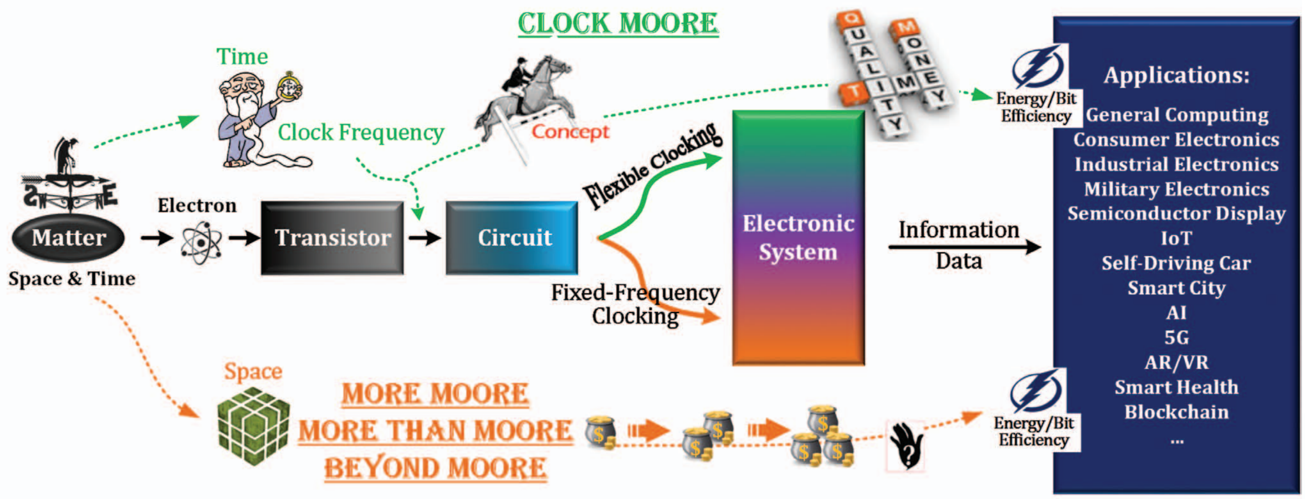
Liming Xiu

Time Moore



Exploiting Moore's law from the perspective of time

Moore's law has served as a goal for the semiconductor industry for more than 50 years. After decades of relentlessly racing forward, convincing new evidence now shows that the end of conventional Moore's law scaling is really near. Besides the More Moore, More Than Moore, and Beyond Moore approaches, which concentrate most of their efforts on squeezing processing power from the perspective of space, it is useful to inspect Moore's law from the time perspective since matter exists in both space and time. In microelectronics, time is reflected as clock frequency. Clocking is another direction to exploit Moore's law in greater depth: we will "play with frequency," which is a "forgotten" opportunity that remains open for a potentially significant profit. This article investigates a more sophisticated strategy to manipulate frequency for clocking an electronic system. The aim is to obtain higher compute energy efficiency from given silicon real estate, not by packing more transistors into it, not by blindly increasing clock speed, but by adjusting clock frequency dynamically wherever and whenever possible. This approach is termed *Clock Moore*. Furthermore, we propose an idea of using the rate of switching as a computational variable to encode information, which is termed *Rate Moore*. Together, Clock Moore and Rate Moore make up "Time Moore," which enhances transistors' collective computing capability by using time more efficiently.

Digital Object Identifier 10.1109/MSSC.2018.2882285
Date of publication: 6 February 2019

Moore's Law

Brief Review

Moore's law is a symbol for technology innovation, and its terminology is rooted in the semiconductor industry. In 1965, Gordon Moore observed that the number of components per IC doubles every year [1]. In 1975, the forecast was revised to doubling every two years [2]. The period is often quoted as 18 months when the combined effect of more transistors and transistors being faster is considered. Although Moore's law is an observation or projection, not a physical or natural law, its predictions have proven to be rather accurate for over five decades. It has been used in the semiconductor industry to guide long-term planning and to set targets for research and development [3]–[5]. Advancement in world economic growth in past decades is strongly linked to Moore's law; it describes a driving force of technological change, social change, productivity increase, and economic growth [6]–[8].

Moore's law has been reinvented multiple times. In the 1970s and 1980s of Moore's Law 1.0, the focus was on scaling up the number of components on a single chip. Higher transistor densities allow more functions to be integrated into a single CPU die. Moore's Law 2.0 came into play in the mid-1990s. It

Moore's law describes a driving force of technological change, social change, productivity increase, and economic growth.

focused on scaling performance up by increasing clock speed relentlessly. But clock speed stabilized around 2000 because faster speeds induce more heat dissipation than a chip could withstand. Moore's Law 3.0 focuses on integrating other functional components. Initially, this means additional CPU cores. Later, as the system on chip (SoC) became common, it included features like onboard graphics processing units (GPUs), cellular and Wi-Fi radios, PCI Express lanes, and so on. Moore's Law 1.0 provided the mainframe computer and the minicomputer, while Moore's Law 2.0 offered microcomputers in both desktop and laptop incarnations. Moore's Law 3.0 is giving us technologies such as smartphones, tablets, the wearable industry, the self-driving car, and renewing artificial intelligence (AI).

Key inventions that made Moore's law possible are the IC in 1958; CMOS in 1963, dynamic random-access memory (DRAM) technology in 1967; and flash memory, chemically amplified photoresist, and deep ultraviolet excimer laser photolithography in the 1980s. Moreover, the interconnect innovation of the late 1990s is another enabling factor, including chemical and mechanical polishing/planarization, trench isolation, and copper interconnect. Although interconnect is not a direct factor in creating a smaller transistor, it enables improved wafer yield, additional layers for metal wiring, closer spacing of devices, and lower electrical resistance.

The force of Moore's law is now able to produce a monster chip with 21 billion transistors onboard: Nvidia's Volta GV100 GPU in 2017 [9]. However, when the wire and gate get too small, electrons begin to stray from their dictated paths and short-circuit the chip, which makes shrinking transistors further a futile endeavor. We are now hitting the brick wall of physics: transistors are reaching their atomic size limit, and Moore's law is breaking down [10], [11]. Since the 1990s, the semiconductor industry has released the research road map International Technology Roadmap for Semiconductors (ITRS) [3], [4] every two years to coordinate the industry-wide effort for moving forward. In 2017, the update on ITRS stopped, a clear sign that the end of conventional CMOS transistor scaling is near.

Three Paths: More Moore, More Than Moore, and Beyond Moore

Moore's law is not about semiconductors, computers, performance, physics, or electronics. It is mostly about economics, and it is ending in a literal sense because the exponential growth in transistor count cannot continue forever. However, from the consumer's perspective, Moore's law simply means "user value doubles every two years." In this form, the law will continue as long as the industry can keep stuffing its devices with new functionalities. This continuing effort around Moore's law can be captured in three phrases: *More Moore, More Than Moore,* and *Beyond Moore*, each is accompanied by a profound insight in its way of pursuing the "doubled-value every two years."

More Moore is the strategy of continually scaling the transistor down. It evolves from constant-field scaling to constant-voltage scaling to equivalent scaling. Miniaturization is its characteristic. The issues involved are: lithography, power supply and threshold voltage, short-channel effect, gate oxide, high-field effect, dopant number fluctuation, and interconnect delay [12], [13]. In the early stage of Moore's law, scaling transistors down also improves speed and reduces energy consumption, which is known as *Dennard scaling* [14], [15]. While Moore's law states that more transistors could be packed into the same area from generation to generation, Dennard scaling ensures that each individual transistor in a new generation would be cooler and draw less power. These triple benefits led to the rise of affordable PCs in the 1980s. Moore's miniaturization and Dennard's scaling are artificially tied together. However, the breakdown of Dennard scaling in the mid-2000s stopped the continuous trend of clock speed increase. The heat-removal problem has prevented clock-based scaling. Instead, compute capability is enhanced by adding more CPU cores (the multicores architecture) and improving single-threaded CPU performance. Moore's law hence continued once more from 2005 to 2014: transistors' speed gain might not be greater than that of their predecessors but they were more power efficient and less expensive to build; chips might have more transistors on board but not all of them are able to be turned on simultaneously (dark silicon) [16], [17]. In the meantime, semiconductor manufacturers have innovated with technologies such as strained silicon, hi-k metal gate, fin field-effect transistor, and fully depleted silicon on insulator. However, none of them have re-enabled the continuous geometrical scaling that Dennard scaling offers.

The More Than Moore strategy is next. It takes the challenge from the other direction: rather than making the chip better and letting the application follow, it begins with application. From smartphones and supercomputers to data centers in the cloud, it works downward to see what chips are needed to support them. The idea of More Than Moore is not to focus solely on the computing power of a single chip but also to observe the efficiency of the whole system from a higher perspective. It encourages functional diversification, which refers to integrating functionalities that do not necessarily scale according to Moore's law but provide additional value to the end application in different ways. It changes from a single technology transition to the integration of various technologies. Moore's law was initially proposed and verified in the development of the logic and memory circuits. More Than Moore examines the opportunity of integrating myriad functions at the system level, which typically includes nondigital functionalities such as analog, radio frequency, sensor, actuator, embedded DRAM, microelectromechanical systems, high-voltage circuit, power control, and passive components. From new types of transistor structures and process compatibility of various types of circuits to advanced packaging technologies, More Than Moore improves the overall integration efficiency, makes a system capable of supporting more functions, and, at the same time, reduces the overall system cost. In essence, it evolves from the "cheaper, better, faster" of More Moore to "better and more comprehensive" [3], [4].

Currently, we are reaching the limit of silicon-based CMOS. The fundamental physical size limit of an atom will cause a hard stop (the gap between two silicon atoms is approximately 0.5 nm and the width of nine silicon crystal unit cells is essentially 5 nm). New technologies such as multiple patterning, immersion lithography, and 3D tri-gate transistors can probably support chips with a 5–7-nm process. Therefore, what happens next, when the quantum effect comes into play and continued scaling is no longer possible?

This is the domain of Beyond Moore. One option is to use 3D, which is an architectural approach: continue to use silicon but configure it in a new way [18]. Rather than simply etching flat circuits onto the surface of a silicon wafer, we can stack many thin layers of silicon with circuitry etched into each of them. In principle, this should make it possible to pack more computational power into the same space. In practice, however, it works only with memory chips that do not have serious heat problem (it consumes power only when a memory cell is accessed, which is not often). On the other hand, there are several prospects to replace the CMOS transistor. Many of these alternative devices operate on state variables other than charge, and some of them may offer functionalities beyond those binary devices, which could be useful for operations that are more complex. Examples of such novel devices include novel materials such as an FET with III-V, Ge, carbon nanotubes, and graphene; SpinFET; spin-torque; spin-wave; tunneling transistor; piezoelectric transistor; molecular switch; nanoelectromechanical systems; and thermal transistor. Beyond Moore can also include ideas of biologically inspired ways to compute (neuromorphic computing, which aims to model processing elements on neurons in the brain), approximate computing, and superconducting computing [3], [4], [19]–[23].

From the Perspective of Time

Figure 1 shows the 40-year trend for microprocessor development [24]. From the 1970s to now, the number of onboard transistors grew at an exponential rate. In the mid-2000s, Dennard scaling failed and, as a result, clock rate stops increased. The top speed stabilizes at approximately 3–4 GHz, and power consumption peaks in the range of a few hundred watts. Clock frequency reaches its limit, which forces us to investigate a question. Besides the fixed-frequency-clocking design strategy, which has dominated the IC design community for decades, are there other more creative ways to use clock frequency? Is it beneficial to use dynamic-frequency-clocking in applications, wherever and whenever possible? It is fair to say that More Moore, More Than Moore, and Beyond Moore all focus on space: getting more compute capability from ever-smaller space and using ever-less matter. Play-with-frequency is a strategy to obtain more computing power from the perspective of time. Can more juice be squeezed out from transistors by marching along this new path? This is called *Clock Moore.*

During the past five decades of semiconductor industry growth, the mainstream computational paradigm is to use electric charges (electrons' movement, collectively represented as voltage and/or current) for encoding information. Using charges requires matter and subsequently requires space: the amount of charge is virtually proportional to the size of space. As Moore's law is running out of steam, is it possible to use time for encoding information? This is called *Rate Moore*. Figure 2 illustrates three directions for future growth [23]: 1) create new devices, 2) build new architectures with or without new devices, and 3) develop new computational paradigms. The roles of Clock Moore and Rate Moore are marked on this map. They will be discussed in the sections "Clock Moore" and "Rate Moore," respectively.

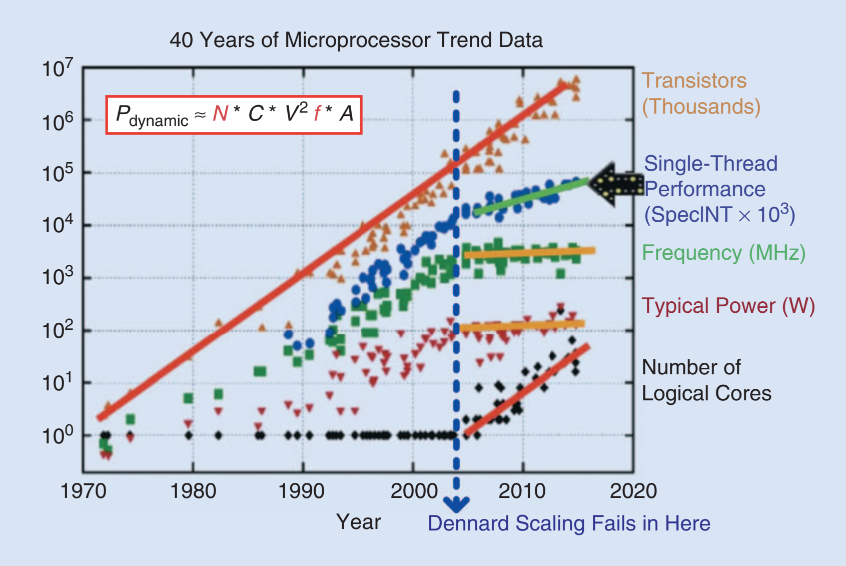


FIGURE 1: The Dennard scaling failed around the middle of the 2000s [24].

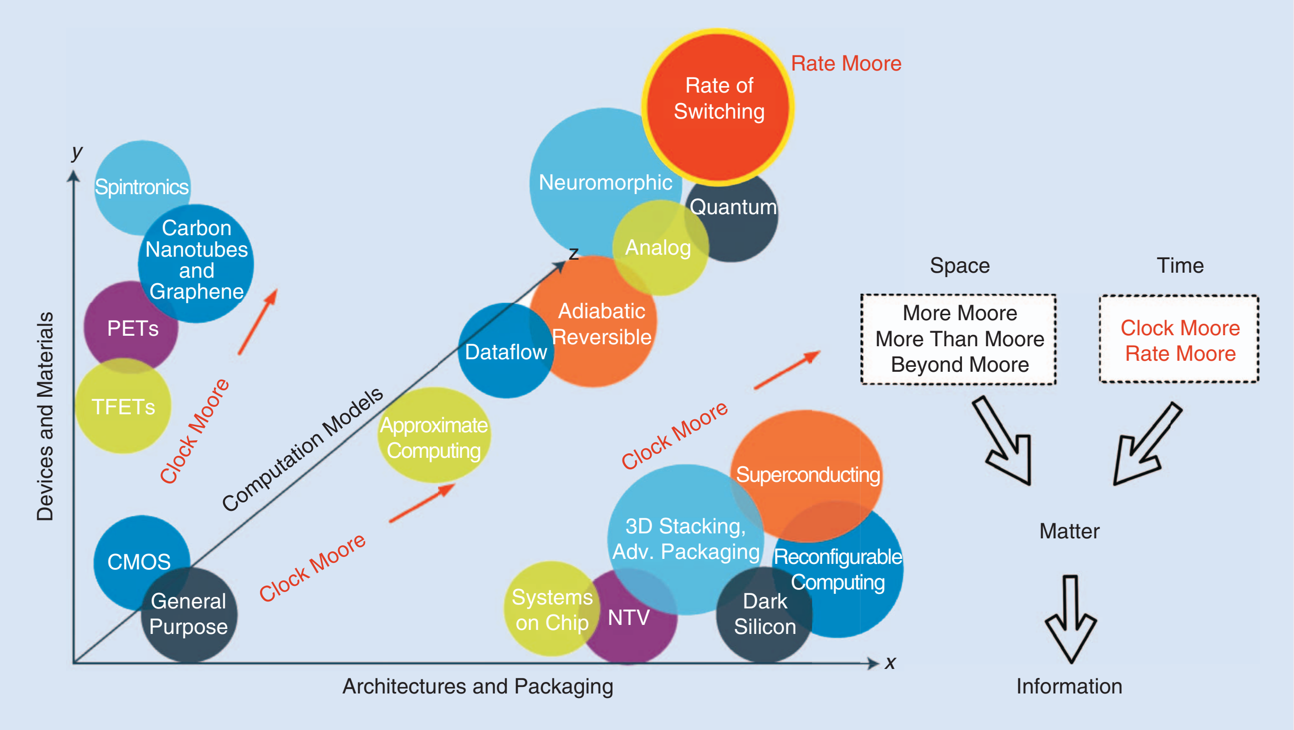


FIGURE 2: The roles of Clock Moore and Rate Moore on microelectronics' future growth [23]. PETs: piezoelectric transistors; TFETs: tunneling field-effect transistors; Adv.: advanced; NTV: near-threshold voltage.

Clock Moore

Space and Time: The Real Estate in Microelectronics

In the physical world, everything exists in the 4D frame of space and time, and microelectronic devices are no exception. The microelectronics skyscraper can be viewed as being built from layers. On the very top is the application layer, which includes items such as communication, self-driving cars, AI, big data, industrial control, smart health, display, smart city, artificial reality/virtual reality, desktop and mobile computing, and blockchain applications. The task of processing information, in the form of digital data, is in the center of all those applications. Below the application layer is the task layer, which has four subtasks: generating, processing, moving, and using data. Below the task layer is the data layer, where data is the product of ICs. The two basic variables that an IC designer can use when designing his or her chip are voltage/current and clock frequency. Voltage and current are created from matter (electrons) and their levels (or strengths) are used to encode information. In microelectronics, the level is proportional to the number of electrons involved in the process (i.e., the movement of electrons is collectively represented as the voltage or current level). On the other front, clock frequency indicates the passage of time and controls the speed of all actions. Matter exists in both space and time so the two fundamental cornerstones of our universe, space and time, are central for microelectronics. When we create devices for computing, both space and time are real estates that are usable and valuable. In past practice, we have focused heavily on the space side of the story. Now it is time to pay more attention on the time side.

Play With Voltage: Electrical Signal and Silicon Space

Since the beginning of microelectronics, electrical voltage (or current) has been used to encode information. Its strength, or level, represents the exact meaning of the intended information. Level versus time, S(t), is defined as electrical signal. In analog design methodology, every point in the level scale possesses its divine value (the ability to differentiate neighbor points is only limited by noise). The state-of-the-art analog-to-digital converter can achieve 24-b resolution that could distinguish voltage level in a few microvolts' range. In digital design, we only care about two levels: high and low (1-b resolution). Microelectronics relies on this principle of level being the information differentiator. In physical implementation, level is created from a plurality of electrons, which, in turn, is proportional to the amount of matter used in the process. Matter occupies space, so an electrical signal uses silicon space. For a given signal standard and given material, a large signal requires more silicon. The continuous technology transition from old to new generation is the effort of enhancing material property and improving signal standard so that a smaller signal in a new generation (less electrons, less energy, less matter, less space) can be used to represent the same amount of information as a larger signal in previous generations. The entire history of Moore's law is virtually to exploit space, to squeeze more information processing power from given space: building better and smaller switches (i.e., transistors), stuffing more transistors in a 2D surface and then packing even more transistors by going vertical. In short, we have used space well.

> ***The entire history of Moore's law is virtually to exploit space, to squeeze more information processing power from given space.***

The Difficulty in Dealing With Time

Time is a very special entity that people cannot directly sense with their five senses. They can feel the passage of time only indirectly, with the help of some natural phenomena or physical mechanisms. Such mechanisms include the movement of the sun by using a sundial, the flow of water or sand by using a water or sand clock, and cyclical movement generated by a mechanical device such as with a mechanical clock or mechanical vibrations when using a crystal oscillator. Other mechanisms include periodic oscillation measured by electrical pulse and the transition between energy states of a certain element by using an atomic clock [25], [26]. Inside electronic devices, flow of time is established through electrical oscillation in the form of a chain of electrical pulses. This chain of pulses is termed *clock signal* (*clock* for short), as illustrated in Figure 3. Since voltage/current is the only media for which the circuit designer has direct control, the sense of time has to be created from it. As depicted in Figure 3(b), a moment is generated from voltage transition, and flow of time is created from indexing each of those moments. Compared to the level that is used to directly differentiate information, time is a secondary product induced from levels. For this reason, it is inherently difficult to deal with time, and it is, hence, difficult to make precise, accurate, and stable timing control devices. In most cases, the electrical pulse train of the clock signal has to be periodically calibrated against a better timing source, such as a crystal oscillator [27], [28].

The clock pulse train is a marker system for marking events and subsequently establishing the sequence of order for computing activities. Frequency is used to gauge the speed of the clock pulse train. In Figure 3, the moments $t_0, t_1, t_2, \ldots$ are used as index points to mark functional events. The quality of the clock pulse train is highly dependent on the accuracy of the locations of those moments. It requires those moments to occur exactly at their designated locations as deviation from the designated value can lead to loss of operating margin for functional circuit (digital design) or degraded accuracy for signal processing (analog design). The degree of this deviation is quantitatively described by jitter (time domain) or phase noise (spectrum domain) [29], [30]. Therefore, the key requirement for clock quality is precision: all the moments must position as closely as possible to their designated locations.

Two Long-Lasting Problems in Manipulating Clock Frequency

For the past several decades of circuit design practice, clock signals have been mostly used in the form of a fixed frequency with high-frequency stability. For a given application, a clock generator is only required to generate a few select frequencies. Furthermore, fast switching between frequencies is not considered to be a high priority for design. Under this scenario, the clock signal is conveniently created as a pulse train made of identical pulses. In other words, the value of all the pulses' lengths is constant, and all the spans between the moments are of equal length. For implementation, this kind of clock signal is most suitable for high-precision (low jitter, pure spectrum). It is also structurally beautiful since all the pulses are identical. However, it is rigorous (or rigid) for this type of clock to be used in practice since it is difficult to accommodate a large variety of values with the pulse's length, and it is difficult to change the pulse length from one value to another quickly. For a future complex system of a dynamic nature, this rigid clock is no longer adequate. We want a clock pulse train wherein

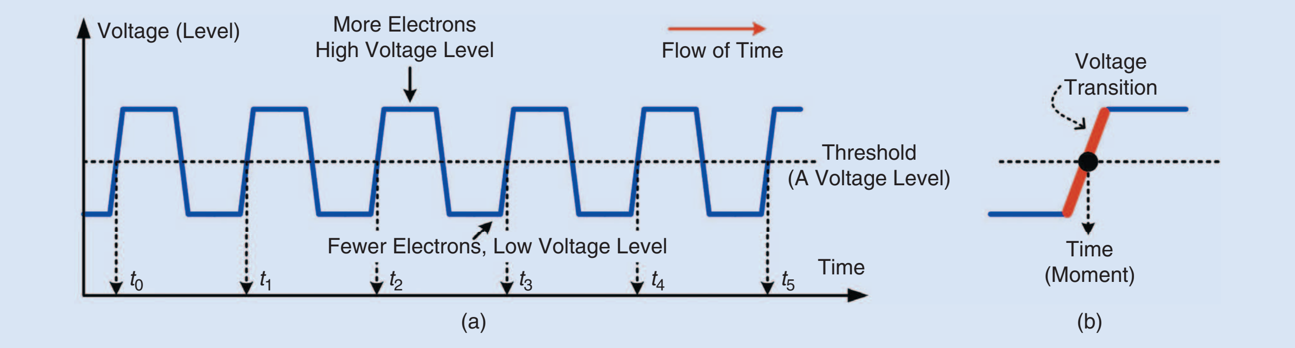


FIGURE 3: Time is created from the voltage transition.

This rebirth of the clock frequency concept can free us to attack the problem of arbitrary frequency generation.

the moments of $t_0, t_1, t_2, \ldots$ can be dynamically adjusted in operation: the span can be created at any length we want, and the span can be changed from one value to another quickly. Our aim is to control the functional events' sequence of order with greater freedom so that compute-power efficiency can be improved. This type of clock signal with such flexibility is termed *flexible clock*. It is more suitable for a modern system in which clock flexibility has a higher priority than clock precision.

There are two reasons that a flexible clocking style has not been pursued seriously before. One is that previous application requirements are relatively simple, and the fixed-frequency clocking strategy is able to handle most tasks (there were also other more demanding requirements than clocking). However, the more profound reason is that time (frequency) is indirectly created from voltage. None of the four foundational elements (resistor, capacitor, inductor, and mersister) recognize the concept of time, and they only respond to electrons' movement. This difficulty can be symbolized by these two challenges: arbitrary frequency generation and instantaneous frequency switching. This is the counterpart problem in voltage domain, arbitrary voltage generation, and instantaneous voltage switching, which have been addressed thoroughly. To be precise, arbitrary frequency generation refers to the demand of "frequency being adjusted in very fine step (granularity in parts per billion range)" and instantaneous frequency switching is "the switching of clock frequency being accomplished in a rapid and quantifiable fashion (one or two cycles)" [52]. Moreover, for any given design, these two features have to be achieved simultaneously. The widely used circuit for generating clock signal is the phase-locked loop (PLL) [31]–[33]. However, it is not a solution to these challenges due to its use of the compare-then-correct feedback mechanism. Thus, these two issues lurked in the road ahead, as illustrated in Figure 4.

Re-Investigating the Concept of Clock Frequency

When the pulse train of Figure 3 is inspected, it appears that arbitrary frequency generation refers to the capability of creating pulses with a size of any length in time that could be demanded by users. This, as discussed, is a difficult task. Is this the only way to fulfill users' demands? To investigate this problem at a higher level, we have to look beyond the locality of each pulse. In past practice, one way commonly adopted by people for defining frequency is to use the inverse of the length of the span between the two moments that makes up a particular pulse (the so-called instantaneous frequency). In this view, it implies that all the pulses in a clock pulse train must have the same length in time. This local-oriented view is straightforward but not the full picture. In a higher view, frequency is defined in a large frame of one second. It is the number of clock pulses occurring within the time window of one second when the clock signal itself is concerned. It is the number of operations executed within the time window of one second when functional operation is concerned. This broader view leads to a radical rethinking: the very constraint of "all cycles must have same length in time" could be removed. It is not an essential element in the definition of clock frequency but a convenience for implementation. This rebirth of the clock frequency concept can free us to attack the problem of arbitrary frequency generation. In 2008, a new concept, time average frequency (TAF), was introduced [34]. This concept removes the "equal-length" constraint and defines the clock frequency solely on activities that occurred in the time window of 1 s. Although TAF allows the use of multiple types of cycles in a clock pulse train that makes the occurrence of the moments seem "irregular" (refer

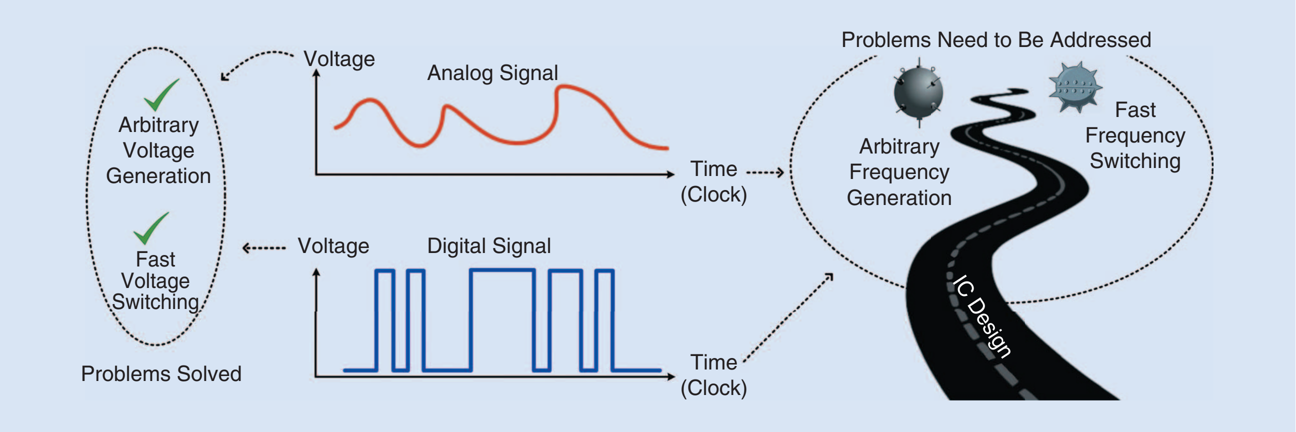


FIGURE 4: The two big remaining challenges in IC design.

to Figure 3), the precision of those moments' locations is not blemished since the locations are designated and known by their creator. Thus, TAF and jitter are two different concepts.

On the other front, to make instantaneous frequency switching possible, we have to abandon the compare-then-correct mechanism used in PLL. (Note: PLL is a beautiful blend of analog and digital circuits. It is one of the foundational components in IC design. It will surely stay in this field forever. The word "abandon" is not used in its literal sense.) One approach is to construct each individual pulse directly, the so-called direct period synthesis (DPS) method [35]–[50]. Together, from the joint effort of a new concept and a new method, a new discipline in frequency synthesis emerges: time-average frequency direct period synthesis time-average frequency direct period synthesis (TAF-DPS) [51]. Its aim is the aforementioned two challenges.

> **The circuit technique of TAF-DPS provides a chip architect with the tool of exploring frequency for higher information processing efficiency.**

Clock Moore: Play With Frequency

The circuit technique of TAF-DPS provides a chip architect with the tool of exploring frequency for higher information processing efficiency. This school of thought, termed *Clock Moore*, is illustrated in Figure 5. Moore's law (Dennard scaling) drives the transistor feature size smaller and makes the degree of integration higher. Consequently, the chip's processing power becomes stronger, and its energy footprint for processing each bit of information decreases. This achievement was accomplished under the fixed-frequency design philosophy. Hence, it is reasonable to state that, so far, the gain in computing power mostly comes from exploring space. As the potential of space is becoming exhausted, it is wise to check the other piece of real estate: time. Dynamic frequency design style, enabled by TAF-DPS, is the approach to explore time, to gain even higher processing efficiency from a given area of silicon [52]. We will briefly investigate its beneficial potentials from several key design concerns, such as data movement (accuracy), data movement (architecture), processor instruction set architecture, field-programmable gate array (FPGA) and frequency, Von Neumann bottleneck, network time synchronization, clock distribution, clock spectrum, pulse-width modulation (PWM), frequency for sensor design, and frequency as a software programmable variable.

Data Movement: Accuracy of Data Flow

Data processing is the centerpiece of all modern applications. In fact, the sole function of a chip is to process information, including the tasks of generating, processing, moving, and using data. All those tasks rely on one basic operation: transferring data between two places, symbolically labeled as Tx (transmitter) and Rx (receiver), respectively, in Figure 6(b). Frequency controls the data flow, which is a digital stream of ones and zeros. As shown, each module has its own clock with an associated clock frequency. Moreover, each module works in its own environment, including conditions of voltage, temperature, and loading. For a successful data transfer (i.e., no data loss or cycle slip), the data flows on both sides (controlled by clock frequencies) must be equal in average. For various levels of operations, the period for "averaging" is different. It could be in the range of a second, millisecond, microsecond, or even a nanosecond. The shorter this period of "achieving equal data flow on average" is, the higher the processing efficiency will be. This representative operation is so omnipresent in microelectronic design that it can be found everywhere: between small circuit blocks, between functional modules, between chips, and among networks [53]. As we can

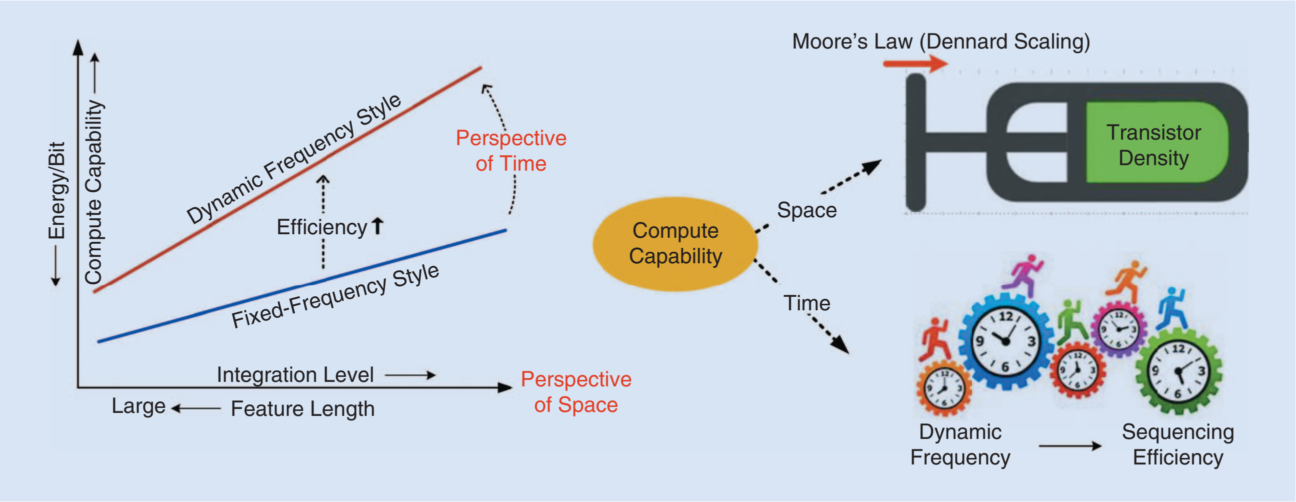


FIGURE 5: The exploitation of computing capability from space and from time.

see from Figure 6, frequency is the key in achieving this no-loss data transfer. The more flexible the clock sources CLK-T and CLK-R are, the more fluently they can adjust their frequencies to match each other as quickly as possible. This adaptiveness is supported by arbitrary frequency generation and instantaneous frequency switching at the circuit level and exploited by system architect at the system level.

Data Movement: Chip Architecture

When Dennard scaling failed in around 2004, computer architecture moved from a single CPU to a multicores structure. Processors of this style can contain many, even hundreds, of computing cores (e.g., Intel Xeon Phi family). IBM's TrueNorth is another example of a very large number of cores, but it uses neuron- and synapse-like structures instead of conventional cores. Under such a scenario, the issue of data exchange among the cores presents a sizeable opportunity for improving overall compute power efficiency. For best performance, each core runs at its own clock frequency. An interface network controlled by flexible clock generators is ideally suitable for handling the data transport problem (please refer to [52, sec. III.2]). In this case, clock frequency plays a key role.

Frequency Scalable Instruction Set Architecture

Dynamic voltage and frequency scaling (DVFS) is a technique for energy management [54]–[56]. However, with a PLL-based clock generator, in most cases this task can only be carried out in a coarse-grained fashion because the PLL output is sparse in frequency and its frequency switching is slow. TAF-DPS is an ideal solution to this challenge, thanks to its features of arbitrary frequency generation and instantaneous frequency switching. Using TAF-DPS, the DVFS can be implemented in a fine-grained style. The submicrosecond timescale power management, presented in [57], can be realized without much difficulty since TAF-DPS can switch its output frequency in two clock cycles. To improve the power efficiency further, it is beneficial that we play the frequency at the instruction set level. A frequency scalable instruction set architecture (FS-ISA) can methodically include frequency as a dimension in the design of an instruction set. For different operations (such as data handling and memory operations, arithmetic and logic operations, control flow operations, coprocessor instructions, and complex instructions), the clock speed will be adjusted accordingly, and this information can be part of the instruction. The idea of multicore architecture is to process more instructions per clock at a lower clock frequency. Its drawback is that application software requires parallelization to run on multiple cores simultaneously, but software applications vary greatly in the extent to which they can be easily parallelized. Furthermore, improving software is more costly than simply adopting the cheaper hardware delivered by new technology nodes. FS-ISA provides an option that might ease some of the problems involved and can be supported by TAF-DPS from the circuit level. In this case, for any instruction, the architect can select the accompanying clock period as one particular value out of a group of values for optimal performance. With this circuit-level support, ISA architect has one more freedom to balance the pipeline depth and the logic per clock cycle (usually measured in a unit of the number of FO4 inverters) and, hence, is better able to address the efficiency issue as discussed in [58].

Flexible Clock Source on FPGAs

When necessary, the flexible clock generator TAF-DPS can be implemented completely in the digital domain, as exemplified in [49]. This method allows for use in FPGA-based applications. Due to its flexibility, an FPGA has becoming an important asset in heterogeneous computing or reconfigurable computing. The TAF-DPS on an FPGA method can provide FPGA users with the features of fine frequency granularity and fast frequency switching at a very low cost. Furthermore, users can reprogram the clock generator's configuration whenever they feel that it is appropriate. This is "play with frequency" in its literal sense. For example, in the case of convolutional neural network acceleration, we often encounter a sparse network. Flexible and low-cost TAF-DPS clock sources could be a tool for improving performance or reducing power consumption since clock frequencies can be dynamically and geographically adjusted based on the sparsity of the network.

Memory-Driven Computing

The Internet of Things (IoT) and the big data scenario have produced data at an exponential rate. This seriously challenges our ability to turn such

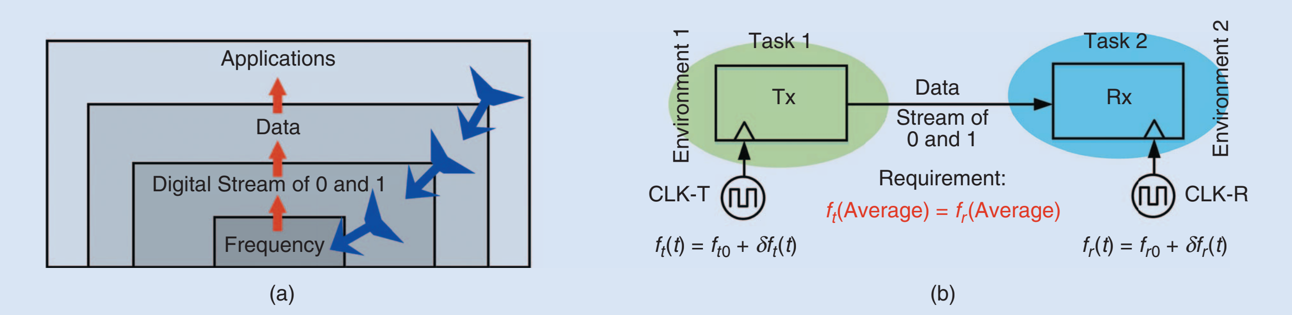

Applications
Data
Digital Stream of 0 and 1
Frequency
(a)
Environment 1
Task 1
Tx
Data
Stream of
0 and 1
Task 2
Rx
Environment 2
CLK-T
$f_t(t) = f_{t0} + \delta f_t(t)$
Requirement:
f_t(Average) = f_r(Average)
CLK-R
$f_r(t) = f_{r0} + \delta f_r(t)$
(b)

FIGURE 6: Frequency is the key in all applications and (a) and (b) frequency controls data movement. Tx: transistor; Rx: receiver; CLK: clock.

large amounts of data into useful information. Instead of traditional high-performance computing that requires a precision arithmetic to calculate an answer to a specific problem or enterprise computing devoted to performing a logic operation as a system of record for business transactions, today's dominant workloads often involve searching through terabytes of data to find something of significance. The metric to classify a computer's performance is changing from floating-point operations per second to traversed edges per second or giga updates per second. Actual computation is probably not the limiting factor in terms of speed and energy consumption; rather, it is the task of moving data back and forth among processors, cache levels, main memory, and storage as well as the energy required to hold the state in DRAM chips. The shared bus between the program memory and data memory leads to the Von Neumann bottleneck. Since CPU speed and memory size have increased much faster than the throughput between them, this bottleneck has become more of a problem, and its severity increases with every newer generation of CPUs. One potential solution to this challenge is memory-driven computing [59], as illustrated in Figure 7. In this new computing paradigm, memory becomes central and nonvolatile, while computation becomes peripheral and ephemeral. In this new environment, there is little hope that improvements in general-purpose processors will eventually yield improvements to all existing software. The only way to significantly improve computational throughput and efficiency will most likely be to design a machine that is optimized for a particular task. In such a scenario, frequency flexibility becomes important. For each particular task, clock frequency has to be tuned for optimized performance in terms of computation and data movement. Frequency granularity and frequency-switching speed will be powerful tools for obtaining the best result. The FS-ISA (discussed previously) can play a role in memory-driven computing as well.

Network Time Synchronization

Figure 6 describes the data communication problem in a local area where there are direct links for clock information to pass among communicating parties. This scenario most often occurs among blocks within a chip or systems networked in a small local area network. For a problem of this nature, frequency synchronization among the clock sources is the object. For a large packet-oriented network (such as the Internet), a direct frequency link is impossible. In this case, time synchronization, rather than frequency synchronization, is the goal. It has to be achieved through synchronization message exchange. In a time-division multiple access network, synchronization messages are exchanged in guaranteed time slots. In a packet network, messages are exchanged as regular communication packets [53]. Time synchronization depends on the statistical characteristic of the network (such as network delay). Based on the cost and the required synchronization target, the task of synchronization can

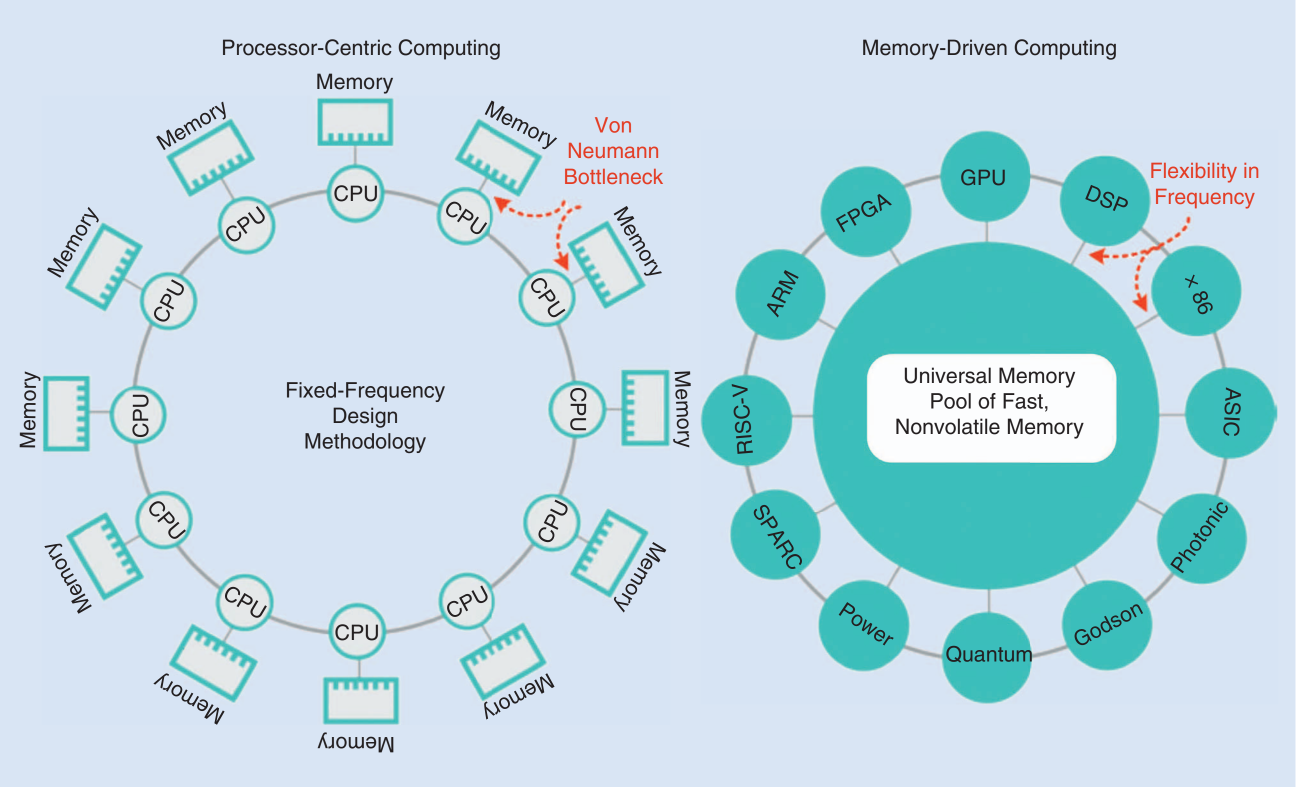


FIGURE 7: Memory-driven computing requires flexible clocking [59]. DSP: digital signal processor. ASIC: application-specific integrated circuit; RISC-V: reduced instruction set computing.

The most characteristic feature of a clock network is its large capacitive loading presented to the clock source.

be carried out in software, hardware, or a hybrid mode. In all of these methods, the fundamental building blocks of the time-synchronization mechanism are the synchronization event detection, remote clock estimation, and local clock correction techniques. They all affect the achievable synchronization precision.

Parameters influence the synchronization precision as shown in $\pi = c1 \cdot \varepsilon + c2 \cdot P + c3 \cdot G + c4 \cdot u + c5 \cdot GS$, where π is the precision; ε is the transmission delay uncertainty when reading the remote clock; P is the clock drift (due to the local oscillator frequency drift); G is the clock reading granularity; u is the rate adjustment granularity; GS is the clock-setting granularity; and $c1$, $c2$, $c3$, $c4$, and $c5$ are the weighing factors [60]. The granularities G, GS, and u are related to the size of the clock tick directly or implicitly. Flexible clock generators, equipped locally at each node with the capabilities of arbitrary frequency generation and instantaneous frequency switching, can provide finer frequency granularity and a better capability to manage frequency drift. This can surely help improve the network time synchronization accuracy ([61, Sec. 5.6]).

Clock Distribution

The end of Dennard scaling from around 2004 has given rise to the multicores processor architecture. The key reason for this shift is high power consumption and the heat-removal problem associated with clock frequency increase. However, another reason is the clock signal distribution problem. Modern SoCs can be regarded as many on-chip micronetworks communicating with each other all the time. A clock is the key signal that makes this happens. Distributing a high-frequency clock signal to a large physical area is difficult. From the clocking perspective, chip architecture can be classified as globally asynchronous, locally synchronous, globally synchronous, and locally synchronous (GSLS). In the GSLS approach, the clock signal drives all the on-chip modules running at the same frequency with fixed phase relationships among each other.

There are several design elements to consider when distributing a clock signal globally: the skew caused by different distribution paths, the jitter accumulated along the distribution path, the silicon and metal resource required for routing the clock signal, and the power used by the distribution network. In practice, there are several distribution methods, including conventional trees, delay/skew-compensated H-tree, clock mesh, and distributed PLL array. They all have advantages and disadvantages. The TAF-DPS flexible clock source provides another possibility: distributing a global clock signal in low frequency and boosting it to user-desired values at the destinations ([61, Sec. 6.1]).

The most characteristic feature of a clock network is its large capacitive loading presented to the clock source. During operation, the clock source is responsible for the charge and discharge of this large capacitance. Therefore, an effective way to lower the power is to recycle the energy used by charging and discharging (charge recovery clock distribution). This approach is realized by using the principle of LC resonance in clock distribution. The large capacitance associated with the clock distribution network functions as the C of the LC oscillator. During the charge and discharge process, energy is stored and released periodically. Ideally, 100% of the energy can be recycled, and the electrical oscillation (the clock waveform) can be self-sustaining. In practice, due to the parasitic resistance associated with the inductor and clock sinks, some portion of the energy is lost as generated heat. Hence, compensation circuitry has to be incorporated on chips to provide energy supporting the oscillation. Using LC resonance for clock distribution, rather than $C \cdot V^2 \cdot f$, the consumed power now is $I^2 R$ where R is the total parasitic resistance. This power is frequency independent so this approach is a good candidate for distributing the clock in a high-frequency (gigahertz) range. In LC resonance clock distribution, there are standing-wave [62] and traveling-wave [63] methods. They all, however, lack the frequency flexibility that the processor operation requires because the oscillation frequency is determined by the physical structure of the network (i.e., the natural frequency of the LC resonator). As a circuit technique in supporting LC resonance clock generation and distribution, TAF-DPS can be used to enhance their frequency flexibility ([61, Sec. 6.2]).

Clock Spectrum

Because of its periodic nature, a clock signal has sharply focused frequency tones in its spectrum. A perfect clock signal would have all of its energy concentrated at the desired frequency and its odd harmonics and would, therefore, radiate energy with very high efficiency that can exceed the regulatory limit for electromagnetic interference. Spread-spectrum clocking (SSCG) is a technique that can be used to alleviate this problem. Instead of one frequency, this technique generates a group of frequencies around a center value and reshapes the system's electromagnetic emission profile [64]–[66]. TAF naturally spreads the clock energy since it uses two or more frequencies (periods) to mimic one virtual average frequency by assigning appropriate weights to these component frequencies.

One of the primary concerns among the many issues associated with SSCG is the risk that modifying the system clock runs the danger of clock and logic circuit misalignment. In other words, with the conventional voltage-controlled oscillator (VCO)-adjustment-based SSCG, the short-term jitter could be out of control since VCO is a complex nonlinear component. When TAF-DPS is used for SSCG, its open-loop style ensures operation precision. Only two types of periods are used whether the spread spectrum feature is on or off. When the spread spectrum function is turned on, only the weight and the occurrence pattern of the periods need to be adjusted. Therefore, no additional timing risk is added ([51, Sec. 6.14]).

> **TAF-DPS is a circuit technique that can also function as a PWM generator since it directly constructs each pulse in its output pulse train.**

PWM

PWM is an important circuit technique that has many applications, such as power delivery, voltage regulation, class-D audio amplifier, and pulse code modulation digital sound. A PWM pulse train can be generated by a microcontroller-controlled counter or by an RC delay-based analog PWM modulator. TAF-DPS is a circuit technique that can also function as a PWM generator since it directly constructs each pulse in its output pulse train. It can produce three types of pulse trains: type I of fixed duty cycle with a varying period, type II of varying duty cycle with a fixed period, and type III of fixed pulse length with a varying period. Compared to the conventional approaches, the TAF-DPS PWM technique is highly flexible. Since the PWM circuit is mainly used in low-megahertz applications, the TAF-DPS PWM generator can be constructed purely from digital standard cells. This can lead to extremely low-cost and low-power implementation, and it can be a very useful tool for implementing designs in many emerging applications, such as the IoT ([61, Sec. 5.12.3]).

Frequency Source for Senor Design: Time of Flight

Sensors are crucial building blocks of the IoT. To detect the myriad types of changes that occurred in our surrounding environment, many different types of sensors are required. The environmental changes of interest are the electromagnetic field, voltage and current, temperature, pressure, liquid and gas flow, light intensity, time of flight (TOF), and chemical composition, among others. Any such change can only present itself through one of the following types of mediums: electromagnetic radiation (light and radio wave), acoustic radiation (sound and ultrasound), particle radiation, mechanical force, heat, or a type of material. The output from any sensing element must be in one of the following forms: change of voltage, current, resistance, or capacitance or mechanical vibration. Following the sensing element is the sensing circuit, which converts the amount of change or the vibration into voltage or frequency or time. For an electrical circuit, only voltage, frequency, and time can be directly manipulated in quantifiable fashion and used for signal processing.

Circuit design offers two approaches for signal sensing: voltage and time based. In voltage-based systems, voltage represents the information and time is just for indexing. In time-based sensing, however, time is specifically used to convey messages. In practice, an important method for manipulating time is through frequency source. For high-precision time sensing, two frequency sources of slightly different frequencies can be employed (the Vernier method). This technique is useful for measuring TOF, which is necessary in many sensor applications of high measurement accuracy. In this application, small frequency granularity $\xi = f_2 - f_1$ leads to small-time granularity $\sigma = T_1 - T_2$, which subsequently leads to finer time resolution in the TOF measurement ([52, Sec. III.4]). When voltage is used as the medium for high-precision information sensing, the resolution is limited by voltage headroom and noise floor. For a time-based approach, however, there is no such limitation since there is no end in time.

Frequency as a Software-Programmable-Variable

A digital-to-analog converter (DAC) is an important component in signal processing. When a DAC is in operation, the output voltage can be considered as a programmable variable. Its input control, a digital value, can be programmed through hardware or software to achieve a multitude of functional effects. Similarly, it is desirable that the frequency is a programmable variable as well and the counterpart of the DAC, digital-to-frequency converter (DFC). A DAC's resolution is ultimately limited by its voltage noise level. For a DFC, there is virtually no limit; the eventual limit is Heisenberg uncertainty. The features of arbitrary frequency generation and instantaneous frequency switching enabled by TAF-DPS makes the DFC concept feasible. With a DFC, we want to give application engineers and software programmers a tool to use to explore new opportunities in higher application levels. This is a new school of thought, and we expect some pleasant surprises when this tool becomes available to thousands of creative minds.

Summary

The issues discussed in the previous sections are examples of what we called *play with frequency*, Its value can be appreciated from two aspects: solving traditional problems from a new perspective and handling emerging problems with a more powerful tool. This new practice is not just circuit inventions. It starts with a revolutionary concept, which is an anomaly in the long history of several decades of using clock frequency in IC design. It serves as a paradigm shift in the field of microelectronics design. When Newton and Leibniz were nurturing calculus in the 17th

In the history of microelectronics, using electric charges is the dominant method of encoding information.

century, they invented the concept of *infinitesimal*, which is both powerful and confusing. This concept reflects the most significant change in mathematical culture since the ancient Greeks, who required that concepts make logical sense. The use of the infinitesimal concept, for the first time in the history of mathematics, abandoned logical rigor in favor of practical usefulness. Using the TAF concept, similarly for the first time in decades of electrical engineering practice, trades the clock pulse train's structural beauty with its functional power. Figure 8 depicts a big picture of Clock Moore: play with frequency.

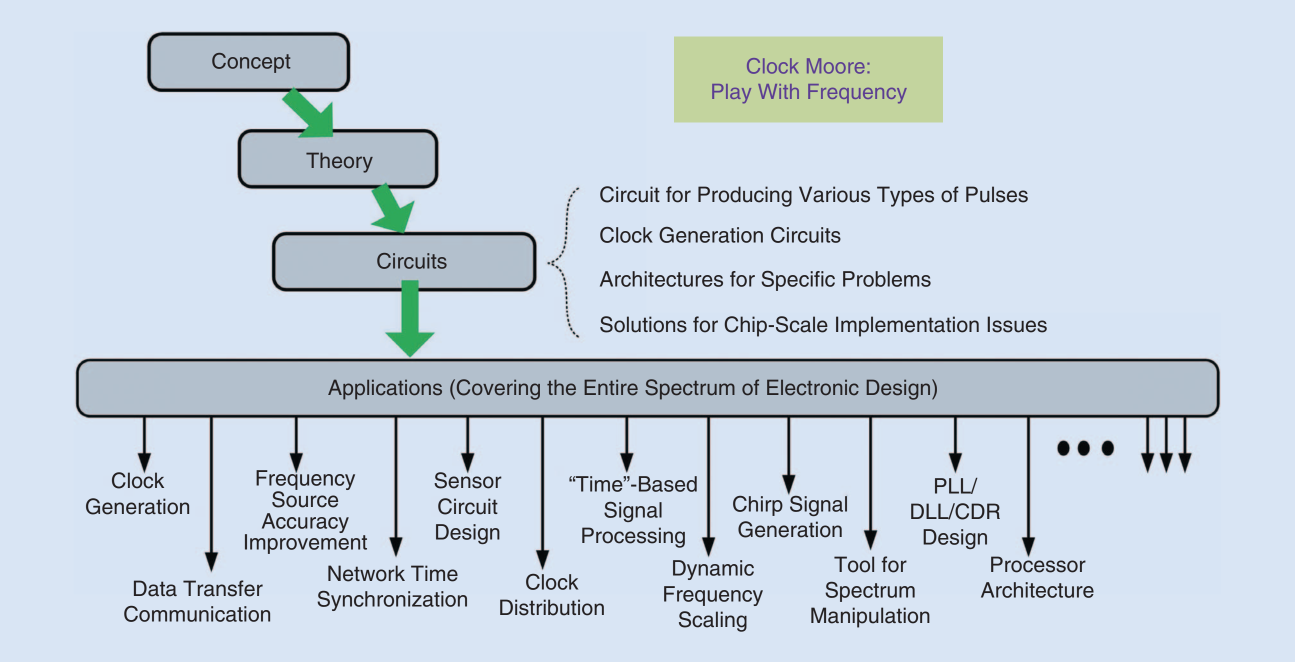
Rate Moore

Information-Encoding Methods in Microelectronics

In the good old time of Moore's law (especially when Dennard scaling held), process shrink prowess overwhelmed innovations in all other technological areas, forcing the market to adapt to a fixed technology rather than having technologies support market needs. After five decades, however, the semiconductor industry's landscape starts to change. Research and development now shifts its focus from the miniaturization of long-established CMOS technology to the coordinated introduction of new devices, new integration technologies, and new architectures for computing, as illustrated in Figure 9 [3]. Myriad activities can be roughly classified into the following five categories [19]–[23]:

- architecture and software advance (advanced energy management, advanced circuit design, SoC specialization, logic specialization, dark silicon, and near-threshold voltage operation [67])
- 3D integration and packaging (3D chip stacking through silicon vias, metal layers, and active layers)
- resistance reduction (superconductors and crystalline metals)
- millivolt switches or better transistors (tunnel field-effect transistors, heterogeneous semiconductors, strained silicon, carbon nanotubes and graphene, and piezo-electric transistors)
- beyond transistor new logic paradigm (spintronics [68], [69], topological insulators, nanophotonics, and biological and chemical computing).

In the history of microelectronics, using electric charges is the dominant method of encoding information. Electric charge as a state variable lies at the core of Moore's law. Eventually, however, fundamental physics limitations will determine the conclusion of CMOS scaling. In the quest for new routes, an alternative would be to find an entirely new information processing state variable based on different physics. It could use electron spin, magnetic dipoles, phase state, electron state, and photons (please refer to the dark blue bottom portion of Figure 9) to improve the computation performance and reduce the switching energy for devices with the smallest features in the order of a few nanometers. Those nanoscale structures pass tokens in the spin, excitonic, photonic, magnetic, quantum, or even heat domains. The emerging physical



Concept
Theory
Circuits
Clock Moore: Play With Frequency
Circuit for Producing Various Types of Pulses
Clock Generation Circuits
Architectures for Specific Problems
Solutions for Chip-Scale Implementation Issues
Applications (Covering the Entire Spectrum of Electronic Design)
Clock Generation
Data Transfer Communication
Frequency Source Accuracy Improvement
Network Time Synchronization
Sensor Circuit Design
Clock Distribution
"Time"-Based Signal Processing
Dynamic Frequency Scaling
Chirp Signal Generation
Tool for Spectrum Manipulation
PLL/ DLL/CDR Design
Processor Architecture
• • •

FIGURE 8: The big picture of Clock Moore. DLL: dynamic linked library; CDR: clock data recovery.

behaviors and idiosyncrasies of these switches can execute specific algorithms by enabling unique architectures. Alternative tokens also require new transport mechanisms to replace the conventional chip wire interconnect schemes of charge-based computing. Ultimately, exploiting those novel techniques could extend throughput in high-performance computing. New intrinsic limits to scaling in post-CMOS technologies are likely to ultimately be bounded by thermodynamic entropy and Shannon noise.

For example, quantum computing (QC) is a computing method that uses quantum-mechanical phenomena of superposition and entanglement [70], [71]. It is completely different from traditional digital electronic computing that is based on binary logic. Instead of using a bit, which is always in one of the two definite states of zero or one, quantum computing uses qubit that can be in superposition of states. Hence, quantum computing is not in the category of binary logic. The Rate Moore design methodology, which will be discussed in the next section, utilizes rate of switching as a state variable to represent information. In essence, it directly uses time as information (not just an indexing tool as used in a clock signal). As depicted in Figures 2 and 9, Clock Moore is an effort carried out in the direction of circuit and architecture innovations while Rate Moore is a novel approach introduced at the state variable level.

Rate Moore: Rate of Switching for Information

Figure 10 illustrates the engineering practices used by a circuit designer to design an electrical circuit. Figure 10(a) is the analog design approach in which signal processing is based on the voltage value. Every value has its divine meaning, and every value is meaningful only when associated with its time stamp. The digital design method is depicted in Figure 10(b). In this domain, voltage value is only differentiated by two

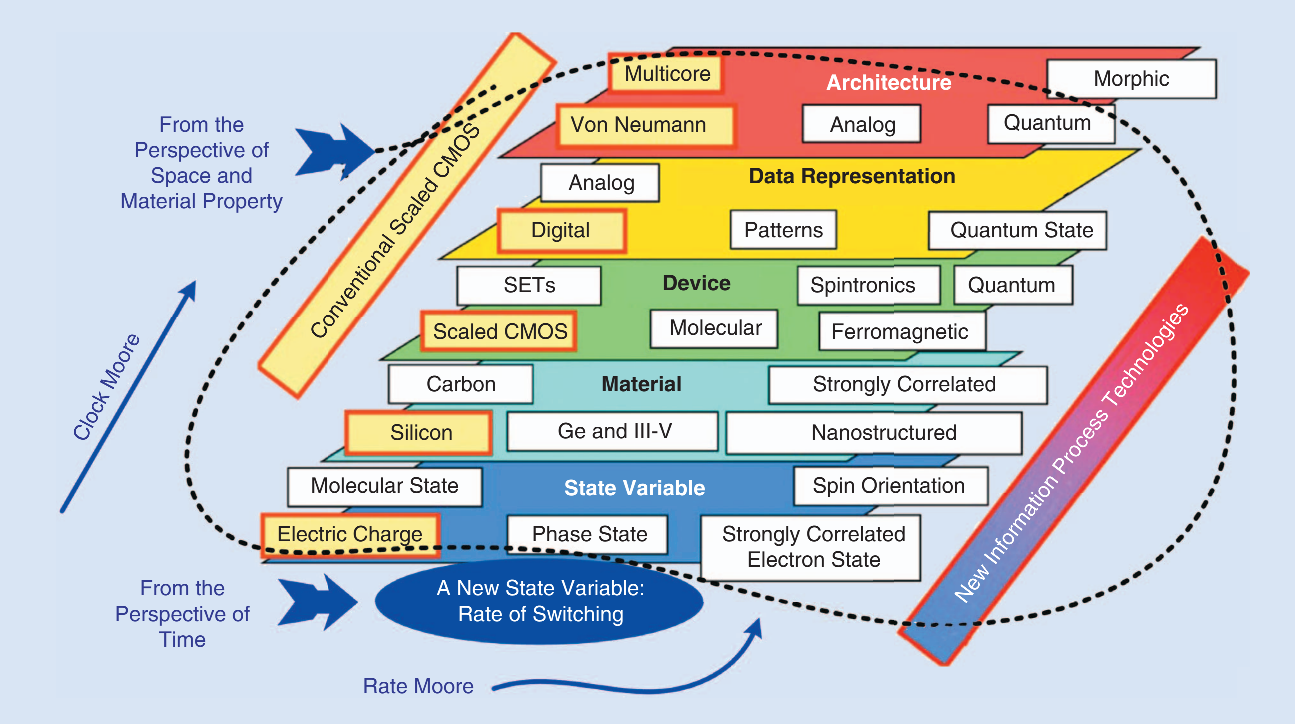


FIGURE 9: Introducing a new computing variable from the perspective of time: rate of switching.

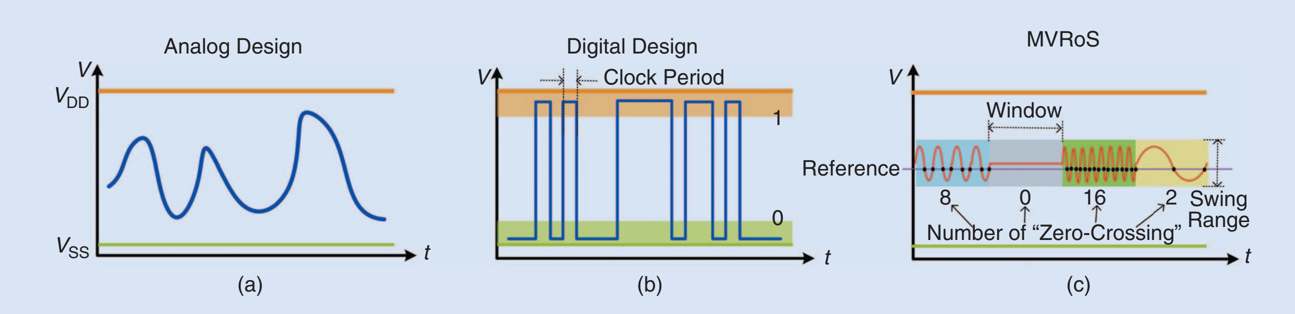


FIGURE 10: Circuit design methodologies: (a) analog, (b) digital, and (c) MVRoS.

The four computational variables (electric charge, electric dipole, magnetic dipole, and orbital state) all rely on material property and, therefore, space (since matter occupies space).

levels, high and low, and all other values have no meaning. A clock signal is often used to mark the timing locations of events. In Figure 10(c), the idea of the rate of switching design approach is illustrated. In this method, information is represented by the rate of switching activities. Instead of the clock period, the frame of time is divided into a plurality of windows, and the switching activities that occur within each window are considered as information. In other words, the number of zero crossings is counted and used for signal processing. As can be appreciated, the number of activities within any given window could be larger than two. Therefore, this method is called *multivalue rate of switching* (*MVRoS*).

Figure 11 compares traditional binary logic and MVRoS. First, the computing variable in binary logic is electric charge while the number of zero-crossing is used as the computing variable in MVRoS. Second, in MVRoS, the "window" is structurally the same as a traditional clock period, but the size of a window is usually larger than the clock period. The window is used to count embedded switching activities and can be utilized not only as an index format but also as a set of events within an index. From a clock signal perspective, in conventional digital circuit design, the set of switching activities that occurred within one period is marked as a group, and the internal details are of no concern (they have been resolved by logic gates). For the MVRoS method, the internal switching activities are meaningful. Similar to a conventional clock signal in which each clock pulse can be numbered for indexing, each window is numbered for the same purpose. For conventional digital logic, the information within each clock period is a high/low voltage state plus clock index number. Meanwhile, in MVRoS, the information within each window is the number of zero-crossings plus window index number.

There are at least five advantages in MVRoS. First, unlike the full voltage swing from V_{SS} to V_{DD} in the conventional digital approach (the open and close of a switch), the voltage swing in MVRoS can be much smaller since we are only concerned about zero-crossing. This could potentially lead to very low voltage V_{DD} design and thus reduce the power consumption to a great degree (power consumption is becoming an issue of higher concern than speed in modern designs). The second advantage is that a larger number of states is achievable (larger than the two possibilities of zero and one). The use of multivalued logic can improve computing efficiency in a great stride (for example, imagine the scenario of using a decimal system directly in a computer). Third, the efficient use of time resources is improved two-fold since the window is used not only for indexing but also for information processing. Fourth, although the speed gain in each new generation of CMOS technology is not as large as in previous generations, the transistor's absolute switching speed is still increasing. This fact favors the rate of switching approach as it is cheaper to take advantage of the transistor's speed gain when using rate of switching processing than traditional level-based processing. Finally, as Dennard scaling was, MVRoS could potentially be scalable with technology advancement (new transistor feature sizes, new devices, and new computation variables).

The four computational variables (electric charge, electric dipole, magnetic dipole, and orbital state) all rely on material property and, therefore, space (since matter occupies space). MVRoS introduces a new dimension, time, but it still depends on the electric charge for its implementation. It does not introduce any new kind of device but simply provides a new perspective for electric charge. Its control variable is a voltage or current (electric charge), and its output is still voltage or current. However, its state variable is rate of switching

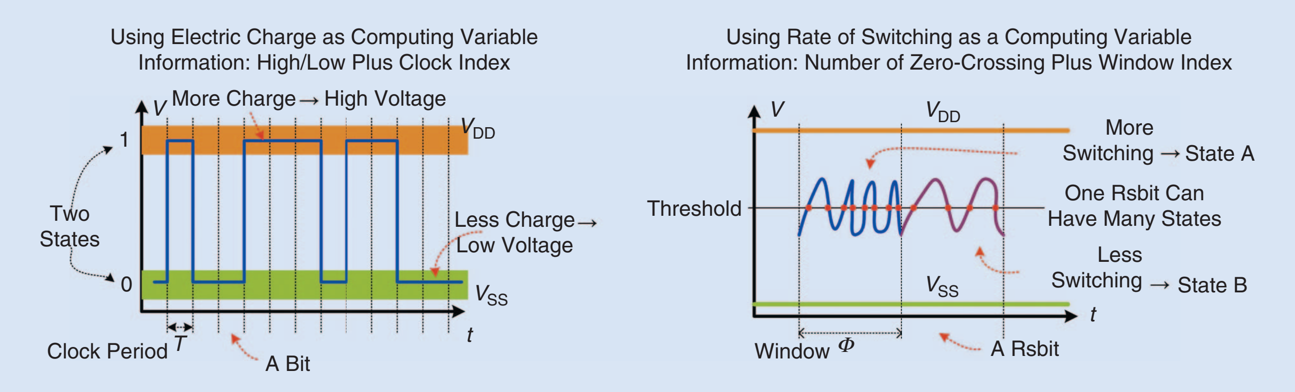

Using Electric Charge as Computing Variable
Information: High/Low Plus Clock Index
V
More Charge → High Voltage
1
V_{DD}
Two States
0
Less Charge → Low Voltage
V_{SS}
t
Clock Period T
A Bit
Using Rate of Switching as a Computing Variable
Information: Number of Zero-Crossing Plus Window Index
V
V_{DD}
More Switching → State A
Threshold
One Rsbit Can Have Many States
Less Switching → State B
V_{SS}
t
Window Φ
A Rsbit

FIGURE 11: A detailed comparison between traditional binary logic and MVRoS.

(time). At this stage, MVRoS is just an idea at its infant stage. Two key challenges are ahead: the development of a many-valued logic system and the creation of associated fundamental logic circuits. The field of many-valued logic started in the 1920s and has enjoyed great advancements since then. However, for MVRoS, more work is definitely required. For the task of creating fundamental logic circuits to handle multivalued logic (similar to the creation of binary logic gates of NOT, AND, and OR), it is a new research field.

> ***Time Moore is at the intersection of many disciplines and a challenge that requires an overwhelmingly collaborative effort.***

Modern computing models can be roughly classified into four categories [72]: classical digital computing (CDC), analog computing (AC), neuro-inspired computing (NC), and QC. CDC includes all the binary digital electronics that form the basis for the microelectronics industry. AC includes nonbinary devices that implement computation through direct physical principles. NC comprises devices based on the principles of brain operation and general neuronal computation. QC is designed to solve some problems with combinatorial complexity through the selection of a desired state from a superposition of all possible answers to a problem. There are strengths and weaknesses for each of these computation paradigms. AC could be simpler than some digital approximations but does not lend itself to general-purpose computing because the devices are specialized for dedicated computation. The computational precision is problematic to maintain and can be sensitive to its environment. CDC is good at deterministic and algorithmic calculation but poor at simple reasoning and recognition. NC devices are proven to be inherently resilient and are very good at solving problems for which CDC is not. Many unexplored opportunities exist for NC, but much is still not understood about how the brain actually computes. QC theoretically could enable the efficient solution of some combinatorial and NP-hard problems or could be used to simulate the electronic state of complex molecules. However, it surely is not a suitable replacement for CDC in domains where CDC excels.

Those options create possibilities for approaches that could go beyond what traditional digital electronics have effectively performed. However, they are not suitable as replacements for digital electronics in tasks that digital computing already performs well. MVRoS can be regarded as an extension of CDC or a new technological implementation of the CDC model. This potential candidate could be considered to continue the phenomenon of Dennard scaling. As the transistor feature size shrinks, it switches more quickly; but conventional charge-based transistor operation does not enjoy a supply voltage decrease after the Dennard scaling breakdown. MVRoS can take advantage of the supply voltage decrease and thus facilitate the power consumption reduction, since it uses the rate of switching instead of a switch's fully open and fully closed states. Our view is that MVRoS could be the most immediately relevant option to boost computation efficiency after traditional digital computation.

Conclusion: At the Crosspoint of Multidiscipline

Alan Turing demonstrated how to describe the solution to computable problems by using a method of programming.

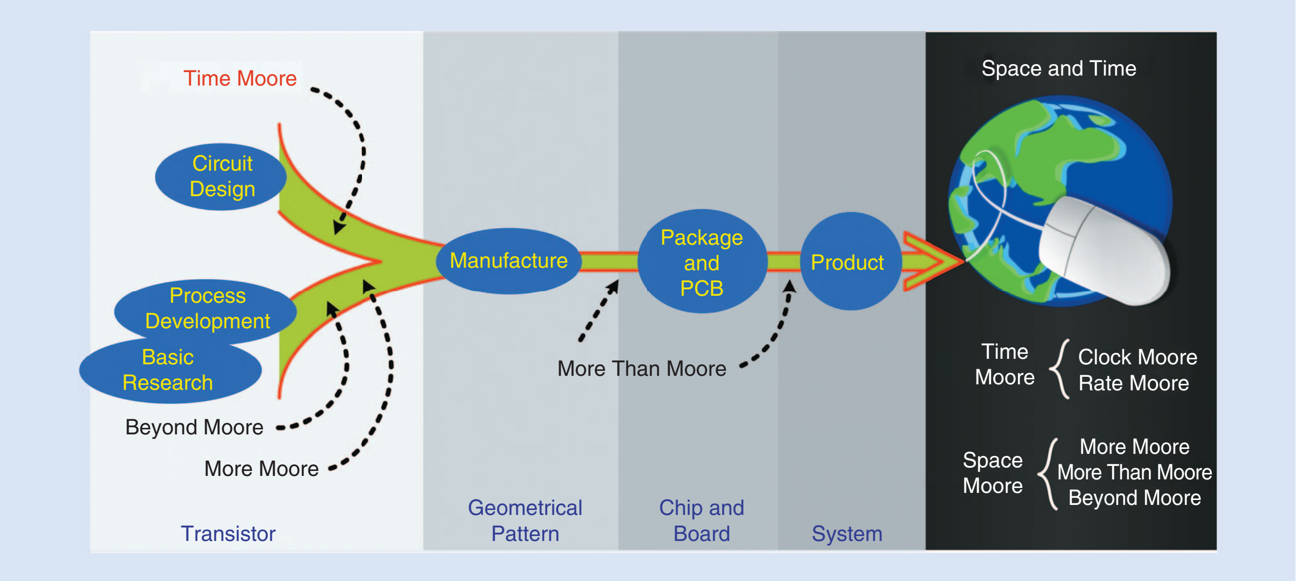


FIGURE 12: Time Moore: it is now time for circuit and system professionals to play a more important role. PCB: printed circuit board.

John von Neumann designed a computer for running the program. Gordon Moore described how semiconductor scaling makes the computer grow exponentially more capable over time. Along this route, the ingenuity of determined and creative engineers has created the modern society of electronics. After five decades of brilliant work on Moore's law, it is not unreasonable to say that now there is little left from the direction of space (the geometrical scaling). The meaning of the phrase *Moore's law*, however, is likely to morph once again, to capture the different trends still driving toward the same goal: the long-term improvement of computation performance. Moore's law will surely endure as a concept since it stands for something much bigger than the performance of a transistor. Time Moore is one such new ideology. A prevalent view among semiconductor professionals is that time only serves the supportive role of event indexing. However, with recent technological advances and the more demanding requirements of modern applications, this will no longer be the case. To rise electronics' information-processing capability to the next stage, it is now the time to explore *time* in greater depth, at a much higher sophistication.

As illustrated in Figure 12, in the path of Space Moore (which includes More Moore, More Than Moore, and Beyond Moore), the driving force mostly comes from material science and an equipment maker. The trendsetters are physicists, chemists, and other scientists, and they define how the transistors look. Circuit designers play a secondary role of using the transistors to create functions. In the era of Time Moore, however, we have a whole new world of time to explore. Time Moore is at the intersection of many disciplines and a challenge that requires an overwhelmingly collaborative effort from a very large-scale IC designer, system architect, process researcher, computer scientist, network architect, and software engineer, among others. It will be a rich process of cross-fertilization since time is one of the fundamental pieces of our universe. History has shown that, once unleashed to the world, a big idea sets into motion and is rarely confined to a single discipline. For the case of exploiting time and subsequently improving the information processing efficiency of electronics, the conceptual and technological pieces have already come together to make this campaign imaginable. In this battlefield, circuit and system professionals are in a unique position that no other professional is able to enjoy. Hence, it is now their responsibility to lead. The purpose of this article is to make more people aware of this new trend. As Charles Darwin said more than 100 years ago, "it is not the strongest of the species that survives, nor the most intelligent, but the one most responsive to change." Along this line of thought, the most effective clock is not the one with the highest frequency or the one with the purest spectrum, but the one that is most responsive to change.

References

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electron.*, vol. 38, no. 8, pp. 114–117, 1965.

[2] G. E. Moore, "Progress in digital integrated electronics," in *Proc. Tech. Dig. Int. Electron Devices Meeting*, 1975, pp. 11–13.

[3] International Technology Roadmap for Semiconductors (ITRS). [Online]. Available: www.itrs.net

[4] International Technology Roadmap for Semiconductors 2.0, "2015 edition, executive report." [Online]. Available: https://www.semiconductors.org/clientuploads/Research_Technology/ITRS/2015/0_2015%20ITRS%202.0%20Executive%20Report%20(1).pdf

[5] R. R. Schaller, "Technological innovation in the semiconductor industry: A case study of the International Technology Roadmap for Semiconductors (ITRS)," Ph.D. dissertation, George Mason Univ., 2004.

[6] P. K. Bondyopadhyay, "Moore's law governs the silicon revolution," *Proc. IEEE*, vol. 86, no. 1, pp. 78–81, 1998.

[7] K. Rupp and S. Selberherr, "The economic limit to Moore's law," *Proc. IEEE*, vol. 98, no. 3, pp. 351–353, 2010.

[8] K. Flamm, "Has Moore's law been repealed? An economist's perspective," *Comput.* Sci. Eng., vol. 19, no. 2, pp. 29–40, 2017.

[9] T. P. Morgan, "Nvidia's Tesla Volta GPU is the beast of the datacenter," May 10, 2017. [Online]. Available: https://www.nextplatform.com/2017/05/10/nvidias-tesla-volta-gpu-beast-datacenter

[10] V. V. Zhirnov, R. K. Cavin, III, J. A. Hutchby, and G. I. Bourianoff, "Limits to binary logic switch scaling—A Gedanken model," *Proc. IEEE*, vol. 91, no. 11, pp. 1934–1939, Nov. 2003.

[11] J. R. Powell, "The quantum limit to Moore's law," *Proc. IEEE*, vol. 96, no. 8, pp. 1247–1248, 2008.

[12] Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S. H. Lo, G. A. Sai-Halasz, R. G. Viswanathan, H. Wann, S. J. Wind, and H. S. Wong, "CMOS scaling into the nm regime," *Proc. IEEE*, vol. 85, no. 4, pp. 486–504, 1997.

[13] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H. S. P. Wong, "Device scaling limits of Si MOS FETs and their application dependencies," *Proc. IEEE*, vol. 89, no. 3, pp. 259–288, 2001.

[14] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFETs with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, Oct. 1974.

[15] M. Bohr, "A 30 year retrospective on Dennard's MOSFET scaling paper," *IEEE Solid-State Circuits Society Newslett.*, vol. 12, no. 1, pp. 11–13, 2007.

[16] P. A. Gargini, "How to successfully overcome inflection points, or long live Moore's law," *Comput. Sci. Eng.*, vol. 19, no. 2, pp. 51–62, 2017.

[17] T. N. Theis and H. S. P. Wong, "The end of Moore's law: A new beginning for information technology," *Comput. Sci. Eng.*, vol. 19, no. 2, pp. 41–50, 2017.

[18] E. P. DeBenedictis, M. Badaroglu, A. Chen, T. M. Conte, and P. Gargini, "Sustaining Moore's law with 3D chips," *Computer*, vol. 50, no. 8, pp. 69–73, 2017.

[19] R. K. Cavin III, P. Lugli, and V. V. Zhirnov, "Science and engineering beyond Moore's law," *Proc. IEEE*, vol. 100, pp. 1720–1749, May 2012.

[20] D. E. Nikonov and I. A. Young, "Overview of beyond-CMOS devices and a uniform methodology for their benchmarking," *Proc. IEEE*, vol. 101, no. 12, pp. 2498–2533, 2013.

[21] K. Bernstein, R. K. Cavin, W. Porod, A. Seabaugh, and J. Welser, "Device and architecture outlook for beyond CMOS switches," *Proc. IEEE*, vol. 98, no. 12, pp. 2169–2184, Dec. 2010.

[22] T. N. Theis and P. M. Solomon, "In quest of the 'next switch': Prospects for greatly reduced power dissipation in a successor to the silicon field-effect transistor," *Proc. IEEE*, vol. 98, no. 12, pp. 2005–2014, Dec. 2010.

[23] J. M. Shalf and R. Leland, "Computing beyond Moore's law," *Computer*, vol. 48, no. 12, pp. 14–23, 2015.

[24] T. Conte, "IEEE rebooting computing initiative & international roadmap of devices and systems," in *Proc. IEEE Rebooting Computer Architecture 2030 Workshop.* [Online]. Available: https://arch2030.cs.washington.edu/slides/arch2030_tom_conte.pdf

[25] D. W. Allan, N. Ashby, and C. C. Hodge, "The science of timekeeping," Hewlett-Packard, Englewood, CO, Application Note 1289, 1997. [Online]. Available: http://allanstime.com/Publications/DWA/Science_Timekeeping/TheScienceOfTimekeeping.pdf

[26] T. Jones, *Splitting the Second: The Story of Atomic Time.* Philadelphia, PA: Inst. Physics Publishing, 2000.

[27] E. A. Gerber and R. A. Sykes, "Quartz frequency standards," *Proc. IEEE*, vol. 55, no. 6, pp. 783–791, 1967.

[28] J. R. Vig, "Quartz crystal resonators and oscillators for frequency control and timing applications: A tutorial," U.S. Army Communications-Electronics Command, Fort Monmouth, NJ, Tech. Rep. AD-A328861, Jan. 2000. [Online]. Available: https://www.am1.us/wp-content/uploads/Documents/U11625_VIG-TUTORIAL.pdf

[29] A. Hajimiri and T. H. Lee, "General theory of phase noise in electrical oscillators," *IEEE J. Solid-State Circuits*, vol. 33, no. 2, pp. 179–194, 1998.

[30] R. Poore. (2001). Overview on phase noise and jitter. Agilent Technologies. Santa Clara, CA. [Online]. Available: http://cp.literature.agilent.com/litweb/pdf/5990-3108EN.pdf

[31] F. M. Gardner, *Phaselock Techniques*, 3rd ed. Hoboken, NJ: Wiley, 2005.

[32] W. F. Egan, *Phase-Lock Basics*, 2nd ed. Hoboken, NJ: Wiley, 2007.

[33] R. B. Staszewski and P. T. Balsara, *All-Digital Frequency Synthesizer in Deep-Submicron CMOS*. Hoboken, NJ: Wiley, 2006.

[34] L. Xiu, "The concept of time-average-frequency and mathematical analysis of flying-adder frequency synthesis architecture," *IEEE Circuits Syst. Mag.,* vol. 8, no. 3, pp. 27–51, Sept. 2008.

[35] H. Mair and L. Xiu, "An architecture of high-performance frequency and phase synthesis," *IEEE J. Solid-State Circuits*, vol. 35, pp. 835–846, June 2000.

[36] D. E. Calbaza and Y. Savaria, "A direct digital periodic synthesis circuit," *IEEE J. Solid-State Circuits*, vol. 37, no. 8, pp. 1039–1045, Aug. 2002.

[37] L. Xiu and Z. You, "A Flying-Adder architecture of frequency and phase synthesis with scalability," *IEEE Trans. VLSI*, vol. 10, pp. 637–649, Oct., 2002.

[38] L. Xiu, W. Li, J. Meiners, and R. Padakanti, "A novel all-digital PLL with software adaptive filter," *IEEE J. Solid-State Circuit*, vol. 39, no. 3, pp. 476–483, Mar. 2004.

[39] L. Xiu, "A Flying-Adder based on-chip frequency generator for complex SoC," *IEEE Trans. Circuit System II*, vol. 54, pp, 1067–1071, Dec. 2007.

[40] L. Xiu, "A Flying-Adder PLL technique enabling novel approaches for video/graphic applications," *IEEE Trans. Consumer Electron.*, vol. 54, pp. 591–599, May 2008.

[41] P. Sotiriadis, "Theory of Flying-Adder frequency synthesizers, Part I: Modeling, signals' periods and output average frequency," *IEEE Trans. Circuits Syst. I*, vol. 57, no. 8, pp. 1935–1948, 2010.

[42] P. Sotiriadis, "Theory of Flying-Adder frequency synthesizers, Part II: Time and frequency domain properties of the output signal," *IEEE Trans. Circuits Syst. I*, vol. 57, no. 8, pp. 1949–1963, 2010.

[43] P. Sotiriadis, "Exact spectrum and time—Domain output of Flying-Adder frequency synthesizers," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 57, no. 9, pp. 1926–1935, 2010.

[44] T. Rapinoja et al., "A digital frequency synthesizer for cognitive radio spectrum sensing applications," *IEEE Trans. Microw. Theory Tech*, vol. 58, no. 5, pp. 1339–1348, 2010.

[45] S. A. Talwalkar, "Quantization error spectra structure of a DTC synthesizer via the DFT axis scaling property," *IEEE Trans. Circuits Syst. I*, vol. 59, no. 6, pp. 1242–1250, 2012.

[46] S. A. Talwalkar, "Digital-to-time synthesizers: Separating delay line error spurs and quantization error spurs," *IEEE Trans. Circuits Syst. I*, vol. 60, no. 10, pp. 2597–2605, 2013.

[47] L. Xiu, W. T. Lin, and K. Lee, "A Flying-Adder fractional-divider based integer-N PLL: The 2nd generation Flying-Adder PLL as clock generator for SoC," *IEEE J. Solid-State Circuits*, vol. 48, pp. 441–455, Feb. 2013.

[48] L. Xiu, "Direct period synthesis for achieving sub-PPM frequency resolution through time average frequency: The principle, the experimental demonstration, and its application in digital communication," *IEEE Trans. VLSI*, vol. 23, no. 7, pp. 1335–1344, 2015.

[49] L. Xiu and P. L. Chen, "A reconfigurable TAF-DPS frequency synthesizer on FPGA achieving 2 ppb frequency granularity and two-cycle switching speed," *IEEE Trans. Ind. Electron.*, vol. 64, pp. 1233–1240, Feb. 2017.

[50] "TMS320DM816x DaVinci digital video processors," Texas Instruments, Technical Reference Manual, 2013. [Online]. Available: http://www.ti.com/lit/ug/sprugx8b/sprugx8b.pd

[51] L. Xiu, *Nanometer Frequency Synthesis Beyond the Phase-Locked Loop*. Hoboken, NJ: Wiley, 2012.

[52] L. Xiu, "Clock technology: The next frontier," *IEEE Circuit Syst. Mag.,* vol. 37, pp. 27–46, May 2017.

[53] J. L. Ferrant et al., *Synchronous Ethernet and IEEE 1588 in Telecoms*. Hoboken, NJ: Wiley, 2013.

[54] E. G. Larsson and O. Gustafsson, "The impact of dynamic voltage and frequency scaling on multicore DSP algorithm design," *IEEE Signal Process. Mag.*, vol. 28, no. 3, p. 3, 2011.

[55] S. Park et al., "Accurate modeling of the delay and energy overhead of dynamic voltage and frequency scaling in modern microprocessors," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 32, no. 5, pp. 695–708, 2013.

[56] X. Lin, Y. Wang, Q. Xie, and M. Pedram, "Task scheduling with dynamic voltage and frequency scaling for energy minimization in the mobile cloud computing environment," *IEEE Trans. Serv. Comput.*, vol. 8, no. 2, pp. 175–186, 2015.

[57] B. Keller et al., "A RISC-V processor SoC with integrated power management at submicrosecond timescales in 28 nm FD-SOI," *IEEE J. Solid-State Circuits*, vol. 52, no. 7, pp. 1863–1875, 2017.

[58] J. Aragón, J. González, and A. González, "Control speculation for energy-efficient next-generation superscalar processors," *IEEE Trans. Comput.*, vol. 55, no. 3, pp. 281–291, 2006.

[59] R. S. Williams, "What's next? [The end of Moore's law]," *Comput. Sci. Eng.*, vol. 19, no. 2, pp. 7–13, 2017.

[60] M. Horauer, "Clock synchronization in distributed systems," Ph.D dissertation, Vienna Univ. Technol., 2004.

[61] L. Xiu, *From Frequency to Time-Average-Frequency: A Paradigm Shift in the Design of Electronic system*. Hoboken, NJ: Wiley, 2015.

[62] F. O'Mahony, C. P. Yue, M. A. Horowitz, and S. S. Wong, "A 10-GHz global clock distribution using coupled standing-wave oscillators," *IEEE J. Solid-State Circuits*, vol. 38, no. 11, pp. 1813–1820, 2003.

[63] J. Wood, T. C. Edwards, and S. Lipa, "Rotary traveling-wave oscillator arrays: A new clock technology," *IEEE J. Solid-State Circuit*, vol. 36, no. 11, pp. 1654–1665, Nov. 2001.

[64] Y. Matsumoto, K. Fujii, and A. Sugiura, "Estimating the amplitude reduction of clock harmonics due to frequency modulation," *IEEE Trans. Electromagn. Compat.*, vol. 48, no. 4, pp. 734–741, 2006.

[65] N. Keskin and H. Liu, "Practical considerations for electromagnetic interference suppression rate with spread spectrum clocking," *IEEE Electromagn. Compat. Mag.*, vol. 5, no. 2, pp. 57–60, 2016.

[66] H. Ryu, S. Park, E. T. Sung, S. G. Lee, and D. Baek, "A spread spectrum clock generator using a programmable linear frequency modulator for multipurpose electronic devices," *IEEE Trans. Electromagn. Compat.*, vol. 57, no. 6, pp. 1447–1456, 2015.

[67] R. G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, "Near-threshold computing: Reclaiming Moore's law through energy efficient integrated circuits," *Proc. IEEE*, vol. 98, no. 2, pp. 253–266, 2010.

[68] S. A. Wolf et al., "Spintronics: A spin-based electronics vision for the future," *Science*, vol. 294, issue no. 5546, pp. 1488–1495, Nov. 16, 2001.

[69] I. Zutic, J. Fabian, and S. D. Sarma, "Spintronics: Fundamentals and applications," *Rev. Mod. Phys*, vol. 76, no. 2, pp. 323–410, April 2004.

[70] M. Nielsen and I. Chuang, *Quantum Computation and Quantum Information*. Cambridge, MA: Cambridge Univ. Press, 2000.

[71] L. Gomes, "Quantum computing: Both here and not here," *IEEE Spectr.*, vol. 55, no. 4, pp. 42–47, Apr. 2018.

[72] L. Joneckis, D. Koester, and J. Alspector, "An initial look at alternative computing technologies for the intelligence community," Inst. Defense Anal., Alexandria, VA, Tech. Rep. P-5114, Jan. 2014. [Online]. Available: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA610103

About the Author

Liming Xiu (limingxiu@ieee.org) received his B.S. and M.S. degrees in physics from Tsinghua University, China, in 1986 and 1988, respectively, and his M.E.E.E. degree from Texas A&M University, College Station, in 1995. From 1995 to 2009, he was a senior member technical saff with Texas Instruments, Dallas. From 2009 to 2012, he was the chief clock architect with Novatek Microelectronics, Taiwan. From 2012 to 2015, he was the vice president of Kairos Microsystems, Dallas. Since 2015, he has been the chief scientist of IC technology with the BOE Technology Group, Beijing, China. He was the vice president of the IEEE Circuits and Systems Society from 2009 to 2010. He invented the Flying-Adder frequency synthesis architecture and is the promoter of the time-average-frequency concept and theory. He has 25 U.S. patents and published numerous papers and articles and three books: *VLSI Circuit Design Methodology Demystified*, *Nanometer Frequency Synthesis Beyond Phase-Locked Loop*, and *From Frequency to Time-Average-Frequency: A Paradigm Shift in the Design of Electronic System*.