Proceedings of the 4th International Conference on Materials Chemistry and Environmental Engineering
DOI: 10.54254/2755-2721/84/20240895

The evolution of silicon integrated circuit technology —from scaling to lower power consumption

Heng Pu

School of Material Science and Engineering, Tongji University,
4800 Caoan Road, Shanghai, 201804, China

Herrypu@outlook.com

Abstract. With the rapid development of silicon-based microelectronics industry, the silicon-based microelectronics industry are reaching a point where microelectronics can no longer simply scaled to increasingly small sizes, and power consumption controlling are becoming increasingly important. The article starts from the early development of silicon integrated circuit and addresses how the focus transferred from scaling to scaling to lower power consumption. Based on that, the article makes a systematic review of all the technologies of contributing to lower power consumption, from designing to manufacturing, from architecture to circuit level. The article also introduces new devices and architectures adapting 3D integrated circuit and stacking technology as well as the FE-FET transistor and FE-RAM. Meanwhile, the development of lithography technology is also discussed. And a brief outlook on the future development of silicon integrated circuit technology is given.

Keywords: silicon integrated circuit, scaling, lower power consumption

1. Introduction

1.1. The development of silicon integrated circuit technology

1.1.1. The trend of silicon integrated circuit technology

With the continuous maturity and development of integrated methodology and micro-machining technology, as well as the continuous expansion of the application field of integrated technology, integrated circuits show a trend of miniaturization, systematization and correlation. Since 1965, the number of transistors in an integrated circuit doubles every 18 months [1]. The manufacturing technology is updated every 2 to 3 years, which is based on the result of the continuous reduction of the gate length, and the reduction of the device gate length is basically in accordance with the principle of equal proportion reduction, while promoting the improvement of other process parameters.

1.1.2. The Importance of low-power consumption

However, with the consistent scaling of transistors, making smaller sizes has become increasingly challenging due to the laws of physics. Although the manufacturing process has been very precise, there are still certain errors and uncertainties, which limit the transistor size to further reduction. In addition,

DOI: 10.54254/2755-2721/84/20240895

as the size of the transistor shrinks, its quantum effects will become more and more obvious, which further increases the difficulty and uncertainty of manufacturing.

Moreover, due to the shrinking size of transistors, electronic devices face problems of heat dissipation and power consumption. As transistors get smaller, their power density increases, causing the chip to overheat and possibly damage. In addition, as transistors shrink in size, their electrical performance also changes, resulting in increased power consumption and decreased performance. In order to solve these problems, people began to study low-power design technology, which aims to reduce the power consumption and heat of circuit systems, improve the stability of the system, extend the service life of equipment and reduce the impact on the environment.

2. Advances and Innovations for Low-Power Consumption

2.1. Low power design method

Integrated circuit power consumption can be divided into three parts, one part is the dynamic power consumption caused by the circuit charging and discharging the load capacitor, the other part is the power consumption caused by the short circuit current between the power supply and the ground caused by the CMOS transistor in the short time during the jump process, the P tube and the N tube are switched on at the same time, and the third part is the static power consumption caused by the leakage current. Formula (1) and (2) are classical formulas for power analysis of integrated circuits [2]:

$$P = P_{switching} + P_{switching} + P_{switching} \tag{1}$$

$$= ACV^2 f + \tau AVI_{short} + VI_{leak} \tag{2}$$

Where f is the frequency of the system; A is the jump factor, that is, the average inversion ratio of the entire circuit; C is the total capacitance of the gate circuit; V is the supply voltage; τ is the time from the beginning of the level signal change to stabilize.

2.1.1. Low-Power-Consumption (LPC) Technology on Process-Level

At the current process level, the power consumption of SoC is mainly caused by switching power consumption, which can be reduced by reducing the supply voltage. This is why integrated circuits have transitioned from a 5V supply voltage to 3.3V, then to 1.8V, 1.3V, and even lower voltages. However, reducing the supply voltage can lead to some issues because if the threshold voltage remains unchanged, the noise margin will decrease and the ability to resist interference will weaken, leading to a decrease in signal transmission accuracy. To maintain a considerable noise margin, the threshold voltage should also be reduced correspondingly with the decrease in supply voltage. However, after entering the 0.13-micron process the decrease in threshold voltage can lead to static power consumption increasing exponentially. Therefore, to reduce power consumption by reducing voltage, other methods must be used to compensate for the corresponding delay loss to avoid a decrease in system performance. One method is to develop system parallelism and pipelining; the second method is to dynamically control the clock frequency and power voltage through the operating system based on different user requirements for circuit performance, achieving the goal of ensuring performance requirements while saving power consumption; the third method is to dynamically change the supply voltage based on performance requirements. At the critical path of the system, a higher voltage is maintained to ensure the system performance, while at the non-critical path, the voltage is reduced to reduce power consumption.

2.1.2. LPC Technology on Circuit-Level

The data and address lines of the bus in the SoC are generally numerous and long, and each line needs to drive the load, usually accounting for 15~20% of the total power consumption, and in some cases even more than 70%. Therefore, bus low-power technology is one of our focuses.

Reducing the amplitude is currently a mature technology whose principle is that when the output voltage of the output terminal is V_{swing}, the power consumption of the voltage jump is [2]:

DOI: 10.54254/2755-2721/84/20240895

$$P_s = AVCV_{swing}f \tag{3}$$

Therefore, reducing V_{swing} can achieve the purpose of reducing power consumption.

2.1.3. LPC Technology on Gate-Level

In the era of deep submicron for SoC, low-voltage technology is mainly used to achieve low-power technology. Complementary CMOS has great advantages in many aspects and various EDA manufacturers also provide complete support, so in most cases, complementary CMOS is selected. The transmission gate has its superiority in a limited range, such as full adder (Full Adder) has lower power consumption than complementary CMOS when the supply voltage is high, and when using CPL to implement multiplier, it also has great advantages.

2.1.4. LPC Technology on Register Transfer (RTL) Level

RTL low-power technology mainly starts from reducing unwanted jumps (glitch--Spurious switch, hazards). Although this kind of jump has no negative impact on the logical function of the circuit, it will lead to an increase in the switching factor A, thereby leading to an increase in power consumption.

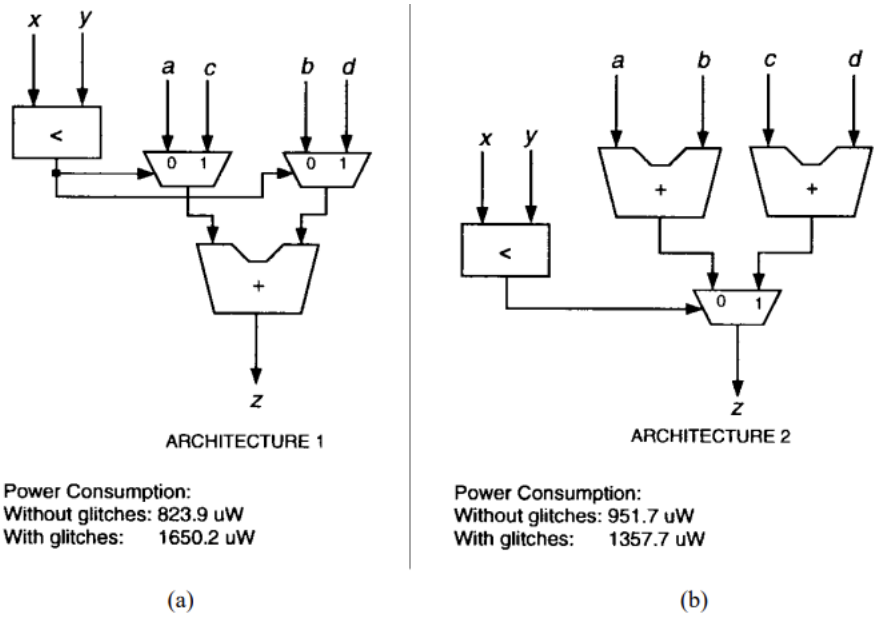

Figure 1. Alternate architectures that implement the same function: effect of glitching [3].

Methods to reduce glitches mainly include eliminating their conditions, such as delay path balance, using clock signal synchronization to reduce glitches, and structural reconstruction. Reducing the number of stages of delay imbalance in delay paths can greatly reduce glitches. By inserting a trigger controlled by a clock signal into the circuit to synchronize the signal to be transmitted, the glitch in front of the trigger can be blocked at the trigger to avoid layer upon layer transmission leading to glitch explosion. However, it is necessary to weigh whether the increased power consumption and area of the introduction of clock trees and triggers are worth the improvement.

2.1.5. LPC Technology on Architecture Level

Parallel technology (parallel) can reduce power consumption. Taking different multipliers as an example, one is a normal 32-bit multiplier with one path, and the other is a 32-bit multiplier with two parallel paths. The latter can achieve the same performance with half the frequency, and the voltage can also be reduced, resulting in a significant overall power consumption reduction. Pipeline technology (PIPELINE) was originally used to increase the main frequency of processors, but it also has significant benefits in reducing power consumption.

DOI: 10.54254/2755-2721/84/20240895

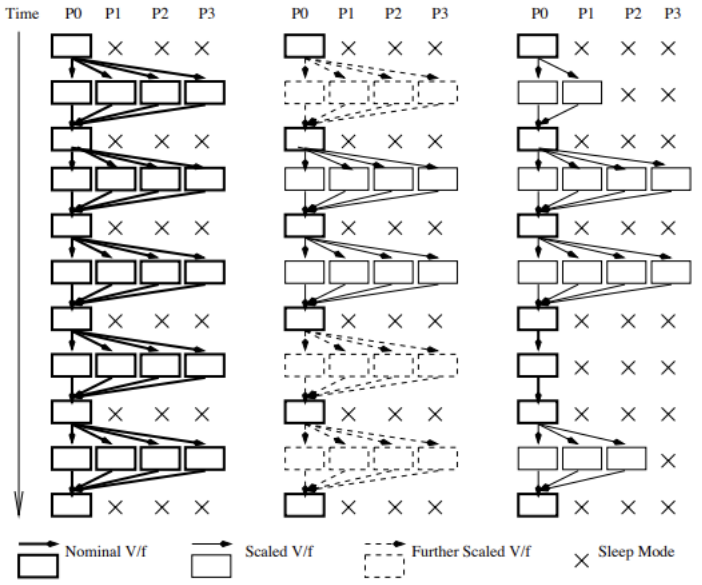

Figure 2. Execution of an imaginary parallel code on a CMP with four processors under three different scenarios: full throttle parallel execution (left), power-aware parallel execution regulated with DVFS exclusively (center), and power-aware parallel execution regulated with DVFSand adaptive parallelism (right). Regions are not drawn toscale with respect to each other [4].

2.1.6. LPC Technology on Algorithm Level

The capacitance of on-chip buses compared to off-chip buses in SoC can be reduced by several orders of magnitude, but it still accounts for a significant proportion in the entire design. Therefore, to reduce overall power consumption, it is necessary to reduce their switching probability. Hamming distance refers to the number of bits that are different between adjacent two binary data. If the Hamming distance exceeds half, inverted code transmission can be adopted. This bus flip decoding technique can greatly reduce the chance of jumping and is especially suitable for data buses, because the data on the data bus is usually not correlated. The cost of bus flip decoding is one more transmission line for indicating whether data is flipped, and circuits for determining Hamming distances and receiving flipped data at the receiving end must be added to the area.

In addition, data transmitted on address buses usually has strong continuity. In the case of continuous switching, using Gray coding technology can reduce switching by about 50%, but Gray coding and binary coding must be converted to each other, thus increasing circuit area.

2.1.7. LPC Technology on System-Level

System-level low-power technologies mainly include clock gating technology and asynchronous circuits, etc. Clock gating technology is currently the most effective low-power technology. Without clock gating technology, the same value would be repeatedly loaded into each succeeding register at every clock cycle rising edge, which would generate unnecessary power consumption in the following registers, clock network, and multiplexer. Inserting clock gating circuits into register slots and clock networks can control and eliminate these unnecessary register activities, greatly reducing power consumption.

The clock tree in synchronous circuits consumes a significant amount of energy. Although clock gating technology can be used to alleviate this problem, it still cannot fundamentally solve it. As clock frequency needs to meet the normal operation of all modules, there is speed waste in low-speed modules. The working mode of asynchronous circuits is “event driven,” and circuits only work when needed. Therefore, speed waste is eliminated because there is no global clock tree power consumption.

DOI: 10.54254/2755-2721/84/20240895

2.2. New devices and architectures

2.2.1. 3D integrated circuit and stacking technology

2.5D packaging and 3D packaging are two commonly used wafer-level multi-layer stacking technologies. 2.5D package is to package the chip to the Si intermediary layer, and use the high-density wiring on the Si intermediary layer for interconnection. Since there is no active device on the Si intermediary layer, this technology is to interconnect multiple chips on the same plane through the Si intermediary layer, without forming a three-dimensional stack between chips, so it is called 2.5D packaging. At present, the representative technology of 2.5D packaging is TSMC's chip-on-wafer-on-substrate, CoWoS) package and Intel's embedded multi-die interconnect bridge (EMIB) package. 3D packaging is truly a vertical interconnection between chips. That is, Foveros technology for logic chip stacking and Co-EMIB technology which is a combination of EMIB and Foveros packaging technology. The packaging process of Foveros is similar to that of CoWoS. The difference is that the intermediary layer of CoWoS is a bare wafer, so it is a passive intermediary layer. However, the intermediary layer of Foveros is a functional chip, which belongs to the active Si intermediary layer. In addition to the horizontal physical layer interconnection and vertical interconnection of Foveros technology, the horizontal interconnection between Foveros 3D stacks is realized by EMIB packaging. The complex interconnection structure puts forward extremely high requirements on the level of wiring design. The packaging process of Co-EMIB is to first use Foveros method to form different blocks of chips, and then use EMIB packaging method to connect these blocks through Si bridge. Whether it is 2D horizontal interconnect or 3D stacked interconnect, low power consumption, high bandwidth and high performance of almost SOC level integration can be achieved between single chip and single chip, bringing excellent flexibility to chip packaging.

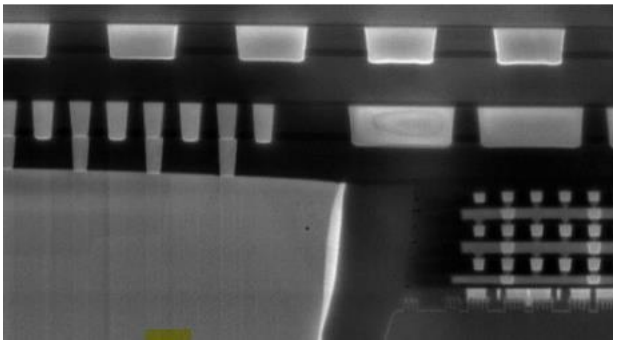

Figure 3. Through Silicon Vias with 3D tri-gate transistors and 22FL interconnect stack [5]

2.2.2. New transistor and new non-volatile memory

In addition, ferroelectric field effect transistors (FE-FETs) are also one of the answers to these challenges. This is a type of field-effect transistor with ferroelectric properties. It takes advantage of the non-volatile memory properties of ferroelectric materials, implanting field effects and charge accumulation in them to achieve long-term stable memory effects. Compared with traditional memory, it has the advantages of low power consumption, high speed and high density. Therefore, a successful FE-FET design can greatly reduce the size and energy usage thresholds of conventional devices, and increase speed.

Deep Jariwala and Kwan-Ho Kim [6] have developed a new FE-FET design that demonstrates record-breaking performance in both computing and storage. The new transistor is covered with a two-dimensional semiconductor called molybdenum disulfide (MoS2) on the ferroelectric material aluminium-scandium nitride (AlScN), demonstrating for the first time that the two materials can be effectively combined to create transistors that are attractive for industrial manufacturing. The device is said to be known for its unprecedented thinness, allowing each individual device to operate with a minimal surface area. In addition, these tiny devices can be manufactured in large arrays that can scale to industrial platforms.

DOI: 10.54254/2755-2721/84/20240895

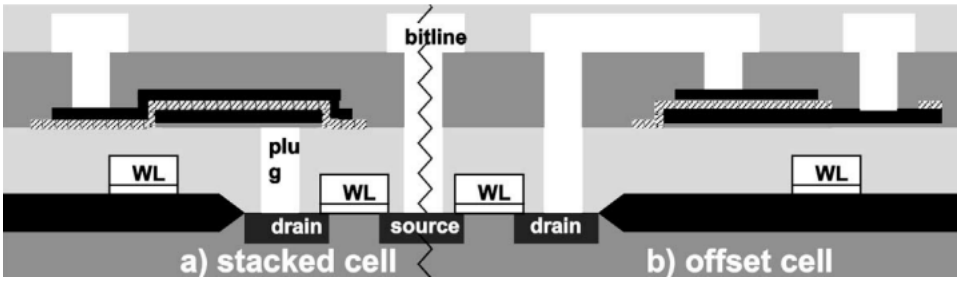

Figure 4. Principle approaches to integrate a ferroelectric capacitor [7]

Developed from this, FeRAM (ferroelectric RAM) is a non-volatile memory with fast write speeds. Compared to traditional non-volatile memory (such as EEPROM, flash memory), FeRAM has higher read/write durability, faster write speed operation and lower power consumption, so our FeRAM is suitable for use where data is frequently rewritten. For example, FeRAM memory has been adopted as a common memory for recording information about meters, measuring instruments, industrial robots, and automobiles. Thanks to the fast write speed, FeRAM can save write data even in the event of a sudden power outage. Not only that, FeRAM can record data more frequently than EEPROM and flash memory. When writing data, EEPROM and flash memory require high voltage, which consumes more power than FeRAM. Therefore, by using FeRAM, battery life can be extended in small, battery-powered devices.

2.3. Innovative manufacturing process

2.3.1. Deep ultraviolet lithography

Deep ultraviolet lithography is a technique used to manufacture microelectronic devices, especially integrated circuits and semiconductor devices. It uses deep ultraviolet light to transfer the designed microelectronic pattern onto a photosensitive material.

Deep ultraviolet lithography works by using deep ultraviolet light (usually with a wavelength between 150-300 nanometres) emitted by the lithography machine to shine on the photosensitive material through a mask (that is, a designed pattern) in a vacuum environment. The part of the photosensitive material that is illuminated by light will undergo a chemical reaction, and its properties will change. During the post-exposure processing, the illuminated part will contrast sharply with the rest, resulting in the desired pattern.

Deep ultraviolet lithography is one of the most used lithography techniques because it allows for higher resolution and smaller device sizes. This allows designers to fit more transistors into the same chip area, resulting in increased integration. Higher integration can reduce power consumption, as fewer transistors mean fewer switches and lower static power consumption. In addition, higher device densities can also be achieved using deep ultraviolet lithography, which provides designers with more freedom to optimize circuit architecture and design. By using more efficient circuit layout and optimized circuit design, power consumption can be reduced and performance improved.

2.3.2. Extreme ultraviolet lithography

With the development of technology, more and more researchers have found that the traditional Deep ultraviolet lithography production process has many defects, one of the most prominent problems is that the bending degree of the crystal is not enough, resulting in the formation of ultraviolet crystals on the crystal surface. Therefore, many people began to pay attention to the problems existing in Deep ultraviolet lithography technology, hoping to find a new production technology to solve these problems. So, extreme ultraviolet lithography technology came into being.

DOI: 10.54254/2755-2721/84/20240895

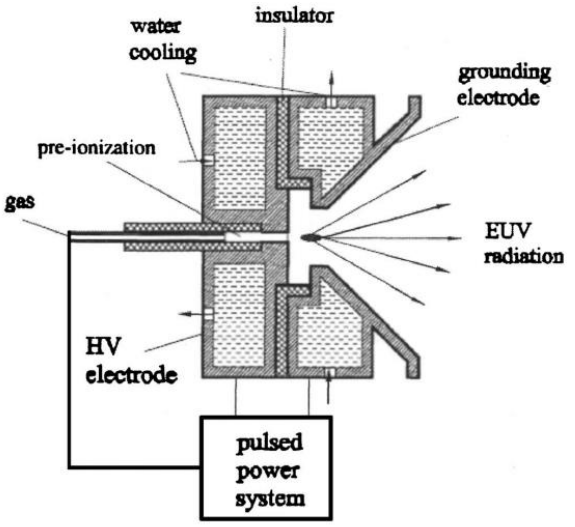

Figure 5. Schematics of a discharge-produced plasma EUV source [8].

However, extreme ultraviolet lithography technology still faces huge challenges such as the production and control of light source, the transmittance of its optical system etc.

Despite these challenges above, extreme ultraviolet lithography continues to make breakthroughs and advances. At present, only the Dutch company ASML can provide commercial ultra-ultraviolet lithography machine, its latest model NXE:3400C can achieve 170 wafers per hour capacity, each wafer can characterize more than 10 billion transistors [9]. The company, whose major customers include chip giants such as Samsung, TSMC, and Intel, has begun mass production of chip products that are 7 nm or more advanced.

3. Conclusion and Outlook

3.1. Based on new materials and technologies other than silicon

At present, more than 90% of semiconductor devices on the market are made of silicon (Si) material [10], one of the first generation of elemental semiconductor materials, which has the advantages of high integration, good stability, low power consumption and low cost. But in the post-Moore era, in addition to the development direction of higher integration, better performance through different materials on integrated circuits is one of the development directions. At the same time, with the development of 5G, new energy vehicles and other industries, the demand for high-frequency, high-power, high-voltage semiconductors, silicon-based semiconductors are difficult to fully meet due to material characteristics, and the second and third generation semiconductors represented by GaAs, GaN, SiC usher in development opportunities.

According to Yole development's forecast [10], the global GaN RF device market size will reach $2 billion in 2025 from $740 million in 2019. The global SiC power device market will grow from $370 million in 2018 to nearly $1.4 billion in 2024, with a CAGR of more than 30%.

3.2. The impact of artificial intelligence and quantum computing

Artificial intelligence is one of the major trends in semiconductor manufacturing in the future. Ai technology requires more advanced processors and chips for efficient data processing and computation. At present, semiconductor companies represented by Intel, Nvidia and AMD are actively developing products with artificial intelligence capabilities. For example, Intel launched the Myriad X chip, which is optimized for deep learning networks and can achieve real-time analysis and reasoning of deep neural networks, which can be widely used in robotics, intelligent monitoring and other fields. Nvidia introduced the Turing architecture, a graphics processor that fully supports artificial intelligence and

DOI: 10.54254/2755-2721/84/20240895

deep learning processing. AMD has also introduced a real-time voice command solution based on artificial intelligence that offers comprehensive security and functionality extensions, as well as the ability to provide a more responsive experience.

Quantum computing is also one of the important development directions of semiconductor manufacturing in the future. Quantum computers are computers that use qubits as registers, and they can perform certain computational operations at very fast speeds, which is much faster than traditional computers. This technique can be applied to the training and reasoning of artificial intelligence, with significant implications for future fields such as machine learning and data mining. Google has successfully achieved the milestone of more than 150 qubits of computing, predicting that quantum computers will become the core technology to create the next generation of computers.

In terms of technology trends, artificial intelligence, quantum computing and 5G network technology will become the main development directions of integrated circuits in the future. These areas require chips and modules that are more efficient, faster, more reliable and lower energy consumption to support their applications.

In brief, with the boost of artificial intelligence, quantum computing and 5G network technology the need of integrated circuits, I believe, will keep a rapidly growing. And the power consumption controlling will be a vital concern of the microelectronics industry.

References

[1] P. Hagouel, "Semiconductor material and device characterization [Book Review]," *IEEE Circuits Devices*, vol. 15, no. 3, p. P.36-36, 1999.

[2] J. M. Rabaey and M. Pedram, *Low Power Design Methodologies*. 1996. Accessed: Oct. 06, 2023. [Online]. Available: http://www.researchgate.net/publication/283374662_Low_Power

[3] A. Raghunathan, S. Dey, and N. K. Jha, "Register transfer level power optimization with emphasis on glitch analysis and reduction," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 18, no. 8, pp. 1114–1131, 1999.

[4] J. Li and J. F. Martinez, "Dynamic power-performance adaptation of parallel computation on chip multiprocessors," in *The Twelfth International Symposium on High-Performance Computer Architecture, 2006.*, IEEE, 2006, pp. 77–87.

[5] Foveros: 3D Integration and the use of Face-to-Face Chip Stacking for Logic Devices." Accessed: Nov. 07, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8993637/

[6] X. Liu, D. Wang, J. Zheng, P. Musavigharavi, and D. Jariwala, "Post-CMOS Compatible Aluminum Scandium Nitride/2D Channel Ferroelectric Field-Effect-Transistor," 2020, doi: 10.1021/acs.nanolett.0c05051.

[7] T. Mikolajick *et al.*, "FeRAM technology for high density applications," *Microelectron. Reliab.*, vol. 41, no. 7, pp. 947–950, 2001.

[8] B. Wu and A. Kumar, "Extreme ultraviolet lithography: A review," *J. Vac. Sci. Technol. B Microelectron. Nanometer Struct. Process. Meas. Phenom.*, vol. 25, no. 6, pp. 1743–1761, 2007.

[9] "TWINSCAN NXE:3400C." Accessed: Oct. 06, 2023. [Online]. Available: https://www.asml.com/en/products/euv-lithography-systems/twinscan-nxe3400c

[10] "ROHM speeding ahead in the GaN race – An interview by Yole Group." Accessed: Oct. 06, 2023. [Online]. Available: https://www.yolegroup.com/player-interviews/rohm-speeding-ahead-in-the-gan-race-an-interview-by-yole-group/