IPSJ Transactions on System and LSI Design Methodology Vol.17 2–6 (Feb. 2024)

[DOI: 10.2197/ipsjtsldm.17.2]

Invited Paper

Slashing IC Power and Democratizing IC Access for the Digital Age

Tadahiro Kuroda[1,a)]

Received: November 17, 2023

Abstract: The continuous growth of the semiconductor industry, driven by the use of AI to fuse the physical with virtual space, requires drastic improvement in IC power efficiency, memory capacity, and memory bandwidth. This paper describes two solutions to slash IC power using 3D integration and specialized chips. In addition, it proposes a novel sliced bread memory stacking scheme that enables more than a 10-fold increase in the number of memory chips and hence capacity per stack, as well as memory bandwidth. Furthermore, it elaborates on an agile development platform that enables designing chips like writing software and prototyping chips in days. The ease of use of the platform and its 10-fold reduction of development time and costs are expected to democratize access to specialized chips to accelerate innovation by increasing the number of developers. It will also accelerate the transition of society to the digital age. Finally, the author will discuss the need for the global IC industry to move away from its reliance on competition to co-existence and co-evolution to sustain its growth.

Keywords: 3D integration, More than Moore, IC democratization, agile development, ASIC

1. Introduction

In the last 40 years, the semiconductor (IC) industry experienced an annualized growth rate of 9.5%, characterized by two waves of growth. During the first wave, semiconductors were used in appliances to deliver comfort in the physical space, creating a market size which was about 0.2% of the world's nominal GDP. The second wave started around 1995 on the heels of the launch of Windows 95, which is considered one of the most important products of the personal computer (PC) industry. Semiconductors are used in PCs and later also in smartphones to create the virtual space and make it portable, resulting in the IC market growing to 0.4% of the world's nominal GDP in size.

The market now looks to be on the verge of a third wave, propelling its size to 0.6% of the world's nominal GDP. Since the surge started during the COVID-19 pandemic, it may or may not be sustainable. If it is, then the industry is on its third wave of growth stimulated by the adoption of ICs to fuse the physical with virtual space using artificial intelligence (AI). Specifically, sensors are used to collect data from the physical space to create a digital twin, which is then analyzed using AI in the virtual space to determine what action to take through actuators, in order to deliver value to the users. Automated driving and smart factories represent two such examples.

The adoption of deep learning a decade ago delivered a quantum leap in AI performance and ushered in the new AI age, driving rapid advance in AI technology and exponential growth in its adoption. The introduction of generative AI, exemplified by ChatGPT, created another catalyst for explosive growth by making AI accessible to the masses. As a result, AI is becoming ubiquitous to improve various facets of our life. And since AI requires a lot of processing power and memory, it is great for the industry. However, this comes with a hefty cost, in terms of skyrocketing power consumption.

In the decade after the introduction of AlexNet in 2012, the amount of AI processing grew by more than four orders of magnitude [1]. However, around the same time, power efficiency of general-purpose processors used for such processing, including CPUs and GPUs, increased by only an order of magnitude [2]. The result is an explosive growth in power consumption of IT equipment. According to one estimate, using the total power generated today as the baseline, IT equipment power consumption alone is expected to grow to twice the baseline by 2030, and 200 times by 2050.

That was all before generative AI came to the scene. The upgrade from ChatGPT-2 to ChatGPT-3 alone increased the total amount of training compute needed by two orders of magnitude, and we are still at the infancy of generative AI. It is therefore imperative that the semiconductor industry develops technologies to drastically slash IC power consumption in order to sustain such exponential growth in the digital age without destroying the environment.

In addition to scaling, a More Moore solution, various solutions are being developed to cut IC power consumption by orders of magnitude, including a technology solution in the form of 3D IC integration, and a design solution using specialized instead of general-purpose chips.

2. 3D Integration Technology

The AI market is rapidly shifting from training to inference. In

[1] The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan

[a)] tsldm-submit@sig-sldm.org

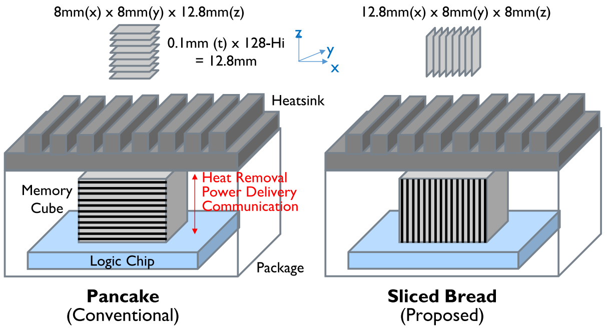


Fig. 1 Memory cube stacking schemes.

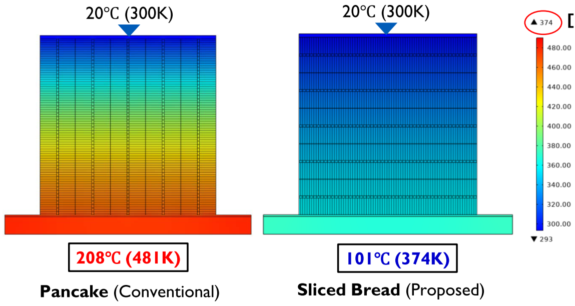
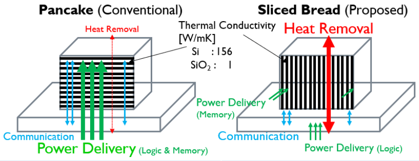


Fig. 2 Simulated thermal performance comparison.



	Pancake (Conventional)	Sliced Bread (Proposed)
Heat Removal	High resistance through SiO$_2$ (requires many thermal TSVs)	Good conductivity through Si
Power Delivery	Same path for all chips (requires large logic chip area)	Separate paths to individual chips (requires technology for delivering power from chip edge)
Communication	Multidrop Bus ⇒ large delay/power (can use entire chip area but requires ESD	TCI Point-to-Point ⇒ small delay/power (+ propagation across chip − ESD delay = lower cost overall)

Fig. 3 Merits of sliced bread vs. pancake scheme.

2023, the market size for inference is $26 billion vs. $6 billion for training. For generative AI inference, the performance bottleneck is in memory capacity and bandwidth. Generative AI inference requires real time response to each piece of input data. For a large model with 1 trillion parameters, real time inference requires access to 1 TB of weight data within 30 ms. In other words, a memory capacity of 1 TB and bandwidth of 30 TB/s are necessary. However, even Fugaku, one of the world's fastest supercomputers equipped with 30 GB of memory accessible at 1 TB/s, must increase both its memory capacity and bandwidth 30-fold to deliver such performance. Since no such memory solution exists today, ChatGPT3.5, which has 175 billion parameters in its model, relies on 128 units of the most advanced GPU from nVidia to perform each inference. At a cost of a few million dollars, this is not a cost-effective solution. Since the computing power required for inference per GPT3.5 token is only about 350 GFLOPs, one GPU can process thousands of tokens. Therefore, the solution has excess processing power while being constrained in memory capacity and bandwidth.

If there is a solution capable of performing each ChatGPT3.5 inference with just one processing node (package), it can reduce the cost to a few hundred thousand dollars, which is about 1/10 that of the nVidia solution.

One such memory solution is 3D-DRAM which integrates a large number of DRAM chips in 3D to create a memory cube. But heat removal is a big challenge. In Hot Chips 2023, Samsung reported difficulty in stacking more than 12 chips.

The Systems Design Lab (d.lab) at the University of Tokyo is developing the Hyperscale Cube 3D integration technology that achieves high heat removal efficiency to enable a large capacity while delivering a high bandwidth. As shown in **Fig. 1**, in our proposed scheme, the integrated memory chips are turned on their edge like a loaf of sliced bread before being stacked on top of the logic chip, as opposed to being laid flat like pancakes in the conventional stacking scheme. Since silicon oxide has low thermal conductivity, silicon oxide layers in the memory chips act like blankets in the pancake scheme which impede heat flow to the heatsink on top. Meanwhile, since silicon has 100 times better thermal conductivity, the silicon substrates of the memory chips, which are aligned with the direction of heat flow in the sliced bread scheme, enable efficient heat removal through the heatsink.

The drastic improvement in thermal performance is quantified in a thermal simulation of 50 chips stacked to create a cube with a total thermal density of 5 W/mm^2. The cube is placed on top of a logic chip with a thermal density of 1 W/mm^2. In addition, the chips are populated with thermal through-silicon vias (TSVs) in 5% of their area to facilitate heat removal. As shown in **Fig. 2**, the sliced bread scheme lowers the junction temperature at the logic chip by more than 100°C for the same ambient temperature.

Two other challenges in enabling a 3D memory cube are power delivery and data communication. In the conventional scheme, both must go through all intervening chips from the bottom up. As a result, they become more and more difficult as the number of chips increases. We are developing technology to deliver power from the sides of the cube through the package directly to the individual chips. Furthermore, since all chips in the cube stand on their edge on top of the logic chip, each can communicate directly with the logic chip. As a result, a point-to-point (P2P) instead of multidrop bus topology can be employed, resulting in lower delay and power. In addition, the adoption of the magnetically coupled ThruChip Interface (TCI) [3] for contactless communication simplifies assembly as well as ensures high reliability.

The merits of the sliced bread vs pancake scheme are summarized in **Fig. 3**.

We can thus use the sliced bread scheme to integrate hundreds of DRAM chips to create a Hyperscale D-Cube. In addition, we can integrate logic chips to create a Hyperscale L-Cube, and molecular dynamics processor chips to create a Hyperscale MD-Cube.

3. Agile-X Platform

The second solution to the IC power consumption problem shifts the chip design from using general-purpose circuits, as in a CPU or GPU, to using specialized circuits, as in an application-specific integrated circuit (ASIC). By definition, a general-purpose chip must support a wide range of applications, which requires it to have a lot of unnecessary circuits and wiring from the standpoint of any specific application. This results in not only wasted area, but also power consumption. A specialized chip can reduce power consumption by orders of magnitude. This is why leading system companies such as GAFA and Tesla are developing processors in-house to establish a competitive edge through low power consumption. However, specialized chips development is costly, both in terms of manpower and time. That is why it has been limited to only a few major corporations. Innovations in development methodology and tools are therefore necessary for wide adoption of this solution.

In fact, it was the design automation innovation driven by academia, in particular, the University of California at Berkeley, which ushered in the last ASIC age in the 1990s. The academia is therefore expected to take the lead once again in enabling the new ASIC age.

The University of Tokyo is creating a platform for agile development of specialized chips as shown in **Fig. 4**. It consists of an agile design platform which allows users to design ICs like writing software by using a silicon compiler. The users create the design by writing code in a high-level description language, which is then synthesized by the silicon compiler into a physical design ready for manufacturing. At the same time, an agile prototyping platform is being developed to shorten prototyping time by using a semi-custom manufacturing process. Using direct patterning and wafer splitting, wafers pre-made with general-purpose transistors are customized through wiring to produce specialized chips. The result is an order-of-magnitude reduction in the development time and costs of specialized ICs.

The reduced development time will enable not only fast time-to-market, but also tighter integration of hardware with software. Traditionally, ICs must be verified to be fully functional before being released into the market, and they are not revised afterwards due to the long development cycle. By contrast, software bugs can be patched after-the-fact. In addition, software can be upgraded on a continuous basis. By shortening the IC development cycle to be comparable to that of software, both can continue to be upgraded and stay tightly integrated. In fact, a system that tightly integrates newly developed hardware and software can be quickly released to win market share, and then undergo rapid iterations of improvement to create exponential growth in performance. This is especially important for the digital society which demands fast time-to-market and has rapidly evolving market needs and technology landscape.

In addition, the agile development platform will enable chip design without highly specialized skills. The population of IC designers is an order-of-magnitude smaller than that of software developers, due to the difficulty of IC design. Therefore, by enabling designing IC like writing software, the platform can potentially increase 10-fold the population of IC designers, which can in turn unleash the power of the collective brain.

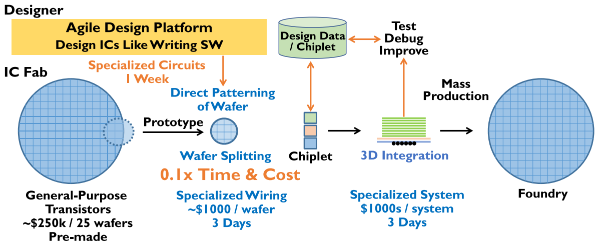

Designer
Agile Design Platform
Design ICs Like Writing SW
IC Fab
Specialized Circuits
1 Week
Design Data / Chiplet
Test
Debug
Improve
Direct Patterning of Wafer
Prototype
Mass Production
Wafer Splitting
Chiplet
3D Integration
0.1x Time & Cost
General-Purpose Transistors ~$250k / 25 wafers Pre-made
Specialized Wiring ~$1000 / wafer 3 Days
Specialized System $1000s / system 3 Days
Foundry

Fig. 4 Agile development platform.

Since innovation comes from people and is catalyzed by exchange and collision of different ideas, it can be greatly accelerated by creating a collective brain, or a large community of collaborating developers. The evidence is found in research on the evolution of fishing tools on multiple islands in the South Pacific. Researchers found a correlation between the number of variations in fishing tools and the island population [4]. Home sapiens are also believed to be able to create a larger variety of tools than Neanderthals despite having a smaller brain due to their ability to form large communities.

The potential of a collective brain in accelerating innovation for the digital society is the reason behind the recent movement in the industry to democratize access to ICs. By eliminating the need for specialized hardware design skills, the agile development platform will do just that. To speed up the process, the University of Tokyo launched the Agile-X Center for the Democratization of Innovative Semiconductor Technology in 2022. The goal is to use the agile development platform to bring the world's brains together to create a collective brain. Anyone with innovative ideas can design their own chips to realize their ideas, while interacting with each other to promote additional innovation.

In the words of physicist Richard Feynman, "there [wa]s plenty of room at the bottom" in the first half century of the IC, when the industry pursued microelectronics to generate tremendous growth. With integration level fast approaching a trillion transistors in a single package, there is now plenty of room at the top, for exploration of system ideas to take advantage of the massive IC processing power to enrich the digital society.

4. From Greedy to Green

The digital society is the result of a paradigm shift from the previous industrial and information society, as shown in **Fig. 5**. The industrial society is characterized by mass production and mass consumption of standardized products. The value chain consists of materials at the bottom as the resource, which are used to make parts such as ICs, which in turn are used to make products. Value is delivered by things through the consumption of products. Infrastructure consists of roads, harbors, and railways for moving materials, products, and everything in between. Since ICs are standardized and low in the value chain, the key performance metric is cost.

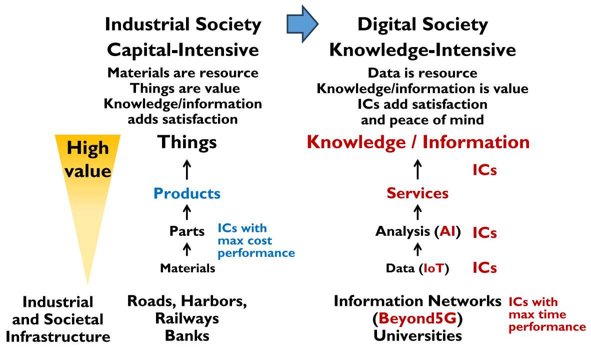

Industrial Society
Capital-Intensive
Materials are resource
Things are value
Knowledge/information adds satisfaction
Digital Society
Knowledge-Intensive
Data is resource
Knowledge/information is value
ICs add satisfaction and peace of mind
High value
Things
Products
Parts
ICs with max cost performance
Materials
Knowledge / Information
ICs
Services
Analysis (AI)
ICs
Data (IoT)
ICs
Industrial and Societal Infrastructure
Roads, Harbors, Railways Banks
Information Networks (Beyond5G) Universities
ICs with max time performance

Fig. 5 Paradigm shift from industrial to digital society.

By contrast, the digital society is characterized by the consumption of services. customized by AI for individual users. The value chain consists of data collected by ICs as the resource, which is analyzed by ICs using AI to create services. Value is found in the knowledge and information embedded in the services. Infrastructure consists of information networks for moving data. Since ICs are needed to extract knowledge and information for the services, they move up the value chain. At the same time, with the explosion of data and rapidly rising sophistication of AI, IC performance is increasingly limited by power, as discussed in the Introduction. In addition, as ICs are the core components of information networks, time-to-market has become critical. Unlike consumer products, infrastructure has a long replacement cycle. Therefore, once an IC is adopted, it will not be readily replaced even if better ICs become available later. Furthermore, since the digital society is knowledge-intensive, and knowledge is created by people for people, it is a human-centric society that benefits directly from more people participating in its growth.

In the industrial society, large capital investments are made to massively integrate more and more transistors, in order to both increase the performance of general-purpose chips and reduce their costs. Growth is driven by a greedy strategy since more is better. In the digital society, more and more massive data processing is needed to create high value-added services. Since power is increasingly limiting IC processing power, growth is driven by a green strategy to dramatically slash IC power consumption, including the shift to 3D integration of specialized ICs. The shift is also imperative in order to achieve carbon neutrality which has become urgent due to the burdens on the environment created by the industrial society.

In short, in parallel to society's shift from an industrial to a digital society, a paradigm shift is occurring in the IC industry from massive integration in general-purpose chips to 3D integration of specialized chips, and from a greedy to a green strategy.

5. Concluding Remarks

The dominant theory for the evolution of nature has been Darwin's theory of natural selection. However, if indeed evolution has been solely driven by survival of the fittest, it should not have resulted in such a rich diversity of life forms. A complementary theory of super-evolution is being proposed which adds co-existence and co-evolution to competition to explain nature's evolution [5].

There are 10 times more species of life forms today than during the Cretaceous Period (from approximately 145 to 66 million years ago). The difference seems to be attributable to the emergence of flowering plants. In the forest, before there were flowers, plants simply served as food for insects. However, since flowers rely on insects to spread their pollen, insects became important to the reproduction of and hence survival of plants, resulting in a mutually beneficial relationship. Flowers started to evolve to become more attractive to insects, such as changing into more vibrant colors. In turn, insects evolved so as to be able to reach flowers with increasing diversity in shapes and forms. Therefore, the evolution of one drives the evolution of the other. As a result, the relationship between plants and insects is not one of competition, but co-existence and co-evolution. Such is the theory of super-evolution of nature.

The co-evolution of plants and insects was further accelerated when the life cycle of flowers from pollination to fertilization was shortened from a year to a few hours, resulting in a rapid increase in diversity. This in turn accelerated the evolution and diversification of not only insects, but also mammals and primates further up the food chain. Therefore, the Earth today that is rich in life forms and full of colors has flowers to thank.

The IC industry has been characterized by intense competition, both between companies and between nations. The Japan-US trade war in the 1980s was a key reason for the decline of Japan's semiconductor industry. Japan's semiconductor industry has been stagnant for the last quarter century, while the world market has enjoyed continuously strong growth close to double digits. During this time, Japan has fallen so much behind in process technology that it is close to impossible for it to catch up. Meanwhile, Japan still maintains a competitive edge in materials and manufacturing equipment technologies. Since such technologies are critical for the development of 3D integration, there is an opportunity for Japan to collaborate with leaders in process technology, including the US, Europe, and Taiwan, in a complementary relationship. In fact, IC development and manufacturing has grown into such a complex collection of advanced technologies that it is challenging for any one country to be self-sufficient. This has been highlighted by the recent global chip shortage. The global IC industry therefore needs to rethink its reliance on competition for growth.

Taking hints from nature, co-existence and co-evolution may be what it needs. There are parallels between the IC industry and nature, with plants replaced by chips and insects by chip users. Flowers have contributed to the emergence of rich forests. The IC industry needs to find its "flowers" to trigger a rapid evolution into a rich IC forest with diverse value-added services. We are on the verge of the super-evolution of ICs, fueled by dramatic power reduction through 3D integration of specialized ICs, and more people innovating through the democratization of IC access.

Acknowledgments The author thanks Wai-Yeung Yip for valuable discussions.

References

[1] Gholami, A.: AI and Memory Wall, Medium (online), available from ⟨https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8⟩ (accessed 2023-10-16).

[2] Sun, Y. et al.: Summarizing CPU and GPU Design Trends with Product Data, Northeastern Comp Architecture Res Gr (NUCAR), arXiv (2019), available from ⟨https://arxiv.org/abs/1911.11313⟩ (accessed 2023-10-16).

[3] Kuroda, T. and Yip, W.-Y.: Wireless Interface Technologies for 3D IC and Module Integration, *Cambridge University Press* (2021).

[4] Kline, M.A. and Boyd, R.: Population size predicts technological complexity in Oceania, *Proc. Biological Sciences*, Vol.277, No.1693, pp.2559–2564 (Aug. 2010).

[5] Kuroda, T.: The Super-Evolution of Semiconductors, in Japanese, *Nikkei Business Publications, Inc.* (2023).

Tadahiro Kuroda received his Ph.D. degree in electrical engineering from the University of Tokyo, Tokyo, Japan, in 1999. In 1982, he joined Toshiba Corporation. From 1988 to 1990, he was a Visiting Scholar with the University of California, Berkeley, where he conducted research in the field of VLSI CAD. In 1990, he was back to Toshiba, and engaged in the research and development of BiCMOS ASICs, ECL ASICs, and high-speed low-power CMOS LSIs. He invented a Variable Threshold-voltage CMOS (VTCMOS) technology to control VTH through substrate bias, and applied it to a DCT core processor in 1995. He also developed a Variable Supply-voltage scheme to control VDD by an embedded DC-DC converter, and employed it to a microprocessor core and an MPEG-4 chip in 1997. He left Toshiba to join Keio University in 2000, and became a full professor in 2002. He was the Mackay Professor at the University of California, Berkeley, in 2007. He invented a ThruChip Interface (TCI) by using magnetic coupling for communications among stacked chips in 2008, and a Transmission Line Coupler (TLC) by using electromagnetic coupling for communications among stacked PCBs in 2012. He left Keio to join the University of Tokyo in 2019. He is the director of Systems Design Lab (d.lab) at the University of Tokyo, and the chairperson of Research Association for Advanced Systems (RaaS). He has published more than 500 papers, including 40 ISSCC papers, 30 VLSI Symposia papers, 19 CICC papers and 19 A-SSCC papers. He wrote 32 books/chapters and filed more than 200 patents. He is an IEEE Fellow, an IEICE Fellow, and a chair of Symposium on VLSI Technology and Circuits. He was an elected AdCom member of two terms. He was a recipient of the 2005 P&I Patent of the Year Award, the 2007 ASP-DAC Best Design Award, the 2009 IEICE Achievement Award, and the 2011 IEICE Society Award. He served as a Steering Committee Chair for A-SSCC, a Vice Chair for ASP-DAC, sub-committee chairs for A-SSCC, ICCAD, SSDM, and VLSI-DAT, and TPC members for ISSCC, Symposium on VLSI Circuits, CICC, DAC, ASPDAC, ISLPED, SSDM, ISQED, and other international conferences. He was a Distinguished Lecturer and a representative of Region 10 for the IEEE Solid-State Circuits Society.

(Invited by Editor-in-Chief: *Tohru Ishihara*)