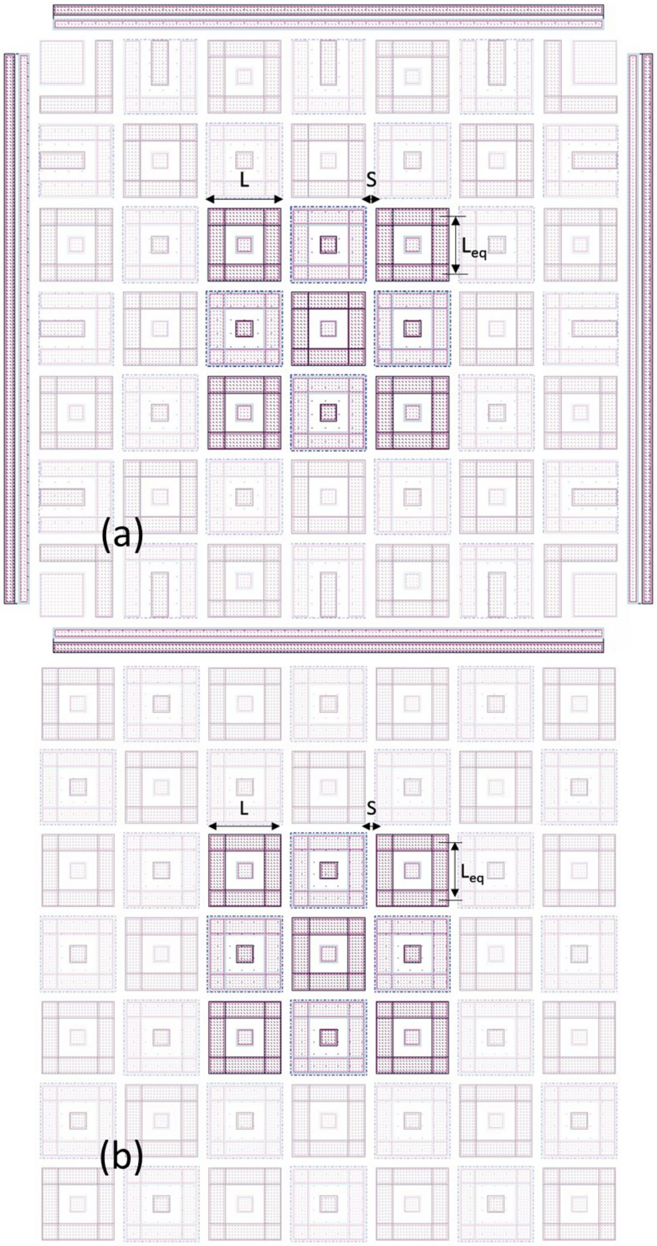


Figure 5.27 Illustration of *N*x*N* scalable array containing the 3x3 device fabricated in a 22nm FDSOI CMOS technology for (a) sudoku DTSCR ESD array structure, and (b) sudoku SCR ESD array structure.

conduction path width per side (L_{eq}). We hence propose a sudoku ESD array design strategy as depicted in the figure. 5.27a and figure 5.27b for *N*x*N* sudoku DTSCR and sudoku SCR ESD arrays, containing the (smaller) 3x3 Sudoku devices fabricated. L and S are cell dimension and cell-to-cell isolation, L_{eq} is estimated from the two ends bounded by the inner/outer edges of P+/N+ rings considering the cell corner effect, $W_{eq\text{-}total}$ is the total equivalent ESD discharging width for a Sudoku ESD array, and $J_{t2\text{-}N}$ is ESD failure current density (I_{t2} per area) of a sudoku ESD array of NxN (i.e., the sudoku DTSCR ESD area efficiency).

Below is the equation used to derivate the area-efficiency:

$$W_{eq-total} = \frac{4L_{eq}\,N^2-4L_{eq}N}{2} = 2N(N-1)L_{eq} \tag{1}$$

$I_{t2\text{-}N}$ is measured by TLP for an NxN Sudoku array, which is normalized to $W_{eq\text{-}total}$ as,

$$I_0 = \frac{I_{t2-N}}{W_{eq-total}} = \frac{I_{t2-N}}{2N(N-1)L_{eq}} \tag{2}$$

The sudoku ESD area efficiency is then derived as,

$$J_{t2-N} = \frac{I_{t2-N}}{A} = \frac{2N(N-1)L_{eq}}{(NL+NS-S)^2} I_0 \tag{3}$$

where the NxN sudoku ESD array layout size is,

$$A = (NL + NS - S)^2 \tag{4}$$

Eq (3) implies that the cell dimensions (L_{eq}) will directly affect ESD area efficiency, making designing sudoku DTSCR ESD arrays very subtle and involving. However, if choosing $N\rightarrow\infty$ for a very large sudoku DTSCR array, it gives,

$$J_{t2}|_{N\to\infty} = \frac{2L_{eq}}{(L+S)^2} I_0 = \frac{I_{t2-N}}{N(N-1)(L+S)^2} \tag{5}$$

indicating that the design uncertainty in the accuracy of selecting L_{eq}, i.e., cell dimension design, will no longer be a sensitive design factor. Figure 5.28 shows the normalized ESD area efficiency versus the array size for the NxN Sudoku-DTSCR ESD array structures. It is readily observed that a larger sudoku ESD array improves its ESD area efficiency as expected, but the improvement will saturate for a very large array. It is found that after the N is above the 10, there is no big difference improvement, and considering the two complexities of the big layout, it is a balance between area efficiency and the complexity of the layout. This gives quantitative design guidelines for optimizing ESD area efficiency of scalable Sudoku-DTSCR ESD arrays for LV ICs, removing the subtle design trade-off complexity including cell size, cell layout, and array size.

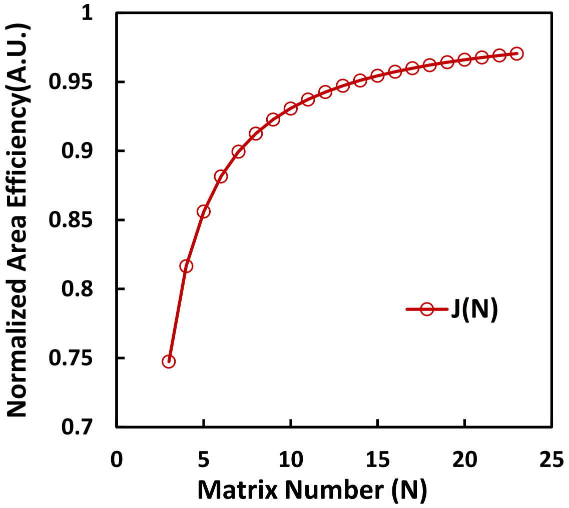


Figure 5.28 Normalized ESD area efficiency versus the array size (N) for scalable NxN Sudoku-DTSCR ESD array structures in a 22nm FDSOI CMOS.

Chapter 6 Graphene MEMS ESD switch design and evaluation

6.1 Single crystalline gNEMS ESD switch structure

6.1.1 Overview of the gNEMS ESD switch function

As CMOS technology continuously advances into nanometer nodes, ESD protection design becomes much more challenging due to the inherent ESD-oriented design overhead. Nevertheless, since the never-zero ESD-induced overhead forever exists in the traditional in-Si PN-junction based ESD protection structures, it rapidly becomes unacceptable to large, sophisticated, high-performance chips designed in near-nm technologies due to its inherent induced parasitic capacitance (C_{ESD}), leakage current (I_{leak}) and noises, and ESD-consumed Si areas. It hence calls for revolutionary ESD protection solutions for future ICs. Graphene has excellent electrical properties like high electrical and thermal conductivity and good mechanical properties like extremely flexible and lightweight. Graphene gNEMS ESD switch is a totally mechanical switch structure above-Si and we improve its performance a lot via the better graphene material like single crystalline graphene. compared with previously used poly crystalline graphene.

Figure 6.1 depicts the new single crystalline graphene gNEMS switch structures for on-chip ESD protection. The gNEMS switch is a two-terminal ESD device used for on-chip ESD protection which can be integrated into the CMOS BEOL process to save the silicon area. It features a single-crystal graphene membrane over a cavity and the suspended graphene film serves as one electrode and the heavily doped p-type Si bottom is the other electrode. Previously, we fabricated prototype gNEMS devices using poly-crystalline graphene that demonstrated basic ESD switching function by TLP testing, confirming dual-

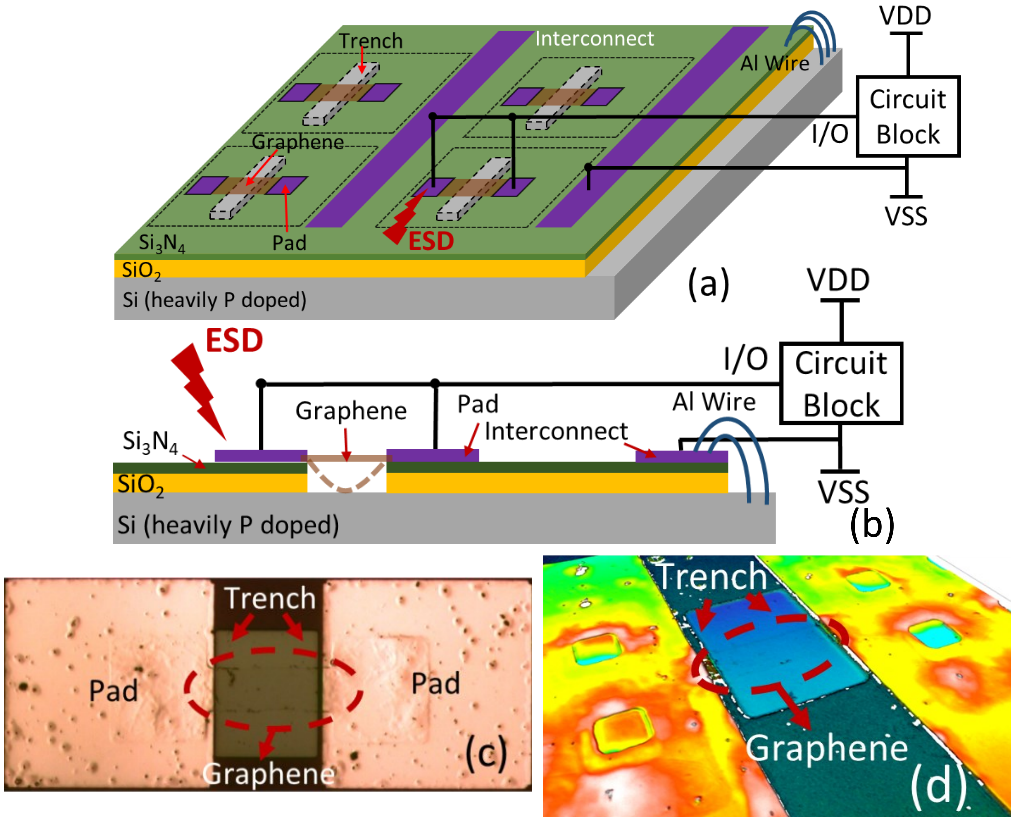


Figure 6.1 Illustration of above-IC graphene gNEMS switch for on-chip ESD protection (a) a 3D top view, and (b) a cross-section view, (c) a 2D top view of optical microscope, and (d) a 3D confocal view showing the vertical trench.

polarity ESD discharging I-V behaviors. In this work, new gNEMS ESD switch devices were designed and fabricated using single-crystalline graphene monolayer films aiming to dramatically improve the gNEMS ESD robustness [41]. Figure 6.1a is a 3D view of the single crystalline gNEMS ESD switch with the suspended graphene. over the dielectric layer composed of SiO_2 and Si_3N_4. We also use PVD to deposit the metal as an interconnection to test the performance under the TLP and VFTLP stressing. Figure 6.1b is the cross-section of a single crystalline gNEMS ESD switch showing the suspended graphene ribbon and the cavity trench. Figure 6.1c is the 2D top view from the optical microscope. We can see the trench and the suspended graphene ribbon between the two metal pads. Figure 6.1d is the 3D confocal view showing the very shallow depth of the trench. The graphene MEMS ESD switch device functions as follows: The two terminals are connected to an I/O pad (anode of graphene membrane) and a GND pad (cathode of P-Si) in an IC, respectively. During an ESD event, when an ESD pulse appears at the I/O pad, the strong transient electrostatic field induced will pull down the suspended graphene ribbon, which eventually touches the P-Si in the cavity, hence, turns on the MEMS switch, forms a conducting path to discharge the ESD pulse and protects the IC.

We envision that single-crystal graphene is one of the main factors to enhance the ESD capability of gNEMS devices because single-crystal graphene has many advantageous properties over poly-crystal graphene, such as better conductivity, fewer defects, and higher mechanical strength, due to no poly-grain boundaries. Figure. 6.2 shows the Raman spectrum comparison for both single crystalline and poly crystalline graphene films grown in this work where the graphene characteristic peaks (D, G, and 2D) are clearly observed. It is known that the relative G and 2D peak intensity can confirm the monolayer's property

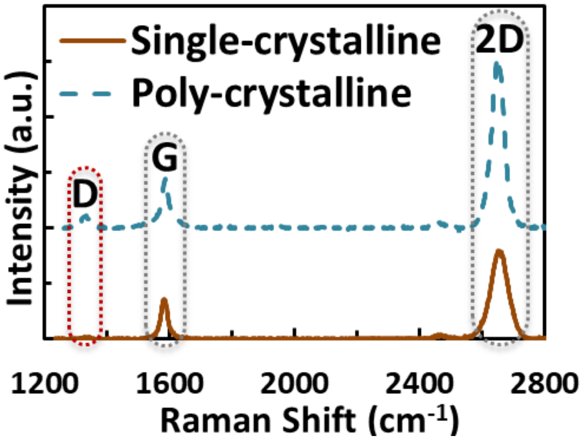


Figure 6.2 Raman spectrum for poly-crystalline and single-crystalline graphene monolayer films grown in this work confirms monolayer graphene.

belonging to grown graphene films. The D peak, reflecting graphene crystal structure, is higher for poly-crystal graphene due to poly grain boundaries [42].

6.1.2 The fabrication process of the gNEMS ESD switch

Single-crystalline graphene films were grown using the chemical vapor deposition (CVD) method and fine-polished copper foil stated in the previous chapter. Briefly, it starts with a heavily P-doped Si wafer. A SiO_2 layer of about 250nm was deposited on the Si substrate by PECVD at 400°C, followed by a Si_3N_4 film of about 100nm on SiO_2 by the same

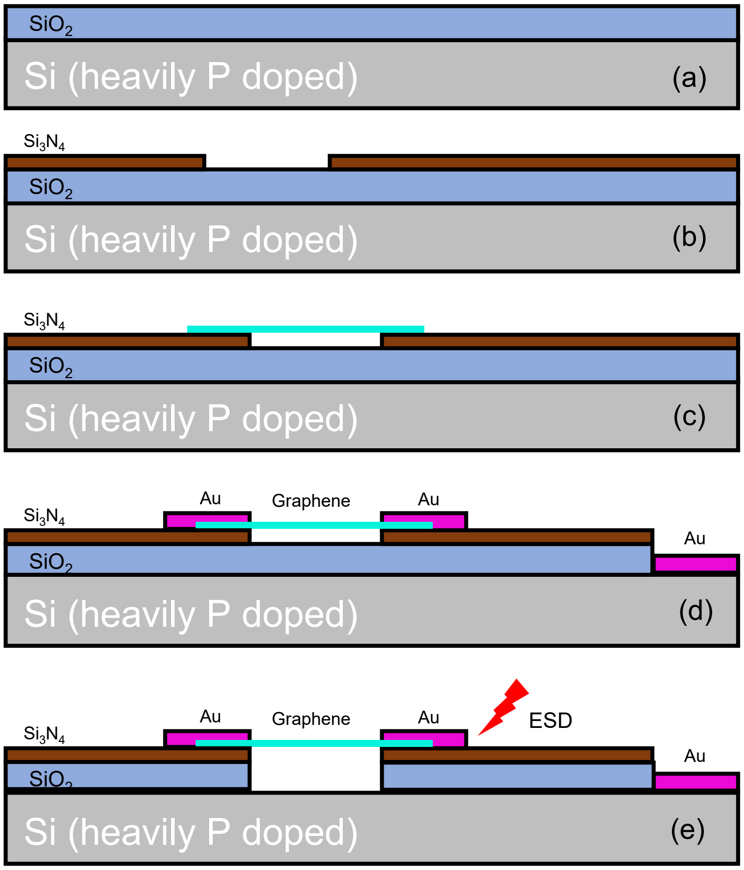


Figure 6.3 A CMOS-compatible process flow for making gNEMS devices: (a) SiO_2 deposition on P-Si, (b) opening in Si_3N_4 above SiO_2, (c) graphene transfer onto Si_3N_4, (d) forming PVD Au pads, and (e) creating an air cavity in SiO_2 masked by Si_3N_4 using HF etching lift-off, resulting in a gNEMS switch.

PECVD process. A cavity window pattern is defined by lithography and the cavity was formed by RIE etching into the Si_3N_4 layer, which is used as the hard mask for the final HF etching. The graphene films used were grown on a copper substrate [43] [44]. Next, a graphene film was transferred onto the Si substrate as follows: First, a polymethyl methacrylate (PMMA) was spin-coated onto the graphene membrane on a copper substrate. Second, the graphene on copper was placed into a configured iron chloride solution to dissolve the copper. This procedure was repeated to ensure that the copper substrate was removed completely, which was followed by cleaning any residual copper ion in deionized water. Next, the graphene/PMMA film was transferred onto the silicon containing the cavity structure. The PMMA layer was then dissolved in acetone, resulting in the graphene membrane on the patterned silicon. Lithography was used to pattern the graphene membrane into graphene ribbons covering the Si_3N_4 windows created, which was followed by using RIE oxygen to form the patterned graphene ribbons. Next, an e-beam was used to deposit 90nm, Au, at 10nm Pa and the Au pads were formed by lift-off. Finally, with the Si_3N_4 as the blocking mask, hydrogen fluoride (HF) vapor was applied to etch into the patterned SiO_2 to release the graphene ribbon instead of using the buffered HF due to the large liquid tension damaging the suspended graphene membrane. The above procedures are depicted in figure 6.3.

Besides the TLP and DC pads, the ground-signal (GS) pads with a 150μm pitch were designed to allow the VFTLP testing. In the following testing, a large number of new single-crystal gNEMS devices of various dimensions were designed and fabricated for a systematic study in this work. The gNEMS switch has different dimensions from the various length ((L = 3/5/7/10/15/20µm) to the various width (W = 3/5/7/10/15µm) at the fixed 350nm depth. Figure 6.4 presents the images of the fabricated gNEMS ESD switch samples showing the graphene ribbon, the cavity, and two gold pads. Figure 6.4a shows the optical image after the graphene membrane patterned by oxygen plasm and figure 6.4b shows the optical image after the metal pad deposition and lifting. The SEM image after HF vapor etching is taken to show the gNEMS switch structure in figure 6.4c.

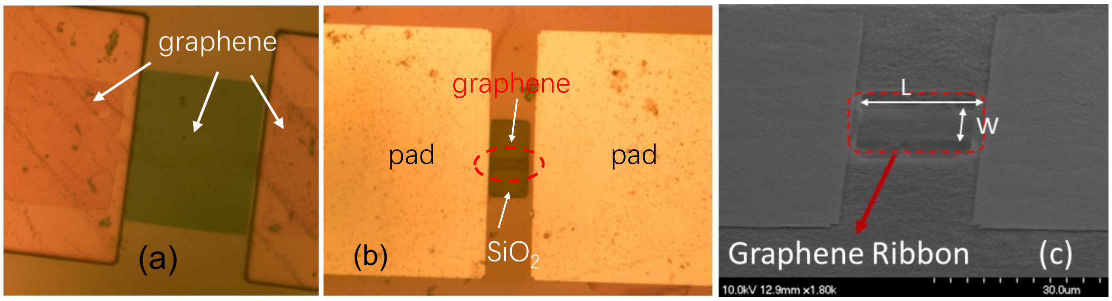


Figure 6.4 Optical top views of graphene ESD devices fabricated: (a) post oxygen RIE showing the patterned graphene ribbon, (b) after e-beam pad deposition, and (c) SEM image shows single-crystal graphene gNEMS ESD switch fabricated. With dimension defined as L = length and W = width.

6.2 gNEMS ESD switch behaviors on poly and single crystalline graphene

6.2.1 Temperature influence on poly crystalline gNEMS

Figure 6.5 shows that the graphene MEMS ESD device can respond to the fast TLP pulse and provide the ESD discharge function over a wide operating temperature range from 10°C to -10°C [45]. We can observe that the atmosphere temperatures can largely affect the triggering voltage V_{t1} and breakdown current I_{t2}. It is readily found that at 110°C, the measured ESD current-handling capability, i.e., thermal breakdown current, is I_{t2}~0.99mA, and the ESD triggering voltage is V_{t1}~10.6V. At room temperature of 30°C, I_{t2} ~2.53mA and V_{t1}~8.84V are obtained. While at the -10°C, the performance can be enlarged with observed I_{t2} ~3.85mA and V_{t1} ~5.03V. Clearly, the graphene MEMS ESD switch can handle much higher ESD pulses at low temperature, mainly due to the better heat dissipation and higher electron mobility of the graphene ribbon, which also causes lower ESD triggering voltage V_{t1}.

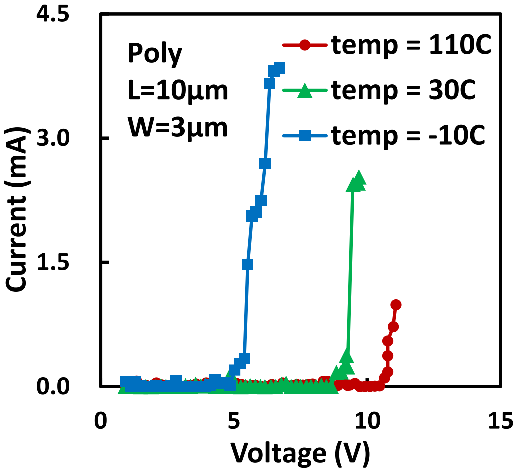


Figure 6.5 Temperature effect of a sample poly-crystal gNEMS device by TLP testing.

6.2.2 Performance comparison between poly and single crystalline graphene

Though both poly crystalline and single crystalline graphene are both monolayer graphene, mostly, the single crystalline graphene has better electrical, thermal and mechanical properties such as high electrical conductivity, good heat dissipation, and fairly flexible strength, which in total makes it a better candidate for gNEMS ESD switch. The reason is that in general, poly crystalline graphene has many grain boundaries which will not only affect conduction properties but also easily induce cracks in graphene films under the fast and strong ESD zapping. Figure. 6.6 compares the DC and TLP testing results for both single-crystal and poly-crystal gNEMS devices of varying dimensions [46] [47]. It is readily observed that single-crystal gNEMS devices outperform poly-crystal gNEMS switches dramatically in both DC sweeping and TLP zapping tests, e.g., for DC,

I_{t2}~0.37mA for single-crystal gNEMS over I_{t2}~0.14mA for poly-crystal gNEMS (W/L = 5μm/7μm); for TLP, I_{t2}~31.1mA for single-crystal gNEMS over I_{t2}~5.88mA for poly-crystal gNEMS (W/L = 5μm/7μm). Not only these two specific samples, but the dramatic improvement in ESD discharging current handling capability was also broadly observed via statistical measurement.

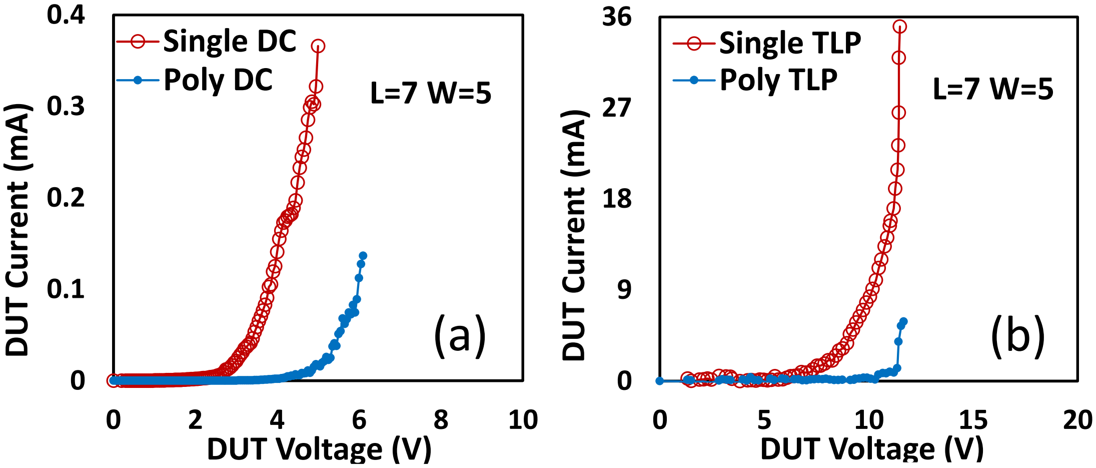


Figure 6.6 Comparison between single-crystal and poly-crystal graphene gNEMS switches: (a) DC sweeping I-V curves, and (b) transient ESD discharging I-V curves by TLP zapping

6.2.3 Comprehensive ESD measurement on single crystalline gNEMS

Fully ESD testing including the DC, TLP, and VFTLP on the DUT in order to evaluate the function and performance under the different working situations [48]. Figure 6.7a shows the DC testing using the Agilent Precision Semiconductor Parameter Analyzer with the sample dimension of length 5μm and width 3μm. The DC switching function is clearly observed with a turn-on voltage of ~2.45V and a thermal breakdown current of ~0.30mA. We use the Barth 4002 instrument with pulse width 100ns and rise time 10ns to evaluate the HBM ESD protection performance with a sample of length 5μm and width 3μm, which behaves very well. In figure 6.7b, we can read that the turn-on voltage is 7.79V and the thermal breakdown current is 30.3mA on the condition of low leakage current 2pA. Figure 6.7c presents the transient ESD discharging I-V curve for a sample gNEMS device with a length of 3μm and width of 3μm measured by ultrafast VFTLP. The Barth 4012 instrument with pulse width 1ns and rise time 100ps to evaluate the CDM ESD protection performance, which shows the designed ESD discharging function with V_{t1}~4.2V and I_{t2}~31.3mA. Overall, the new single crystalline gNEMS devices were validated for ESD protection not

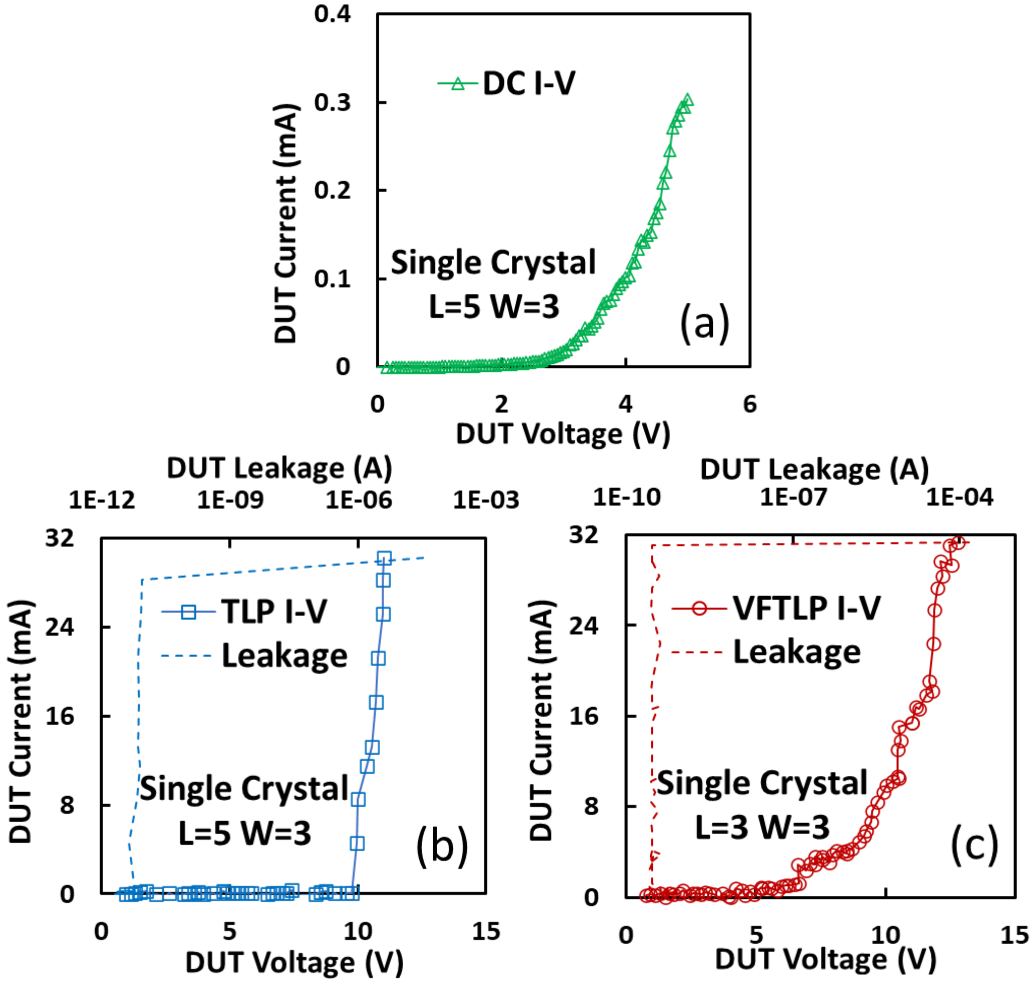


Figure 6.7 Measured switching I-V behaviors for different gNEMS devices of varying sizes under different testing methods: (a) DC sweeping, (b) TLP stressing for HBM ESD, (c) VFTLP stressing for ultrafast CDM ESD.

only for HBM ESD models but also, and for the first time, for ultrafast CDM ESD model, confirming its broader ESD protection functions as expected.

6.3 gNEMS switch reliability evaluation on ESD performance

The most important criterion is repeatability which is the main parameter for the material product. Via the single crystalline graphene utilizing, we expect to exaggerate the reliability of the gNEMS switch by DC, TLP, and VFTLP testing. Figure 6.8a depicts the DC sweeping I-V curves for a gNEMS device with length 20μm and width 15μm under an 11-times repeating DC stressing routine where the DC sweeping voltage was clamped to

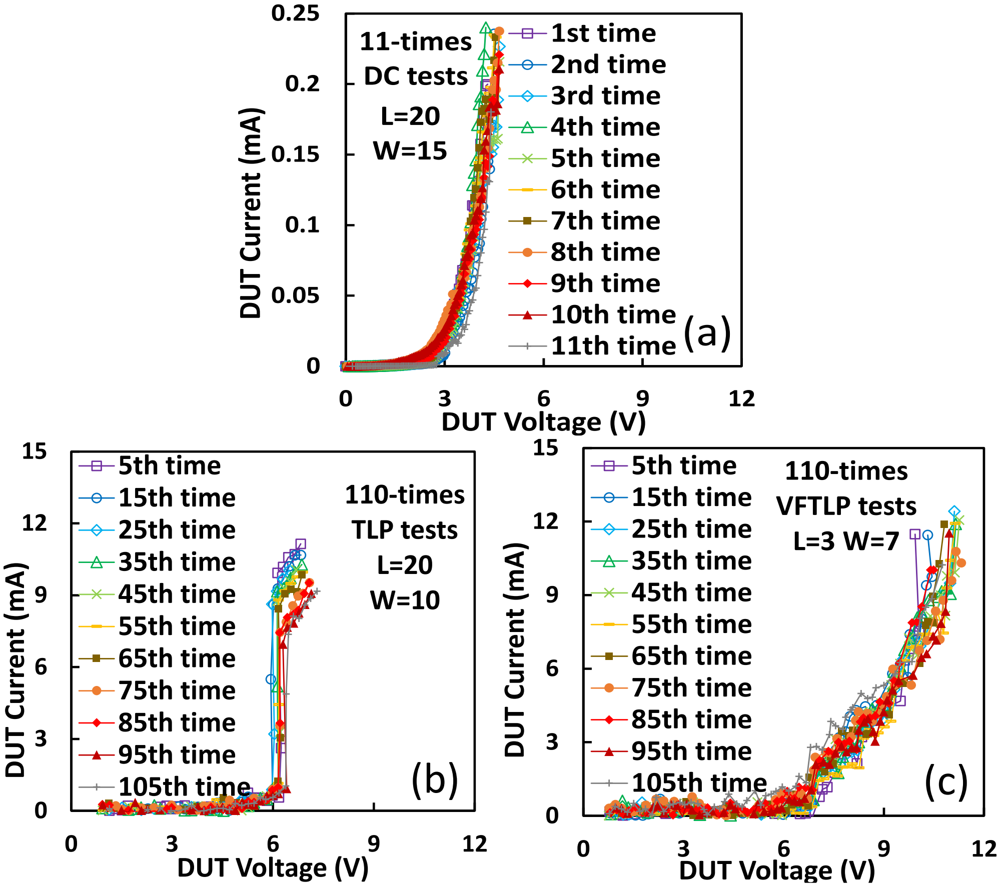


Figure 6.8 Repeating stressing tests for different single-crystal gNEMS devices of varying sizes confirm that the gNEMS structures are extremely stable: (a) 11-times repeating DC sweeping stress test, (b) 110-times repeating TLP zapping, and (c) 110-times repeating VFTLP zapping test.

below the thermal breakdown current threshold (I~0.24mA) to avoid device failure so that the same gNEMS device can be stressed repeatedly with triggering voltage about 2.65V to evaluate possible aging effects. The testing clearly shows that the gNEMS device is very stable under repeatable DC stresses. Figure. 6.8b presents transient ESD discharging I-V behaviors for a gNEMS switch with a sample of length 20μm and width 15μm under a vast 110-times repeating TLP zapping procedure where the TLP pulse train was set to below the thermal breakdown threshold to avoid ESD thermal breakdown to the same device with the current compliance of 10mA. For presentation clarity, I-V curves from every 10 sequential repeating TLP tests are grouped together and the I-V curve of the middle test run (i.e., 5th for the group of 1st to 10th tests) is selected to be shown in the chart in Figure 6.8b, representing 10 sequential repeatings TLP zaps per I-V curve. It is readily observed that the gNEMS switch is very stable after 110 times repeating TLP zapping tests, and the V_{t1} stays at 6.2V. Similarly, for the VFTLP testing, a sample of length 3μm and width 7μm is picked to run the 110-times VFTLP repeatable stressing shown in figure 6.8c. It is readily observed that the gNEMS works well on the triggering voltage of 6.1V with the current compliance of 10mA in case of thermal breakdown for the same device uniform testing to evaluate the CDM performance. Overall, the comprehensive and vast repeating stressing tests clearly show that the new single crystalline gNEMS switch is extremely reliable and functions uniformly in ESD protection, mainly due to the quality of the superior materials of single-crystalline graphene films fabricated.

6.4 The influence of ESD pulsing condition on gNEMS

6.4.1 Rise time condition under TLP and VFTLP measurement

Since the rise time of TLP pulse can be adjusted among the 0.2ns, 2ns, and 10ns as well as the VFTLP among the 0.1ns, 0.2ns, and 0.4ns, we can study the influence of rising time on the behavior of the single crystalline gNEMS ESD switch. Figure 6.9a and figure 6.9c show the observed pulse rise time effect on gNEMS devices under a 9-times repeating TLP and VFTLP testing routine. We pick two samples of length 10μm and width 10μm under

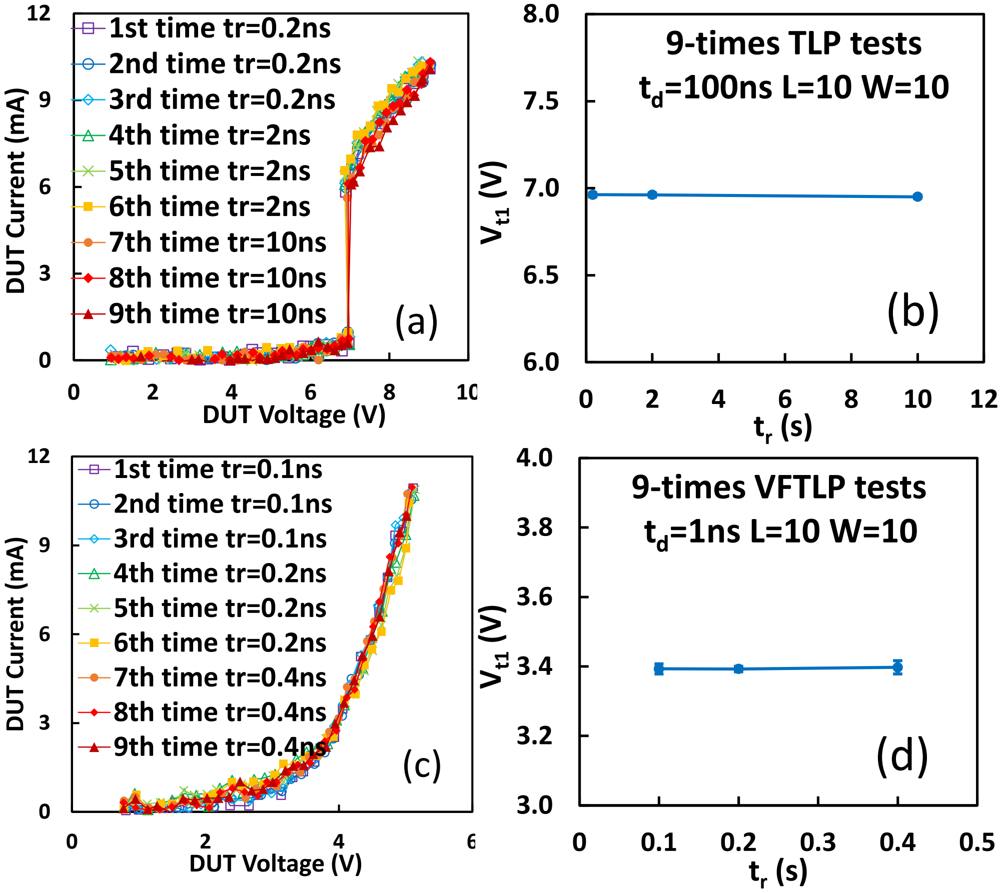


Figure 6.9 Evaluation of possible t_r effect on different single-crystal gNEMS devices by 9-times repeating TLP and VFTLP ESD zapping tests: (a) 9-times repeating TLP zapping with fixed t_d~100ns and varying t_r~0.2/2/10ns, (b) TLP V_{t1}~t_r statistics, (c) 9-times repeating VFTLP zapping with fixed t_d~1ns and varying t_r~100/200/400ps, and (d) TLP V_{t1}~t_r statistics. Tight error bars in (b/d) indicate testing stability.

9 repeating stresses (3 times for each given t_r), respectively, which have the fixed pulse width of 100ns for TLP and 1ns for VFTLP. It is readily observed for the same device limited to the current of 10mA before the thermal breakdown to keep the uniformity, the I-V curves have the same on-resistance R_{on} and triggering voltage V_{t1}, which clearly shows that no matter on the TLP or VFTLP, single crystalline gNEMS device is very stable and varying t_r has almost no noticeable impact on the ESD V_{t1}. Figure 6.9b and figure 6.9d

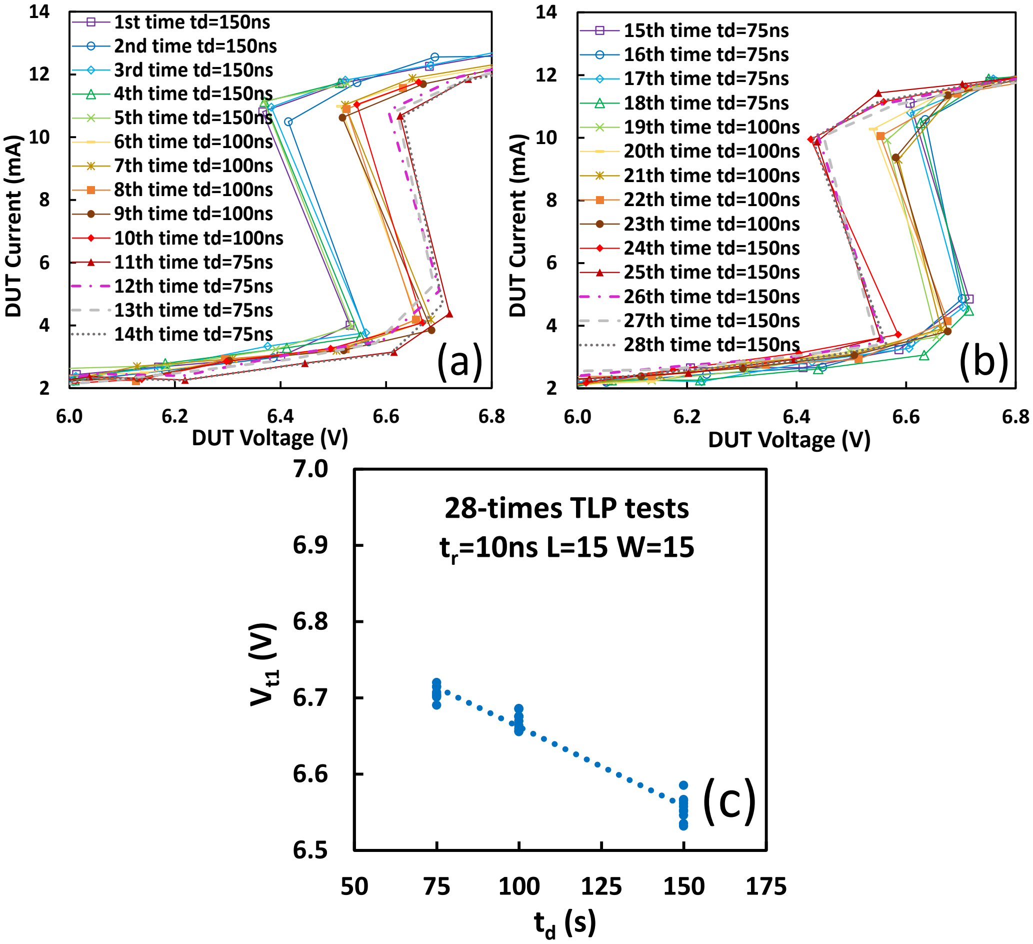


Figure 6.10 Evaluation of possible t_d effect on different single-crystal gNEMS devices by repeating TLP ESD zapping test routines: (a) 28-times repeating TLP zapping with fixed t_r~10ns and decreasing t_d~150/100/75ns routine, (b) 28-times repeating TLP zapping with fixed t_r~10ns and increasing t_d~75/100/150ns routine, (c) TLP V_{t1}~t_d statistics.

show the details of the set of curves on the tiny error bar, a straight line, which can be just thought of as the testing errors on both TLP and VFTLP.

6.4.2 Pulse duration condition under TLP and VFTLP measurement

In order to test the influence of different pulse duration, we pick two samples of length 15μm and width 15μm for TLP stressing as well as length 20μm and width 10μm for VFTLP stressing. Figure. 6.10a and 6.10b depict TLP zapping with a fixed t_r=10ns and varying pulse duration (t_d=75ns/100ns/150ns in both decreasing and increasing zapping sequences, respectively) for 28 repeating stresses (4-5 times for each given t_d) on the same

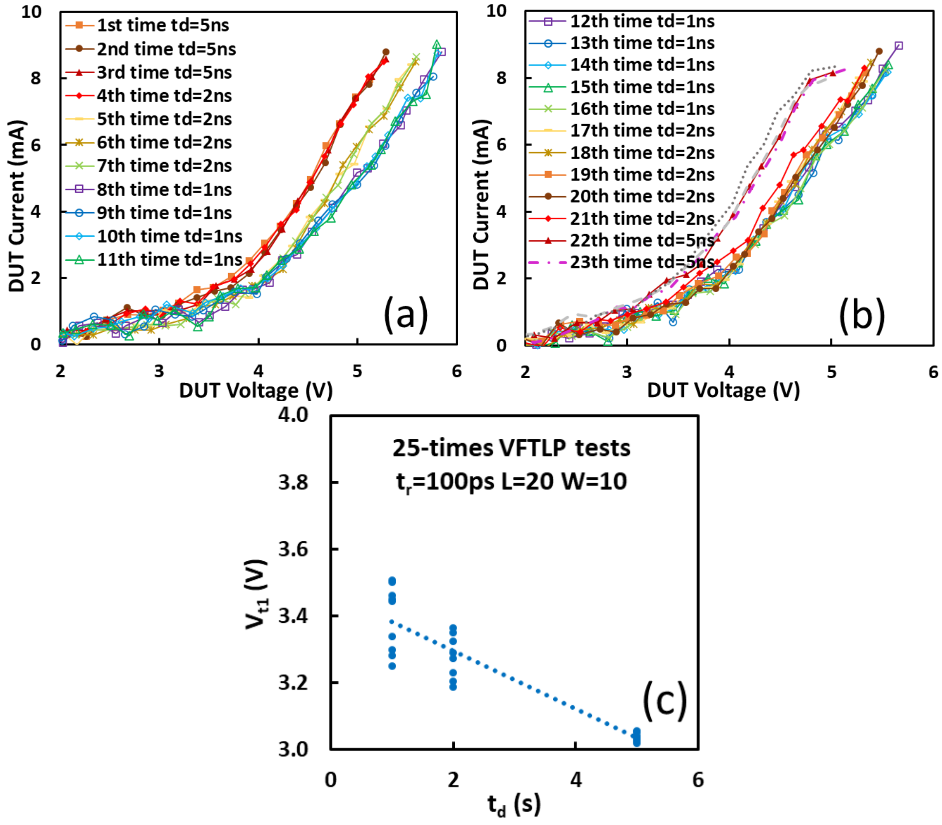


Figure 6.11 Evaluation of possible t_d effect on different single-crystal gNEMS devices by repeating VFTLP ESD zapping test routines: (a) 25-times repeating VFTLP zapping with fixed t_r~100ps and decreasing t_d~5/2/1ns routine, (b) 25-times repeating VFTLP zapping with fixed t_r~100ps and increasing t_d~1/2/5ns routine, and (c) VFTLP V_{t1}~t_d statistics.

sample. It is readily observed that the I-V curves have different triggering voltages and similar on-resistance showing the fairly stable gNEMS device. Figure 6.10c shows the statistical V_{t1}~t_d relationship for the single-crystal gNEMS under the 28-times repeating TLP zapping routine. It is readily observed that with the pulse duration increasing, the triggering voltage decrease because the longer pulse duration will have larger power from the square TLP pulse, which will make the gNEMS switch easier to be turned on.

In terms of the VFTLP, figure 6.11a and figure 6.11b display the statistical V_{t1}~t_d relationship for the single-crystal gNEMS under the 28-times repeating TLP zapping routine. Similarly, a fixed t_r=100ps and varying pulse duration (t_d=100ps/200ps/400ps in both descending and ascending stressing manners, respectively) for 25 repeating stresses (4-5 times for each given t_d) were issued. It is readily observed that the I-V curves have different triggering voltages and similar on-resistance showing the fairly stable gNEMS device. Figure 6.11c shows the statistical V_{t1}~t_d relationship for the single-crystal gNEMS under the 25-times repeating VFTLP stressing routine. The slight V_{t1} decrease with longer td may be attributed to the fact that a longer pulse induces a stronger electrostatic field that excises a bigger pull-in force to the graphene ribbon, hence reducing ESD V_{t1}.

6.5 The influence of graphene ribbon dimension on gNEMS

6.5.1 gNEMS dimension effect on triggering voltage

The influences on device dimensions on ESD operations of single-crystal gNEMS switches were studied statistically in this work by applying TLP zapping pulses to different gNEMS devices of various dimensions. Figure 6.12a depicts the statistics for the measured V_{t1}~W relationship for different gNEMS devices of varying graphene width (W =

5/7/10/15μm) at a fixed length of L=10μm from TLP testing with t_r=10ns and t_d=100ns. It is readily observed that the ESD triggering voltage (V_{t1}~6.9V) is insensitive to the width of the graphene ribbon. This may be explained as follows: the graphene stiffness, resisting the pulling force, is closely related to the film dimensions (~W/L^3) [49], hence increase of W makes it harder to pull in the graphene ribbon; meanwhile, a wider W means larger graphene size that induces more electrostatic pulling force. These two opposite effects may cancel each other against W, hence making V_{t1} insensitive to W at a fixed L. Figure 6.12b presents the statistics for measured V_{t1}~L relationship for different gNEMS devices of varying graphene length (L = 5/7/10/15/20μm) at a fixed width of W=7μm under the same TLP condition. It clearly shows a monotonous V_{t1}~L relationship at a fixed W. This can be attributed to two factors: As L increases, the graphene stiffness decreases dramatically, leading to a much weaker elastic force resisting the pulling force; while the increase in graphene size due to longer L produces a stronger electrostatic pulling force. Together, ESD V_{t1} decreases as L increases substantially under TLP stressing.

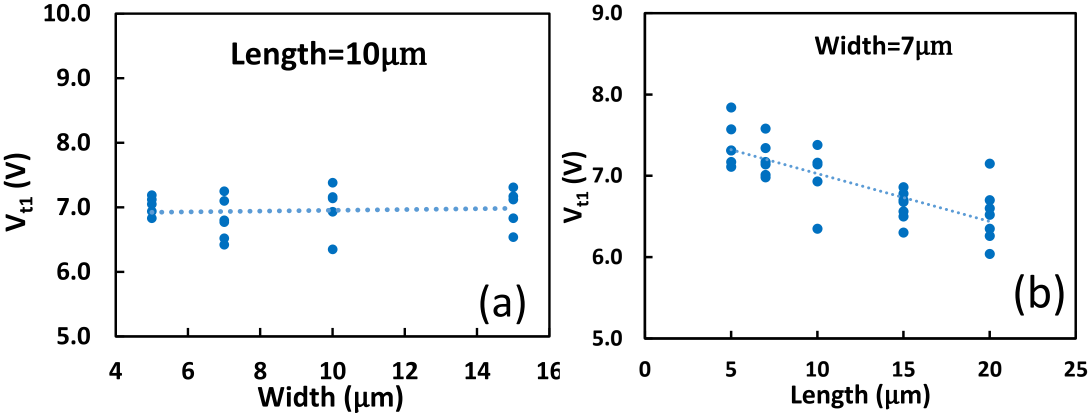


Figure 6.12 Statistics for possible graphene dimension impacts on ESD V_{t1} under TLP stressing: (a) V_{t1}~W for a fixed L is almost flat, and (b) V_{t1}~L for a given W showing a monotonous trend.

In terms of VFTLP stressing, a similar study was done in figure 6.13. It is depicted that among different graphene widths (W = 3/5/7/10μm) at a fixed length of L=3μm under VFTLP testing (t_r=100ps and t_d=1ns) the statistics for measured V_{t1}~W relationship for different gNEMS devices were measured. Similarly, it is readily observed in figure 6.13a that the measured ESD V_{t1}~4.65V is stable, insensitive to W of graphene ribbons, which is again attributed to the balance between increases in both elastic force and electrostatic pulling force, in opposite directions. On the other hand, Figure 6.13b depicts the statistics for measured V_{t1}~L relationship for different gNEMS devices of various graphene length (L = 3/5/10/15/20μm) at a fixed width of W=10μm under the same VFTLP testing condition, which clearly shows a monotonous V_{t1}~L relationship at a fixed W. Again, this is due to the same reason as that of TLP. Hence, ESD V_{t1} decreases as L increases under VFTLP zapping as shown in Figure 6.13b. Overall, V_{t1} can be adjusted by design variation of gNEMS dimensions (e.g., L, W, and cavity depth) and graphene films (e.g., graphene film quality and layer numbers) in practical device designs to meet IC specifications. In this work, varying V_{t1} of 6.04V~7.79V by TLP and 3.02V~6.1V by VFTLP were obtained for the prototype gNEMS devices.

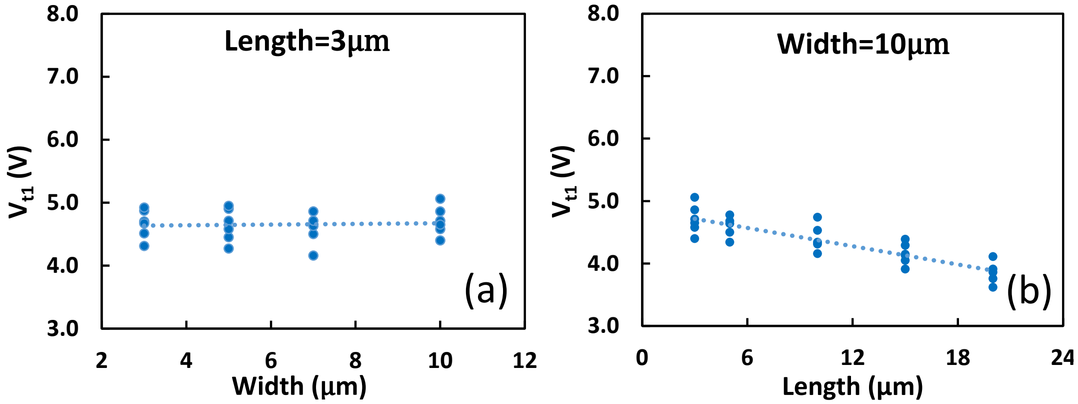


Figure 6.13 Statistics for possible graphene dimension impacts on ESD V_{t1} under VFTLP stressing: (a) V_{t1}~W for a fixed L is almost flat, and (b) V_{t1}~L for a given W shows a monotonous trend.

6.5.2 gNEMS dimension effect on current handleability

In order to evaluate the current handling capability, we take a statistical measurement on lots of gNEMS samples with various sizes of length = 3/5/7/10/15/20μm and width = 3/5/7/10/15μm, which are characterized under the TLP with t_r=10ns and t_d=100ns and VFTLP with t_r=100ps and t_d=1ns. In figure 6.14a and figure 6.14b, the measured ESD

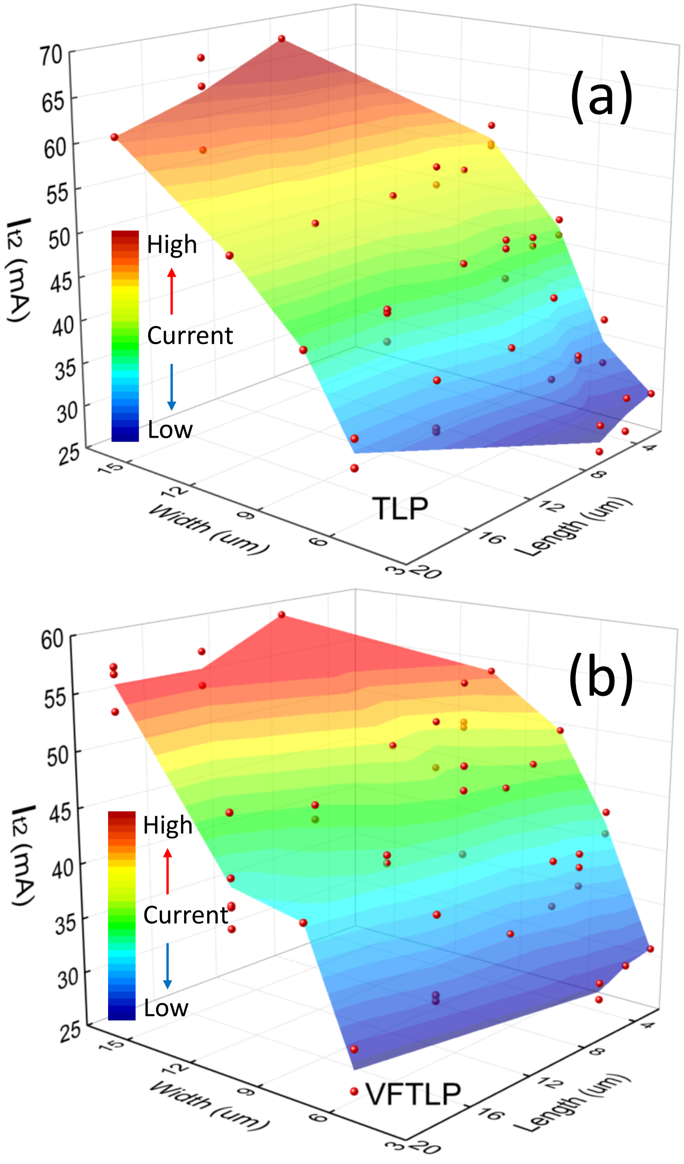


Figure 6.14 Statistics for measured ESD I_{t2} capability of different gNEMS of varying dimensions under TLP and VFTLP stressing: (a) by TLP zapping, and (b) under VFTLP stressing.

thermal breakdown current is 25.5mA~69mA under TLP zapping and 27.6mA~59.9mA under VFTLP stressing, respectively. It is readily observed that as width increases at a fixed length, I_{t2} increases substantially, indicating increased ESD current handling capability for a wider width which is mainly because a wider graphene ribbon can certainly conduct more ESD current. On the other hand, at a given width, I_{t2} decreases slightly for longer length, though seems to be not sensitive to length. This might be attributed to the fact that a longer graphene ribbon means a larger graphene size that may have more defects. It is worth noting in the figure 6.14 shows a statistical scaling trend for I_{t2}, a firm quantitative I_{t2} scaling factor could be drawn from the prototype devices yet.

In the meantime, we can make the statistical measurements of leakage current for varying dimensions under TLP stressing with t_r=10ns and t_d=100ns. It is depicted in figure 6.15 that the measured leakage currents for different gNEMS devices are extremely low, about a few pA. It is believed that the leakage may be associated with the contacts and interconnects of the prototype devices, not directly through the cavity gap, hence, not scaled to the gNEMS graphene film sizes. Such low leakage current can validation the

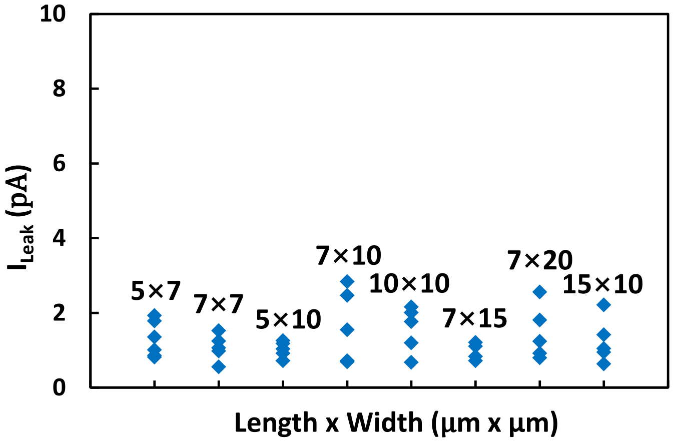


Figure 6.15 Statistics show that the measured leakage currents I_{leak} for different gNEMS devices of varying dimensions under TLP stressing are extremely low.

advantage of the mechanical ESD switch compared with the PN junction based ESD devices.

6.6 Quality control on the single crystalline gNEMS

6.6.1 Single crystalline gNEMS being robust

In order to demonstrate the excellent ESD robustness of the single crystalline gNEMS devices, Figure 6.16 depicts the measured transient ESD I-V curves for the specific sample gNEMS switches under both TLP and VFTLP tests to their upper limits of ESD I_{t2}. Figure 6.16a shows the I-V curve under TLP ESD based on the sample gNEMS device of length 20μm and width 7μm, which achieves an extremely high ESD I_{t2} of ~293mA until the thermal breakdown, which means an ESD current handling capability of J_{t2} ~ 1.19×10^{10} A/cm^2. In a nutshell, this is equivalent to an HBM ESD capability of ~178kV/μm^2 for the measured single-crystal gNEMS device, which is much improved from the reported J_{t2}~1.5kV/μm^2 for the poly-crystal gNEMS switch; further, it is orders of magnitudes more ESD robust than ~7.5V/μm^2 for a typical Si PN-based SCR ESD device. Figure. 6.16b depicts the I-V curve under VFTLP ESD based on the sample gNEMS device length 15μm and width 7μm, which achieves an extremely high ESD I_{t2} of ~149mA, meaning an ESD current handling capability of J_{t2} ~ 6.09×10^9 A/cm^2. It is noteworthy that, the exploratory gNEMS device was validated, characterized the improvement a lot by using single-crystal graphene. Furthermore, to our best knowledge, these test results set the records for ESD protection capability of any ESD structures reported as of today, which suggests that the novel single-crystal gNEMS switches are very promising future ESD protection solutions.

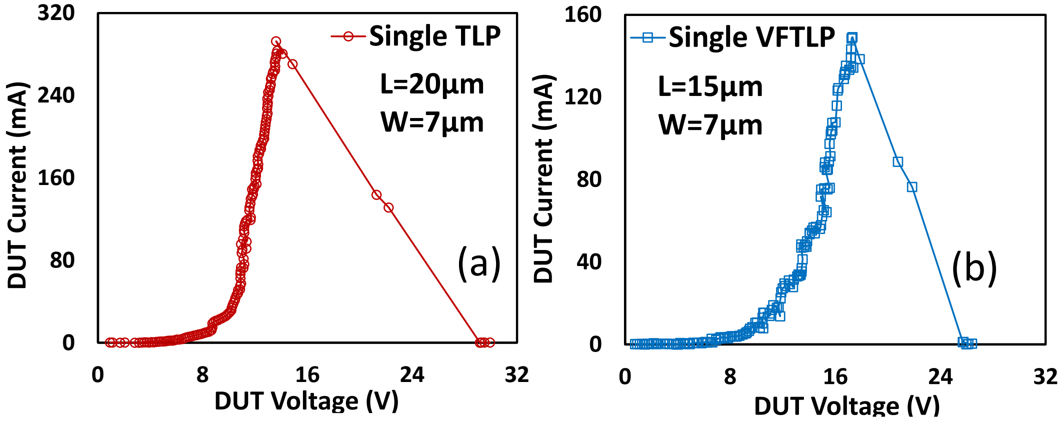


Figure 6.16 Measurements of sample single-crystal gNEMS devices set the record for ESD protection robustness, i.e., I_{t2}: (a) TLP zapping, and (b) VFTLP stressing.

6.6.2 Single crystalline gNEMS failure analysis

Since any of gNEMS ESD devices will eventually fail under high ESD stressing beyond its ESD protection capability. Figure 6.17 is used to study the possible ESD failures in single-crystal gNEMS ESD switches. This depicts the SEM image for a sample gNEMS device observed and shows two possible ESD failure signatures: a crack across the suspended graphene ribbon in the lower red dashed circle and a hole generated in the graphene film in the upper red dashed circle. The crack will totally impact the mechanical property of gNEMS when being pulled down and bouncing off back, which will influence the triggering voltage. Meanwhile, the hole will induce the hotspot during the ESD discharging current that will quickly accumulate the heat eventually resulting in the thermal breakdown, which decreases the gNEMS current handling capability. Though single crystalline graphene has normally better quality than poly crystalline graphene, there still exist some defects during the growth and device fabrication phases, which need to be further optimized in future work.

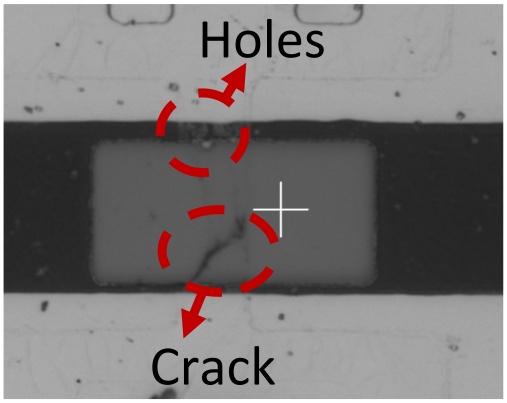


Figure 6.17 Two ESD failure signatures were observed for single-crystal gNEMS device: a hole in the graphene (Upper) and a crack across the graphene ribbon (Lower).

6.7 Transient gNEMS behaviors via TLP testing

During the ESD zapping, the suspended graphene ribbon will experience two forces: one is the ESD-induced electrostatic force that pulls down the graphene membrane and the other is the elastic force that restores the bent shape of the graphene ribbon, which the two opposite forces balancing with each other to determine the on and off states of a gNEMS switch device. The standard TLP ESD testing applies a string of well-defined ESD square waveforms to stress the gNEMS samples, and the instantaneous DUT I-V characteristics are obtained for the gNEMS ESD switch devices. During TLP stressing test, pulse trains of square waveforms are generated with the voltage step from a low level to higher levels with a voltage increasing, and transient voltage and current waveforms and instantaneous I-V curves are monitored to understand the gNEMS ESD discharging behaviors.

Figure 6.18a shows a well-behaved gNEMS ESD discharging I-V curve, featuring V_{t1} about 6V and negligible leakage current of a few of pA levels. Corresponding to figure

6.18a, figure 6.18b depicts transient I and V waveforms across the gNEMS device at selected TLP pulse heights corresponding to a few points in the ESD discharging I-V curve shown in Figure 6.18a. It is readily observed that, at the ESD triggering threshold (red point), the DUT current is almost zero (no discharging) when gNEMS stays off and ready to be turned on. As TLP pulses step up (from yellow to green to blue), the gNEMS are triggered progressively, and the DUT current increases when gNEMS gradually turns on and starts to discharge the ESD pulses. It is found that the mechanical gNEMS switch is like a non-snapback ESD device shown in figure 6.18a. As seen from figure 6.18b, the time latency of gNEMS triggering is reversely proportional to the height of the incident TLP pulse waveform which takes average I and V values by integration over a later window of 70%~90% in the DUT I/V waveforms obtained. This is the standard TLP procedure and we can find that after the gNEMS is turned on, the transient current will dramatically increase which means the small on-resistance. Furthermore, with the voltage step increasing, the time when current increases are earlier and a little larger because a higher

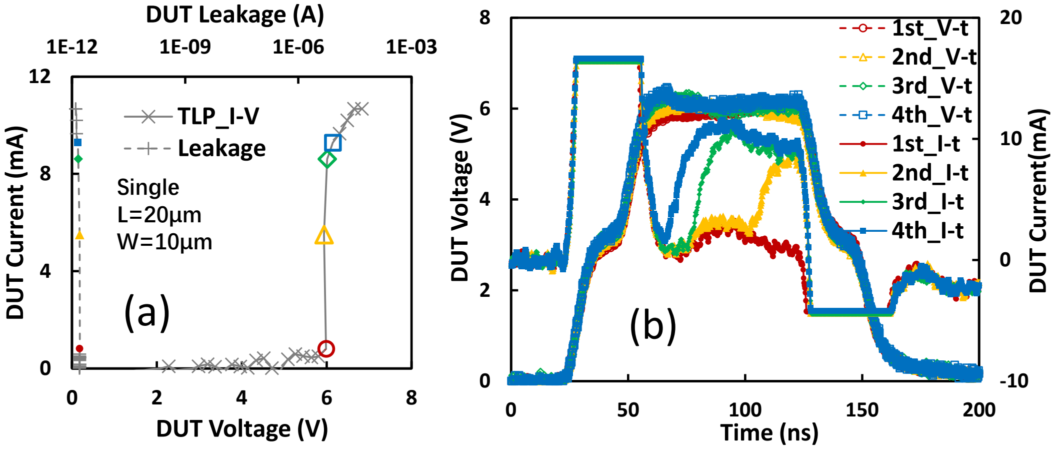


Figure 6.18 Sample gNEMS ESD discharging behaviors by TLP testing reveals dynamic gNEMS triggering and discharging characteristics: (a) instantaneous ESD discharging I-V curve, and (b) transient DUT V-t and I-t waveforms corresponding to the gNEMS triggering threshold (red) and full ESD conduction with increased TLP pulse heights (from yellow to green to blue).

TLP pulse height will gradually increase the electrostatic pull-down force to overcome the elastic restoring force, therefore, it takes time for the center part of the graphene film to touch the bottom electrode, hence to have the different time of turn-on.

6.8 gNEMS triggering mechanism by FEM simulation

6.8.1 gNEMS transient movement analysis on triggering

The actual transient behaviors such as bending and displacement of the suspended graphene ribbon of gNEMS devices under ESD stressing are extremely complicated and could not be experimentally monitored yet. Comprehensive FEM simulation can be conducted to investigate the transient ESD triggering characteristics of gNEMS devices under the same TLP square waveform pulse trains with $t_r = 10$ns and $t_d = 100$ns [50]. Figure. 6.19 depicts the 3D FEM simulation module for a gNEMS device with length and width of the graphene ribbon in X-axis and Y-axis, and physical displacement (bending) of and induced net contact force upon the graphene film in Z-axis, respectively. The colored area is the suspended graphene film with its two ends pinned to the Au pads at the two far ends on X-axis where the TLP pulse is applied to the graphene ribbon vertically via the top pads. The transparent black framework is the whole gNEMS device and the colored area is the suspended graphene membrane length 20μm, width 10μm, and the cavity depth 350nm, which shows the distribution of physical displacement of and the induced net contact force upon the suspended graphene ribbon.

As seen from figure 6.19a, under TLP stressing, the graphene ribbon is pulled downward starting from the center part of the membrane. There is a net induced pulling force proportional to the deformation degree, the electrostatic force minus the restoring elastic force, which becomes the net contact force when the graphene membrane touches the bottom Si. We can see the different colors standing for the different Z-axis displacement never beyond the cavity depth (maximum 350nm). It is depicted in figure 6.19b that the induced net contact force upon the graphene film when being pulled down completely and touching the bottom Si to trigger the gNEMS on, which is color coded to state the net force

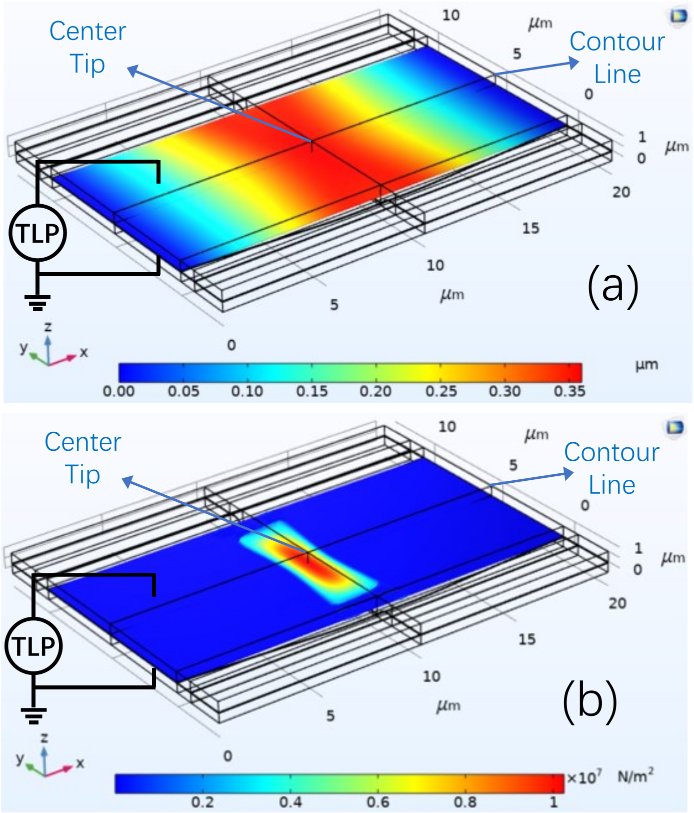


Figure 6.19 3D transient FEM simulation of gNEMS during TLP ESD stressing: (a) displacement of suspended graphene film (physical displacement in Z-axis with zero displacement at z=0 and "touching" when z=-350nm), and (b) contact force upon the graphene ribbon induced by the net pulling force when the graphene membrane touches the bottom Si (gNEMS = on).

strength, i.e., Blue for lowest magnitude and Red for highest magnitude. Figure 6.19b also suggests that when the suspended graphene ribbon starts to touch the bottom electrode and beyond, the graphene-Si contacting area originated in the center will increase (spreading from a contacting tip at the center to a growing larger center contacting area), proportional to the net contacting force induced onto the graphene ribbon, which varies in X and Y directions.

Figure 6.20 depicts the simulated vertical displacement of the graphene ribbon for the same sample gNEMS switch under TLP zapping of a square waveform of the height of 7.2V. The t-domain behavior reveals that the suspended graphene membrane has the largest bending at the center and the physical displacement increases as the TLP pulse continues until touching the bottom at V_{t1} ~ 6V (i.e., ESD triggering). As the TLP stressing continues beyond the ESD triggering threshold, the net pulling force will increase the contacting area of the graphene film upon the bottom Si, enhancing the ESD discharging area. The ESD trigging time can also be studied that requires complicated calibration

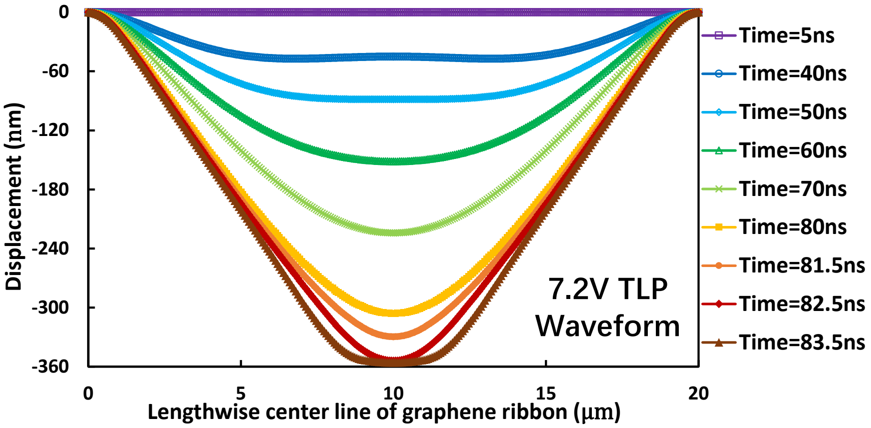


Figure 6.20 Simulated graphene ribbon vertical displacement in the time domain. As the time flow, graphene ribbon bends more and more until touching the bottom.

6.8.2 TLP voltage step influence on the gNEMS triggering by FEM simulation

The dynamic ESD triggering characteristics for gNEMS is further studied by applying a TLP pulse train of varying waveform heights from 3.6V, 4.8V, 6.0V, 7.2V to 8.4V with a voltage step of 1.2V. Figure 6.21 presents the transient graphene ribbon displacement of a sample gNEMS device (d=350nm, L=20µm, W=10µm) under different TLP pulses with the same rise time 10ns and pulse width 100ns.

It clearly shows that the gNEMS switch cannot be turned on under TLP pulses of a voltage lower than 6V. Then gNEMS is just turned on to discharge the ESD pulse at V_{t1} ~ 6V when the graphene membrane starts to touch the bottom Si. Finally, as the pulses voltage beyond 6V, the net pulling force will continue to enlarge the graphene-Si contact area, resulting in enhanced ESD discharging current. It is also observed that a higher voltage of TLP pulse reduces the initial graphene-Si touching time, hence accelerating the ESD triggering time. Meanwhile, with the pulse arising, the higher voltage of the TLP

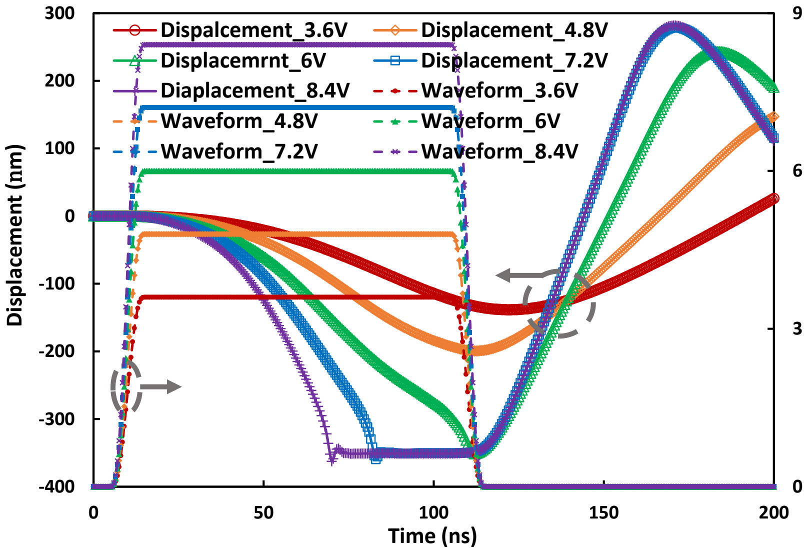


Figure 6.21 Transient ESD triggering behaviors for a sample gNEMS (d=350nm, L=20µm, W=10µm), stressed by a TLP pulse train of varying pulse heights (3.6V, 4.8V, 6V, 7.2V & 8.4V) with a step of ΔV = 1.2V, show the ESD triggering threshold at V_{t1} ~ 6V.

pulse will start to pull the graphene membrane downwards and keep the bottom-Si touching time duration longer until the pulse elapsing. Obviously, the graphene ribbon displacement is directly proportional to the strength of ESD-induced electrostatic pulling force.

6.9 Pulse condition and graphene dimension impact analysis

6.9.1 Pulse condition influence on gNEMS switch by FEM simulation

Figure 6.22 studies the pulse condition of the TLP on the gNEMS switch with length 20µm, width 10µm, and the depth of cavity 350nm. Figure 6.22a shows the transient graphene ribbon displacement under TLP waveforms of varying pulse durations from 75ns, 100ns, to 150ns with a fixed pulse height of 6V and rise time of 10ns. It is observed that, in addition to confirming triggering voltage about 6V under standard TLP of pulse width 100ns, a shorter TLP pulse of 75ns has not had enough energy and net pulling force during this time to pull the graphene ribbon contacting the bottom to trigger the gNEMS. However, a longer TLP pulse of 150ns can have enough time to pull the graphene ribbon downwards to touch the bottom-Si and keep a while until the TLP pulse elapsing, which turns on the gNEMS device and further enlarge the graphene-Si contacting area to enhance ESD discharging. We can validate this in the other method like in figure 6.22b. It shows that, by varying the stimulating TLP pulse heights, the ESD triggering voltage can be discovered for the gNEMS device under varying TLP pulse duration, i.e., V_{t1} = 7V, 6V, and 5.6V for

t_d = 75ns, 100ns, and 150ns, respectively. It is obvious that a longer TLP pulse duration

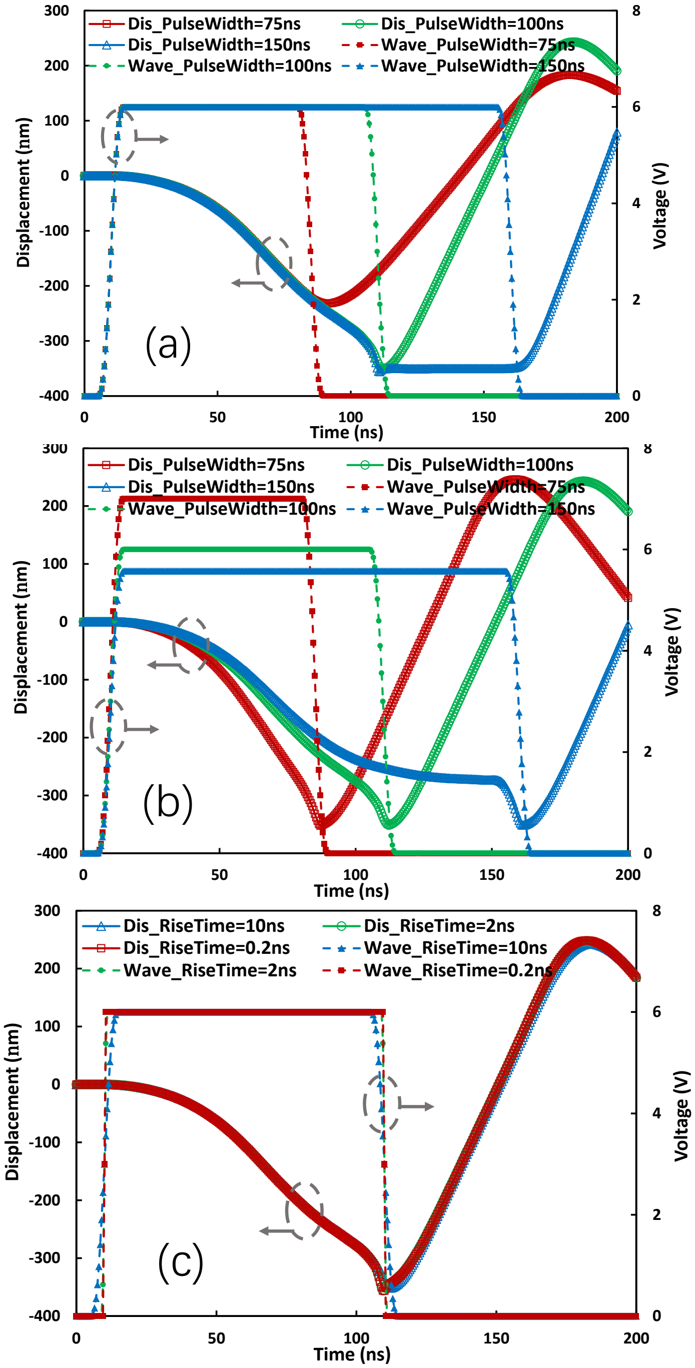


Figure 6.22 Transient ESD response under TLP stressing: (a) varying t_d of 75ns, 100ns and 150ns at fixed pulse height (6V) and rise time (t_r = 10ns), (b) varying t_d of 75ns, 100ns and 150ns and TLP pulse height at fixed t_r = 10ns, leading to different ESD V_{t1} of 7V, 6V and 5.6V, and (c) varying t_r of 0.2ns, 2ns and 10ns at fixed pulse height (6V) and t_d = 100ns showing no change in ESD triggering.

reduces the ESD triggering voltage due to more electrostatic force on the graphene membrane. It is readily found that even with the short pulse duration, we can increase the pulse height inducing a larger electrical force to make the graphene ribbon touch down. Therefore, it is a complex outcome that combines the net pulling force and sustainable time. In terms of the varying rise time, figure 6.22c reveals that, under fixed TLP pulse height of 6V and duration of t_d = 100ns, the variation in TLP pulse rise time, i.e., t_r = 0.2ns, 2ns, and 10ns, has almost no visible change in displacement of the graphene ribbon, apparently because the difference in pulse wave rise time is too small to have any impact on the net pulling force, hence does not affect gNEMS ESD triggering behavior.

6.9.2 Graphene dimension influence on gNEMS switch by FEM simulation

Impacts of gNEMS device dimensions on its triggering behaviors where device splits include length 10μm, 15μm, and 20μm with fixed width 10μm and depth 350nm as illustrated in figure 6.23. Figure 6.23d depicts the three samples of gNEMS with different lengths under TLP pulse of t_r = 10ns and t_d = 100ns. From figure 6.23a, figure 6.23b to figure 6.23c, we can find the different length dimensions of the graphene ribbon under the TLP height of 6V. It is obvious that gNEMS devices with shorter graphene ribbons (L = 10μm and 15μm) could not be triggered off at 6V due to stronger restoring force in a shorter graphene ribbon. It means that ESD V_{t1} increases as L decreases for gNEMS devices because the net pulling force is strongly related to the graphene ribbon length, i.e., a weaker electrostatic force and a stronger elastic restoring force induced onto a shorter graphene membrane.

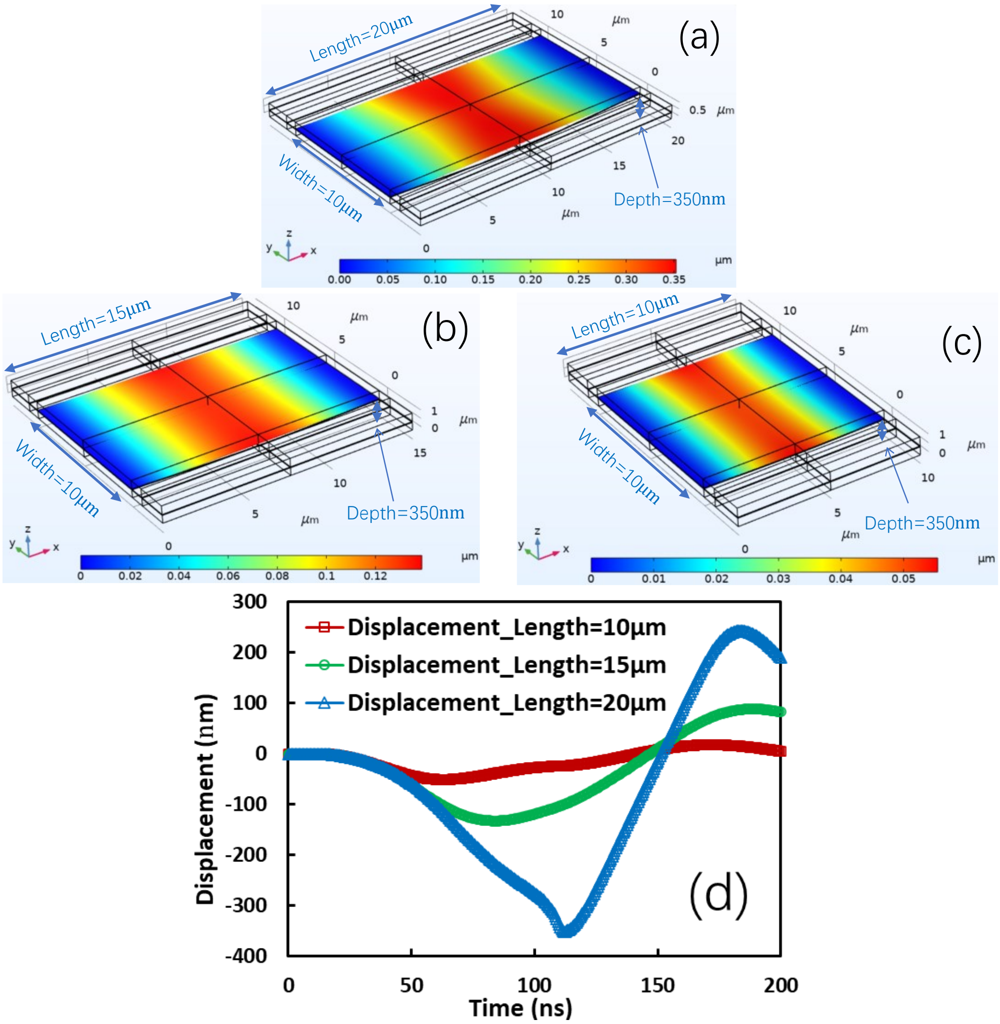


Figure 6.23 gNEMS device splits of different dimensions stressed by TLP pulse of 6V show changing graphene displacements: (a) L = 20µm, W = 10µm, d = 350nm (base device), (b) L = 15µm, W = 10µm, d = 350nm, (c) L = 10µm, W = 10µm, d = 350nm, and (d) transient ESD triggering behaviors by graphene ribbon length effect.

In terms of the varying width, it shows triggering voltage influenced by the width dimension where device splits include width 5µm, 10µm, and 15µm with fixed length 20µm and depth 350nm as illustrated in figure 6.24. From figure 6.24a, figure 6.24b to figure 6.24c, we can find the different length dimensions of the graphene ribbon under the TLP height of 6V. Figure 6.24d depicts the three samples of gNEMS with different lengths under TLP pulse of t_r = 10ns and t_d = 100ns. It is found that the graphene ribbon width seems to have little effect on gNEMS triggering, mainly because, a wider film induces more electrostatic force, while also increasing the restoring elastic force, hence balancing each other. The wider the graphene ribbon, it will be attracted by the larger electrical field force while the larger elastic force, which thus shows no difference of the net pulling force.

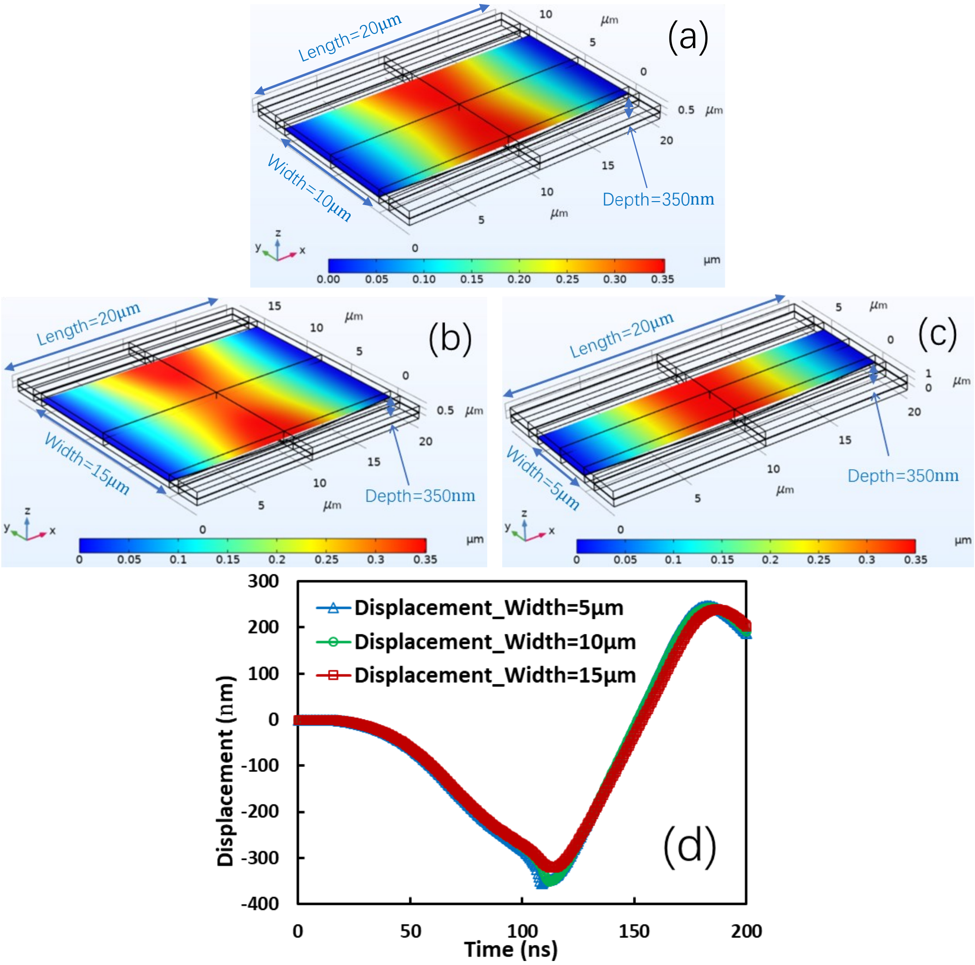


Figure 6.24 gNEMS device splits of different dimensions stressed by TLP pulse of 6V show changing graphene displacements: (a) L = 20μm, W = 10μm, d = 350nm (base device), (b) L = 20μm, W = 15μm, d = 350nm, (c) L = 20μm, W = 5μm, d = 350nm, and (d) transient ESD triggering behaviors by graphene ribbon width effect.

The depth influence on the triggering voltage can be displayed in figure 6.25 where device splits include depth 200nm, 350nm, and 500nm with fixed length 20μm and width

10μm stressed by TLP pulses of 6V. Clearly, gNEMS of d = 350nm turns on at V_{t1} ~ 6V. ESD triggering is easier for gNEMS of d = 200nm and the graphene-Si contact area continues increasing beyond the ESD triggering threshold. On the other hand, gNEMS of d = 500nm could not be turned on by the TLP pulse of 6V because the deep depth for the net pulling force induced by the electrical field force with the same duration time. The deeper the cavity is, the larger the triggering voltage results.

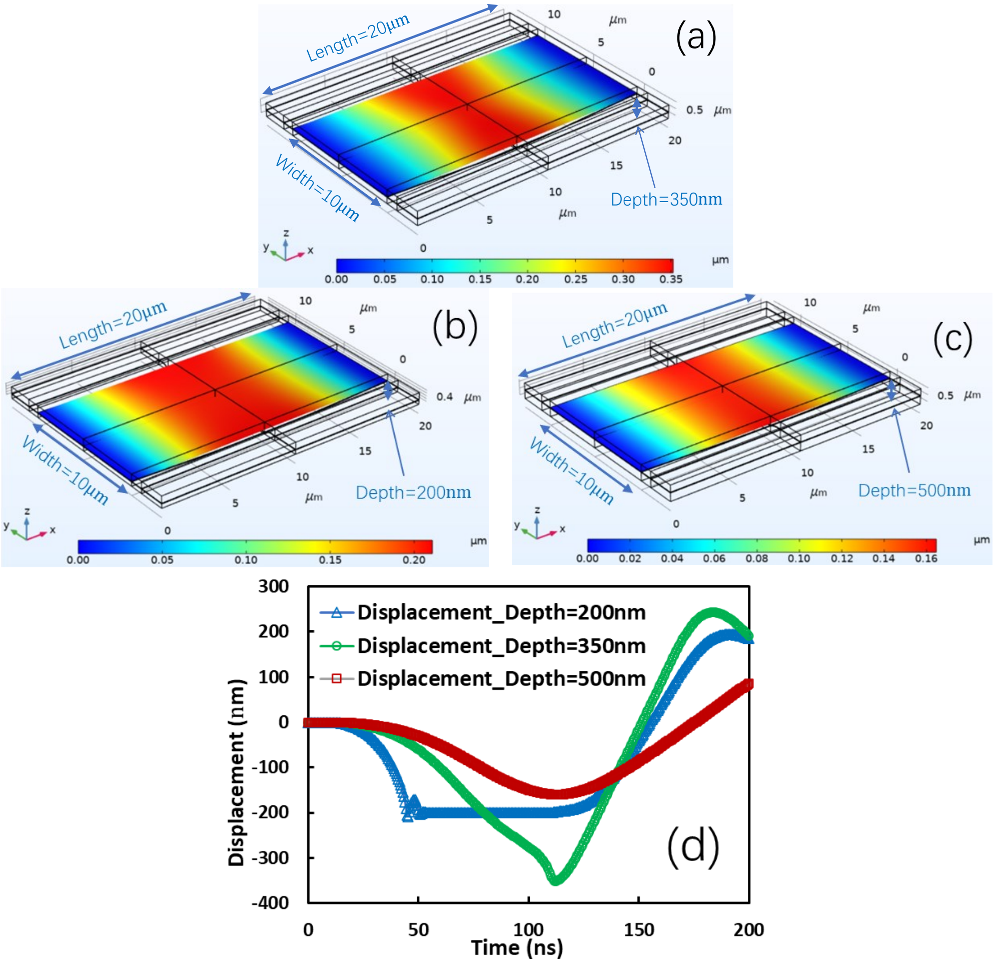


Figure 6.25 gNEMS device splits of different dimensions stressed by TLP pulse of 6V show changing graphene displacements: (a) L = 20μm, W = 10μm, d = 350nm (base device), (b) L = 20μm, W = 10μm, d = 220nm, (c) L = 20μm, W = 10μm, d = 5000nm, and (d) transient ESD triggering behaviors by graphene ribbon depth effect.

6.10 The shape of graphene membrane impact analysis

Here we evaluate the graphene membrane shape impacts in gNEMS ESD triggering behaviors. Figure 6.26a, figure 6.2b and figure 6.26c illustrate the three gNEMS device splits including a baseline rectangular (L =20µm, W = 10µm, d = 350nm), a dumbbell-shaped (central handle of 8µm long and 5µm wide) and a spindle-shaped (pulling arms of 6µm long and 5µm wide) graphene membranes. Figure 6.26d depicts the transient ESD triggering behaviors under ~6V TLP pulse stressing when the graphene ribbon just touches the bottom Si. It is observed that the dumbbell-shaped gNEMS could not be turned on because the reduced center graphene membrane substantially reduces the electrostatic pulling force. On the other hand, it is found that the spindle-shaped gNEMS can be turned

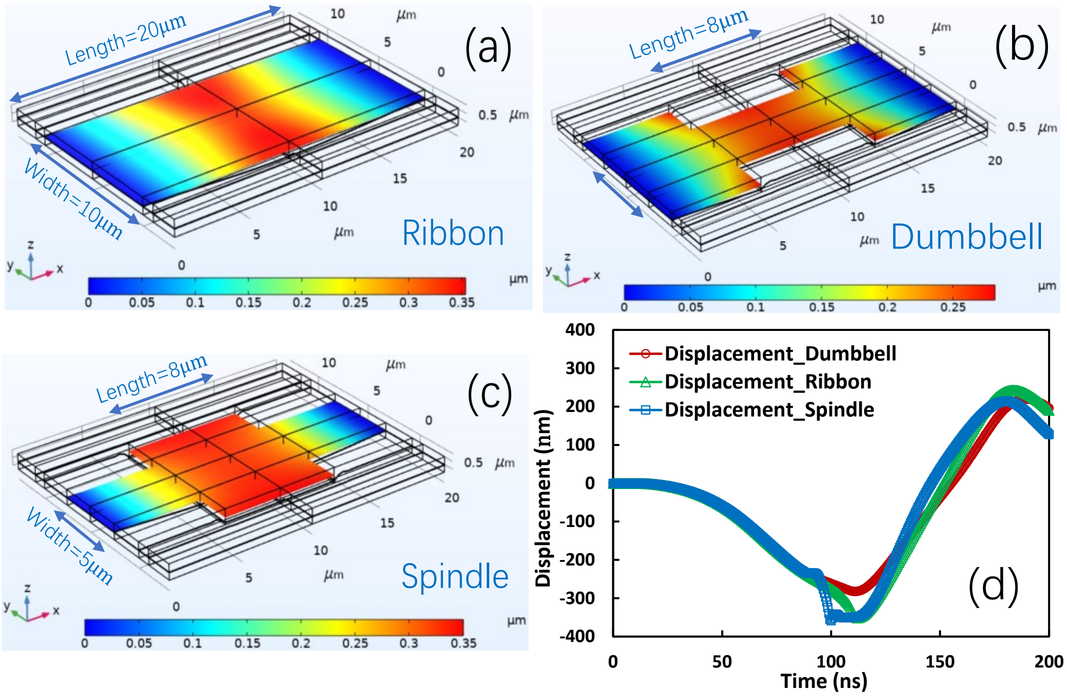


Figure 6.26 gNEMS devices (L =20µm, W = 10µm, d = 350nm) with different shapes of graphene membranes experience different net pulling force under TLP stressing: (a) baseline rectangular graphene ribbon, (b) dumbbell-shaped with the center part of 8µm long and 5µm wide, (c) spindle-shaped with the pulling arms being 6µm long and 5µm wide, and (d) transient response zapped by TLP pulses of ~6V.

on faster and at a lower threshold, and the net pulling force continues to enlarge the graphene-Si contacting area beyond the ESD triggering threshold. This is attributed to the that the narrower pulling arms have weaker elastic restoring force, while the electrostatic pulling force remains almost unchanged, resulting in a stronger net pulling force on the graphene membrane.

Figure 6.27a, figure 6.27b and figure 6.27c compare the spindle-shaped gNEMS devices (W = 10μm, d = 350nm, and a reduced center L =8μm) featuring varying pulling arm width of graphene ribbons such as 3μm, 5μm and 7μm width, which are names spindle side width (SW) 3, spindle-SW5 and spindle-SW7. Figure 6.27d depicts that, under TLP zapping of ~6V pulses, the observed ESD V_{t1} decreases slightly for gNEMS devices with narrower arms, mainly because of reduced elastic restoring force associated with a narrower

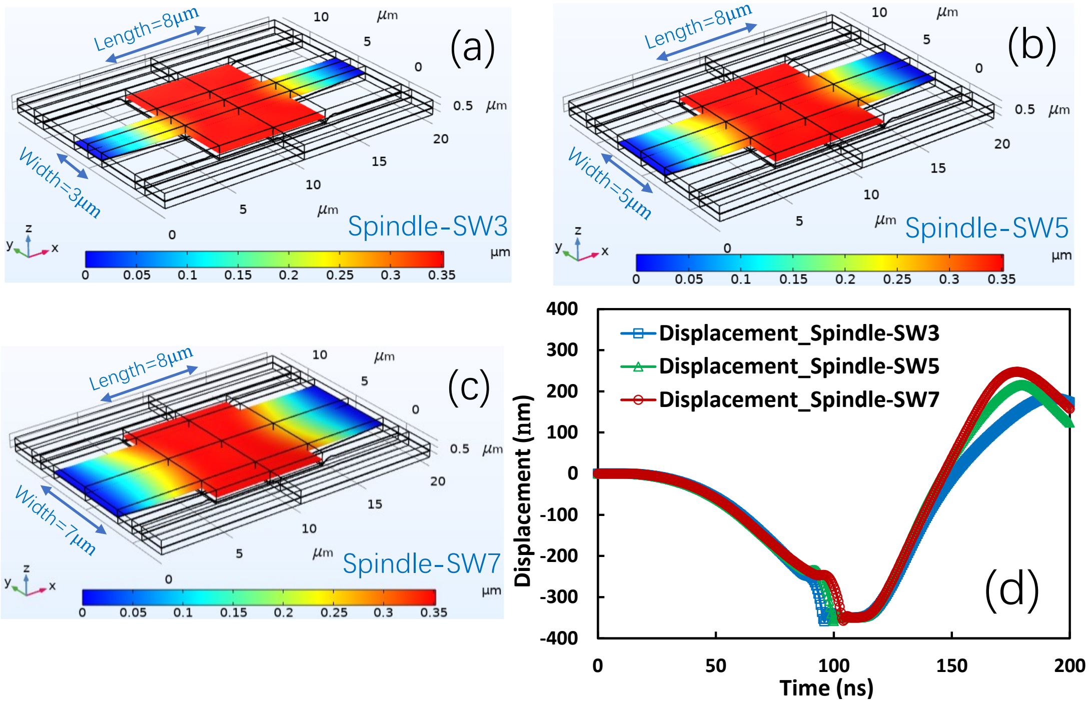


Figure 6.27 Spindle-shaped gNEMS devices (L =8μm, W = 10μm, d = 350nm) with different pulling arms under TLP stressing: (a) Spindle-SW3 of 3μm wide arms, (b) Spindle-SW5 of 5μm wide arms, (c) Spindle-SW7 of 7μm wide arms, and (d) transient displacement of gNEMS splits stressed by TLP pulse of ~6V.

graphene pulling arm while no change of the electrostatic pulling force from the center part. Therefore, the wider arm will provide less net pulling force resulting larger triggering voltage.

Similarly, figure 6.28a, figure 6.28b and figure 6.28c studies impacts of graphene ribbon shapes of spindle-shaped gNEMS devices (W = 10µm, d = 350nm, and arm width of 5µm) with a varying center length of 5µm, 8µm and 11µm under TLP stressing with time duration of 100ns and rise time 10ns. They have separately named spindle central length (CL) 5, spindle-CL8, and spindle-CL10. It is believed that a longer center graphene membrane induces a stronger electrostatic pulling force, however, the corresponding shorter pulling arms also increases the elastic restoring force substantially; as a result, the net pulling force may stay almost stable, hence may not impact ESD triggering of gNEMS devices, which is confirmed in Figure 6.28d.

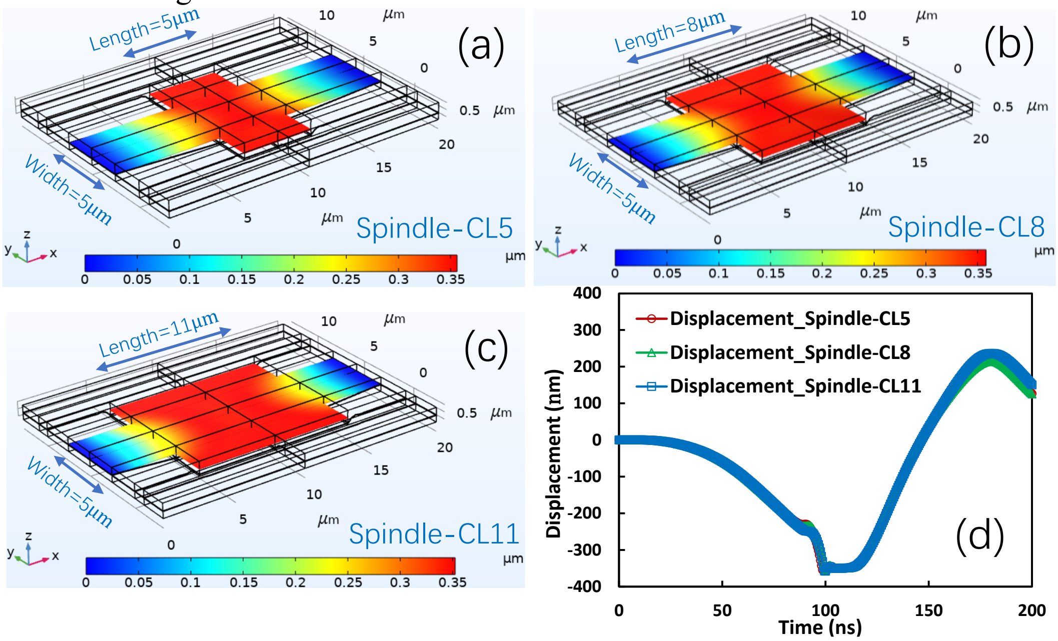


Figure 6.28 Spindle-shaped gNEMS devices (W = 10µm, d = 350nm) with varying length of 5µm, 8µm and 11µm under TLP zapping: (a) Spindle-CL5 of L = 5µm, (b) Spindle-CL8 of L = 8µm, (c) Spindle-CL11 of L = 11µm, and (d) transient response stressed by TLP pulses of ~6V.

Figure 6.29 further studies the graphene shape impacts on ESD triggering behaviors using three new gNEMS shapes of spindle and hammock shaped graphene membranes (L = 8µm, W = 10µm, d = 350nm) under TLP stressing of ~6V. The three splits are spindle-F1 (single pulling arm width of 6µm), hammock-F1 (two pulling arms of 3µm wide each), and hammock-F2 (three pulling arms of 2µm wide each) as depicted in figure 6.30a, figure 6.30b and figure 6.30c, respectively. In general, it is believed that the same central graphene membrane size introduces equal electrostatic pulling forces in the three gNEMS devices. However, it is observed that the induced electrostatic force is not uniform across the whole suspended graphene membrane, being stronger in the center and reducing gradually from the center to the outer edge. On the other hand, the contribution of the

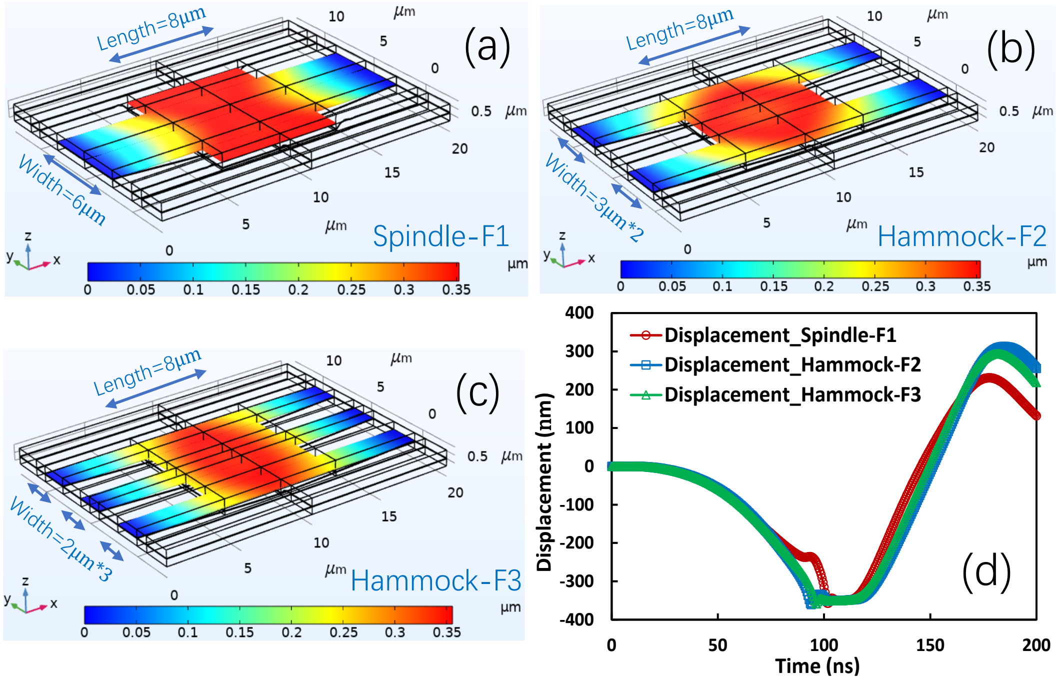


Figure 6.29 Comparing spindle and hammock shaped gNEMS devices (L =8µm, W = 10µm, d = 350nm) under TLP zapping: (a) Spindle-F1 featuring arm width of 6µm, (b) Hammock-F2 having two pulling arms of 3µm wide each, (c) Hammock-F3 having three pulling arms of 2µm wide each, and (d) transient response of gNEMS splits stressed by TLP pulse of ~6V.

elastic restoring force is directly related to the arm position (located in the center or on the edge), which results in a higher restoring force were corresponding to the center of the graphene membrane. Consequently, spindle-F1 shall induce a much stronger restoring force over its hammock-shaped counterparts. Meanwhile, hammock-F3 shall have somewhat more elastic force than hammock-F2 because of the middle arm percentile. This understanding is validated in Figure 6.30d, which shows that spindle-F1 is harder to trigger than its hammock counterparts; while hammock-F3 seems to be turned on slightly easier than hammock-F2.

The last shape influence study is involving three hammocks-shaped gNEMS splits of the same center graphene membrane (L = 8µm, W = 10µm, d = 350nm) and two pulling arms but with vary arm width i.e., hammock-SW2 (arm width of 2µm), hammock-SW3 (arm

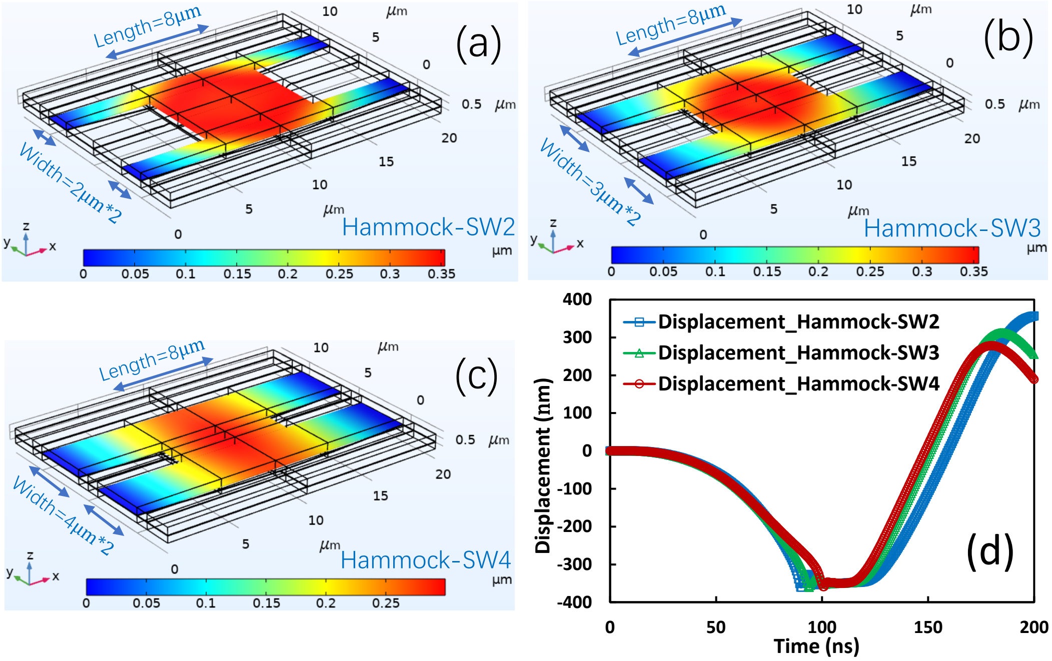


Figure 6.30 Influence of pulling arm locations on elastic restoring forces of gNEMS devices (L = 8µm, W = 10µm, d = 350nm) under TLP stressing: (a) hammock-SW2 having two arms of 2µm wide each, (b) hammock-SW3 featuring two arms of 3µm wide each, (c) hammock-SW4 having two arms of 4µm wide each, and (d) transient response of gNEMS splits stressed by TLP pulse of ~6V.

width of 3μm) and hammock-SW4 (arm width of 4μm), shown in figure 6.30a, figure 6.30b and figure 6.30c, respectively. Since the initiating electrostatic force is roughly the same due to the same central part, the wider pulling arms should induce a stronger elastic restoring force, and also closer to the center will add the effective elastic force. Hence, the net pulling force will be stronger for hammock-SW4 and decreases from hammock-SW3 to hammock-SW2, which is confirmed in figure 6.30d, which shows the earliest triggering voltage belonging to hammock-SW4 and the latest one belonging to the hammock-SW2. Overall, these case studies reveal the influences of graphene ribbon shapes on ESD triggering characteristics of varying gNEMS devices.

Chapter 7 Conclusion

In this work, we report the thermal management belonging to the design reliability. This dissertation reports the design, fabrication and analysis of a novel under-transistor in-hole thermal sensor diode structure. Measurement confirms that the fabricated under-FET in-hole diode functions properly as a thermal sensor. With the new concept validated experimentally, the new under-FET sensor can be a potential solution to achieving full-chip dynamic thermal mapping with a fine spatial resolution down to transistor level for accurate real-time chip-scale thermal management for future ICs. Within thermal management, the novel under-FET thermal sensor is the big contribution to solving the full-chip mapping problem with many advantages like no extra silicon area consumption, being close to the heat generation source, and CMOS process compatibility. Use the exemplary multi-PA module chip in 40nm CMOS as the victim MOSFETs and utilize the temperature sensing and control circuits to demonstrate the feasibility of thermal management via the auto feedback control circuits. Combining the intelligent post-data processing guidance opens a door for ML-enabled real-time full-chip intelligent thermal management for future ICs. What's more, the under-FET structure can also be used as the TSV-embedded ESD diode, which still has similar advantages like less silicon area are taken and CMOS compatible. Meanwhile, the same fabricated structure can be a novel concept of vertical TSV-like diode ESD structure which are simulated by HBM and CDM as well as silicon validated by TLP testing, which confirms full ESD discharge I-V characteristics. Furthermore, the new vertical TSV-like diode offers a disruptively new way for robust ESD protection without consuming precious Si area that is a potential solution for the distributed ESD protection especially for CDM, which will be the full-chip ESD protection floorplan for the next-generation ICs.

The systematic TCAD ESD mixed-mode flow is a necessity for the ESD design guidelines. TCAD calibration is the milestone for the quantitative simulation to acquire accurate data instead of the guideline trend. In this dissertation, we optimize the flow by using the calibration kit designing silicon devices compared with the validated silicon TLP testing on a whole batch of ESD devices. It is also noted that the TCAD fabrication process needs to accord with the real fabrication procedure ahead. Through the optimized TCAD ESD calibration flow, the data matches a good agreement, which can give accurate predictions based on the calibration to prevent future silicon ESD failure. It is also reported in this dissertation that true 3D TCAD ESD simulation overcomes the inaccuracy problem associated with 2D and pseudo-3D TCAD ESD simulation that has been commonly used by the semiconductor industry. Through the comprehensive comparison of 3D ESD structure using true 3D TCAD ESD simulation with 2D and 2.5D ESD simulation, we can prove that 3D simulation is more accurate than 2D and 2.5D especially more evident in some complexity unsymmetric structures though computing hungry. Since 3D TCAD simulation can accurately reveal the uneven current flowing due to the corner/edge factors, which results in the real heat dissipation phenomenon to lead the real-world ESD design. Based on the 3D TCAD simulation, we propose the design and 3D analysis of the first scalable sudoku DTSCR ESD array structure fabricated in a 22nm FDSOI CMOS technology. Silicon validation testing and TCAD 3D ESD simulation give a comprehensive confirmation showing the dramatic improvement in the ESD discharging area efficiency while offering the desired low ESD triggering voltage for LV ICs. ESD stressing measurements confirm that sudoku DTSCR ESD arrays achieve a very higher J_{t2} of ~6.47mA/μm^2 in TLP testing and J_{t2} of ~27.1mA/μm^2 in VFTLP zapping, equivalent to ESD area efficiency improvement over its finger-DTSCR ESD device counterparts by ~64.6% (TLP) and ~55.5% (VFTLP), respectively. In terms of the sudoku SCR ESD array, it achieves a higher J_{t2} of ~12.1mA/μm^2 in TLP testing and J_{t2} of ~43.9mA/μm^2 in VFTLP zapping, equivalent to ESD area efficiency improvement over its finger-DTSCR ESD device counterparts by ~105% (TLP) and ~84.8% (VFTLP), respectively. The measured ESD triggering voltage drops from V_{t1} ~ 10.8V for Sudoku-SCR ESD structures to V_{t1} ~ 1.92V for Sudoku-DTSCR ESD arrays, showing the desired ESD triggering tunability. Furthermore, a design scalability model was derived as design guidelines for optimizing scalable sudoku DTSCR for low voltage ICs.

Compared with the traditional in-Si PN junction based ESD protection device, the ultrafast and miniaturized mechanical graphene ESD switch may be the ideal and revolutionary future ESD protection device. Thru the single crystalline graphene utilization, the gNEMS has dramatic improvement, whose design, fabrication and characterization are presented in this dissertation. Systematic and statistical measurements show that the new single-crystal gNEMS devices dramatically outperform their poly-crystal counterparts in ESD protection robustness by TLP and VFTLP testing, as well as in device reliability and stability per accelerated aging evaluation. Furthermore, it is firstly reported on the VFTLP silicon validation to evaluate the CDM performance. Meanwhile, over 110-time reputable testing over TLP and VFTLP validates the reliability of gNEMS as a mature product in the future. ESD testing sets the records of the current density of J_{t2}~1.19×10^{10} A/cm^2 under TLP zapping and J_{t2}~6.09×10^9 A/cm^2 under VFTLP stressing, which is largely beyond other ESD devices like poly crystalline MEMS switch and PN junction device, which demonstrates the overall on-chip boosted robustness of ESD protection for future ICs. Then the transient 3D FEM simulation was conducted to investigate the device structural influences of gNEMS devices on dynamic ESD triggering characteristics. The mechanism of the gNEMS switch turn-on state is studied to reveal the counterbalance of two forces (electrostatic field force and elastic force), which determine the triggering occasion and transient contacting state. Through the comprehensive FEM 3D simulation, both the gNEMS device dimensions and the shapes of the suspended graphene membranes can be adjusted to optimize the ESD triggering behaviors, whose discoveries provide guidelines for the design optimization of gNEMS ESD switch structures. In total, this dissertation investigates the two main areas containing thermal management and ESD protection to fulfill the improvement of the IC design reliability.

Bibliography

[1] C. Prasad, S. Ramey and L. Jiang, "Self-heating in advanced CMOS technologies," *IEEE International Reliability Physics (IRPS) Symposium,* pp. 6A-4, 2-6 April 2017.

[2] K. Matsumoto, S. Ibaraki, S. K. Sueoka, K. Sakuma, H. Kikuchi, Y. Orii and F. Yamada, "Experimental thermal resistance evaluation of a three-dimensional (3D) chip stack," *IEEE Semiconductor Thermal Measurement and Management (Semi-Therm) Symposium,* pp. 8-13, 20-24 March 2011.

[3] P. K. Ramamoorthy and A. Bono, "Measurement and characterization of die temperature sensor," *IEEE Semiconductor Thermal Measurement and Management (Semi-Therm) Symposium,* pp. 41-44, 9-13 March 2014.

[4] S. Pan and K. Makinwa, "3.6 A CMOS resistor-based temperature sensor with a 10fJ·K2 Resolution FoM and 0.4°C (30) inaccuracy from -55°C to 125°C after a 1-point trim," *IEEE International Solid-State Circuits Conference (ISSCC),* pp. 68-70, 16-20 February 2020.

[5] B. Li, Q. Wang, Y. Xiao, X. Jiang, Y. Li, L. Xiao and Q. Gong, "On chip, high-sensitivity thermal sensor based on high-Q polydimethylsiloxane-coated microresonator," *Applied Physics Letters,* vol. 96, no. 25, p. 1109, 2010.

[6] C. Tsamis, A. Nassiopoulou and A. Tserepi, "Thermal properties of suspended porous silicon micro-hotplates for sensor applications," *Sensors and Actuators B: Chemical,* vol. 95, no. 1-3, pp. 78-82, 2003.

[7] R. Quan, U. Sonmez, F. Sebastiano and K. Makinwa, "4600μm2 1.5°C (3σ) 0.9 kS/s thermal-diffusivity temperature sensor with VCO-based readout," *IEEE International Solid-State Circuits Conference (ISSCC),* pp. 1-3, 22-26 February 2015.

[8] R. Zhang, K. Yang, T. Liu and L. Milor, "Impact of front-end wearout mechanisms on the performance of a ring oscillator-based thermal sensor," *IEEE International Workshop on Advances in Sensors and Interfaces (IWASI),* pp. 258-263, 13-14 June 2019.

[9] S. Pan, C. Gürleyük, M. Pimenta and K. Makinwa, "A 0.12 mm2 Wien-bridge temperature sensor with 0.1°C (3σ) inaccuracy from -40°C to 180°C," *IEEE International Solid-State Circuits Conference (ISSCC),* pp. 184-186, 17-21 February 2019.

[10] M. A. Pertijs, K. Makinwa and J. Huijsing, "A CMOS smart temperature sensor with a 3σ inaccuracy of ±0.1°C from −55°C to 125°C," *EEE J. Solid-State Circuits,* vol. 40, no. 12, pp. 2805-2815, 2005.

[11] C. Zhao, Y. Wang, D. Genzer, D. Chen and R. Geiger, "A CMOS on-chip temperature sensor with -0.21°C 0.17°C inaccuracy from -20°C to 100°C," *IEEE*

International Symposium on Circuits and Systems (ISCAS), pp. 2621-2625, 19-23 May 2013.

[12] S. H. Pan, N. Chang and T. Hitomi, "3D-IC dynamic thermal analysis with hierarchical and configurable chip thermal model," *IEEE EEE International 3D Systems Integration (3DIC) Conference,* pp. 1-8, 2-4 October 2013.

[13] A. Bakker and J. H. Huijsing, "Micropower CMOS smart temperature sensor," *IEEE European Solid-State Circuits (ESSCIRC) Conference,* pp. 238-241, 19-21 September 1995.

[14] L. Luh, J. Choma, D. J and H. Chiueh, "A high-speed CMOS on-chip temperature sensor," *IEEE European Solid-State Circuits Conference (ESSCIRC),* pp. 290-293, 21-23 September 1999.

[15] C. Li, Q. Chen, M. Di, Z. Pan and A. Wang, "In-hole diodes for on-chip thermal sensing," *IEEE Electron Devices Technology and Manufacturing (EDTM) Conference,* pp. 683-686, 6-21 April 2020.

[16] C. Li, Q. Chen, F. Zhang, M. Di, Z. Pan, F. Lu and A. Wang, "Under-FET Thermal Sensor Enabling Smart Full-Chip Run-Time Thermal Management," *IEEE Journal of the Electron Devices Society,* vol. 8, pp. 1242-1248, 2020.

[17] A. Wang, Practical ESD Protection Design, John Wiley & Sons, 2021.

[18] F. Lu, R. Ma, Z. Dong, L. Wang, C. Zhang, C. Wang, Q. Chen, X. Wang, F. Zhang, C. Li and H. Tang, "A systematic study of ESD protection co-design with high-speed and high-frequency ICs in 28 nm CMOS," *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 63, no. 10, pp. 1746-1757, 2016.

[19] Z. F, L. C, D. M, P. Z, L. C, C. N and W. A., "Design and Analysis of a 28GHz 9KV ESD-Protected Distributed Travelling-Wave TRx Switch in 22nm FDSOI," *IEEE Journal of the Electron Devices Society,* vol. 8, pp. 655-661, 2020.

[20] W. C., L. F., C. Q., Z. F., L. C., W. D. and W. A., "A study of impacts of ESD protection on 28/38GHz RF switches in 45nm SOI CMOS for 5G mobile applications," *IEEE Radio and Wireless Symposium (RWS),* pp. 157-160, 15-18 Janurary 2018.

[21] Z. F., W. C., L. F., C. Q., L. C., W. XS., L. D. and W. A., "A full-chip ESD protection circuit simulation and fast dynamic checking method using SPICE and ESD behavior models," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 38, no. 3, pp. 489-498, 2018.

[22] D. M., L. C., P. Z. and W. A., "Pad-based CDM ESD protection methods are faulty," *IEEE Journal of the Electron Devices Society,* vol. 8, pp. 1297-1304, 2020.

[23] D. M., P. Z., L. C. and W. A., "ESD Design Verification Aided by Mixed-Mode Multiple-Stimuli ESD Simulation," *IEEE Journal of the Electron Devices Society,* vol. 9, pp. 1194-1201, 2021.

[24] L. C., W. C., C. Q., Z. F., L. F., S. X., Y. Y., L. H., C. G., L. T. and F. D., "Characterization and analysis of diode-string ESD protection in 28nm CMOS by

VFTLP," *IEEE 24th International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA),* pp. 1-4, 4-7 July 2017.

[25] L. C., Z. F., W. C., Q. C., L. F., W. H., D. M., C. Y., Z. H. and W. A., "Temperature Dependence of Diode and ggNMOS ESD Protection Structures in 28nm CMOS," *14th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT),* pp. 1-3, 1-3 November 2018.

[26] A. K. Geim and S. N. Konstantin, "The rise of graphene," *Nanoscience and Technology: a collection of reviews from nature journals,* pp. 11-19, 2010.

[27] L. C., C. Q., N. J., Z. F., W. H., D. M., W. C., X. Y. and W. A., "Graphene-based on-chip esd protection," *Electron Devices Technology and Manufacturing Conference (EDTM),* pp. 176-178, 12-15 March 2019.

[28] C. Li, Q. Chen, J. Ng, M. Di, Z. Pan, T. Jones, Y. Xie, J. Hopkins and A. Wang, "Emerging Graphene-Based on-Chip ESD Protection," *AAAFM Material Comunication,* pp. 86-91, 2021.

[29] C. Q., M. R., Z. W., L. F., W. C., L. O., Z. F., L. C., T. H., X. YH. and W. A., "Systematic characterization of graphene ESD interconnects for on-chip ESD protection," *IEEE Transactions on Electron Devices,* vol. 63, no. 8, pp. 3205-3212, 2016.

[30] C. Q., L. C., N. J., L. F., W. C., Z. F., M. R., X. YH. and W. A., "Transient characterization of graphene NEMS Switch ESD protection structures," *IEEE Electron Devices Technology and Manufacturing Conference (EDTM),* pp. 95-96, 1-2 March 2017.

[31] D. M., L. C., P. Z. and W. A., " Misconception with pad-based CDM ESD protection," *IEEE Electron Devices Technology & Manufacturing Conference (EDTM),* pp. 1-4, 6-21 April 2020.

[32] C. Li, M. Di, Z. Pan and A. Wang, "Enabling 3D Heterogeneous Structures Towards Smart Chips: A Review," *Advances in Science, Technology and Engineering Systems Journal (ASTESJ),* vol. 5, no. 1, pp. 267-273, 2020.

[33] L. C., D. M., P. Z., W. H. and W. A., "Vertical TSV-Like Diode ESD Protection," *IEEE Electron Devices Technology & Manufacturing Conference (EDTM),* pp. 1-3, 8-11 April 2021.

[34] D. M., L. C., P. Z. and W. A., "Non-Pad-Based in Situ in-Operando CDM ESD Protection Using Internally Distributed Network," *IEEE Journal of the Electron Devices Society,* vol. 9, pp. 1248-1256, 2021.

[35] L. C., P. Z., D. M., Z. F., L. Z., J. N. and W. A., "ESD device layout design guidelines by 3D TCAD simulation," *IEEE Electron Devices Technology & Manufacturing Conference (EDTM),* pp. 1-4, 6-21 April 2020.

[36] P. Z., L. C., D. M., Z. F. and W. A., "3D TCAD Analysis Enabling ESD Layout Design Optimization," *IEEE Journal of the Electron Devices Society,* vol. 8, pp. 1289-1296, 2020.

[37] H. Xie, H. Feng, R. Zhan, A. Wang, D. Rodriguez and D. Rice, "A New Low-Parasitic Polysilicon SCR ESD Protection Structure for RF ICs," *IEEE Electron Device Letters,* vol. 26, no. 2, pp. 121-123, 2005.

[38] M. P. Mergens, C. C. Russ, K. G. Verhaege, J. Armer, P. C. Jozwiak, R. Mohn, B. Keppens and C. Trinh, "Diode-Triggered SCR (DTSCR) for RF-ESD Protection of BiCMOS SiGe HBTs and CMOS Ultra-Thin Gate Oxides," *IEEE International Electron Devices Meeting (IEDM),* pp. 21-23, 8-10 December 2003.

[39] L. C., Z. F., P. Z., D. M., W. C. and W. A., "Sudoku DTSCR ESD Array in 22nm FDSOI," *IEEE Electron Devices Technology & Manufacturing Conference (EDTM),* pp. 1-3, 8-11 April 2021.

[40] L. C., Z. F., W. C., P. Z., D. M. and W. A., "Analyze Scalable Sudoku-Type DTSCR ESD Protection Array Structures in 22nm FDSOI," *IEEE Journal of the Electron Devices Society,* vol. 9, pp. 1137-1144, 2021.

[41] C. Q., L. C., L. F., W. C., Z. F., W. T., X. X., Z. K., L. X., N. J. and X. YH., "Characterization of single-crystalline graphene ESD interconnects," *IEEE 12th International Conference on ASIC (ASICON),* pp. 977-980, 25-28 October 2017.

[42] A. F. A. M. L. B. R. K. K. N. a. C. C. A. Eckmann, "Probing the nature of defects in graphene by Raman spectroscopy," *Nano Letters,* vol. 12, no. 8, pp. 3925-3930, 2012.

[43] T. Wu, X. Zhang, Q. Yuan, J. Xue, G. Lu, Z. Liu, H. Wang, H. Wang, F. Ding, Q. Yu and X. Xie, "Fast growth of inch-sized single-crystalline graphene from a controlled single nucleus on Cu–Ni alloys," *Nature Materials,* vol. 15, no. 1, pp. 43-47, 2016.

[44] T. Wu, G. Ding, H. Shen, H. Wang, L. Sun, D. Jiang, X. Xie and M. Jiang, "Triggering the continuous growth of graphene toward millimeter-sized grains," *Advanced Functional Materials,* vol. 23, no. 2, pp. 198-203, 2013.

[45] C. Li, M. Di, Z. Pan and A. Wang, "A Study of Materials Impacts on Graphene Electrostatic Discharge Switches," *IEEE Electron Devices Technology & Manufacturing Conference (EDTM),* pp. 1-3, 8-11 April 2021.

[46] Q. Chen, J. Ng, C. Li, F. Lu, C. Wang, F. Zhang, Y. Xie and A. Wang, "Systematic transient characterisation of graphene NEMS switch for ESD protection," *Micro & Nano Letters,* vol. 12, no. 11, pp. 875-880, 2017.

[47] Q. Chen, C. Li, F. Lu, C. Wang, F. Zhang, A. Wang, J. Ng and Y. Xie, "TLP measurement and analysis of graphene NEMS switches for on-chip ESD protection," *International Conference on Nano/Micro Engineered and Molecular Systems (NEMS),* pp. 370-374, 9-12 April 2017.

[48] L. C., C. Q., N. J., Z. F., W. H., D. M., P. Z., W. T., Z. K. and X. X. Y., "Design, Fabrication and Characterization of Single-Crystalline Graphene gNEMS ESD Switches for Future ICs," *IEEE Transactions on Device and Materials Reliability,* vol. 21, no. 3, pp. 331-337, 2021.

[49] R. P. K. B. a. H. A. T. S. Pamidighantam, "Pull-in voltage analysis of electrostatically actuated beam structures with fixed–fixed and fixed–free end conditions," *J. Micromechanics and Microengineering,* vol. 12, no. 4, p. 458, 2012.

[50] C. Li, M. Di, Z. Pan, F. Zhang, Q. Chen and A. Wang, "Investigating Graphene gNEMS ESD Switch for Design Optimization," *IEEE Journal of the Electron Devices Society,* vol. 9, pp. 1172-1180, 2021.