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.

Figure 1. Forty years of microprocessor trend data showing the increase of transistor counts for GPU (white triangles) and CPU (blue dots).
Early Mask Making and the Founding of BACUS

By Jim Wiley, ASML US, Inc.

BEFORE BACUS

In late 1979, a small group of Silicon Valley mask makers got together for dinner and drinks at the Bacuus Inn on El Camino Real in Santa Clara, California and decided to organize a one day symposium focusing on common problems with chrome blanks: uniformity, flatness, and defects. By 1980, almost all Silicon Valley mask makers were using the same APT 914 chrome positive resist process (spin to develop, rinse, flood expose, develop to strip, rinse and dry) and the same Ultratech high pressure water mask cleaning process. Now, there was not much difference in the specific details of the mask making processes. The full adoption of projection printing for wafers created an urgent demand for zero defect chrome masks. Meanwhile, the KLA automatic defect detection systems were finding more defects than expected. So, same defects, defects. We needed a name for the organization and after a few more drinks, settled on “Bay Area Chrome Users Society,” BACUS for short. We invited chrome blank suppliers, all located outside of the San Francisco Bay Area to show up and present on an assigned topic. (The topics were usually the suppliers perceived “problem”) The first symposium was a success, so we repeated it on an annual basis. We made enough profit from the symposia to publish the “Bay Area Chrome Users Society” BACUS News. The full adoption of projection printing for wafers created an urgent demand for zero defect chrome masks. Meanwhile, the KLA automatic defect detection systems were finding more defects than expected. So, same process, same tools, same problem. The easiest solution? Blame the supplier! In this case the chrome blanket suppliers.

THE FOUNDING OF BACUS

In 1970, a Boeing 747 made its first commercial passenger trip, the first commercially available DRAM IC was introduced – and I joined the mask engineering community at National Semiconductor in Santa Clara, California. At that point, mask making was an analog process. Not even one computer was involved. Each mask geometry was defined at 400 times final size by highly skilled operators precisely scribing the edges into a thin layer of red plastic that was lightly adhered to a mylar substrate. After all edges were scribed, the now isolated thin layer of red plastic was manually lifted from the mylar with surgical tweezers. The most complex IC’s had about 1000 transistors, so it took several days to “cut and peel” the artwork for each of the 5 layers. A three micron thick resolution photographic emulsion coated on a 2” square thin glass plate was exposed to the artwork with a 40x reduction camera. After dunking in developer, short stop and fixer, the 10x reticle was dried by a dip in methanol. A 2.5” square emulsion “master” was exposed to the single-die 10x reticle using a David Mann photorepeater. Emulsion “sub-masters” and “working plates” were produced by contact printing. Manually eyeballed to the underlying patterns under a microscope, the 2” silicon wafers were contact printed too!

By 1974, the mask engineering community had embraced CAD (Computer Aided Digitizing at that point,) DRC’s, computer driven artwork generators, chrome photomasks and laser defect repair. Wafers were now printed using automatic alignment and proximity printing. Still, specific details of the mask making processes and home brew processing tools was trade secret.

From 1970 to 1980, Bay Area mask makers from more than ten companies held semi-periodic (monthly) informal meetings with invited speakers including tool suppliers, material suppliers and other mask making topics. But things were about to change.

In the late 70’s a few major innovations caused the mask industry to come together and establish the first BACUS symposium. These mask tool innovations created common processes and common problems. By 1980, almost all Silicon Valley mask makers were using the same APT 914 chrome positive resist process (spin to develop, rinse, flood expose, develop to strip, rinse and dry) and the same Ultratech high pressure water mask cleaning process. Now, there was not much difference in the specific details of the mask making processes. The full adoption of projection printing for wafers created an urgent demand for zero defect chrome masks. Meanwhile, the KLA automatic defect detection systems were finding more defects than expected. So, same process, same tools, same problem. The easiest solution? Blame the supplier! In this case the chrome blanket suppliers.

Four dozen years after joining the mask community in 1970, the Boeing 747 is leaving commercial passenger service and I am retiring from full-time employment. I have enjoyed every moment, but nothing has been more satisfying than what evolved after that dinner at the Bacuus Inn. After my retirement, you can find me via my SPIE Profile or via the various social networks.
Why 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 found 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 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.

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 platform.

- 815 mm\(^2\) 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
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 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. Then, in the 2012 SPIE keynote speech, 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 maskmaking 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 = Gate length + S/D CT width + 2xSpacer 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 TDDB (Time Dependent Dielectric Breakdown) related reliability problem. The following example il-
lectures 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: $\sigma_{total} = \sqrt{\sigma_{Gate\,CD}^2 + \sigma_{CT\,CD}^2 + \sigma_{Gate\,-CT\,Overlay}^2}$

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

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

**How Precise We Need in the Next Technology Node?**

It was proposed that CGP scales down to 32 - 42nm in 5 nm technology node\(^1\). Let’s assume CGP = 40nm for ~0.7x linear scaling. With $L_g = 15$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 $6.5 \sigma$ to ensure every transistor works, then the total budget for $1\sigma_{total} = 0.46$nm (3nm/6.5) to account for these 3 imprecisions: Gate CD, CT CD and Gate-CT misalignment. For Gate or CT CD, if controlled at $3\sigma = 0.5$nm and the misalignment controlled $3\sigma = 1.5$nm, then $\sigma_{total} = 0.51$nm $> 0.46$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.

**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.

**References**

3. Andrew Ng, Baidu keynote in GPC, 2015.
2018 BACUS Lifetime Achievement Award

This award is given in recognition of an individual who has, during their lifetime, through inventions or other activities over the course of their career, made a significant impact on the technology of mask making.

Presented to
Frank E. Abboud
SPIE Fellow, Intel Corp.

For his thirty+ years of significant contributions in all areas of photomask technology, more specifically for his leadership in the development of advanced e-beam pattern generation. He has been President and Vice President of BACUS and Chair of the BACUS Symposium.

Presented by
Brian Grenon
2018 BACUS Member

September 2018
Monterey, California

2018 BACUS Prize Award

Presented to
Tsuneo Terasawa
Key Technology Development Leader

Hidehito Watanabe
HVM Prototype Manager

Hiroki Muyai
HVM Development Leader

In recognition of their contributions in advancing the Mask Industry through Innovation: Concept, Development, and Commercialization of the EUV Actinic Blank Inspection (ABI) System

Presented by
Frank Abboud
2018 BACUS Vice-President

September 2018
Monterey, California
Industry Briefs

- **eBeam Initiative Surveys Report 27% Growth in Photomasks Delivered, Continued Confidence in EUV**

The eBeam Initiative announced the completion of its 7th annual eBeam Initiative perceptions survey. Nearly 40 companies including photomasks, electronic design automation (EDA), chip design, equipment, materials, manufacturing and research participated. The eBeam Initiative also completed its 4th annual mask makers’ survey with feedback from 10 captive and merchant photomask manufacturers.

Results from the mask makers’ survey indicate that mask output grew 27 percent compared to last year, while overall mask yields remained steady. Respondents were optimistic about the state of the photomask market, which grew 4.1 percent in 2017, and predicted continued growth at a compound annual growth rate (CAGR) of 4.1 percent or more between 2018 and 2020. Confidence and optimism in EUV lithography continue to remain high, while the perceived need for multi-beam mask writing (MBMW) continues to grow. Perceptions on the use of inverse lithography technology (ILT) at the leading edge also increased.

http://www.ebeam.org/home

- **GlobalFoundries Halts 7-Nanometer Chip Development**

IEEE Spectrum

In a major shift in strategy, GlobalFoundries is halting its development of next-generation chipmaking processes. It had planned to move to the so-called 7-nm node, then begin to use extreme-ultraviolet lithography (EUV) to make that process cheaper. From there, it planned to develop even more advanced lithography that would allow for 5- and 3-nanometer nodes. Despite having installed at least one EUV machine at its Fab 8 facility in Malta, N.Y., all those plans are now on indefinite hold, the company announced. The move leaves only three companies reaching for the highest rungs of the Moore's Law ladder: Intel, Samsung, and TSMC.

https://electroiq.com/insights-from-leading-edge/2018/08/iftle-395-and-then-there-were-3-ic-history-for-the-younger-generation/

- **Global Semiconductor Sales Increase 17.4% Year-to-Year in July**

Solid State Technology

The Semiconductor Industry Association (SIA), representing U.S. leadership in semiconductor manufacturing, design, and research, announced worldwide sales of semiconductors reached $39.5 billion for the month of July 2018, an increase of 17.4 percent compared to the July 2017 total of $33.6 billion. Global sales in July 2018 were 0.4 percent higher than the June 2018 total of $39.3 billion.

Regionally, sales increased compared to July 2017 in China (29.4 percent), the Americas (20.7 percent), Europe (11.7 percent), Japan (11.5 percent), and Asia Pacific/All Other (5.7 percent). Sales were up compared to last month in China (1.7 percent) and the Americas (0.4 percent), held flat in Asia Pacific/All Other, and decreased slightly in Japan (-0.1 percent), and Europe (-2.4 percent).

Join the premier professional organization for mask makers and mask users!

About the BACUS Group

Founded in 1980 by a group of chrome blank users wanting a single voice to interact with suppliers, BACUS has grown to become the largest and most widely known forum for the exchange of technical information of interest to photomask and reticle makers. BACUS joined SPIE in January of 1991 to expand the exchange of information with mask makers around the world.

The group sponsors an informative monthly meeting and newsletter, BACUS News. The BACUS annual Photomask Technology Symposium covers photomask technology, photomask processes, lithography, materials and resists, phase shift masks, inspection and repair, metrology, and quality and manufacturing management.

Individual Membership Benefits include:
- Subscription to BACUS News (monthly)
- Eligibility to hold office on BACUS Steering Committee

Corporate Membership Benefits include:
- 3-10 Voting Members in the SPIE General Membership, depending on tier level
- Subscription to BACUS News (monthly)
- One online SPIE Journal Subscription
- Listed as a Corporate Member in the BACUS Monthly Newsletter

2019

Photomask Japan
16-18 April 2019
PACIFICCO Yokohama
Yokohama, Japan
www.photomask-japan.org

SPIE Advanced Lithography
24-28 February 2019
San Jose Marriott and
San Jose Convention Center
San Jose, California, USA

SPIE is the international society for optics and photonics, an educational not-for-profit organization founded in 1955 to advance light-based science, engineering, and technology. The Society serves nearly 264,000 constituents from 166 countries, offering conferences and their published proceedings, continuing education, books, journals, and the SPIE Digital Library in support of interdisciplinary information exchange, professional networking, and patent precedent. SPIE provided more than $4 million in support of education and outreach programs in 2017. www.spie.org

You are invited to submit events of interest for this calendar. Please send to lindad@spie.org; alternatively, email or fax to SPIE.