A New Smart-MicroSystems Age Enabled by Heterogeneous Integration of Silicon-Centric and AI Technologies - My Personal View

Nicky Lu
Etron Technology, Inc. and Invention & Collaboration Lab.
Hsinchu Science-Based Industrial Park, Taiwan, R.O.C.

***Abstract*—After 60 years of development efforts since the 1960s to the current Giga-/Tera-Scale-Integration or System-on-a-Chip era [1-3], it is expected that Monolithic Silicon IC products using 2-nm CMOS devices will appear soon. The subsequent challenge is whether more novel device structures using heterogeneous materials and 3D-structures will be invented to realize manufacturable 1-nm ICs. On the other hand, through 20 years of efforts since 1999, many Heterogeneous Integration (HI) [4-6] products, each of which is composed of silicon and non-silicon materials/ dice/chiplets, diversified devices/circuits, innovative architectures and multi-dimensional arrangements of dice and other components inside either Chip-package or Module, have been increasingly emerging, especially recently benefiting from a strong driving force stimulated by the IEEE HI Roadmap unveiled in 2018 [5]. This paper presents an exciting, powerful and new Trend of Semiconductors, *Intelligent Grand Scale Integration (IGSI)*, which is optimally utilizing Mixed Integration of Monolithic and HI Technologies (Si-4.0 [6]) with embedded 3A's (Algorithm, Architecture and AI) Design-Intelligences. A key target of IGSI technologies is to drive much higher energy efficiency of managing electronic information for more-effective/ intelligent future systems with better performance, lower power, higher reliability and smaller form-factor than those of our current systems. One effective way as proposed is to network multiple *Self-Smart MicroSystems (S-SmS)* each of which is designed with 3A's to a complete system level which can handle huge data processing smartly in its own compact multi-dimensional form factor like in a versatile solid-state micro-universe which has abundant self-contained intelligent functions with maximized speed-power efficiency due to close proximity of electronic/photonic/ micro-mechanical operations. It is projected that in such an S-SmS each Joule (energy unit) be able to operate more than 10^20 devices per die per joule allowed by thermodynamics (on the other hand, its performance can reach over hundreds of thousands of TOPS - Tera Operations Per Second) inside and/or across these MicroSystems to complete the final system need. Then how powerful a future system can be by *networking enough S-SmS units* and furthermore how many unprecedented and unexpected applications will be unleashed! To use AI computing systems as an example, it is expected that S-SmS be quickly applied to AI's edge, device or wearable applications. Moreover, just like the experiences of migrating from a Mainframe computer to networked PC Servers, Data servers used in AI Clouds may use such a networked S-SmS architecture to build large systems in order to optimize the energy efficiency and heat dissipation. The trend equally adds values to system's transformation and optimization in Autonomous Car areas, Industrial 4.0 Factory areas, Telecommunication and Computing areas and so forth.**

Keywords—Heterogeneous Integration, Smart MicroSystem, Intelligent Grand Scale Integration, HI Roadmap, Silicon Age 4.0

I. Personal Experiences & Observations on How HI of Si-Centric Technologies Has Been Developed

After the invention of transistors in 1947, humanity's capability to pursue high performance computing was greatly improved by establishing computer systems using *Solid-State Microelectronics* which utilized multiple transistors and passive components on motherboards to replace vacuum tubes. Subsequently, integrated circuits (IC) were invented in the 1960s, and the progresses are best portrayed by Moore's Law (i.e. doubling numbers of transistors made on silicon dice in monolithic forms every two years) [1] from Small-Scale-Integration, LSI, VLSI to GSI (Giga-Scale Integration) from the 1970s to 2000s. Meindl [3] showed a FoM (Figure-of-Merit) of Digital Circuits by a straight line with exponentially increased number of transistors per die per joule versus years. References [7,8] have shown that an ultimate range of energies to allow bit-information transfer may be around 10^-21 Joule. This kind of energy efficiency of GSI (eg. based on some published data on 10-nm transistors [9] which shows around 10^-16 Joule per transistor On-Off) is still 5 to 6 orders-of-magnitude away from its limit. So with such GSI to TSI (Tera-Scale Integration), implementing a sub-system on a die, named as SOC (System on a Chip), has been set as an ultimate goal.

In another school of thought, however, due to different characteristics and FoMs among Digital Circuits/Devices (C/D), Analog C/D , Memory C/D, Radio-Frequency C/D implemented in silicon [4], a SiP (System in Package) concept has been proposed and realized as another integration alternative. Since 1999, a concept of different memory dice stacked in 3D within a package has been evolved and realized for commodity memory Combos [10]. An important breakthrough was to achieve *Known-Good-Die (KGD)* – that is, the IC Bare Die without package can be made through a Wafer-Level-Burn-In (WLBI) process to guarantee lifetime quality to be as good as standards for a Packaged Die as Chip [6]. Applying the KGD process to memories has resulted in a proven *Known-Good-Die-Memory (KGDM)* to be useful and popular [4]. Figure 1 shows the early stage of a concept explicitly disclosed and named firstly as *Heterogeneous Integration System Chip (HISC)*; its implementation was sketched in the early 2000s [4,6]. At that time the significance of giving a new name

HISC emphasized the new *architecture* of arranging various *KGDs* (executing various functions such as digital, analog, memory, RF, power and so on) to be all integrated simultaneously in a Multi-Dimensional (mD) geometry with a complicated and advanced *multi-layer substrate* structure(s) which integrate different passive components *within a package or a module*. Furthermore it is important to articulate beginning a new *Holistic Topside-down Design Approach* with systematic turnkey integration methodologies from dices to packaging, ie. *HIDAS (HI Design, Architecture and System)* [4,6]. All key design parameters like performance, power, thermal dissipation, energy efficiency, endurance, quality, yield, cost and so forth must be coherently and synergistically designed and executed to result in a well-functional mD-integrated turnkey system.

II. Expansion of Monolithic and Heterogeneous Integrations in Ages of Silicon-Centric Technologies from Versions 1.0 to 4.0

From 2000 to 2015, despite both monolithic and heterogeneous integrations were progressing well (even including a new encouraging thought was promoted as More than Moore [11]), many pessimistic views on the Silicon industry's future were espoused, with predictions such as "The end of Moore's law (by 2025)" [12]. This resulted in a low tide in stock market valuations and venture-capital investments in the Silicon industry. Fortunately, more inventions and innovations have been continuously pushing the IC industry to move ahead and upwards, leading to optimistic views on the Silicon industry and extremely positive forecasts for industry growth (one piece of evidence is that the market value of a silicon manufacturer with a pure-foundry business model – TSMC [Taiwan Semiconductor Manufacturing Co.] – rose to make TSMC the world's 11th largest publicly-traded company in November 2020).

Why and what has really happened for this major change can be understood fundamentally by analysing humanity's technology breakthroughs through a hierarchically organized historical-to-futuristic description of the Semiconductor industry development from years 2000 to the present and then future forecasts. Figure 2 (updated from [6]) illustrates a simplified Roadmap in sketched block diagrams of Silicon Ages 1.0 to 4.0, which was created for ease of description and explanation.

From the 1960s to 2010s, minimum dimensions of physical features on Silicon dies for devices evolved from several micrometers down to ~20 nanometers (shrunk by ~500 times), and the number of devices on a die increased by ~500 million times. The scaling methodology was derived by Dennard et al. [2] based on a *Linear shrink* of both the *Line-width* and *vertical dimensions* and thus gave a clear guideline of achieving Moore's Law direction of doubling the number of monolithic transistors every two years. The power-density (Power/Area) per transistor's function could be unchanged through many shrunk processing nodes/generations, and the cost of IC products with faster increasing functions was reduced (in addition to productivity gains by increasing wafer size from 2 to 12 inches in diameters) to create effective Microelectronics for system industry growth, colloquially termed as "Moore's law economy booms". This *Line Scaling* methodology on monolithic planar transistors is named as *Si1.0 (Silicon Age 1.0).*

During the 2010s to the present, a *Silicon Age 2.0* has migrated technology nodes to scale steadily from 20nm down to 5nm in current volume production. While the device scaling methodology could not depend only on Line-width shrinking, there has been a successful transition to an *Area Scaling* by using both a vertical Gate transistor structure, named as either Tri-Gate or FinFET Transistor structure and complicated interconnection technologies. The result is that the number of transistors on die can still be doubled by Monolithic integration every two years.

At the same time, more and more systems are using HI and SiP for dense integration; multiple dices are integrated in 3D and/or mD and fusion with more advanced substrate and packaging technologies (including many new and cost effective ways of doing Silicon Substrate as explored in 1990s). The clear purpose is to integrate more functions in a package or a module with smaller form factor and affordable power/energy efficiency. This methodology is named as *Volumetric* or *Volume Scaling* by going from monolithic planar technologies to 3D/mD technologies, catalyzing *Silicon Age 3.0*.

In 2016, a major piece of news stimulated further growth in both microsystem and semiconductor/IC industries: it is believed that Apple's iPhone 7 started using a HI solution shipped by TSMC [13]. Actually, an innovative technology called the 3D Wafer-based System Integration (WSI) was realized by Yu and his team in TSMC [14], where unique advantages were created for implementing HI such as CoWoS (Chip-on-Wafer-on-Substrate) and InFO (Integrated Fan-Out) for higher speed and integration to achieve another level of microsystem performance, noise immunity, and energy efficiency. This breakthrough has marked a critical milestone for Si3.0 and showing that "Integration of Si2.0 and 3.0 kills not only the worries about the end of Moore's law but also strengthens further economic growth through a Virtual Moore's Law enabled by HI [6, 15, 16]".

Around the same time, another big event during 2016 to 2018 was that IEEE agreed with and endorsed "Heterogeneous Integration Roadmap (HIR)" with strong participation of IEEE Electronics Packaging Society (EPS), SEMI, the IEEE Electron Devices Society (EDS), the IEEE Photonics Society and The American Society of Mechanical Engineers (ASME EPPD Division). This historical landmark should be credited to Bill Chen of ASE and Bill Bottoms of 3MTS, both of whom believe in that HI is important not only for Semiconductor industries, but also for Photonic, Mechanical, System, Aerodynamic/Space and many more industries, and that the best way to accelerate its growth technically and economically is to formulate an IEEE Roadmap activity like ITRS [17] to promote collaborations of all necessary participants and stimulate more innovations to be synchronized in coordinated movements. The afore-mentioned historical development of HI is summarized in Figure 3.

Looking forward to *Silicon Age 4.0*, as shown in Fig. 2 the major enhanced methodologies from Si2.0/Si3.0 into Si4.0 focus on two areas: the first one is to combine further technology advancements of both Monolithic and Heterogeneous Integrations to create a New *Smart MicroSystems (SmS)* Age enabled by Nanometer-Silicon-Centric and AI technologies, and the second one is not only continuing to drive higher number of devices/circuits into an *mD-integration Scaled-down form factor* occupying a smaller footprint but more importantly *Scaling-up more Function* x *Value* results for SmS. Currently more HI products are continuously moving into *Silicon Age 4.0* and present more diversified contents and progresses than those in Si3.0. The new development trend is to increase *more mixtures of silicon and non-silicon* materials/dice/chiplets, diversified devices/circuits, innovative architectures and multi-dimensional arrangements of dice and other components inside either Chip-package or Module. Moreover monolithic integration is continuing to explore new directions which are not only improving the devices/technologies used in Si2.0 but also may incorporate some new methodologies which can enhance the Area Scaling principle particularly on silicon dices via smart designs of C/D [18].

III. New Self-Smart Microsystems and/or Their Network for Intelligent System Integration

As Si4.0 is providing much more powerful technologies, the goal of using them is to achieve individual SmS whose energy efficiency of managing huge data/information is greatly improved while not exceeding the ultimate limit predicted by Von Neumann, Landauer and Meindl [3,7,8]. With the recent major progress of AI technologies and applications, it is highly expected that a new synchronized and synergistic technology named as ***ISA** (**I**ntelligent Design, Embedded **S**oftware and 3A's - **A**lgorithm, **A**rchitecture and **A**I)* can be used effectively to generate various *Self-Smart MicroSystems (S-SmS)* with superior performance and function while still under controllable energy/power/heat/noise boundaries.

Figure 4 shows an example of extending these concepts of Si4.0 and SmS to a big AI Eco-System consisting of numerous AI Edges and Devices (AED) connected to a centralized Cloud composed of many servers. It is expected that many AED will soon adopt these S-SmS enabled by Si4.0. Furthermore, it is expected that, like past experiences of migrating Mainframe computers to networked PC Servers, Data servers used in AI Clouds may need to use such a networked S-SmS architecture to build large systems in order to optimize energy efficiency and heat dissipation. By networking enough S-SmS units, future systems can be tremendously more powerful, resulting in many unprecedented and unexpected applications to be unleashed.

IV. Si4.0 Inducing Pervasive Intelligence^{N} Applications & Exponentially Growing S-SmS Economy for Another Semiconductor Boom

Figure 5 shows a Sketch of Future Semiconductor Economy (SE) Potential based on my expectation initially conceived in the year 2010 [6,19]. Developments since have only increased my confidence in this projection. If merely based on Si1.0 growth, then SE may reach \$500B by mid-2030s. The formula of growing number of transistors is [16]:

$$\underset{\uparrow\ \text{Number of Transistors}}{2}^{\,30 \longleftarrow \frac{\text{60 years}}{\text{2 years}}} \tag{1}$$

With the combined forces of Si2.0 and Si3.0, SE may reach \$900B. By adding all growth potential from Si1.0/2.0/3.0/4.0, SE may reach \$1200B. An interesting formula is proposed here for this growth projection [16]:

$$\sum_{\text{years}} \Big(\underset{\uparrow\ \text{Number of HI Sub-systems (HI = IC}^{\text{H}})}{\text{HI}}^{A \longleftarrow \frac{\text{New applications}}{\text{year}}} \Big)^{\underset{\uparrow\ \text{Contents \& Clouds}}{\text{C\&C}}} \tag{2}$$

Further updates on recent progresses in HI technology development will be presented [5]. More presentations will reveal that both of the aforementioned enhanced system-capabilities and energy-efficiency created by prevailing Smart MicroSystems of using IC/HI/IGI technologies are accelerating the growth of many intelligence-driven fields, such as AI, Cell/Gene Intelligence, Aging/Environment Intelligences, Data Security/Privacy and Space & Earth Interaction Intelligence, etc. These *Pervasive Intelligences (PI)* technologies [17] have been widely applied to recently exponentially growing *Intelligence^{N}* Applications, which is confidently believed will stimulate another exponential *SmS-Economy boom* enabled by IGSI succeeding semiconductor's world-changing *Moore's Law Economy*.

Acknowledgment

The author would like to express thanks to Dr. Bill Chen, Dr. Bill Bottoms, Dr. Justin J.-Y. Chueh, Dr. C. Y. Lu, and Justine Tsai for their contributions.

References

[1] Moore, Electronics, 1965; ISSCC 2003 Plenary Talk.

[2] Dennard, JSSC, 1974.

[3] Meindl, ISSCC 1993 Plenary Talk; Proceedings of the IEEE 1995; JSSC, 2000.

[4] Lu, ISSCC 2004 Plenary Talk; FSA Magazine 2005.

[5] Bill Chen and Bill Bottoms, HIR https://eps.ieee.org/technology/heterogeneous-integration-roadmap.html.

[6] Lu, A-SSCC 2006 Plenary Talk.

[7] Von Neumann, Theory of Self-Reproducing Automata, 1966.

[8] Landauer, IBM journal of R&D, 1961; Berut, Nature, 2012.

[9] Auth, IEDM, 2017.

[10] Memory KGDM, Etron's Product Realization, Documents and Specifications, and Other KGDM Manufacturers' Products (which appeared firstly, to this author's knowledge, around 1999-2000s).

[11] Zhang, ICEPT, 2005.

[12] L.S., The Economist, April 19th, 2015.

[13] Report on iPhone 7 Technologies, https://www.techinsights.com/blog/apple-iphone-7-teardown.

[14] Yu, IEEE IITC, 2013; IEDM, 2014; SEMICON West, 2016; SEMICON Taiwan, 2016.

[15] Yoshida, https://www.eetimes.com/after-moores-law-what ; https://www.eetimes.com/chip-industry-maps-heterogeneous-integration; Patterson, https://www.eenewseurope.com/news/anticipating-more-virtual-moores-law.

[16] Lu, Keynote speech at 2018 MIT Technology Review / DeepTech Semiconductor Industry Trend, Beijing; Keynote speech at 2018 EE Times Double Summits, Shenzhen.

[17] ITRS: http://www.itrs2.net/.

[18] Private Communications with William Chen.

[19] Lu, Keynote Speech of 2010 GSA Memory+ Conference, San Jose, CA, USA.

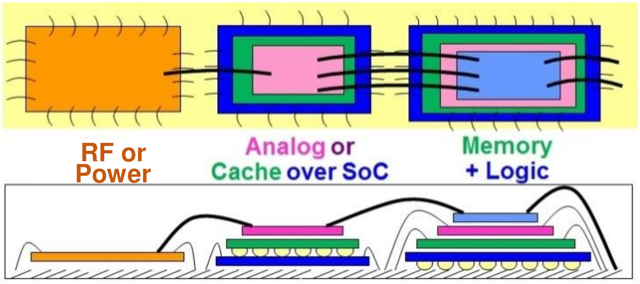

Fig. 1. A Schematic illustration of an mD HISC (Multi-Dimensional Heterogeneous Integration System Chip) Structure: Cross Section View [4].

Fig. 4. An AI Cloud Composed of Servers and Many AI Edge/Devices.

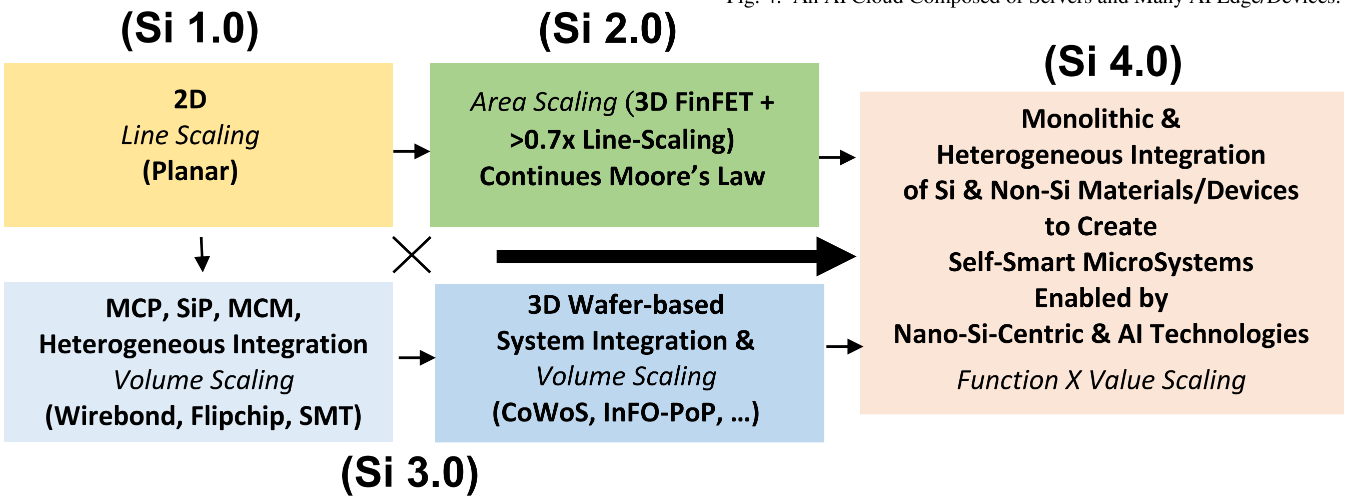
(Si 1.0)
2D
Line Scaling
(Planar)
(Si 2.0)
Area Scaling (3D FinFET + >0.7x Line-Scaling) Continues Moore's Law
(Si 4.0)
Monolithic & Heterogeneous Integration of Si & Non-Si Materials/Devices to Create Self-Smart MicroSystems Enabled by Nano-Si-Centric & AI Technologies
Function X Value Scaling
MCP, SiP, MCM, Heterogeneous Integration
Volume Scaling
(Wirebond, Flipchip, SMT)
3D Wafer-based System Integration &
Volume Scaling
(CoWoS, InFO-PoP, …)
(Si 3.0)

Fig. 2. A Block Diagram Description of Silicon Ages from 1.0 to 4.0 for Past 60 Years.

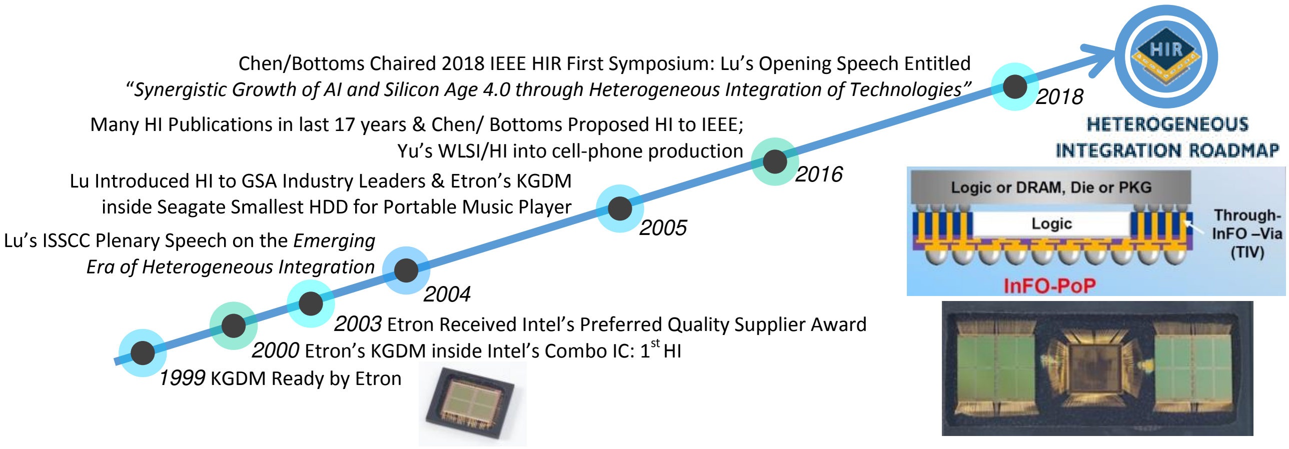
Chen/Bottoms Chaired 2018 IEEE HIR First Symposium: Lu's Opening Speech Entitled
"Synergistic Growth of AI and Silicon Age 4.0 through Heterogeneous Integration of Technologies"
2018
Many HI Publications in last 17 years & Chen/ Bottoms Proposed HI to IEEE;
Yu's WLSI/HI into cell-phone production
2016
Lu Introduced HI to GSA Industry Leaders & Etron's KGDM
inside Seagate Smallest HDD for Portable Music Player
2005
Lu's ISSCC Plenary Speech on the Emerging
Era of Heterogeneous Integration
2004
2003 Etron Received Intel's Preferred Quality Supplier Award
2000 Etron's KGDM inside Intel's Combo IC: 1st HI
1999 KGDM Ready by Etron
HIR
HETEROGENEOUS INTEGRATION ROADMAP
Logic or DRAM, Die or PKG
Logic
Through-InFO –Via (TIV)
InFO-PoP

Fig. 3. A Brief Historical Summary of HI Development (My Personal Experiences).

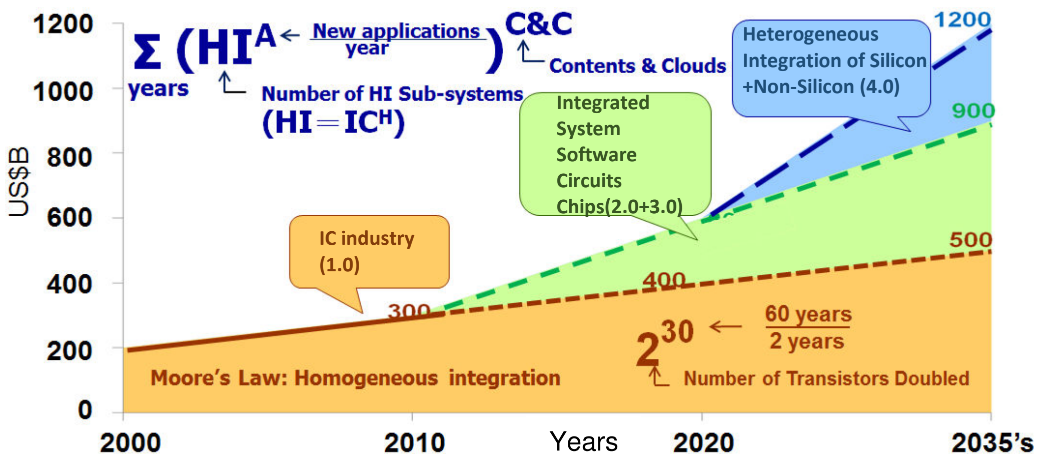

Fig. 5. A Sketch of Past, Present, and Future Semiconductor Economy in term of Revenues versus Years.