A New Silicon Way: Generating Semiconductor -Intelligence Paradigm with a Virtual Moore's Law Economics and Heterogeneous Technologies

Nicky C.C. Lu
Etron Technology, Inc. and TSIA (Taiwan Semiconductor Industry Association)
Hsinchu Science-Based Industrial Park, Taiwan, R.O.C.

Abstract—The future of the silicon-based economy will not be as pessimistic as some commentators have argued, given their predictions of the end of Moore's Law Economy (ME) by the early 2020s. On the contrary, a Virtual Moore's Law Economy (VME) will develop and thrive, advancing innovation by a new Silicon Way of producing various application-driven Heterogeneous Integrated (HI) Nano-systems by optimization of physics, materials, devices, circuits/chips, software and systems to enable exciting applications for business growth. The semiconductor industry will enjoy sufficient financial returns from new application and system-product sales, even considering more expensive silicon investment. Such a technological approach based on a (Function × Value)-Scaling Down-Plus-Up Methodology, in addition to Linear- Scaling, Area-Scaling and Volumetric-Scaling Methodologies, can fundamentally change the way of thinking and execution toward optimizing coherently both technology definition and final system design with an holistic HIDAS (HI Design/Architecture/System) method. This will drive IC scaling to an effective 1-Nanometer Realm, stimulating a thriving silicon industry which can have at least 30 more years of growth toward a 1 trillion-dollar size.

I. Introduction: Silicon Age 1.0 (Si1.0)

One of the most unexpected and consequential developments in human history, generating huge impact on economic progress and civilization, is the explosive computing power created by the silicon industry, whose contributions to the modern information-technology world can be termed as launching the "Silicon Age". Nowadays a handheld smartphone with cloud computing infrastructure can perform so much previously unbelievable computing, communication and entertainment.

But when the transistor was first created in 1947, it was made with Germanium and used a point-contact device. After years of continuous inventions, R&D and accumulated manufacturing learnings, silicon technology using MOSFET finally succeeded and continues to grow today as the cornerstone and mainstream technology of ICs. In 1965 Gordon Moore projected that the number of transistors made on a silicon die would double every two years and that such silicon technology would create growing economic value in microelectronics and reward investors trusting this industry with significant economic returns on investments (ROI) [1]. Now, after celebrating the 50th anniversary of "ME"[2], the silicon industry has improved its technology from a minimum feature size of 30 micrometers to 16/14 nanometers, achieving 22-billion transistors onto a monolithic fabricated CMOS IC die, resulting in industry revenues >US$330B! But recently some commentators and analysts have been raising the topic, "The end of Moore's Law?!", challenging that ME could come to an end by 2025 [3]. But is the Silicon Age so close to the end, and what is our industry's own view (a BIG question)?

In 1974 Dennard published a MOSFET scaling theory [4] to elaborate on Moore's Law. The theory stated: let the device dimensions of oxide thickness, channel length and width be shrunk by a factor of κ, then the area of a transistor and the Power-Delay-Product (=switching energy) per circuit can be reduced by κ^2 and κ^3, respectively. So if κ is 1.4, then the area is shrunk by 2X, thus doubling the number of transistors per previous unit area while the switching energy per information-transfer is still acceptable. This Line-Scaling (L-S) methodology has proven very effective in steering IC growth and has substantiated ME using planar MOSFETs from 30-um down to 28-nm nodes over 20 generations. This stage of our industry can be categorized as one of Homogeneous Integration or Silicon Age 1.0 (Si1.0). But what is happening below 28-nm nodes and what will happen below 10-nm? Will the L-S methodology be sustainable? Will ME run out of steam due to the loss of advanced technology by scaling? (BIG questions).

II. Goals and Scope of This Work

Now, in the year 2016, marks a critical time to examine how Si1.0 has evolved, where silicon industry stands now, and what the future may hold to address the above BIG questions. Our industry must gather efforts to respond to questions on how much further can silicon go, which will definitely require new methodologies for technology enhancements and innovations.

This paper describes my personal views in attempting to answer these BIG questions by a systematic methodology study to explain why strong confidence remains in the silicon industry's future [5]. An analytical and semi-quantitative assessment has been developed for estimating the current/future fates and economic value of both the IC and Microelectronics industries, providing a new way of looking into what value can be created by ICs and if there is a sufficient ROI as IC trends toward the molecular/atomic scale which will demand huge investments in manufacturing. An insightful and executable methodology is derived: "Volumetric-Scaling (V-S) multiplied by Function/Value-Creation Scaling (FV-S)", where not only are geometries scaled down, but also values originated from or associated with the unit die area must be scaled up, thus achieving higher-valued nano-system products by using HI [6] of various dice and/or components inside either an intelligent package or module [7].

III. Si2.0 + Si3.0 Now: Area-Scaling, HI, V-S

Figure 1 presents data on the line-width featured by Gate Length versus major process Nodes. For the recent two generations of 22-nm and 14-nm,the gate lengths are longer than process-node lengths, though process nodes were still declared following the 0.7X scaling rule per generation. From Fig. 2, an Area Scaling (A-S; measured by Gate Pitch × Metal Pitch) methodology actually replaced L-S as a new Figure-of-Merit for logic technology advancement [8]. By further examining SRAM cell areas versus process nodes (Fig. 3), such an A-S methodology was further proven in that the area-scaling factor for SRAM cells does follow a ~0.5X scaling parameter pergeneration. Figure 4 sketches how silicon technology has currently scaled from 22/20-nm nodes to the 10-nm node: by using 3D Areal Density as a scaling guideline [8,9,10] the ME-like effect can be continued by several technology revolutions, thus continuing an Effective Moore's

Plenary Speech Published in 2016 IEEE Asian Solid-State Circuits Conference
(A-SSCC 2016), Nov 2016 at Toyama, Japan

Law Economy (EME). Starting from the 22/20-nm node, both A-S and EME have been replacing L-S and ME, respectively, effectively achieving positive economic growth, and the conventional planar-transistor structure has been replaced by a 3D Tri-Gate Transistor structure [8], which relaxes the stringent 0.7X L-S rule due to smartly utilizing the 3D over the planar surface so that each die area can still effectively sustain the increase of the number of transistors by 2X.The other successful example for A-S is the invention of the 3D NAND structure (Bit Cost Scaling) [10] which has been realized with an areal 3D stacked NAND cell structures, now reaching a stack of over 64 layers vertically [10]. This new silicon way results in a semiconductor industry revolution termed as Si2.0 by using either 3D transistors [8] or 3D cell structures [10].

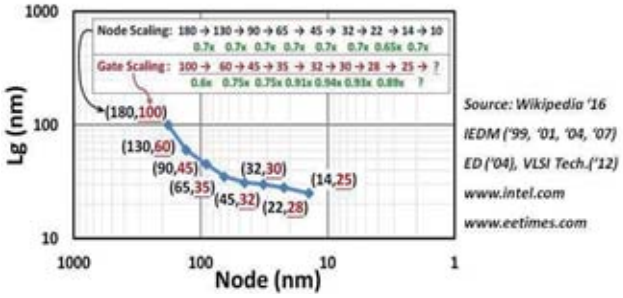


Fig.1. Gate Length vs. Process Node with Insets of Key Parameters

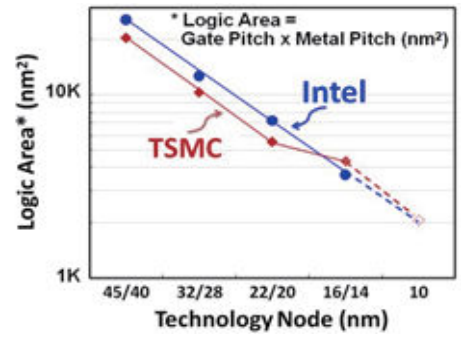


Fig.2. Area-Scaling Dimensions vs. Process Nodes in Logic Circuits [8]

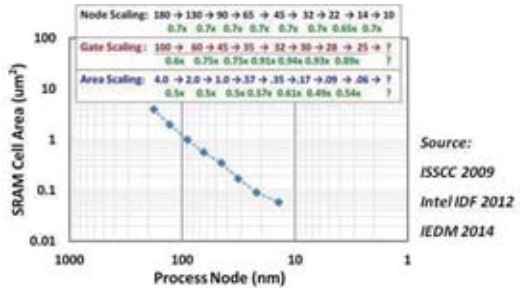


Fig.3. Area-Scaling Dimensions of SRAM Cells vs. Process Nodes

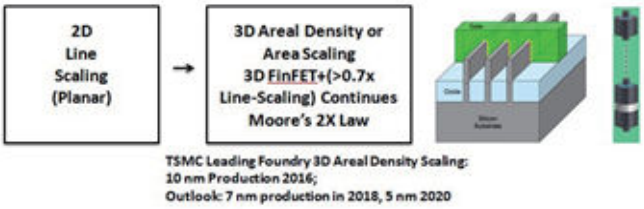


Fig.4. A Sketch of Silicon Age 2.0 due to an Area-Scaling Methodology [8, 9, 10] Which Creating an Effective Moore's Law Economy (EME)

Around the year 2000, another path became possible to grow capacity and function onto restricted silicon dice areas and into a constrained package volume (Fig. 5). The memory industry took the lead in achieving KGDM (Known-Good-Die Memory) technologies allowing 3D stacked dice in a package with proven enduring reliability, such as stacking RAM dice on top of Flash dice in an MCP (Multi-Chip Package) [6,7]. After KGDM could be manufactured, mass-market electronics like HDD (Hard Disk Drive), Digital Displays (TVs, monitors), and communication devices (networks, routers, mobile phones) started integrating multiple dice in a package [7], e.g. an innovative system-chip design/architecture can use an n-dimensional design (n>=3) to integrate mixed SOC die, DRAM dice, Flash die and analog dice located either vertically or horizontally (3D+2D aggregated stacks) within a BGA package as SiP (System in Package). This System-Chip (SC) technology has greatly enhanced the homogeneous integration of planar transistors to heterogeneous integration (HI) of multiple dice mixing different silicon technologies within a package. Such technology breakthrough plus architecture/design innovations optimizes the economic value per packaged chip [6,7] and marks the beginning of another silicon revolution by this V-S methodology [6,7,11] to enhance EME, which is termed as Si3.0.

Another way of generating higher value in a silicon micro-system is to optimize HI by integrating different non-silicon materials either in the transistor level or by using different components inside modules [7]. But since year 2000 it has been a great challenge to find the most-suitable technological micro-platform to accomplish the aforementioned HI architecture (HIA) in order to bridge to future nD-system-chip's needs which require versatile design in terms of flexibility and adaptability, low cost, high performance, high thermal conductivity, low standby current, reduced thickness of the dice-stack and easy-to-accommodate varieties of technologies/materials, etc. Recently an innovative technology called the 3D Wafer-based System Integration (WSI) has been proposed and realized by Yu [11], where unique advantages were created for implementing HI. WSI includes CoWoS (Chip-on-Wafer-on-Substrate) and InFO (Integrated Fan-Out) for high speed and high integration, respectively, to achieve another level of micro-system performance and stacking. An InFO platform through a RDL (Re-Distributed Layer) technology [12] can enable silicon dice to connect directly to the PCB level without another substrate layer. In addition, Yu's TIVs (Through Interconnect Via) can provide pillars to connect different dice or components using mixed vertical and horizontal interconnect technologies [12]. The HI advantages are amplified by this InFO to facilitate a turnkey completion of IC+assembly at once, which thus results in a thinner but higher performance smartphone (which have been shipping in volume recently) with higher clock frequency, wider bandwidth, higher routing density, lower power dissipation and lower profile/thickness [13]. This makes V-S more effective [6,7,9,11], further growing the EME and indicating that Si3.0 can continue to deliver ROI, not only for silicon manufacturers but also for system makers utilizing IC+3D-assembly technologies to optimize their system products'uniqueness and value [13].

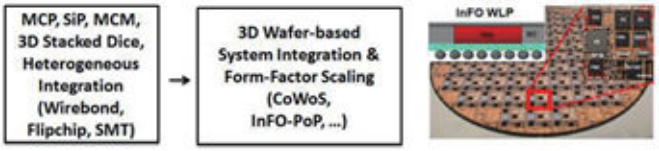


Fig.5. A Sketch of Silicon Age 3.0 due to a Creative Volumetric-Scaling Methodology [6,7] and a 3D Wafer-based System Integration Achieving Form-Factor Scaling [9,11] to Create EME

IV. Emerging Si4.0: New HIDAS Integration Method Creating the VME Growth

Economic demands, application pull and technology push are three constructive and compelling power forces that have continuously driven the Silicon Age. Both A-S and V-S are being used for pushing technologies from 22/20-nm down to 10-nm/7-nm nodes and will hence enlarge the EME. But recently the IC industry's revenues have reduced by -0.8% and-3% in the yearly comparison of 2015/2014 and 2016/2015, respectively. Besides the aforementioned BIG questions on

silicon technology's life, the other question that has been often raised is whether there are Killer Applications whose volumes are big enough to pull the silicon business upward after matured PC-Age and Cellphone-Age?

Due to technological breakthroughs in Si2.0 and 3.0, many significant innovative applications have recently emerged. Looking forward 5- to 15-years, we foresee the following applications to arise and demand more powerful ICs: smarter phones, clouds/servers/networks, higher performance computing, IoT and various wearable devices, VR/AR/M-R/RR(Real Reality), autonomous cars and smart vehicles, digital-grid power meters, biomedical and remote-health-care devices, drones, robots for Industry 4.0, etc. Looking further 15- to 30-years from today, there will be more novel applications demanding ICs strongly, such as AI (Artificial Intelligence), human-robots, One-World Language, Same-Day Global Traveling, Microbiology ICs strengthening human bodies, Outer-Space and Under-Sea Living Devices, etc . So the IC industry must prepare its way toward the Si4.0 to satisfy so many applications.

Figure 6 illustrates this new direction, that is, (3D transistors or 3D cells) times (3D WSI platform) integration [6,8,9,10,11] to achieve a silicon-centric nano-system with HI Design/Architecture/System (HIDAS), a new vertical-design method involving holistic final system-product design deep down to device-level. Further system-performance optimization must be designed from forming a HI Architecture (HIA) to study synthetically how to fully utilize down to advantages provided by the nanometer silicon-based technologies and vice versa. J. D. Meindl described a very useful hierarchical methodology to study scaling limits [14]. Those principles can be used to systematically define and design a complex HIDAS nano-system by holistic analyses plus syntheses from physics, materials, devices, circuits, and systems with software optimization focused on application perspectives. For instance, at the material and transistor level, perhaps some non-silicon materials can be integrated to enhance new device-structure performance such as using GeSn, III-V materials, or a surrounding gate made by carbon nanotube structures with different channel materials [15], etc. At the circuit level, perhaps inserting some super-performance nanometer materials/devices/circuits can provide special functions, either on analog, RF, digital or embedded/stacked memories or even spintronics devices like MRAMs, which will be heterogeneously fabricated among lower-cost silicon devices/circuits to optimize special needs with unique values demanded by a specific application. The InFO technology platform can be used as micro-assembly structures to host stacked or distributed heterogeneous dices, either RF/analog dice, DRAM chips, storage devices, sensors, MEMS, etc., by using either package modes or KGDMs or both mixed. The smart silicon center could consist of multiple CPU/GPU cores in either SOC forms or by various dice aggregated, embedded memories or passive components on some InFO as carriers which can be much smaller than today's PCB due to scalable silicon-carriers. Furthermore, in order to miniaturize and add higher value to the system, different non-silicon components such as imaging sensors, pressure sensors, MEMS, or micro lenses can be connected through TIVs [11] and more importantly, various improved miniaturized thermal-dissipation cooling sinks could be built on top or at bottom of InFO's to directly connect to outside heat sinks on external PCB carriers. So many exciting application possibilities due to nano-system integration, with more software/algorithm advancements, could be created by this HIDAS method.

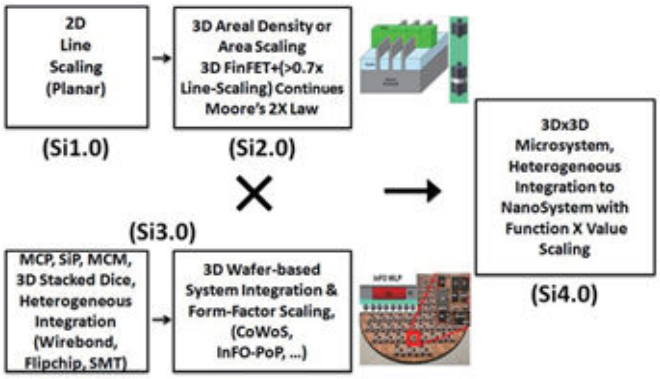


Fig.6. An Illustration on the Roadmap toward Future 3Dx3D [9] Heterogeneous Integrated Nano-system for Enlarging Silicon Values with a (Function × Value)-Scaling Methodology to Create VME (Silicon Age 4.0)

More importantly, as the value generated from HIDAS can be reflected over the silicon core area, an equivalent economic value can be created and enlarged based on performance advantages in addition to technology scaling, thus resulting in a Virtual Moore's Law Economy (VME)! There is no doubt that silicon feature sizes will be continuously scaled down, which can somewhat achieve ME-like scaling, but HIDAS provides more special functions and value to justify an advance process node, thus relaxing the Line-Scaling factor to >0.8~0.9X as long as the final HIDAS results in bandwidth increase, power reduction, higher efficiency of energy consumption for signal transfer, lower noises, etc. Therefore, for the growing VME era, the Line-Width scaling may not follow the 0.7X rule, but process nodes can still effectively follow the 0.7X nomenclature rule just as before: equivalently effective 5.0-nm, 3.5-nm, 2.8-nm, 2.0-nm, 1.4-nm to 1.0-nm nodes, as long as the high value provided by IC chip can justify the necessary ROI.

In reality, a nano-system-chip design can optimize both the power-delay-product reduction being derived from technology scaling-down and more values being created from the HIA scaling-up, which results in VME to justify huge investments. The HIA can further generate a higher value product by utilizing scaled silicon technologies with an essential total integration of non-silicon components such as MEMS, vision sensing, pressure and thermal sensing, or micro-fluid biological detection, etc., which enables many new applications that are not yet conceived by our current limited imagination.

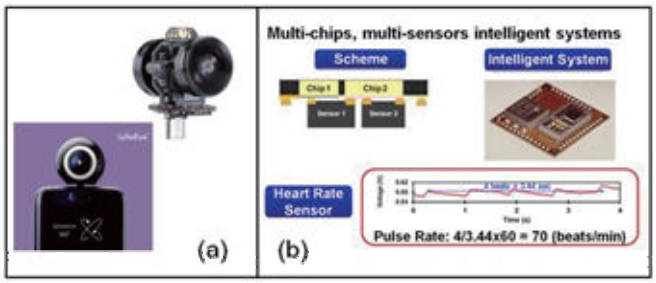
IV. Opportunities in the Semiconductor-Intelligence Paradigm

Silicon technologies have created and will continue to create intelligences at the heart of many existing, emerging and foreseeable applications. These applications in the next 5 to 30 years were briefly described in Section III. It is again emphasized that many more new IC products needed to provide intelligence and miniaturization for future even unforeseeable applications emerging due to the Si4.0 HIA technology.

Reference 13 shows how fusion of an advanced silicon technology into an InFO platform can create an innovative HI-system solution to achieve the currently most advanced smartphone ino mass market in 2016, which is the best precursor example to illuminate the technology inflection point from the Si3.0 to 4.0. Figure 7 illustrates two other precursor examples illuminating silicon's revolution toward HIA. Even though the technologies used were not in the most advanced node, these new products can realize novel applications that humans have not experienced until using a HIDAS-like design. Figure 7a shows a very small (3.0 cm diameter), light (18 grams) and cost-effective semiconductor platform enabling a spherical 360-degree video capture device which can be plugged into a mobile phone. This product transforms one's eyesight with limited FOV (Field of View), which for humans only covers

about 160-degree of looking ahead, to a new way of video recording all scenery around in a spherical 360 ° FOV.
Moreover, this "LyfieEye" expands humans' analog eyesight capability to digital video reality (RR/VR) which can thus be shared with the world through platforms like YouTube and Facebook and the spherical contents can be viewed under full control by either finger-/gyro- or VR-mode. Figure 7b shows a smart biomedical wearable device for measuring heart rates. The device can be so small and sensitive primarily due to using the InFO structure to integrate sensors with silicon chips [17].

In Fig. 8 the author attempts to sketch the potential growth of the silicon economy if the industry can be revolutionized from ME, through EME, to VME. If VME and Si4.0 does not happen, then the silicon industry may find it hard to attract investments due to recently reduced ROI because IC product prices continue to decline. Although recently more M&A deals happened, it is hardly effective to turnaround the unattractive ROI trend for such an invention-driven silicon industry. The best way for the silicon industry to avoid its economic slow-down is improving infrastructures like better education for more talents to join the industry and continuing innovations to increase values provided by HI of both application and silicon technology.



(a) a Spherical 360° Video Capture Micro-System as LyfieEye (Selfie+Life) [16]; (b) a Heart-Rate Measurement Wearable Micro-System Using the InFO Platform [12]

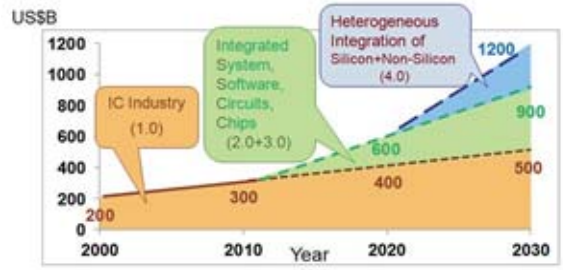


Fig.8. A Sketch of Future Silicon Economy Potential

V.Conclusions

In the past 5 decades, Moore's Law has provided a Rule-of-Thumb for the silicon IC economy that can be macroscopically sketched as "Just a Matter of a 2^x Scaling, which simply increases the number of transistors to create more function from a silicon die". The Si1.0 was thusly defined based on Line-Scaling Methodology by a rigid 0.7X scaling parameter from 30-um down to 28-nm for a total of 20 nodes. Technical productivity increased by 10^8 times from 300 transistors to $3x10^{10}$ transistors on a die, owing to three factors: the number of transistors increased by 10^6 times following 2^{20}, the die area increased by 10^2 times, and the wafer diameter increased from 3- to 12-inches. As a result, the industry's revenues grew over 250 billion dollars, and humans have benefited from so many revolutionary electronics applications such as mainframe computers, PCs, notebooks, mobile phones, 3G, 4K TV, digital music/-photos, the Internet, search engines, etc.

Now the Si2.0/3.0 age is featured by mixed Monolithic 3D transistors, 3D NAND or HI of multiple dice within a package to generate EME from 2011 to at least 2025, from 22/20-nm nodes scaled down to 10/7-nm. The underpinning technologies are either A-S or V-S or both. Productivity will increase 300 times from 10^{10} to effective 3×10^{12} transistors over one die basis
owing to these factors: the number of transistors increasing, a magnified effect due to 3D devices and dice-stacking constructions as well as an increase of die area. Humans started enjoying more intelligent electronics applications such as supercomputing/server/cloud, smart-phones, 4G, IOT/IOC, personal care, industry 4.0 (automation and smart robots), VR/AR/MR/RR, 3D scanning/printing, etc.

Starting from now on through 2025 to 2045 (as projected by the author), the Si4.0 starts using a value-added HI structure of composing both silicon Tera-scale ICs [14] and non-silicon components to construct a nano-system. This age emphasizes a (Function × Value) Scaling methodology with an effective 0.7X
nomenclature based on area scaling <1.0x plus more HI values to effectively achieve an equivalent 0.7X scaling. The major nodes show an equivalent shrinking from 5.0- to 1.0-nm node. Productivity can increase >1000 times owed to three factors: the number of transistors increasing, more accumulated die areas increasing due to multiple dice, and effective returns by value-added HIDAS design! The most critical challenge in HIDAS is how to reduce and effectively dissipate the heat generated by operational thermal power in such a small form-factor. The positive news is that our current energy needed for transferring information by an individual silicon switch is still tens of thousands of times larger than the Landauer's Limit [14,17,18]. The author projects that our silicon industry will continuously disclose how technologies in Si2.0+3.0+4.0 ages propel civilization into a new exciting era with many amazing applications, which thrives a Virtual Moore's Law Economy!

Acknowledgment

The author would like to express thanks to D. Yu, J. Sun, CY Lu, S. Pan, M. Ken, R. Crisp, E. Hsu, J. Lu and T. Lu for their valuable discussions and contributions.

[1] Moore, "No Exponential is Forever: But Forever Can Be Delayed", ISSCC, 2003.
[2] Holt, "Moore's law: A path going forward", ISSCC, 2016.
[3] L. S., "The End of Moore's Law," The Economist, April 19, 2015.
[4] Dennard, et al, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," IEEE JSSC, Oct. 1974.
[5] Lu, TSIA Report at WSC Seoul Statement, Korea, May 26, 2016. [6]Lu, "Emerging Technology and Business Solutions for System Chips," ISSCC, 2004.
[7] Lu, "Emerging Era of Heterogeneous Integration for System Chips: Technology and Business Solutions," FSA Magazine, June 2005; Products by Intel/Etron and Seagate/Agere/STM/Etron, respectively, 2000-2006.
[8] Bohr, "A 4.6GHz 162Mb SRAM Design in 22nm Tri-Gate CMOS Technology," ISSCC, 2012 ; Intel Investor Meeting, Jan. 14, 2014; Nenni, "TSMC Responds to Intel's 14nm Density Claim", Semiwiki, Jan. 21, 2014. [9] Sun, "Semiconductor Innovation into the Next Decade," A-SSCC 2014; TSMC Investor Conference, April 14, 2016.
[10] Y. Fukuzumi, et al, "BiCS (Bit Cost Scaling) NAND,"; IEDM, 2007; R. Smith and J. Jeong, "3D NAND Is the Leadership Technology for Server Storage," Flash Memory Summit, Aug. 2016.
[11] Yu, "Innovative Wafer-based Interconnect Enabling System Integration and Semiconductor Paradigm Shifts," IEEE IITC, 2013; "WSI for SiP," IEDM 2014.
[12] Yu, "WSI for Heterogeneous Integration," SEMICON West, July 2016; "WLSI Extends Si Processing and Supports Moore's Law," SEMICON, Taiwan, Aug. 2016.
[13] Wang, "New Packaging May Spur TSMC Growth," Taipei Times, Apr. 18, 2016; Report on iPhone7 Technologies, http://www.chip-works.com/about-chip works/overview/blog/apple-iphone-7-teardown
[14] Meindl et al., "Circuit Scaling Limits for Ultra-large-Scale Integration," ISSCC,1981 ; IEDM 1983; IEEE T-ED, Nov. 1984; IEEE JSSC, Oct. 2000.
[15] Desai, Javey, et al, "MoS2 Transistors with 1-nanometer gate lengths", Science, Oct. 7, 2016.
[16] LyfieEye, Product by eCapture Tech, Inc., 2016; www.eCapture-Tech.com.
[17] Landauer, "Irreversibility and Heat Generation in the Computer Process," IBM journal of R&D, 1961.
[18] Berut, et al., "Experimental Verification of Landauer's Principle Linking Information and Thermodynamics," Nature, March 2012.