INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS™

2022 UPDATE

MORE MOORE

THE IRDS IS DEVISED AND INTENDED FOR TECHNOLOGY ASSESSMENT ONLY AND IS WITHOUT REGARD TO ANY COMMERCIAL CONSIDERATIONS PERTAINING TO INDIVIDUAL PRODUCTS OR EQUIPMENT.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

MORE MOORE TEAM

U.S.A. & Canada	
Anshul A. Vyas	Applied Materials
Arvind Kumar	IBM
Bich-Yen Nguyen	Soitec
Charles Kin P. Cheung	NIST
Chorng-Ping Chang	AMAT
Christopher Henderson	Semitracks
Gennadi Bersuker	Aerospace Corporation
Gerhard Klimeck	Purdue Univ.
Huiming Bu	IBM
James Stathis	IBM
Jim Fonseca	Purdue Univ.
Joe Brewer	Univ. Florida
Joel Barnett	TEL
Kwok Ng	SRC
Lars Liebmann	Intel
Masako Kodera	Mosos Lake Industries
Matthias Passlack	TSMC
Mehdi Salmani	Boston Consulting Group
Philip Wong	Stanford Univ.
Prasad Sarangapani	Synopsys
Qi Xiang	Xilinx
Rich Liu (co-chair)	**Macronix**
SangBum Kim	IBM
Saumitra Mehrotra	NXP
Sergei Drizlikh	Samsung
Siddharth Potbhare	NIST
SungGeun Kim	Microsoft
Takeshi Nogami	IBM
Wilman Tsai	Stanford Univ.
Witek Maszara	V-tek Consulting
Yanzhong Xu	Intel Corporation
Yuzo Fukuzaki (co-chair)	**Tech Insights**

ASIA	
Atsushi Hori	Tokyo Inst of Technology
Digh Hisamoto	Hitachi
Hajime Nakabasyashi	TEL
Hiroshi Iwai	Tokyo Inst of Technology
Hitoshi Wakabayashi	Tokyo Inst of Technology
Jiro Ida	Kanazawa IT
Kazuyuki Tomida	Sony
Kosuke Tsukamoto	Shinko Electric Industries
Kunihiko Iwamoro	ROHM
Kuniyuki Kakushima	Tokyo Inst of Technology
Masahiko Ikeno	Hitachi High-Tech
Masami Hane	Renesas
Shinichi Ogawa	AIST
Shinichi Takagi	University of Tokyo
Takashi Matsukawa	AIST
Tatsuya Ohguro	Toshiba
Tetsu Tanaka	Tohoku University
Tesuo Endoh	Tohoku University
Toshiro Hiramoto	University of Tokyo
Yasuo Kunii	Analysis Atelier
Yasushi Akasaka	TEL
Yoshihiro Hayashi	Keio University
Jongwoo Park	Samsung
Moon-Young Jeong	Samsung
Cheng-tzung Tsai	UMC
Geoffrey Yeap	TSMC

EUROPE	
Christiane Le Tiec	MKS Instruments
Francis Balestra	IMEP Grenoble
Fred Kuper	NXP
Gerben Doornbos	TSMC
Jurgen Lorenz	Fraunhofer IISB
Karim Cherkaoui	Tyndall National Inst.
Krishna Bhuwalka	Huawei
Laurent Le-Pailleur	ST
Malgorzata Jurczak	LAM Research
Mark van Dal	TSMC
Mustafa Badaroglu (chair)	**Qualcomm Technologies**
Olivier Faynot	CEA-LETI
Paul Mertens	IMEC
Peter Ramm	Fraunhofer EMFT
Ray Duffy	Tyndall National Inst.
Robert Lander	NXP
Thierry Poiroux	CEA-LETI
Yannick Le Tiec	CEA-LETI

MORE MOORE

1. INTRODUCTION

System scaling enabled by Moore's scaling is increasingly challenged by the scarcity of resources such as power and interconnect bandwidth. This has become more challenging under the requirements of seamless interaction between big data and instant data (Figure MM-1). Instant data generation requires ultra-low-power devices with an "always-on" feature at the same time with high-performance devices that can generate the data instantly. Big data requires abundant computing, communication bandwidth, and memory resources to generate the service and information that clients need.

The More Moore International Focus Team (IFT) of the International Roadmap of Devices and Systems (IRDS) provides physical, electrical, and reliability requirements for logic and memory technologies to sustain power, performance, area, cost (PPAC) scaling for big data, mobility, and cloud (e.g., Internet-of-Things (IoT) and server) applications. This is done over a time horizon of 15 years for mainstream/high-volume manufacturing (HVM).

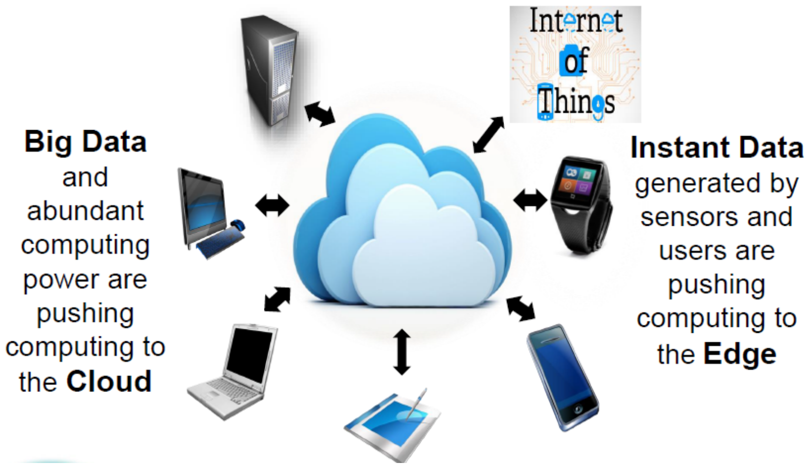

Internet of Things
Big Data and abundant computing power are pushing computing to the Cloud
Instant Data generated by sensors and users are pushing computing to the Edge

Figure MM-1 Big data and instant data

Following systems drivers is forecasted to impact the More Moore logic technologies:

Mobile

- Heterogenous integration
- Edge computing
- Extreme reality (VR/AR)
- AI enhanced edge computing and connectivity (mobile phone, 6G, cellular, IoT, Wi-Fi, wireless connectivity, smart cameras and speakers) driving any data, any place, highest speed and lowest power with content-rich data

Data and HPC servers – cache integration, memory, IO

- AI accelerators in enterprise/cloud
- Codec ASICs - 24/7/365 continuous run of video and audio (codec), 5 years minimum time
- Networking – Always-on, 500W power envelope
- Advanced driver assistance system (ADAS) chips – Autonomous driving
- Memory and IO solutions for AI, graphics, HPC

Novel compute fabrics

- Neural processing unit

- Fine-pitch 3D stacking
- Reconfigurable compute fabrics
- Smart 2.5D interposers

1.1. CURRENT STATE OF TECHNOLOGY

A major portion of semiconductor device production is devoted to digital logic that needs to support a technology platform for two device types: 1) high-performance logic, and 2) low-power/high-density logic. Key considerations for this technology platform are speed, power, density, cost, capacity, and time-to-market. The More Moore roadmap provides an enablement view for continued scaling of MOSFETs in order to maintain historical trends of improved device performance at reduced power and cost and at high volume.

1.2. DRIVERS AND TECHNOLOGY TARGETS

The following applications drive the requirements of More Moore technologies that are addressed in the IRDS [1]:

- High-performance computing—more performance at constant power density (constrained by thermal)
- Mobile computing—more performance and functionality at constant energy (constrained by battery) and cost
- Autonomous sensing and computing (IoT)—targeting reduced leakage and variability

Technology drivers include following focal items:

- Logic technologies
- Ground rule scaling
- Performance boosters
- Performance-power-area (PPA) scaling
- 3D integration
- Memory technologies
- DRAM technologies
- Flash technologies
- Emerging non-volatile-memory (NVM) technologies

More Moore targets bringing PPAC value for node scaling every 2−3 years [2]:

- (P)erformance: >10% more operating frequency at scaled supply voltage
- (P)ower: >20% less energy per switching at a given performance
- (A)rea: >30% less chip area footprint
- (C)ost: <30% more wafer cost – 15% less die cost for scaled die.

System scaling considers the co-integration of logic, memory and IO solutions bringing the following targets:

- TOPS (Tera Operations per Second): Throughput
- TOPS/W (TOPS per Watt): Energy efficiency
- TOPSxTOPS/W/Area is an indicator of Energy-Area-efficient performance (aka EDP: Energy Delay Product at Unit Area)
- 2.0-2.4x scaling of TOPSxTOPS/W/Area for node scaling per frame, per inference, per training, and/or per pocket

These scaling targets have driven the industry toward a number of major technological innovations, including material and process changes such as high-κ gate dielectrics and strain enhancement, and in the near future, new structures such as gate-all-around (GAA); alternate high-mobility channel materials, and new 3D integration schemes allowing heterogeneous stacking/integration. These innovations will be introduced at a rapid pace, and hence understanding, modeling, and implementation into manufacturing in a timely manner is crucial for the industry.

It is important to note that both cost metric (15% less die cost) and market cadence necessitating new products at high volume every year are becoming more important targets in the mobile and high-performance computing industry. As the applications strictly requiring all figure-of-merits (FoMs) are concurrently met, it is necessary to advance an effective list of process technologies for sustaining certain device architectures to their limits, such as pushing the finFET architecture until 2025 while ensuring a swift transition to gate-all-around devices, which will sustain more than another decade. This approach will also help in sustaining the cost at reduced risk while moving from one logic generation to another. This becomes more difficult whenever the cost of wafer processing becomes more expensive with the increased number of steps because of multiple patterning lithography steps. However, it is necessary to reduce the cost by more than 15% at every logic generation for the same of number of transistors, which can only be enabled by pitch scaling enabled by new advancements in channel material, device architecture, contact engineering, and device isolation. Increased process complexity must also be taken into account for the overall die yield. In order to compensate the cost of complexity, acceleration in design efficiency is needed to further scale the area to reach the die-cost scaling targets. These design-induced scaling factors were also observed in the earlier work of the System Drivers Technology Workgroup of ITRS and those were used as calibration factors to match the area scaling trends of the industry [2]. The design scaling factor is now considered as one of the key elements in the More Moore technology roadmap.

2. Summary and Key Points

The following are forecasted in the projected IRDS More Moore roadmap:

- Ground rule scaling is expected to slow down and saturate around 2028. Extreme-ultraviolet (EUV) technology will be the enabler of ground rule scaling while keeping the cost under control and providing process complexity reduction. Transition to 3D integration and use of beyond CMOS devices for complementary System-on-Chip (SoC) functions are projected after 2028.
- Ground-rule scaling needs to be accompanied with the design-technology-co-optimization (DTCO) constructs that accommodate the area reduction as well as tightening the critical design rules that limit the overall SoC area scaling.
- A main challenge in 3D integration is how to partition the system to come up with better utilization of devices, interconnect, and sub-systems such as memory, analog, and input/output (I/O). Parasitics improvement will become a major knob for performance improvement for nodes spanning between 2022 and 2028, such as with the introduction of low-κ device spacer.
- SiGe and 2D material channels are gaining importance complementing the Si channels.
- It becomes increasingly difficult to control interconnect resistance, electromigration (EM), and time-dependent-dielectric-breakdown (TDDB) limits. Interconnect resistance has now entered an exponential increase regime because of non-ideal scaling of barriers for Cu bringing less metallization volume and increased scattering at the surface and grain-boundary interfaces. Therefore, there is a need for new barrier materials, atomic layer deposition (ALD) based barrier deposition, and/or non-Cu metallization solutions. In addition to the resistance scalability, TDDB is putting a limit on the minimum space between the adjacent lines for a given low-κ dielectric, forcing a slow-down in the permittivity (κ-value) scaling.
- Performance across six nodes spanning from 2022 to 2037 is forecasted to improve on average for wireloaded datapaths, most of improvements taking place as transition from 3 to 4 GAA devices around 2031.
- System-on-chip (SoC) level area across six nodes spanning from 2022 to 2037 is forecasted to improve, but less than 30%, node-to-node on average.
- Power density poses a significant challenge for scaling, particularly as a result of 3D integration after 2031. Therefore, it is necessary to factor in thermal considerations in device and architectures.
- Energy per switching reduction is expected to be limited less than 20% in a node-to-node basis on average. This is a critical challenge of scaling because of a slow-down in capacitance and supply voltage reduction.
- DRAM needs to maintain sufficient storage capacitance and adequate cell transistor performance is required to keep the retention time characteristic in the future. If efficiency of cost scaling becomes poor in comparison with introducing the new technology, DRAM scaling will be stopped and 3D DRAM cell stacking structure will be adopted. Alternatively, a new DRAM concept could be adopted.
- 2D Flash memory density cannot be increased indefinitely by continued scaling of charge-based devices because of controllability limits of threshold voltage distribution. Flash density increase will continue by stacking memory layers vertically, leading to adoption of 3D Flash technology. Decrease in array efficiency due to increased interconnection and yield loss from complex processing are challenges for further reducing the cost-per-bit benefit.

Currently, 96 layers are already at volume production and there is optimism that 128 layers are achievable, with 192 and 256 layers possible.

- Ferroelectric RAM (FeRAM) is a fast, low power, and low voltage non-volatile memory (NVM) alternative and thus is suitable for radio frequency identification (RFID), smart card, ID card, and other embedded applications. Processing difficulty limits its wider adoption. Recently, HfO_2-based ferroelectric field-effect transistor (FET), for which the ferroelectricity serves to change the threshold voltage (Vt) of the FET and thus can form a 1T cell similar to Flash, has been proposed. If developed to maturity, this may serve as a low-power and very fast, Flash-like memory.
- Spin-transfer torque-magnetic RAM (STT-MRAM) to replace the stand-alone NAND Flash seems remote. STT-MRAM is now mostly considered not as a standalone memory but an embedded memory. STT-MRAM would also be a potential solution for embedded Flash (NOR) replacement. This may be particularly interesting for low-power IoT applications. On the other hand, for other embedded systems applications using higher memory density, NOR Flash is expected to continue to dominate, since it is still substantially more cost-effective and well established for being able to endure the printed circuit board (PCB) soldering process (at ~250°C) without losing its preloaded code.
- 3D cross-point memory has been demonstrated for the storage class memory (SCM) to improve I/O throughput and reduce power and cost. Since the memory including the selector device is completely fabricated in the back-end-of-line (BEOL) process it is relatively inexpensive to stack multiple layers to reduce bit cost.
- High-density resistive RAM (ReRAM) development has been limited from the lack of a good selector device, since simple diodes have limited operation ranges. Recent advances in 3D cross point memory, however, seem to have solved this bottleneck and ReRAM could make rapid progress if other technical issues, such as erratic bits, are solved.
- PCM provides a good scaling trajectory overcoming the burdens such as high variability in RRAM and low Ron/Roff ratios in MRAM.
- More convergence of embedded memories with computing is expected giving the performance and energy losses by moving data from/to the memory to/from the compute, which is the so-called memory-wall problem. This will bring in compute-in-memory (CIM) arrays to evolve in particular edge-AI applications. CIM arrays will also take advantage of embedded NVM devices that could be integrated in the BEOL stack bringing a better area footprint for improvement of the TOPS/mm2 metric.

The links to the tables of technology roadmaps for Logic Core Device, DRAM, Flash, and NVM are below:

Table MM-1	*More Moore—Logic Core Device Technology Roadmap*
Table MM-2	*More Moore—DRAM Technology Roadmap*
Table MM-3	*More Moore—Flash Technology Roadmap*
Table MM-4	*More Moore—NVM Technology Roadmap*

3. CHALLENGES

The goal of the semiconductor industry is to be able to continue to scale the technology to improve overall performance at reduced power and cost. The performance of the components and the final chip can be measured in many different ways: higher speed, higher density, lower power, form factor reduction, bill-of-material reduction, more functionality, etc. Traditionally, dimensional scaling had been adequate to bring about these aforementioned performance merits, but it is no longer the case. Processing modules, tools, material properties, etc., are presenting difficult challenges to continue scaling. We have identified these difficult challenges and summarized in Table MM-5 and Table MM-6. These challenges are divided into near-term 2022-2028 (Table MM-5) and long-term 2029-2037 (Table MM-6).

3.1. NEAR-TERM CHALLENGES

Table MM-5 Difficult Challenges—Near-term

Near-Term Challenges: 2022-2028	**Description**
Power scaling	Voltage and capacitance scaling slow down and lack of solutions for power reduction. Introduction of gate-all-around (GAA) devices is a remedy to reduce the supply voltage, but not in a sustained manner that allows continuous scaling. Power scaling is also limited because of slow-down of loading capacitance scaling, impacted by increasing portion of parasitic components with scaling. Therefore, an introduction of low-κ materials, design-technology-co-optimization (DTCO) introducing new contact access schemes, novel power landing vias, is needed.
Parasitics scaling	Maintaining control of increased parasitics in stacked devices. Stacked devices require high-aspect ratio contacts to access the bottom contact. This will increase both the contact resistance as well as the fringe capacitance between the gate and drain/source. Interface resistance will also require new silicidation schemes that conformally wrap the source/drain.
Cost reduction	Cost-effective area scaling through EUV, DTCO, and 3D stacking. Throughput and yield challenges of EUV necessitate a careful selection of ground rules that optimize the die cost as a significant part of cost is determined by the middle-of-line (MOL) and BEOL stack. Therefore, new process-enhanced design constructs that tighten the secondary design rules such as tip-to-tip and the P-N separation rule are necessary to allow a further shrink of the standard cell and bitcell area on top of ground rule scaling for low-cost die. Process integration of those design constructs might require new materials to allow better etch selectivity and self-deposition. Slow-down in SRAM density scaling brings disintegration of last-level-cache and/or buffer memories as a stacked die on logic.
Integration enablement for SRAM-cache applications	Bitcell scaling is slowing down because of the slow-down of the device (e.g., fin) pitch and gate pitch (i.e., contacted poly pitch (CPP)). New device schemes such as P-over-N stacked device (CFET) bring an opportunity to significantly reduce the SRAM area. This is enabled because of optimized layouts that eliminate the critical design rules impacting the area. Option of embedded NVM in high-performance logic. Being able to integrate most of emerging memories (e.g., MRAM) at the interconnect stack also bring an opportunity for high-density memories. However, the stack as well as the materials should be compatible with BEOL.
Interconnect scalability	Maintaining control of interconnect resistance and EM and TDDB limits. Interconnect resistance has now entered an exponential increase regime because of non-ideal scaling of the barrier for Cu and increased scattering at the surface and grain-boundary interfaces. Therefore, there is a need for new barrier materials and Cu alternative solutions. In addition to resistance scalability, TDDB is putting a limit on the minimum space between the adjacent lines for a given low-κ dielectric.

3.2. Long-term Challenges

Table MM-6 Difficult Challenges—Long-term

Long-Term Challenges: 2028-2037	**Description**
Power scaling	Power scaling — no solutions are left besides steep-subthreshold (SS) devices to enable complementary SoC functions bringing power reduction but replacing mainstream CMOS. However, most of steep-SS device candidates do not bring an adequate performance comparable to CMOS at nominal supply voltages. New architectures are necessary to attain the performance through parallelization and also more fine-pitch 3D stacking with increased stacked memory capacity is required to reduce the amount of less energy-efficient external memory accesses.
Thermal issue due to increased power density	Thermal challenges (e.g., power density and dark silicon) of 3D stacking. Gate-all-around (GAA) devices have limited heat conductance due to confined architecture. Increased pin density due to aggressive standard cell height scaling and increased drive by stacked devices put a significant pressure on the power density.
Cost reduction with 3D integration	Managing cost, yield, and process complexity of 3D integration. Using vertical devices separated by the interconnect significantly increases the wafer cost and the number of masks (i.e., process complexity) adding pressure to the defectivity (e.g., D0) control. Architectures need to be refined for reducing the interconnect complexity between tiers as well as simplified integration and function per tier (e.g., I/O in one tier, SRAM in another tier, etc.).
Integration of non-Cu metallization to replace Cu	Adoption of non-Cu interconnects for low-resistance, meeting EM/TDDB, and temperature budget compatibility with devices used in 3D integration. Introduction of buried rail for power distribution will require careful assessment of material, thermal, and stress considerations as this metallization will be in close proximity to logic devices.

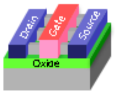
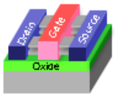
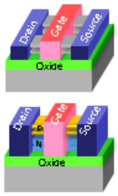
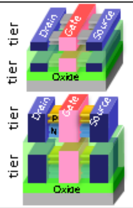
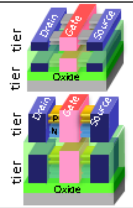
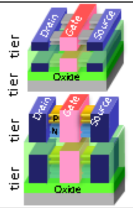
4. Technology Requirements—Logic Technologies

4.1. Ground Rules Scaling

The More Moore roadmap focuses on effective solutions to sustain the performance and power scaling at scaled dimensions and scaled supply voltage. Ground rule scaling drives die-cost reduction. However, this scaling increases the portion of parasitics in the total loading and brings diminishing returns of scale in performance and power scaling. Therefore, it is necessary to focus on technology scaling solutions that also scale the parasitics of device and interconnect. Ground-rule scaling needs to also enable DTCO constructs that accommodate the area reduction as well as tighten the critical design rules that limit the area scaling. Due to the rising costs and process complexity of multiple patterning, EUV is used as a remedy to pattern-tight ground rules in fewer process steps. The projected roadmap of ground rules as well as device architectures is shown in Table MM-7. Evolution of ground rules in shown Figure MM-2. There is not yet a consensus on the node naming across different foundries and integrated device manufacturers (IDMs); however, the projected rules give an indication of technology capabilities in line with the PPAC requirements. Key parameters in the ground rules are the gate pitch, metal pitch, fin pitch, gate length, and 3D tier stacking capability, which are important factors in core logic area scaling.

Table MM-7 Device Architecture and Ground Rules Roadmap for Logic Devices.

Note: GxxMxx/Tx notation refers to Gxx: contacted gate pitch, Mxx: tightest metal pitch in nm, Tx: number of tiers. This notation illustrates the technology pitch scaling capability. On top of pitch scaling there are other elements such as cell height, number of stacked devices, DTCO constructs, 3D integration, etc. that define the target area scaling (gates/mm^2).

YEAR OF PRODUCTION	2022	2025	2028	2031	2034	2037
	G48M24	G45M20	G42M16	G40M16/T2	G38M16/T4	G38M16/T6
Logic industry "Node Range" Labeling	"3nm"	"2nm"	"1.5nm"	"1.0nm eq"	"0.7nm eq"	"0.5nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI
Logic device structure options	finFET LGAA	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM
Platform device for logic	finFET	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D
						
LOGIC DEVICE GROUND RULES						
Mx pitch (nm)	32	24	20	16	16	16
M1 pitch (nm)	32	23	21	20	19	19
M0 pitch (nm)	24	20	16	16	16	16
Gate pitch (nm)	48	45	42	40	38	38
Lg: Gate Length - HP (nm)	16	14	12	12	12	12
Lg: Gate Length - HD (nm)	18	14	12	12	12	12
Channel overlap ratio - two-sided	0.20	0.20	0.20	0.20	0.20	0.20
Spacer width (nm)	6	6	5	5	4	4
Spacer k value	3.5	3.3	3.0	3.0	2.7	2.7
Contact CD (nm) - finFET, LGAA	20	19	20	18	18	18
Device architecture key ground rules						
Device lateral pitch (nm)	24	26	24	24	23	23
Device height (nm)	48	52	48	64	60	56
FinFET Fin width (nm)	5.0					
Footprint drive efficiency - finFET	4.21					
Lateral GAA vertical pitch (nm)		18.0	16.0	16.0	15.0	14.0
Lateral GAA (nanosheet) thickness (nm)		6.0	6.0	6.0	5.0	4.0
Number of vertically stacked nanosheets on one device		3	3	4	4	4
LGAA width (nm) - HP		30	30	20	15	15
LGAA width (nm) - HD		15	10	10	6	6
LGAA width (nm) - SRAM		7	6	6	6	6
Footprint drive efficiency - lateral GAA - HP		4.41	4.50	5.47	5.00	4.75
Device effective width (nm) - HP	101.0	216.0	216.0	208.0	160.0	152.0
Device effective width (nm) - HD	101.0	126.0	96.0	128.0	88.0	80.0
PN seperation width (nm)	45	40	20	15	15	10

Acronyms used in the table (in order of appearance): LGAA—lateral gate-all-around-device (GAA), CFET (Complementary Field Effect Transistor), 3DVLSI—fine-pitch 3D logic sequential integration.

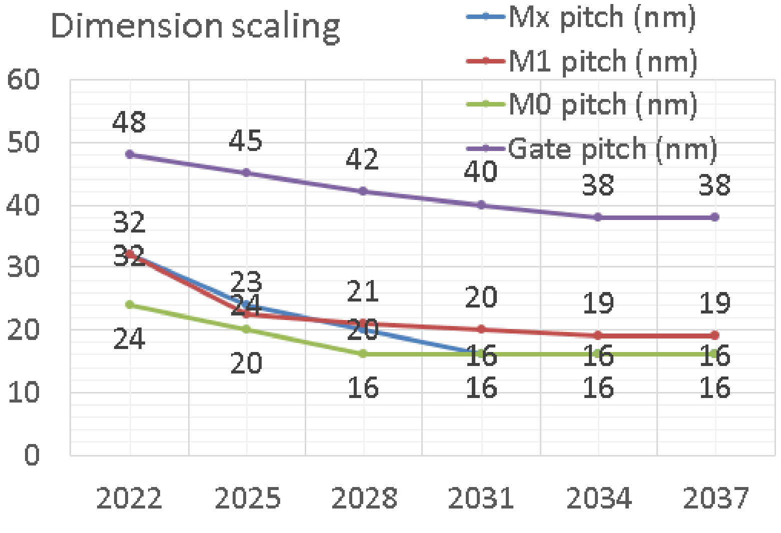


Figure MM-2 Projected scaling of key ground rules.

Ground rule scaling alone is not adequate to scale the cell height. It is necessary to bring the design scaling factor into practice [2][3]. For example, standard cell height will be further reduced by scaling the number/width of active devices in the standard cell as well as scaling the secondary rules such as tip-to-tip, extension, P-N separation, and minimum area rules. Similarly, the standard cell width can be reduced by focusing on critical design rules such as fin termination at the edge fin, etc., and enabling structures such as contact-over-active [4]-[6]. Also, the contact structure needs to be carefully selected to reduce the risk of increased current density at the junctions. It is expected that beyond 2028 P and N devices could be stacked on top of each other allowing a further reduction [7]. This trend in standard cell scaling is shown in Figure MM-3.

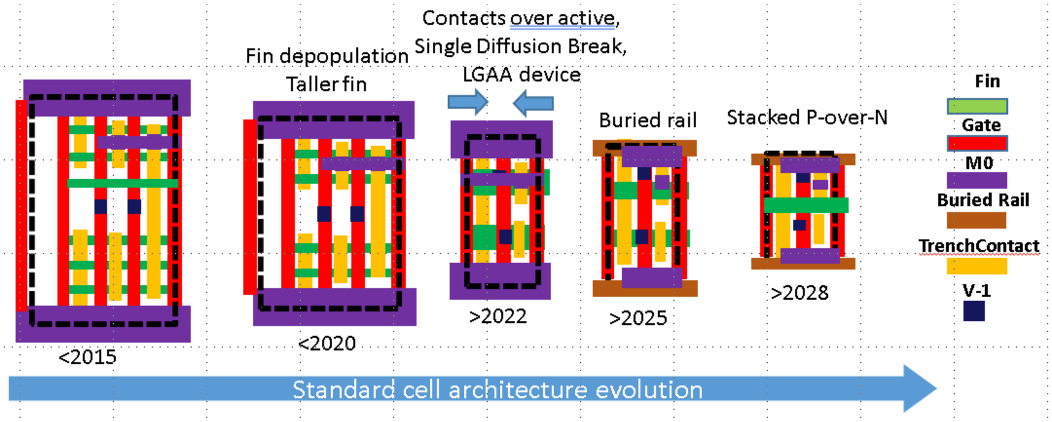


Figure MM-3 Scaling of standard cell height and width through fin depopulation and device stacking.

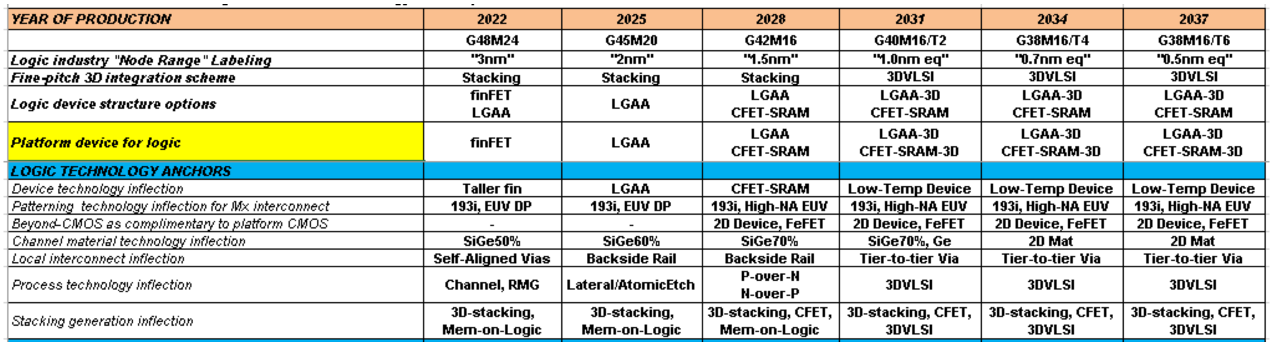
After 2031 there is no room for 2D geometry scaling, where 3D very large scale integration (VLSI) of circuits and systems using sequential/stacked integration approaches will be necessary. This is due to the fact that there is no room for contact placement as well as worsening performance as a result of gate pitch scaling and metal pitch scaling. It is projected that physical channel length would saturate around 12nm due to worsening electrostatics while gate pitch reduction reserving sufficient width (~14nm) for the device contact, providing acceptable parasitics. This drawback of pitch scaling has been compromised with dual-gate-pitch processing where relaxed pitch devices are used for high-performance cells while tight-pitch devices are used for high-density cells [9][10]. 3D VLSI expects to bring PPAC gains for the target node as well as to pave ways for heterogeneous and/or hybrid integration. The challenge of such integration in 3D is how to partition the system to come up with better utilization of devices, interconnects, and sub-systems such as memory, analog, and I/O. That is why the functional scaling and/or significant architectural changes are required after 2031. This would potentially be the time where Beyond CMOS and specialty technology devices/components would bring up the system scaling towards high system performance at unit power density and at unit cube volume.

4.2. Performance Boosters

In order to maintain the scaling at low voltages, scaling in recent years focused on additional solutions to boost the performance such as the use of introducing strain to channel; stress boosters; high-κ metal gate; lowering contact resistance, and improving electrostatics. This was all done in order to compensate the gate drive loss while supply voltage needs to be scaled down for high-performance mobile applications.

A roadmap overview of device architecture, key modules, and performance boosters is shown in Table MM-8.

Table MM-8 Device Roadmap and Technology Anchors for More Moore Scaling.



YEAR OF PRODUCTION	2022	2025	2028	2031	2034	2037
	G48M24	G45M20	G42M16	G40M16/T2	G38M16/T4	G38M16/T6
Logic industry "Node Range" Labeling	"3nm"	"2nm"	"1.5nm"	"1.0nm eq"	"0.7nm eq"	"0.5nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI
Logic device structure options	finFET LGAA	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM
Platform device for logic	finFET	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D
LOGIC TECHNOLOGY ANCHORS						
Device technology inflection	Taller fin	LGAA	CFET-SRAM	Low-Temp Device	Low-Temp Device	Low-Temp Device
Patterning technology inflection for Mx interconnect	193i, EUV DP	193i, EUV DP	193i, High-NA EUV	193i, High-NA EUV	193i, High-NA EUV	193i, High-NA EUV
Beyond-CMOS as complimentary to platform CMOS	-	-	2D Device, FeFET	2D Device, FeFET	2D Device, FeFET	2D Device, FeFET
Channel material technology inflection	SiGe50%	SiGe60%	SiGe70%	SiGe70%, Ge	2D Mat	2D Mat
Local interconnect inflection	Self-Aligned Vias	Backside Rail	Backside Rail	Tier-to-tier Via	Tier-to-tier Via	Tier-to-tier Via
Process technology inflection	Channel, RMG	Lateral/AtomicEtch	P-over-N N-over-P	3DVLSI	3DVLSI	3DVLSI
Stacking generation inflection	3D-stacking, Mem-on-Logic	3D-stacking, Mem-on-Logic	3D-stacking, CFET, Mem-on-Logic	3D-stacking, CFET, 3DVLSI	3D-stacking, CFET, 3DVLSI	3D-stacking, CFET, 3DVLSI

Mx—tight-pitch routing metal interconnect.

FinFET still remains the key device architecture that could sustain scaling until 2025 [4][6][11]. Electrostatics and fin depopulation (i.e., increasing fin height while reducing number of fins at unit footprint area) remain as the two effective solutions to improve performance. Parasitics improvement is expected to continue as a major knob for performance improvement as a result of tightening design rules. It is forecasted that the parasitics will remain as a dominant term in the performance of critical paths. For reduced supply voltage, a transition to GAA structures such as lateral nanosheets would be necessary to sustain the gate drive by improved electrostatics [12]. Sequential integration would allow stacking of devices on top of each other with the adoption of monolithic 3D (M3D) integration [13]. Scaling focus will shift from single-thread performance gain to power reduction and then evolve onto highly-parallel 3D architectures allowing low Vdd operation and more functions embedded at unit cube volume. Squeeze of active area portion will make other design rules bottleneck in design scaling such as area reserved for the power rail is expected to be buried through backside rail below the device contact layer to allocate this for additional intra-cell routing [14][15][16].

While device architectures are seeing changes, subsequent modules are expected to also evolve. These may include: channel material evolving from Si to SiGe, Ge, 2D materials; contact module evolving from silicides to novel materials providing lower Schottky barrier height (SBH), wrap-around contact integration schemes to increase the contact surface area. Below is a list of these schemes.

4.2.1. Transition to new device architectures

As mentioned earlier finFET is likely to be sustained until 2025. Beyond 2022 a transition to lateral GAA devices is expected to start and potentially evolve to cover applications such as 3D hybrid memory-on-logic applications. This situation would be due to the limits of fin-width scaling (saturating the Lgate scaling to sustain the electrostatics control) and contact width. Parasitic capacitance penalty, effective drive width (Weff), and replacement metal gate (RMG) integration pose challenges in GAA adoption [17]. Projected evolution of device architectures is shown in Figure MM-4. FinFET and GAA architectures are leading not only to fully depleted channels but also to fully inverted channels (volume inversion). It is projected that complementary FET (CFET) will be the subsequent evolution of lateral GAAs in the 3D form where P devices will be stacked over N devices [7].

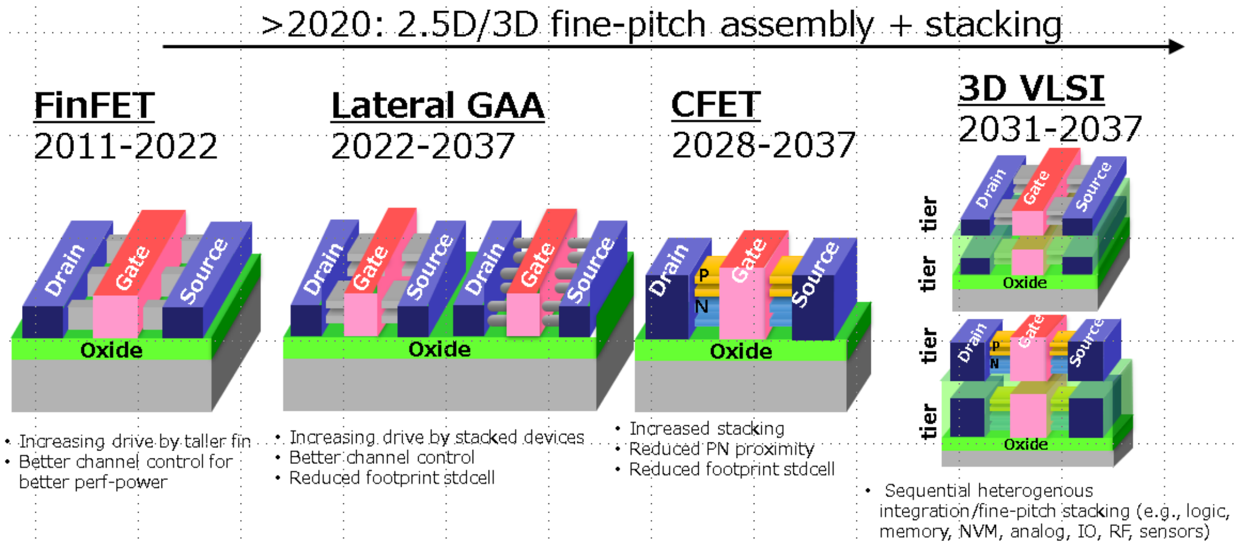

>2020: 2.5D/3D fine-pitch assembly + stacking
FinFET
2011-2022
Drain
Gate
Source
Oxide
• Increasing drive by taller fin
• Better channel control for better perf-power
Lateral GAA
2022-2037
Drain
Gate
Source
Drain
Gate
Source
Oxide
• Increasing drive by stacked devices
• Better channel control
• Reduced footprint stdcell
CFET
2028-2037
Drain
P
N
Gate
Source
Oxide
• Increased stacking
• Reduced PN proximity
• Reduced footprint stdcell
3D VLSI
2031-2037
tier
tier
tier
tier
Drain
Gate
Source
Oxide
Drain
P
N
Gate
Source
Oxide
• Sequential heterogenous integration/fine-pitch stacking (e.g., logic, memory, NVM, analog, IO, RF, sensors)

Figure MM-4 Evolution of device architectures in the IRDS More Moore roadmap

4.2.2. Starting Substrate

Bulk silicon will still remain the mainstream substrate while silicon-on-insulator (SOI) and strain-relaxation-buffer (SRB) will be used to support better isolation (e.g., RF co-integration) and defect-free integration of high-mobility channels, respectively.

4.2.3. High-mobility channels

High-mobility materials such as Ge and III-V bring promise in increasing drive current by means of an order of magnitude increase in intrinsic mobility. With the scaling in gate length, the impact of mobility on drain current becomes limited because of the velocity saturation. On the other hand, whenever gate length further scales down, the carrier transport becomes ballistic. This allows velocity of carriers, also known as “injection velocity,” scaling with the mobility increase. Having drain current mostly ballistic increases the injection velocity because of lower effective mass, therefore results in increase of the drain current. However, low effective mass for the high mobility device can actually cause high tunneling current at higher supply voltage. This may degrade the effective performance of III-V devices at short channel after work function tuning (e.g., threshold voltage increase) to lower the leakage current (Ioff) to compensate for the tunneling current. Another consideration for high mobility channel is the lower density of states. The current is proportional to the multiplication of drift velocity and carrier concentration in the channel [18]. This requires a correct selection of gate length (Lg), supply voltage (Vdd), and device architecture in order to maximize this multiplication, where the selection of those parameters will be different for the type of channel material used. This all needs to be holistically tackled [19][20].

It is likely that high-mobility channels will occur in the form of 3D stacked tiers dedicated to high-performance functions, such as high-speed IOs, high-current analog drivers, RF, photonics devices, power management, etc, which does not need to follow aggressive dimensional scaling. Improved performance and enablement of new features in the total system needs to be weighed against the cost, determined by substantial investment in new tools and fab infrastructure. On the other hand, increasing the number of vertical stacked nanosheets employing high mobility channels allow very high performance at reduced footprint [21].

4.2.4. Strain engineering

Strain engineering has been used as one of the most effective solutions in the last decade, as illustrated for the 32nm node and earlier [22]. However, the effect of those stressors may not extrapolate intuitively into newer nodes. With the scaling down of gate pitch, SiGe on the source/drain epitaxial (S/D EPI) contact and strain relaxation buffer (SRB) remain as effective boosters to scale mobility more than double on top of high-mobility channel material [23]. SiGe channel for PMOS and strained Si channel for NMOS has been successfully demonstrated on a 7nm CMOS platform using SRB [24] and on gate-all-around devices [25]. Other strain engineering techniques also contain gate stressor and ground plane stressors, which adopt the beneficiary vertical stress components for NMOS. Reducing parasitic device resistance

Controlling source/drain series resistance within tolerable limits will become increasingly difficult. Due to the increase of current density, the demand for lower resistance with smaller dimensions at the same time poses a great challenge. It is estimated that in current technologies, series resistance degrades the saturation current by 40% or more. External resistance impact on the drive current is expected to become worse with the gate pitch scaling. In addition, increasing interconnect resistance by scaling is expected to necessitate much lower resistance values for the device contact. In order to maximize the benefits of high-mobility channels in the drain current, it becomes much more important to reduce the contact resistance. Silicide contacts are failing to maintain the required reduction of contact resistance with the gate pitch scaling and decreasing channel resistance with improved drive. One promising reduction is achieved by metal-insulator-semiconductor (MIS) contacts, which utilize an ultra-thin dielectric between the metal and semiconductor interface. This reduces the Fermi-level pinning and therefore reduces the Schottky Barrier Height (SBH) [26][27]. This SBH reduction occurs from the exponential decay of the metal induced gap states (MIGS) inducing charge density accumulation in the bandgap of the dielectric.

4.2.5. Reducing parasitic device capacitance

Parasitic capacitance between gate and source/drain terminal of the device is expected to increase with technology scaling. In fact, this component is getting more important than channel-capacitance-related loading whenever the standard cell context is considered and even more elevated in the GAA structures as a result of unused space between stacked devices. There is a need to focus on low-κ spacer materials and even air spacer. Those still need to provide good reliability and etch selectivity for S/D contact formation [28][29]. Also, there are significant limits in increasing finFET or lateral GAA device AC performance by increasing the height of the device (fin/nanosheet stack). Energy per switch vs. delay relationship seems to quickly saturate and then decline with increasing device height. Scaling trend of key parasitic improvements is shown Figure MM-4.

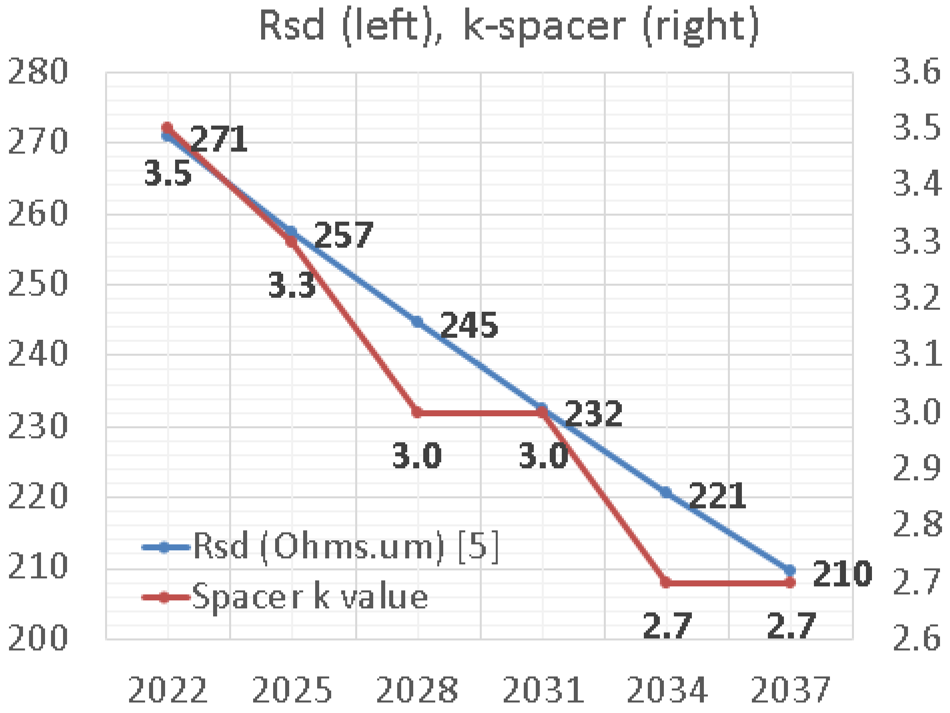


Figure MM-5 Scaling trend of device S/D access resistance (Rsd) and k-value of device spacer

Note [5]: Rsd is the total parasitic series resistance (source plus drain) per micron of MOSFET width. These values include all components such as accumulation layer, spreading resistance, sheet resistance, and contacts. It is assumed that there is 5% improvement per each node cycle (2 years or 3 years).

4.2.6. Increasing drive per footprint

FinFET and lateral GAA devices enable a higher drive at unit footprint (by enabling drive in the third dimension) if device pitch can be aggressively scaled while increasing the fin height or number of stacked GAA devices [28][30]. This will then increase drive at unit footprint but bringing a trade-off between fringing capacitance between gate and contact, and series resistance. This trend in reducing the number of fins while balancing the drive with increased fin height is defined as fin depopulation strategy, which also simulataneuously reduces the standard cell height, and therefore, the overall chip area. Complementary FET will further scale the drive per footprint by stacking P devices over N, or vice versa. This will substantially increase the amount of devices at unit footprint.

4.2.7. Improving electrostatics

FinFET provides good electrostatics integrity due to its tall narrow channel that is controlled by a gate from three-sides that allows relaxing the scaling requirements of fin thickness. Junction formation engineering, EOT scaling and density of interface traps (Dit) reduction are potential solutions to maintain the electrostatics control in the channel [31][32]. LGAA devices bring better electrostatics than finFET by providing a gate control from all sides of device channel. Since the devices are stacked on top of each other, the spacing between the devices needs to kept smaller to reduce the parasitic capacitance between the source/drain and gate while still leaving adequate space forthe gate dielectric and Vt-tuning work function metals deposition.

4.2.8. Improving device isolation

Besides the channel leakage induced by electrostatics, there are potentially other leakage sources such as sub-fin leakage or punchthrough current. This leakage current flows through the bottom part of the fin from source to drain. This gets more problematic in SiGe and Ge channels because of low effective mass of Ge. Ground plane doping, dielectric isolation, and quantum well below the channel would potentially solve this leakage problem; therefore improving the electrostatics [33].

4.2.9. Reducing process and material variations

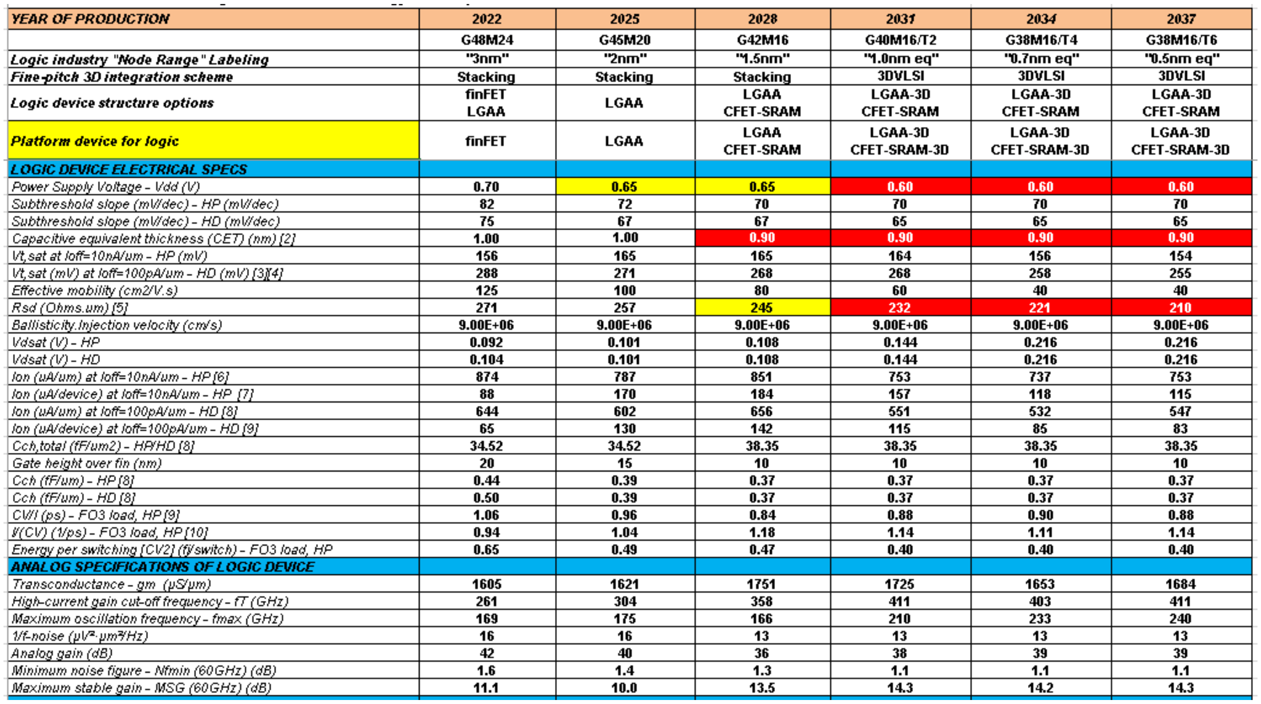
Reducing variability would further allow supply voltage (Vdd) scaling. Controlling channel length and channel thickness are important to maintain the electrostatics in the channel. This would require, for example, controlling the profile of the fin and lithography processes to reduce the CD uniformity (CDU), line width roughness (LWR), and line edge roughness (LER). Dopant-free channel and low-variability work-function metals would reduce the variations in the threshold voltage. With the introduction of high-mobility materials, gate stack passivation is needed to reduce the interface-related variations as well as maintaining the electrostatics and mobility.

4.2.10. Beyond CMOS for application-specific functions and architectures

MOSFET scaling will likely become ineffective and/or very costly for the complementary SoC functions, such as memory selector, cross-bar switch, etc. Completely new, non-CMOS types of logic devices and maybe even new circuit architectures are potential solutions (see the Beyond CMOS chapter for detailed discussions). Such solutions ideally can be integrated onto the Si-based platform to take advantage of the established processing infrastructure, as well as being able to include Si devices, such as memories, onto the same chip. Even early adoption of Beyond CMOS technology and/or computing are likely to be adopted around 2028 by ferroelectric-FET, BEOL oxide transistors, IGZO, and/or 2D materials for ultra-low power applications and also memristors for neuromorphic applications [34].

The projected roadmap for the electrical specifications of logic core device is listed in Table MM-9. This edition of More Moore roadmap includes both logic and analog specifications of More Moore platform device. Analog specifications are derived from the device targets of a logic device but this would potentially require relaxation of contacted poly pitch on the same wafer to allow longer channel lengths. There would also be considerations such as reliability and matching where performance targets need to be derated in an effort to meet those concurrent goals, e.g. increasing overdrive voltage by stacking devices.

Table MM-9 Projected Electrical Specifications of Logic Core Device



YEAR OF PRODUCTION	2022	2025	2028	2031	2034	2037
	G48M24	G45M20	G42M16	G40M16/T2	G38M16/T4	G38M16/T6
Logic industry "Node Range" Labeling	"3nm"	"2nm"	"1.5nm"	"1.0nm eq"	"0.7nm eq"	"0.5nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI
Logic device structure options	finFET LGAA	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM
Platform device for logic	finFET	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D
LOGIC DEVICE ELECTRICAL SPECS						
Power Supply Voltage - Vdd (V)	0.70	0.65	0.65	0.60	0.60	0.60
Subthreshold slope (mV/dec) - HP (mV/dec)	82	72	70	70	70	70
Subthreshold slope (mV/dec) - HD (mV/dec)	75	67	67	65	65	65
Capacitive equivalent thickness (CET) (nm) [2]	1.00	1.00	0.90	0.90	0.90	0.90
Vt,sat at Ioff=10nA/um - HP (mV)	156	165	165	164	156	154
Vt,sat (mV) at Ioff=100pA/um - HD (mV) [3][4]	288	271	268	268	258	255
Effective mobility (cm2/V.s)	125	100	80	60	40	40
Rsd (Ohms.um) [5]	271	257	245	232	221	210
Ballisticity.Injection velocity (cm/s)	9.00E+06	9.00E+06	9.00E+06	9.00E+06	9.00E+06	9.00E+06
Vdsat (V) - HP	0.092	0.101	0.108	0.144	0.216	0.216
Vdsat (V) - HD	0.104	0.101	0.108	0.144	0.216	0.216
Ion (uA/um) at Ioff=10nA/um - HP [6]	874	787	851	753	737	753
Ion (uA/device) at Ioff=10nA/um - HP [7]	88	170	184	157	118	115
Ion (uA/um) at Ioff=100pA/um - HD [8]	644	602	656	551	532	547
Ion (uA/device) at Ioff=100pA/um - HD [9]	65	130	142	115	85	83
Cch,total (fF/um2) - HP/HD [8]	34.52	34.52	38.35	38.35	38.35	38.35
Gate height over fin (nm)	20	15	10	10	10	10
Cch (fF/um) - HP [8]	0.44	0.39	0.37	0.37	0.37	0.37
Cch (fF/um) - HD [8]	0.50	0.39	0.37	0.37	0.37	0.37
CV/I (ps) - FO3 load, HP [9]	1.06	0.96	0.84	0.88	0.90	0.88
I/(CV) (1/ps) - FO3 load, HP [10]	0.94	1.04	1.18	1.14	1.11	1.14
Energy per switching [CV2] (fJ/switch) - FO3 load, HP	0.65	0.49	0.47	0.40	0.40	0.40
ANALOG SPECIFICATIONS OF LOGIC DEVICE						
Transconductance - gm (μS/μm)	1605	1621	1751	1725	1653	1684
High-current gain cut-off frequency - fT (GHz)	261	304	358	411	403	411
Maximum oscillation frequency - fmax (GHz)	169	175	166	210	233	240
1/f-noise (μV²·μm²/Hz)	16	16	13	13	13	13
Analog gain (dB)	42	40	36	38	39	39
Minimum noise figure - Nfmin (60GHz) (dB)	1.6	1.4	1.3	1.1	1.1	1.1
Maximum stable gain - MSG (60GHz) (dB)	11.1	10.0	13.5	14.3	14.2	14.3

4.3. Performance-Power-Area (PPA) Scaling

An important speed metric for the transistor is the intrinsic delay (CV/I) where C includes the gate capacitance plus the gate fringing capacitances. These fringing capacitances have been found to be larger than the intrinsic capacitance over the channel region. This requires a modeling of parasitic components in the device [35]. The ratio of total fringing capacitances to the gate capacitance over the channel is increasing with scaling.

In order to capture the behavior of a wireloaded datapath to connect the device parameters to SoC, we use a ring-oscillator-based circuit model where each stage is implemented with a D4 inverter driving a wireload with its branches driving three D4 inverters.

In this datapath model the delay of each stage is approximated by the Elmore expression given below [2]:

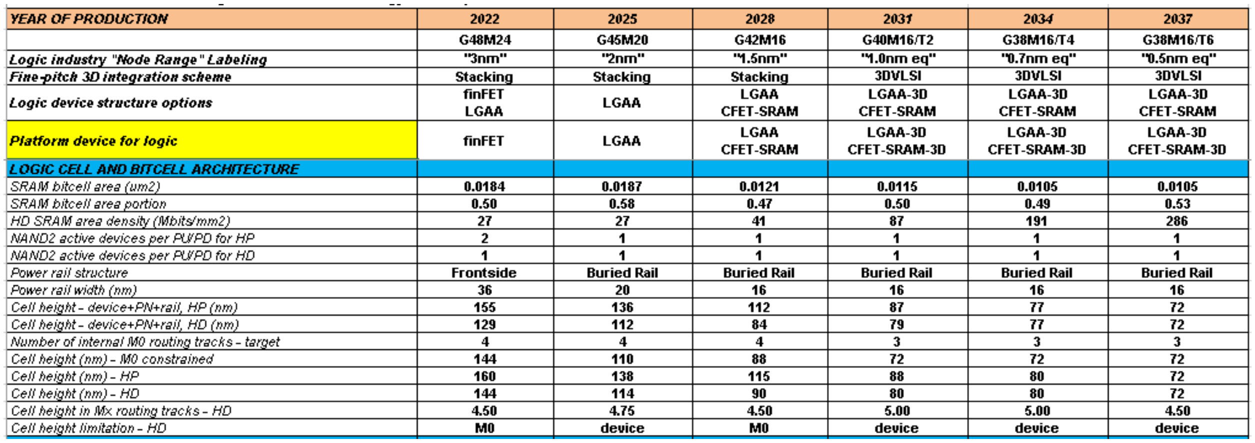
$$Tdel=0.69*Rdr*Cint + (0.69*Rdr+0.38*Rw)*Cw+0.69*(Rdr+Rw)*Cout$$

where Rdr is the resistance of driver, Cint is the capacitance seen at the output of driver, Rw is the wire resistance, Cw is the wire capacitance, and Cout is the load capacitance due to the gates connected to the load. For logic technologies beyond 10nm the dominant term is typically found to be Rw*Cout [2]. This means that increasing the driver strength does not help if there is no improvement in the parasitic resistance of interconnect and/or a reduction in the parasitic loading of standard cell.

It is also possible to extract circuit-level parameters such as delay and power per stage with the use of targeted compact models, e.g., virtual source model (VSM), which is open source distribution from MIT [36]. Details of this modeling and how interconnect is coupled with the device in the standard-cell context are explained in [2][37].

Projected scaling of PPA metrics as well as the standard cell and bitcell layout characteristics (e.g., number of active devices, Weff, etc) are shown in Table MM-10.

Table MM-10 Logic cell and bitcell architecture.



YEAR OF PRODUCTION	2022	2025	2028	2031	2034	2037
	G48M24	G45M20	G42M16	G40M16/T2	G38M16/T4	G38M16/T6
Logic industry "Node Range" Labeling	"3nm"	"2nm"	"1.5nm"	"1.0nm eq"	"0.7nm eq"	"0.5nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI
Logic device structure options	finFET LGAA	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM
Platform device for logic	finFET	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D
LOGIC CELL AND BITCELL ARCHITECTURE						
SRAM bitcell area (um2)	0.0184	0.0187	0.0121	0.0115	0.0105	0.0105
SRAM bitcell area portion	0.50	0.58	0.47	0.50	0.49	0.53
HD SRAM area density (Mbits/mm2)	27	27	41	87	191	286
NAND2 active devices per PU/PD for HP	2	1	1	1	1	1
NAND2 active devices per PU/PD for HD	1	1	1	1	1	1
Power rail structure	Frontside	Buried Rail	Buried Rail	Buried Rail	Buried Rail	Buried Rail
Power rail width (nm)	36	20	16	16	16	16
Cell height - device+PN+rail, HP (nm)	155	136	112	87	77	72
Cell height - device+PN+rail, HD (nm)	129	112	84	79	77	72
Number of internal M0 routing tracks - target	4	4	4	3	3	3
Cell height (nm) - M0 constrained	144	110	88	72	72	72
Cell height (nm) - HP	160	138	115	88	80	72
Cell height (nm) - HD	144	114	90	80	80	72
Cell height in Mx routing tracks - HD	4.50	4.75	4.50	5.00	5.00	4.50
Cell height limitation - HD	M0	device	M0	device	device	device

Performance scaling across six nodes spanning from 2022 to 2037 is projected to have a mild increase for datapaths with wireload because of the negative impact of wire resistance on performance, particularly after 2028. We also take into account the wirelength reduction as function of area scaling translating into the reduction of wire-related loading capacitance and resistance. Wirelength is expected to further reduce as a result of 3DVLSI after 2031.

Energy per switching reduction is forecasted to become limited. This is mostly achieved by fin/device depopulation, which also enables the cell height reduction bringing a scaling of wire and cell related capacitances. We also consider that DTCO constructs such as contact-over-active, single diffusion break, di-electric spacer between N and P, etc., as described in [4][17][38][39], will further reduce the standard cell width. Routed gate density is improved until 2028. After 2031 it is expected that 3D scaling by sequential/stacked integration (full-scale 3DVLSI) would further maintain the scaling of the number of functions per unit cube.

4.4. SYSTEM-ON-CHIP (SOC) PPA METRICS

Due to standard cell and bitcell density improving on a node-to-node basis, it is possible to integrate more functions in a given SoC footprint. The footprint for mobile SoC integration is assumed to increase across generations because of newly added functions exceeding the scaling reduction. Therefore, the amount of memory as well as graphical processing unit (GPU) processors and neural processing units (NPU) follow the density scaling of SRAM and standard cell, respectively, and if the trend for more parallel architectures continues. On the other hand, the number of central processing units (CPUs) per node is determined based on assumed node-to-node throughput scaling of 1.7×. In other words less improvement in the system clock frequency will mean a need for more CPUs to reach the throughput target. Thanks to advances in DTCO, lateral nanosheets, followed by device-over-device stacking (e.g. P-over-N) and 3D VLSI, SoC footprint scaling factor for the same function can still be maintained.

Integration capacity of logic technology is shown in Figure MM-6 (amount of NAND2-equivalent standard cell density as well as bitcell density).

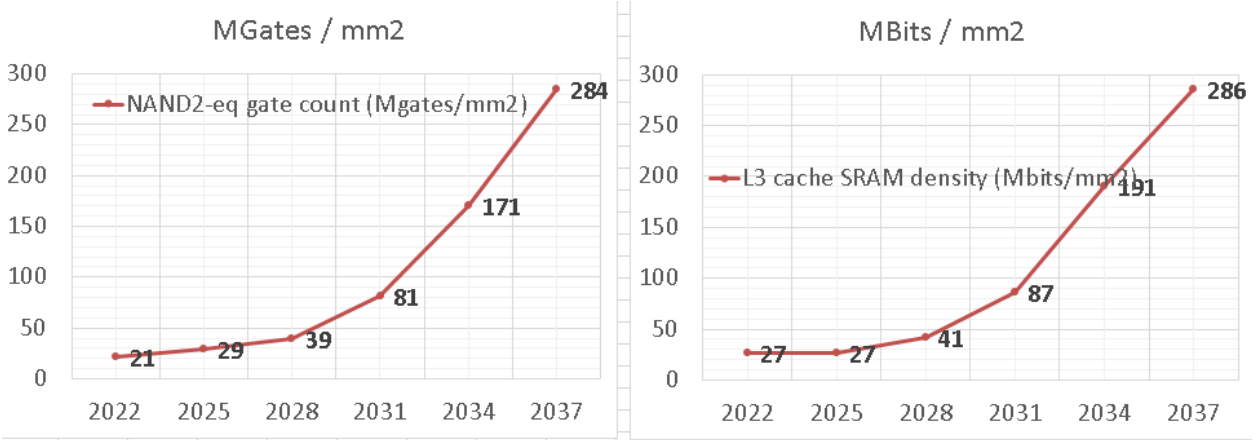
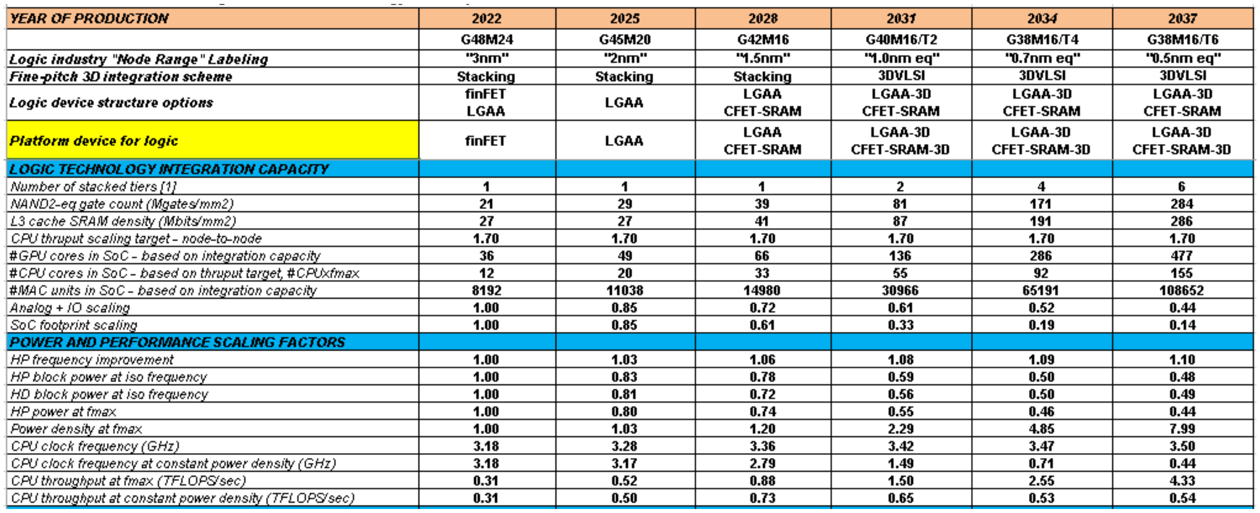


Figure MM-6 NAND2-equivalent standard cell count (left) and 111-bitcell (right) scaling.

Projected power and performance scaling of SoC is given in Table MM-11.

Table MM-11 Area, Power, and Performance Scaling of SoC



YEAR OF PRODUCTION	2022	2025	2028	2031	2034	2037
	G48M24	G45M20	G42M16	G40M16/T2	G38M16/T4	G38M16/T6
Logic industry "Node Range" Labeling	"3nm"	"2nm"	"1.5nm"	"1.0nm eq"	"0.7nm eq"	"0.5nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI
Logic device structure options	finFET LGAA	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM
Platform device for logic	finFET	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D
LOGIC TECHNOLOGY INTEGRATION CAPACITY						
Number of stacked tiers [1]	1	1	1	2	4	6
NAND2-eq gate count (Mgates/mm2)	21	29	39	81	171	284
L3 cache SRAM density (Mbits/mm2)	27	27	41	87	191	286
CPU thruput scaling target - node-to-node	1.70	1.70	1.70	1.70	1.70	1.70
#GPU cores in SoC - based on integration capacity	36	49	66	136	286	477
#CPU cores in SoC - based on thruput target, #CPUxfmax	12	20	33	55	92	155
#MAC units in SoC - based on integration capacity	8192	11038	14980	30966	65191	108652
Analog + IO scaling	1.00	0.85	0.72	0.61	0.52	0.44
SoC footprint scaling	1.00	0.85	0.61	0.33	0.19	0.14
POWER AND PERFORMANCE SCALING FACTORS						
HP frequency improvement	1.00	1.03	1.06	1.08	1.09	1.10
HP block power at iso frequency	1.00	0.83	0.78	0.59	0.50	0.48
HD block power at iso frequency	1.00	0.81	0.72	0.56	0.50	0.49
HP power at fmax	1.00	0.80	0.74	0.55	0.46	0.44
Power density at fmax	1.00	1.03	1.20	2.29	4.85	7.99
CPU clock frequency (GHz)	3.18	3.28	3.36	3.42	3.47	3.50
CPU clock frequency at constant power density (GHz)	3.18	3.17	2.79	1.49	0.71	0.44
CPU throughput at fmax (TFLOPS/sec)	0.31	0.52	0.88	1.50	2.55	4.33
CPU throughput at constant power density (TFLOPS/sec)	0.31	0.50	0.73	0.65	0.53	0.54

Clock frequency is projected to only mildly improve because of increasing parasitics and limited gate drive (Vgs-Vt) as function of scaling. After 2028 increasing of number of stacked devices, low-k materials, and 3D-VLSI help to reduce the wirelengths through the split of cells in 3D. Also, thermal (increasing power density) constraints reduce the average frequency if the chip needs to be operated at constant power density. Basically, if nothing is done for the mitigation of thermal issues, the CPU needs to be throttled more frequently to maintain the same power density. The rate of power reduction tends to flatten because of slow-down in supply voltage (Vdd) and slow-down of capacitance scaling towards the end of roadmap. This view on power-constrained CPU throughput scaling was also discussed by the ITRS System Drivers Technology Workgroup [42]. Impact of those trends in frequency, area, and energy to the system metrics such as area-efficient performance (TOPS/mm2) and energy-efficient performance (TOPS/W) is shown in Figure MM-7.

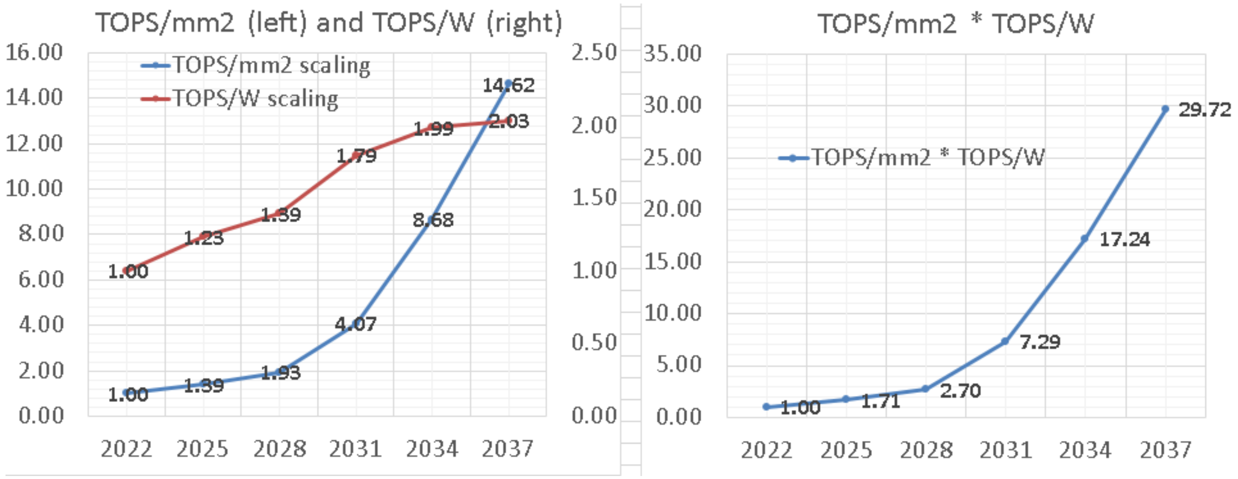


Figure MM-7 Scaling projection of computation throughput of CPU cores at the maximum clock frequency and at thermally constrained average frequency.

4.5. Interconnect Technology Requirements

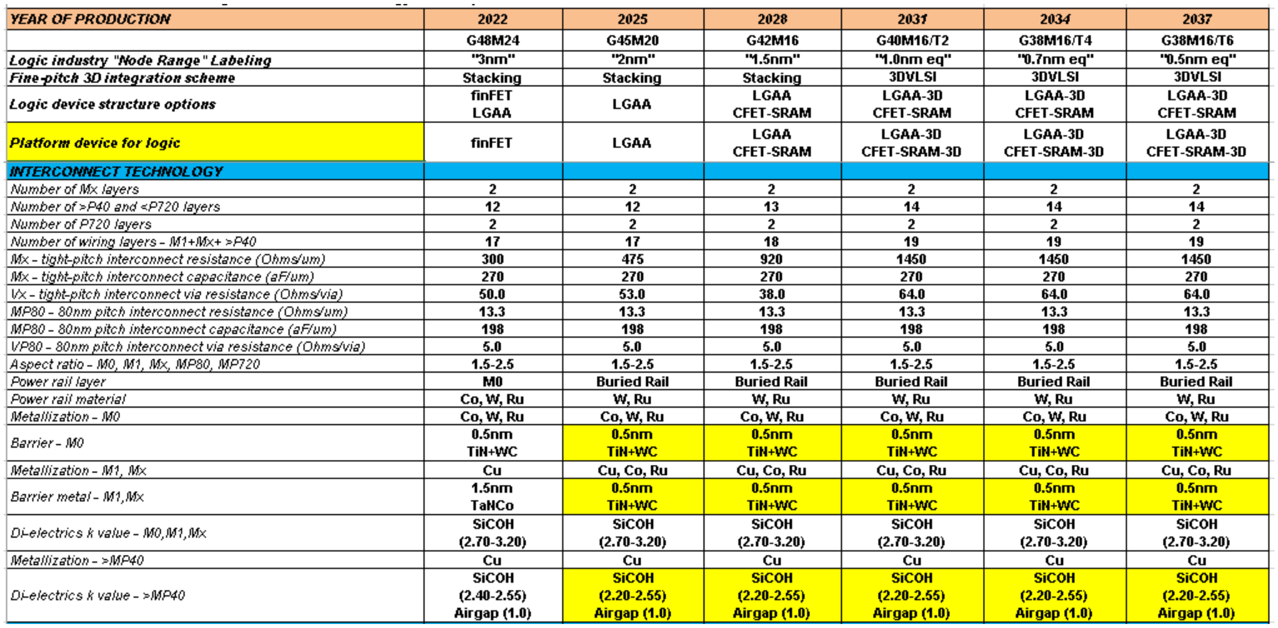
The most difficult challenge for interconnects is the introduction of new materials that meet the wire conductivity requirements, reduce dielectric permittivity, and meet reliability requirements. As for the conductivity, the impact of size effects on interconnect structures must be mitigated. Future effective κ requirements preclude the use of a trench etch stop for dual damascene structures. Dimensional control is a key challenge for present and future interconnect technology generations and the resulting difficult challenge for etch is to form precise trench and via structures in low-κ dielectric material to reduce variability in resistance-capacitance (RC). The damascene scheme used for integration requires tight control of pattern, etch, and planarization. To extract maximum performance, interconnect structures cannot tolerate variability in profiles without producing undesirable RC degradation. These dimensional control requirements place new demands on high-throughput imaging metrology for measurement of high aspect ratio structures. New metrology techniques are also needed for inline monitoring of adhesion and defects. Larger wafers and the need to limit test wafers will drive the adoption of more in-situ process control techniques. Table MM-12 highlights and differentiates the top key challenges while Table MM-13 shows the interconnect scaling roadmap.

Table MM-12 Interconnect Difficult Challenges

Critical Challenges	**Summary of Issues**
Materials—Mitigate impact of size effects in interconnect structures	Line and via sidewall roughness, intersection of porous low-κ voids with sidewall, barrier roughness, and copper surface roughness will all adversely affect electron scattering in copper lines and cause increases in resistivity.
Metrology—Three-dimensional control of interconnect features (with its associated metrology) will be required	Line edge roughness, trench depth and profile, via shape, etch bias, thinning due to cleaning, CMP effects. The multiplicity of levels, combined with new materials, reduced feature size and pattern dependent processes, use of alternative memories, optical and RF interconnect, continues to challenge.
Process—Patterning, cleaning, and filling at nano-dimensions	As features shrink, etching, cleaning, and filling high aspect ratio structures will be challenging, especially for low-κ dual damascene metal structures and DRAM at nano-dimensions.
Complexity in Integration—Integration of new processes and structures, including interconnects for emerging devices	Combinations of materials and processes used to fabricate new structures create integration complexity. The increased number of levels exacerbate thermomechanical effects. Novel/active devices may be incorporated into the interconnect.

Critical Challenges	Summary of Issues
Practical Approach for 3D—Identify solutions that address 3D interconnect structures and other packaging issues	Three-dimensional chip stacking circumvents the deficiencies of traditional interconnect scaling by providing enhanced functional diversity. Engineering manufacturable solutions that meet cost targets for this technology is a key interconnect challenge.
Growing gap between area-efficient signaling and power distribution in the local interconnects	Increasing trade-off between area-efficient wiring within the cell versus low-resistance power delivery to the standard cell, necessitating decoupling of power rails from the signal routing such as introduction of buried power rails.

Table MM-13 Interconnect Roadmap for Scaling



YEAR OF PRODUCTION	2022	2025	2028	2031	2034	2037
	G48M24	G45M20	G42M16	G40M16/T2	G38M16/T4	G38M16/T6
Logic industry "Node Range" Labeling	"3nm"	"2nm"	"1.5nm"	"1.0nm eq"	"0.7nm eq"	"0.5nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI
Logic device structure options	finFET LGAA	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM
Platform device for logic	finFET	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D
INTERCONNECT TECHNOLOGY						
Number of Mx layers	2	2	2	2	2	2
Number of >P40 and <P720 layers	12	12	13	14	14	14
Number of P720 layers	2	2	2	2	2	2
Number of wiring layers - M1+Mx+ >P40	17	17	18	19	19	19
Mx - tight-pitch interconnect resistance (Ohms/um)	300	475	920	1450	1450	1450
Mx - tight-pitch interconnect capacitance (aF/um)	270	270	270	270	270	270
Vx - tight-pitch interconnect via resistance (Ohms/via)	50.0	53.0	38.0	64.0	64.0	64.0
MP80 - 80nm pitch interconnect resistance (Ohms/um)	13.3	13.3	13.3	13.3	13.3	13.3
MP80 - 80nm pitch interconnect capacitance (aF/um)	198	198	198	198	198	198
VP80 - 80nm pitch interconnect via resistance (Ohms/via)	5.0	5.0	5.0	5.0	5.0	5.0
Aspect ratio - M0, M1, Mx, MP80, MP720	1.5-2.5	1.5-2.5	1.5-2.5	1.5-2.5	1.5-2.5	1.5-2.5
Power rail layer	M0	Buried Rail	Buried Rail	Buried Rail	Buried Rail	Buried Rail
Power rail material	Co, W, Ru	W, Ru	W, Ru	W, Ru	W, Ru	W, Ru
Metallization - M0	Co, W, Ru	Co, W, Ru	Co, W, Ru	Co, W, Ru	Co, W, Ru	Co, W, Ru
Barrier - M0	0.5nm TiN+WC	0.5nm TiN+WC	0.5nm TiN+WC	0.5nm TiN+WC	0.5nm TiN+WC	0.5nm TiN+WC
Metallization - M1, Mx	Cu	Cu, Co, Ru	Cu, Co, Ru	Cu, Co, Ru	Cu, Co, Ru	Cu, Co, Ru
Barrier metal - M1,Mx	1.5nm TaNCo	0.5nm TiN+WC	0.5nm TiN+WC	0.5nm TiN+WC	0.5nm TiN+WC	0.5nm TiN+WC
Di-electrics k value - M0,M1,Mx	SiCOH (2.70-3.20)	SiCOH (2.70-3.20)	SiCOH (2.70-3.20)	SiCOH (2.70-3.20)	SiCOH (2.70-3.20)	SiCOH (2.70-3.20)
Metallization - >MP40	Cu	Cu	Cu	Cu	Cu	Cu
Di-electrics k value - >MP40	SiCOH (2.40-2.55) Airgap (1.0)	SiCOH (2.20-2.55) Airgap (1.0)	SiCOH (2.20-2.55) Airgap (1.0)	SiCOH (2.20-2.55) Airgap (1.0)	SiCOH (2.20-2.55) Airgap (1.0)	SiCOH (2.20-2.55) Airgap (1.0)

4.5.1. Conductor

Copper (Cu) is expected to remain to be the preferred solution for the interconnect metal, at least until 2028 while non-Cu solutions (e.g. Co and Ru) are expected to be used for the local interconnect (M0). On the other hand, due to limits of electromigration, the local interconnect (middle-of-line (MOL)), M1, and Mx levels will embed non-Cu solutions such as Cobalt (Co), particularly for the via, due to its better integration window to fill the narrow trenches on top of the EM performance as well as its lower resistance compared to Cu at scaled dimensions. Although a resistivity increase due to electron scattering in Cu or higher bulk resistivity in non-Cu solutions (e.g., Co) are already apparent, a hierarchical wiring approach such as scaling of line length along with that of the width still can overcome the problem.

4.5.2. Barrier Metal

Cu wiring barrier materials must prevent Cu diffusion into the adjacent dielectric but also must form a suitable, high quality interface with Cu to limit vacancy diffusion and achieve acceptable electromigration lifetimes. Ta(N) is a well-known industry solution. Although the scaling of Ta(N) deposited by plama vapor deposition (PVD) is limited, other nitrides such as Mn(N) that can be deposited by chemical vapor deposition (CVD) or atomic layer deposition (ALD) have recently attracted attention. As for the emerging materials, self-assembled monolayers (SAMs) are researched as the candidates for future generation.

4.5.3. Inter-metal Dielectrics (IMD)

Reduction of the IMD κ value is slowing down because of problems with manufacturability. The poor mechanical strength and adhesion properties of low-k materials are obstructing their incorporation. Delamination and damage during CMP are major problems at early stages of development, but for mass production, the hardness and adhesion properties needed to sustain the stress imposed during assembly and packaging must also be achieved. Difficulties associated with the integration of highly porous ultra-low-κ (κ ≤ 2) materials become clearer, and air-gap technologies is the alternative path to lower the inter-layer capacitance. As the emerging materials, metal organic framework (MOF) and carbon organic framework (COF) could be advocated.

4.5.4. Reliability—Electromigration

An effective scaling model has been established in the earlier editions of roadmap where it assumes that the void is located at the cathode end of the interconnect wire containing a single via with a drift velocity dominated by interfacial diffusion. The model predicts that lifetime scales with w*h/j, where w is the linewidth (or the via diameter), h the interconnect thickness, and j the current density. Whereas the geometrical model predicts that the lifetime decreases by half for each new generation, it can also be affected by small process variations of the interconnect dimensions. Jmax (maximum equivalent DC current density) and JEM (DC current density at the electromigration limit) are limited by the interconnect geometry scaling. Jmax increases with scaling due to reduction in the interconnect cross-section and increase in the maximum operating frequency. The practical solutions to overcome the lifetime decrease in the narrow linewidths have been discussed actively over the past years. Recent studies show an increasingly important role of grain structure in contributing to the drift velocity and thus the EM reliability beyond the 45nm node. Process solutions with Cu alloys seed layer (e.g., Al or Mn) have shown to be an optimum approach to increase the lifetime. Other approaches are the insertion of a thin metal layer (e.g., CoWP or CVD Co) between the Cu trench and the dielectric SiCN barrier and the usage of the short length effect. The short length effect has effectively been used to extend the current carrying capability of conductor lines and has dominated the current density design rule for interconnects.

4.5.5. Reliability—Time Dependent Dielectric Breakdown

Basically, the dielectric reliability can be categorized according to the failure paths and mechanisms as shown in Figure MM-8. While a large number of factors and mechanisms have already been identified, the physical understanding is far from complete. For instance, it is necessary to correctly account for LER, voltage dependence, etc in modeling TDDB lifetime that directly impacts the estimate of Vmax (or minimum dielectric spacing).

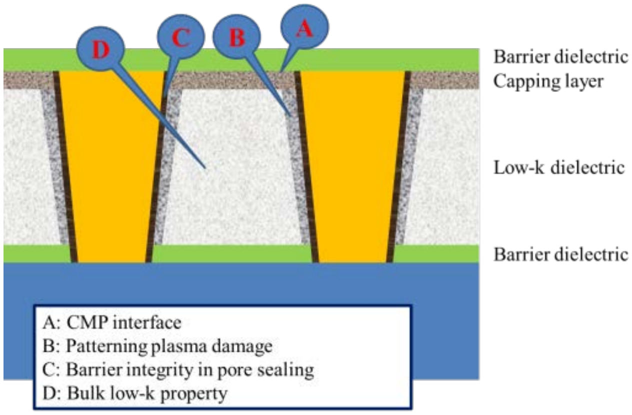


Figure MM-8 Degradation paths in low-κ damascene structure

4.6. 3D HETEROGENEOUS INTEGRATION

Every logic generation needs to add new functions in each node to keep unit price constant (to preserve profit margins). This is getting more difficult due to the following challenges:

- Fewer functions left on board/system to co-integrate
- Heterogeneous cores specialized per function—specialized performance improvement requirements needed per each dedicated core
- Off-package memories—costly to co-integrate with logic, technology not compatible with baseline CMOS (where wafer/die-level stacking might be needed)

Die cost reduction has been enabled so far by concurrent scaling of gate pitch, metal pitch, and cell height scaling. This is expected to continue until 2028 and this will accompanied by fine-pitch 3D stacking assembly such as ubump stacking and hybrid bonding [39][41]. Cell height scaling would likely be pursued by 3D devices (e.g., finFET, lateral GAA, and CFET) and DTCO constructs in cell and physical design. However, this scaling route is expected to be more challenged by diminishing electrical/system benefits and also by diminishing area-reduction/$ at SoC level. Therefore, it is necessary to pursue 3D integration routes such as device-over-device stacking, fine-pitch layer transfer, and/or monolithic 3D (or sequential integration). These pursuits will maintain system performance and power gains while potentially maintaining the cost advantages such as treating expensive non-scaled components somewhere else and using the best technology fit per tier functionality. 3D stacking routes should factor in known-good-die sorting and test methodologies to improve the stacking yield where wafer-to-wafer stacking would require a very high yielding process on each of the stacked wafers because of testing and wafer sorting challenges. Adding more heterogeneity in die stacking would require careful planning how tiers are partitioned, for instance having a smaller I/O die on top of logic die would require a lot of 2D routing in the logic die to fanin the connections from the corresponding logic block on the logic tier to the I/Os in the IO tier above. This routing would itself introduce the some area penalty in the logic tier. The overall trade-off should also include the assembly/stacking yield and additional wafer process steps such as TSV, wafer thinning, Cu pad/uBump processing.

3DVLSI can be routed either at gate or transistor levels. 3DVLSI offers the possibility to stack tiers enabling high-density contacts at the tier level (up to several million vias per mm²). The partitioning at the gate level allows IC performance gain due to wire length reduction while partitioning at the transistor level by stacking nFET over pFET (or the opposite), enabling the independent optimization of both types of transistors (customized implementation of channel material/substrate orientation/channel and raised source/drain strain, etc. [12][42]) while enabling reduced process complexity compared to a planar co-integration, for instance the stacking of III-V nFETs above SiGe pFETs [27][43]. These high mobility transistors are well suited for 3DVLSI because their process temperatures are intrinsically low. 3DVLSI, with its high contact density, can also enable applications requiring heterogeneous co-integration with high-density 3D vias, such as NEMS with CMOS for gas sensing [44][45] or highly miniaturized imagers [46]. There is significant momentum on integrating device-on-device stacking (e.g. P device over N) to decouple the channel engineering (e.g. Ge channel for PMOS) for better performance [47]. Better performance of higher tiers achieved by a freedom to select a better substrate should however factor in the potential degradation of performance due to processing them at lower temperature budget compared to the devices at the most bottom tier.

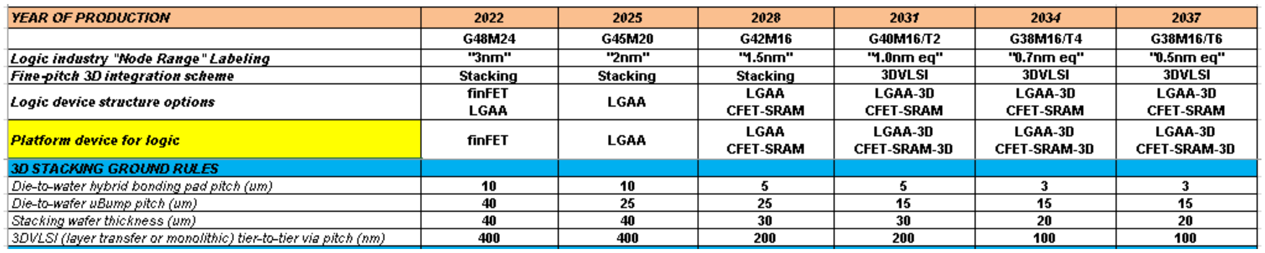
In order to address the transition from 2D to 3DVLSI, the following generations are projected in the roadmap:

- Die-to-wafer and wafer-to-wafer stacking (Table MM-15)
  - Approach: Fine-pitch dielectric/hybrid bonding and/or flip-chip assembly
  - Opportunities: Reducing bill-of-materials on the system, heterogeneous integration, high-bandwidth, and low latency memory on logic
  - Challenges: Design/architecture partitioning, power distribution network, thermal
- Device-on-device (e.g., P-over-N stacking)
  - Approach: Sequential integration
  - Opportunities: Reducing 2D footprint of standard cell and/or bit cell
  - Challenges: Minimizing interconnect overhead is key between N&P enabling low-cost
- Adding logic 3D SRAM and/or MRAM stack (embedded/stacked)
  - Approach: Sequential integration and/or wafer transfer
  - Opportunities: 2D area gain, better connection between logic and memory enabling system latency gains.
  - Challenges: Solving the thermal budget of interconnect at the lower tier if stacking approach is used, Revisiting the cache hierarchy and application requirements, power, and clock distribution
- Adding Analog and I/O
  - Approach: Sequential integration and/or wafer transfer
  - Opportunities: Giving more freedom to designer and allows integration of high-mobility channels, pushing non-scaling components to another tier, IP re-use, scalability, IO voltage enablement in advanced nodes
  - Challenges: Thermal budget, reliability requirements, power and clock distribution
- True-3D VLSI: Clustered functional stacks
  - Approach: Sequential integration and/or wafer transfer
  - Opportunities: Complementary functions other than CMOS replacement such as neuromorphic, high-bandwidth memory or pure logic applications incorporating new data-flow schemes favoring 3D connecting. Application examples include image recognition in neuromorphic fabric, wide-IO sensor

interfacing (e.g., DNA sequencing, molecular analysis), and highly parallel logic-in-memory computations.

- Challenges: Architecting the application where low energy at low frequency and highly parallel interfaces could be utilized, mapping applications to non-Von Neumann architectures.

Table MM-14 3D stacking technology ground rules.

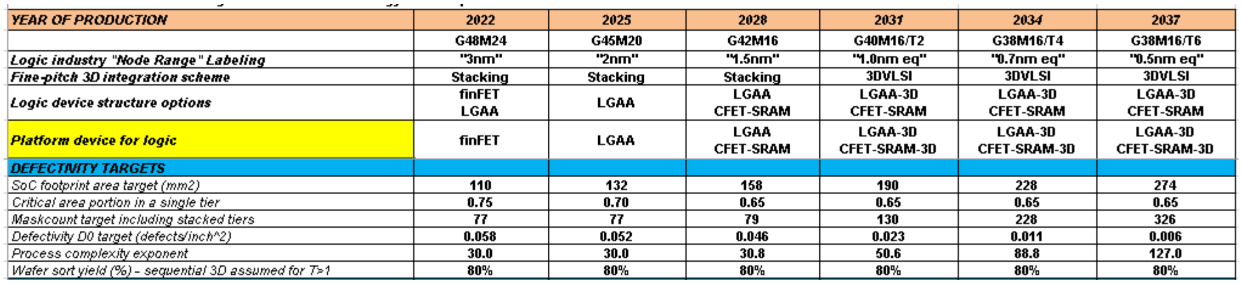


YEAR OF PRODUCTION	2022	2025	2028	2031	2034	2037
	G48M24	G45M20	G42M16	G40M16/T2	G38M16/T4	G38M16/T6
Logic industry "Node Range" Labeling	"3nm"	"2nm"	"1.5nm"	"1.0nm eq"	"0.7nm eq"	"0.5nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI
Logic device structure options	finFET LGAA	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM
Platform device for logic	finFET	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D
3D STACKING GROUND RULES						
Die-to-water hybrid bonding pad pitch (um)	10	10	5	5	3	3
Die-to-wafer uBump pitch (um)	40	25	25	15	15	15
Stacking wafer thickness (um)	40	40	30	30	20	20
3DVLSI (layer transfer or monolithic) tier-to-tier via pitch (nm)	400	400	200	200	100	100

4.7. DEFECTIVITY REQUIREMENTS

More Moore scaling necessitates an increase in the number of metallization layers, therefore an increase in the mask count if no advancement is done in the patterning technology. The expected transition from the 193i lithography to EUV will potentially save masks. However, the mask count is expected to escalate after 2031 because of increased need for the metallization and repeated masks used for the front-end-of-line (FEOL) and middle-of-line (MOL) integration for 3D integration. This will in turn increase the process complexity, therefore the defectivity (D0) requirements. The required D0 level is expected to significantly scale down (Table MM-16).

Table MM-15 Defectivity (Bose-Einstein D0) Requirements for mobile processor.



YEAR OF PRODUCTION	2022	2025	2028	2031	2034	2037
	G48M24	G45M20	G42M16	G40M16/T2	G38M16/T4	G38M16/T6
Logic industry "Node Range" Labeling	"3nm"	"2nm"	"1.5nm"	"1.0nm eq"	"0.7nm eq"	"0.5nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI
Logic device structure options	finFET LGAA	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM
Platform device for logic	finFET	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D
DEFECTIVITY TARGETS						
SoC footprint area target (mm2)	110	132	158	190	228	274
Critical area portion in a single tier	0.75	0.70	0.65	0.65	0.65	0.65
Maskcount target including stacked tiers	77	77	79	130	228	326
Defectivity D0 target (defects/inch^2)	0.058	0.052	0.046	0.023	0.011	0.006
Process complexity exponent	30.0	30.0	30.8	50.6	88.8	127.0
Wafer sort yield (%) - sequential 3D assumed for T>1	80%	80%	80%	80%	80%	80%

4.8. DEVICE RELIABILITY

Reliability is an important requirement for almost all users of integrated circuits. The challenge of realizing the required levels of reliability is increasing due to (1) scaling, (2) new materials and devices, (3) more demanding mission profiles (higher temperatures, extreme lifetimes, high currents), and (4) increasing constraints of time and money. These reliability challenges will be exacerbated by the need to introduce multiple major technology changes in a brief period of time. Interactions between changes can increase the difficulty of understanding and controlling failure modes. Furthermore, having to deal simultaneously with several major issues will tax limited reliability resources.

Reliability requirements are highly application dependent. For most customers, current overall chip reliability levels (including packaging reliability) need to be maintained over the next fifteen years in spite of the reliability risk inherent in massive technology changes. However, there are also niche markets that require reliability levels to improve. Applications that require higher reliability levels, harsher environments, and/or longer lifetimes are more difficult than the mainstream office and mobile applications. Note that a constant overall chip reliability level requires a continuous improvement in the reliability per transistor because of scaling. Meeting reliability specifications is a critical customer requirement and failure to meet reliability requirements can be catastrophic.

4.8.1. Device reliability difficult challenges

Table MM-14 indicates the top near-term reliability challenges. The first near-term reliability challenge concerns failure mechanisms associated with the MOS transistor. The failure could be caused by breakdown of the gate dielectric or degradation of device parameters, like threshold voltage and leakage current, change beyond the acceptable limits. The time to failure is decreasing with scaling. Depending on the circuit it may take more than one soft breakdown to produce an IC failure, or the circuit may function for longer time until the initial degradation spot has progressed to a "hard" failure. Threshold voltage-related failure is primarily associated with the negative bias temperature instability observed in p channel transistors in the inversion state and the analogous positive bias temperature instability in n channel transistors. Burn-in options to enhance reliability of end-products may be impacted, as it may accelerate negative bias temperature instability (NBTI) shifts.

ICs are used in a variety of different applications. There are some special applications for which reliability is especially challenging. First, there are the applications in which the environment subjects the ICs to stresses much greater than found in typical consumer or office applications. For example, automotive, military, and aerospace applications subject ICs to extremes in temperature and shock. In addition, aviation and space-based applications also have a more severe radiation environment. Furthermore, applications like base stations require IC's to be continuously on for tens of years at elevated temperatures, which makes accelerated testing of limited use. Second, there are important applications (e.g., implantable electronics, safety systems) for which the consequences of an IC failure are much greater than in mainstream IC applications. In general, scaled-down ICs are less "robust" and this makes it harder to meet the reliability requirements of these special applications. Memories, energy-harvesting and energy-storage devices exhibit their specific degradation modes, which may be rather different from those of transistors, in particular, abrupt breakdowns with no signs of preceding degradation. New computing paradigms such as neuromorphic and quantum computing, impose additional requirements on stability/drift of device characteristics.

At the heart of reliability engineering is the fact that there is a distribution of lifetimes for each failure mechanism. With low failure rate requirements, we are interested in the early-time range of the failure time distributions. There has been an increase in process variability with scaling (e.g., distribution of dopant atoms, chemical mechanical polishing (CMP) variations, and line-edge roughness). At the same time the size of a critical defect decreases with scaling. These trends will translate into an increased time spread of the failure distributions and, thus, a decreasing time to first failure. We need to develop reliability engineering software tools (e.g., screens, qualification, and reliability-aware design) that can handle the increase in variability of the device physical properties, and to implement rigorous statistical data analysis to quantify the uncertainties in reliability projections. The use of Weibull and log-normal statistics for analysis of breakdown reliability data is well established, however, the shrinking reliability margins require a more careful attention to statistical confidence bounds in order to quantify risk. This is complicated by the fact that new failure physical mechanisms, for instance, correlated defect generation, may lead to significant deviations from Weibull distribution, making error analysis non-straightforward. Statistical analysis of several reliability processes such as bias temperature instability (BTI) and hot carrier degradation is not currently standardized in practice but may be needed for accurate modeling of circuit failure rate.

The single long-term reliability difficult challenge concerns novel, disruptive changes in devices, structures, materials, and applications. For such disruptive solutions there is at this moment little, if any, reliability knowledge (as least as far as their application in ICs is concerned). This will require significant efforts to investigate, model (both a statistical model of lifetime distributions and a physical model of how lifetime depends on stress, geometries, and materials), and apply the acquired knowledge (new building-in reliability, designing-in reliability, screens, and tests). It also seems likely that there will be less-than-historic amounts of time and money to develop these new reliability capabilities. Disruptive material or devices therefore lead to disruption in reliability capabilities, and it will take considerable resources to develop those capabilities.

Table MM-16 Device Reliability Difficult Challenges

Near-Term 2022-2028	**Summary of issues**
Reliability due to material scaling, process, and structural changes, and novel applications.	• TDDB, negative BTI (NBTI), positive BTI (PBTI), hot carrier injection (HCI), random telegraphic noise (RTN) in scaled non-planar devices • Gate to contact breakdown • Increasing statistical variation of intrinsic failure mechanisms in scaled non-planar devices • 3D device structure reliability challenges • Reduced reliability margins drive need for improved understanding of reliability at circuitry operation conditions • Reliability of embedded electronics in extreme or critical environments (medical, automotive, grid...)
Long-Term 2029-2037	**Summary of issues**

Reliability of novel devices, structures, and material stacks.	• Understand and control the failure mechanisms associated with new combined set of materials and device structures • Shift to system level reliability perspective with unreliable devices • Muon induced soft error rate

4.8.2. Device reliability potential solutions

The most effective way to meet requirements is to have complete built-in-reliability and design-for-reliability solutions available at the start of the development of each new technology generation. This would enable finding the optimum reliability/performance/power choice and would enable designing a manufacturing process that can consistently have adequate reliability. Unfortunately, there are serious gaps in these capabilities today and these gaps are likely to grow even larger in the future. The penalty will be an increasing risk of reliability problems and a reduced ability to push performance, cost and time-to-market.

It is commonly thought that the ultimate nanoscale device will have a high degree of variation and high percentage of non-functional devices right from the start. This is viewed as an intrinsic nature of devices at the nanoscale. As a result, it will not be possible any longer for designer to take into account a 'worst case' design window, because this would jeopardize the performance of the circuits too much. To deal with it, a complete paradigm change in circuit and system design will therefore be needed. While we are not there yet, the increase in variability is clearly already a reliability problem that is taxing the ability of most manufacturers. This is because variability degrades the accuracy of lifetime projection, forcing a dramatic increase in the number of devices tested. The coupling between variability and reliability is squeezing out the benefit of scaling. At some point, perhaps before the end of the roadmap, the cost of ensuring each and every one of the transistors in a large integrated circuit to function within specification may become too high to be practical. As a result, the fundamental philosophy of how to achieve product reliability may need to be changed. This concept is known as resilience, the ability to cope with stress and catastrophe. One potential solution would be to integrate so-called solutions and monitors in the circuits that are sensing circuit parts that are running out of performance and then during runtime can change the biasing of the circuits. Such solutions need to be further explored and developed. Ultimately, circuits that can dynamically reconfigure itself to avoid failing and failed devices (or to change/improve functionality) will be needed.

The growing complexity of a reliability assessment due to proliferation of new materials; gate stack compositions tuned to a variety of specific applications; as well as shorter cycle for process development, may be alleviated to some degree by greater use of the physics-based microscopic reliability models, which are linked to material structure simulations and consider degradation processes on atomic level. Such models, a need for which is slowly getting wider recognition, will reduce our reliance on statistical approach, which is both expensive and time consuming, as discussed above. These models can provide additional advantage due to the fact that they can be incorporated in compact modeling tools with relative ease and require only a limited calibration prior to being applied to a specific product.

Some small changes may already be underway quietly. A first step may be simply to fine-tune the reliability requirements to trim out the excess margin, perhaps even having product-specific reliability specifications. More sophisticated approaches involve fault-tolerant design, fault-tolerant architecture, and fault-tolerant systems. Research in this direction has increased substantially. However, the gap between device reliability and system reliability is very large. There is a strong need for device reliability investigation to address the impact on circuits. Recent increase in using circuits such as SRAM and ring oscillator to look at many of the known device reliability issue is a good sign, as it addresses both the issues of circuit sensitivity as well as variability. More device reliability research is needed to address the circuit and perhaps system aspects. For example, most of the device reliability studies are based on quasi-DC measurements. There is no substantial research on the impact of degradation on devices at circuit operation speed. This gap in measurement speed makes modeling the impact of device degradation on circuit performance difficult and risky.

In the meantime, we must meet the conventional reliability requirements. That means an in-depth understanding of the physics of each failure mechanism and the development of powerful and practical reliability engineering tools. Historically, it has taken many years (typically a decade) before the start of production for a new technology generation to develop the needed capabilities (R&D is conducted on characterizing failure modes, deriving validated, predictive models and developing design for reliability and reliability TCAD tools.) The ability to qualify technologies has improved, but there still are significant gaps.

For the reliability capabilities to catch up requires a substantial increase in reliability research-development-application and cleverness in acquiring the needed capabilities in much less than the historic time scales. Work is needed on rapid characterization techniques, validated models, and design tools for each failure mechanism. The impact of new materials like alternate channel material needs particular attention. Breakthroughs may be needed to develop design for reliability tools that can provide a high-fidelity simulation of a large fraction of an IC in a reasonable time. As mentioned above, increased reliability resources also will be needed to handle the introduction of a large number of major technology changes in a brief period of time.

The needs are clearly many, but a specific one is the optimal reliability evaluation methodology, which would deliver relevant long-term degradation assessment while avoiding excessive accelerated testing that may produce misleading results. This need is driven by the decreasing process margin and increasing variability, which greatly degrades the accuracy of lifetime projection from a standard sample size. The ability to stress a large number of devices simultaneously is highly desirable, particularly for long term reliability characterization. Doing it at manageable cost is a challenge that is very difficult to meet and becoming more so as we migrate to more advanced technology nodes. A break-through in testing technology is badly needed to address this problem.

5. Technology Requirements—Memory Technologies

CMOS logic and memory together form the predominant majority of semiconductor device production. The types of memory considered in this chapter are DRAM and non-volatile memory (NVM). The emphasis is on commodity, stand-alone chips, since those chips tend to drive the memory technology. However, embedded memory chips are expected to follow the same trends as the commodity memory chips, usually with some time lag. For both DRAM and NVM, detailed technology requirements and potential solutions are considered.

5.1. DRAM

For DRAM, the main goal is to continue scaling the footprint of the 1T-1C cell, to the practical limit of 4F2. The challenges are vertical transistor structures, high-κ dielectrics to improve the capacitance density, while keeping the leakage low. In general, technical requirements for DRAMs become more difficult with scaling. In the past several years, DRAM was introduced with many new technologies (e.g., 193 nm argon fluoride (ArF) immersion high-NA lithography with double patterning technology, improved cell FET technology including fin type transistor [48]-[50], buried word line/cell FET technology [51] and so on).

Since the DRAM storage capacitor gets physically smaller with scaling, the equivalent oxide thickness (EOT) must scale down sharply to maintain adequate storage capacitance. To scale the EOT, dielectric materials having high relative dielectric constant (κ) will be needed. Therefore, metal-insulator-metal (MIM) capacitors have been adopted using high-κ ($ZrO_2/Al_2O/ZrO_2$) [52] as the capacitor of DRAMs having the ground rules between 48nm and 30nm half-pitch. And this material evolution and improvement are continued until 20 nm HP and ultra high-κ (perovskite $\kappa > 50 \sim 100$) material are released. Also, the physical thickness of the high-κ insulator should be scaled down to fit the minimum feature size. Due to that, capacitor 3-D structure will be changed from cylinder to pillar shape.

On the other hand, with the scaling of peripheral CMOS devices, a low-temperature process flow is required for process steps after formation of these devices. This is a challenge for DRAM cell processes that are typically constructed after the CMOS devices are formed, and therefore are limited to low-temperature processing. The DRAM peripheral device requirement can relax Ioff but demands more Ion of low standby power (LSTP) devices. But, in the future, high-κ metal gate will be needed for sustaining the performance [53].

The other important topic is the migration from 6F2 to 4F2 cell. As the half-pitch scaling becomes very difficult, it is impossible to sustain the cost trend. The most promising way to keep the cost trend and increasing the total bit output by generation is changing the cell size factor (a) scaling (where a = [DRAM cell size]/[DRAM half pitch]2). Currently 6F2 (a = 6) is the most common. For example, vertical cell transistor is required but still a couple of challenges are remaining. Other alternative is the use of 3D DRAM.

All in all, maintaining sufficient storage capacitance and adequate cell transistor performance are required to keep the retention time characteristic in the future. And their difficult requirements are increasing to continue the scaling of DRAM devices and to obtain the bigger product size (i.e., >16 Gb). In addition to that, if efficiency of cost scaling becomes poor in comparison with introducing the new technology, DRAM scaling will be stopped, and 3D cell stacking structure will be adopted, or a new DRAM concept will be adopted. 3D cell stacking and new concept DRAM are discussed but there is no clear path for further scaling beyond the 2D DRAM.

5.2. NVM—Flash

There are several intersecting memory technologies that share one common trait—non-volatility. The requirements and challenges differ according to the applications, ranging from RFIDs that only require Kb of storage to high-density storage of hundreds of Gb in a chip. Nonvolatile memory may be divided into two large categories—Flash memories (NAND Flash and NOR Flash), and non-charge-based-storage memories. Nonvolatile memories are essentially ubiquitous, and a lot of applications use embedded memories that typically do not require leading edge technology nodes. The More Moore nonvolatile memory tables only track memory challenges and potential solutions for leading edge standalone parts.

Flash memories are based on simple one transistor (1T) cells, where a transistor serves both as the access (or cell selection) device and the storage node. At this time Flash memory serves more than 99% of applications.

When the number of stored electrons reaches statistical limits, even if devices can be further scaled and smaller cells achieved, the threshold voltage distribution of all devices in the memory array becomes uncontrollable and logic states unpredictable. Thus memory density cannot be increased indefinitely by continued scaling of charge-based devices. However, effective density increase may continue by stacking memory layers vertically.

The economy of stacking by completing one device layer then another and so forth is questionable. As depicted in Figure MM-9 [54], the cost per bit starts to rise after stacking several layers of devices. Furthermore, the decrease in array efficiency due to increased interconnection and yield loss from complex processing may further reduce the cost-per-bit benefit of this type of 3D stacking. In 2007, a ‘punch and plug’ approach was proposed to fabricate the bit line string vertically to simplify the processing steps dramatically [54]. This approach makes 3D stacked devices in a few steps and not through repetitive processing, thus promised a new low-cost scaling path to NAND flash. Figure MM-9 illustrates one such approach. Originally coined bit-cost-scalable, or BiCS, this architecture turns the NAND string by 90 degrees from a horizontal position to vertical. The word line (WL) remains in the horizontal planes. As depicted in Figure MM-9, this type of 3D approach is much more economical than the stacking of complete devices, and the cost benefit does not saturate up to quite high number of layers.

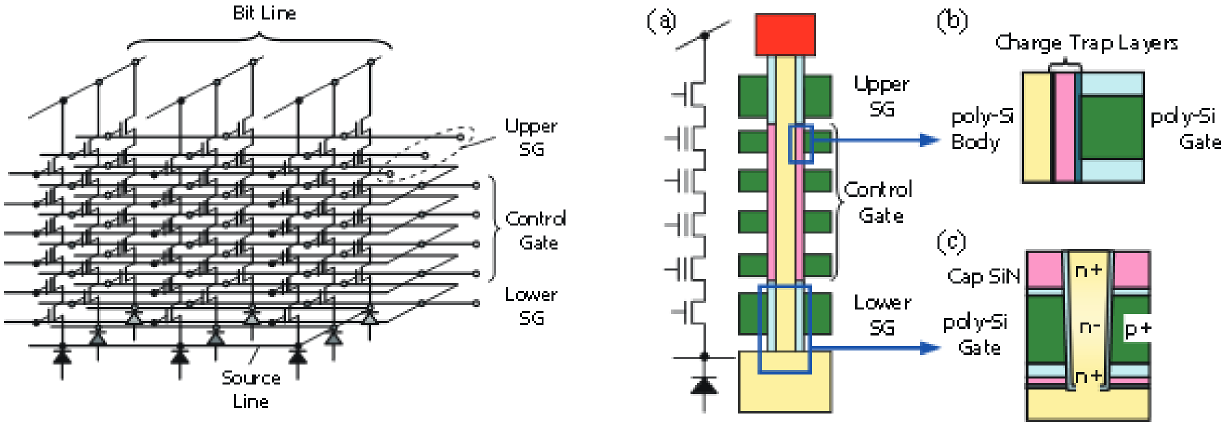


Figure MM-9 (left) A 3D NAND array based on a vertical channel architecture. (right) BiCS (bit cost scalable) – a 3D NAND structure using a punch and plug process [54].

A number of architectures based on the BiCS concept have been proposed since 2007 and several, including some that use floating gate instead of charge trapping layer for storage, have gone into volume production in the last 2−3 years. In general, all 3D NAND approaches have adopted a strategy of using much larger areal footprints than the conventional 2D NAND. The x- and y- dimensions (equivalent to cell size in 2D) of 3D NAND are in the range of 100nm and higher compared to ~15nm for the smallest 2D NAND. The much larger “cell size” is made up by stacking a large number of memory layers to achieve competitive packing density.

The economics of 3D NAND is further confounded by its complex and unique manufacturing needs. Although the larger cell size seems to relax the requirement for fine line lithography, to achieve high data rate it is desirable to use large page size and this in turn translates to fine pitched bit lines and metal lines. Therefore, even though the cell size is large metal lines still require ~20nm half-pitch that is only achievable by 193i lithography with double patterning. Etching of deep holes is difficult and slow, and the etching throughput is generally very low. Depositing many layers of dielectric and/or polysilicon, as well as metrology for multilayer films and deep holes all challenge unfamiliar territories. These all translate to large investment in new equipment and floor space and new challenges for wafer flow and yield.

The ultimate unknown is how many layers can be stacked. There seems no hard physics limit on the stacking of layers. Beyond certain aspect ratio (100:1 perhaps?) the etch-stop phenomenon, when ions in the reactive ion etching process are bent by electrostatic charge on the sidewall and cannot travel further down, may limit how many layers can be etched in one operation. However, this may be bypassed by stacking fewer layers, etching, and stacking more layers (at higher cost). Stacking many layers may produce high stress that bends the wafer and although this needs to be carefully engineered it does not seem to be an unsolvable physics limit. Even at 200 layers (at ~50nm for each layer) the total stack height is about 10µm, which is still in the same range as 10−15 metal layers for logic IC's. This kind of layer thickness does not significantly affect bare die thickness (thinnest is about 40µm so far) yet. However, at 1000 layers the total layer thickness may cause thick dies that do not conform to the form factor for stacking multiple dies (e.g., 16 or 32) in a thin package. At this time, 176 layers are in volume production and there is optimism that 300+ layers are achievable and even 500 and 800 layers are possible. In addition to processing challenges stacking more layers also increases the area overhead required to making contacts to more word lines. This area overhead, plus the added processing complexity, will eventually diminish the cost benefit by adding more layers.

Renewed shrinking of the areal x-y footprint may eventually start when stacking more layers proves to be too difficult. However, such a trend is not guaranteed. If the hole aspect ratio is the limitation, shrinking the footprint would not reduce the ratio and would thus not be helpful. Furthermore, the larger cell size seems to at least partially contribute to the better performance of 3D NAND (speed and cycling reliability) compared to tight-pitch 2D NAND. Whether x-y scaling can still deliver such performance is not clear. Therefore, the roadmap projections for future generations stay the same as the current node in 2022. On the other hand, increasing the number of storage bits per memory cell, although technically challenging, seems to make progress. This is partly to take advantage that 3D NAND devices are intrinsically larger thus with more stored electrons and easier to make into more logic levels. At this time 4-bits/cell devices (QLC) are in volumn production, and there is optimism that 5-bits/cell and more may become viable in the near future. Higher number of storage bits in a cell requires some trade off in performance since it takes longer to program and read, and relliability suffers when squeezing logic levels close together. Yet for many read-intensive applications such trade-offs are acceptable for lower cost.

5.3. NVM—Emerging

Since 2D NAND Flash scaling is limited by statistical fluctuation due to too few stored charges, several non-conventional non-volatile memories that are not based on charge storage (ferroelectric or FeRAM, magnetic or MRAM, phase-change or PCRAM, and resistive or ReRAM) are being developed and form the category often called "emerging" memories. Even though 2D NAND is being replaced by 3D NAND (that is no longer subject to the drawback of too few electrons) some characteristics of non-charge based emerging memories, such as low voltage operation, or random access, are attractive for various applications and thus continue to be developed. These emerging memories usually have a two-terminal structure (e.g., resistor or capacitor) thus are difficult to also serve as the cell-selection device. The memory cell generally combines a separate access device in the form of 1T-1C, 1T-1R, or 1D-1R.

5.3.1. FeRAM

FeRAM devices achieve non-volatility by switching and sensing the polarization state of a ferroelectric capacitor. To read the memory state the hysteresis loop of the ferroelectric capacitor must be traced and the stored datum is destroyed and must be written back after reading (destructive read, like DRAM). Because of this 'destructive read' it is a challenge to find ferroelectric and electrode materials that provide both adequate change in polarization and the necessary stability over extended operating cycles. Many ferroelectric materials are foreign to the normal complement of CMOS fabrication materials, and can be degraded by conventional CMOS processing conditions. FeRAM is fast, low power, and low voltage and thus is suitable for RFID, smart card, ID card, and other embedded applications. Processing difficulty limits its wider adoption. Recently, HfO_2 based ferroelectric FET, for which the ferroelectricity serves to change the Vt of the FET and thus can form a 1T cell similar to Flash memory, has been proposed. If developed to maturity this new memory may serve as a low power and very fast Flash-like memory.

5.3.2. MRAM

Magnetic RAM (MRAM) devices employ a magnetic tunnel junction (MTJ) as the memory element. An MTJ cell consists of two ferromagnetic materials separated by a thin insulating layer that acts as a tunnel barrier. When the magnetic moment of one layer is switched to align with the other layer (or to oppose the direction of the other layer) the effective resistance to current flow through the MTJ changes. The magnitude of the tunneling current can be read to indicate whether a ONE or a ZERO is stored. Field switching MRAM probably is the closest to an ideal "universal memory" since it is non-volatile and fast and can be cycled indefinitely. Thus, it may be used as NVM as well as SRAM and DRAM. However, producing a magnetic field in an IC circuit is both difficult and inefficient. Nevertheless, field switching MTJ MRAM has successfully been made into products. The required magnetic field for switching, however, increases when the storage element scales while electromigration limits the current density that can be used to produce higher H field. Therefore, it is expected that field switch MTJ MRAM is unlikely to scale beyond 65nm node. Recent advances in "spin-transfer torque (STT)" approach where a spin-polarized current transfers its angular momentum to the free magnetic layer and thus reverses its polarity without resorting to an external magnetic field offer a new potential solution. During the spin transfer process, substantial current passes through the MTJ tunnel layer and this stress may reduce the writing endurance. Upon further scaling the stability of the storage element is subject to thermal noise, thus perpendicular magnetization materials are projected to be needed at 32nm and below. Perpendicular magnetization has been recently demonstrated.

With rapid progress of NAND Flash and the recent introduction of 3D NAND that promises to continue the equivalent scaling, the hope of STT-MRAM to replace NAND seems remote. However, its SRAM-like performance and much smaller footprint than the conventional 6T-SRAM have gained much interest in that application, especially in mobile devices that do not require high cycling endurance, as in computation. Therefore, STT-MRAM is now mostly considered not as a standalone memory but an embedded memory [55][56], and is not tracked in the standalone NVM table. STT-MRAM would be a potential solution not only for embedded SRAM replacement but also for embedded Flash (NOR) replacement. This may be particularly interesting for IoT applications since low power is the most important. On the other hand, for other embedded systems applications using higher memory density, NOR Flash is expected to continue to dominate since it is still substantially more cost effective. Furthermore, Flash memory is well established for being able to endure the PCB board soldering process (at ~ 250°C) without losing its preloaded code, which many emerging memories have not been able to demonstrate yet.

5.3.3. PCRAM and Crosspoint Memory

PCRAM devices use the resistivity difference between the amorphous and the crystalline states of chalcogenide glass (the most commonly used compound is $Ge_2Sb_2Te_5$, or GST) to store the logic levels. The device consists of a top electrode, the chalcogenide phase change layer, and a bottom electrode. The leakage path is cut off by an access transistor (or diode) in series with the phase change element. The phase change write operation consists of: (1) RESET, for which the chalcogenide glass is momentarily melted by a short electric pulse and then quickly quenched into amorphous solid with high resistivity, and (2) SET, in which a lower amplitude but longer pulse (usually >100ns) anneals the amorphous phase into low resistance crystalline state. The 1T-1R (or 1D-1R) cell is larger or smaller than NOR Flash, depending on whether MOSFET or BJT (or diode) is used. The device may be programmed to any final state without erasing the previous state, thus providing substantially faster programming throughput. The simple resistor structure and the low voltage operation also make PCRAM attractive for embedded NVM applications. The major challenges for PCRAM are the high current (fraction of mA) required to reset the phase change element, and the relatively long set time and high temperature tolerance to retain the preloaded code during solder reflow (at ~250°C). Thermal disturb is a potential challenge for the scalability of PCRAM. However, thermal disturb effect is non-cumulative (unlike Flash memory in which the program and read disturbs that cause charge injection are cumulative) and the higher temperature RESET pulse is short (10ns). Interaction of phase change material with electrodes may pose long-term reliability issues and limit the cycling endurance and is a major challenge for DRAM-like applications. Like DRAM, PCRAM is a true random access, bit alterable memory.

The scalability of PCRAM device to < 5nm has been demonstrated using carbon nanotubes as electrodes [57], and the reset current followed the extrapolation line from larger devices. In at least one case, cycling endurance of 1E11 was demonstrated [58]. Phase change memory has been used in feature phones to replace NOR Flash since 2011, and has been in volume production at ~45nm node since 2012, but no new product has been introduced since then. PCM memories have been also targeted in the last years as potential candidates for eFlash replacement for embedded applications [59][60]. In these works alloying of phase change materials of different classes allowed to obtain memory compliant to soldering reflow; however, such high temperature stability has come at the expense of slower write speed.

Recently, a 3D cross point memory (3D XP) has been reported [61]. Details are still lacking but it is speculated that the threshold switching ovonic threshold switching (OTS) property of chalcogenide-based phase change material constitutes the core of the selector device responsible for the cross point cell, which was first reported in 2009 [62]. This is the first commercial realization of the widely published storage class memory (SCM) [63][64]. Computer systems badly needimproved I/O throughput and reduce power and cost, and it is a promising candidate to change the entire memory hierarchy not only for high-end computation but for mobile systems as well. In addition, since the memory including the selector device is completely fabricated in the BEOL process it is relatively inexpensive to stack multiple layers to reduce bit cost.

3D cross point memory (3D XP) consists of a selector element made of ovonic threshold switching (OTS) (or an equivalent device) in series with a storage element. The selector device has a high ON/OFF ratio and is at OFF state at all times except when briefly turned on during writing or reading. The storage element is programmed to various logic states. Since the selector is always off, with high resistance the memory array has no leakage issue even if all storage elements are at low resistance state. During write or read operation the selector is temporarily turned on (by applying a voltage higher than its threshold voltage) and the OTS characteristic suddenly reduces its resistance to a very low vaue, allowing reading (or programming) current to be dominated by the resistance of the storage element. The storage element may be a phase-change material and in that case the memory cell is a phase-change RAM (PCRAM) switched by OTS. The storage element may also be a resistive memory material. Although bipolar operation makes the circuitry and operation more complicated, the array structure is very similar to that using PCRAM.

3D XP memory has entered commercial application for several years now, providing SCM functiion for servers and at some time also for PCs. Up to this date, however, it is not widely adopted since existing products use a proprietary interface. Yet, its unique advantage of random access and low cost making it a good choice for augmenting DRAM function in a memory system thus its potential for wider adoption still exists. Its eventual success will depend on whether it can continue to hold a cost advantage over DRAM.

Note that SLC 3D NAND can also serve SCM function with the proper interface, at much lower cost. But its considerably slower latency makes it more suitable for storage SCM, and 3D XP is more suitable for memory-type SCM (i.e., more DRAM-like).

5.3.4. Resistive Memory (ReRAM)

A large category of two-terminal devices, in which memory state is determined by resistivity of a metal-insulator-metal (MIM) structure, are being studied for memory applications. Many of these resistive memories are still in research stage and are discussed in more detail in the Beyond CMOS roadmap chapter. Because of their promise to scale below 10nm, and operate at extremely high frequencies (< ns) with low power consumption, the focused R&D efforts in many industrial labs in the last decade make this technology widely considered a potential successor to NAND (including 3D NAND). Being a two-terminal device, high-density ReRAM development has been limited by the lack of a good selector device. Recent advances in 3D XP memory, however, seem to have solved this bottleneck and ReRAM could make rapid progress if other technical issues such as erratic bits are solved. In addition to 3D XP array (similar to PCRAM-based 3D XP memory) high-density ReRAM products may be fabricated using a 2D array and small word-line (WL), and small bit-line (BL) half pitch. Furthermore, if eventually the OTS type of selector device is adopted it seems feasible to fabricate BiCS type 3D ReRAM using a transistor in the bottom and OTS selector for each ReRAM device in the 3D array, as depicted in Figure MM-10 [65]. Although no high-density ReRAM product has been introduced yet since the bottleneck of bipolar selector devices seems solved by the introduction of 3D XP memory, progress in ReRAM may be reasonably expected. Recently, however, the passion for developing high-density ReRAM seems to have dissipated. This may be due to two reasons. (1) the success of 3D NAND Flash has increased the entrance barrier and (2) difficulty in meeting the reliability requirements for large arrays. (Note that several announcements were made for successful development of ReRAM for smaller, Mb-size, arrays for embedded applications.)

In the last several years these above issues seem to doom large scale application of high-density ReRAM. The original argument that ReRAM, because it consists of thousands of atoms, is free from statistical fluctuation seems questionable now. It seems the filament that operates the ReRAM is made up by only a few atoms (ions). There seems evidence that even relatively large ReRAM device is subject to statistical fluctuation. Therefore, we stop tracking ReRAM for high-density application in this roadmap release.

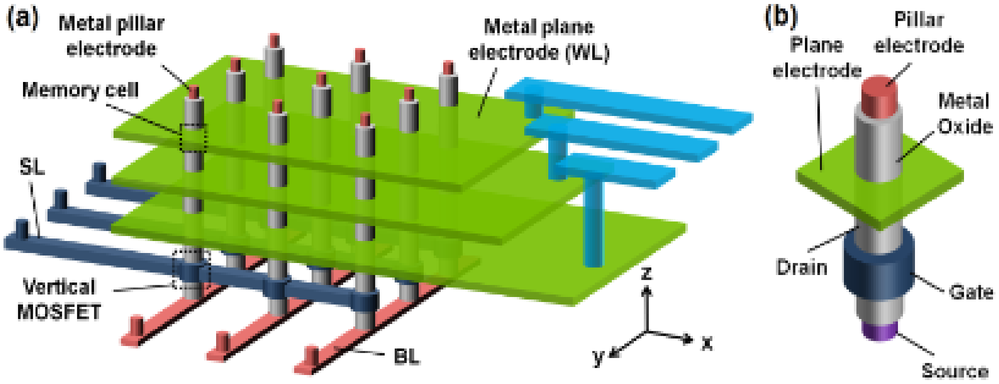


*Figure MM-*10 *Schematic view of (a) 3D cross-point architecture using a vertical RRAM cell and (b) a vertical MOSFET transistor as the bit-line selector to enable the random access capability of individual cells in the array [65].*

6. Potential Solutions

Below are the potential solutions to address the scaling challenges that were addressed in section 3 towards the targets described in section 1.2. Near-term (2022-2028) potential solutions are listed in Table MM-16 while long-term (2029-2037) potential solutions are listed in Table MM-17.

Table MM-17 Potential Solutions—Near-term

Near-Term Potential Solutions: 2022-2028	Description
Performance	• Increasing fin height to match performance • Reduce interface contact resistance through new materials and wrap-around contact • Introduce low-κ device spacer • Reduce interconnect resistance through barrier and liner scaling
Power	• Introduce GAA architectures • Reduce device parasitics
Area and Cost	• Adoption of EUV for single and double patterning • DTCO enhancement • Introduction of high-density emerging memory as cache applications

Table MM-18 Potential Solutions—Long-term

Long-Term Potential Solutions: 2029-2037	Description
Performance	• Reduce wirelength and drive-per-footprint through 3D stacking • Employ vertically stacked memory-logic sub-systems within highly parallel new computational schemes and heterogonous stacking
Power	• Increase parallelism by the introduction of large bandwidth access to the memory • Fine-grain power gating
Area and Cost	• High regularity • 3D integration/stacking with each tier adopted minimal cross-tier interconnect and integration overhead

These potential solutions are mostly targeting improvement of the PPAC value of logic technologies. It should be noted that emergence of application drivers such as 6G and extreme reality (e.g., AR/VR) brings new potential solutions for the analog/RF/sensor co-integration enablement with the use of those technology platforms. Examples include co-integration of III-V technologies with Si logic through layer transfer and/or selective growth for the enablement of products in small form factor. Si technologies, developed on low-loss SOI substrates, are expected to push the envelope of mm-wave communications where high transition frequency (Ft) and low insertion loss will be traded with a relatively lower output power in comparison to non-Si counterparts.

Si photonics is gaining momentum in short-to-medium distance connectivity applications such as chip-to-chip communications in data server racks and back-haul network of radio access cells; and also merging of energy-efficient computing with the optics in augmented reality products. Those solutions require highly integrated interposer incorporating optical modulators, laser source, photo diodes, photonic waveguides, wave-division-multiplexors, and assembly interfaces coupling fiber to the waveguide. The requirements, challenges, and potential solutions are described in the Outside System Connectivity roadmap report.

Another growing solution is the trend of miniaturizing personalized healthcare with the co-integration of heterogeneous technologies. Those products are expected to co-integrate sensors, battery, high-endurance/high-speed non-volatile memory, RF connectivity components, and ultra-low-power processing augmented with machine learning capability in the same package. More Moore technologies are helping in this context to reduce the power consumption of those devices as well as bringing new memories (e.g., MRAM, FeRAM) required for these applications.

7. CROSS TEAMS

Through cross-functional team interaction with other IFTs, the More Moore team incorporated valuable inputs in our roadmap both in terms of requirements as well as technology capability limits:

- Systems and Architectures (SA) IFT—computational datapath/fabric such as number of CPU, GPU, and NPU cores per a given footprint as well as latency/bandwidth for data access
- Application Benchmarking (AB) IFT—performance and energy scaling targets, chip-level power (active, static, sleep), thermal envelope
- Lithography IFT—Pitch limits of 193i and EUV lithography, CDU/LER capability, timeline of EUV in HVM adoption
- Yield IFT—unit-step related defect impact on material quality, infrastructural constraints such as CD and defect density on filtration and detection, and impact of defects on reliability.
- Metrology IFT—Extendibility of metrology of 3D devices such as lateral-GAA
- Outside System Connectivity (OSC) IFT—I/O and integration requirements for 5G and high-speed memory for data server
- Packaging Integration IFT—form factor and hetero-technology needs for mobile, 5G, and automotive
- Beyond CMOS (BC) IFT—3D memories such as RRAM and PCM, memristor for neuromorphic applications

8. CONCLUSIONS AND RECOMMENDATIONS

In this chapter we proposed a roadmap that could sustain More Moore scaling for concurrent enablement of performance, power, and area/cost. We identified the following inflection points:

- GAA is expected to become a mainstream device in 2025 with early introduction in 2022 and requires a significant attention on the capacitance reduction to maintain performance scaling target.
- Slow-down in pitch scaling tackled with design technology co-optimization enables the SoC area reduction where this might require process-related dimension control is necessary besides lithography.
- We identified that 3D integration is needed beyond 2028. Thermal is becoming a significant challenge in 3D adoption and needs to revisit the architecture get back the performance scaling through parallelization.
- We identified that significant reduction of defectivity level as well as careful split of technology and architecture across tiers are required to maximize the adoptability of 3D.
- Stacked SRAM is likely to become a mainstream alternative to fully integrated SRAM and/or eDRAM cache applications, probably around 2025.

9. References

[1] J.-A. Carballo, et al., “ITRS 2.0: towards a re-framing of the semiconductor technology roadmap”, Proc. ICCD, October 2014.

[2] W.-T. J. Chan, A. Kahng, S. Nath, and I. Yamamoto, “The ITRS MPU and SoC system drivers: calibration and implications for design-based equivalent scaling in the roadmap,” Proc. IEEE Int. Computer Design (ICCD), pp. 153-160, October 2014.

[3] M. Badaroglu and J. Xu, “Interconnect-aware device targeting from PPA perspective”, ICCAD, November 2016.

[4] C. Auth et al., “A 10nm high performance and low-power CMOS technology featuring 3rd-generation finFET transistors, self-aligned quad patterning, contact overactive gate and Cobalt local interconnects,” IEDM, Session 2.9, December 2017.

[5] X. Wang, et al., “Design-technology co-optimization of standard cell libraries on Intel 10nm process”, IEDM, Session 28.2, December 2018.

[6] G. Yeap, et al., “5nm CMOS production technology platform featuring full-fledged EUV, and high mobility channel FinFETs with densest 0.021 um2 SRAM cells for mobile SoC and high-performance computing applications,” IEDM, Section 36.7, December 2019.

[7] C.-Y. Huang, et al., “3-D self-aligned stacked NMOS-on-PMOS nanoribbon transistors for continued Moore’s Law scaling”, IEDM, pp. 20.6.1-20.6.4, December 2020.

[8] L. Liebmann, et al., “CFET design options, challenges, and opportunities for 3D integration”, IEDM, pp. 3.1.1-3.1.4, December 2021.

[9] J. Jeong, et al., “Performance-power management aware state-of-the-art 5nm FinFET design (5LPE) with dual CPP from mobile to HPC application”, IEDM, pp. 20.1.1-20.1.4, December 2020.

[10] Y. Yasuda-Masuoka, et al., “High performance 4nm FinFET platform (4LPE) with novel advanced transistor level DTCO for dual-CPP/HP-HD standard cells”, IEDM, pp. 13.3.1-13.3.4, December 2021.

[11] S.-W. Wu, et al., “A 7nm CMOS platform technology featuring 4th generation finFET transistors with a 0.027um2 high density 6-T SRAM cell for mobile SoC applications”, IEDM, Session 2.6, December 2016.

[12] G. Bae, et al., “3nm GAA technology featuring multi-bridge-channel FET for low power and high-performance applications”, IEDM, Session 28.7, December 2018.

[13] P. Batude, et al., “Advances in 3D CMOS sequential integration”, IEDM, Section 14.1, p. 1-4, December 2009.

[14] A. Gupta, et al., “Buried power rail scaling and metal assessment for the 3nm node and beyond”, IEDM, pp. 20.3.1-20.3.4, December 2020.

[15] A. Gupta, et al., “Buried power rail metal exploration towards the 1nm node”, IEDM, pp. 22.5.1-22.5.4, December 2021.

[16] B. Cline, et al., “Power from below: buried interconnects will help save Moore’s law”, IEEE Spectrum, Vol. 58, No. 9, pp. 44-51, September 2021.

[17] M. Badaroglu, et al., “PPAC scaling enablement for 5nm mobile SoC technology,” ESSDERC, September 2017.

[18] A. Veloso, et al., “Challenges and opportunities of vertical FET devices using 3D circuit design layouts”, IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), 2016.

[19] T. P. Ma, “Beyond Si: opportunities and challenges for CMOS technology based on high-mobility channel materials”, Sematech Symposium Taiwan, September 2012.

[20] T. Skotnicki and F. Boeuf, “How can high mobility channel materials boost or degrade performance in advanced CMOS”, VLSI, pp. 153-154, June 2010.

[21] C.-E. Tsai, et al., “Highly stacked 8 Ge0.9Sn0.1 nanosheet pFETs with ultrathin bodies (~3nm) and thick bodies (~30nm) featuring the respective record Ion/Ioff of 1.4x10e7 and record Ion of 92uA at Vov=Vds=-0.5V by CVD epitaxy and dry etching”, IEDM, pp. 569-572, December 2021.

[22] K. Kuhn, et al. “Past, present and future: SiGe and CMOS transistor scaling”, Electrochemical society trans., Vol. 33, No. 6, pp. 13-17, 2010.

[23] G. Eneman, et al., “Stress simulations for optimal mobility group IV p- and nMOS finFETs for the 14nm node and beyond”, IEDM, pp. 6.5.1-6.5.4, December 2012.

[24] R. Xie, et al., “A 7nm finFET technology featuring EUV patterning and dual strained high mobility channels”, IEDM, Section 2.7, December 2016.

[25] A. Agrawal, et al., “Gate-all-around strained Si0.4Ge0.6 nanosheet PMOS on strain relaxed buffer for high-performance low power logic application”, IEDM, pp. 2.2.1-2.2.4, December 2020.

[26] K.-W. Ang, et al., “Effective Schottky barrier height modulation using dielectric dipoles for source/drain specific contact resistivity improvement”, IEDM, pp. 18.6.1-18.6.4, December 2012.

[27] O. Gluschenkov, et al., “FinFET performance with Si:P and Ge: group-III-metal metastable contact trench alloys”, IEDM, December 2016.

[28] S.C Song, et al., “Holistic technology optimization and key enablers for 7nm mobile SoC,” VLSI, pp. T198-T199, June 2015.

[29] K. Cheng, et al., “Air spacer for 10nm finFET CMOS and beyond,” IEDM, December 2016.

[30] A. Keshavarzi, et al., “Architecting advanced technologies for 14nm and beyond with 3D FinFET transistors for the future SoC applications”, IEDM, pp. 4.1.1-4.1.4, December 2011.

[31] J. Mitard, et al., “15nm-wfin high-performance low-defectivity strained-germanium pFinFETs with low temperature STI-last process”, VLSI, pp. 1-2, June 2014.

[32] R. Xie, et al., “A 7nm finFET technology featuring EUV patterning and dual strained high mobility channels”, IEDM, December 2016.

[33] G. Eneman, et al., “Quantum barriers and ground-plane isolation: a path for scaling bulk-finFET technologies to the 7nm node and beyond”, IEDM, pp. 12.3.1-12.3.4, December 2013.

[34] F.-K. Hsueh, et al., “First demonstration of ultrafast laser annealed monolithic 3D gate-all-around CMOS logic and FeFET memory with near-memory-computing macro”, IEDM, pp. 40.4.1-40.4.4, December 2020.

[35] M.-G. Bardon, et al., “Extreme scaling enabled by 5 tracks cells: Holistic design-device co-optimization for finFETs and lateral nanowires”, IEDM, December 2016.

[36] A. Khakifirooz and D. A. Antoniadis, “Transistor performance scaling: The role of virtual source velocity and its mobility dependence,” IEDM, pp. 667–670, December 2006.

[37] J. Wang, et al., “Challenges and opportunities for stacked transistor: DTCO and device”, VLSI, paper T15-4, June 2021.

[38] S.C. Song, et al., “System design technology co-optimization for 3D integration at <5nm nodes”, IEDM, pp. 22.3.1-22.3.4, December 2021.

[39] R. Ritzenthaler, et al., “Comparison of electrical performance of co-integrated forksheets and nanosheets transistors for the 2nm technological node and beyond”, IEDM, pp. 26.2.1-26.2.4, December 2021.

[40] Y.-K. Cheng, et al., “Next-generation design and technology co-optimization (DTCO) of system on integrated chip (SoIC) for mobile and HPC application”, IEDM, pp. 41.3.1-41.3.4, December 2020.

[41] D.C.H. Yu, et al., “Foundry perspectives on 2.5D/3D integration and roadmap”, IEDM, pp. 3.7.1-3.7.4, December 2021.

[42] K. Jeong and A. Kahng, “A power-constrained MPU roadmap for the International Technology Roadmap for Semiconductors (ITRS),” Proc. Int. SoC Design Conf. (ISOCC), pp. 49-52, March 2010.

[43] P. Batude, et al., “GeOI and SOI 3D monolithic cell integrations for high density applications”, VLSI, A9-1, p.166-167, June 2009.

[44] I. Ouerghi, et al., « High performance polysilicon nanowire NEMS for CMOS embedded nanosensors”, IEDM, Section 22.4, p. 1-4, December 2014.

[45] P. Batude, et al., “3-D sequential integration: a key enabling technology for heterogeneous co-integration of new function with CMOS”, Journal on Emerging and Selected Topics in Circuits and Systems 2, p. 714-722, 2012.

[46] P. Coudrain, et al., “Setting up 3D sequential integration for back-illuminated CMOS image sensors with highly miniaturized pixels with low temperature fully depleted SOI transistors”, IEDM, December 2008.

[47] W. Rachmady, et al., “300mm heterogeneous 3D integration of record performance layer transfer germanium PMOS with silicon NMOS for low power high performance logic applications”, IEDM, Section 29.7, December 2019.

[48] J. Y. Kim, et al., "The breakthrough in data retention time of DRAM using recess-channel-array transistor (RCAT) for 88 nm feature size and beyond", VLSI, p.11, June 2003.

[49] J. Y. Kim, et al., "S-RCAT (sphere-shaped-recess-channel-array transistor) technology for 70nm DRAM feature size and beyond", VLSI, p.34, June 2005.

[50] S.-W. Chung, et al., "Highly scalable saddle-Fin (S-Fin) transistor for sub-50 nm DRAM technology", VLSI, p.32, June 2006.

[51] T. Schloesser, et al., "6F2 buried wordline DRAM cell for 40 nm and beyond", IEDM, p. 809, December 2008.

[52] D.-S. Kil, et al., “Development of new TiN/ZrO2/Al2O3/ZrO2/TiN capacitors extendable to 45nm generation DRAMs replacing HfO2 based dielectrics”, VLSI, p.38, June 2006.

[53] M. Sung, et al, “Gate-first high-k/metal gate DRAM technology for low power and high-performance products”, IEDM, December 2015.

[54] H. Tanaka, et al., "Bit cost scalable technology with punch and plug process for ultra-high-density flash memory", VLSI, pp. 14-15, June 2007.

[55] Y. Lu, et al., “Fully functional perpendicular STT-MRAM macro embedded in 40 nm logic for energy-efficient IoT applications”, IEDM, pp. 660-663, December 2015.

[56] O. Golonzka, et al., “MRAM as embedded non-volatile memory solution for 22FFL FinFET technology”, IEDM, Session 18.1, December 2018.

[57] J. Liang, et al., "A 1.4uA reset current phase change memory cell with integrated carbon nanotube electrodes for cross-point memory application", VLSI, 5B-4, June 2011.

[58] I.S. Kim, et al., "High-performance PRAM cell scalable to sub-20nm technology with below 4F2 cell Size, extendable to DRAM applications", VLSI, 19-3, June 2010.

[59] V. Sousa, et al., “Operation fundamentals in 12Mb phase change memory based on innovative Ge-rich GST materials featuring high reliability performance”, VLSI, June 2015.

[60] W.-C. Chien, et al., “Reliability study of a 128Mb phase change memory chip implemented with doped Ga-Sb-Ge with extraordinary thermal stability”, IEDM, S21.1, December 2016.

[61] H. Castro, "Accessing memory cells in parallel in a cross-point array", Publication 2015/0074326 A1 US Patent Office, March 12, 2015.

[62] DC Kau, et al., "A stackable cross point phase change memory", IEDM, pp. 617-620, December 2009.

[63] R. Freitas and W. Wilcke, "Storage class memory, the next storage system technology", 52(4/5), 439, IBM Journal of Research and Development, 2008.

[64] G.W. Burr, et al., "An overview of candidate device technologies for storage class memory", 52(4/5), 449, IBM Journal of Research and Development, 2008.

[65] H.Y. Chen, et al., "HfOx based vertical resistive random-access memory for cost-effective 3D cross-point architecture without cell selector", IEDM, pp. 497-500, (20.7.1-20.7.4), December 2012.