IEEE
Advancing Technology for Humanity

INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS

INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS

2017 EDITION

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

Acknowledgments – More Moore Team iii
1. Introduction 1
1.1. Current State of Technology 1
1.2. Drivers and Technology Targets 1
2. Summary and Key Points 2
3. Challenges 4
3.1. Near-term Challenges 4
3.2. Long-term Challenges 5
4. Technology Requirements—Logic Technologies 5
4.1. Ground Rules Scaling 5
4.2. Performance Boosters 7
4.3. Performance-Power-Area (PPA) Scaling 10
4.4. System-On-Chip (SoC) PPA Metrics 12
4.5. Interconnect Technology Requirements 14
4.6. Device Reliability 17
4.7. 3D Heterogeneous Integration 19
4.8. Defectivity Requirements 21
5. Technology Requirements—Memory Technologies 21
5.1. DRAM 21
5.2. NVM—Flash 22
5.3. NVM—Emerging 23
6. Potential Solutions 27
7. Cross Teams 28
8. Conclusions and Recommendations 28
9. References 29

List of Figures

Figure MM-1 Big data and instant data 1
Figure MM-2 Scaling of standard cell height and width through fin depopulation and device stacking 7
Figure MM-3 NAND2-eq standard cell count (left) and 111-bitcell (right) scaling in an 80mm^2 die 13
Figure MM-4 Number of CPU and GPU core in an 80mm^2 die 13
Figure MM-5 CPU clock frequency and datapath power at the iso frequency (referenced to 2017) scaling 14
Figure MM-6 Scaling projection of computation throughput of CPU cores at the maximum clock frequency and at thermally-constrained average frequency 14
Figure MM-7 Degradation paths in low-κ damascene structure 17
Figure MM-8 Defectivity (D0) requirements for >80% wafer sort yield target of an 80mm^2 die 21
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[38]. 23
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[48]. 26

List of Tables

Table MM-1 More Moore—Logic Core Device Technology Roadmap 3
Table MM-2 More Moore—DRAM Technology Roadmap 3
Table MM-3 More Moore—Flash Technology Roadmap 3
Table MM-4 More Moore—NVM Technology Roadmap 3
Table MM-5 Difficult Challenges—Near-term 4
Table MM-6 Difficult Challenges—Long-term 5
Table MM-7 Device Architecture and Ground Rules Roadmap for Logic Device Technologies 6
Table MM-8 Device Roadmap Enabling More Moore Scaling: 1) Device Architecture, 2) Performance Boosters, 3) Device Modules 7
Table MM-9 Projected Electrical Specifications of Logic Core Device 10
Table MM-10 Projected Performance-power-area (PPA) Metrics 11
Table MM-11 Integration Capacity of Logic Technology 12
Table MM-12 Power and Performance Scaling of SoC 13
Table MM-13 Interconnect Difficult Challenges 15
Table MM-14 Interconnect Roadmap for Scaling 15
Table MM-15 Device Reliability Difficult Challenges 18
Table MM-16 Potential Solutions—Near-term 27
Table MM-17 Potential Solutions—Long-term 27

Link to More Moore Excel Tables

ACKNOWLEDGMENTS – MORE MOORE TEAM

U.S.A.	
Anshul A. Vyas	Applied Materials
Arvind Kumar	IBM
Bhagawan Sahu	Global Foundries
Charles Kin P. Cheung	NIST
Chorng-Ping Chang	AMAT
Christopher Henderson	Semitracks
Eric Snyder	MKS Inst
Gennadi Bersuker	Aerospace Corporation
Gerhard Klimeck	Purdue Univ.
Huiming Bu	IBM
James Stathis	IBM
Jim Fonseca	Purdue Univ.
Jim Hutchby	SRC
Joe Brewer	Univ. Florida
Joel Barnett	TEL
Kirk Prall	Micron
Kwok Ng	SRC
Mehdi Salmani	Boston Consulting Group
Paul Mertens	IMEC
Philip Wong	Stanford Univ.
Prasad Sarangapani	Purdue Univ.
Qi Xiang	Xilinx
Rich Liu	Macronix
SangBum Kim	IBM
Saumitra Mehrotra	NXP
Saurabh Sinha	ARM
Siddharth Potbhare	NIST
Sung Geun Kim	Microsoft
Takeshi Nogami	IBM
Terry Hook	IBM
Witek Maszara	Global Foundries
Yanzhong Xu	Intel PSG

ASIA	
Atsushi Hori	Kyocera Corporation
Digh Hisamoto	Hitachi
Hajime Nakabasyashi	TEL
Hitoshi Wakabayashi	Tokyo Inst of Technology
Jiro Ida	Kanazawa IT
Kunihiko Iwamoro	ROHM
Masami Hane	Renesas
Satoshi Kamiyama	TEL
Shinichi Ogawa	AIST
Shinichi Takagi	University of Tokyo
Takashi Matsukawa	AIST
Tesuo Endo	Tohoku University
Tetsu Tanaka	Tohoku University
Toshiro Hiramoto	University of Tokyo
Yasuo Kunii	Hitachi
Yasushi Akasaka	TEL
Yuzo Fukuzaki	Sony
Jongwoo Park	Samsung
Moon-Young Jeong	Samsung
Sang Hyun Oh	SK Hynix
Cheng-tzung Tsai	UMC
Geoffrey Yeap	TSMC
Samuel C. Pan	TSMC
Tony Oates	TSMC
Wilman Tsai	TSMC

EUROPE	
Alex Burenkov	Fraunhofer
Christiane Le Tiec	MKS Instruments
Dan Mocuta	IMEC
Fred Kuper	NXP
Francis Balestra	IMEP Grenoble
Gerben Doornbos	TSMC
Herve Jaouen	ST
Jurgen Lorenz	Fraunhofer IISB
Kristin DeMeyer	IMEC
Laurent Le-Pailleur	ST
Malgorzata Jurczak	ASM Int
Mark van Dal	TSMC
Matthias Passlack	TSMC
Michel Haond	ST
Mustafa Badaroglu	Qualcomm
Olivier Faynot	LETI
Robert Lander	NXP
Thierry Poiroux	LETI
Yannick Le Tiec	LETI

MORE MOORE

1. INTRODUCTION

System scaling enabled by Moore's scaling is increasingly challenged with the scarcity of resources such as power and interconnect bandwidth. This is particularly due to the emergence of cloud, seamless interaction of big data, and instant data that have become a necessity (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 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 More Moore power, performance, area, cost (PPAC) scaling for big data, mobility, and cloud (e.g., Internet-of-Things (IoT) and server) applications. The IFT then forecasts logic and memory technologies over the roadmap time horizon of 15 years for main-stream/high-volume manufacturing (HVM).

Figure MM-1 Big data and instant data

1.1. CURRENT STATE OF TECHNOLOGY

A major portion of semiconductor device production is devoted to digital logic. Both high-performance logic and low-power logic that is typically for mobile applications are included. Detailed technology requirements and potential solutions are considered for both types in the same logic platform. Key considerations are speed, power, density requirements, and the targets for each. One key theme is the continued scaling of MOSFETs for leading-edge logic technology in order to maintain historical trends of improved device performance at reduced power and cost.

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: >15% more operating frequency at scaled supply voltage
- (P)ower: >35% less energy per switching at a given performance
- (A)rea: >35% less chip area footprint
- (C)ost: <30% more wafer cost – 20% less die cost for scaled die.

These scaling targets have driven the industry toward a number of major technological innovations, including material and process changes such as higher-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 heterogenous stacking/integration. These innovations are expected to be introduced at a rapid pace, and hence understanding, modeling, and implementation into manufacturing in a timely manner is expected to be a major issue for the industry.

It is important to note that the cost metric (20% less die cost) and market cadence necessitating new products every year are becoming more important targets with the mobile industry. As the applications strictly requiring all figure-of-merits (FoMs) are concurrently met, it is necessary to advance an effective list of knobs for sustaining certain device architectures to their limits, such as pushing the finFET architecture for the next five years. 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 as an outcome of the multiple patterning lithography steps. However, we need to reduce the cost by more than 20% for the same of number of transistors, which can only be enabled by pitch scaling due to 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, an 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 used those as calibration factors to match the area scaling trends of the industry[2]. The design scaling factor is also addressed as a key element of this 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 2027. Extreme-ultraviolet (EUV) technology is now started to slow down this saturation trend by having the cost under control due to process complexity reduction. Transition to 3D integration and use of beyond CMOS devices for complementary System-on-Chip (SoC) functions are projected after 2027.
- Ground-rule scaling needs to also enable design-technology-co-optimization (DTCO) constructs that accommodate the area reduction as well as tighten the critical design that limits 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 the major knob for performance improvement for nodes spanning between 2017 and 2024, due to the introduction of low-κ device spacer.
- SiGe and Ge channels are gaining importance as the high-mobility channels. III-V channel faces challenges of variability, band-to-band tunneling, and large investments in fab infrastructure.
- 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 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, forcing a slow-down in the permittivity (κ-value) scaling. Use of non-Cu solutions is necessary to cope with the increased electro-migration risk due to the decreased metal volume ramping up the current density.

- Performance scaling across 7 nodes spanning from 2017 to 2033 is a 9% node-to-node improvement for datapaths without wireload while it also incurs a 10% node-to-node penalty for datapaths loaded with tight pitch metal routing. If wireload routing is done with intermediate metal (at 80nm pitch), 8% node-to-node performance improvement can be realized.
- Clocking frequency at nominal supply voltage is expected to be improved from 2.5GHz (in 2017) to 4.2 GHz (in 2033). Power density poses a significant challenge for scaling where average clocking will stall to 0.5 GHz in 2033 if the same chip is operated at constant power density across nodes. Therefore, it is necessary to factor in thermal considerations in device and architectures.
- Energy per switching reduction for logic devices is expected to be more or less on track, about 19% reduction 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 tremendously low in comparison with introducing the new technology, DRAM scaling will be stopped and 3D cell stacking structure such as 3D-NAND 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, 64 layers are starting at volume production and there is optimism that 128 layers are achievable, with 192 and 256 layers possible.
- Ferroelectric RAM (FeRAM) is fast, low power, and low voltage and thus is suitable for radio frequency identification (RFID), smart card, ID card, and other embedded applications. Processing difficulty and high cost (compared to Flash memories) limit wider adoption. Recently, HfO2-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 NAND Flash 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. Therefore, 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 crosspoint 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.

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 in overall performance. The performance of the components and the final chip can be measured in many different ways: higher speed, higher density, lower power, 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 2017-2024 (Table MM-5) and long-term 2025-2033 (Table MM-6).

3.1. NEAR-TERM CHALLENGES

Table MM-5 Difficult Challenges—Near-term

Near-Term Challenges: 2017-2024	**Description**
Power scaling	Voltage and capacitance scaling slow-down and lack of knobs 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. Loading capacitance is now mostly due to the parasitic components of the device that, with continuous scaling of ground rules, cause those components to dominantly form a significant portion of overall capacitance. Therefore, an introduction of low-κ materials, design-technology-co-optimization introducing new contact access schemes as well as local interconnect schemes that allow lower parasitics is needed.
Parasitics scaling	Maintaining control of increased parasitics in vertically stacked devices. Vertical 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 and design-technology-co-optimization (DTCO). Throughput and yield challenges of EUV necessitate a careful selection of ground rules that optimizes the die cost as most of the cost is determined by the middle-of-line (MOL) and BEOL stack. Therefore, new 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.
Integration enablement for SRAM-cache applications	Bitcell scaling is slowing down because of the slow-down of device vertical (e.g. fin pitch) and horizontal pitch (contacted poly pitch). New device schemes such as P-over-N stacked device or vertical devices bring an opportunity to significantly reduce the SRAM area. This is due to the 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 the BEOL stack.
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 need for new barrier materials and Cu alternative solutions. In addition to resistance scalability, time-dependent-dielectric-breakdown (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: 2025-2033	Description
Power scaling	Power scaling—no knobs are left besides steep-subthreshold (SS) device as complementary and architecture. However, most of steep-SS device candidates do not bring an adequate performance comparable to CMOS at nominal supply voltages. In order to maximize the performance of steep-SS device, current architectures need to attain the performance through parallelization.
Use cases of vertical device structures	Performance scaling and functional diversification with vertical devices and new architectures. Using vertical devices at conventional logic and architectures will raise routing congestion and increased parasitics. There is a need for new logic schemes that maximize the advantage of 3D capability.
Thermal issue due to increased power density	Thermal challenges (e.g., power density and dark silicon) of 3D stacking. Gate-all-around devices have limited heat conductance due to confined architecture. Increased pin density due to aggressive standard cell height scaling and increasing drive by vertically stacked devices put a significant pressure to the power density.
Cost reduction with 3D integration	Managing cost, yield, and process complexity of3D 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.

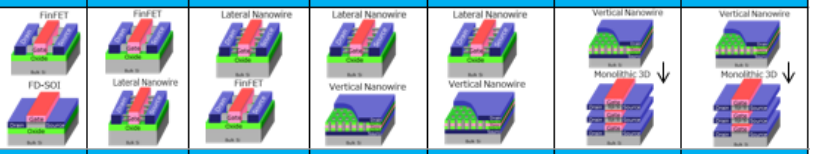
4. TECHNOLOGY REQUIREMENTS—LOGIC TECHNOLOGIES

4.1. GROUND RULES SCALING

The More Moore roadmap focuses on effective knobs to sustain the performance scaling at scaled dimensions and scaled supply voltage. Ground rule scaling drives the die cost reduction while maintaining the reduction of parasitics as a function of geometric scaling. On the other hand, increasing portion of parasitics in the total loading result in diminishing returns of scale. Therefore, it is necessary to focus on technology scaling knobs that also scale the parasitics of device and interconnect. Ground-rule scaling needs to also enable design-technology-co-optimization (DTCO) constructs that accommodate the area reduction as well as tighten the critical design that limits area scaling. Due to the rising costs and process complexity of multiple patterning, EUV is used to enable single-exposure patterning of tight ground rules. The projected roadmap of ground rules as well as device architectures is shown in Table MM-7. 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 contacted poly pitch, metal pitch, fin pitch, and gate length, which are important factors in core logic area scaling.

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

Note: PxxMxxTx notation refers to Pxx: contacted poly pitch, Mxx: metalx pitch in nm, Tx: number of tiers. This notation illustrates the technology capability. On top of pitch scaling there are other elements such as cell height, vertical integration, fin depopulation, DTCO constructs, etc. that define the target area scaling (gates/mm2).

YEAR OF PRODUCTION	2017	2019	2021	2024	2027	2030	2033
	P54M36	P48M28	P42M24	P36M21	P32M14	P32M14T2	P32M14T4
Logic industry "Node Range" Labeling (nm)	"10"	"7"	"5"	"3"	"2.1"	"1.5"	"1.0"
IDM-Foundry node labeling	i10-f7	i7-f5	i5-f3	i3-f2.1	i2.1-f1.5	i1.5-f1.0	i1.0-f0.7
Logic device structure options	finFET FDSOI	finFET LGAA	LGAA finFET	LGAA VGAA	LGAA VGAA	VGAA, LGAA 3DVLSI	VGAA, LGAA 3DVLSI
Logic device mainstream device	finFET	finFET	LGAA	LGAA	LGAA	VGAA	VGAA
DEVICE STRUCTURES							
							
LOGIC DEVICE GROUND RULES							
MPU/SoC Metalx ½ Pitch (nm)[1,2]	18.0	14.0	12.0	10.5	7.0	7.0	7.0
MPU/SoC Metal0/1 ½ Pitch (nm)	18.0	14.0	12.0	10.5	7.0	7.0	7.0
Contacted poly half pitch (nm)	27.0	24.0	21.0	18.0	16.0	16.0	16.0
L_g: Physical Gate Length for HP Logic (nm) [3]	20	18	16	14	12	12	12
L_g: Physical Gate Length for LP Logic (nm)	22	20	18	16	14	14	14
Channel overlap ratio - two-sided	0.80	0.80	0.80	0.80	0.80	0.80	0.80
Spacer width (nm)	8	7	6	5	5	5	5
Contact CD (nm) - finFET, LGAA	18	16	14	12	10		
Contact CD (nm) - VGAA						12	12
Device architecture key ground rules							
FinFET Fin Half-pitch (nm)	16.0	14.0					
FinFET Fin Width (nm)	8.0	7.0					
FinFET Fin Height (nm)	45	50					
Footprint drive efficiency - finFET	3.06	3.82					
Lateral GAA lateral half-pitch (nm)			12.0	10.5	9.0		
Lateral GAA vertical half-pitch (nm)			8.0	8.0	8.0		
Lateral GAA (nanosheet) thickness (nm)			5.0	5.0	5.0		
Lateral GAA (nanosheet) minimum width (nm)			7.0	7.0	6.0		
Number of vertically stacked nanosheets			3	4	5		
Device height (nm)			47	63	79		
Footprint drive efficiency - lateral GAA			3.00	4.57	6.11		
Vertical GAA lateral half-pitch (nm)						7.0	7.0
Vertical GAA width (nm)						6.0	6.0
Contact-gate enclosure (nm)						2.0	2.0
Footprint drive efficiency - vertical GAA						1.7	1.7
Defice effective width (nm)	98.0	107.0	72.0	96.0	110.0	24.0	24.0
Device lateral half pitch (nm)	16.0	14.0	12.0	10.5	9.0	7.0	7.0
Device height (nm)	45.0	50.0	47.0	63.0	79.0	24.0	24.0
Minimum device width (fin, nanosheet) or diameter (nm)	8.0	7.0	7.0	7.0	6.0	6.0	6.0

Acronyms used in the table (in order of appearance): FDSOI: Fully-Depleted Silicon-On-Insulator (FDSOI), LGAA: Lateral Gate-All-Around-Device (GAA), VGAA: Vertical GAA, 3DVLSI: Fine-pitch 3D logic sequential integration.

Ground rule scaling stand-alone is not adequate enough to scale the cell height. It is necessary to bring the design scaling factor into practice[2]. For example, standard cell height will be further reduced by scaling the number of active devices in the standard cell as well as scaling the secondary rules such as tip-to-tip, extension, P-N seperation, 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 allowing processes such as contact-over-active[4]. Also, the contact structure needs to be carefully selected to reduce the risk of increased current density at the junctions. It is expected that in 2024 P and N devices could be stacked on top of each other allowing a further reduction. This trend in standard cell scaling is shown in Figure MM-2.

After 2027 there is no room for 2D geometry scaling where 3D very large scale integration (VLSI) of circuits and systems using sequential/stacked integration approaches. This is due to the fact that there is no room for contact placement as well as worsening performance as a result of contacted poly pitch (CPP) scaling and metal pitch scaling. It is projected that physical channel length would saturate around 12nm due to worsening electrostatics while CPP would saturate at 32nm to reserve sufficient CD (~12nm) for the device contact, providing acceptable parasitics. For the vertical GAA physical gate length could be kept less tight as the gate length is determined by the thickness of stack instead of footprint space. But this relaxation of gate length in vertical GAA is constrained by the power penalty as a result of increase in the channel capacitance. 3D VLSI expects to bring PPAC gains for the target node as well as to pave ways for heterogeneous 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 is required after 2027. 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.

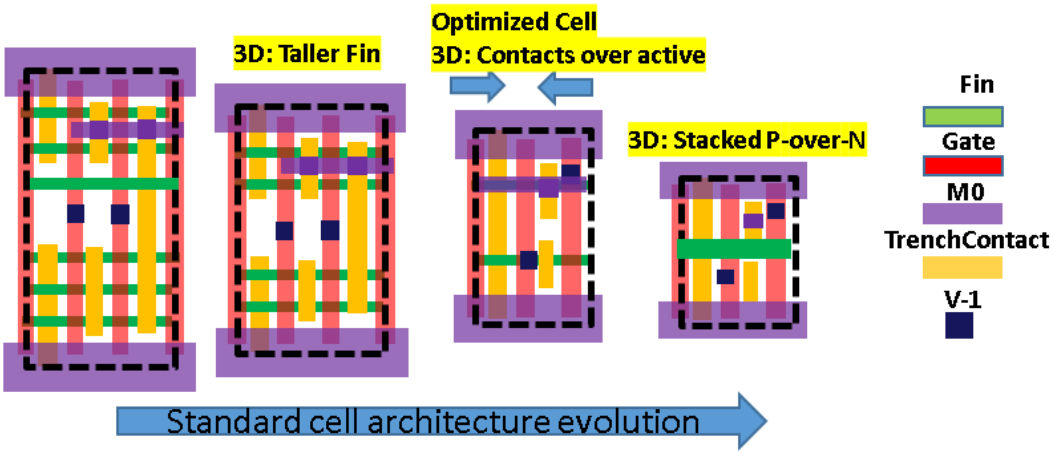


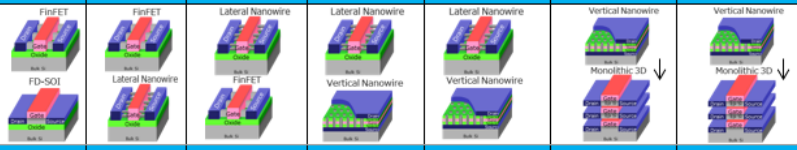
Figure MM-2 Scaling of standard cell height and width through fin depopulation and device stacking

4.2. PERFORMANCE BOOSTERS

In the early years before the 130nm node, transistors enjoyed Dennard scaling where equivalent oxide thickness (EOT), transistor gate length (Lg) and transistor width (W) were scaled by a constant factor in order to provide a delay improvement at constant power density. Nowadays there are numerous input parameters that can be varied, and the output parameters are complicated functions of these input parameters. Other sets of projected parameter values (i.e., different scaling scenarios) may be found to achieve the same target. In order to maintain the scaling at low voltages, scaling in recent years focused on additional knobs 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 Enabling More Moore Scaling: 1) Device Architecture, 2) Performance Boosters, 3) Device Modules

YEAR OF PRODUCTION	2017	2019	2021	2024	2027	2030	2033
	P54M36	P48M28	P42M24	P36M21	P32M14	P32M14T2	P32M14T4
Logic industry "Node Range" Labeling (nm)	"10"	"7"	"5"	"3"	"2.1"	"1.5"	"1.0"
IDM-Foundry node labeling	i10-f7	i7-f5	i5-f3	i3-f2.1	i2.1-f1.5	i1.5-f1.0	i1.0-f0.7
Logic device structure options	finFET FDSOI	finFET LGAA	LGAA finFET	LGAA VGAA	LGAA VGAA	VGAA, LGAA 3DVLSI	VGAA, LGAA 3DVLSI
Logic device mainstream device	finFET	finFET	LGAA	LGAA	LGAA	VGAA	VGAA
DEVICE STRUCTURES							
							
LOGIC TECHNOLOGY ANCHORS							
Patterning technology inflection for Mx interconnect	193i, EUV	193i, EUV DP	193i, EUV DP	193i, High-NA EUV	193i, High-NA EUV+(DSA)	193i, High-NA EUV+(DSA)	193i, High-NA EUV+(DSA)
Channel material technology inflection	Si	SiGe25%	SiGe50%	Ge, IIIV (TFET?), 2D Mat	Ge, IIIV (TFET?), 2D Mat	Ge, IIIV (TFET?), 2D Mat	Ge, IIIV (TFET?), 2D Mat
Process technogy inflection	Conformal deposition	Conformal Doping, Contact	Channel, RMG	Stacked-device Non-Cu Mx	Stacked-device Non-Cu Mx	Steep-SS, 3D	Steep-SS, 3D
Stacking generation inflection	2D	2D	3D-stacking: W2W D2W	3D-device: P-over-N Hetero	3D-device: Mem-on-Logic Hetero	3D-device: Mem-on-Logic Hetero	3D-device: Logic-on-Logic Hetero

FDSOI: Fully-Depleted Silicon-On-Insulator (FDSOI), LGAA: Lateral Gate-All-Around-Device (GAA), VGAA: Vertical GAA, TFET: Tunneling FET, SS: Subthreshold Slope, Mx: Tight-pitch metal interconnect.

FinFET still remains the key device architecture that could sustain scaling until 2021 for high-performance logic applications[4]. Electrostatics and fin depopulation, increasing fin height while reducing number of fins at unit footprint area remain as the two effective knobs to improve performance. Beyond 2019 parasitics scaling becomes the major knob

as a result of tightening design rules. It is forecasted that the parasitics will be more a dominant term in the performance of critical paths. For reduced supply voltage a transition to GAA structures such as lateral nanowires would be necessary to sustain the gate drive by improved electrostatics. Lateral GAA structure would eventually evolve to vertical GAA structure to gain back the performance loss due to increasing parasitics at tighter pitches. Thanks to evolution of those vertical GAA structures, sequential integration would allow stacking of devices on top of each other with the adoption of monolithic 3D (M3D) integration, the so-called sequential/stacked integration approaches[6]. Scaling focus will shift from 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.

While device architectures are seeing changes, subsequent modules should also evolve. Those could be for instance: 1) starting substrates such as Si to silicon-on-insultator (SOI) and strain-relaxation-buffer (SRB); 2) channel material evolving from Si to SiGe, Ge, IIIV; 3) contact module evolving from silicides to novel materials providing lower Schottky barrier height (SBH) and to 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 sustain until the end of 2023. Beyond 2019 a transition to GAA will start and potentially a transition to vertical nanowires devices will be needed when there is no room left for the gate length scale down. 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. One compromise solution could be the electrically GAA nanowire (EGAA NW) architecture with much reduced parasitic capacitance and increased effective width for better short channel control and stronger drive [7].

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. SOI also provides a knob for threshold voltage (Vt) tuning, allowing to tune a device to either high-performance or low-leakage, thanks to the backgate control.

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’, scale 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 bring high tunneling current at higher supply voltage. This may degrade 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 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[7]. This requires a correct selection of 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[9]. A shift in the centroid of charge away from the gate potential adds to the equivalent oxide thickness (EOT), reducing the inversion capacitance, particularly in III-V high-mobility channels. Despite the fact that drive current of III-V might not be that high, the overall delay merit (CV/I) can result better than the ones of Si and other high-mobility channels (e.g. Ge). On the other hand, Vt variability due to channel dimensions and composition appears to become a major impediment in using IIIV channel material in scaled devices. Band gap and thus Vt seem highly modulated by body thickness due to quantum confinement effects for device with body thickness/diameter around 5-6 nm. Si and Ge appear to have much less sensitivity to such channel dimension variations. Also the impact of chemical composition variation in ternary III-V, like InGaAs, might also cause Vt variation. Indium % change impacts band gap, which also impacts Vt. The cost factor should also be taken into account, such as the requirement of new tools as well as an infrastructure for dealing with potentially toxic waste requiring substantial investment in new fabs. Thus, the improved performance needs to be weighed against the cost, as this could be a greater factor compared to other options.

4.2.4. Strain engineering

This knob has been used as one of the most effective knobs in the last decade as illustrated for the 32nm node and earlier [10]. However, affect of those stressors may not extrapolate intuitively into newer nodes. With the scaling down of contacted

poly pitch, SiGe on the soure/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[11]. SiGe channel for PMOS and strained Si channel for NMOS has been successfully demonstrated on a 7nm CMOS platform using SRB[12]. On the other hand, SRB or S/D stressors may not be useful for channel stress generation in vertical devices, which appear in the roadmap around 2021. Other strain engineering techniques also contain gate stressor and ground plane stressors, which adopt the beneficiary vertical stress components for NMOS. Compressively strained SiGe channel is also shown in ultra-thin body and buried oxide fully depleted SOI (UTBB FDSOI) in order to boost pFET performance[13][14]. A high level of stress is maintained in the channel thanks to the planar configuration (with low aspect ratio, compared to finFET). Combined with the use of back-bias (to reduce Vdd and thus the dynamic power), it enables high performance, low power circuits on UTBB FDSOI.

4.2.5. Reducing parasitic device resistance

Controlling source/drain series resistance within tolerable limits will become much more 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 contacted poly pitch (CPP) scaling. On top of this 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 getting off-stream in maintaining the required reduction of contact resistance with the poly 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)[15][16]. This SBH reduction happens by the exponential decay of the metal induced gap states (MIGS) inducing charge density accumulation in the bandgap of the dielectric.

4.2.6. 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 elevated in the GAA structures as a result of unused space between consecutive devices. There is a need to focus on low-κ spacer materials and even air spacer that still provide good reliability and etch selectivity for S/D contact formation[17][18]. It appears that there are significant limits in increasing finFET or lateral GAA device AC performance by increasing the height of the device (fin/nanowire stack). Energy per switch vs. delay relationship seems to quickly saturate and then decline with increasing height.

4.2.7. Increasing drive per footprint

FinFET and lateral GAA devices enable a higher drive at unit footprint (by enabling drive in the third dimension) if fin pitch can be aggressively scaled while increasing the fin height[17][19]. This will then increase drive at unit footprint by scaling the fin pitch comes at 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, therefore, the overall chip area.

4.2.8. Improving electrostatics

FinFET has better 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 (i.e., body thickness) compared to UTBB FDSOI. In UTBB FDSOI electrostatic control could be done by using silicon (i.e., body) thickness and buried oxide (BOX) thickness where convergent scaling of both silicon thickness and BOX thickness enables electrostatics scaling (i.e., drain-induced barrier lowering (DIBL) <100 mV/V) down to Lgate beyond 10 nm. Thick buried oxide (Tbox) and thin Si (Tsi) scalings are typically kept at compromise between manufacturability and short-channel-effects control. Junction implantation engineering, EOT scaling and density of interface traps (Dit) reduction are potential solutions to maintain the electrostatics control in the channel[20][21].

4.2.9. 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 Ge channels because of low effective mass of Ge. Ground plane doping and quantum well below the channel would potentially solve this leakage problem; therefore improving the electrostatics[22].

4.2.10. Reducing process and material variations

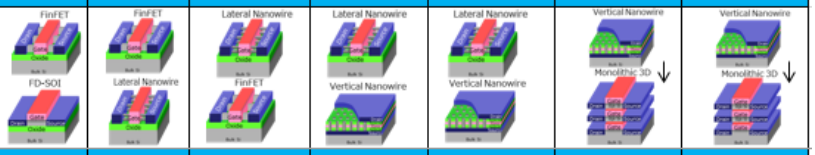
Reducing variability would further allow Vdd scaling. Controlling channel length and channel thickness are important to maintain the electrostatics in the channel. This would require, for example, controlling the profile of the fin and lithography processes to reduce the CD uniformity (CDU), line width roughness (LWR), 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.

4.2.11. Beyond CMOS for application-specific functions and architectures

Finally, beyond the roadmap range of this edition (beyond 2030), MOSFET scaling will likely become ineffective and/or very costly. Completely new, non-CMOS types of logic devices and maybe even new circuit architectures are potential solutions (see Emerging Research Devices section 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 2024 frame by tunneling field-effect transistor (TFET) for ultra-low power applications and memristors for neuromorphic applications.

The projected roadmap for the electrical specifications of logic core device is listed in Table MM-9.

Table MM-9 Projected Electrical Specifications of Logic Core Device

YEAR OF PRODUCTION	2017	2019	2021	2024	2027	2030	2033
	P54M36	P48M28	P42M24	P36M21	P32M14	P32M14T2	P32M14T4
Logic industry "Node Range" Labeling (nm)	"10"	"7"	"5"	"3"	"2.1"	"1.5"	"1.0"
IDM-Foundry node labeling	i10-f7	i7-f5	i5-f3	i3-f2.1	i2.1-f1.5	i1.5-f1.0	i1.0-f0.7
Logic device structure options	finFET FDSOI	finFET LGAA	LGAA finFET	LGAA VGAA	LGAA VGAA	VGAA, LGAA 3DVLSI	VGAA, LGAA 3DVLSI
Logic device mainstream device	finFET	finFET	LGAA	LGAA	LGAA	VGAA	VGAA
DEVICE STRUCTURES							
							
DEVICE ELECTRICAL SPECS							
Logic device mainstream device	finFET	finFET	LGAA	LGAA	LGAA	VGAA	VGAA
Power Supply Voltage - Vdd (V)	0.75	0.70	0.65	0.65	0.65	0.60	0.55
Subthreshold slope - [mV/dec]	68	75	67	72	75	50	40
Inversion layer thickness - [nm] [4]	1.10	1.10	1.00	1.00	1.00	1.00	1.00
Vt,sat (mV) at Ioff=10nA/um - HP Logic [5][6]	190	213	194	212	226	151	121
Vt,sat (mV) at Ioff=100pA/um - LP Logic [5][6]	323	360	324	352	371	248	198
Effective mobility (cm2/V.s)	150	125	125	100	100	100	100
Rext (Ohms.um) - HP Logic [7]	300	285	271	257	244	232	221
Ballisticity,Injection velocity (cm/s)	1.32E-07	1.39E-07	1.46E-07	1.46E-07	1.46E-07	1.46E-07	1.46E-07
Vdsat (V) - HP Logic	0.141	0.160	0.149	0.163	0.140	0.140	0.140
Vdsat (V) - LP Logic	0.155	0.177	0.168	0.186	0.163	0.163	0.163
Ion (uA/um) at Ioff=10nA/um - HP logic w/ Rext=0 [8]	2027	1770	1917	1788	1777	1905	1805
Ion (uA/um) at Ioff=10nA/um - HP logic, after Rext [9]	972	869	897	872	878	960	937
Ion (uA/device) at Ioff=100nA/um - HP logic, after Rext [9]	95	93	84	84	97	23	22
Ion (uA/um) at Ioff=100pA/um - LP logic w/ Rext=0 [8]	1447	1094	1214	1028	991	1361	1359
Ion (uA/um) at Ioff=100pA/um - LP logic, after Rext [9]	596	441	468	404	387	604	637
Ion (uA/device) at Ioff=100pA/um - LP logic, after Rext [9]	58	47	39	39	43	14	15
Cch,total (fF/um2) - HP/LP Logic [9]	31.38	31.38	34.52	34.52	34.52	34.52	34.52
Spacer k value	4.50	4.00	3.50	3.50	3.00	2.50	2.50
Fringe capa scaling factor	1.00	0.90	1.11	1.17	0.94	0.85	0.85
Cgate,total (fF/um) - HP Logic [10]	1.51	1.27	1.42	1.29	0.95	0.89	0.89
Cgate,total (fF/um) - LP Logic [10]	1.66	1.41	1.60	1.47	1.11	1.04	1.04
CV/I (ps) - FO3 load, HP Logic [11]	3.49	3.07	3.09	2.88	2.12	1.67	1.57
I/(CV) (1/ps) - FO3 load, HP Logic [12]	0.29	0.33	0.32	0.35	0.47	0.60	0.64
Energy per switching [CV2] (fJ/switching) - FO3 load, HP Logic	2.54	1.87	1.80	1.63	1.21	0.96	0.81

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[23]. As shown in the logic core technology table, 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 invertor driving a star wireload with its branches driving

three D4 invertors. Wireload model is in pi2 configuration to account the distributed RC effect. Details of this modeling how interconnect is coupled with the device in the standard-cell context are explained in[2]. For circuit-level transient simulations we used a virtual source model (VSM) environment, which is open source distribution from MIT[24]. We used virtual source model (VSM) to capture the circuit-level parameters such as delay and power per stage from a ring oscillator. Its inputs are validated in technology computer aided design (TCAD) with the support from the NanoHub Team of Purdue University[25]. There are also analytical modeling tools such as MASTAR[26], which is an analytical modeling tool to capture the major device characteristics such as Ion, Ieff, and Ioff. MASTAR was used in the editions of ITRS before 2013.

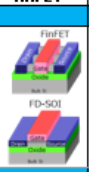
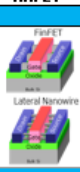
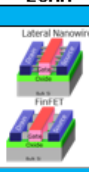
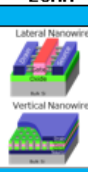
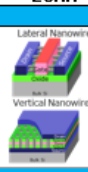
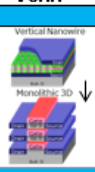
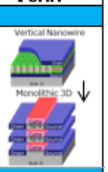
In this datapath model the delay of each stage is approximated by the Elmore expression given below[2]:

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

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, Cout is the load capacitance due to the gates connected to the load, and WL is the wire length. For logic technologies beyond 2017 the dominant term is typically found to be Rw*Cout[2]. This means that increasing the driver strength does not really help if there is no improvement in the parasitic resistance of interconnect and/or a reduction in the parasitic loading of standard cell.

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 Projected Performance-Power-Area (PPA) Metrics

YEAR OF PRODUCTION	2017	2019	2021	2024	2027	2030	2033
	P54M36	P48M28	P42M24	P36M21	P32M14	P32M14T2	P32M14T4
Logic industry "Node Range" Labeling (nm)	"10"	"7"	"5"	"3"	"2.1"	"1.5"	"1.0"
IDM-Foundry node labeling	i10-f7	i7-f5	i5-f3	i3-f2.1	i2.1-f1.5	i1.5-f1.0	i1.0-f0.7
Logic device structure options	finFET FDSOI	finFET LGAA	LGAA finFET	LGAA VGAA	LGAA VGAA	VGAA, LGAA 3DVLSI	VGAA, LGAA 3DVLSI
Logic device mainstream device	finFET	finFET	LGAA	LGAA	LGAA	VGAA	VGAA
DEVICE STRUCTURES							
							
LOGIC CELL AND FUNCTIONAL FABRIC TARGETS							
Logic device mainstream device	finFET	finFET	LGAA	LGAA	LGAA	VGAA	VGAA
SRAM height in device grids	10	10	10	10	10	8	8
SRAM bitcell area (um2)	0.0346	0.0269	0.0202	0.0151	0.0115	0.0072	0.0072
SRAM 111 bitcell area density - Mbits/mm2	29	37	50	66	87	140	140
NAND2 active devices per pull-up and pull-down for datapath cell	2	2	1	1	1	4	4
NAND2 active devices per pull-up and pull-down for low power	2	1	1	1	1	1	1
NAND2 width at CPP multiples	3	3	3	3	3	2	2
Average cell width scaling factor	1.00	0.90	0.90	0.90	0.90	0.90	0.90
Dummy devices in standard cell	4	3	3	3	3	2	2
Width increment (nm)	-	-	15	6	0	-	-
Active width (nm)	-	-	22	13	6	-	-
Cell height (nm)	256	196	160	115	90	140	140
NAND2 equivalent raw-gate density - Mgates/mm2	24	39	55	89	129	124	124
Effective drive width of D4 ND2 cell (nm) - datapath cell	784	856	648	576	440	384	384
Effective drive width of D1 ND2 cell (nm) - low power cell	196	107	72	96	110	24	24
Cell drive at saturation (Ohms)	1969	1881	2237	2589	3367	3256	3058
Cell related routing (FO3) capacitance (fF)	1.50	1.29	1.05	0.76	0.76	0.92	0.92
Cell related datapath loading (FO3) device-related capacitance	7.08	6.52	5.52	4.45	2.52	2.06	2.06
Cell related low-power loading (FO3) device-related capacitance	1.77	0.81	0.61	0.74	0.63	0.13	0.13
Cell related datapath loading (FO3) capacitance (fF)	8.58	7.81	6.57	5.21	3.28	2.98	2.98
Cell related low-power loading (FO3) capacitance (fF)	3.27	2.11	1.67	1.50	1.39	1.05	1.05
Wirelength (um) - 30x(8*CPP*WidthScale+CellHeight)	14.16	11.06	9.34	7.34	6.16	7.66	7.66
Wireload resistance (Ohms) - WireResistance+4 vias - tight	1961	4072	4908	6884	15670	19420	19420
Wireload capacitance (fF) - tight pitch	2.83	2.21	1.87	1.47	1.23	1.53	1.53
Wireload resistance (Ohms) - WireResistance+6 vias - 80nm	284	284	281	275	260	280	280
Wireload capacitance (fF) - 80nm pitch	2.97	2.32	1.96	1.54	1.29	1.61	1.61
FO3+Wireload stage delay (ps) - no wireload	11.66	10.14	10.14	9.31	7.62	6.69	6.29
FO3+Wireload stage delay (ps) - tight pitch wireload	29.24	38.38	38.76	40.53	53.27	61.36	60.74
FO3+Wireload stage delay (ps) - 80nm pitch wireload	17.71	14.94	14.65	13.22	11.34	11.05	10.42
FO3+Wireload stage dynamic power at 1GHz clock, 80nm pitch wireload (mW)	3.51	2.17	1.53	1.29	1.13	0.96	0.80
FO3+Wireload stage IDDQ at 1GHz clock - nW	235.20	239.68	168.48	149.76	114.40	92.16	84.48

Performance scaling across 7 nodes spanning from 2017 to 2033 is 9% node-to-node improvement for datapaths without wireload while it becomes 10% node-to-node penalty for datapaths loaded with tight pitch metal routing. If wireload routing

is done with intermediate metal (at 80nm pitch), 8% node-to-node performance improvement can be realized This scheme then requires an effective reduction of vertical resistance (which is the cumulative sum of via resistances) in order to retrieve the performance gains whenever the routing of critical paths is done in the intermediate metallization. We also factor in the wirelength reduction as function of CPP and cell height scaling where scaling helps for the reduction of wire-related loading capacitance.

Energy per switching reduction becomes limited, about 19% reduction in a node-to-node basis in average. This is majorly achieved to fin/device depopulation, which also enabled the cell height reduction bringing a scaling of wire and cell related capacitances. We also assume that design-technology-co-optimization (DTCO) constructs such as contact-over-active, single diffusion break, etc, as described in[4][7], will further reduce the standard cell width, which is assumed at a scale of ×0.9 in a node-to-node basis on top of conventional CPP scaling. Raw gate density is improved by around ×1.4 in a node-to-node basis until 2027 with a slow-down in area scaling between 2024 and 2027. Switching to vertical GAA structure (VGAA) will also allow CPP-grid reduction for the NAND2 standard cell from 3 to 2 thanks to the fact that the source/drain nodes of the device can be placed in different levels due to efficient layout. Similarly, SRAM height can be efficiently laid-out in 8 device grids using a vertical device architecture instead of 10 device grids using a lateral device architecture. After 2030 it is expected that 3D scaling by sequential/stacked integration (full-scale 3DVLSI) would maintain the scaling of the number of functions per unit cube.

4.4. System-On-Chip (SoC) PPA Metrics

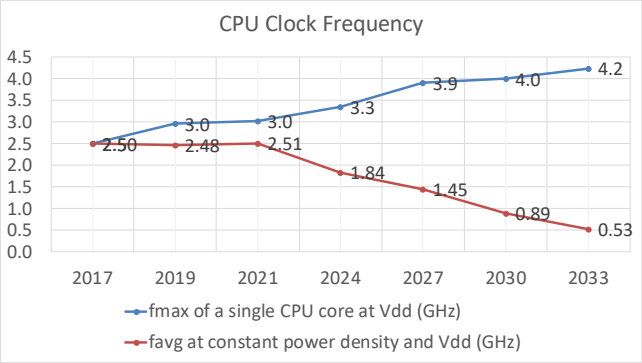
Thanks to standard cell and bitcell density improving in a node-to-node basis, it a possible to integrate more function in a given SoC footprint. The footprint for integration is assumed to be fixed at 80mm^2 across generations. Therefore, amount of memory as well as graphical processing unit (GPU) processors follow the density scaling of bitcell and standard cell, respectively, and provided the trend for more parallel architectures. On the other hand, number of central processing units (CPUs) per node is determined to reach the assumed node-to-node throughput scaling of 1.7×, which is a target provided by the Systems and Architectures IFT. In other words less improvements in the system clock frequency will mean a need for more CPUs to reach the thruput target. The number of GPUs and CPUs in 2017 are also provided by the Systems and Architectures IFT.

Thanks to device-over-device stacking as well as 3D VLSI, SoC footprint scaling factor for the same function can still be maintained at a scale of ×0.72 in average in a node-to-node basis.

Integration capacity of logic technology is shown in Table MM-11. The amount of NAND2-eq standard cell density as well as bitcell density are shown in Figure MM-3. Number of CPU and GPU cores are shown in Figure MM-4.

Table MM-11 Integration Capacity of Logic Technology

YEAR OF PRODUCTION	2017	2019	2021	2024	2027	2030	2033
	P54M36	P48M28	P42M24	P36M21	P32M14	P32M14T2	P32M14T4
Logic industry "Node Range" Labeling (nm)	"10"	"7"	"5"	"3"	"2.1"	"1.5"	"1.0"
IDM-Foundry node labeling	i10-f7	i7-f5	i5-f3	i3-f2.1	i2.1-f1.5	i1.5-f1.0	i1.0-f0.7
Logic device structure options	finFET FDSOI	finFET LGAA	LGAA finFET	LGAA VGAA	LGAA VGAA	VGAA, LGAA 3DVLSI	VGAA, LGAA 3DVLSI
Logic device mainstream device	finFET	finFET	LGAA	LGAA	LGAA	VGAA	VGAA
LOGIC TECHNOLOGY INTEGRATION CAPACITY							
Design scaling factor for standard cell	-	0.98	1.09	0.96	1.03	2.00	1.00
Design scaling factor for SRAM (111) bitcell	-	1.00	1.00	1.00	1.00	1.25	1.00
Number of stacked devices in single tier	1	1	3	4	5	1	1
Number of tiers	1	1	1	1	1	2	4
Tier utilization efficiency	0.80	0.80	0.80	0.80	0.70	0.70	0.70
SoC footprint area target - mobile (mm2)	80	80	80	80	80	80	80
NAND2 gatecount in single tier (50% digital) (Mgates)	772	1260	1764	2863	3601	3472	3472
NAND2-eq gate count in SiP (50% digital) (Mgates)	772	1260	1764	2863	3601	6944	13889
node-to-node NAND2-eq gate count scaling	-	1.63	1.40	1.62	1.26	1.93	2.00
>L2 cache bitcell in single tier (15% 111 SRAM) (MB)	43	56	74	99	130	209	209
>L2 cache SRAM bitcell in SiP (15% 111 SRAM) (MB)	43	56	74	99	130	419	837
node-to-node bitcell scaling	-	1.29	1.33	1.33	1.31	3.21	2.00
CPU thruput scaling factor (SA target) - relative	1.70	1.70	1.70	1.70	1.70	1.70	1.70
CPU thruput - relative	1.00	1.70	2.89	4.91	8.35	14.20	24.14
#GPU cores in SiP (SA target)	16	24	32	64	128	256	512
#GPU cores in SiP (integration capacity)	16	26	37	59	75	144	288
#CPU cores in SiP (SA target)	8	10	12	18	25	28	30
#CPU cores in SiP (derived from the SA thruput target, #CPUxfmax)	8	12	19	30	43	71	114
Analog + IO scaling - relative	1.00	0.85	0.72	0.61	0.52	0.44	0.38
SoC footprint scaling (50% digital, 35% SRAM, 15% analog+IO) - relative	1.00	0.71	0.53	0.38	0.30	0.16	0.10
node-to-node SoC footprint scaling	-	0.71	0.75	0.72	0.80	0.52	0.65



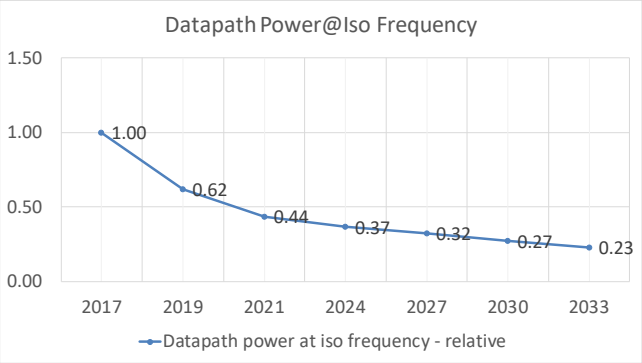


Figure MM-3 NAND2-eq standard cell count (left) and 111-bitcell (right) scaling in an 80mm^2 die

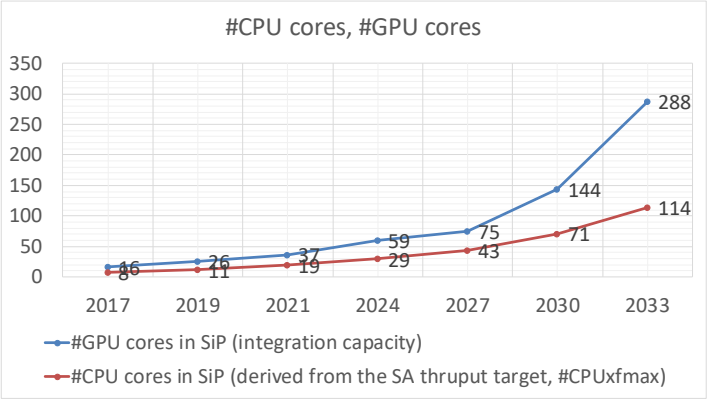


Figure MM-4 Number of CPU and GPU core in an 80mm^2 die

Projected power and performance scaling of SoC is given in Table MM-12. Frequency saturates around 4.2GHz (Figure MM-5) at the end of roadmap timeframe because of increasing parasitics and limited gate drive (Vgs-Vt) as function of scaling. Thermal (increasing power density) constraints reduces the average frequency down to 0.53GHz at the end of roadmap. 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. Power reduction slows down because of slow-down in supply voltage (Vdd) and capacitance towards the end of roadmap (Figure MM-5). Potential solutions of thermal challenges raise an opportunity to maintain an overall computational throughput scaling of ×24 over 7 node generations until 2033 instead of ×3 if the system is fully thermal constrained (Figure MM-6). This view on power-constrained CPU throughput scaling has also been discussed by the ITRS System Drivers Technology Workgroup[27].

Table MM-12 Power and Performance Scaling of SoC

YEAR OF PRODUCTION	2017	2019	2021	2024	2027	2030	2033
	P54M36	P48M28	P42M24	P36M21	P32M14	P32M14T2	P32M14T4
Logic industry "Node Range" Labeling (nm)	"10"	"7"	"5"	"3"	"2.1"	"1.5"	"1.0"
IDM-Foundry node labeling	i10-f7	i7-f5	i5-f3	i3-f2.1	i2.1-f1.5	i1.5-f1.0	i1.0-f0.7
Logic device structure options	finFET FDSOI	finFET LGAA	LGAA finFET	LGAA VGAA	LGAA VGAA	VGAA, LGAA 3DVLSI	VGAA, LGAA 3DVLSI
Logic device mainstream device	finFET	finFET	LGAA	LGAA	LGAA	VGAA	VGAA
POWER AND PERFORMANCE SCALING FACTORS							
Vdd (V)	0.75	0.70	0.65	0.65	0.65	0.60	0.55
Physical gate length for HP Logic (nm)	20.0	18.0	16.0	14.0	12.0	12.0	12.0
Datapath speed improvement at Vdd - relative	1.00	1.19	1.21	1.34	1.56	1.60	1.70
node-to-node	-	0.19	0.02	0.11	0.17	0.03	0.06
Datapath power at iso frequency - relative	1.00	0.62	0.44	0.37	0.32	0.27	0.23
node-to-node	-	-0.38	-0.29	-0.16	-0.12	-0.16	-0.16
Datapath power at fmax - relative	1.00	1.21	0.87	0.85	0.76	0.63	0.61
Datapath power at constant power density - relative	1.00	1.01	0.72	0.47	0.28	0.14	0.08
Power density of logic path cube at fmax - relative	1.00	1.20	1.21	1.82	2.69	4.49	8.00
fmax of a single CPU core at Vdd (GHz)	2.5	3.0	3.0	3.3	3.9	4.0	4.2
favg at constant power density and Vdd (GHz)	2.50	2.48	2.51	1.84	1.45	0.89	0.53
CPU SiP throughput at fmax (TFLOPS/sec)	0.16	0.27	0.46	0.79	1.34	2.27	3.86
node-to-node	-	1.70	1.70	1.70	1.70	1.70	1.70
CPU SiP throughput at constant power density (TFLOPS/sec)	0.16	0.23	0.38	0.43	0.50	0.51	0.48
node-to-node	-	1.42	1.69	1.13	1.15	1.02	0.95

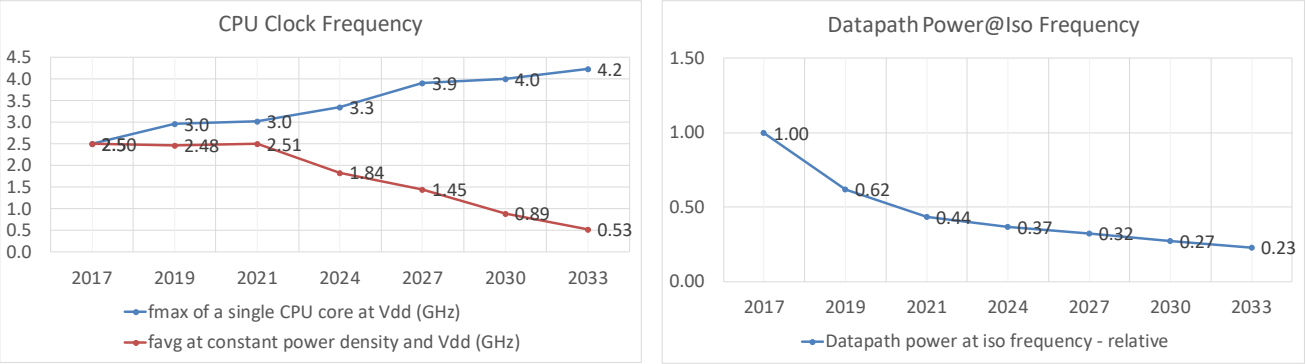


Figure MM-5 *CPU clock frequency and datapath power at the iso frequency (referenced to 2017) scaling*

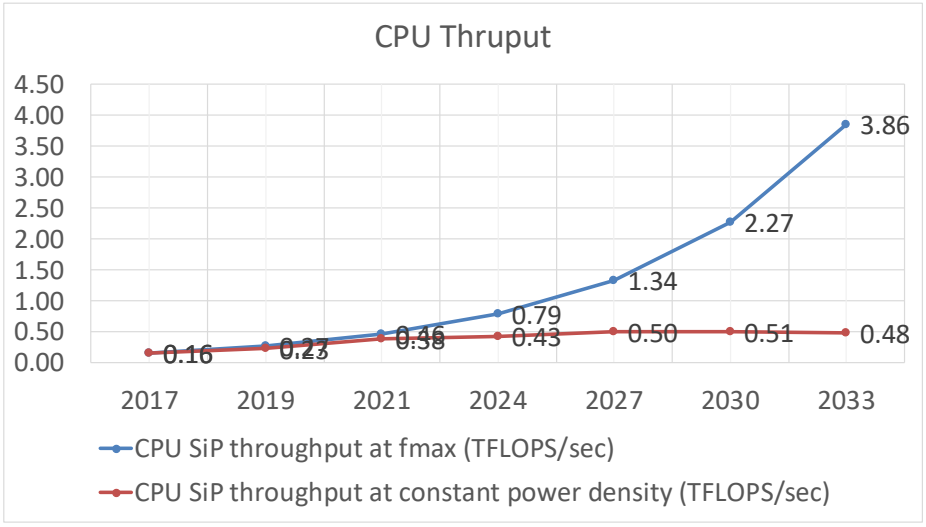


Figure MM-6 *Scaling projection of computation throughput of CPU cores at the maximum clock frequency and at thermally-constrained average frequency*

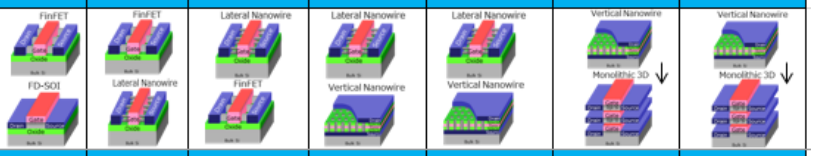
4.5. INTERCONNECT TECHNOLOGY REQUIREMENTS

The most difficult challenge for interconnects is the introduction of new materials that meet the wire conductivity requirements and reduce dielectric permittivity. 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 dominant architecture, damascene, 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 in-line 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-13 highlights and differentiates the top key challenges while Table MM-14 shows the interconnect scaling roadmap.

Table MM-13 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.
Practical Approach for 3D—Identify solutions which 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.

Table MM-14 Interconnect Roadmap for Scaling

YEAR OF PRODUCTION	2017	2019	2021	2024	2027	2030	2033
	P54M36	P48M28	P42M24	P36M21	P32M14	P32M14T2	P32M14T4
Logic industry "Node Range" Labeling (nm)	"10"	"7"	"5"	"3"	"2.1"	"1.5"	"1.0"
IDM-Foundry node labeling	i10-f7	i7-f5	i5-f3	i3-f2.1	i2.1-f1.5	i1.5-f1.0	i1.0-f0.7
Logic device structure options	finFET FDSOI	finFET LGAA	LGAA finFET	LGAA VGAA	LGAA VGAA	VGAA, LGAA 3DVLSI	VGAA, LGAA 3DVLSI
Logic device mainstream device	finFET	finFET	LGAA	LGAA	LGAA	VGAA	VGAA
DEVICE STRUCTURES							
							
INTERCONNECT TECHNOLOGY							
Conductor	Cu, non-Cu	Cu, non-Cu	Cu, non-Cu	Cu, non-Cu	Cu, non-Cu	Cu, non-Cu	Cu, non-Cu
Number of wiring layers	14	16	18	20	20	20	20
Barrier metal - tight pitch	Ta(N) Mn(N)	Ta(N) Mn(N)	Ta(N) Mn(N)	TiN (non-Cu)	TiN (non-Cu)	TiN (non-Cu)	TiN (non-Cu)
Barrier thickness - tight pitch	2.5	2.0	1.5	1.0	0.5	0.5	0.5
Inter-metal dielectrics (IMD) and k value - intermediate wire	SiCOH (2.40-2.55) Airgap	SiCOH (2.40-2.55) Airgap (1.0)	SiCOH (2.40-2.55) Airgap (1.0)	SiCOH (2.40-2.55) Airgap (1.0)	SiCOH (2.70-3.20) Airgap (1.0)	SiCOH (2.70-3.20) Airgap (1.0)	SiCOH (2.70-3.20) Airgap (1.0)
Di-electric Young Modulus							
Mx - tight-pitch interconnect resistance [Ohms/um]	130	350	500	900	2500	2500	2500
Mx - tight-pitch interconnect capacitance [aF/um]	200	200	200	200	200	200	200
Vx - tight-pitch interconnect via resistance [Ohms/via]	30	50	60	70	70	70	70
ARx -tight-pitch interconnect aspect ratio	2.0	2.0	2.0	2.0	2.0	2.0	2.0
TDDB Emax - tight-pitch interconnect (MV/cm)							
Jmax - tight-pitch interconnect (MA/cm2 at 105ºC)							
MP80 - 80nm pitch interconnect resistance [Ohms/um]	13	13	13	13	13	13	13
MP80 - 80nm pitch interconnect capacitance [aF/um]	210	210	210	210	210	210	210
VP80 - 80nm pitch interconnect via resistance [Ohms/via]	10	10	10	10	10	10	10
ARP80 - 80nm pitch interconnect aspect ratio	2.0	2.0	2.0	2.0	2.0	2.0	2.0
TDDB Emax - 80nm pitch interconnect (MV/cm)							
Jmax - 80nm pitch interconnect (MA/cm2 at 105ºC)							

4.5.1. Conductor

Copper (Cu) is expected to remain to be the preferred solution for the interconnect metal, at least until 2021. On the other hand, due to limits of electromigration, the local interconnect (medium-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 the non-Cu materials, two directions are proposed. One is the usage of the metals with less size effect e.g., silicides and the other is the introduction of materials that have different conductance mechanism e.g., carbon and collective excitations. The latter materials are still in R&D phase to implement to the semiconductor. 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 ILD 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 ($\kappa \leq 2$) 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) are advocated.

4.5.4. Reliability—Electromigration

An effective scaling model has been established assuming 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 are 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 options 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-7. 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).

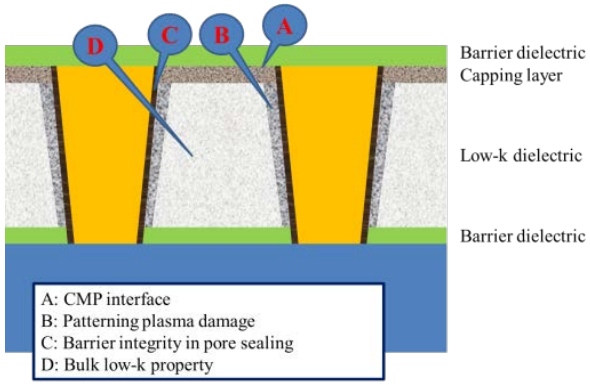


Figure MM-7 Degradation paths in low-κ damascene structure

4.6. 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 levels 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.

4.6.1. Device reliability difficult challenges

Table MM-15 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 either breakdown of the gate dielectric or threshold voltage change beyond the acceptable limits. The time to a first breakdown event is decreasing with scaling. This first event is often a “soft” breakdown. However, 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 “soft” breakdown 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. It has grown in importance as threshold voltages have been scaled down and as silicon oxy-nitride has replaced silicon dioxide as the gate insulator. Burn-in options to enhance reliability offend-products may be impacted, as it may accelerate negative bias temperature instability (NBTI) shifts. Introduction of high-κ gate dielectric may impact both the insulator failure modes (e.g., breakdown and instability) as well as the transistor failure modes such as hot carrier effects, positive, and negative bias temperature instability. The replacement of polysilicon with metal gates also impacts insulator reliability and raises new thermo-mechanical issues. The simultaneous introduction of high-κ and metal gate makes it even more difficult to determine reliability mechanisms.

As mentioned above, the move to copper and low κ has impacted front end reliability due to poorer thermal conductivity of low-κ dielectrics, leading to higher on-chip temperatures and higher localized thermal gradients.

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.

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 physics may lead to significant and important deviations from the traditional statistical distributions, making error analysis non-straightforward. Statistical analysis of other reliability data 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.

Table MM-15 Device Reliability Difficult Challenges

Near-Term 2017-2024	***Summary of issues***
Reliability due to material, process, and structural changes, and novel applications.	• TDDB, negative BTI (NBTI), positive BTI (PBTI), hot carrier injection (HCI), random telegraphic noise (RTN) in scaled and non-planar devices • Gate to contact breakdown • Increasing statistical variation of intrinsic failure mechanisms in scaled and non-planar devices • 3D device structure reliability challenges • Reduced reliability margins drive need for improved understanding of reliability at circuit level • Reliability of embedded electronics in extreme or critical environments (medical, automotive, grid...)
Long-Term 2025-2033	***Summary of issues***
Reliability of novel devices, structures, and materials.	• Understand and control the failure mechanisms associated with new materials and device structures Shift to system level reliability perspective with unreliable devices • Muon induced soft error 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.

4.6.2. 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 knobs 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.

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 a relative ease and required 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 have 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 preventing 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.

4.7. 3D HETEROGENEOUS INTEGRATION

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

- Little 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 fitting to baseline CMOS (where wafer/die-level stacking might be needed)

Die cost reduction has been enabled so far by concurrent scaling of poly pitch, metal pitch, and cell height scaling. This would like to continue until 2024. Cell height scaling would likely to be pursued by 3D device (e.g., finFET and lateral GAA), device stacking, 3DVLSI, and design-technology-co-optimization (DTCO) constructs in cell and physical design. However, this scaling route will become 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 and/or monolithic 3D (or sequential integration) These pursuits will maintain system performance and power gains while maintaining the cost advantages such as treating expensive non-scaled components somewhere else and using the best technology fit per tier functionality.

3DVLSI offers the possibility to stack devices enabling high-density contacts at the device level (up to 100 million vias per mm^2 with N14 ground rules). 3DVLSI can be routed either at gate or transistor levels. 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.[6][27]) while enabling reduced process complexity compared to a planar co-integration, for instance the stacking of III-V nFETs above SiGe pFETs[27]. 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[29][30] or highly miniaturized imagers[31].

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

- Die-to-wafer and wafer-to-wafer stacking
  - Approach: Fine-pitch di-electric/hybrid bonding and/or flip-chip assembly
  - Opportunities: Reducing bill-of-materials on the system, heterogenous integration
  - Challenges: Design/architecture partitioning
- N&P stacking
  - Approach: Sequential integration
  - Opportunities: Reducing 2D footprint of standard 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, beyond CMOS adoption
  - Approach: Sequential integration and/or wafer transfer
  - Opportunities: Complementary functions other than CMOS replacement such as neuromorphic, high-bandwidth memory. Application examples include image recognition in neuromorphic fabric and wide-IO sensor interfacing (e.g., DNA sequencing, molecular analysis).
  - Challenges: Architecting the application where low energy at low frequency and highly-parallel interfaces could be utilized, mapping applications to non-Von Neumann architectures.

4.8. 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. Expected transition from the 193i lithography to EUV in 2019 will potentially save masks because of multiple 193i masks used to print one metal and one via would now perform patterning using 1-2 EUV masks for the metal and potentially single mask for the via. However, the mask count is expected to escalate after 2027 because of more need for the metallization and repeated masks used for the front end of line (FEOL) integration in the 3D integration. This will in turn increase the process complexity, therefore the defectivity (D0) requirements. The required D0 level is expected to scale down by 2.2× in 2033 to keep the yield under control for an 80mm^2 mobile die (Figure MM-8).

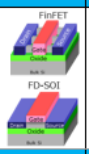
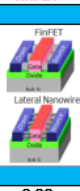
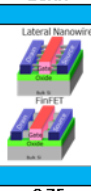
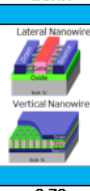
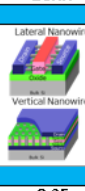
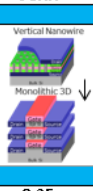
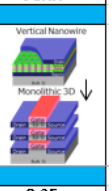
YEAR OF PRODUCTION	2017	2019	2021	2024	2027	2030	2033
	P54M36	P48M28	P42M24	P36M21	P32M14	P32M14T2	P32M14T4
Logic industry "Node Range" Labeling (nm)	"10"	"7"	"5"	"3"	"2.1"	"1.5"	"1.0"
IDM-Foundry node labeling	i10-f7	i7-f5	i5-f3	i3-f2.1	i2.1-f1.5	i1.5-f1.0	i1.0-f0.7
Logic device structure options	finFET FDSOI	finFET LGAA	LGAA finFET	LGAA VGAA	LGAA VGAA	VGAA, LGAA 3DVLSI	VGAA, LGAA 3DVLSI
Logic device mainstream device	finFET	finFET	LGAA	LGAA	LGAA	VGAA	VGAA
DEVICE STRUCTURES							
							
DEFECTIVITY TARGETS							
Critical area portion in a single tier	0.85	0.80	0.75	0.70	0.65	0.65	0.65
Maskcount target	73	69	73	67	79	124	218
Defectivity D0 target (defects/inch^2)	0.100	0.120	0.120	0.140	0.130	0.080	0.046
Process complexity exponent	20.0	18.9	20.0	18.4	21.6	34.0	59.7
Wafer sort yield (%)	81%	80%	80%	80%	80%	80%	80%

Figure MM-8 Defectivity (D0) requirements for >80% wafer sort yield target of an 80mm^2 die

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 to scale the footprint of the 1T-1C cell, to the practical limit of 4F2. The issues 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 of 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[32]-[34], buried word line/cell FET technology[35] and so on). Due to new technologies, DRAM will continue to scale with 2−3 year cycle and 20 nm half-pitch (HP), minimum feature size, DRAM will be available by 2018.

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 ($ZrO_2/Al_2O/ZrO_2$)[36] 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 κ > 50 ~ 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[37].

The other big topic is 4F2 cell migration. 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 majority. To migrate 6F2 to 4F2 cell is very challenging. For example, vertical cell transistor must be needed but still a couple of challenges are remaining.

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 become tremendously low in comparison with introducing the new technology, DRAM scaling will be stopped, and 3D cell stacking structure like as 3D-NAND 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

Non-volatile memory consists of several intersecting 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. Up to now 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, 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[38], 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[38]. This approach maked 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.

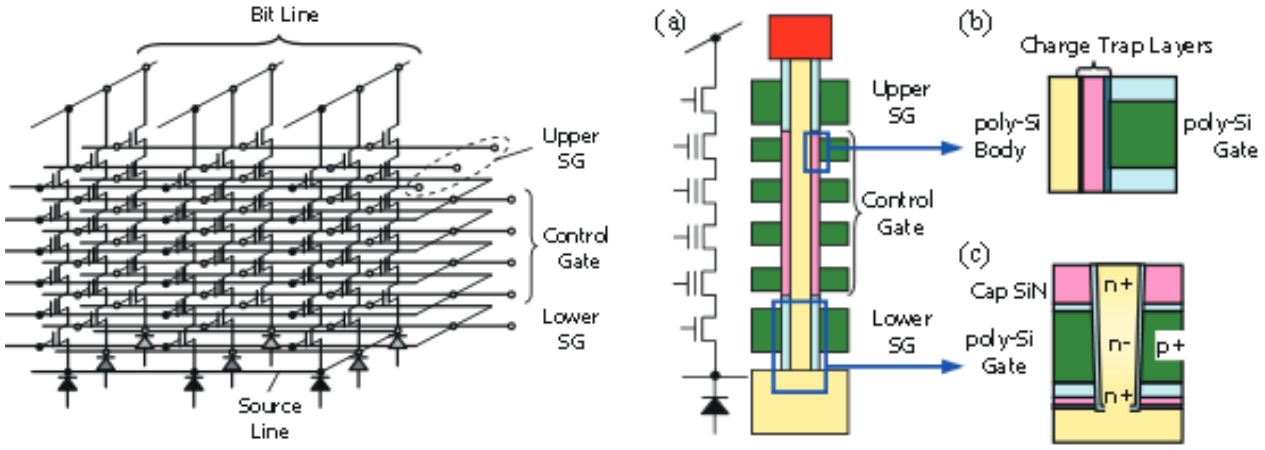


Figure MM-9 *(left) A 3D NAND array based on a vertical channel architecture. (right) BiCS (bit cost scalable) – a 3D NAND structure using a punch and plug process[38].*

A number of architectures based on the BiCS concept have been proposed since 2007 and several, including some that uses 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 x-y 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. And depositing of 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, 64 layers are beginning volume production and there is optimism that 128 layers are achievable and even 192 and 256 layers are possible.

Shrinking of 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 thus not 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. Probably new innovation or a more powerful emerging memory will be needed to further reduce bit cost.

5.3. NVM—EMERGING

Since 2D NAND Flash scaling is limited by statistical fluctuation due to too few stored charge, 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 developed and form the category of 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. The 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 and high cost (compared to Flash memories) limit 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 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 stressing 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[39], 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, for 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[40], and the reset current followed the extrapolation line from larger devices. In at least one case, cycling endurance of 1E11 was demonstrated[41]. 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 [42][43]. 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[44]. 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[45]. This is the first commercial realization of the widely published storage class memory (SCM)[46][47]. Computer systems badly needed improved 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 very low, 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.

PCRAM has the advantage of being unipolar in operation, is more product proven, and has high-cycling endurance. ReRAM, on the other hand, promises higher temperature operation and in some cases faster switching. At this time, high-density ReRAM is still in the development stage. Once developed, there seems little barrier prohibiting it from achieving 3D XP structure.

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 emerging device (beyond CMOS) roadmap chapter. Because of their promise to scale below 10nm, and operate at extremely high frequencies (< ns) and low power consumption, the focused R&D efforts in many industrial labs make this technology widely considered a potential successor to NAND (including 3D NAND). Being a resistor requiring either bi- or unipolar operation high-density ReRAM development has been limited by the lack of a good selector device, since simple diodes have limited operation ranges. 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. ReRAM trends are shown in several tabulation forms. 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[48] 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 should be expected Thus, these speculative trends are included in the potential solutions.

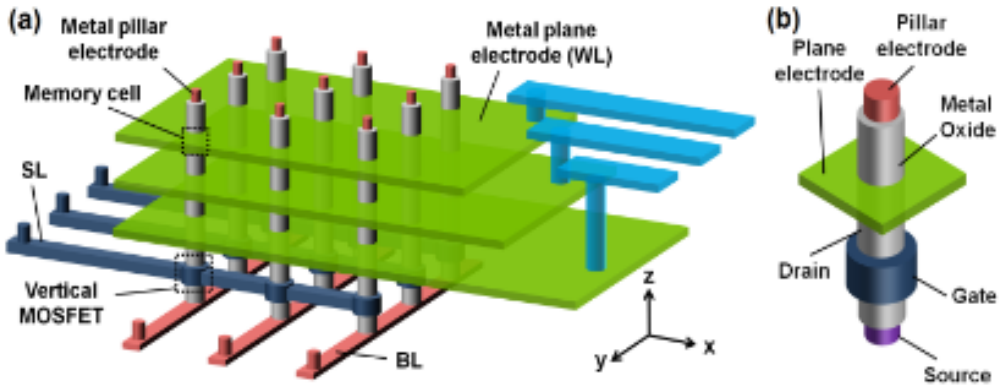


Figure MM-10 Schematic view of (a) 3D cross-point architecture using a vertical RRAM cell and (b) a vertical MOSFET transistor as the bit-line selector to enable the random access capability of individual cells in the array[48].

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 (2017-2024) potential solutions are listed in Table MM-16 while long-term (2025-2033) potential solutions are listed in Table MM-17.

Table MM-16 Potential Solutions—Near-term

Near-Term Potential Solutions: 2017-2024	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-17 Potential Solutions—Long-term

Long-Term Potential Solutions: 2025-2033	Description
Performance	• Reduce wirelength through 3D stacking • Employ NVM and beyond-CMOS devices within highly-parallel new computational schemes (e.g., neuromorphic) and heterogonous stacking
Power	• Increase parallelism while reducing Vdd through steep-SS device adoption • Fine-grain power gating
Area and Cost	• Introduce vertical-GAA • 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 5G brings new potential solutions for the analog and RF 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 versatile radios 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. 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 heterogenous 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 improve 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 we incorporated other teams' valuable inputs in our roadmap both in terms of requirement as well as technology capability limits:

- System Architecture (SA) IFT—computational datapath/fabric such as number of CPU and GPU cores per a given footprint
- 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.
- Metrology IFT—Extendibility of metrology of 3D devices such as lateral-GAA and vertical-GAA
- Outside System Connectivity (OSC) IFT—I/O and integration requirements for 5G and high-speed memory for data server
- Packaging 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 adoption is expected in 2021 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 2027. 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.
- Emerging memories is likely to become a potential alternative to SRAM-based and/or eDRAM cache applications, probably around 2021.

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 over active gate and Cobalt local interconnects," IEDM, Session 2.9, December 2017.

[5] S.-W. Wu, "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.

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

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

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

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

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

[11] 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.

[12] R. Xie, "A 7nm finFET technology featuring EUV patterning and dual strained high mobility channels", IEDM, Section 2.7, December 2016.

[13] R. Berthelon et al., "A novel dual isolation scheme for stress and back bias maximum efficiency in FDSOI technology", IEDM, Section 17.7, December 2016.

[14] R. Carter et al., "22nm FDSOI technology for emerging mobile, internet-of-things, and RF applications", IEDM, Section 2.2, December 2016.

[15] 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.

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

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

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

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] Nanohub website (http://nanohub.org) and ITRS tools on Nanohub (https://nanohub.org/tools/itrs/).

[26] MASTAR tool (http://www.itrs.net/Links/2011ITRS/MASTAR2011/) and downloading and installation instructions at: (http://www.itrs.net/Links/2011ITRS/MASTAR2011/MASTARDownload.htm).

[27] 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.

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

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

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

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

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

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

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

[47] 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.

[48] 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.