INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS

2016 EDITION

MORE MOORE WHITE PAPER

1. IRDS MORE MOORE MISSION

System scaling enabled by Moore's scaling is more and more challenged with the scarcity of resources such as power and interconnect bandwidth. Particularly due to the emergence of cloud, seamless interaction of big-data and instant data have become a necessity (Figure 1). Instant data generation require ultra-low-power device with "always-on" feature at the same time with high-performance device that can generate the data instantly. Big data require abundant computing and memory resources to generate the service and information that clients need.

More Moore focus team in the International Roadmap of Devices and Systems (IRDS) provides physical, electrical and reliability requirements for logic and memory technologies to sustain More Moore (PPAC: power, performance, area, cost) scaling for big data, mobility, and cloud (IoT and server) applications and forecast logic and memory technologies (15 years) in main-stream/high-volume manufacturing (HVM).

Figure 1: Big data and Instant Data.

Following applications drive requirements of More Moore technologies that are addressed in IRDS [1][2]:

- 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 & computing (Internet-of-Things: IoT) – targeting reduced leakage & variability

In this white paper we will discuss device, interconnect, memory roadmap; and the upcoming inflection points, roadblocks, and potential solutions as it will be addressed in the 2016 edition of IRDS roadmap. Below are some take-aways from the projected IRDS More Moore roadmap:

- Ground rule scaling slows down and saturates around 2024. EUV (extrame-ultraviolet) technology now started to slow down this saturation trend by having the cost under control thanks to process complexity reduction. Transition to 3D integration and use of beyond CMOS devices for complementary System-on-Chip (SoC) functions are projected after 2024.
- Ground-rule scaling need to also enable design-technology-co-optimization (DTCO) constructs that accommodate the area reduction as well as tighten the critical design that limits for overall SoC area scaling.
- 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 I/O. Parasitics improvement will become the major knob for performance improvement for nodes spanning between 2017 and 2024.
- SiGe and Ge channels are gaining importance as the high-mobility channels. IIIV channel faces challenges of variability, band-to-band tunneling, and large investments in fab infrastructure.
- Interconnect technology sees the use of non-Cu options, particularly in addressing the electromigration risks of Cu. On the other hand, metal resistance exponentially increases in both Cu and non-Cu options, which makes a

careful selection of BEOL stack for SoC not to face performance loss due to the high resistance of tight-pitch layers.

- Performance scaling across 7 nodes spanning from 2015 to 2030 is 22% node-to-node improvement for datapaths without wireload while it becomes 9% node-to-node improvement for datapaths loaded with tight pitch metal routing. If wireload routing is done with intermediate metal (at 80nm pitch), the node-to-node performance improvement becomes 16%, which seems to be relatively on track.
- Energy per switching reduction for logic devices is on track, about 36% reduction in a node-to-node basis in average.
- DRAM needs to maintain sufficient storage capacitance and adequate cell transistor performance are required to keep the retention time characteristic in the future. 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 new DRAM concept will 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. 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.
- 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, HfO2 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, has been proposed. If developed to maturity, this may serve as a low power and very fast Flash-like memory.
- 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 which 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 dominate since it is still substantially more cost effective and well established for being able to endure the PCB board soldering process (at ~ 250C) without losing its preloaded code.
- 3D cross point memory has been demonstrated for the storage class memory (SCM) to improve I/O throughput and reduce power and cost. Since the memory including the selector device is completely fabricated in the BEOL process it is relatively inexpensive to stack multiple layers to reduce bit cost.
- 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 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.

2. LOGIC TECHNOLOGIES

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

More Moore platform targets at bringing Power-Performance-Area-Cost (PPAC) value for node scaling (every 2-3 years) [3][4]:

- (P)erformance: >30% more maximum operating frequency at constant energy
- (P)ower: >50% less energy per switching at a given performance
- (A)rea: >50% area reduction
- (C)ost: <30% wafer cost – 30-35% less die cost for scaled die.

Those scaling targets have driven the industry toward a number of major technological innovations, including material and process changes such as higher-K gate dielectrics and strain enhancement, and in the near future, new structures such as gate-all-around (nanowire), alternate high-mobility channel materials, and new 3D integration schemes allowing heterogonous 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 notice that cost metric (30-35% less die cost) and market cadence necessitating new product every year are becoming more important targets with the mobile industry. As the applications strictly requiring all figure-of-merits (FoMs) concurrently met, it is necessary to come along effective list of knobs for sustaining certain device architecture to its 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. Particularly this gets more difficult whenever the cost of wafer processing is getting 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 35-40% for the same of number of transistors, which can only be enabled by aggressive pitch scaling due to new advancements in channel material, device architecture, contact engineering, and device isolation. However, increased process complexity must also be taken into account for the overall die yield.

1.1 GROUND RULES SCALING

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 end up with 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 need to also enable design-technology-co-optimization (DTCO) constructs that accommodate the area reduction as well as tighten the critical design that limits for 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. Projected roadmap of ground rules as well as device architectures are shown in Table 1. There is not yet a consensus on the node naming across different foundries and IDMs; however, the projected rules gives 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 1: Device architecture and ground rules roadmap for logic device technologies. PxxMxx notation refers to Pxx: contacted poly pitch and Mxx: metalx pitch in nm. This shows the technology capability. On top of pitch scaling there are other elements such as cell height, vertical integration, fin depopulation, DTCO contructs, etc define the target area scaling (gates/mm2).

YEAR OF PRODUCTION	2015	2017	2019	2021	2024	2027	2030
Logic device technology naming	P70M56	P54M36	P42M24	P32M20	P24M12G1	P24M12G2	P24M12G3
Logic industry "Node Range" Labeling (nm)	"16/14"	"11/10"	"8/7"	"6/5"	"4/3"	"3/2.5"	"2/1.5"
Logic device structure options	finFET FDSOI	finFET FDSOI	finFET LGAA	finFET LGAA VGAA	VGAA, M3D	VGAA, M3D	VGAA, M3D
	FinFET / FDSOI	FinFET / FDSOI	Lateral Nanowire / FinFET	Vertical Nanowire / Lateral Nanowire	Vertical Nanowire / Monolithic 3D	Vertical Nanowire / Monolithic 3D	Vertical Nanowire / Monolithic 3D
LOGIC DEVICE GROUND RULES							
MPU/SoC Metalx ½ Pitch (nm)[1,2]	28.0	18.0	12.0	10.0	6.0	6.0	6.0
MPU/SoC Metal0/1 ½ Pitch (nm)	28.0	18.0	12.0	10.0	6.0	6.0	6.0
Contacted poly half pitch (nm)	35.0	24.0	21.0	16.0	12.0	12.0	12.0
L_g: Physical Gate Length for HP Logic (nm) [3]	24	18	14	10	10	10	10
L_g: Physical Gate Length for LP Logic (nm)	26	20	16	12	12	12	12
Channel overlap ratio - two-sided	0.80	0.80	0.80	0.80	0.80	0.80	0.80
Spacer width (nm)	12	8	6	5	4	4	4
Contact CD (nm) - finFET, LGAA	22	14	16	12	11	11	11
Device architecture key ground rules							
FinFET Fin Half-pitch (new) =0.75 or 1.0 M0/M1 (nm)	21.0	18.0	12.0				
FinFET Fin Width (nm)	8.0	6.0	6.0				
FinFET Fin Height (nm)	42.0	42.0	42.0				
Footprint drive efficiency - finFET	2.19	2.50	3.75				
Lateral GAA Lateral Half-pitch (nm)			12.0	10.0			
Lateral GAA Vertical Half-pitch (nm)			12.0	9.0			
Lateral GAA Diameter (nm)			6.0	6.0			
Footprint drive efficiency - lateral GAA, 3x NWs stacked			2.4	2.8			
Vertical GAA Lateral Half-pitch (nm)				10.0	6.0	6.0	6.0
Vertical GAA Diameter (nm)				6.0	5.0	5.0	5.0
Footprint drive efficiency - vertical GAA, 3x NWs stacked				2.8	3.9	3.9	3.9
Defice effective width - [nm]	92.0	90.0	56.5	56.5	56.5	56.5	56.5
Device lateral half pitch (nm)	21.0	18.0	12.0	10.0	6.0	6.0	6.0
Device width or diameter (nm)	8.0	6.0	6.0	6.0	5.0	5.0	5.0

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, M3D: Monolithic-3D.

As it can be noted from the roadmap, after 2024 there is no headroom for 2D geometry scaling where 3D VLSI integration 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. It is projected that physical channel length would saturate around 12nm due to worsening electrostatics while CPP would saturate at 24nm to reserve sufficient CD (~11nm) 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. Challenge of such integration in 3D is how to partition the system to come up with better

utilization of devices, interconnect and sub-systems such as memory, analog, and I/O. That's why the functional scaling is required after 2024. This would potentiality 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.

1.2 PERFORMANCE BOOSTERS

In the early years before 130nm node, transistors enjoyed Dennard scaling where oxide thickness (EOT), transistor 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-k 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.

FinFET still remains the key device architecture that could sustain scaling until 2021 for high-performance logic applications [5]. In electrostatics and fin depopulation (increasing fin height while reducing number of fins at unit footprint area) remain as the major knob 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 gate-all-around (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.

A roadmap overview of device architecture, key modules, and performance boosters is shown in Table 2.

While device architecture seeing changes subsequent modules should also evolve. Those could for instance be: 1) starting substrates such as Si to SOI and 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:

Transition to new device architectures: As mentioned earlier finFET will likely to sustain until the end of 2023. Beyond 2019 a transition to gate-all-around (GAA) will start and potentially a transition to vertical nanowires devices will be needed when there will be no room left for the gate length scale down due to the limits of fin width scaling (saturating the Lgate scaling to sustain the electrostatics control) and contact width.

Starting Substrate: Bulk silicon will still remain the mainstream substrate while silicon-on-insulator (SOI) and strain-relaxation-buffer (SRB) would 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 Vt tuning, allowing to tune a device to either high-performance or low-leakage, thanks to the backgate control.

High-mobility channels: High-mobility materials such as Ge and IIIV bring promise in increasing drive current by means of an order of magnitude increase in intrinsic mobility (Figure 2). 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, which is the so-called 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 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, 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 [8]. 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 IIIV high-mobility channels. Despite the fact that drive current of IIIV 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 IIIV, 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.

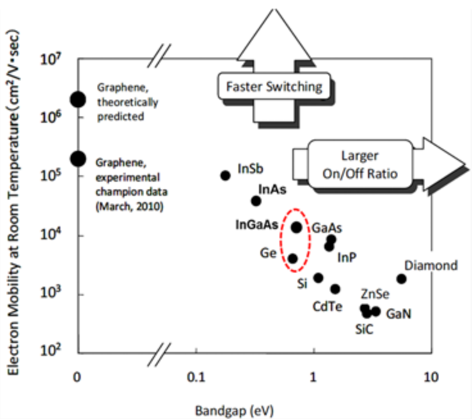
Table 2: Device roadmap enabling More Moore scaling: 1) Device architecture, 2) Device module characteristics, 3) Performance boosters.

YEAR OF PRODUCTION	2015	2017	2019	2021	2024	2027	2030
Logic device technology naming	P70M56	P54M36	P42M24	P32M20	P24M12G1	P24M12G2	P24M12G3
Logic industry "Node Range" Labeling (nm)	"16/14"	"11/10"	"8/7"	"6/5"	"4/3"	"3/2.5"	"2/1.5"
Logic device structure options	finFET FDSOI	finFET FDSOI	finFET LGAA	finFET LGAA VGAA	VGAA, M3D	VGAA, M3D	VGAA, M3D
	FinFET / FDSOI	FinFET / FDSOI	Lateral Nanowire / FinFET	Vertical Nanowire / Lateral Nanowire	Vertical Nanowire / Monolithic 3D	Vertical Nanowire / Monolithic 3D	Vertical Nanowire / Monolithic 3D
DEVICE ARCHITECTURE & MODULES							
Starting substrate	Si, SOI	Si, SOI	Si,SOI, SRB, QW	Si,SOI, SRB, QW	Si,SOI, SRB, QW	Si,SOI, SRB, QW	Si,SOI, SRB, QW
N-channel	Si	sSi	sSi, Ge	sSi, sGe, IIIV	sSi, sGe, IIIV	sSi, sGe, IIIV	sSi, sGe, IIIV
P-channel	Si	Si,SiGe	Si,SiGe	Si,SiGe	Ge	Ge	Ge
Channel formation	Etch	Etch, EPI	Etch, EPI	Etch, EPI	Etch, EPI	Etch, EPI	Etch, EPI
Contact material	Silicide	Low-SBH	Low-SBH	Low-SBH	Low-SBH	Low-SBH	Low-SBH
Contact integration	EPI	EPI	EPI WAC	WAC	WAC	WAC	WAC
DEVICE PERFORMANCE BOOSTERS							
Main performance booster	SCE finHeight Vt	SCE finHeight Vt	Parasitics finHeight	Parasitics finHeight	Low Vdd 3D	Low Vdd 3D	Low Vdd 3D
Scaling focus	Perf	Power	Power	Power	Function	Function	Function
Channel strain	Yes	Yes	Yes	Yes	Yes	Yes	Yes
S/D strain	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Transport scheme	DD	Quasi Ballistic	Quasi Ballistic	Ballistic	Ballistic TFET, JFET, NCMOS	Ballistic TFET, JFET, NCMOS	Ballistic TFET, JFET, NCMOS, Spin

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, M3D: Monolithic-3D, SRB: Strain-Relaxation-Buffer, QW: Quantum well, SBH: Schottky Barrier Height, WAC: Wrap-around-contact, DD: Drift-diffusion, TFET: Tunneling FET, JFET: Junctionless FET, NCMOS: Negative-capacitance MOSFET.

Strain engineering: This knob has been used as one of the most effective knobs in the last decade (Figure 3) [9]. However, effective of those stressors may not extrapolate intuitively into newer nodes as illustrated for the 32nm node and earlier (Figure 3). With the scaling down of contacted poly pitch, SiGe on the 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 [10]. SiGe channel for PMOS and strained Si channel for NMOS has been successfully demonstrated on an N7 CMOS platform using SRB[11]. 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 UTBB FDSOI in order to boost pFET performance [12][13]. 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.

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% and more from that of ideal case. This proportion will likely become harder to maintain or worse with the poly pitch scaling and also increasing interconnect resistance by scaling, will all leaving less headroom for the device contact itself. In order to maximize the benefits of high-mobility channels in the drain current, it gets 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 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) [14][15]. This SBH reduction happens by the exponential decay of the metal induced gap states (MIGS) induced charge density in the bandgap of the dielectric.



	Si	Ge	GaAs	$In_{0.53}Ga_{0.47}As$	InAs
Eg (eV)	1.1	0.66	1.4	0.75	0.35
μ_n (cm^2/v-s)	1,350	3,900	4,600	>8,000	40,000
μ_p (cm^2/v-s)	480	1,900	500	350	<500
m^*/m_o	0.165	0.12	0.067	0.041	0.024

Figure 2: Intrinsic mobility of different materials.

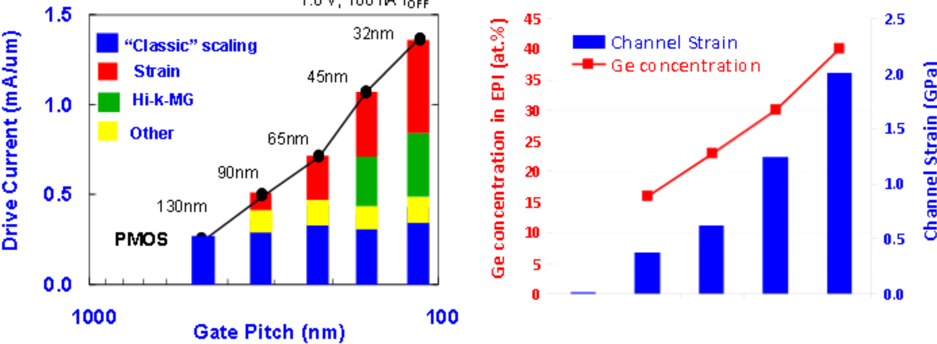


Figure 3: Impact of strain engineering on device performance [9].

Reducing parasitic device capacitance: Parasitic capacitance between gate and source/drain terminal of the device is increasing with technology scaling and exceed the channel capacitance as the poly pitch is scaled down. 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-k spacer materials and even air spacer that still provide good reliability and etch selectivity for S/D contact formation [16][17]. 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.

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 [16][18]. This increased drive at unit footprint by scaling the fin pitch comes at a trade-off between fringing capacitance between gate and contact and series resistance.

Improving electrostatics: FinFET has better electrostatics integrity due to its tall narrow channel that is controlled by a gate from three-sides where this 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 BOX thickness where convergent scaling of both silicon thickness and BOX thickness enables electrostatics scaling (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 [19][20].

Improving device isolation: Besides the channel leakage induced by electrostatics, there are potentially other leakage sources such as sub-fin leakage. 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 will potentially solve this leakage problem; therefore improving the electrostatics [21].

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 instance 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 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.

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 type of logic devices and maybe even new circuit architecture 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 TFET for ultra-low power applications and memristors for neuromorphic applications.

Projected roadmap for the electrical specifications of logic core device is listed in Table 3.

Table 3:Projected electrical specifications of logic core device.

YEAR OF PRODUCTION	2015	2017	2019	2021	2024	2027	2030
Logic device technology naming	P70M56	P54M36	P42M24	P32M20	P24M12G1	P24M12G2	P24M12G3
DEVICE PHYSICAL&ELECTRICAL SPECS							
Power Supply Voltage - Vdd (V)	0.80	0.75	0.70	0.65	0.55	0.45	0.40
Subthreshold slope - [mV/dec]	75	70	68	65	40	25	25
Inversion layer thickness - [nm] [4]	1.10	1.00	0.90	0.85	0.80	0.80	0.80
Vt,sat (mV) at Ioff=100nA/um - HP Logic [5][6]	129	129	133	136	84	52	52
Vt,sat (mV) at Ioff=100pA/um - LP Logic [5][6]	351	336	333	326	201	125	125
Effective mobility (cm2/V.s)	200	150	120	100	100	100	100
Rext (Ohms.um) - HP Logic [7]	280	238	202	172	146	124	106
Ballisticity.Injection velocity (cm/s)	1.20E-07	1.32E-07	1.45E-07	1.60E-07	1.76E-07	1.93E-07	2.13E-07
Vdsat (V) - HP Logic	0.115	0.127	0.136	0.128	0.141	0.155	0.170
Vdsat (V) - LP Logic	0.125	0.141	0.155	0.153	0.169	0.186	0.204
Ion (uA/um) at Ioff=100nA/um - HP logic w/ Rext=0 [8]	2311	2541	2782	2917	3001	2670	2408
Ion (uA/um) at Ioff=100nA/um - HP logic, after Rext [9]	1177	1287	1397	1476	1546	1456	1391
Ion (uA/um) at Ioff=100pA/um - LP logic w/ Rext=0 [8]	1455	1567	1614	1603	2008	1933	1582
Ion (uA/um) at Ioff=100pA/um - LP logic, after Rext [9]	596	637	637	629	890	956	821
Cch,total (fF/um2) - HP/LP Logic [9]	31.38	34.52	38.35	40.61	43.14	43.14	43.14
Cgate,total (fF/um) - HP Logic [10]	1.81	1.49	1.29	0.97	1.04	1.04	1.04
Cgate,total (fF/um) - LP Logic [10]	1.96	1.66	1.47	1.17	1.24	1.24	1.24
CV/I (ps) - FO3 load, HP Logic [11]	3.69	2.61	1.94	1.29	1.11	0.96	0.89
I/(CV) (1/ps) - FO3 load, HP Logic [12]	0.27	0.38	0.52	0.78	0.90	1.04	1.12
Energy per switching [CV2] (fj/switching) - FO3 load, HP Logic	3.47	2.52	1.89	1.24	0.94	0.63	0.50

1.3 PERFORMANCE-POWER-AREA (PPA) SCALING

An important speed metric for the transistor is the intrinsic speed *I/CV* 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 [22]. 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 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 is explained in [3]. For circuit-level transient simulations there is a need for compact-model based software such as BSIM CMG or open source models such as Virtual Source Model (VSM) from MIT [23]. We used VSM models to capture the circuit-level parameters such as delay and power per stage from a ring oscillator. Its inputs were transparently validated in TCAD with the support from the NanoHub Team of Purdue University [24]. There are also analytical modeling tools such as MASTAR [25], 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 [3]:

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

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 [3]. 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 (performance, power, and area) metrics is shown in Table 4.

Table 4: Projected power-performance-area (PPA) metrics of functional datapath.

YEAR OF PRODUCTION	2015	2017	2019	2021	2024	2027	2030
Logic device technology naming	P70M56	P54M36	P42M24	P32M20	P24M12G1	P24M12G2	P24M12G3
LOGIC CELL AND FUNCTIONAL FABRIC TARGETS							
SRAM height in fins	10	10	10	10	8	8	8
SRAM 111 bitcell area density - Mbits/mm2	17	29	50	78	217	217	217
NAND2 active fins per pull-up and pull-down	4	3	3	2	5	5	5
NAND2 width at CPP multiples	3	3	3	3	2	2	2
NAND2 equivalent raw-gate density - Mgates/mm2	9	19	33	78	87	87	87
Effective width of ND2 cell (nm) - datapath cell	368	270	170	113	79	79	79
Cell drive at saturation (Ohms)	3695	4315	5909	7788	9059	7872	7325
Cell related loading capacitance at saturation (fF) - FO3	3.99	2.42	1.31	0.66	0.49	0.49	0.49
Wireload resistance (Ohms) - 100xCPP tight pitch, 4 vias	320	684	2180	3020	7340	7340	7340
Wireload capacitance (fF) - 100xCPP, tight pitch	1.40	0.91	0.76	0.58	0.43	0.43	0.43
Wireload resistance (Ohms) - 100xCPP, 80nm pitch, 6 vias	151	132	135	152	141	141	141
Wireload capacitance (fF) - 100xCPP, 80nm pitch	1.47	1.01	0.88	0.67	0.50	0.50	0.50
FO3 wireloaded stage delay (ps) - no wireload	10.18	7.19	5.35	3.55	3.05	2.65	2.47
FO3 wireloaded stage delay (ps) - tight pitch wireload	14.80	11.28	11.03	8.69	9.43	8.67	8.33
FO3 wireloaded stage delay (ps) - 80nm pitch wireload	14.42	10.46	9.11	7.27	6.28	5.46	5.09
FO3 wireloaded stage dynamic power at 1GHz clock, 80nm pitch wireload (mW)	3.49	1.93	1.07	0.56	0.30	0.20	0.16
FO3 wireloaded stage IDDQ at 1GHz clock - nW	117.76	81.00	47.50	29.41	17.28	14.14	12.57

Performance scaling across 7 nodes spanning from 2015 to 2030 is 22% node-to-node improvement for datapaths without wireload while it becomes 9% node-to-node improvement for datapaths loaded with tight pitch metal routing. If wireload routing is done with intermediate metal (at 80nm pitch), the node-to-node performance improvement becomes 16%. 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. Energy per switching reduction is on track, about 36% reduction in a node-to-node basis in average. This is achieved thanks to fin depopulation, which also enabled the cell height reduction. Raw gate density also improved by around x2 in a node-to-node basis until 2024. After 2024 it is expected that 3D scaling by sequential/stacked integration would maintain the scaling of the number of functions per unit cube. Current roadmap edition assumes that the raw gate density is a metric to represent the scaling of number of functions.

3. 3D HETEROGENOUS CO-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, M3D, 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 stacking and monolithic 3D (or sequential integration) to 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.

M3D 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). M3D 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) enables the independent optimization of both types of transistors (customized implementation of channel material / substrate orientation / channel and raised source/drain strain, etc. [6][26]) 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 M3D because their process temperatures are intrinsically low. M3D, 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 [27][28] or highly miniaturized imagers [29].

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

- Generation-1 (2 tiers): N&P stacking, 2-tier
  - Approach: Sequential integration
  - Opportunities: Reducing 2D footprint of standard cell
  - Challenges: Minimizing interconnect overhead is key between N&P enabling low-cost
- Generation-2 (3 tiers): 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
- Generation-3 (4 tiers): adding Analog and IO
  - 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
- Generation-4 (>4 tiers): clustered functional stacks, beyond CMOS adoption

- Approach: Sequential integration and/or wafer transfer
- Opportunities: Complementary functions other than CMOS replacement – neuromorphic, high-bandwidth memory, application examples - 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. INTERCONNECT SCALING

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 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 5 highlights and differentiates the top key challenges while Table 6 shows the interconnect scaling roadmap.

Table 5: 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.

Conductor: Cu will be the preferred solution. On the other hand, due to limits of electromigration the local interconnect (MOL), M1, and Mx levels will embed non-Cu solutions such as 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.

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 PVD is limited, other nitrides such as Mn(N) which can be deposited by CVD or ALD have recently attracted attention. As for the emerging materials, SAM (Self-Assembled Monolayers) are researched as the candidates for future generation.

IMD (Inter-metal Dielectrics): Reduction of the ILD k 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-k ($k \leq 2$) materials become clearer, and air-gap technologies is the alternative path to lower the inter-layer capacitance. As the emerging materials, MOF (Metal Organic Framework) and COF (Carbon Organic Framework) are advocated.

Reliability-EM (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 as shown in Figure 4. The model predicts that life time 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. J_{max} (The maximum equivalent dc current density) and J_{EM} (The maximum current density) limited by the interconnect geometry scaling is shown in Figure 5. J_{max} 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.

Table 6: Interconnect roadmap for scaling.

YEAR OF PRODUCTION	2015	2017	2019	2021	2024	2027	2030
Logic device technology naming	P70M56	P54M36	P42M24	P32M20	P24M12G1	P24M12G2	P24M12G3
Logic industry "Node Range" Labeling (nm)	"16/14"	"11/10"	"8/7"	"6/5"	"4/3"	"3/2.5"	"2/1.5"
INTERCONNECT TECHNOLOGY							
Conductor	Cu	Cu	Cu	Cu Silicides Carbon Collective Excitations	Cu Silicides Carbon Collective Excitations	Cu Silicides Carbon Collective Excitations	Cu Silicides Carbon Collective Excitations
Number of wiring layers	13	14	15	16	18	22	30
Barrier metal - intermediate wire (tight pitch)	Ta(N)	Ta(N), Mn(N)	Ta(N), Mn(N)	Ta(N),Mn(N), SAM	Ta(N), Mn(N),SAM	Ta(N), Mn(N),SAM	Ta(N), Mn(N),SAM
Barrier thickness - intermediate wire							
Inter-metal dielectrics (IMD) and k value - intermediate wire	SiCOH (2.55)	SiCOH (2.40-2.55) Airgap (1.0)	SiCOH (2.20-2.55) Airgap (1.0)	SiCOH (2.20-2.55) Airgap (1.0) MOF, COF	SiCOH (2.00-2.55) Airgap (1.0) MOF, COF	SiCOH (2.00-2.55) Airgap (1.0) MOF, COF	SiCOH (2.00-2.55) Airgap (1.0) MOF, COF
Di-electric Young Modulus							
Mx - tight-pitch interconnect resistance [Ohms/um]	40	130	500	900	3000	3000	3000
Mx - tight-pitch interconnect capacitance [aF/um]	200	190	180	180	180	180	180
Vx - tight-pitch interconnect via resistance [Ohms/via]	10	15	20	35	35	35	35
ARx -tight-pitch interconnect aspect ratio	2.0	2.0	2.0	2.0	2.0	2.0	2.0
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

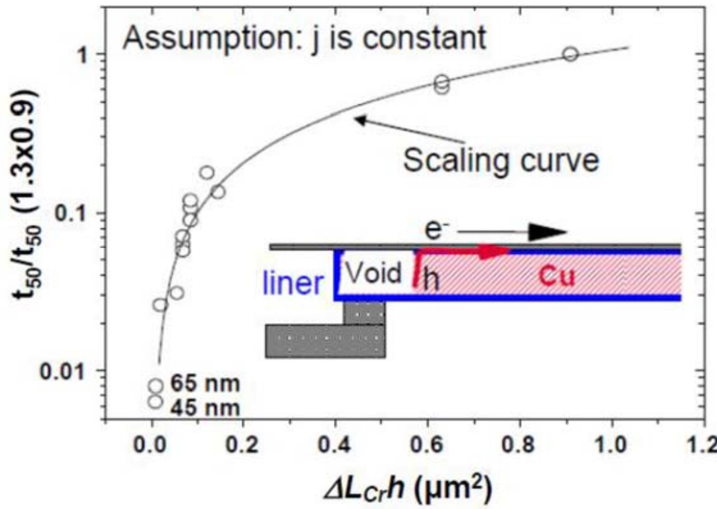


Figure 4: Experiment and model of lifetime scaling versus interconnect geometry.

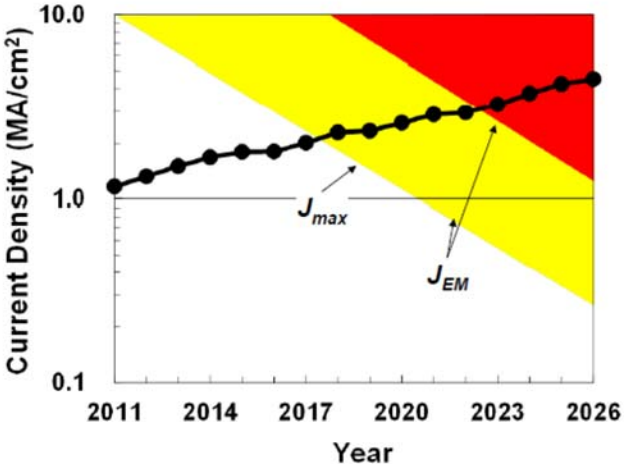


Figure 5: Evolution of Jmax (from device performance) and J_{EM} (from targeted lifetime).

Reliability - TDDB (Time Dependent Dielectric Breakdown): Basically, the dielectric reliability can be categorized according to the failure paths and mechanisms as shown in Figure 6. 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 which directly impacts the estimate of Vmax (or min spacing).

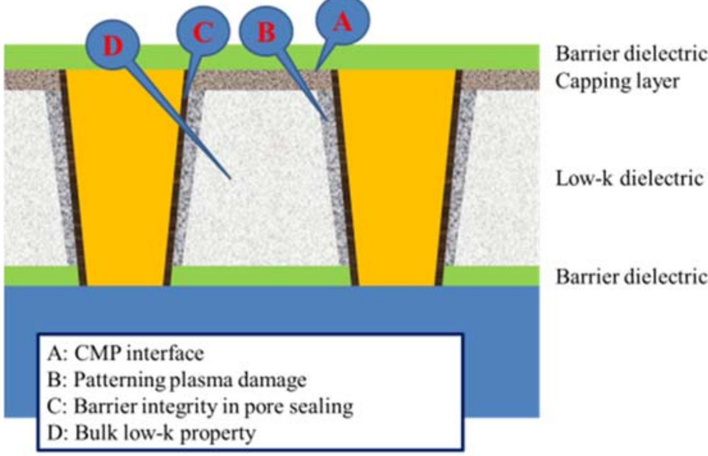


Figure 6: Degradation paths in low-k damascene structure.

5. 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.

1.4 DRAM

For DRAM, the main goal is to continue to scale the foot-print of the 1T-1C cell, to the practical limit of $4F^2$. The issues are vertical transistor structures, high-κ dielectrics to improve the capacitance density, and meanwhile 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 [30]-[32], buried word line/cell FET technology [33] and so on). Due to new technologies, DRAM will continue to scale with 2-3 year cycle and 20 nm HP (minimum feature size) DRAM will be available by 2017.

Since the DRAM storage capacitor gets physically smaller with scaling, the 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 MIM (metal-insulator-metal) capacitors have been adopted using high κ ($ZrO2/Al_2O/ZrO2$) [34]as the capacitor of 40-30's nm half-pitch DRAM. And this material evolution and improvement are continued until 20 nm HP and ultra high-K (perovskite $\kappa > 50 \sim 100$) material will be released in 2016. 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 which are typically constructed after the CMOS devices are formed, and therefore are limited to low-temperature processing. DRAM peripheral device requirement can relax I_{off} but demands more I_{on} of LSTP device. But, in the future, high-κ metal gate will be needed for sustaining the performance [35].

The other big topic is $4F^2$ cell migration. As the half-pitch scaling become 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 $6F^2$ (a = 6) is the majority. To migrate $6F^2$ to $4F^2$ 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 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.

1.5 NVM - FLASH

Non-volatile memory (NVM) 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.

However, the economy of stacking by completing one device layer then another and so forth is questionable. As depicted in Figure 7 [36], 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 is proposed to fabricate the bit line string vertically to simplify the processing steps dramatically [36]. This approach makes 3D stacked devices in a few steps and not through repetitive processing, thus promises a new low cost scaling path to NAND flash. Figure 7 illustrates one such approach. Originally coined BiCS, or Bit Cost Scalable, 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 7, 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.

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, but 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’s 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.

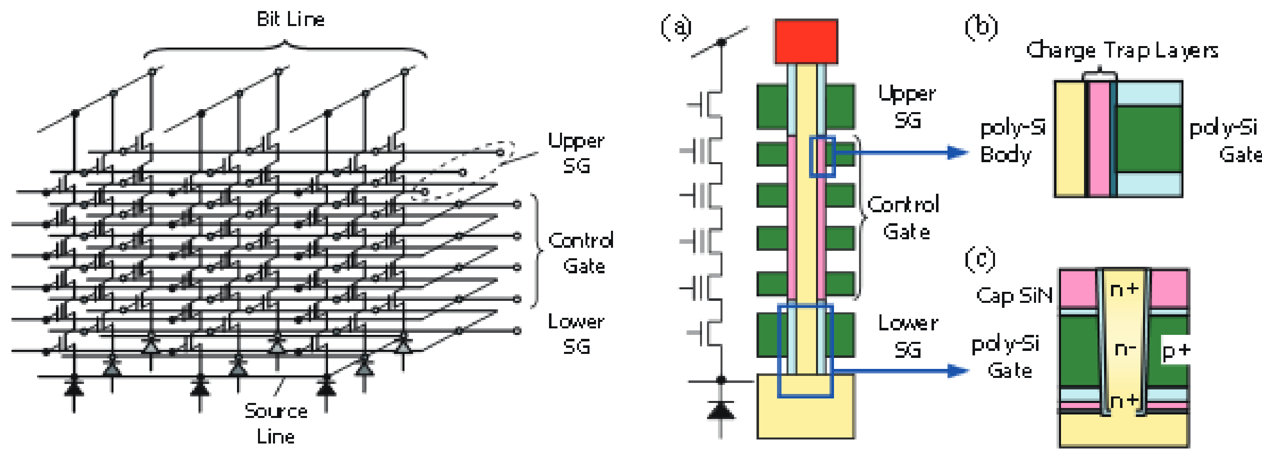


Figure 7: (left) A 3D NAND array based on a vertical channel architecture. (right) BiCS (Bit Cost Scalable) – a 3D NAND structure using a punch and plug process [36].

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 10um, which is still in the same range as 10-15 metal layers for logic IC’s. This kind of layers thickness does not

significantly affect bare die thickness (thinnest about 40um 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 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.

1.6 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 which is no longer subject to the drawback of too few electrons some characteristics of non-charge based emerging memories, such as low voltage operation, random access, are attractive for various applications 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.

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, HfO2 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.

MRAM - 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 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 which 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 [37], 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 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 ~ 250C) without losing its preloaded code, for which many emerging memories have not been able to demonstrate yet.

PCRAM and Cross Point 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, for 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, and the device may be programmed to any final state without erasing the previous state, thus provides 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 ~250C). Thermal disturb is a potential challenge for the scalability of PCRAM. However, thermal disturb effect is non-cumulative (unlike Flash memory for which 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 [38], and the reset current followed the extrapolation line from larger devices. In at least one case, cycling endurance of 1E11 was demonstrated [39]. 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 candidate for eFlash replacement for embedded applications [40][41]. 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 is reported [42]. 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 [43]. This is the first commercial realization of the widely published storage class memory (SCM) [44][45]. Computer systems badly needed improve 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 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 thus 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 this case the memory cell is a PCRAM switched by OTS. The storage element may also be a resistive memory material. Although bipolar operation makes the circuitry and operation more complicated but the array structure is very similar to that using PCRAM.

PCRAM has the advantage of being unipolar in operation, more product proven, and 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 a development stage. Once developed, there seems little barrier prohibiting it from achieving 3D XP structure.

Resistive Memory (ReRAM) - A large category of two-terminal devices, which memory state is determined by resistivity of a 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) chapter. Because of their promise to scale below 10nm, 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 WL/BL half pitch. Furthermore, if eventually 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 8 [46] although no high-density ReRAM product has been introduced. Yet since the bottleneck of bipolar selector device 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.

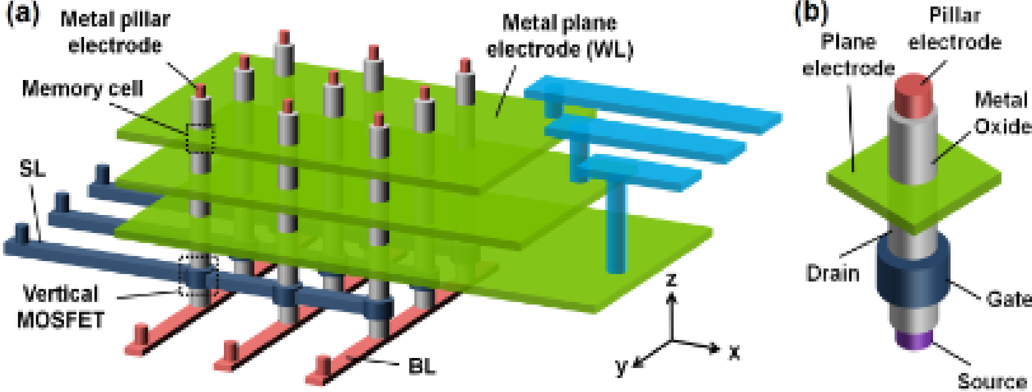


Figure 8: Schematic view of (a) 3D cross-point architecture using a vertical RRAM cell and (b) a vertical MOSFET transistor as the bit-line selector to enable the random access capability of individual cells in the array [46].

6. REFERENCES

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

[2] ITRS 2015 edition. Source: http://standards.ieee.org/develop/indconn/irds/index.html.

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

[4] W. Huang et al, "Scaling with design constraints: predicting the future of big chips", IEEE Micro, Vol. 31, No. 4, pp. 16-29, Jul-Aug 2011.

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

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

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

[10] 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.

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

[12] 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.

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

[14] 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.

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

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

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

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

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

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

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

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

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

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

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

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

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

[45] 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.

[46] 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.

7. ACKNOWLEDGMENTS

Alan Allan, Alex Burenkov, An Chen, Andrew Kahng, Atsunobu Isobayashi, Bhagawan Sahu, Carlos Beita, Charles Kin P. Cheung, Cheng-tzung Tsai, Chorng-Ping Chang, Christiane Gottschalk, Christopher Henderson, Dan Herr, Dan Mocuta, Digh Hisamoto, Eric Snyder, Fred Kuper, Frederic Boeuf, Gennadi Bersuker, Geoffrey Yeap, Gerben Doornbos, Gerhard Klimeck, Heiko Feldmann, Herve Jaouen, Hidekazu Oda, Hirofumi Inoue, Hitoshi Wakabayashi, Ichiro Mizushima, Ines Thurner, James Stathis, Jeff Butterbaugh, Jim Fonseca, Jim Hutchby, Jiro Ida, Joe Brewer, Joel Barnett, Jongwoo Park, Jurgen Lorenz, Kirk Prall, Kristin DeMeyer, Kunihiko Iwamoro, Kwok Ng, Laurent Le-Pailleur, Linda Wilson, Lothar Pfitzner, Malgorzata Jurczak, Mark Neisser, Mark van Dal, Matthias Passlack, Mehdi Salmani, Michel Haond, Mike Garner, Moon-Young Jeong, Mustafa Badaroglu, George Orji, Olivier FaynotPaolo Gargini, Patrick Cogez, Philip Wong, Prasad Sarangapani, Qi Xiang, Rich Liu, Robert Lander, Samuel C. Pan, Sang Hyun Oh, SangBum Kim, Saumitra Mehrotra, Saurabh Sinha, Shinichi Takagi, Siddharth Potbhare, SungGeun Kim, Tatsuya Ohguro, Tetsu Tanaka, Thierry Poiroux, Tohru Mogami, Tom Conte, Tony Oates, Toshiro Hiramoto, Toshiro Sugii, Wilman Tsai, Witek Maszara, Yannick Letiec, Yanzhong Xu, Yasushi Akasaka, Yasushi Gohou, Yuzo Fukuzaki