1.2 Fertilizing AIoT from Roots to Leaves

Kou-Hung Lawrence Loh

Senior Vice President & Corporate Strategy Officer
MediaTek, Hsinchu, Taiwan

1.0 Introduction

IoT with artificial intelligence, AIoT, enriches everything around the world. The application space is unlimited, ranging from fundamental research, enterprise, industry, transportation, services, and personal daily life. The impact of AIoT is far reaching. By 2030, with an average of 40 connected "things" per person and 8.6 billion population in the world, it is expected to have ~350 billion devices in operation, reaching $16 trillion, or 14% of global GDP, as illustrated in Figure 1.2.1.

AIoT comprises a wide spectrum of technologies that can be categorized into three pillars, as shown in Figure 1.2.2. A broad view of multimedia consists of sensors, human-machine interface, and actuators interacting between the digital and real worlds. It can include microphones and speakers, cameras and displays, vital-sign monitoring devices, robotaxis, and so forth. The massive amount of data is processed both locally and in the cloud, requiring 5G and other advanced communication technologies to reliably support high network capacity and high data throughput at low latency. The inputs to the cloud and edge AI units are analyzed, classified, and perceived to synthesize cognitive actions with a capability greatly surpassing human ability by 2030. A large variety of technologies such as novel device components, new AI platforms, heterogeneous computing units, advanced wireless and wireline communication, and package assembly are required to anchor the pillars and to support the rich leaf-applications at the top. In addition, excellent power efficiency to prolong battery life is a must for edge devices, which also helps to reduce carbon emission. In the following sections, nurturing and expanding each pillar from roots to leaves will be discussed with focuses on AI for multimedia, communications technology to bridge the edge and cloud, and essential circuit-level technologies.

2.0 The World with Artificial Intelligence

The availability of big data in conjunction with the advancement of Deep Neural Networks (DNN) and high compute-power VLSI drives the emergence of modern AI. One of the most deployed applications is computer vision. Compared to the conventional image processing algorithm approach, accuracy has improved so much that by 2015, a human can no longer beat a computer in image classification. AI has expanded everywhere, including into the high volume consumer space such as smartphones, advanced driver assistance systems (ADAS), voice assistants, surveillance cameras, and smart remote healthcare systems.

The early development of DNNs before 2016 mainly focused on accuracy improvement without considering computation complexity. While accuracy for image classification improved 24% between 2012 and 2016, the demand on computing resources increased >10× as illustrated by the bubble sizes in Figure 1.2.3. Starting from 2017, significant attention was paid to improve computation efficiency in terms of compute power, memory bandwidth, and power consumption, while maintaining accuracy at a similar level to human perception. As a result, on-device DNN inference started to become feasible, opening up the era of edge AI. It is expected that the efficiency of DNNs will continue to improve, enabling more sophisticated edge AI applications.

2.1 Edge AI

The initial success of AI was built on big data and huge amounts of computation resources. It is natural that in those early generations, cloud-based solutions were preferred, not only for training, but also for inference. Taking VGGNet (2014) [1] as an example, a single image classification [2] needed 15.2GMACs (15.2 billion multiply-and-accumulations). It took 250ms to 1250ms to complete, where 200ms to 1200ms was required for data transportation and 50ms was for the actual computation in the cloud. If the same example was run on a high-end smartphone at that time, it would have taken ~3000ms, which is 2.5× slower. In 2016, improvements to DNN models and VLSI computing capability shortened the gap such that a high-end smartphone could perform similar image classification at higher accuracy in less than 50ms without any connection to the cloud. The resulting user experience surpassed cloud-based solutions by more than an order-of-magnitude. The 60× improvement, from 3000ms to 50ms, was the result of a 30× improvement in hardware computing capability and improved compactness of DNN models such as Inception-ResNet (2016) [3] which reduced the computation complexity by 50% compared to earlier models. Such abrupt progress proved that processing DNN tasks locally (the nomenclature of edge AI) is feasible for mission-critical applications in real time. In fact, there is a clear and fast rise of edge AI because of: (1) the need for real-time processing and inference; (2) the increased awareness of privacy and security to keep critical data local; (3) the requirement of offline operations when internet access is sporadic; and (4) the challenge of communication capacity stressed by 350 billion AIoT devices globally by 2030.

2.2 Cloud-Edge Hybrid AI

To date, the exceptional growth of edge AI devices has been mainly in inference engines. Most of the training still resides in the cloud. For example, an inference engine running ResNet50 needs ~3.9GOPs to classify one image, which takes about 10ms on an edge device having 500GOPs capacity. In contrast, training ResNet50 requires processing over 1 million images for 100 epochs (iterations). That corresponds to >2 Exa-OPs, far exceeding the capacity of an ordinary computer. Even running in the cloud, it takes several hours with servers equipped with GPU farms, large arrays of memory, and data storage. Although there is active research on reducing DNN training complexity [4-5], it is still difficult to carry out on an edge device. Therefore, the preferred approach is to rely on the cloud for heavy computation, leaving inference to the edge device, which can interact with the environment in real time, maintaining strong privacy and security. New data collected by edge devices can also be anonymized by AI while preserving key attributes for cloud learning, self-learning, and federated learning. For example, in an autonomous driving ecosystem within a smart city, a connected autonomous vehicle needs to fuse data from its own sensors and those transmitted from the traffic infrastructure in real time to perform safety-critical AI inference. The collected and fused data can be pre-processed to remove private information before being sent to the cloud at a relatively slower pace. The cloud server then uses the aggregated data to update traffic control, HD maps, and to facilitate future city planning. The overall cloud-edge hybrid AI and communication sub-systems form a closed-loop AIoT ecosystem as part of a smart city.

2.3 Edge AI SoC

The broad array of AI application places various demands on hardware design. In the case of image processing, as shown in Figure 1.2.4, since the AI models are mostly non-iterative and need to deal with a large amount of data (pixels), it is more demanding on computation than on memory access. The former can reach up to ~40TOPs, while the latter requires ~15GB/s. On the other hand, for high-end voice applications, although the computation requirement of ~30GOPs is 1000× lower than that for imaging processing, the repetitive operations of recurrent neural network (RNN) and long-short-term memory network (LSTM) call for a higher memory bandwidth of >20GB/s. Simulated compute power, data precision, and memory bandwidth for several computer vision and voice AI-use cases are listed and compared in Figure 1.2.4. The benchmarks range from perception/recognition, object-based construction, and advanced contextual understanding. It can be observed that the compute power and memory bandwidth requirements vary by >3 and >1 orders-of-magnitude, respectively. With a desire for smart handhelds, such as smartphones, to be versatile for most, if not all applications, it becomes a big challenge to design an edge AI SoC that is simultaneously scalable and power efficient.

Beyond elementary AI functions, such as image classification, consider a "follow-me dancing" application, which is easy for humans but takes very heavy concurrent and heterogeneous AI computing, as shown in Figure 1.2.5. Starting from the camera sensor, a live video stream is produced at 30fps. The image signal processing unit (ISP) performs auto-focus, auto-exposure, and auto-white balance (denoted as 3A), as well as producing two types of images, one of which may be enhanced by AI for human viewing, and the other for internal AI perception. The AI processing unit (APU) operates on the internal data to perform face detection (denoted as FD), pose estimation, and 3D reconstruction. The result is simultaneously used to control the robot through Bluetooth and is displayed on a TV through WiFi. A CPU controls the overall scheduling and carries out some of the scalar computation. All operations are concurrently executed by a single smartphone in real time. In essence, an edge SoC needs to be equipped with not only a capable inference engine, but also multiple other processing units to tackle highly complex applications.

The challenges in designing an edge AI SoC may be summarized as follows: (1) flexibility and re-configurability to support various neural networks with dissimilar demands; (2) scalability to fulfill a very wide range of workload; (3) reduction of the amount and distance for data movement, which can consume up to 100× more energy than the ALU alone in certain cases; and (4) low power consumption on the order of 2W to 3W total with ~1W budgeted for a deep learning accelerator (DLA). Architecture wise, it is instructive to first examine the strength and weakness of different types of processors, as shown in Figure 1.2.6. While a CPU has the best flexibility, it is the least efficient compared to a GPU, DSP, or purposely-built APU. For power efficiency, it is the complete opposite – an APU outperforms the rest, but it has the least flexibility. It is important to note that AI tasks consist of not only DNN operations but also traditional signal processing. In fact, for some computer vision (CV) applications, the traditional algorithms can be responsible for up to ~70% of the total computation load. This leads to the architecture of multi-processor system-on-chip (MPSoC) to support concurrent heterogeneous computing.

An APU is purposely built to do AI tasks efficiently. A block diagram is shown in Figure 1.2.7 [6]. It consists of convolution engines, data buffers, and operator engines for neural network operations, such as pooling and activation. A parallel and scalable computing architecture with multiple APU cores provides acceleration for commonly used DNN operations. Adaptive computing is exploited to increase ALU utilization. Distributed local memory and data flow control allow reuse of data between different convolution windows and neural network layers, thereby minimizing the physical data movement. Among APU cores and other heterogeneous processors, tightly-coupled global buffers are used to share data so that relatively slow and inefficient DRAM access can be reduced. Hardware-based data compression is further utilized to free up computing resources and decrease the need for memory bandwidth. To deal with sparsity, joint processing from both hardware and software is developed to perform zero-skipping. All APU cores support multiple numerical representation and native asymmetric quantization to balance power and performance while meeting the target accuracy. The scheduling and synchronization are done by both software and hardware for robustness and optimal performance. Benchmarking using Inception and MobileNet, the APU can achieve more than 6 TOPS/W, performing much better in power efficiency compared to a GPU. Consequently, more than 60% of the DRAM bandwidth requirement can be removed.

2.4 Software Swiss Army Knife

There are a large variety of AIoT applications demanding diverse algorithms and neural network models. The best suited hardware configuration can also be drastically different for those use cases. It is therefore very critical to co-design the hardware and software algorithms starting from the planning stage, and take into account that the application space is only limited by users' imagination. A general platform approach is needed where a software development kit (SDK) needs to support common machine-learning (ML) frameworks while simultaneously optimizing the system performance based on the hardware offering. In addition, to enable developers to focus on system-level AI innovation and fast time-to-market, application libraries, framework API, dynamic hardware arrangement, and toolkits for profiling, and so on are essential. Figure 1.2.8 shows an example platform marketed as NeuroPilot™, from MediaTek. The API, marked in red, supports common frameworks such as TensorFlow, Caffe, Caffe2, ONNX, and MXNet over multiple operating systems, including Android, Linux, and RTOS. The Heterogeneous Runtime module, marked in blue, performs platform-aware inference optimization based on the available computing resources. Hardware configuration and parallelization are dynamically arranged during run time. Network reduction such as pruning, quantization with compensation, zero-skipping for sparse matrices, and weight compression is supported by the toolkits to achieve optimal results and power efficiency. The overall platform enables application-level optimization to be carried out as early as possible during the algorithm design phase, which is vital when developing complex applications.

Neural networks are mostly trained in the cloud today. When deploying the models in edge devices, it typically takes a significant amount of engineering time and effort to select suitable network architectures, re-train the parameters, and trim the networks according to the available energy and hardware constraints. Since the inference optimization is highly complex, it is natural to adopt AI inside the aforementioned platform, such as NeuroPilot™ , to automate each step in the process. From our experience, reinforcement learning with iterative refinement works well for this purpose. An example is shown in Figure 1.2.9. When running InceptionV3 on a CPU, it takes ~210ms. Migrating to an APU, the execution time can be reduced down to 15ms (14× speedup) to further improve the performance and power consumption, it may require ~4 weeks with a group of domain experts to modify and optimize the network. If there are multiple criteria to be achieved, the engineering efforts would explode. In contrast, by deploying fully-automated neural architecture search (NAS), it takes only 3 days to generate a new model that can either speed up the inference from 15ms down to 8ms under the same accuracy, or increase the accuracy from 78% to 82% while keeping the same execution time.

In the next several years, more advances will occur in interconnect technology for network-on-chip (NoC), system-level HW/SW co-design with scalability and re-configurability, as well as device and circuit-level innovation. One of them is near-memory and in-memory computing (or compute-in-memory, CIM) to further reduce data-centric power consumption. The intent is to eliminate the bottleneck in the classic Von Neumann compute-centric architecture. It is expected to achieve significantly higher performance of up to 5TOPs/W to 100TOPs/W, compared to current SRAM-based implementation of 1TOPs/W to 2TOPs/W. From an architectural viewpoint, CIM-SRAM may be treated as a tightly-coupled memory (TCM) for deep-learning accelerator (DLA). Although there has been a lot of active research on CIM, the maturity of this technology may still be a few years out.

3.0 Connecting the Mesh

With the necessity of cloud-edge AI, an integral part of AIoT is the massive amount of aggregated data that is transferred bi-directionally between the digital and real worlds. To support an average data-traffic growth rate of around 150% per year, wireless and wireline communication become very crucial. It is challenging for leaf devices in Figure 1.2.2 to economically fulfill high communication quality because the available energy is constrained and the data from each edge device is often sparse and sporadic. Therefore, innovations in communications standards, communications network architectures, and integrated circuit design are mandatory to facilitate adaptive data traffic at low energy consumption. Local area networks configured in star or mesh structures are bridged by gateways which are finally linked to the cloud through cellular, WiFi, Ethernet, and so forth. Within data centers, high-speed SerDes such as 112 and 224 Gb/s/lane, and optical communication links serve as the communication backbone transporting data at tens of Tb/s among racks. Long-distance fibers and satellites finally interconnect the world to form a global mesh. The whole communication ecosystem has enabled the exponential growth of AIoT. Yet, it is far from sufficient to support the expansion from 50 billion devices in 2020 to 350 billion devices by 2030.

Wireless and wireline communication capacity has also increased at an exponential rate as shown in Figure 1.2.10, which corresponds very well to the number of AIoT devices. The estimated annual data traffic by 2021 is in excess of 3.3 zettabytes (10^{21}) [7]. This densely deployed network comes with new challenges such as security, data traffic management, signal interference, and quality of service, all of which must be fulfilled with low power consumption. In addition to the aforementioned core technology advancement, AI can be pursued to tackle some of the obstacles. For example, AI assistance on GPS signal acquisition for an edge device, as well as resource, security, quality, and interference management for wireless infrastructure have been partially deployed in the field [8]. It is still an active research area. In the following two sections, the current state-of-the-art (SoA), requiring innovation in the next 10 years for wireless and wireline communications will be discussed.

3.1 Advances in Wireless Communication

Edge devices that are connected wirelessly can be divided into two classes: those with low throughput such as IoT sensors, and those with high throughput such as smartphones. Each class can be further categorized according to the distance from the gateways which can be satellites, cellular basestations, or WiFi access points. Various wireless standards exist to service these multiple connection scenarios, such as 3GPP and 802.11.

The main focus for low-throughput IoT devices is on reducing energy consumption per bit (power/throughput = Joules/bit). It is generally achieved by adopting narrow-band communication systems along with extreme optimization on both active and sleep-mode current consumption. It is worth noting that the magnitude of sleep current is usually significantly smaller than that of active current. Yet, it often overwhelms the battery life time for low duty cycled devices. For active mode, wireless standards such as NB-IoT and BLE facilitate the

 978-1-7281-3205-1/20/$31.00 ©2020 IEEE

reduction of average power consumption for both media access control (MAC) and physical (PHY) layers. Analog-light transceiver architectures, low voltage RF, wake-up receivers, and fast transition between sleep and active modes are techniques to minimize the effective Joules/bit. To date, a SoA BLE receiver can achieve an energy efficiency of as low as 2nJ/bit [9]. At the system level, cognitive AI is being employed to make the best use of the available energy by scheduling data collection, processing, and transmission at optimal moments. The ultimate goal is to reduce current consumption so drastically that AIoT devices at the edge can be powered solely by harvesting the energy from their surrounding environment (so-called zero-energy edge devices).

The key demand for high-throughput edge devices is high data rate at high reliability and low latency while balancing within a constrained energy budget. To achieve this goal, the wireless industry has introduced a rich set of technologies to fully utilize the available frequency spectrum in the sub-6GHz range by aggregating fragmented spectrum (carrier aggregation), upgrading to higher order modulation, adopting multi-user MIMO and coordinated multi-point, to name a few. In the latest wireless standards for 5G-NR and 802.11ax, the theoretical throughput can reach 20Gb/s and 14Gb/s, respectively. As we are already close to the limit according to Shannon's theory for a given frequency bandwidth, more frequency spectrum has to be allocated to simultaneously support high throughput and high capacity. Various frequency bands between 6GHz and 100GHz have been opened up for 5G-NR (5.9, 28 and 39 GHz) and next generation WiFi (6GHz to 7GHz).

It is daunting for an edge device, such as a smartphone, to handle those complex techniques, large bandwidths, and multiple frequency bands. Thousands of people over several years are needed to design the highly complex modems and transceivers, not to mention the additional efforts to design them in a power-efficient manner. To gain an appreciation of the challenges, Figure 1.2.11 shows the increase of modem complexity over the past two decades using data from MediaTek. Over the course of the evolution from 2G to 5G between 2005 and 2019, modem gate count has increased 240×. Although Moore's Law and new packaging technologies have helped tremendously, the 5G modem is reaching the limit of power density even the most advanced process technology node has been adopted. Given an average of 5 to 10 years to advance to the next cellular communication generation, a 6G modem is expected to have a gate count on the order of 2000× that of a 2G modem. A similar trend can be formulated for silicon area as shown in Figure 1.2.11. A 5G modem size is ~10× that of a 2G modem, and 6G is expected to be >50×. For RF transceivers, benefits from Moore's Law are less pronounced. A modern 5G transceiver requiring 3 TX and 20 RX elements is shown in Figure 1.2.12 [10]. Compared to the earlier generation with only 1 TX and 1 RX, the size increases 8×. Similarly, a new WiFi 6 transceiver consisting of 4 TX and 4 RX to support 4×4 MIMO in 3 frequency bands (2.4GHz, 5GHz, and 6 GHz) [11] is ~10× larger than the 1×1 single band (2.4GHz) version.

For RF receiver energy efficiency shown in Figure 1.2.13, cellular links have improved by 5 orders of magnitude from 2.5G to 4G between 2005 and 2017. The huge improvement mainly comes from user demands evolving from voice only to large data transportation. As a result, superior communication signal processing techniques such as OFDM, turbo coding, and high order modulation, as well as advanced RF/analog circuit design have been developed. The trend levels off after 2017 as we approach some of the theoretical limits. WiFi, on the other hand, has employed OFDM and relatively higher order modulation than cellular standards due to its focus on data and a shorter distance coverage. Hence, its energy/bit starts at 3 orders-of-magnitude lower than the corresponding cellular generation. The trend is also shown in Figure 1.2.13 with a less steep slope than cellular standards. It can also be observed that after 2017, both cellular and WiFi share similar energy efficiency improvement over time, which is substantially slower than the cellular trend in the past decade. That means significant innovation is needed in the next 10 years to enable the goal of zero-energy edge devices. The challenges lying ahead are deep and broad spanning across many areas such as communications systems and RF/analog engineering, and data science and AI. For example, mmWave offers large bandwidth but requires revolutionary techniques to overcome the insufficient link budget for a large cell size coverage. Similarly, for future WiFi 7 (802.11be) to support 4096QAM using 320MHz channel bandwidth and up to 16 spatial streams (up to 16), it requires an EVM floor as low as –45dBc which is near the level of today's verification equipment.

3.2 Advances in Wireline Communication

All the data that is gathered from the billions of edge, or "leaf", devices eventually ends up in the cloud, carried there by large communication backhauls such as long distance fiber. Once in the data center, moving the data around is itself a massive undertaking. It is estimated that the total data traffic inside data centers (Figure 1.2.14) will reach 20.6 zettabytes (10^{21}) by 2021, imposing humongous efforts to accommodate all the data movement.

Wireline IOs, or SerDes, are the primary building blocks for moving data around. These IOs range from extremely-short-reach (XSR), ultra-short-reach (USR) to long-reach (LR) and can be found in data center switches, ASICs, and various PHYs. In recent years, dramatic advancements in more IOs per component and higher data rate per pin have emerged. As illustrated in Figure 1.2.15 [12], the per-pin data rate has approximately doubled every 4 years across a variety of I/O categories, from DDR to graphics to high-speed Ethernet. What has made this possible is the introduction of more complex modulation schemes, similar to what has occurred in wireless. One such example is PAM4, which has led to the successful mass deployment of 56Gb/s and soon 112Gb/s SerDes. At these data rates, complex modulation schemes suffer significant inter-symbol interference (ISI), which must be mitigated through equalization. Recently, the wireline industry has recognized that DSP-based architectures offer a promising path for reaching 400Gb/s and beyond by taking full advantages of process scaling. For example, Figure 1.2.16 shows that when scaling from 16nm to 7nm, the analog power and area reduced by less than 25% while digital power and area scaled by ~50%. This skew causes a paradigm shift in wireline communication transceivers towards DSP-based architectures. Shown in the inset of Figure 1.2.16 is an 112Gb/s transceiver that takes advantage of this scaling skew [13].

At the same time, however, the increased energy consumption that comes hand-in-hand with increased data rates is leading to cooling and cost problems that must be dealt with. For a complete wireline PHY including transceivers, clock generation, etc., the current SoA energy consumption is 1pJ/bit over 10's mm of distance in the case of 112G XSR [13] and 0.5pJ/b in the case of MediaTek's M-Link, deployed in what is currently the industry's largest dual-die InFO package (see next section). The challenge of increasing data rate while simultaneously reducing the energy per bit will serve as a key goal for the industry, one that cannot be solved by relying on DSP-based architectures and Moore's Law alone.

To this end, new innovations such as silicon photonics co-integration, offer the promise of lowering energy consumption by 60% and being cost effective at the same time. However, challenges related to reliability, testing, and maintenance of silicon photonics compared to electrical I/Os will delay their mass deployment for the foreseeable future. Until then, wireline I/Os will continue to rely on electrical interconnects, with ever more complex modulation schemes likely to be introduced.

4.0 Under the Hood of Circuit Design

During the past decades, Moore's Law has been the key ingredient "fertilizing" advances in computing, and wireless and wireline communications, as illustrated in the previous sections by the examples of AI processors, modem complexity and DSP-based architectures, respectively. From 2007-2017, gate density grew approximately 10× every 5 years (Figure 1.2.17), which is just shy of the predicted rate given by Moore's Law of growing 2× every 1.5 years. The growth rate, or slope, is coincidently similar to the increase in data rate as shown in Figure 1.2.10. However, as process scaling is no longer following a straight line on a logarithmic scale (Figure 1.2.17), the interest in various More-than-Moore techniques has intensified. In particular, advanced packaging technologies to enable heterogeneous integration have received high attention. The following paragraphs will examine some of these advanced techniques being adopted to overcome the limitations of scaling.

4.1 Advances in Package Technology

Very recently, enabled by the development of XSR SerDes, the idea of chiplets - the breaking up and disaggregating of large chip functions into smaller chip functions - has gained significant traction. In the chiplet concept, the smaller chips are assembled together through advanced, 2.5D and 3D packaging technologies such as CoWoS® (chip-on-wafer-on-substrate), InFO (integrated fan-out), and FOWLP (fan-out wafer-level package), all of which are in production. Unlike older generations of multi-chip assembly (Figure 1.2.18), these new technologies have the potential to integrate thousands of high-speed IO's in the same package. Such

high density is achieved by utilizing wafer-level fine-line lithography to form high-quality wires which interconnects smaller chips on top of the silicon interposer. Micro-bumps are used to connect the smaller chips to the interposer. For CoWoS® technology, which is widely used for high performance computing (HPC) products to integrate a SoC and high bandwidth memory (HBM), a 40μm bump pitch and 0.4μm/0.4μm routing layer width/space is standard. A more recent package innovation is a variant of FOWLP [14], which was first pioneered by MediaTek for a 6.4Tb/s network application switch chip now in mass production. Two die are connected via several thousand I/Os to operate as one bigger SoC.

The chiplet concept addresses several issues related to the slow-down of process scaling. One of them is that yield for several small dies is much better than for a single very large die. Another issue addressed is that in the era of complex system-on-chips, it makes more economic sense to break up distinct functions such as CPU, memory, analog, RF, etc. into separate chiplets, which can then be fabricated in the most suitable process nodes and can be more readily re-used. The challenge lies in how to efficiently implement chip-to-chip communication. There are efforts underway to create standards by which the inter-chiplet communication can be done smoothly. There is also an ongoing effort to standardize XSR SerDes, targeting a density of 1Tb/s/mm at 1 to ~1.5pJ/bit energy consumption. Beyond the current SoA 2.5D technologies lies true 3D-IC, where dies will be fabricated or stacked on top of each other. They can be connected through silicon or package, and the distinction between on-die and die-to-die interconnect becomes pointless. The next decade will surely have more new innovations in this area.

4.2 Process Technology

Although process scaling has slowed down and might stall sometime in the next decade, for now the pace of scaling continues unabated. For decades, the industry has kept to the cadence of 50% area shrinking, 30% power reduction, and 20% speed boosting every 18 months as predicted by Moore's Law. Although leakage current threatened to derail this progress a number of years ago, the introduction of FinFET devices has kept the trend on track. However, as the dynamic power continues increasing as shown in Figure 1.2.17 [15], the overall power consumption and thermal management become a nightmare. In our analysis, power density increases 1.7× with every new technology node and is now approaching 3.7W/mm^2. Due to the severe thermal dissipation problems, it is necessary to improve energy efficiency not only for battery operated products, but also for HPC and similar products.

While soon-to-be-introduced to mass production devices like gate all-around (GAA) will help alleviate the power density issue, optimizing circuits for low-voltage operation remains a critical tool for energy efficiency, particularly for leaf devices. Techniques such as sub-threshold operation do help to reduce the system energy consumption, but are susceptible to device variability resulting in a limited maximum operating speed. Hence, more innovation will be needed in the coming years to surmount the issues.

5.0 Summary

IoT has been expanding at a fast pace into every field, from personal to industrial to even military applications. The transformation from IoT to cognitive AIoT requires dealing with big data which is collected, transported, learned, and inferred to activate the interaction with the real world. A variety of roots technologies such as heterogeneous computing, 5G wireless communication, high-speed wireline communication, and advanced packages are needed. The trends of those fundamental technologies and corresponding state-of-the-art performance have been presented. They form three pillars (or trunks) to support all AIoT applications. The 10-year outlook has also been discussed in terms of the challenges and potential directions of development. Current technology is able to ramp up AIoT from roots to leaves. Yet, major innovation is still needed to well position the semiconductor industry to orchestrate over 350 billion connected intelligent devices in year 2030.

Acknowledgements:
The author gratefully thanks the contributions and supports from Central Engineering Group (CEG), Computing and Artificial Intelligence Technology Group (CAI), Wireless Technology Group (WTG), Intelligent Devices Business Group (IDBG), Process Technology & Manufacturing Operations (PTMFO), Central Design Group (CDG), Wireless Communications Business Group (WCP), and Wireless Business Group (WSD) at MediaTek.

References:

[1] K Simonyan and A Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[2] J. Deng, et al., "ImageNet: A large-scale hierarchical image database," *in Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)*, 2009.
[3] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, 2016.
[4] P. Goyal, et al., "Accurate, large minibatch SGD: Training ImageNet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.
[5] T. Akiba, S. Suzuki, and K. Fukuda, "Extremely large minibatch SGD: Training ResNet-50 on ImageNet in 15 minutes," *arXiv preprint arXiv:1711.04325*, 2017.
[6] C.-H. Lin, et al., "A 3.4-to-13.3TOPS/W 3.6TOPS dual-core deep-learning accelerator for versatile AI applications in 7nm 5G smartphone SoC", *ISSCC Dig. Tech. Papers*, pp. 134-135, Feb. 2020.
[7] Cisco Global Cloud Index, 2016-2021.
[8] https://www.gsma.com/futurenetworks/wiki/ai-automation-an-overview/
[9] M. Ding, et al., "A 0.8V 0.8mm2 bluetooth 5/BLE digital-intensive transceiver with a 2.3mW phase-tracking RX utilizing a hybrid loop filter for interference resilience in 40nm CMOS," *ISSCC Dig. Tech. Papers*, pp. 446-448, Feb. 2018.
[10] M.-D. Tsai, et al., "A 12nm CMOS RF transceiver supporting 4G/5G UL MIMO", *ISSCC Dig. Tech. Papers*, pp. 176-177, Feb. 2020.
[11] E. Lu, et al., "A 4x4 dual-band dual-concurrent WiFi6 802.11ax transceiver with integrated LNA, PA and T/R switch achieving +20dBm 1024QAM MCS11 Pout and -43dB EVM floor in 55nm CMOS," *ISSCC Dig. Tech. Papers*, pp. 178-179, Feb. 2020.
[12] F. O'Mahony, "Wireline – 2019 trend," *ISSCC 2019 Press Kit*, pp. 102-105.
[13] T. Ali, et al., "A 460mW 112Gbps DSP-based transceiver with 38dB loss compensation for next generation data centers in 7nm FinFET technology," *ISSCC Dig. Tech. Papers*, pp. 118-119, Feb. 2020.
[14] N.-C. Chen et al., "A novel system in package with Fan-out WLP for high speed SERDES application," *in Proc. IEEE 66th Electronic Components and Technology Conference (ECTC)*, pp. 1495–1501, May 2016.
[15] E. J. Nowak, "Maintaining the benefit of CMOS scaling when scaling bogs down," *IBM Journal of R&D*, vol. 46, pp. 169-180, Mar./May 2002.
[16] H. Mair, et al., "A 7nm FinFET 2.5GHz / 2.0GHz dual-gear, octa-core CPU subsystem with power/performance enhancements for a fully integrated 5G smartphone SoC," *ISSCC Dig. Tech. Papers*, pp. 50-51, Feb. 2020.
[17] M.-D. Tsai, et al., "A multi-band inductor-less SAW-less 2G/3G-TD-SCDMA cellular receiver in 40nm CMOS," *ISSCC Dig. Tech. Papers*, pp. 354-355, Feb. 2014.
[18] C.-S. Chiu, et al., "A 40nm low-power transceiver for LTE-A Carrier Aggregation," *ISSCC Dig. Tech. Papers*, pp. 130-131, Feb. 2017.
[19] T.-M. Chen, et al., "An 802.11ac dual-band reconfigurable transceiver supporting up to four VHT80 spatial streams with 116fsrms-jitter frequency synthesizer and integrated LNA/PA delivering 256QAM 19dBm per stream achieving 1.733Gb/s PHY rate," *ISSCC Dig. Tech. Papers*, pp. 126-127, Feb. 2017.

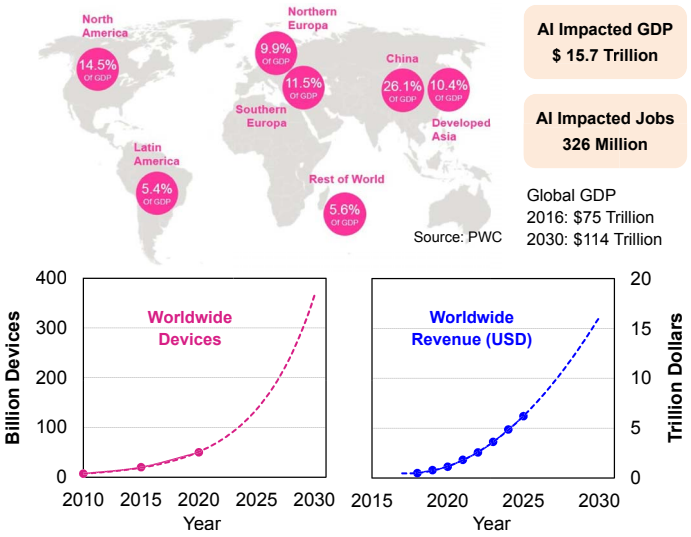

Figure 1.2.1: Impact of Artificial Intelligence in 2030.

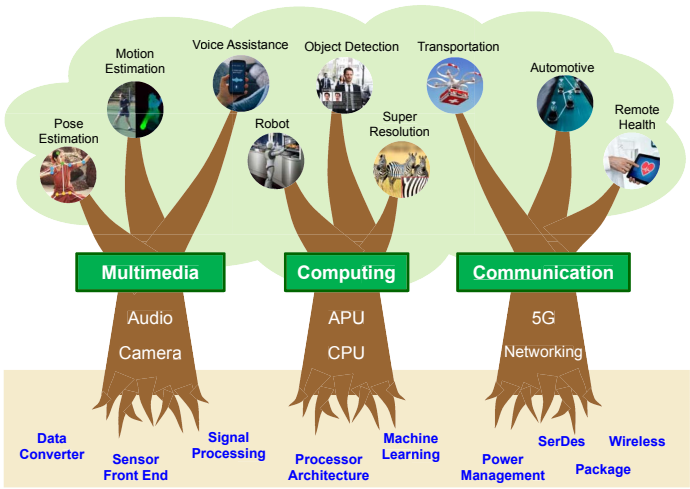

Figure 1.2.2: Empowering AI from root-technologies to leaf-applications.

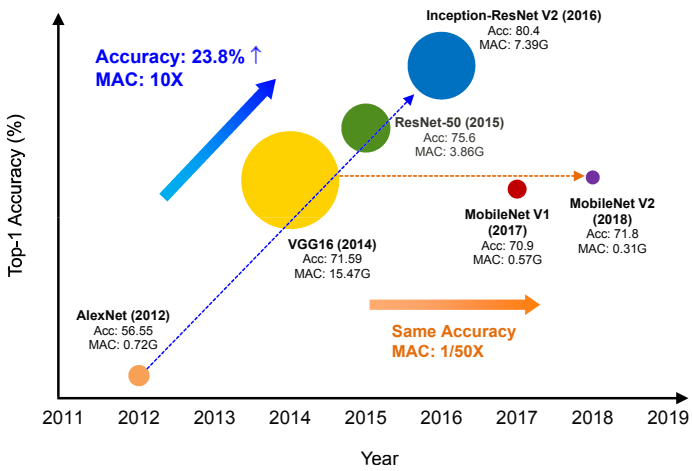

Figure 1.2.3: Evolution of deep neural network architecture.

Vision	Image classification	Object detection	Depth estimation	Image segmentation	Noise reduction	Super resolution
	Perception		Construction		Pixel Quality	
Computation	0.01-8 TOPs		0.1-10 TOPs		0.5-40 TOPs	
Precision	INT8-FP16		INT8-FP16		INT12-FP16	
Memory Bandwidth	~2GB/sec (@100FPS)		~8GB/sec (@30FPS)		>15GB/sec (@2FPS)	

Voice	Speech Recognition	Machine Translation	Natural Language Understanding	Text to Speech
Computation	0.1-1 GOPs	0.8-4 GOPs	2.5-50 GOPs	
Precision	INT8-FP16	FP16	FP16	
Memory Bandwidth	~1GB/sec	~15GB/sec	>20GB/sec	

Figure 1.2.4: Performance requirements for AI vision and AI voice applications.

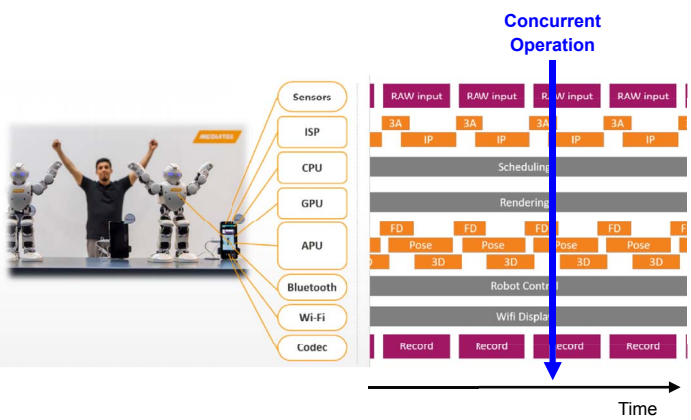

Figure 1.2.5: Concurrent heterogeneous operations in an AI SoC for tracking movements in real-time.

	CPU	GPU	DSP	APU
Main Focus	Control Serial Computing	Graphics Parallel Computing	Signal Processing	Special Purpose
Strength	Low Latency	High Throughput	Power Efficient	Area Efficient Power Efficient
Functionality	CV & DL	CV & DL	CV & DL	DL Only
Flexibility	High	←		Low
Power Efficiency	Low	→		High

Figure 1.2.6: Benchmarks for AI computing with different types of processors.

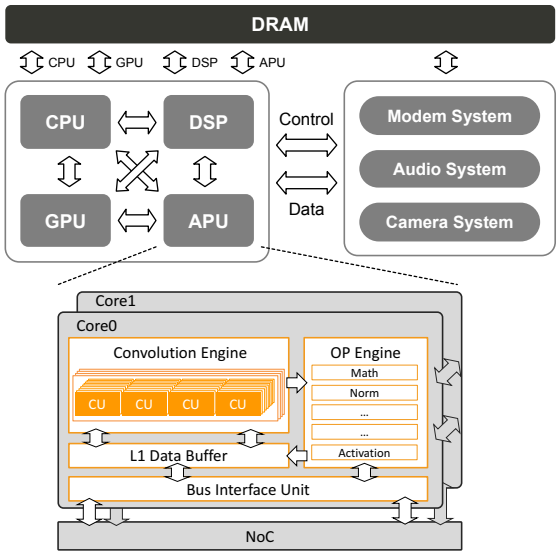

Figure 1.2.7: Tightly coupled system and APU architecture.

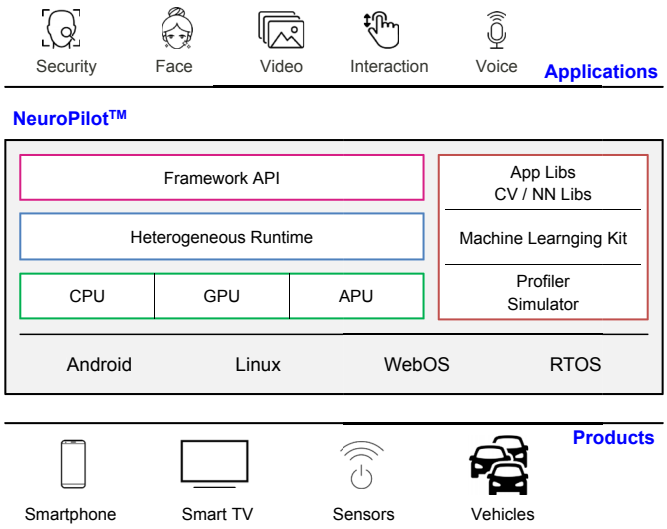

Figure 1.2.8: NeuroPilot™ platform.

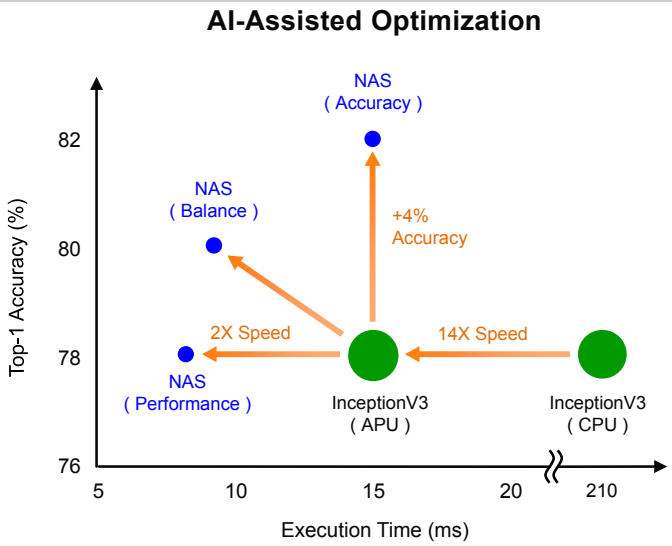

Figure 1.2.9: AI-assisted neural architecture search (NAS) and optimization.

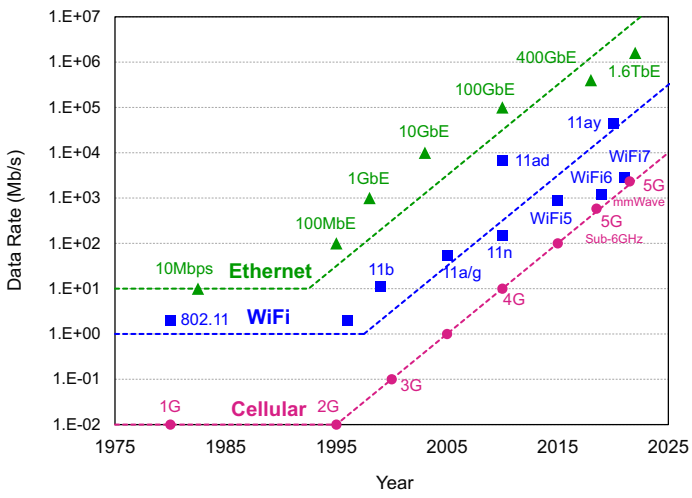

Figure 1.2.10: Evolution of wireless communication standards and data rates.

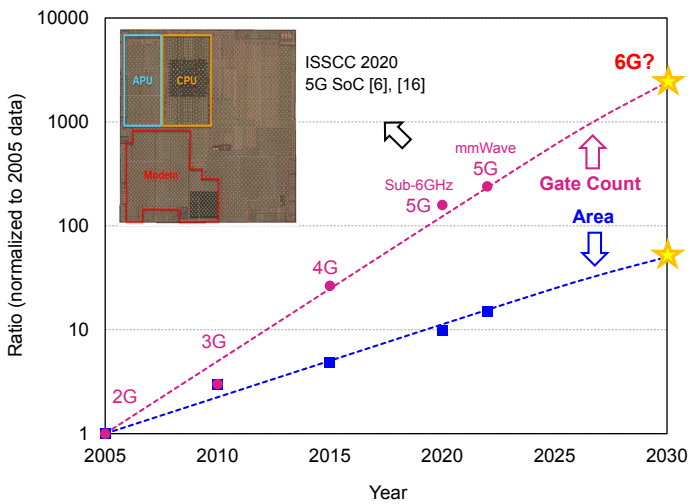

Figure 1.2.11: Gate count and chip area of modems for cellular communication.

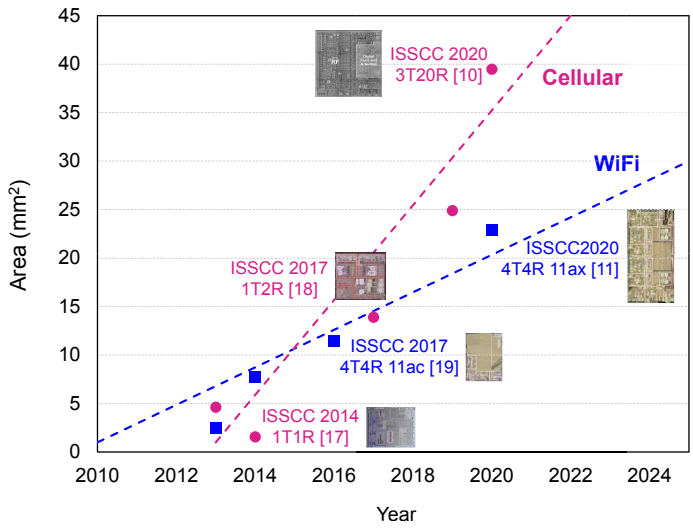

Figure 1.2.12: Complexity of wireless transceivers increases over time.

978-1-7281-3205-1/20/$31.00 ©2020 IEEE

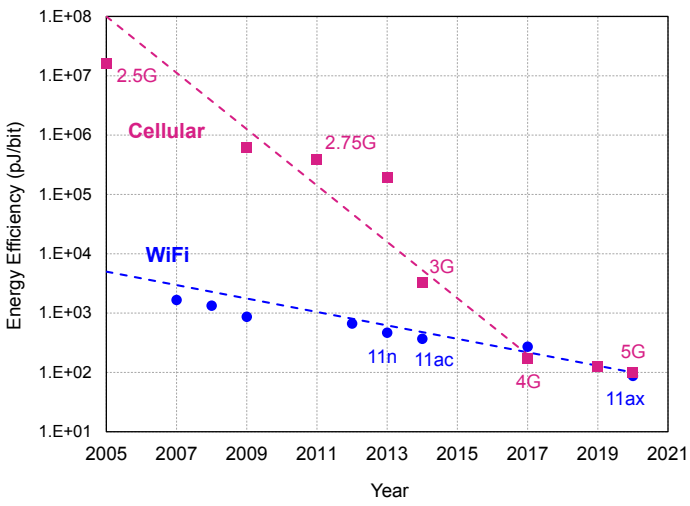

Figure 1.2.13: Improvement of wireless receiver energy efficiency.

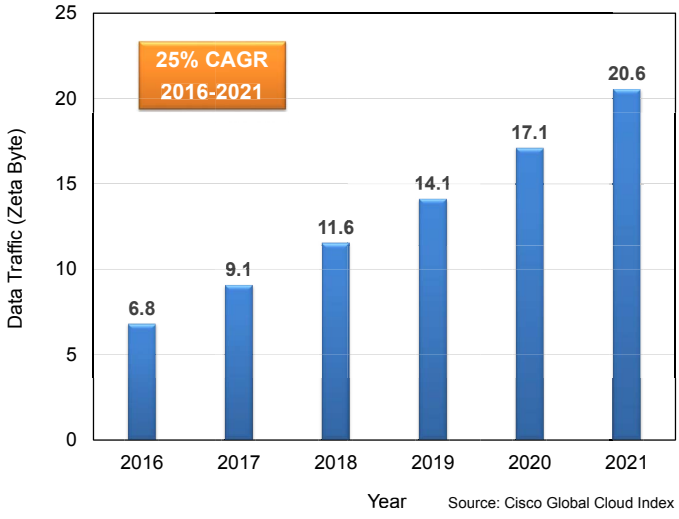

Figure 1.2.14: Data traffic growth triggered by billions of AIoT devices.

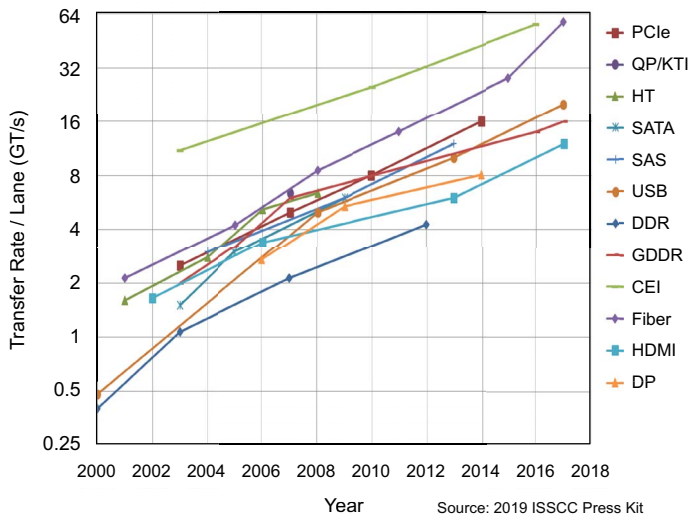

Figure 1.2.15: Evolution of wireline data rates for multiple standards.

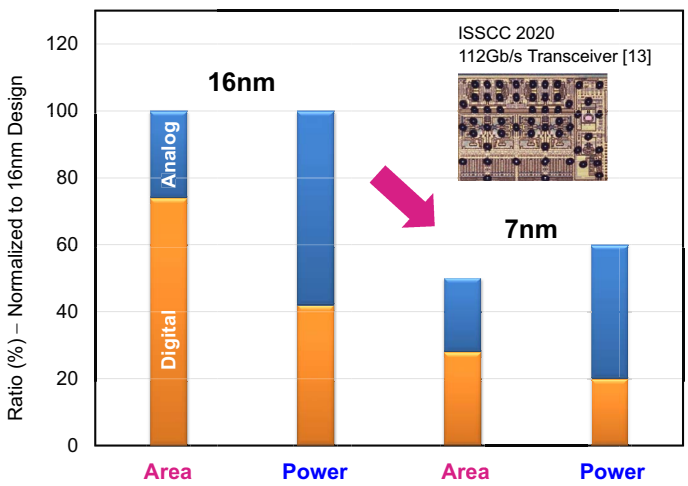

Figure 1.2.16: Power and area scaling in DSP-based SerDes architecture.

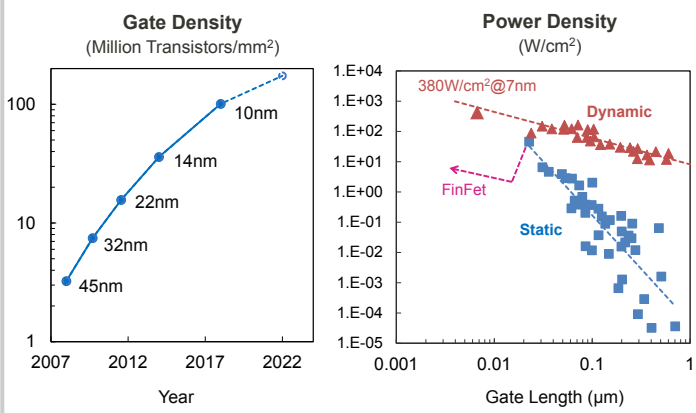

Figure 1.2.17: Advantages of process technology scaling and growth of power density.

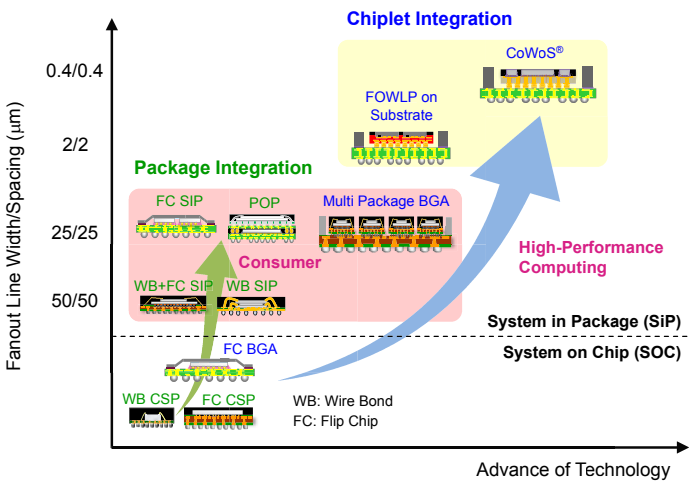

Figure 1.2.18: Package technologies for homogeneous and heterogeneous chip integration.