PROCEEDINGS OF SPIE

SPIDigitalLibrary.org/conference-proceedings-of-spie

Accelerate lithography improvement for high performance computing

John Chen

John Y. Chen, "Accelerate lithography improvement for high performance computing," Proc. SPIE 10809, International Conference on Extreme Ultraviolet Lithography 2018, 1080902 (3 October 2018); doi: 10.1117/12.2504658

SPIE.

Event: SPIE Photomask Technology + Extreme Ultraviolet Lithography, 2018, Monterey, California, United States

Accelerate Lithography Improvement for High Performance Computing

John Y. Chen

NVIDIA Corporation, 2701 San Tomas Expressway, Santa Clara, CA 95050, USA

ABSTRACT

Artificial intelligence (AI) with deep learning is taking off based on High Performance Computing (HPC) engines fueled by “Big Data” in the cloud. NVIDIA’s general-purpose GPU (Graphics Process Unit) is the ideal platform to accelerate computation with its inherent massive parallel processing capability. The Deep Learning machines for AI would be the new driver for the semiconductor industry.

In the past, the minimum feature on a semiconductor chip has greatly shrunk with Moore’s law. From 1971 to 2018, as the feature size scaled from 10 μm to 10 nm, the transistors per chip increased from thousands to billions, and remarkably, its price has gone down to few % of a cent. However, going forward with Moore’s law has finally discontinued in its scaling cadence, the economic benefit of scaling can hardly justify the increased cost of wafer manufacturing unless we can find a way to advance lithography and pack more transistors on a chip. In the near future, the only practical way is EUV including EUV mask, which has made great progress lately even though still challenges ahead.

Illustrated by the latest and most complicated AI chip on this planet, the presenter will describe key lithographic requirements from an end user point of view. An example is given to show how precise the edge placement of a geometry needs to be controlled in order to scale IC density for the future technology nodes.

WHY IS AI TAKING OFF NOW? IS IT FOR REAL?

AI research based on neural network has been around for few decades, why is it happening now? This is simply because IC evolution observed by Moore’s law [1] has made transistors more than the total number of ants in this world and the transistor has become the cheapest product. According to SEMI’s data in 2015 [2], one thousand transistors cost only 0.03 cent; one can hardly find anything cheaper than that nowadays. Meanwhile, the computing power measured as # of trillion operations per second (TOPS) has gone up

tremendously as well. These are the two primary reasons that AI is taking off rapidly now and it’s for real.

Computing power and big data finally caught up with algorithms to Detect, Reasoning, and Action. Just like launching a rocket requiring huge amount of fuel and a powerful engine [3], AI is taking off due to the availability of big data provided by numerous cheap transistors and the High Performance Computing (HPC) operating in many TOPS.

Already, we have seen Google’s AlphaGo AI beats the world’s best human Go player. It can be expected that AI-assisted healthcare would enable earlier detection on Alzheimer and cancer. Deep learning can imitate art, re-creates Monet, Picasso, Vincent van Gogh. Indeed, AI is spreading to everywhere ranging from teaching robots do manufacturing to caring elders to self-driving cars.

WHY GPU (GRAPHIC PROCESS UNIT)?

GPU, not just accelerates CPU, its inherent massive parallel processing, mixed precision capabilities, and new distributed frameworks for managing HPC applications with an unprecedented level of big data is best suited for training and deep learning in AI.

When NVIDIA was first founded, it has focused on GPU for PC gaming, laid its fundamental technology on image processing and rendering. Ten years ago, NVIDIA began to drive GPU for general purpose high speed computations, named GPGPU for General Purpose GPU.

Today, GPGPU has become the natural platform for AI with Deep Learning especially in the area of Image Recognition. We have built VR/AR (Virtual Reality/Augmented Reality), Pro-Visualization, Data Center and Autonomous Driving on the GPGPU platform based on one architecture, CUDA. We are most suited to expand our technologies from Gaming to Robotics, from Virtual Reality to Reality, from Simulation to Detection and Action, and from Smart Cars to Smart City and Intelligent Healthcare.

GPU AT THE END OF MOORE'S LAW

After its profound impact to the IC industry for the past few decades, Moore's law, defined as "*doubling the transistor counts per unit area every two years*" has now ended its cadence. Ending Moore's law's favors the use of GPU much more than that to a CPU. Why?

1. CPU can't use many more transistors (Amdahl's Law), but GPU can.
2. Circuit speed no longer increases (Dennard's scaling ended even earlier), due to power limitation.
3. Performance in TOPS depends more on parallel processing and architecture innovation rather than transistor scaling and engineering.
4. World needs more TOPS.

Shown in the following figure, it is clear that while CPU has leveled off its performance, GPU has been improving with no sign of slowing down.

Fig. 1 Forty years of microprocessor trend data showing the increase of transistor counts for GPU (white triangles) and CPU (blue dots)

One good example is the IBM DEEP LEARNING breakthrough showing that IBM deep learning speeds up linearly with # of GPU's, achieving great results for up to 256 NVIDIA Tesla GP100 GPUs [4].

VOLTA, THE MOST COMPLICATED CHIP ON THE PLANET

Fig.2 shows the GV100 die surrounded by four HBM (High Bandwidth Memory) stacks of DRAM's integrated in a 2.5D interposer with the following features:

- 815 mm² die size
- 21 billion Transistors
- 14 layers of interconnects with
 - >100 billion CT's/Via's
 - 56km of 1x metal length
 - 32nm metal width
- 9.5 km fins in total length

Fig. 2 VOLTA 100 die shown in the center surrounded with 4 stacks of HBM DRAM units on a 2.5D interposer platform.

This new product has 5x improvement over the preceding NVIDIA Pascal GPU accelerators, and delivers the equivalent performance of 100 CPUs for deep learning.

WHAT IT TAKES TO MAKE IT?

In 2009, the author delivered an IEDM plenary speech preaching for the three zeros: **Zero defect, Zero leakage and Zero variation** for the IC industry to meet the requirement of future GPU's [5]. Then, in the 2012 SPIE keynote speech [6], he introduced three *P's* (**Performance, Perfection and Precision**) for the major technological challenges in making complicated IC chips. Arguably, **Precision** is probably the most important requirement out of the three *P's*, and it is indeed the one in the hands of lithography and mask-making people.

First of all, to achieve performance, transistors are often engineered to the extreme with key processes. Transistor performance stands on the steep slope, meaning very sensitive to these process parameters. For a 10nm technology, a 0.5 nm CD variation can lead to

15% transistor drive current degradation. Therefore, the key process parameters must be controlled precisely to keep the transistor stable. Otherwise, one cannot have consistent performance gain. We need to have tight specs for competitive designs, and tight control for better yield. SPC is absolutely necessary to make this happen.

SPC, A NECESSITY FOR PRECISION

SPC (Statistical Process Control) is a quality tool that enable the control of a manufacturing process in a quantitative way. With the spec limits and the actual variations measured in standard deviation or sigma (σ), one can quantify the controllability easily. In the older days, we measure and use 3σ to indicate the control and the precision of key process parameters and so long as the 3σ value is within the spec limits, we were satisfied.

This is no longer sufficient today simply because the sample size needs to be considered is in billions and tens billions as required to make a chip like VOLTA. Statistically, we need to compare 6σ to the spec limit as 6σ represent approximately 1 DPPB (Defective Parts Per Billion), and for chips with billions of transistors and tens or hundreds of billions of Via's, we have to have 6σ within the spec limits. That means that variation expressed in σ must be reduced to less than 1/6 of the spec limit. The precise placement of the geometries is more critical today than ever as shown by the following example.

CONTROLLING GATE TO S/D CONTACT SPACE DOWN TO SUB nm

To increase transistor count per unit area, we need to continue striving for shrinking standard cells. Standard Cell is a unit cell used repeatedly many times in an IC logic design. The total chip area is determined by the standard cell sizes.

In general, a standard cell has its height defined by the metal pitch and its width defined by **CGP (Contacted Gate Pitch)** where $CGP = \text{Gate length} + \text{S/D CT width} + 2 \times \text{Spacer width}$. In order to shrink the CGP, we want shorter Gate length, smaller CT size and narrower space between the Gate and the S/D Contact. While shortening the gate length is limited by the leakage due to MOSFET short-channel effect, and the smallest CT size is constraint by the contact resistance, we are left to narrow the space between the Gate and the S/D CT. However, this very narrow space must be controlled precisely. If the space becomes zero, an electrical short would occur. Even with a finite space, if it's too narrow, it may lead dielectric breakdown over time. A

short kills yield and an almost short is even worse because it manifests itself as a TDDDB (Time Dependent Dielectric Breakdown) related reliability problem. The following example illustrates how precise the CD and overlay need to be when using a 14nm technology to manufacture a chip with 3.3 billion transistors, the smallest chip in our GPU family.

With three fins per transistor in our design, we have about 20 billion this type of spaces per chip, and because the CGP pitch is fixed while other dimensions have variations, the Space dimension in Si is a function of Gate CD, CT CD, and Gate-CT overlay. The total or the net variation for the Space can be expressed as:

$$\sigma_{\text{total}} = \text{SQRT} [(\sigma_{\text{Gate CD}/2})^2 + (\sigma_{\text{CT CD}/2})^2 + \sigma_{\text{Gate-CT Overlay}}^2]$$

Assuming a normal statistical distribution, to get <1/20B (0.05 DPPB) failure rate, we need $6.5 \sigma_{\text{total}}$!

Fig.3 The critical Space between a Gate and a Contact on S/D is in risk if not precisely controlled.

Since the narrowest space in Si to meet reliability requirement on TDDDB must be at least 2 nm, the $6.5 \sigma_{\text{total}}$ must then be <11nm based on the CGP pitch in our design. If we equally divide the σ_{total} to the 3 variations, then $\frac{1}{2} \sigma$ of the CD (for the Gate or S/D Contact) and 1σ the Overlay all need be < 0.98 nm.

HOW PRECISE DO WE NEED TO BE IN THE NEXT TECHNOLOGY NODE?

It was proposed that CGP scales down to 32 - 42nm in 5 nm technology node [7]. Let's assume $CGP = 40\text{nm}$ for $\sim 0.7 \times$ linear scaling. With $L_g = 15\text{nm}$, CT on S/D = 15nm, the nominal Gate-S/D CT space = 5 nm $(40-15-15)/2$. Since we need minimum 2 nm final Space due to

the TDDb spec for reliability, only 3 nm left for the margin to accommodate variations.

For the same chip with only 3.3 billion transistors, again we need 6.5σ to ensure every transistor works, then the total budget for $1\sigma_{\text{total}} = 0.46\text{nm}$ ($3\text{nm}/6.5$) to account for these 3 imprecisions: Gate CD, CT CD and Gate-CT misalignment. For Gate or CT CD, if controlled at $3\sigma = 0.5\text{nm}$ and the misalignment controlled $3\sigma = 1.5\text{nm}$, then $\sigma_{\text{total}} = 0.51\text{nm} > 0.46\text{nm}$, **still fail!** This says that we need do better in **precision**, or we have to relax the scaling.

WHAT ELSE CAN WE DO?

Since the task of searching the sources of these variations is statistical in nature and strongly dependent on big data with empirical results, it suggests that we use AI with deep learning to identify the sources of the variations in a manufacturing process, then eliminate or reduce them to enhance Precision, hence improve performance, yield/reliability and cost for the IC industry.

We can also employ AI and deep learning to make smart tools including EUV scanners and mask making e-beam machines. The AI-assisted tools not only possess automation and robotics, it also has self-diagnosis ability to mitigate variations and achieve ultimate precision in a reproducible way. The AI-enhanced tools can definitely make better GPU chips, hence better AI machines and that again helps make better Lithography and e-beam mask-making tools, smart tools.

CONCLUSION

AI is indeed taking off and is expanding everywhere from autonomous vehicle to robotics, smart city, and intelligent healthcare. We shall see its impact to our life in such a profound way that has never been seen before. General Purpose GPU is an ideal platform for AI and deep learning due to its inherent massive parallel processing capability and its performance continuing to improve regardless the ending of the Moore's law.

AI implemented with advanced GPU's opens huge opportunities in the IC industry, but meanwhile brings challenges in manufacturing complicated GPU chips. Out of the three challenges of *Performance*, *Perfection* and *Precision*, we can argue that **Precision** is most critical and is indeed in our hands. To continue

advancing IC chips for AI applications, we need to minimize CD and overlay variations, making the corresponding standard deviations down to 0.17 nm and 0.5 nm in the next technology node.

Lastly, using AI to find and eliminate variations would make the lithographic tools much more precise and reproducible, hence more capable to make more complicated AI chips. It is author's best wish that we together can generate such a positive reinforcement loop and make the IC industry continue to growth in the era of Artificial Intelligence.

REFERENCES

1. G. Moore, "Cramming more components onto integrated circuits," Electronics, 38(8), April 1965.
2. SEMI Infographic "Why Moore Matters?" 2015.
3. Andrew Ng, Baidu, keynote in GPC, 2015.
4. IBM Deep Learning breakthrough", EE Times, Aug 2017.
5. J.Y. Chen, "GPU Technology Trends and Future Requirements," IEDM Plenary Paper 1.1, Baltimore, Maryland 2009.
6. J.Y. Chen, "Transform designs to chips, an end user point of view on mask making", SPIE keynote, Monterey CA, 2012.
7. P. Raghavan et, al., "5nm: Has the time for a device change come", ISQED, p. 275, 2016.

AUTHOR BIOGRAPHY

The author has been with NVIDIA as the Vice President of Technology & Foundry Management since 2004. Prior to that, he was the Vice President responsible for R&D and E-Beam Mask Making at TSMC, the Vice President of Operations and the Vice President of Business Development at WaferTech in Camas Washington. Earlier in his carrier, he was at Hughes Research Lab and Xerox Palo Alto Research Center working on e-beam lithography and CMOS. He has published more than 100 papers and a book on "*CMOS Devices and Technology for VLSI*" by Prentice Hall. Dr. Chen, an IEEE Fellow, has taught at the EE department in Santa Clara University and Chiao-tung University, Taiwan. He was a Howard Hughes Doctor Fellow and received a Ph.D. in EE and a Master in Executive Engineering Management, both from UCLA. He also holds a M.S. from University of Maine and a B.S. from National Taiwan University, both in E.E. Dr. Chen was a Technical Advisor for ITRI, Taiwan. He now serves on few boards in the IC industry.