To be published and presented as a plenary talk at the 79th Device Research Conference (DRC) June 2021

Optimizing Monolithic and Heterogeneous Integration to Create Intelligent-Grand-Scale-Integration for Smart MicroSystems

Nicky Lu

Etron Technology, Inc. and Invention & Collaboration Lab., Hsinchu Science-Based Industrial Park, Taiwan, ROC.

This paper presents Four perspectives on how the IC Industry based on Silicon technologies continues to grow and is creating a multi-decade Golden Age through synergistic inventions regarding optimizing Monolithic and Heterogeneous Integration *(MHI)* of various new materials, devices, circuits, chips, software, AI and systems *(MDCCSAS)*.

I. Expansion of MI & HI with Silicon-Centric Technologies Using Scaling Methodologies 1.0 to 4.0 (Fig. 1)

From the 1960s to 2010s, the minimum physical dimension on dice shrunk by 500 times down to 20nm and the number of devices increased by 500 million times. This was derived by Dennard's *Line/Voltage-Scaling Methodology*, which keeps both electric field and Power/Area per device unchanged but sharply reduces cost per function [1]. During the 2010s to the present, a Silicon Age 2.0 has accelerated tech-nodes down to 5nm by an *Area-Scaling Methodology*, which uses a 3D Tri-gate/FinFET structure.

Starting from the 2000s, more MicroSystems have used HI (Heterogeneous Integration) [2] for integrating more functions in a package/module with smaller form factor and better power/energy efficiency. This *Volumetric or Volume Scaling Methodology* [2] has ignited a Silicon Age 3.0. For example, in 2016, Apple's iPhone 7 adopted a HI solution manufactured holistically by TSMC using 3D Wafer-based System Integration (WSI) to raise another level of MicroSystem performance, noise immunity and energy efficiency derived from HI.

Looking forward, an exciting Silicon Age 4.0 leverages *a Skyscraper Scaling Methodology* that focuses on two major enhancements: (1) Combine optimally technology advancements of both MI (Monolithic Integration) & HI including smart chiplets to create a new Self-Smart MicroSystem (S-SmS) empowered by Nanometer-Si-ICs with embedded 3A's (Algorithm, Architecture & AI); and (2) Obtain more accumulated number of transistors on Monolithic die and/or on MicroSystem's HI substrate, featuring not only an mD (multi-Dimensional) integration with *scaled-down* form factor but also *scaling-up* productivity designed and driven by a *FV (Function times Value) Methodology* [2].

II. Refining Current MI through Inventions by Holistic MDCCSAS Optimization

"No exponential is forever, but 'forever' can be delayed," said Moore [1]. That inspired my dedicated endeavors to explore deeper on MI, in addition to HI: this has resulted in a conclusion that too rapid migrations from 14nm to 3nm may cause us to lose an ultimate optimization of MI which, by either forward- or backward-scaling compatibility of at least 2.5 generations [3], should achieve better PPACRF (Performance/Power/Area/Cost/ Reliability/Function). To give an example of my findings (Table 1): MHI stacks SOCs (some improved eSRAM caches) and DRAMs, which increases thermal dissipations to reduce DRAM retention time and badly hurts MicroSystem performance. So my MHI studies helped to introduce a new breed *Long-Retention (LR) DRAM* revealing that a DRAM retention-time spec could be >48ms at 125ºC, a 6X improvement, so as to minimize the loss of data bandwidth due to degraded refresh time in the DRAM+SOC MHI.

III. Entering Intelligent-Grand-Scale-Integration (IGSI) to Enable AI-driven Self-Smart MicroSystems and Networks, Which Proceeds to Super-Intelligent Systems

The coming 2~3nm CMOS MI can provide the number of devices on die approaching some trillion, and MHI will increase the number of devices for system chips having potential for another one million-fold increases (to Exa bits). By adding 3A design capabilities, this IGSI enables more S-SmS with immense intelligences to achieve today's supercomputer performance in small form factor with their energy efficiency capable of maneuvering huge data but not exceeding the thermodynamic limit ***kT·ln2*** [4] (~10^{-21} Joule per binary switch at T~300K). It's worthwhile to raise some academic questions: If T is lowered to 0.4K, could the remaining 0.13% (~0.4/300) of energy make information distinguishable? Another different route from MHI may be Quantum Computers using multiple Qubits at T~0.4K. Will both explorations help uncover some mysteries of Information Physics?

IV. Moving toward $1 Trillion Silicon Industry Nourished by Intelligence^N Applications

Figure 2 shows a sketch of future Semiconductor Economy Potential based on MHI. Besides evolutionary growth of current systems and applications, many intelligence-driven fields such as AI/ML, Cell/Gene Intelligence, Aging/Environmental Intelligence, Data Security/Privacy and Space/Earth Interaction Intelligence, etc., are fast growing. These PI (Pervasive Intelligences) enabled by MHI are widely applied to exponentially growing Intelligence^N Applications [2], and this will create another 'exponential S-SmS Economy Boom' enabled by IGSI in parallel to the world-changing 'Silicon Moore's Law and HI Economy'.

References: [1] Dennard, JSSC 1974; Moore, Electronics 1965, ISSCC 2003; Meindl, IEEE Proceedings 1995. [2] Lu, ISSCC 2004, A-SSCC 2016, ISSM 2020. [3] Private communications & Acknowledgments to C. Shiah, B. Rong, Dr. J.-Y. Chueh, W. Chen, LP Huang and W. Huang. [4] Von Neumann, Theory of Self-Reproducing Automata, 1966; Landauer, IBM JR&D, 1961; Meindl in Ref. 1.

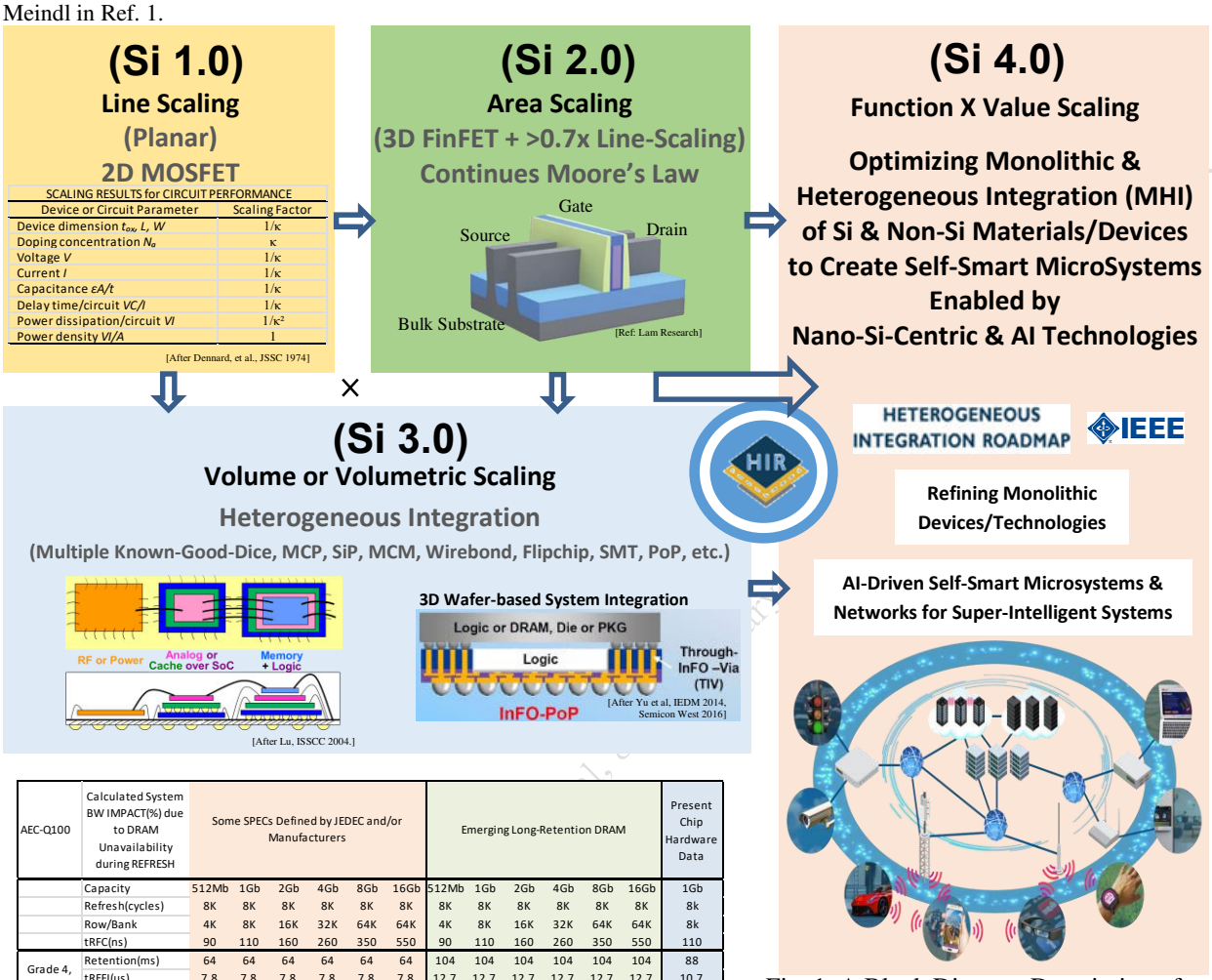

Fig. 1. A Block Diagram Description of Silicon Ages from 1.0 to 4.0.

Table 1. Reduce DRAM Unavailability IMPACT(%) for Optimizing HI Bandwidth between SOC & DRAM by Inventing Long-Retention DRAM Family. (IMPACT(%)=tRFC/tREFI)

AEC-Q100	Calculated System BW IMPACT(%) due to DRAM Unavailability during REFRESH	Some SPECs Defined by JEDEC and/or Manufacturers						Emerging Long-Retention DRAM						Present Chip Hardware Data
	Capacity	512Mb	1Gb	2Gb	4Gb	8Gb	16Gb	512Mb	1Gb	2Gb	4Gb	8Gb	16Gb	1Gb
	Refresh(cycles)	8K	8K	8K	8K	8K	8K	8K	8K	8K	8K	8K	8K	8k
	Row/Bank	4K	8K	16K	32K	64K	64K	4K	8K	16K	32K	64K	64K	8k
	tRFC(ns)	90	110	160	260	350	550	90	110	160	260	350	550	110
Grade 4, -85℃	Retention(ms)	64	64	64	64	64	64	104	104	104	104	104	104	88
	tREFI(μs)	7.8	7.8	7.8	7.8	7.8	7.8	12.7	12.7	12.7	12.7	12.7	12.7	10.7
	IMPACT(%)	1.15	1.41	2.05	3.33	4.48	7.04	0.71	0.87	1.26	2.05	2.76	4.33	1.02
Grade 3, 85-95℃	Retention(ms)	32	32	32	32	32	32	80	80	80	80	80	80	64
	tREFI(μs)	3.9	3.9	3.9	3.9	3.9	3.9	9.8	9.8	9.8	9.8	9.8	9.8	7.8
	IMPACT(%)	2.30	2.82	4.10	6.66	8.96	14.08	0.92	1.13	1.64	2.66	3.58	5.63	1.41
Grade 2, 95-105℃	Retention(ms)	16	16	16	16	16	16	64	64	64	64	64	64	48
	tREFI(μs)	1.95	1.95	1.95	1.95	1.95	1.95	7.8	7.8	7.8	7.8	7.8	7.8	5.9
	IMPACT(%)	4.61	5.63	8.19	13.31	17.92	28.16	1.15	1.41	2.05	3.33	4.48	7.04	1.88
Grade 1, 105-125℃	Retention(ms)	8	8	8	8	8	8	40	40	40	40	40	40	24
	tREFI(μs)	0.98	0.98	0.98	0.98	0.98	0.98	4.9	4.9	4.9	4.9	4.9	4.9	2.9
	IMPACT(%)	9.22	11.26	16.38	26.62	35.84	56.32	1.84	2.25	3.28	5.32	7.17	11.26	3.75

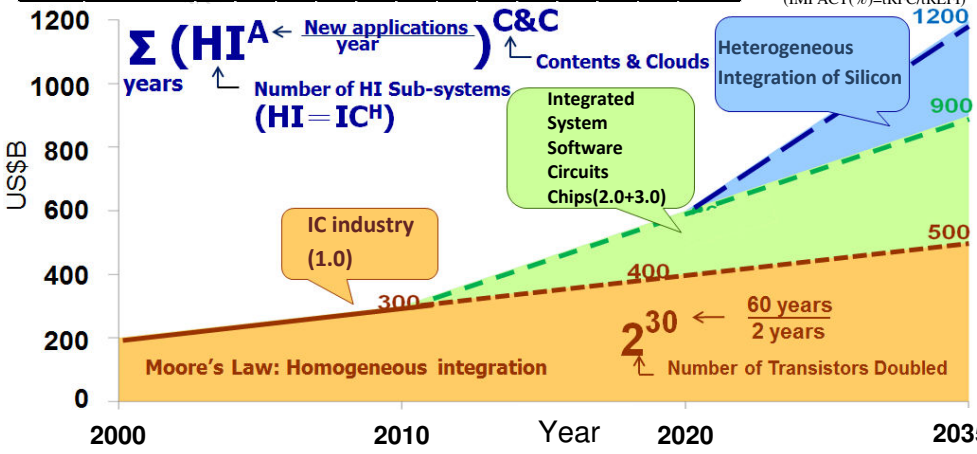

Fig. 2. A Sketch of Past, Present, and Future Semiconductor Economy in terms of Annual Revenues versus Years Based on MHI Technology Advancements from Silicon Ages 1.0 to 4.0.