1.3 The Art of Scaling: Distributed and Connected to Sustain the Golden Age of Computation

Inyup Kang

President, Samsung Electronics, Hwaseong, Korea

1. Introduction

The history of computers was driven by maximizing computing power. In just about thirty years, CPU performance has skyrocketed by 5,000 times [1], taking computers from the level of an almost mechanical calculator to a gadget to run 3D games. This boost was mainly due to the exponential development of process technology under Moore's Law [2]. While this progress also applies to the mobile world, current CPUs in mobile phones are merely comparable to the nervous systems of a jellyfish, not even capable of replicating connectomes [3] of a honeybee. Besides, Moore's Law and Dennard Scaling are already facing their limitations, if not declared is over, widening the gap between endless human imagination and physical computing levels. Despite a number of technical suggestions to tackle this barrier, few papers have managed to capture the essence of its causes and consequences, or to understand that increase in computing power comes from the combined interaction of design and economic factors as well as process technology improvements. This paper takes a general approach to analyze how each area has come to its limitations and review possible suggestions for innovation, to find if there are certain common and general principles under these innovations. Focusing on the law of change hidden in general evolution, we demonstrate how these laws influence the process, design, and economic factors facing limitation, and try to derive the general principles of integration, distribution, and connection.

2 Evolution

2.1 External Force and Internal Optimization

When an external force is applied to a system, the system undergoes a change or evolution reacting to the external force through an internal optimization process. For example, the thermodynamic system evolves to minimize Gibbs free energy under the given external condition of pressure and temperature. Living organisms evolve to maximize their adaptability for survival against changes of natural environment, while the industrial world evolves to maximize profit under given technology and market environment. Although there may be some variations in details, many systems such as cities, living organisms, or even the universe basically try to optimize themselves under given constraints, that is, minimize the cost function.

2.2. Quantitative Changes into Qualitative Innovations

The evolution begins with a quantitative change, but soon faces limitations under its boundary condition. The temperature of water can only reach up to 100°C, the maximum natural size of a group of mammals cannot exceed the Dunbar's number [4], and the size of a sailing ship at around the 15th century that seemed to expand endlessly could not go on forever. When these limitations finally light the fuse on the qualitative change, it can jump up to the next curve of evolution. The boiling water turns into vapors, the number of individuals in a group stretched out with the Cognitive Revolution [5], and the sailing ships were quickly replaced by steamships in the 19th century. These qualitative changes are referred to as 'phase transition', 'paradigm shift', 'disruptive innovation', or in other terms according to the fields of variety, but essentially they carry the same idea.

This qualitative change can take place in two forms. It may come with certain unprecedented breakthrough replacing the traditional one, or with self-evolution into something more complicated. The latter approach can be seen in numerous cases, such as in cities or in the brain of animals. For example, a city first concentrates on expanding its boundaries and population to achieve maximized efficiency. However, as it soon encounters quantitative limitations, it evolves into a metropolitan system that bears multiple satellite cities with specific function, closely connected to each other. Similar metric can be found in the evolution of the brain, for which a simple comparison between the brain of insects and that of humans would suffice.

We will propose "distributed and connected" as key characteristics of this form of innovation, which will be the core foundation of all the discussions set forth herein. As we see it, these principles have also shaped the current evolution in the semiconductor area as well.

3 The Evolution of Semiconductor Chips

3.1 External Force and Internal Optimization

The driving force that pushed forward the semiconductor industry has been, of course, the profit gain. To maximize profit, one should have profound understanding of customer needs, which basically can be summarized as getting 'higher-performance chip within less budget.' Thus, the semiconductor chip-makers have been pursuing low cost and high performance to maximize their profits while satisfying customer needs. In pursuit of this target, the technology changed from simple vacuum tubes and transistors, finally to current Si technology, because of its exponential nature, of Moore's Law. However, Moore's Law is only one of the rules that explain the progress of subsequent technology development. This paper will take a higher-level approach by starting from the principle of low cost and high performance to capture this flow of evolution in the semiconductor industry, and show how it naturally bears Moore's Law on the way.

Taken together, we first define the cost-performance ratio (CPR) as below, a metric to capture the evolution of semiconductor chips

$$CPR = \frac{S}{C\$} \propto S\frac{Y}{A \cdot W\$} \quad (1)$$

The CPR of a chip is proportional to its performance/speedup (S) and inversely proportional to the cost of the chip ($C\$$). The performance is defined at a given power budget and must satisfy a given minimum performance requirement. The cost itself is proportional to the wafer cost ($W\$$), die area (A) at a given technology node and inversely proportional to die yield (Y). The yield itself can be represented as $Y = exp(-D_0 \times A)$ according to Poisson's equation, where D_0 is defect density. The speedup (S) itself can be further broken down into components contributed by architectural change versus those by process technology change as shown below.

$$S = S_d S_i = \frac{S_d}{n_{tr}} S_i n_{tr} \propto S_a \frac{S_i}{a} A \quad (2)$$

S_a is speedup due to design change while S_i is intrinsic speedup due to process technology change. The latter is typically represented by gain in ring oscillator speed. S_a is S_a normalized by the number of transistors (n_{tr}) and can be considered as speedup due to architectural change independent of process technology change. a is the area per transistor which is A/n_{tr}. Substituting (2) in (1) gives the following equation which is a product of three terms. Next, we will show how each of these terms has evolved in the past and how they can be changed in the future to sustain growth in the CPR:

$$CPR \propto S_a \times \frac{S_i}{a} \times \frac{Y}{W\$} \quad (3)$$

3.2. Quantitative Changes

We can illustrate how the CPR has increased in mobile SoCs used in smartphones. As for S_a and S_a, we have seen about +15% growth per year for the former, but –10% degradation per year for the latter (Figure 1.3.1) concerning mobile CPUs. The difference between the two shows that, were it not for the area scaling, the performance growth of CPU would have ceased in 2017.

The other two terms are related to process technology. Trend in S_i/a is represented by the combination of Moore's Law and Dennard Scaling. Here, a has changed by ~0.5× every two years in the past [1]. S_i, on the other hand, has scaled by >1.2× per year [6]. Combining the two gives S_i/a scaling of >1.7× per year, which has been the main thrust that allowed semiconductors to reach its present ubiquitous status. However, Moore's Law and Dennard Scaling have been weakened significantly so that recent S_i/a falls down to 1.4× per year (Figure 1.3.2). Y has remained almost constant because the chip size itself remained almost constant to meet the minimum performance requirement. $Y/W\$$ has almost stayed flat up until 2018 (Figure 1.3.3) but has since fallen sharply due to the increasing number of mask layers, number of process steps, EUV (Extreme Ultraviolet lithography) tool cost, and so on.

Combining the three trends, it is not hard to see that the quantitative scaling is getting saturated (Figure 1.3.4). Beyond this point, it is no longer the norm to expect a better performing chip at the same cost by just waiting a year. Instead, we shall expect unsustainable inflation.

3.3. Qualitative Innovation: Distributed and Connected

Clearly, all three terms in CPR are reaching the limit of quantitative change. The performance upgrade of CPU by pure architectural innovation has already been saturated and the power of Moore's Law, weakened yet still running, has been canceled out by high wafer pricing for 7nm process and beyond.

In this sense, the sustainable growth in CPR can only be possible under qualitative innovation in each area that faces limitations. Among various candidates introduced, ranging from little to substantial, the innovation of S_i/a seems to be the most critical factor concerning that the exponential characteristic of CPR is basically derived from

2022 IEEE International Solid-State Circuits Conference (ISSCC) | 978-1-6654-2800-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISSCC42614.2022.9731536

Moore's Law. Other candidates include, structural change (Gate-All-Around FET, Nanowire, and so on.), material change (Ferro-electric, 2D material, and so on.), neuromorphic device, quantum computer, and so on. Except for the structural change that may be implemented in 3nm node, other technologies seem less likely to replace Si technology in a short amount of time. Moreover, there are currently only few limited research on post-EUV.

This paper will focus more on innovations in design regarding S_a and $\frac{Y}{W\$}$.

3.3.1. Domain Specific Architecture (DSA): Distributed Function and Computation

As mentioned, S_a (that is, performance per given number of transistors) continues to decrease for general-purpose compute architectures. This is mostly because these processors spend more transistors trying to make the computation fast, rather than the compute operation itself [7]. To cope with these challenges, mobile SoCs underwent several changes in structure, from having a single large core to multiple smaller cores, quickly moving up to a heterogeneous model with a few big cores and a large number of smaller, efficient cores (for example, ARM big.LITTLE), and to finer cluster models with a variety of big/mid/little cores.

However, eventually we will have run out of the benefit for adding more cores, as thread-level parallelism is usually limited to a certain degree. Clearly, increasing the number of cores — quantitative growth — is becoming infeasible, and specialization is the way to go. This trend is similar to the way many things evolve with human organizations, cities, or brains. At first, a few general-purpose elements (be it people, buildings, or cells) grow in their size and when numbers start hitting its limitation, each of those general-purpose elements evolve into specialized elements, and they continue to grow until they themselves hit their own limitation. The next step of qualitative change comes from adding different types of cores — which are Domain-Specific Architectures (DSAs).

This trend can be observed from Figure 1.3.5, which shows two die photos from flagship mobile SoCs 12 years apart — C110 [8] for 2010 flagship smartphones, and Exynos 2200 [9] for 2022. We can see that the die size has stayed relatively the same (if the cellular baseband is excluded from the latter), while the area per CPU core has decreased by at least 1/6. Instead, the SoC added a diverse number of processing cores benefiting from all the additional transistors — which includes DSPs for general signal processing, NPUs for deep neural nets and ISPs for image processing for cameras, and so on.

One field that is rapidly improving is NPU, the neural processing unit. Recent advances on deep learning have created a wide variety of applications that required real-time processing of abundant data from sensors. For example, a common technique used for mobile camera applications is to use AI for detecting the scene so that we can direct the ISP to generate and apply scene-specific camera settings tailored to the subject matter [10].

These AI techniques typically involves deep convolution networks [11]. Crunching some practical numbers, with 300mW budget for NPU @ 30fps, we are only given a 10 mJ energy budget per inference. However, assuming 4.2GFLOPs per inference [12] and 70 pJ per instruction at 45 nm process technology [13], we need at least 300mJ of energy, which is at least 30× off the current SoC budget. Even assuming a 4× improvement due to advances in process technology [14], we are at least 7.5× off our energy budget.

Unfortunately, the trend of scaling is not in favor of general-purpose architectures. Figure 1,4,6 shows the S_a trends of mobile CPUs, GPUs, and NPUs. Although the S_a of CPUs is decreasing, which means cores are getting more inefficient in favor of better single-thread performance, S_a of GPU and NPU is improving every year due to the better understanding of the target domain. For example, Samsung's NPU featured adder-tree-based dot-product engines so that it can specialize in convolutions with dilation and kernel decomposition, and feature-map zero-skipping to exploit that convolution neural nets with ReLU-based activation features many zeros [15]. Both hardware improvements and compiler optimizations lead to improvement in S_a when comparing end-to-end performance (normalized by TOPS) over generations of NPUs [15][16][17].

3.3.1.1 Finer DSA – One Size Does Not Fit All: The Curious Case of NPU

The need for highly efficient compute engines for deep neural networks (DNNs) is rapidly growing especially for product domains such as mobile devices, wearables, autonomous driving, and servers [18]. Since the compute requirements for both training and inferencing in DNNs are much higher than those of classical approaches, DSA made perfect sense for product domains with extreme power limitations and stringent real-time constraints. Traditionally, GPUs have been the go-to engine for conducting AI research for many developers. However, since the original purpose of the GPUs is to perform rendering and computer graphics processing, it is clear that it is not optimized for DNN-based compute. NPU has emerged as yet another new processor to gain energy efficiency and provide large amount of compute with relatively small area and power budget.

As more and more applications are switching from conventional approaches to DNN-based approaches, there is a higher demand to support more DNN models to run simultaneously. This puts even higher pressure on the inference/training engines to be even more efficient in terms of area and energy. Modern neural network models vary widely in size (that is, number of parameters in a model), complexity (for example, number of operations per inference), model topology and compute type (for example, various types of ops). For example, at the low end of the spectrum, there are NN models that perform Key Word Spotting (KWS) where the energy per inference is in the order of μJ/inference. At the other end of the spectrum, there are NN models that perform end-2-end Image Signal Processor (ISP) where the number of operations required may be several hundred TOPS per second (for example, 8K30 video signal processing for high quality recording) [19].

The first generation of NPUs focused on performing inferencing for Convolutional Neural Network (CNN). Architectures based on optimizing the compute for 3D tensor operators has been highly effective in performing tasks such as image classification and object detection. While general programmable NPU — whether this is more suitable as CNN or both CNN and RNN is more or less a problem of choice — is suitable for tasks such as face detection, object detection, low-resolution image segmentation, and image classification, it may not be efficient for handling NN models that are 100~1000× smaller in an ultra-low-power environment such as the always-on applications or wearables. For extreme low power and always-on applications, tiny NPU (or micro-NPU) may be a good engineering choice. A very small NPU located in always-on power domain may operate with extremely low power consumption with high utilization rate for audio/speech signal processing. One alternative is to add an always-on operation mode to the general NPU [17] rather than adding a dedicated tiny NPU. Such design choice may be driven by considerations of usage scenarios, area efficiency, and power efficiency.

In particular, NNs that perform simple user context tracking and keyword spotting are often based on Long Short-Term Memory (LSTM) of Recurrent Neural Network (RNN). The nature of compute for such tiny NNs are memory bandwidth bound rather than compute bound. Architectures aimed at balancing the two (compute vs. data transfer) while maximizing the utilization for such small models are heavily desired.

Another example of fine-grained NPU class is an engine aimed at pixel image/video processing. Figure 1.3.7 illustrates the characteristics of various image/video related workload. It is conceptually obvious to see the clusters, but it is interesting enough and may trigger yet another fine grain DSA – it might be prudent to have the NPU designs be specialized even further for additional gain. Unlike CNNs focused on detection and recognition where the output values are discrete classes (for example, face/non-face, foreground/background) or bounding boxes, image-processing NNs focus on generating pixel values as an output. Such examples include super resolution, blind deconvolution, denoising, and tone mapping. Since the amount of compute for a such task is 100~1000× higher than that of detecting faces, a different approach may be a good engineering choice. Feature maps typically have high spatial resolution and the number of model parameters are relatively smaller. In order to make the compute more manageable, shallow networks (for example, smaller number of channels) are often used, which are often derived from general approaches such as model reduction via Neural Architecture Search or distillation [20] or model scaling [21]. In this case, it will be more efficient to use a different class of NPU (for example, Image processing NPU) to handle these types of shallow networks with large feature map sizes.

3.3.2 3D IC: Distributed Si and Connection

Following the S_a innovation in mobile SoC, we will now look into the third factor, Y/W, from equation (3): $Y/W\$ = exp(-D_0 \times A)/W\$$. Although $Y/W\$$ increases as A gets smaller, A has stayed constant due to the minimum (even increasing) performance requirement. Here, instead of one chip with the size of A, putting together two chips with the size of $A/2$ can lead to improvement as in $Y/W\$ = exp(-D_0 \times A/2)/W\$$ while guaranteeing minimum performance level at the same time. This is where 3D packaging comes into the picture (we will not elaborate on the redundancy aspects).

While DSA divides and connects distinctive functions, 3D packaging physically splits the chip and puts together chips, shifting innovation level from intra-chip to inter-chip.

The green circles in Figure 1.3.8 represent $Y/W\$$ ratio of 3D IC over 2D according to chip area A. The bigger the area A, the bigger the gain. For a chip with large area such as server chip, we can earn hundreds of percent of improvement in $Y/W\$$ ratio, and even for the area of 1cm², close to typical mobile SoC, dozens of percent can be

978-1-6654-2800-2/22/$31.00 ©2022 IEEE

expected. In addition, it can offer a possible performance boost by reducing total routing length.

However, there are also limitations to this highly promising 3D IC in terms of interconnection cost, a barrier lying in the course of shifting from integrative qualitative change to distributive innovation. Only when the transition to distribution outperforms quantitative evolution at its limits, even with the additional interconnection cost, can we move on to qualitative innovation. We believe that, right now, we are standing somewhere in the middle of this transition period in terms of 3D IC history.

3.3.2.1 Interconnection Cost

Compared to intra-chip level, interconnection cost gets a stronger presence in inter-chip level distribution. A typical interconnection cost in 3D IC is the additional area needed for connection of each chip. Figure 1.3.8 also shows the $Y/W\$$ gain according to area penalty. Mobile chip whose area is typically ~1cm^2 can get the benefit only when the area penalty is less than 20%, whereas the larger server chip shows the gain up to 30% penalty. In addition to the area, various interconnection cost of 3D IC are as follows: 1) Physical interconnection cost: TSV (Through-Silicon Via) and bonding process; 2) Area overhead: TSV keep-out zone, additional test pads for a separate test; 3) Electrical characteristic: IR drop degradation, signal delay; 4) Thermal coupling: temperature increase because of the closer distance among hot spots

If CPR can be improved in spite of these effects, 3D IC technology would be highly likely to thrive across the semiconductor industry quickly.

Regarding the benefits and losses as above, a 3D IC roadmap can be suggested as Figure 1.3.9 in a mobile chip. 1) SRAM+Logic: SRAM does not offer the greatest $Y/W\$$ improvement since yield is less of a problem when using redundant resources, but it still has the benefit of simple design and low thermal coupling. There are already prototypes available for non-mobile devices, whereas this will be a suitable structure for a AI chip in mobile segment. 2) Partitioning by IPs: There is a high chance that mobile SoC will start in this forme. The key is to minimize performance degradation by dividing and allocating hot spots. We can expect $Y/W\$$ improvement since cutting-edge mobile chips typically starts mass production using the latest process technology whose defect density is still high. 3) Partitioning single IP into upper/lower stacks: Solid gain in performance as well as $Y/W\$$ can be expected, but the technical level of difficulty is extremely high. Placing CPUs top and bottom incurs heavy thermal coupling, and requires a separate DFT design, as well as bump or pad with small pitch due to a large number of interconnections.

3.3.2.2 From Micro to Macro

We have looked over how CPR has evolved in mobile chips, certain challenges it encountered, and other innovations available to tackle those challenges. Yet, there is one critical point left to discuss — the problem of power budget. As mentioned earlier, CPR formula can have meaning only under the power budget. In fact, one of the major reasons that mobiles show lower performance than desktops despite little difference in manufacturing processes — such as their inability to run real 3D or ray tracing scenes (yet) — is power limitations. If only the power budget can be further extended, another quantum jump in mobile chip CPR may be within our reach. One of the technologies that can bring this forward is cloud computing, such as edge computing. Cloud computing requires interconnection technology of the macro world. The interconnection technology has gone through its own evolution and performance-per-cost in communication has had a steady upturn. Likewise, the cost effectiveness of cloud computing can be determined in a similar way. If CPR improves even with the interconnection cost, there seems to be no reason not to use cloud computing.

In order to resolve the limit of scaling in the macro world, inter-device connection should be innovated both in quantitative and qualitative aspects. Such connectivity technology with low-cost, broadband, and low-latency features is also well reflected in the current mobile communication trend. One of the main target objectives of 5G New Radio (NR) in 3rd Generation Partnership Project (3GPP) standards is balanced development of three use cases including URLLC (Ultra Reliability and Low Latency Communication), mMTC (massive Machine Type Communication), and eMBB (enhanced Mobile Broadband). Figure 1.3.10 demonstrates the evolution of data rates and latency from the first generation of wireless cellular communication. For the last few decades up to 5G, it is interesting to see that each generation evolves almost every ten years (why different from Moore's cadence?) with the growth curve fully satisfying the exponential scale. The upcoming 6G communication, whose standards have not been officially set up yet, is also expected to follow this trend [22].

For mobile broadband service, 5G has evolved onto millimeter wave (mmWave) frequency region where wide bandwidth up to 400MHz can be supported per component carrier (CC). The number of CCs supported is showing a scalable increase depending on the computation power of the mobile device. In order to compensate for high signal attenuation throughout radio propagation, mobile devices as well as network equipment are setting up more antennas with analog-digital hybrid beamforming technology. In this regard, innovation in RF technology may play a key role in the successful operation of mmWave bandwidth [23].

Even with this technology, however, it is unlikely for mmWave to achieve the same level of coverage with long-term evolution (LTE) and sub-6 GHz. While mmWave communication has advantages in terms of high transmission rate, it demands higher cost for deploying more base stations than sub-6 GHz to cover all areas. However, as 5G matures, this problem will be settled naturally by seamless connection of both services. The overall wireless cellular systems can utilize sub-6 GHz as a baseline to support general traffic with wide coverage and seamless hand-over, while reinforcing special services that needs high speed and low latency traffic by supporting add-on mmWave technology. Here, connecting the data sent in each range of bandwidth can be completed by dual connectivity (DC) technology. Samsung's 5G modem has been already offering inter-band aggregation that integrates all of LTE, sub-6 GHz, and mmWave, and can provide up to 7.95Gb/s transmission rate in 5G stand-alone network.

Despite the improvements in broadband technology, the latency problem should be solved to satisfy the inter-device connection quality required in cloud or edge computing. The previous LTE consumed several dozens of *ms* latency in the air-interface and several hundreds of *ms* in the wired network, thus just settled for voice or video streaming Quality-of-Service (QoS) [24][25]. However, in the era of 5G, 3GPP managed to define URLLC services achieving the end-to-end (E2E) latency as 5*ms* for V2X (Vehicle-to-Everything) and 10*ms* for Augmented Reality (AR) services [26].

In the wired network, Figure 1.3.11 shows an example of a typical delay in data sharing between 5G User Equipment (UE) and server at the 5G core. First, latency in the wired network assumes distributed edge computing structure that has servers down at 5G core network — only taking into account connection from UE to 5G core, free from the backbone IP network. For the traffic to go all the way up to the public cloud in the backbone network, it may take hundreds of *ms* of latency at worst, but with edge computing E2E latency, it can come down to even 5ms. Edge computing will further evolve in 6G so that a device can heavily participate in computing activity, which will require innovation on the device level to efficiently perform a challenging split computing task.

In the air-interface, 5G supports higher subcarrier spacing (SCS) in addition to 15kHz used in LTE which can realize shorter orthogonal frequency-division multiplexing (OFDM) symbol length. This, combined with mini-slot operation, facilitates low latency operation down to 1ms. Furthermore, Release 16 revealed further advanced latency features which has been adopted as the latest version of Samsung's 5G modem.

Next key factor in 5G is the change in algorithm. Modem performance improvement, which is directly related to transmission rate and latency, is a critical factor for better quality of inter-device cooperation. It is notable that improvement in computing power of a modem chip brought more room to adopt new advanced algorithms. The iterative detection and decoding (IDD) technology applied to Samsung's 5G modem chip can be a good example [27]. Recent silicon performance advancement has enabled the iterative form of detection and decoding technology such as the turbo equalizer. As the detector and decoder of baseband processor iteratively share Log-Likelihood Ratio (LLR) of a coded bit as depicted in Figure 1.3.12, IDD performance approaches the theoretical bound of Maximum A Posteriori (MAP) detector [28][29][30]. The conventional Maximum Likelihood (ML) receiver has less complexity since it assumes *a priori* distribution of all bits of transmitted signal as constant [31], while the the IDD receiver achieves near-MAP performance by delivering LLR feedback from the decoder to the detector and regenerating detector output based on the updated *a priori* distribution computed from LLR feedback. With the repeated iteration, the accuracy of the computed *a priori* distribution value increases and brings out better performance. Figure 1.3.13 shows the performance benefit earned by IDD when transmitting 4-layer multiple-input and multiple-output (MIMO) signals for NR sub-6 GHz. Converted into 5G network performance, this would imply about 5~10% increase of system throughput.

4.0 Conclusions

The Spanish flu in 1918 showed 6% negative growth rate on the world economy, while taking away as much as 20% of the population in some countries [32]. On the other hand, the COVID-19 pandemic revealed 3% negative growth rate in 2020, and it is expected to recover up to 6% in 2021 even with the pandemic still in progress [33]. The pandemic forced us to scale down our society to smaller sections, but nevertheless we have found ways to effectively connect them and adjust to live in this distributed world. We are seeing similar changes in the semiconductor industry.

We have covered multiple barriers that are difficult to overcome, both of technical and economical aspect of scaling. The system architecture is also another barrier—general-

purpose designs inevitably scales poorly. The natural step to survive this scalability problem is to distribute the task to smaller specialized pieces, and to go for technologies that connect and integrate them effectively. Hence, we focused on domain-specific computing, 3D packaging, and communication networks to make that happen.

Naturally, creating a system that connects different components requires connecting various people with different backgrounds—not only with engineering and development backgrounds, but also with governments, educators, and even hobbyists. This requires people who are willing to learn things quickly and solicit communication amongvarious groups of people. Fortunately (or unfortunately), our younger generation has grown naturally accustomed to the ever-connected, distributed world (~ hooked on the Internet, and cell phone). These people have learned how to collaborate with people who they have never met, to embrace their differences, and to live in the ever-connected new world. This has to be the *zeitgeist* until the next epoch!

Acknowledgment:
It has been a great pleasure to work with the brilliant minds of Samsung System LSI co-workers for the paper, kudos to Suknam Kwon, Bogyeong Kang, Hui Won Je, Junhee Yoo, Sung-Boem Park, Sukhwan Lim, Jung Hyun Bae, Daniel D.H. Kim, YoungCheol Min, Eunbi Shin, and Ilryong Kim; from English literature major to math, physics, industrial engineering, and a bunch of EE and CS — distributed and connected.

References:

[1] J. Hennessy and D. Patterson, "Computer Architecture, A Quantitative Approach," *6th edition, Morgan Kaufmann*, p.3, 2017.

[2] G. Moore, "Progress in Digital Integrated Electronics," *IEEE International Electron Devices Meeting Technical Digest*, pp. 11-13, 1975.

[3] S. Seung, "Connectome: How the Brain's Wiring Makes Us Who We Are," *Mariner Books*, 2012.

[4] R. I. M. Dunbar, "Neocortex size as a constraint on group size in primates," *Journal of Human Evolution*, vol. 22, issue 6, pp. 469-493, June 1992.

[5] Y. Harari, "Sapiens: A Brief History of Mankind," *Harper*, 2015.

[6] S. Natarajan et al., "A 32nm logic technology featuring 2nd-generation high-k+ metal-gate transistors, enhanced channel strain and 0.171um^2 SRAM cell size in a 291Mb array," *Proceedings of the IEEE International Electron Devices Meeting (IEDM)*, Dec. 2008.

[7] W. Qadeer, R. Hameed, O. Shacham, P. Venkatesan, C. Kozyrakis, and M. Horowitz, "Convolution engine: Balancing efficiency and flexibility in specialized computing," *Communications of the ACM*, vol. 58, issue 4, pp. 85-93, Apr. 2015.

[8] Samsung Electronics, "Samsung C110 (Exynos 3110)," [Online]. Available: *https://www.samsung.com/semiconductor/minisite/exynos/products /mobileprocessor/exynos-3-single-3110*, [Accessed Oct. 25, 2021].

[9] [Unpublished product, link will be updated when public].

[10] Samsung Electronics, "AI Camera: Redefining mobile photography," [Online]. Available: *https://www.samsung.com/semiconductor/minisite/exynos/technology /ai-camera* [Accessed Oct. 25, 2021].

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, ''Rethinking the inception architecture for computer vision,'' *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, June 2016.

[12] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64270-64277, 2018.

[13] M. Horowitz, "Computing's energy problem (and what we can do about it)," *Proceedings of the IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 10-14, Feb. 2014.

[14] N. P. Jouppi et al., "Ten lessons from three generations shaped Google's TPUv4i: Industrial product," *Proceedings of the ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, pp. 1-14, June 2021.

[15] J. -S. Park et al., "9.5 A 6K-MAC feature-map-sparsity-aware neural processing unit in 5nm flagship mobile SoC," *Proceedings of the IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 152-154, Feb. 2021.

[16] J. Song et al., "An 11.5 TOPS/W 1024-MAC butterfly structure dual-core sparsity-aware neural processing unit in 8nm flagship mobile SoC," *Proceedings of the IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 130-132, Feb. 2019.

[17] J. -S. Park et al., "A multi-mode 8k-MAC HW-utilization aware neural processing unit with a unified multi-precision datapath in 4nm flagship mobile SoC," *to appear at Proceedings of the IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2022.

[18] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "Survey of machine learning accelerators," *Proceedings of the IEEE High Performance Extreme Computing Conference (HPEC)*, Sep., 2020.

[19] W. Yang et al., "Deep learning for single image super-resolution: A brief review" *arXiv preprint* arXiv:1808.03344, Aug. 2018.

[20] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network", *arXiv preprint* arXiv:1503.02531, Mar. 2015.

[21] Mingxing Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *Proceedings of the International Conference on Machine Learning (ICML)*, June 2019.

[22] Samsung Research, "The next hyper connected experience for all," *white paper*, July 2020.

[23] A. Verma et al., "A 16-channel, 28/39 GHz dual-polarized 5G FR2 phased-array transceiver IC with a quad-stream IF transceiver supporting non-contiguous carrier aggregation up to 1.6 GHz BW ," *to appear at Proceedings of the IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2022.

[24] M. Alasti, B. Neekzad, J. Hui, and R. Vannithamby, "Quality of service in WiMAX and LTE networks," *IEEE Communications Magazine*, vol. 48, no. 5, pp. 104–111, May 2010.

[25] N. Ali, A. Taha, and H. Hassanein, "Quality of service in 3GPP R12 LTE advanced," *IEEE Communications Magazine*, vol. 51, no. 8, pp. 103–109, Aug. 2013.

[26] 3rd Generation Partnership Project (3GPP), System Architecture for the 5G System (Release 15), *TS 23.501*, 2017.

[27] H. Kwon, J. Lee, and I. Kang, "Communication system with iterative detector and decoder and method of operation thereof," *US patent*, no. 8897406, Nov. 2014.

[28] B. M. Hochwald and S. ten Brink, "Achieving near-capacity on a multiple-antenna channel," *IEEE Transactions on Communications*, vol. 51, no. 3, pp. 389-400, Mar. 2003.

[29] M. O. Damen, H. E. Gamal, and G. Caire, "On maximum-likelihood detection and the search for the closest lattice point," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2389–2402, Oct. 2003.

[30] S. ten Brink, G. Kramer, and A. Ashikhmin, "Design of low-density parity-check codes for modulation and detection," *IEEE Transactions of Communications*, vol. 52, no.4, pp. 670–678, Apr. 2004.

[31] D. Garrett, L. Davis, S. ten Brink, B. Hochwald, and G. Knagge, "Silicon complexity for maximum likelihood MIMO detection using spherical decoding," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 9, pp. 1544–1552, Sep. 2004.

[32] R. J. Barro, J. F. Ursúa, and J. Weng, "The Coronavirus and the great influenza pandemic: Lessons from the "Spanish flu" for the Coronavirus's potential effects on mortality and economic activity", *Working Papers of National Bureau of Economic Research (NBER)*, No. 26866, Mar. 2020.

[33] International Monetary Fund (IMF), "World Economic Outlook: Real GDP growth," [Online]. Available: *https://www.imf.org/external/datamapper/NGDP_RPCH@WEO/OEMDC/ADVEC/WEOWORLD* [Accessed Apr. 2021].

978-1-6654-2800-2/22/$31.00 ©2022 IEEE

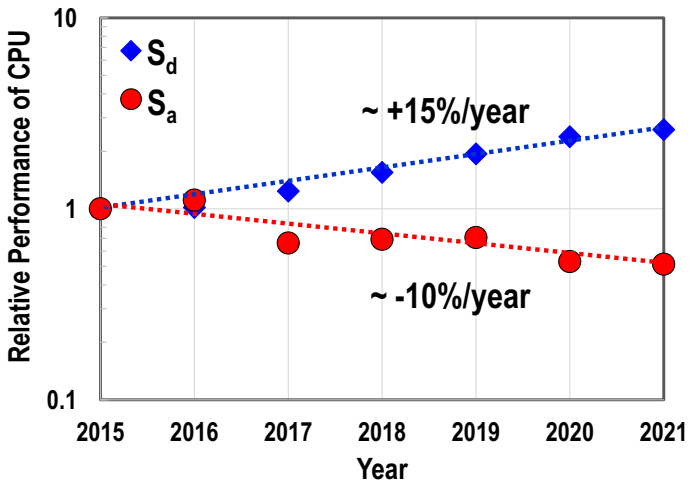

Figure 1.3.1: Relative performance of S_d and S_a.

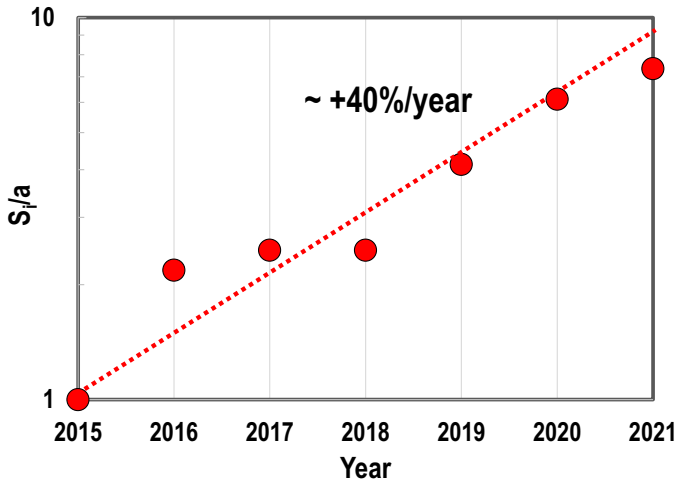

Figure 1.3.2: Recent trend of Moore's Law and Dennard Scaling combination.

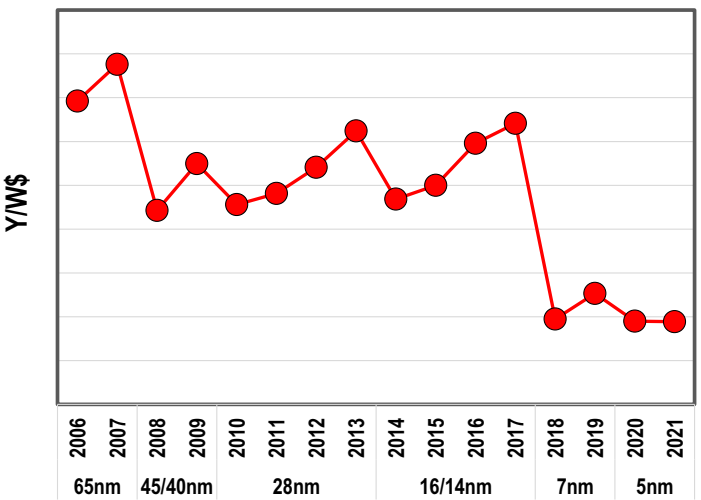

Figure 1.3.3: 1/(Effective Wafer Price) trend.

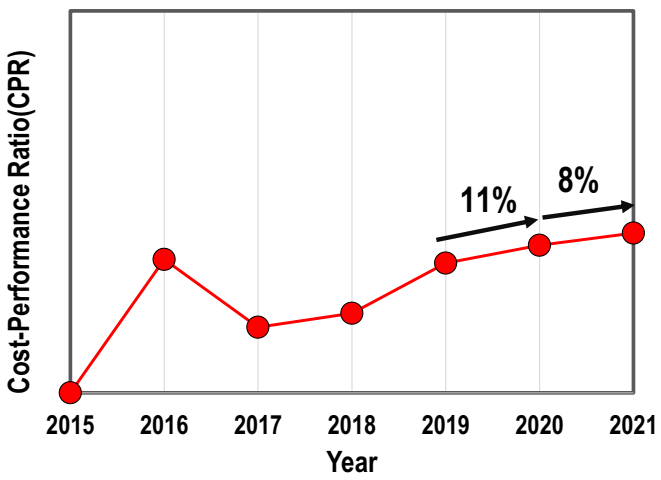

Figure 1.3.4: Cost-Performance Ratio(CPR) trend.

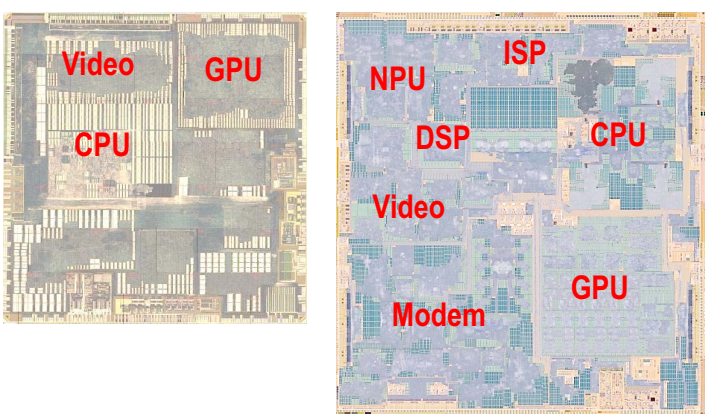

Figure 1.3.5: C110 (Left, 61.7mm² @ 45nm), Exynos 2200 (Right, 99.9mm² @ 4nm).

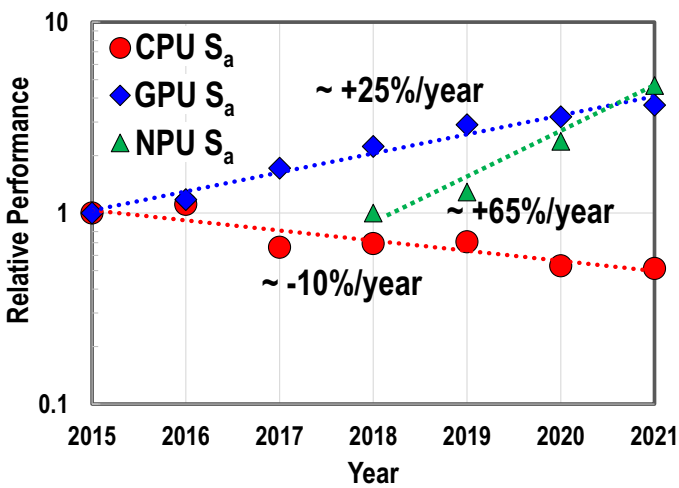

Figure 1.3.6: S_a trend of CPU, GPU and NPU.

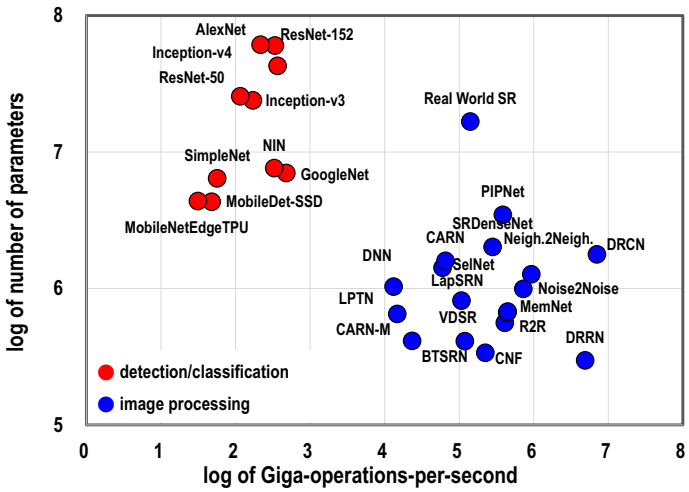

Figure 1.3.7: Deep learning applications – detection/classification vs. signal processing.

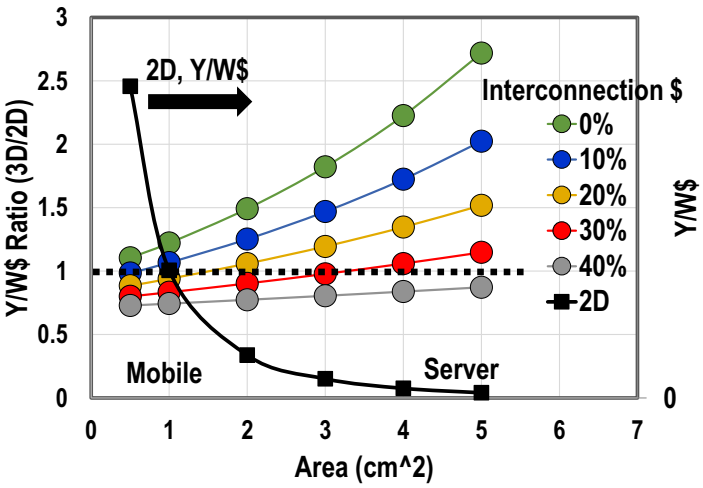

Figure 1.3.8: Interconnection cost vs. Y/W$ gain (defect density: 0.4ea/cm²).

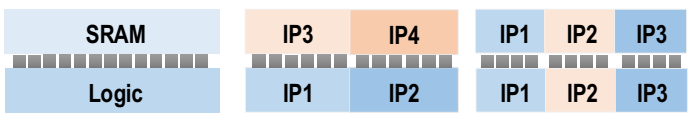

Figure 1.3.9: 3D IC roadmap of mobile chipset.

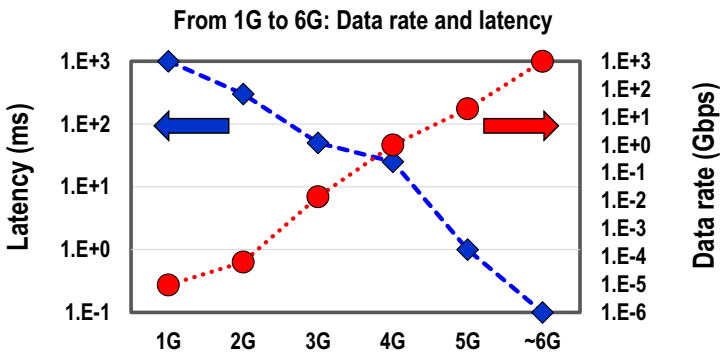

Figure 1.3.10: Peak data rate and air interface latency of cellular networks.

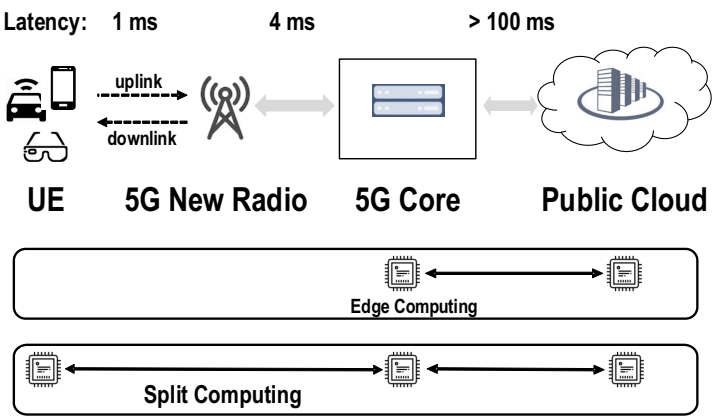

Figure 1.3.11: End-to-end latency between 5G UE and cloud.

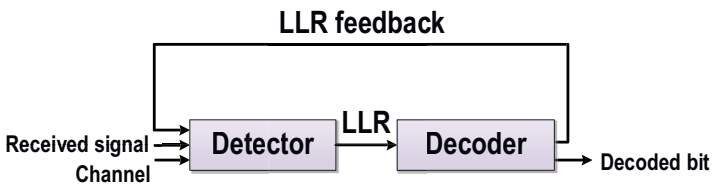

Figure 1.3.12: Iterative detection and decoding.

978-1-6654-2800-2/22/$31.00

Rank	MCS5 (16QAM)	MCS11 (64QAM)	MCS20 (256QAM)
Rank2	0.61dB	1.29dB	0.71dB
Rank4	1.01dB	1.41dB	0.72dB

Figure 1.3.13: SNR gap between IDD on/off of 5G device (@ 10% BLER).