

INSPIRATION PAPER

The future of microchip technology

Contents

Servers & smartphones, the no.1 application drivers 3

Towards high-NA EUV lithography 3

More transistors on a chip thanks to nano & forksheets 4

How to improve the microchip's performance 5

Better systems thanks to heterogeneous integration 6

Memory chips: NAND, DRAM and beyond 7

The biggest trend: edge AI 8

How to evolve towards a greener IC manufacturing 9

Conclusion 10

Disclaimer

The content of this technology paper is intended exclusively for the personal information of the attendee of the Future Summits event 2022. Although imec strives to ensure that the information contained herein is meticulous, correct and complete, it cannot give any guarantee as regards the accuracy, precision and/or the completeness of the information: the information in this technology paper is provided "AS IS". Possible expectations and/or projections elaborated in the topics covered by this paper are based upon the current insights, scientific research and assumptions of imec. Moreover, the provided information cannot be considered in any way as an opinion or a recommendation from imec. Nothing in this paper constitutes professional, financial or investment advice, nor does it take into account individual needs, investment objectives and specific financial circumstances. imec does not accept any liability for any possible damage, loss, costs or expenses that might result from or could come about from the use of this publication and/or information contained in it.

© imec 2022

Servers & smartphones, the no.1 application drivers

In the past few decades, growth of the global semiconductor industry has been driven largely by the demand for cutting-edge electronic devices such as desktops, laptops and wireless communication products, and by the rise of cloud-based computing. Today, two new application drivers are on the horizon.

A first, obvious, application driver, is **the smartphone**. The chips inside, or better: the **systems-on-chips (SOCs)**, will evolve towards ever more functionality on a smaller chip area.

A much faster growing application driver is **the server** market. We know that the amount of data keeps on growing exponentially, but this will be accelerated by the rollout of **5G networks**. Servers will be needed to process and store these data. Also, **supercomputing** and **high-performance-computing-as-a-service** will rely heavily on servers in datacenters as well as emerging applications such as **cloud gaming and AI**.

At the heart of these servers are high-end central processing units (CPUs) and graphical processing units (GPUs). A 2020 Yole report foresees a compound annual growth rate of 29% for these units.

Beyond the traditional dimensional scaling of logic, memory and 3D interconnects, **cross-domain innovations** will be needed to make ever faster and more performant smartphones and servers. There is a need for new modules, new materials, new transistor architectures, as well as changes at SoC level to **realize the benefit at a system level**.

Towards high-NA EUV lithography

CMOS transistor density scaling will roughly continue to follow Moore's Law for the next eight to ten years, largely enabled by advances in **extreme ultraviolet (EUV) lithography**. It uses an extremely short wavelength of 13.5nm which allows to print features with smaller critical dimensions as compared to 193nm (immersion) lithography.

EUV lithography was introduced in the **7nm technology node** to pattern some of the most critical chip structures in one single exposure step. This shifted to multi-patterning EUV lithography beyond the **5nm node**, adding significantly to the wafer cost.

2023 will mark a new milestone in the evolution of EUV lithography. By then, the first new generation of EUV lithography tools is expected to enter the scene: a **high-NA EUV lithography scanner** – projected to print the most critical features of 2nm (and beyond) logic chips in a smaller number of patterning steps. This technique will push the multi-patterning of some of the layers back to single patterning, providing **cost, yield and cycle-time relief**.

The numerical aperture or NA controls the amount of light that is used to form the image, and thus controls the quality of the image. Next to the wavelength of the light used, this is a second' knob' for improving the resolution.

0.55NA EUV lithography will push the patterning towards features smaller than what is possible with current 0.33NA EUV lithography systems. But the road forward is ambitious. The development of EUV lithography systems goes back to the 2000s, with a ten-year time span between the installation of the first pre-production EUV scanners and the recent introduction of EUV lithography in high-volume manufacturing. For high-NA, the ambition is to compress that **time frame to only 3 years**, with a first prototype (the EXE:5000) foreseen for 2023.

Imec is partnering with ASML to open a **joint High-NA Lab** where the high-NA systems will be built, linked to a coat and development track and surrounded with metrology equipment. Together, they will create the ecosystem for the industry to meet the process requirements and establish the infrastructure that goes along with high-NA tool development.

One of the challenges to introduce high-NA EUV in industrial environments, is **stochastic defectivity**. These are printing failures (microbridges, locally broken lines, missing or merged contacts) that are random, non-repeating and isolated. By better understanding and mitigating these defects, lower dose exposures could be used, thus improving throughput and cost.

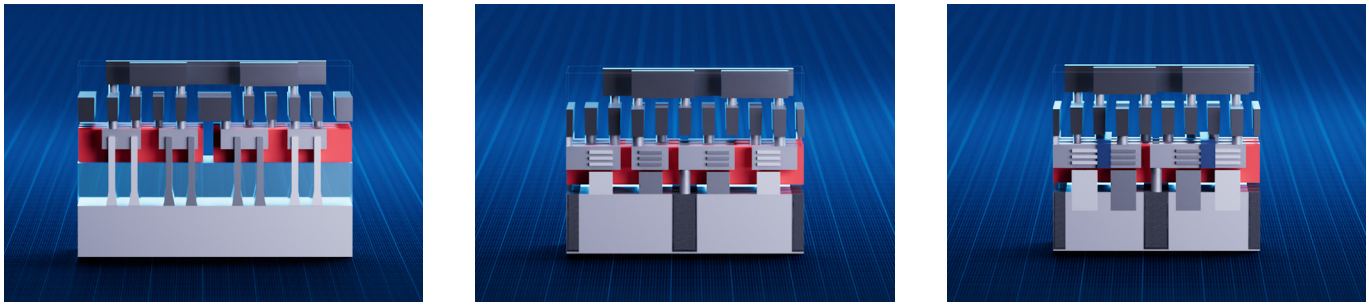
More transistors on a chip thanks to nano & forksheets

Apart from advancements in EUV lithography, Moore's Law cannot continue without **innovations in transistor architecture**. Today, **FinFET** devices are mainstream with the most advanced nodes having two fins an a 6-track (6T) standard cell.

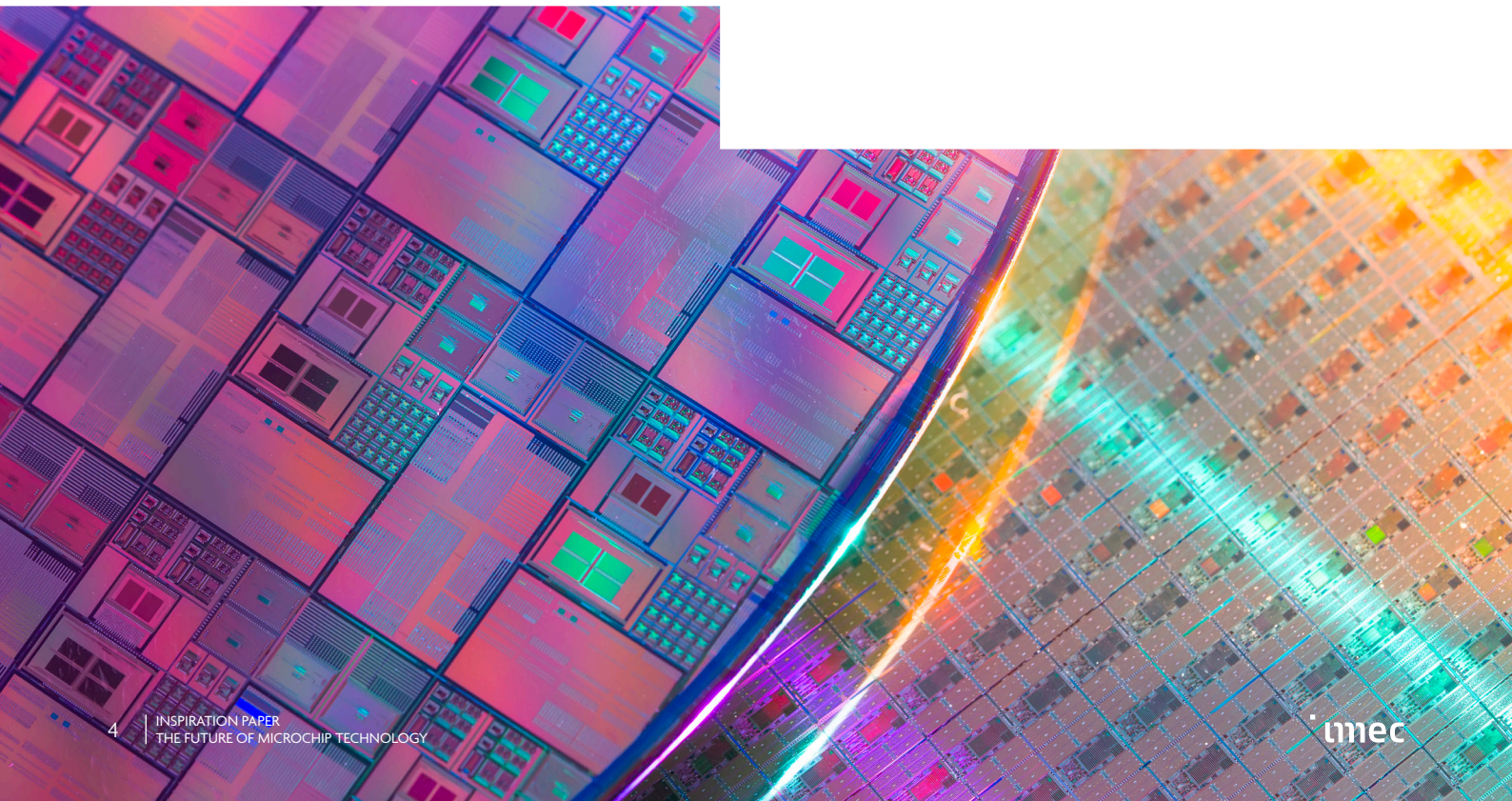
The next-generation device is a vertically stacked **nanosheet** device. In this kind of transistor, current flows through multiple stacks of silicon that are surrounded by the transistor gate. The design reduces avenues for current to leak through, and boosts the amount of current the device can drive.

Scaling nanosheets into the 2nm generation will be limited, and this is where the **forksheet** architecture comes on scene. Unlike nanosheet devices, the sheets are now controlled by a tri-gate forked structure – realized by introducing a dielectric wall in between the p- and nMOS devices before gate patterning. This wall physically isolates the p-gate trench from the n-gate trench, allowing a much tighter n-to-p spacing than what is possible with either FinFET or nanosheet devices.

The complementary FET or **CFET** completes the roadmap as the ultimate compact CMOS structure. The concept of CFET consists in 'folding' the nFET on top of the pFET (either fin-on-fin or sheet-on-sheet) – as such fully exploiting the possibilities of device scaling in 3D.



Natural evolution from FinFET to nanosheet, and to forksheet



How to improve the microchip's performance

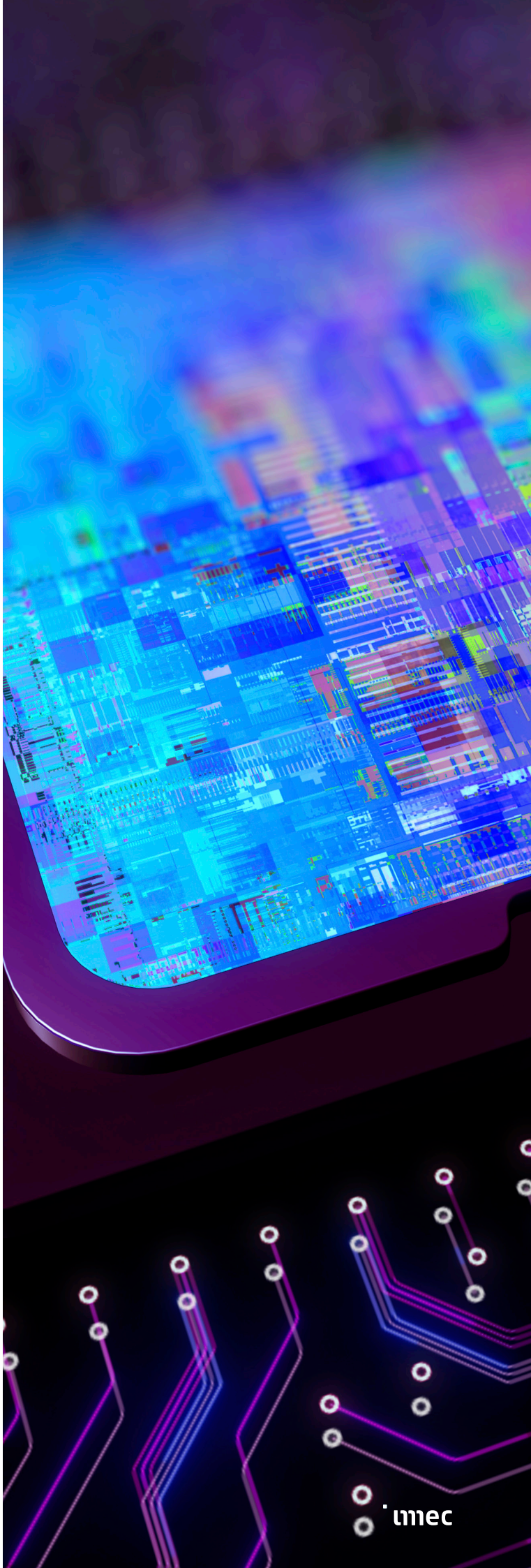
With the above-mentioned innovations, transistor density will follow the path mapped out by Gordon Moore. But node-to-node performance improvements at fixed power – referred to as Dennard scaling – have slowed down due to the inability to scale supply voltage.

One of the 'tricks' to boost performance at system level is the use of **buried power rails (BPRs)**. Power rails, which are part of the power delivery network, are traditionally implemented in the chip's back-end-of-line (BEOL). BPRs on the contrary are buried in the chip's front-end-of-line (FEOL) to help free up routing resources for the interconnects. Moreover, they provide a lower resistive local distribution of the current to a technology that suffers from increasing BEOL resistance with pitch scaling. BPRs improve the system's performance by improving the power distribution. Integrating BPRs within the front-end module is however challenging, as BPR processing may induce stress in the conduction channel – degrading the device mobility – or cause metal contamination issues.

At transistor level, several adjustments are worth considering too.

- Incorporating stress into nanosheet and forksheet devices
- Applying **high-mobility materials** in CFETs
- Using **2D materials** in the transistor's channel. For example, tungsten disulfide promises performance improvements because it enables more aggressive gate length scaling than Si or SiGe

Besides the FEOL, routing congestion and RC delay in the BEOL have become important bottlenecks for performance improvement. Hybrid metallization using Ru or Mo is a way to improve via resistance.



Increased system functionality thanks to 3D & heterogeneous integration

One way to further scale microchips or ICs and add more functionality per area, is to **integrate circuits on top of each other**, leveraging 2.5D or 3D connectivity. Additionally, this may help solve a number of other challenges:

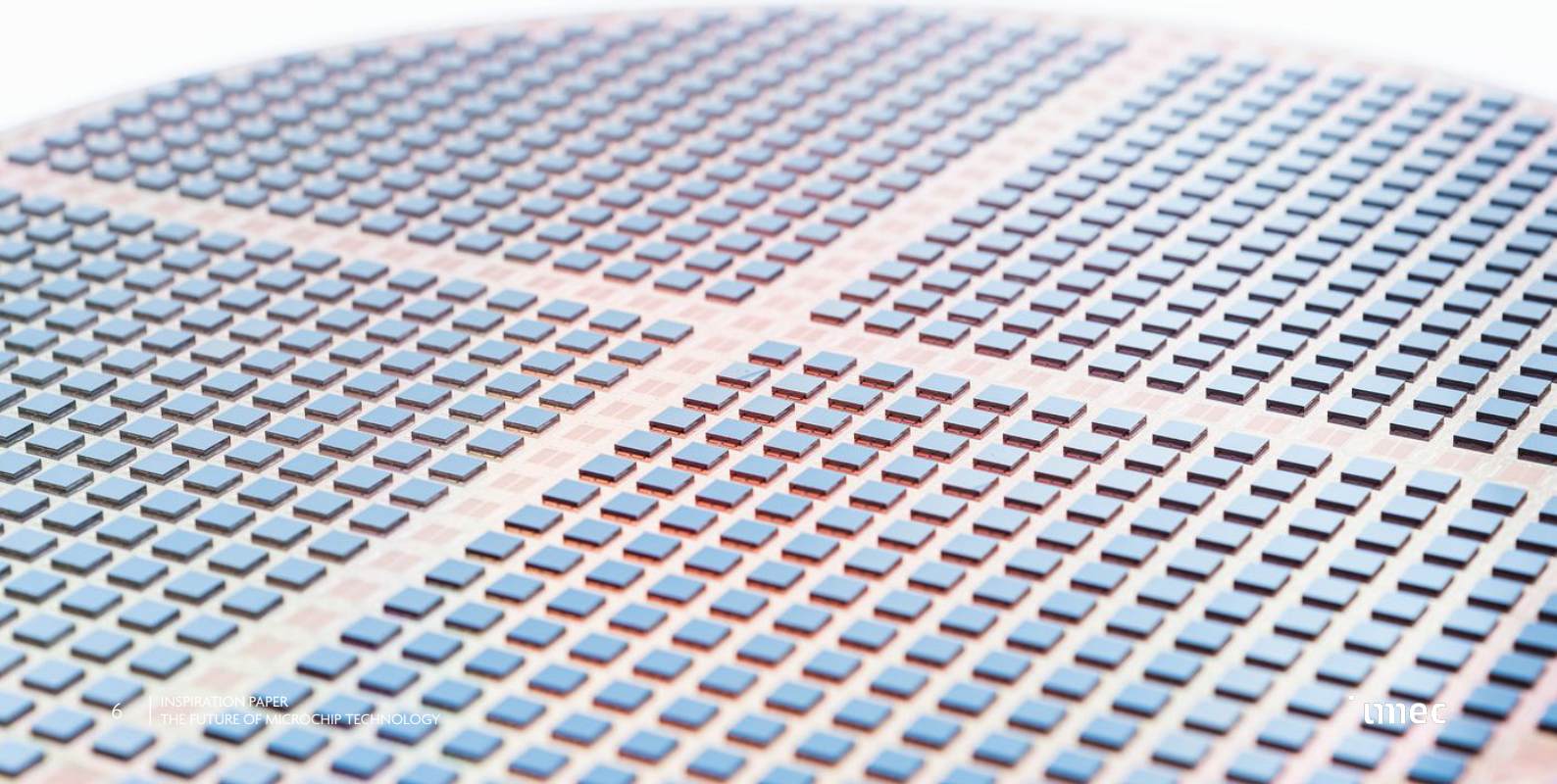
- Provide shorter paths between the processor and its memory, helping to break down the memory wall - the growing disparity of speed between CPU and memory outside the CPU chip.
- Boost functionality in ICs that are limited in size.
- Improve the yield for large, complex systems.

3D integration is also a way to **combine heterogeneous subsystems**, i.e. dies made from different technologies. It may be more advantageous to use different process technologies for subsystems such as logic, memory, I/O interfaces, analog, etc. for example to optimize the design costs and yield. This evolution also answers the needs for more IC diversification and customization.

3D integration can be achieved in several ways. One of these is **die-to-wafer** bonding, using microbumps. The state-of-the-art microbump pitches in production have saturated at about 30µm. Using e.g. material innovations, interconnect pitches down to 7µm could be demonstrated. Such high-density connections leverage the full potential of through-silicon via technology (electrical interconnects that are etched into a silicon wafer), enabling **more than 16 times higher 3D interconnect densities**. Direct bonding of die to silicon is also possible using hybrid Cu bonding. Pitches down to 3µm are possible in this case, with a high tolerance pick-and-place accuracy.

Another 3D integration technology is **wafer-to-wafer** bonding, enabling even more dense interconnect pitches (500nm). Wafer-to-wafer bonding also allows for **extreme thinning of Si substrates** and the realization of connections through the silicon in the 100nm range. These nano-vias allow for direct power delivery to the logic devices from the wafer's back side, freeing-up space for wiring on the front side.

Die-to-wafer bonding



Memory chips: NAND, DRAM and beyond

To enable fast, low-temperature and powerful storage solutions for next-generation smartphones, tablets, smart TVs and other pioneering technologies, research focus is on memory concepts with **increased memory** density, while at the same time **controlling temperature**.

NAND storage will continue to scale incrementally, without disruptive architectural changes in the next few years. Today's most advanced NAND products feature 128 layers of storage capability. The 3D scaling will continue with additional layers potentially enabled by wafer-to-wafer bonding.

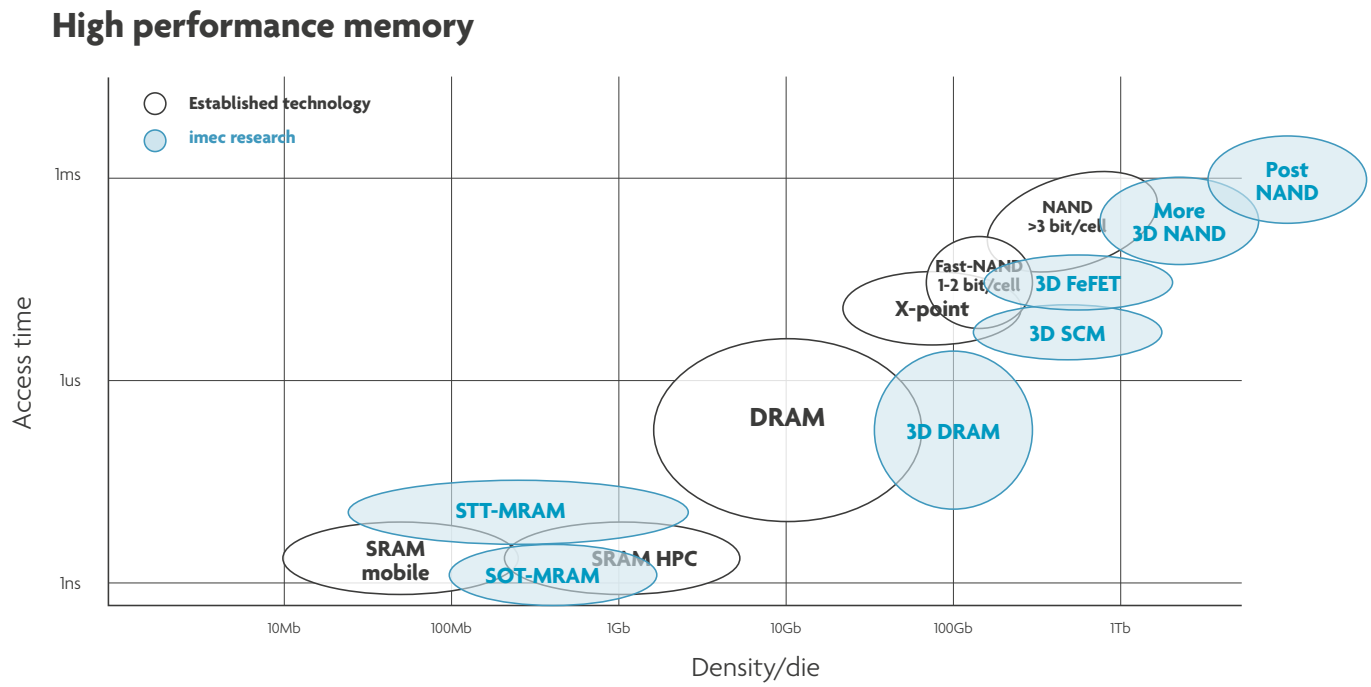
Future developments will focus on using metals like ruthenium, researching alternate memory dielectric stacks, improving channel current and identifying ways to control the stress that evolves due to the growing number of stacked layers. Also, the planar logic transistors in the NAND periphery can be replaced with more advanced FinFET devices.

For **DRAM**, cell scaling is slowing down, and EUV lithography may be needed to improve patterning. True 3D DRAM solutions are also being explored, e.g. by putting the memory array on top of the periphery. Such an architecture requires a low-thermal-budget-deposited semiconductor for the array transistors. And this is where the low-temperature IGZO (or indium-gallium-zinc-oxide) family of transistors enter the scene. Ultimate 3D DRAM implementation will also require these materials to be deposited on topography. This drives the need for atomic layer deposition (ALD) for the layer formation. Finally, just as with NAND, a FinFET-based periphery with high-k/metal gate structures is explored as a replacement for planar transistors with poly-Si gates.

In the **embedded memory** landscape, there are significant efforts to understand – and eventually tear down – the so-called memory wall: how quickly can the CPU access data from DRAM or from SRAM-based caches? How do you ensure cache coherency with multiple CPU cores accessing a shared cache? What are the bottlenecks that limit speed, and how can we improve the bandwidth and data protocols that are used to fetch the data?

Novel memories are being explored in this context, including spin-transfer torque (STT)-MRAM, spin-orbit torque (SOT)-MRAM and voltage controlled magnetic anisotropy (VCMA)-MRAM. Each of these MRAM memories comes with its own benefits and challenges and may help overcome the memory bottleneck by improving speed, power consumption and/or memory density. To further improve density, selector devices can be used, to be integrated with the magnetic tunnel junctions – which are at the heart of these MRAM devices.

A view on the memory landscape and how imec contributes



The biggest trend: edge AI

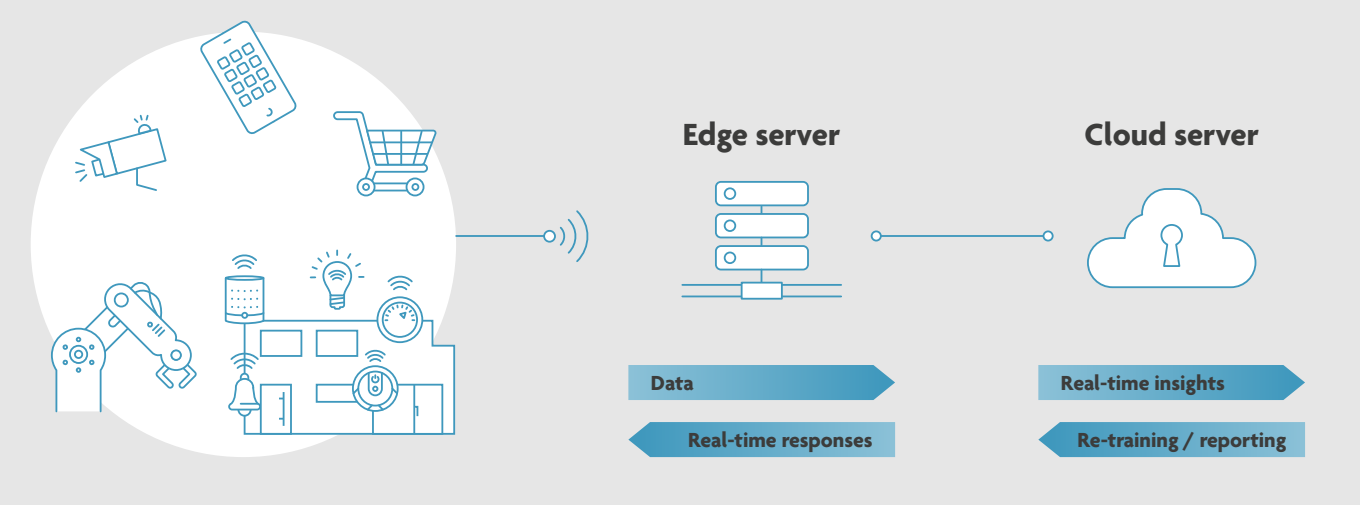
With an expected growth of above 100% in the next five years, **edge AI** is one of the biggest trends in the chip industry. Whereas in cloud AI extremely powerful computers crunch enormous quantities of data, edge AI resides in the **trillions of IoT devices** making our environment smart. The IoT devices communicate wirelessly with an **edge server** that is located relatively close to the IoT device. This server decides what data will be sent to the cloud server (typically, data needed for less time-sensitive tasks, such as re-training) and what data gets processed on the edge server.

Compared to cloud-based AI, in which data needs to move back and forth from the endpoints to the cloud server, edge AI addresses **privacy** concerns more easily since not all the data is sent to the cloud. It also offers advantages of **response speeds** and reduced cloud server workloads.

Most of the IoT devices will be extremely constrained when it comes to **power budget, cost, robustness**, ... In this way, edge AI will need innovation on all levels: from computational concepts down to the implementation technology.

Today, commercially available edge AI chips offer efficiencies in the order of 1 to 100 Tops/W, using fast GPUs or ASICs for computation. For IoT implementations, much higher efficiencies will be needed, in the order of **10,000 Tops/W**.

This could be achieved by looking at **analog compute-in-memory architectures**. This approach breaks with the traditional Von Neumann computing paradigm that is based on sending data from memory to CPU (or GPU) for computations. With analog compute-in-memory, computation is done inside a memory framework, saving a lot of power in moving data back and forth.



The edge AI framework

How to evolve towards a greener IC manufacturing

The semiconductor industry is resource-intensive in energy, water, chemicals, and raw materials. During manufacturing, various classes of emissions are being generated, including greenhouse gases such as CO_2 and fluorinated compounds. To minimize the industry's impact on the environment and to comply with local and global policies, environment, health and safety (EHS) control has become a critical component of every foundry for quite some time.

But due to growing concerns on climate change, resources depletion and global pollution, **fabs and equipment suppliers want to make more efforts** towards greener IC manufacturing. While EHS control is mainly limited to chemicals, abatement and water management, electronics companies want to **know and reduce the full ecological footprint of their product**. A reduced footprint might also guarantee **business continuity** – for example if scarce materials are involved – or give the company a competitive advantage.

The first step towards a 'greener' IC manufacturing is finding a way to quantify that footprint and its constituting elements. Imec has set up a **research program – sustainable semiconductor technologies and systems (SSTS)**. The SSTS program is the first initiative rallying stakeholders from across the IC value chain to anticipate the environmental impact of choices made at chip technology's definition phase. Using concrete and reliable models, and detailed (carbon) footprint analyses, the program helps the IC-making industry cut back on its ecological footprint – as part of the global fight against climate change, resources depletion and pollution.

Within the program a software tool has been developed to estimate energy consumption, water usage and greenhouse gas emissions for present and future logic CMOS technologies. That information makes it possible to take sustainability into account when making technology choices, for instance by compensating the increased use of electricity with more renewable energy use, or steering research in the direction of processes with lower emission levels.

A true **holistic approach** is key. Take for example **EUV lithography**. On the one hand, EUV tools consume more power per tool than 193nm lithography. The throughput of EUV – expressed in wafers per hour – is also lower than for 193nm lithography, enlarging the energy consumption envelop. But on the other hand, EUV single patterning significantly reduces the number of process steps when compared to using 193nm lithography in combination with complex multi-patterning techniques such as self-aligned quadrupole patterning (SAQP). The new framework reveals that a target throughput of 110 wafers per hour for EUV is needed to reach the same electrical energy usage as for SAQP when patterning critical BEOL metal lines.

Imec's SSTS program draws on its widely acclaimed fab expertise and combines the insights in infrastructure, technology and machinery to help the complete IC value chain reduce its ecological footprint. Imec is building a whole **ecosystem of partners**, to inform them of the environmental impact of certain choices made at chip technology's definition phase. As announced in a press release of October 28, 2021, **Apple Inc.** has joined imec's SSTS research program.



imec



Conclusion

Moore's Law is still alive and will continue to deliver ultra-scaled technologies for the next 10 years. Applications such as **smartphones and servers** will benefit most from the next-generation transistors and systems. The main enabler for this continued scaling is EUV lithography, which will evolve towards a higher numerical aperture with **high-NA lithography**. Also, innovations in the front-end-of-line keep on delivering new transistor architectures such as **nano- and forksheets and CFETs**. Next to that, buried power rails, material science (2D and high-mobility materials, stress induction in materials, ...) and **heterogeneous integration** help to increase the system's performance.

New memory concepts and memory innovations are another pilar for future applications, with the aim to tear down the memory wall. And also with **edge AI**, where computation is done near the user at the edge of the network, close to where the data is located, rather than centrally in a cloud computing facility or private data center, a new way of using memory takes the central stage. **Analog compute-in-memory** architectures promise to be the answer for the so much needed power savings for realizing AI in the edge.

Finally, the **green wave** has also hit the chip manufacturing industry with new tools being developed to quantify the footprint and estimate energy consumption, water usage and greenhouse gas emissions for present and future logic CMOS technologies.