IEEE
Advancing Technology
for Humanity

IRDS

INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS™

INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS™

2020 EDITION

EXECUTIVE SUMMARY

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

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

International Roadmap Committee

Europe—Francis Balestra, Mart Graef

Japan—Yoshihiro Hayashi, Hidemi Ishiuchi

U.S.A.—Tom Conte-vice-chair, Paolo Gargini-chair

IEEE

Rebooting Computing and Standards Association, with special thanks to Erik DeBenedictis, Terence Martinez, Rudi Schubert, William Tonti, and Elie Track

Communications Society—Chi-Ming Chen

Electron Devices Society—Fernando Guarin and Terence Hook

The outstanding work by the members of the International Focus Teams is acknowledged in each of their roadmap chapters.

The chairs and co-chairs of these teams are as follows:

Application Benchmarking—Tom Conte

Systems and Architectures—Kirk Bresniker and Stephen Dukes

Outside Systems Connectivity—Michael Garner

More Moore—Mustafa Badaroglu

Lithography—Mark Neisser

Factory Integration—Supika Mashiro and James Moyne

Yield—Slava Libman and Ines Thurner

Beyond CMOS—An Chen, Shamik Das and Matt Marinella

Cryogenic Electronics and Quantum Information Processing—Scott Holmes

Packaging Integration—Dev Gupta

Metrology—George Orji

Environment, Safety, Health, and Sustainability—Leo Kenny and Steve Moffat

More than Moore—Mart Graef

IRDS Project Manager

Linda S. Wilson

Special Acknowledgment

Alan K. Allan

FORWARD

The IEEE International Roadmap for Devices and Systems (IRDS) is the continuation and extension of the National Technology Roadmap for Semiconductors/International Technology Roadmap for Semiconductors (NTRS/ITRS) with a record of uninterrupted publications spanning from 1992 to 2015. The IRDS has followed several of the successful methodologies demonstrated by the NTRS/ITRS with additional expansion to system integration, data centers and Internet communications requirements as well as quantum computing and quantum communications. The IRDS has been previously published in 2016, 2017 and 2018.

Historically the draft of all the chapters of the ITRS used to be collected in the month of October. Afterwards, the Executive Summary was quickly assembled and delivered to the Semiconductor Industry Association (SIA) Board at their mid-November annual meeting.

In the subsequent months major results and forecasts for the following year were delivered in key conferences like the International Electron Devices Meeting (IEDM) (mid-December) and International Solid-State Circuits Conference (ISSCC) (late-February). In addition, multiple company announcements were made in January outlining their vision for the year that had just started. Historically, the publication of ITRS was timed consistently with the SIA board meeting. With the transition to IEEE in May 2016 this constraint no longer existed and the publication schedule evolved towards release of the IRDS at the beginning of 2Q of the following year. As a result, all the information that previously was included as part of the roadmap update and scheduled to be published the following year has now become available for the IRDS publication in real time. As a result, the 2019 IRDS has been renamed the 2020 IRDS.

Additionally, historically the NTRS/ITRS and IRDS have predicted specific problems to occur early on, allowing enough time for research and supplier organizations to develop possible solutions. For instance in 1998 the transformation of the planar silicon gate CMOS to strained silicon, high-κ/metal-gate and FinFET was clearly articulated and these key milestones were accomplished in 2003, 2007 and 2011, respectively. Starting with the 2020 IRDS a section dedicated to Industry Highlights and IRDS messages is being introduced to better show the relations between the two organizations.

OVERVIEW

1. INTRODUCTION

IRDS MISSION

Identify the roadmap of electronic industry from devices to systems and from systems to devices.

IRDS STRUCTURE

This initiative focuses on an International Roadmap for Devices and Systems (IRDS) through the work of International Focus Teams (IFT) closely aligned with the advancement of the devices and systems industries. Led by the International Roadmap Committee (IRC), IFTs collaborated in the development of the 2020 IRDS roadmap, and engaged with other segments of the IEEE, such as the Rebooting Computing Initiative (RCI), Electron Devices Society (EDS), Computer Society (CS), Communication Society (ComSoc), and also with related industry communities like the System and Device Roadmap Japan (SDRJ) and the European SINANO Institute (ESI) in complementary activities to help ensure alignment and consensus across a range of stakeholders, such as:

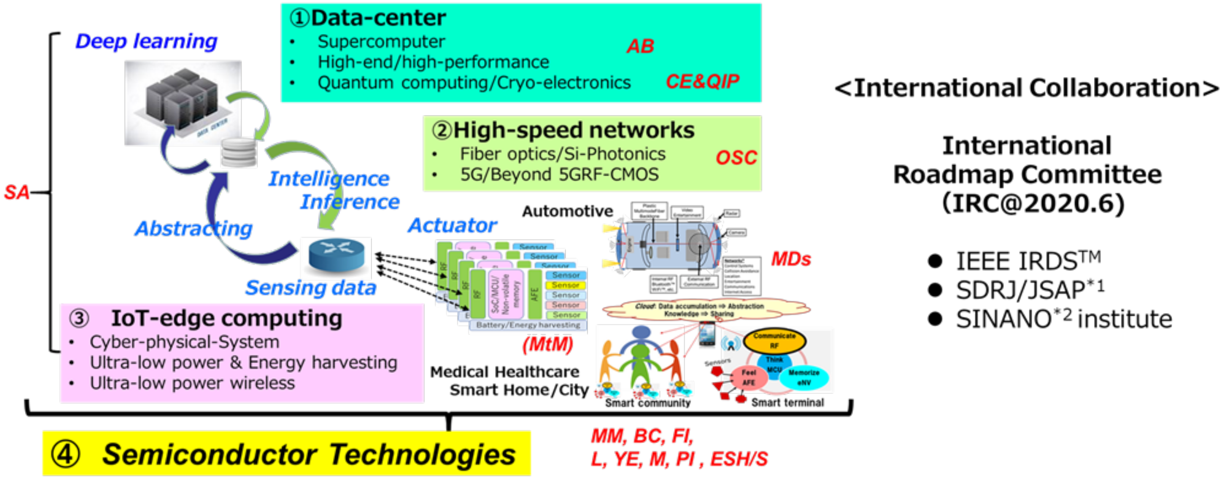
- Academia
- Consortia
- Industry
- National laboratories

IEEE, the world's largest technical professional organization dedicated to advancing technology for humanity, through the IEEE Standards Association (IEEE-SA) Industry Connections (IC) program, supports the IRDS to ensure alignment and consensus across a range of stakeholders to identify trends and develop the roadmap for all the related technologies in the computer industry.

The scope of the IRDS™ of the fundamental building blocks in the electronics industry spanning from devices to systems and from systems to devices, in which the AI-centric IoT-social-infrastructures with high-speed network communication will be built on the semiconductor device and process technologies, as illustrated in Figure ES1.

Currently, IRDS has 13 IFTs with supplemental information on the market drivers (i.e., automobile, medical devices) in the IT eco-systems, as follows: AB: Applications Benchmarking, SA: Systems and Architecture, OSC: Outside system Connectivity, MM: More Moore, BC: Beyond CMOS, CEQIP: Cryogenics Electronics and Quantum Information Processing, PI: Packaging Integration, FI: Factory Integration, L: Lithography, YE: Yield Enhancement, M: Metrology, ESH/S: Environment, Safety, Health, and Sustainability, MtM: More than Moore, MDs: Market drivers (automobile, medical devices).

Each of the IFTs are focusing not only on technical roadmaps in their specific areas but also on cross-boundary areas among the systems and the devices, or the devices and the fabrication process technologies. The major revision of the IRDS™ roadmaps is updated every two years, while the minor revision is done every year if necessary. Here, the IRDS 2020 Edition is the material after a major revision. Under management of the international roadmap committee (IRC) of the international representatives, the technical points in each IFT have been selected and discussed internally, and the results are being released to public through the IEEE IRDS™ homepage.



AB: Applications Benchmarking, ***SA***: Systems and Architecture, ***OSC***: Outside system Connectivity, ***MM***: More Moore, ***BC***: Beyond CMOS, ***CE&QIP***: Cryogenics Electronics and Quantum Information Processing, ***PI***: Packaging Integration, ***FI***: Factory Integration, ***L***: Lithography, ***YE***: Yield Enhancement, ***M***: Metrology, ***ESH/S***: Environment, Safety, Health, and Sustainability, ***MtM***: More than Moore, ***MDs***: Market drivers (automobile, medical devices).

*1: The System Device Roadmap Committee of Japan/The Japan Society of Applied Physics
*2: The European Academic and Scientific Association for Nanoelectronics

The scope of the IRDS™ of the fundamental building blocks in the electronics industry spanning from devices to systems and from systems to devices under the international collaboration to discuss the focus points. See the details in Section 2.1,"Roadmap Process".

Figure ES1 The new ecosystem of the electronics' industry based on semiconductor technologies.

IEEE SPONSORS

The IRDS is sponsored by the IEEE Rebooting Computing (IEEE RC) Initiative in consultation and support from many IEEE Operating Units and Partner organizations, including the following:

CASS—Circuits and Systems Society
CEDA—Council on Electronic Design Automation
CS—Computer Society
CSC—Council on Superconductivity
EDS—Electron Devices Society
EPS— Electronics Packaging Society
MAG—Magnetics Society
NTC—Nanotechnology Council

RS—Reliability Society
SSCS—Solid State Circuits Society
SRC—Semiconductor Research Corporation
IEEE Standards Association

International Sponsors and Cooperation

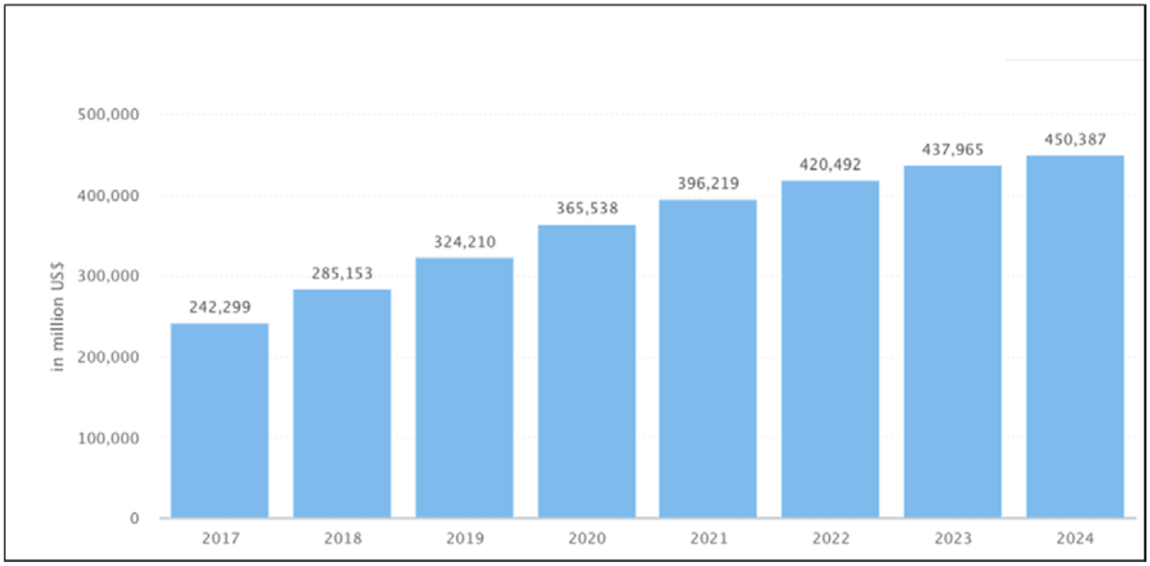
There are several international roadmap efforts directly aligned with the IRDS:

- The SINANO Institute http://www.sinano.eu/
- The System Device Roadmap Committee of Japan (SDRJ) https://sdrj.jp/
- The International Electronics Manufacturing Initiative (iNEMI) http://www.inemi.org/

1.1. 2019/2020 Industry Highlights and Key Messages from the 2020 IRDS

This is a brand new section to better place in context the relation between IRDS projections and the progress of the electronics industry. There are many subjects that could be addressed but this is a simple collection of few key items brought to the attention of the reader.

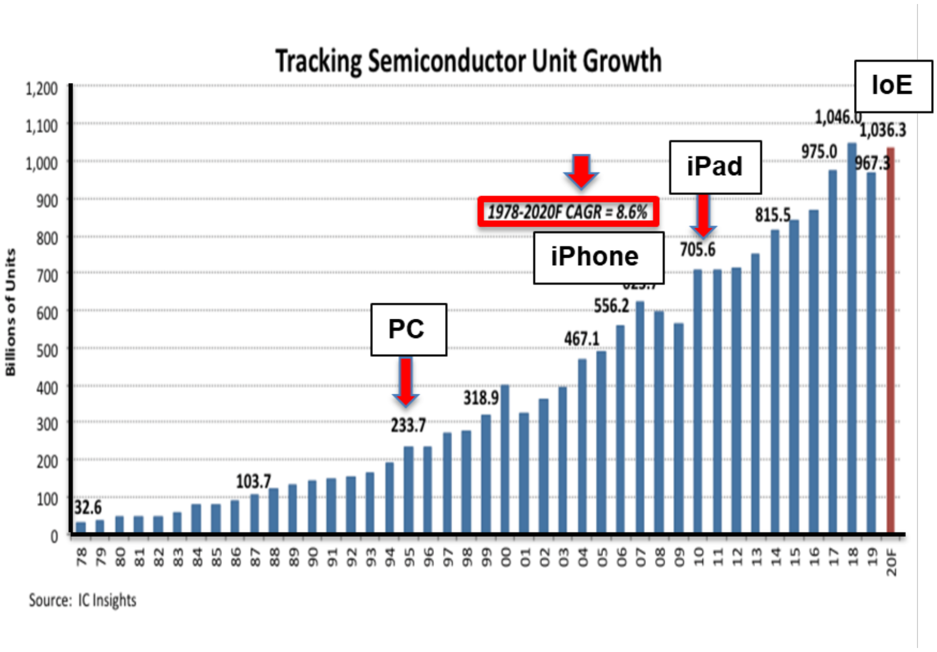
1. The electronics industry and the semiconductor industry remained very healthy in 2019. Consumer electronics grew 10% in 2019 (year over year or YoY) and is forecasted to grow 13% in 2020 to $365B. See Figure ES2.



Source: Statista

Figure ES2 Consumer electronics growth continues

2. The number of units to be shipped in 2020 is forecasted to exceed the 1 billion mark! This represents a 7% growth over 2019 and an average growth of 5%/year over the past 5 years. See Figure ES3.


Tracking Semiconductor Unit Growth

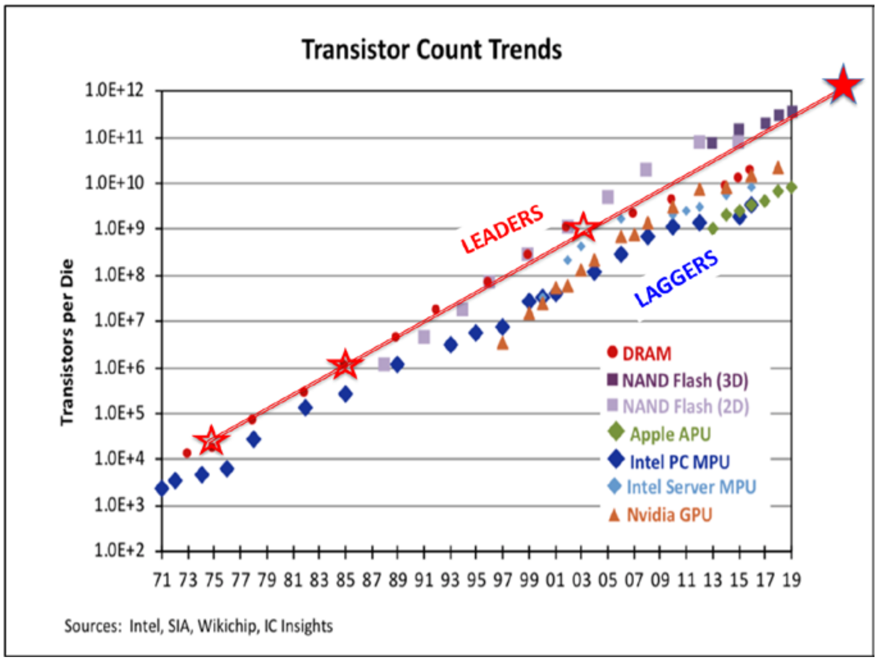
Source: IC Insights

Source: IC Insights

Figure ES3 Unit growth on track at CAGR of 8.6%/year

3. Semiconductor R&D spending grew to 14.6% of revenue in 2019 and a slightly better level of investment is expected in 2020.

4. Moore's Law continues unabated See Figure ES4.


Transistor Count Trends

Sources: Intel, SIA, Wikichip, IC Insights

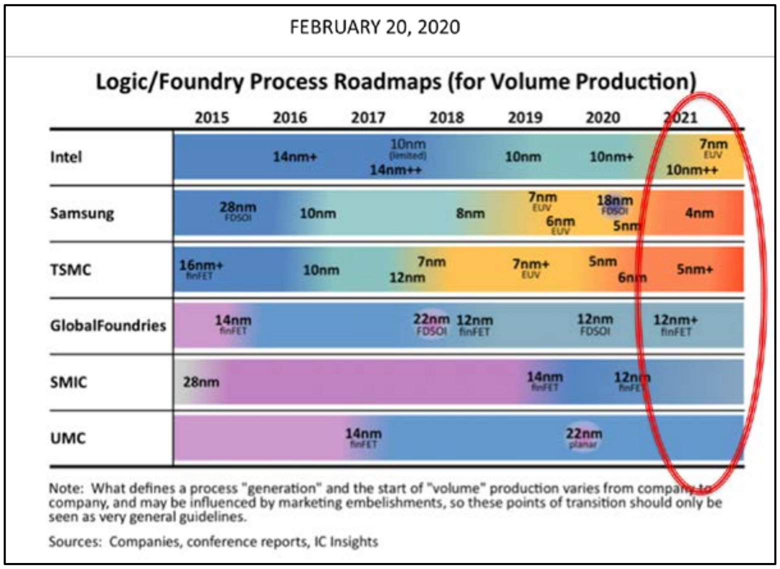
Figure ES4 Moore's Law continues unabated

5. Shipments of logic products with a minimum metal pitch of 36 nm (corresponding to the 18 nm technology node definition according to NTRS/ITRS and IRDS nomenclature) will be introduced in 2020 (Figure ES5). Silicon manufactures still insist in calling this technology “the 5 nm node” even though these node numbers do not correspond to any meaningful and measurable quantity related to transistor density on the wafer! See Figure ES6. (A more detailed explanation is discussed in Section 1.2.3.)

YEAR OF PRODUCTION	2020	2022	2025	2028	2031	2034
	G48M36	G45M24	G42M20	G40M16	G38M16T2	G38M16T4
Logic industry “Node Range” labeling (nm)	“5”	“3”	“2.1”	“1.5”	“1.0 eq”	“0.7 eq”
IDM-Foundry node labeling	i7-f5	i5-f3	i3-f2.1	i2.1-f1.5	i1.5e-f1.0e	i1.0e-f0.7e
Logic device structure options	FinFET	FinFET LGAA	LGAA	LGAA	LGAA-3D	LGAA-3D
Mainstream device for logic	FinFET	FinFET	LGAA	LGAA	LGAA-3D	LGAA-3D
LOGIC TECHNOLOGY ANCHORS						
Patterning technology inflection for Mx interconnect	193i, EUV DP	193i, EUV DP	193i, EUV DP	193i, High-NA EUV	193i, High-NA EUV	193i, High-NA EUV
Beyond CMOS as complimentary to mainstream CMOS	-	-	-	2D Device, FeFET	2D Device, FeFET	2D Device, FeFET
Channel material technology inflection	SiGe25%	SiGe50%	SiGe50%	Ge, 2D Mat	Ge, 2D Mat	Ge, 2D Mat
Process, technology inflection	Conformal doping, Contact	Channel RMG	Lateral/ Atomic Etch	Non-Cu Mx	3D VLSI	3D VLSI
Stacking generation inflection	2D	3D stacking: W2W, D2W Mem-on-Logic	3D stacking: W2W, D2W Mem-on-Logic	3D stacking, Fine-pitch stacking, P-over-N, Mem-on-Logic	3D stacking, 3D VLSI: Mem-on-Logic with Interconnect	3D stacking, 3D VLSI: Logic-on-Logic

Note: Mx—Tight-pitch routing metal interconnect. IDM—independent device manufacturer. FinFET—fin field-effect transistor. LGAA—lateral gate all around. EUV—extreme ultraviolet. NA—numerical aperture. Ge—germanium. SiGe—silicon germanium. RMG—replacement metal gate. VLSI—very large scale integration. W2W—wafer to wafer. D2W—die to wafer. Mem-on-Logic—memory on logic

Figure ES5 IRDS nodes capture essence of technology


FEBRUARY 20, 2020

Logic/Foundry Process Roadmaps (for Volume Production)

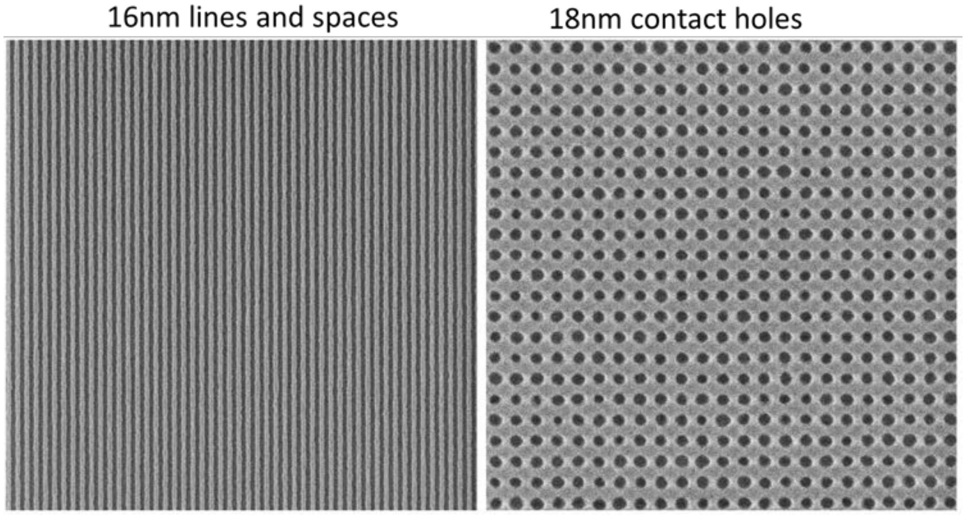
Note: What defines a process "generation" and the start of "volume" production varies from company to company, and may be influenced by marketing embelishments, so these points of transition should only be seen as very general guidelines.

Sources: Companies, conference reports, IC Insights

Figure ES6 Industry unrealistic node nomenclature

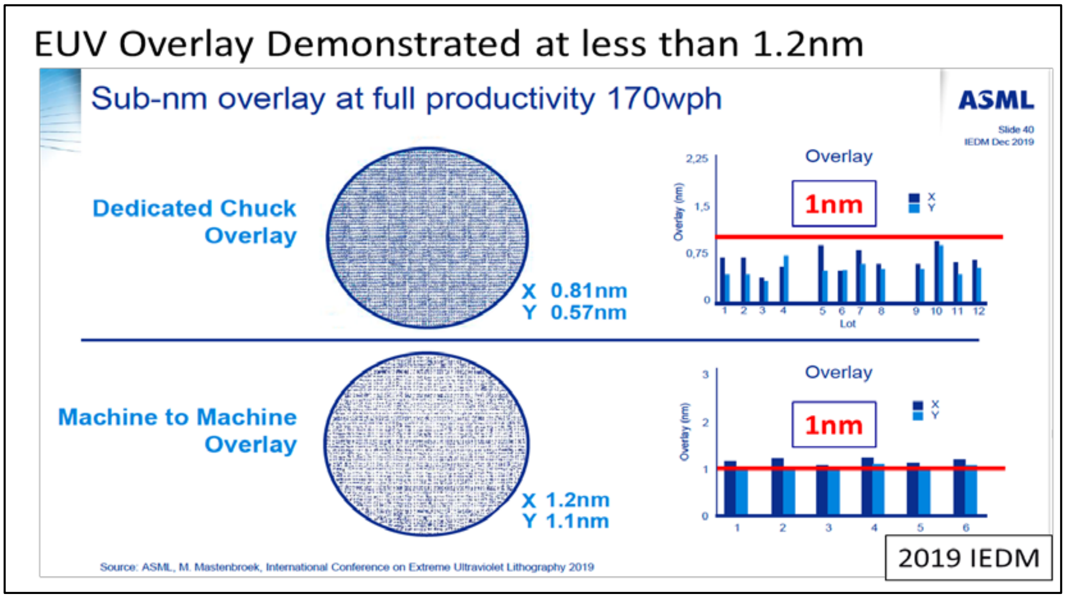
6. EUV LLC started systematic investments in the development of this lithography technology in 1997; after more than 20 years in development EUV lithographic finally made it into high-volume manufacturing in 2019, as demonstrated by more than 30 EUV tools delivered to leading manufactures.

7. Reported experimental results, including presentations made at IEDM 2019, showed that lines and spaces of 16 nm were realized with a single exposure by using a EUV scanner. In addition, full wafer overlay better than 1.2 nm for a cluster of multiple EUV tools was reported. See Figures ES7 and ES8. EUV exposure technology is predicted to drastically reduce the number of masks required to meet design feature goals. See Figure ES9.



Source: IMEC

Figure ES7 Line and holes imaged by single EUV exposure



EUV Overlay Demonstrated at less than 1.2nm

Sub-nm overlay at full productivity 170wph

ASML

Slide 40
IEDM Dec 2019

Dedicated Chuck Overlay

X 0.81nm
Y 0.57nm

Machine to Machine Overlay

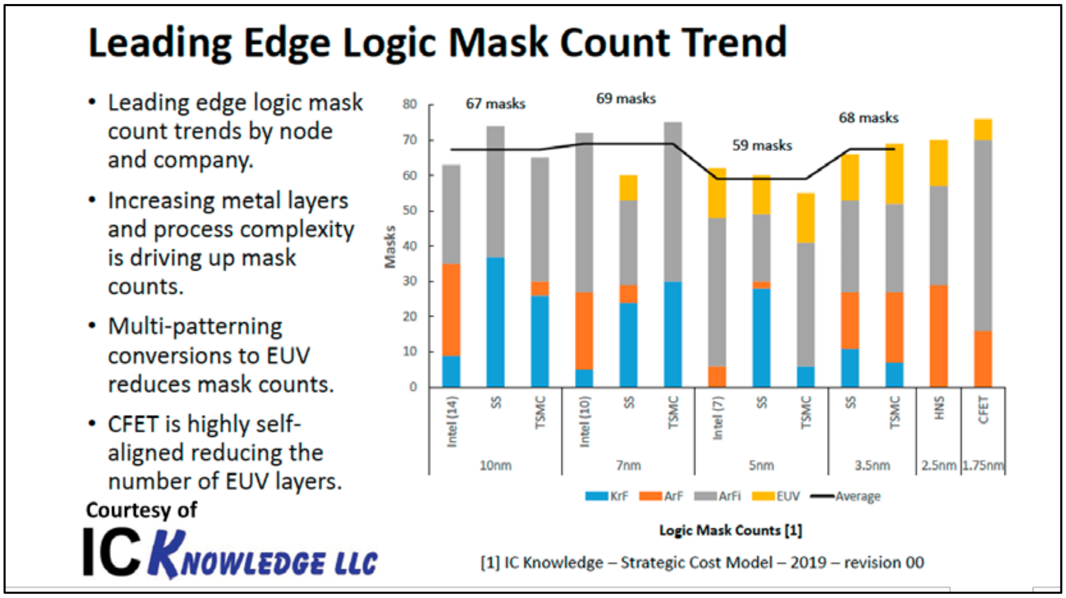
X 1.2nm
Y 1.1nm

Source: ASML, M. Mastenbroek, International Conference on Extreme Ultraviolet Lithography 2019

2019 IEDM

Source: ASML

Figure ES8 Overlays results from multiple EUV scanners



Leading Edge Logic Mask Count Trend

- Leading edge logic mask count trends by node and company.
- Increasing metal layers and process complexity is driving up mask counts.
- Multi-patterning conversions to EUV reduces mask counts.
- CFET is highly self-aligned reducing the number of EUV layers.

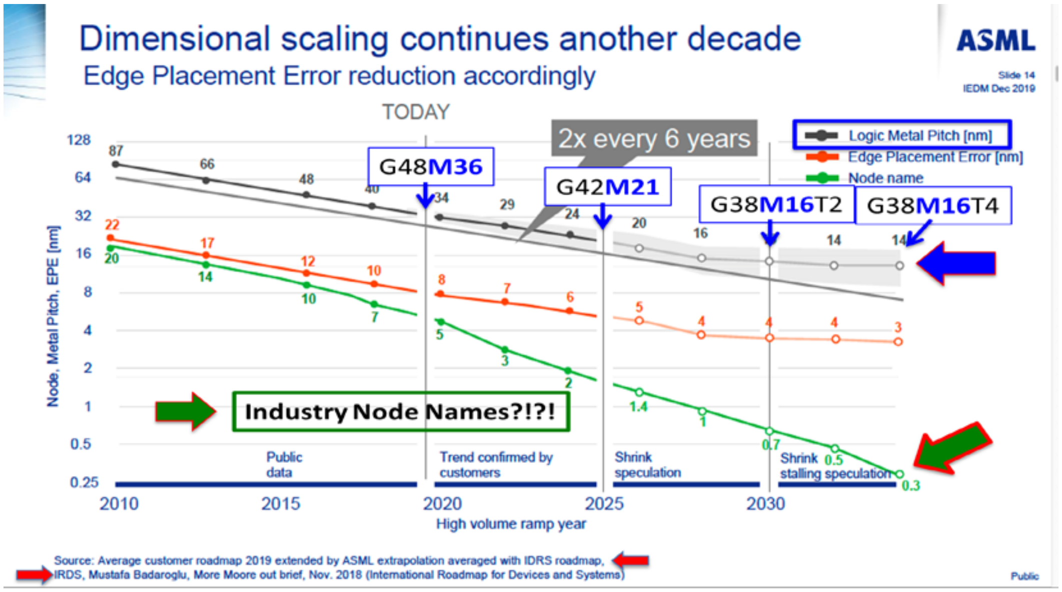
Courtesy of
IC KNOWLEDGE LLC

Logic Mask Counts [1]

[1] IC Knowledge – Strategic Cost Model – 2019 – revision 00

Figure ES9 Reduction in number of masks by EUV exposure

8. Both the 2020 IRDS and ASML predict that the ability to reduce features (line and space) would reach final limits around 7-8 nm by the end of this decade. See Figure ES10 and a more detailed explanation in Section 1.2.3.



Dimensional scaling continues another decade

Edge Placement Error reduction accordingly

Source: Average customer roadmap 2019 extended by ASML extrapolation averaged with IDRS roadmap, IRDS, Mustafa Badaroglu, More Moore out brief, Nov. 2018 (International Roadmap for Devices and Systems)

Courtesy of ASML

Figure ES10 Consistent node timing forecast from IRDS and ASML

9. New device architectures utilizing both vertical monolithic and heterogeneous integration will continue to support increased **functionality** at historical rates for at least 10 more years but expect more innovation beyond that. . See Figure ES11. (Refer to Highlight number 12 below for more details.)

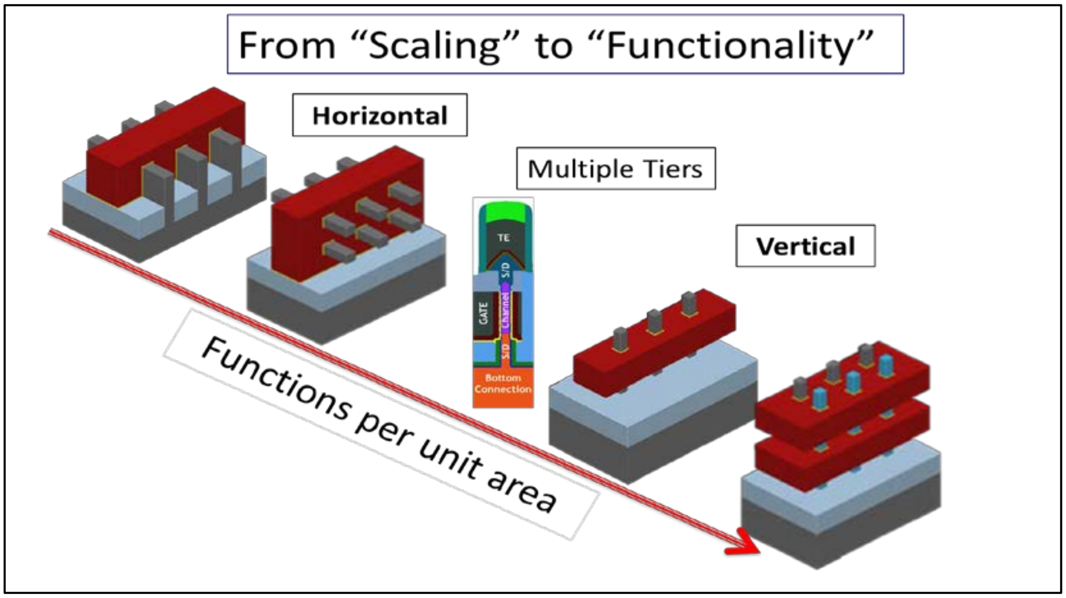
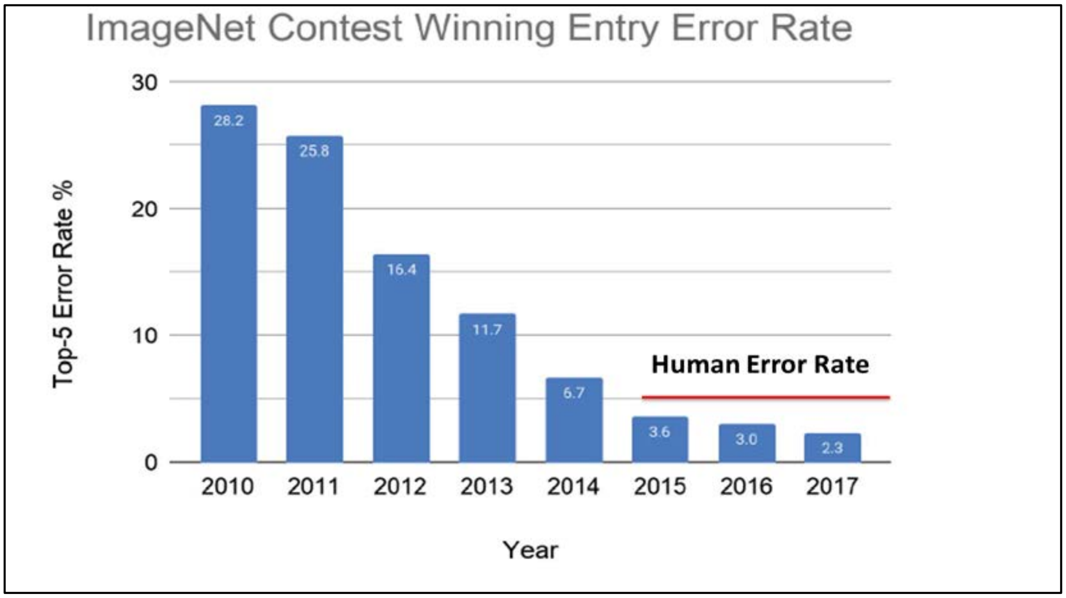


Figure ES11 From 2D transistor scaling to 3D functionality

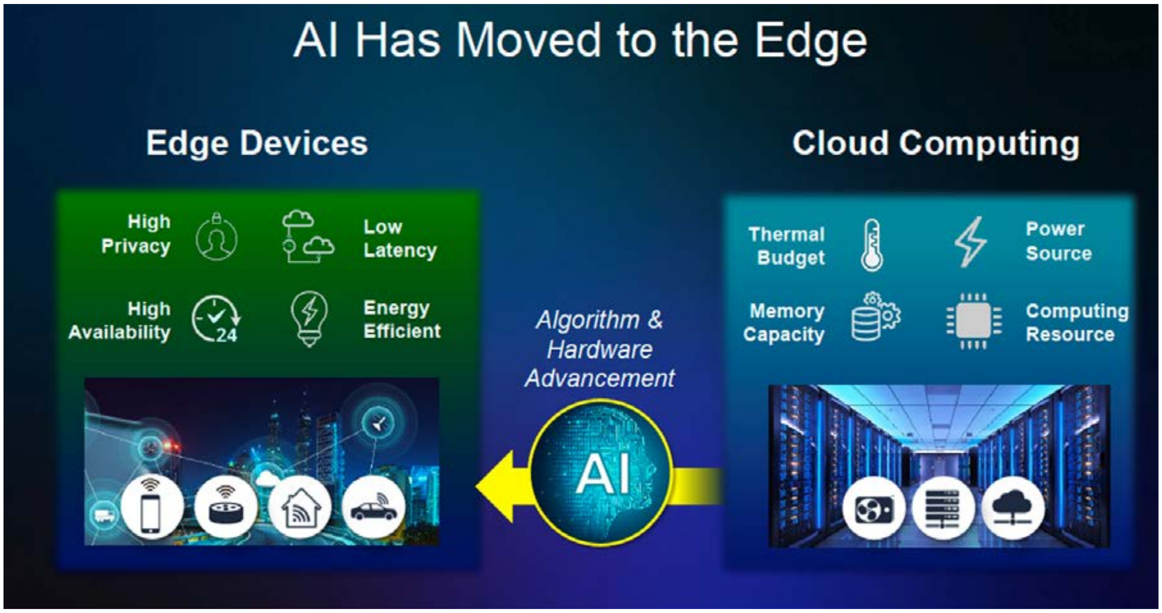
10. Artificial Intelligence and Machine Learning (AI/ML) will continue to revolutionize how computers operate. Multiple successful results using AI/ML were shown in the Google keynote address at ISSCC 2020. In 2017 neuromorphic architecture utilizing extreme parallel processing demonstrated remarkably better ability in image recognition than humans by achieving an error rate of 2.3%; this is way below the typical human error rate of 5%. See Figure ES12.



Source: Google

Figure ES12 AI/ML exceeded human performance

11. Since 2017 AI/ML has been proliferating to multiple applications in the computer field. Speech recognition in 2019 closely followed in the successful streak demonstrating an error rate as low as 4%. AI now dominates new video gaming systems as well. AI/ML has been permeating all the fields of data acquisition. In fact, it was stated in a keynote delivered at ISSCC 2020 by Media Tek that AI/ML is now quickly moving to the "edge". See Figure ES13.



AI Has Moved to the Edge

Edge Devices

High Privacy

Low Latency

High Availability

Energy Efficient

Algorithm & Hardware Advancement

Cloud Computing

Thermal Budget

Power Source

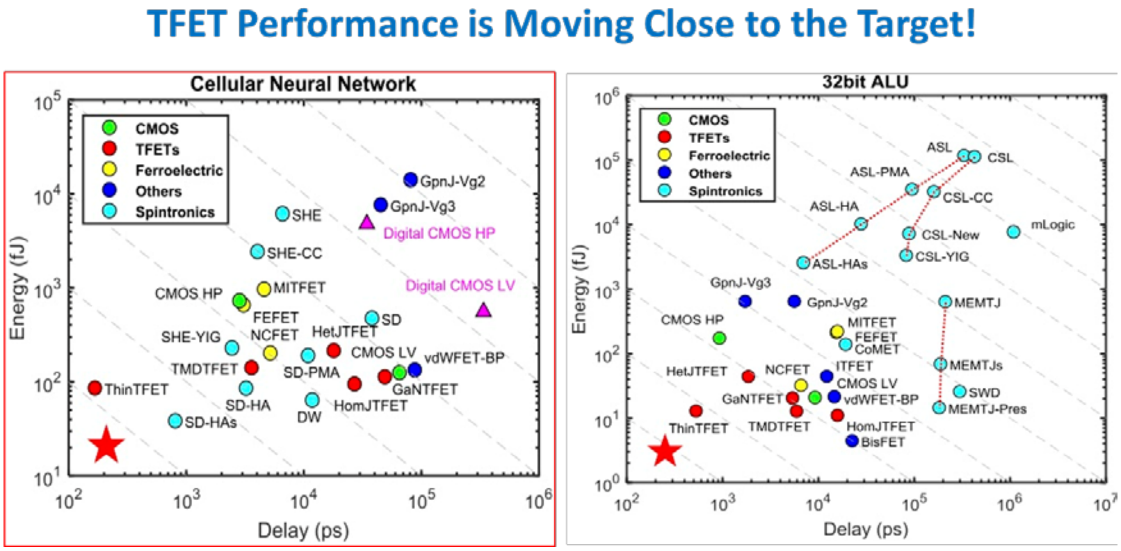
Memory Capacity

Computing Resource

Source: Media Tek, 2020 VLSI Symposium

Figure ES13 Data centers moving closer to the edge

12. Research in Beyond CMOS devices continues to successfully proceed. Performance of several new devices is continuing to improve. See Figure ES14. (Refer to the 2020 Beyond CMOS Chapter for more information.)



TFET Performance is Moving Close to the Target!

Figure ES14 Beyond CMOS continues to improve towards goal

13. Quantum computing (QC) and quantum communications promise unprecedented increases in functionality; from Moore's Law historical exponential growth to "quantum" functionality that grows at an exponential exponent rate but… **will QC be ready for manufacturing in time by 2030?** See Figure ES15.

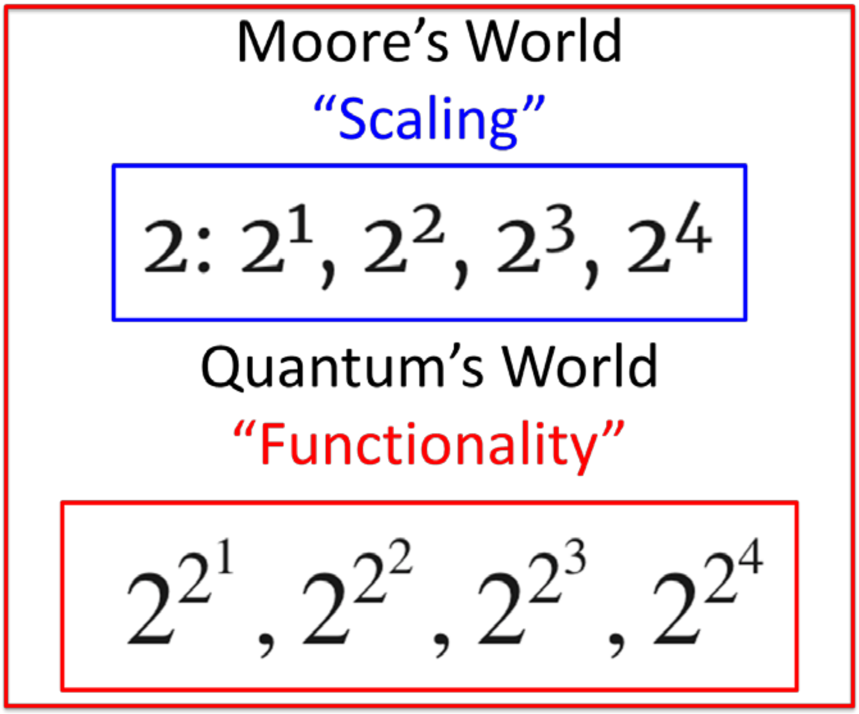
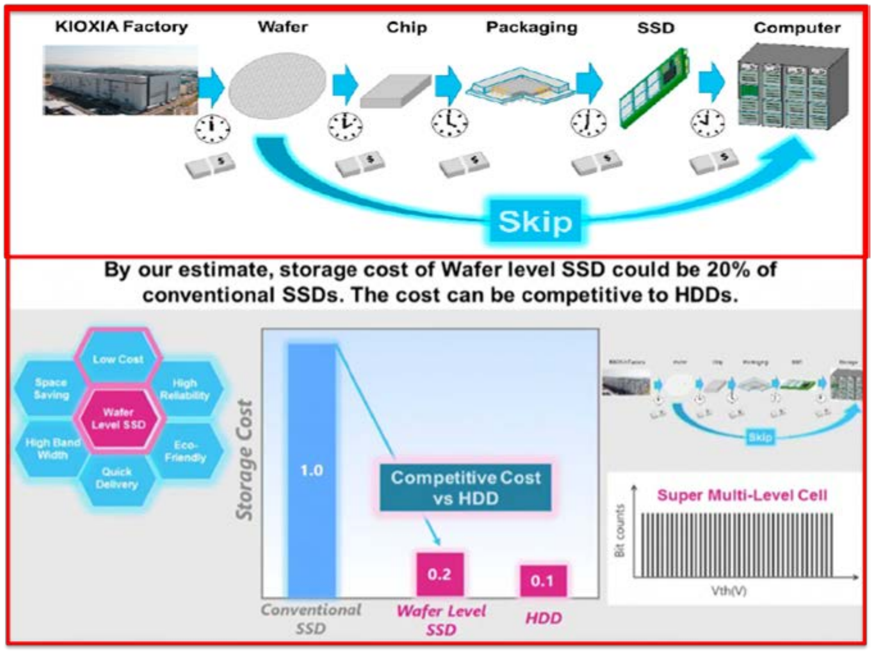


Figure ES15 Quantum promises super exponential functionality

14. NAND manufacturers have continued to add stacks of memory layers since the technology was first announced by the 2013 ITRS and introduced in manufacturing in 2015; products with more than 100 layers have begun shipping in 2020.

15. Several new technology options, including wafer-level solid state drives (SSDs), were proposed at VLSI Symposia 2020 during the Kloxia Keynote address to reduce cost and latency of NAND products. These innovative approaches would reduce the storage cost to 20% of traditional by eliminating the whole packaging and assembly manufacturing steps (Figure ES16).[1] Should this approach be successful in the future it would mean that NAND technology would begin to seriously challenge hard disk drive (HDD) costs.

[1] *https://vlsisymposium.org/*

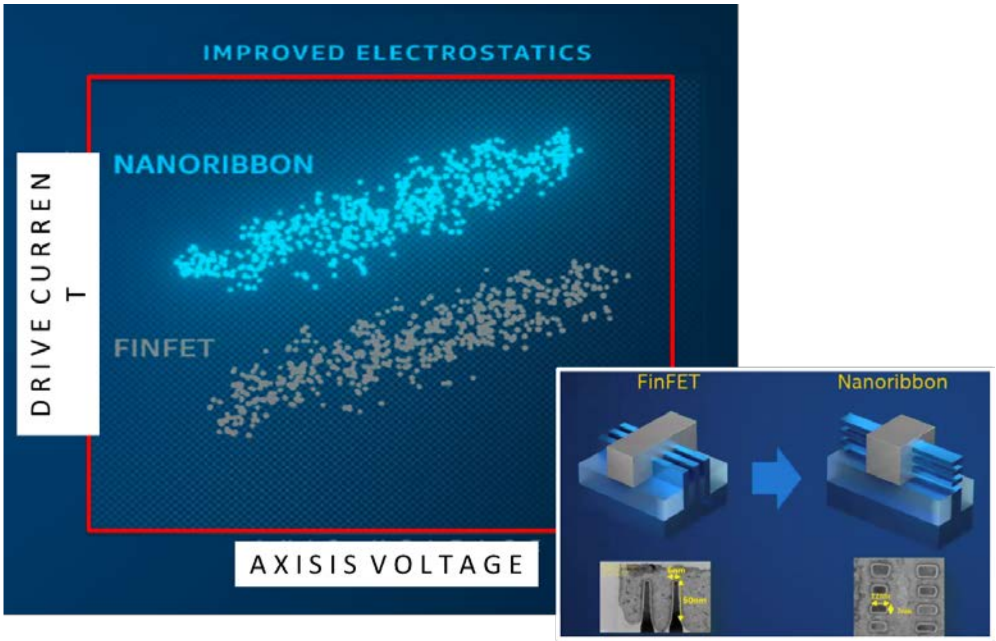


By our estimate, storage cost of Wafer level SSD could be 20% of conventional SSDs. The cost can be competitive to HDDs.

Source: Kloxia, VLSI Symposia 2020

Figure ES16 NAND density/cost challenging HDD

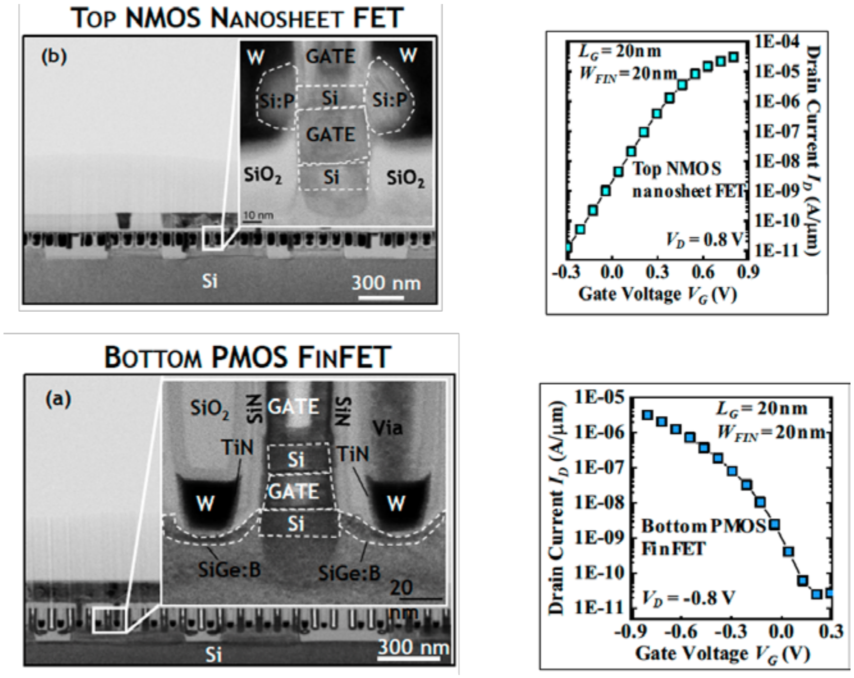
16. Logic companies continue to flirt with the concept of stacked transistors. Intel reported at the VLSI Symposia 2020 that the drive current of nano-stacked ribbons is higher than that of FinFETs (Figure ES17).[2] Several combinations of bottom devices and top devices were also presented by IMEC (Figure ES18). Which logic company or foundry will announce the first product using stacked transistors in the near future?



Source: Intel VLSI Symposia 2020

Figure ES17 Nanoribbons performances exceed FinFETs

[2] *https://vlsisymposium.org*



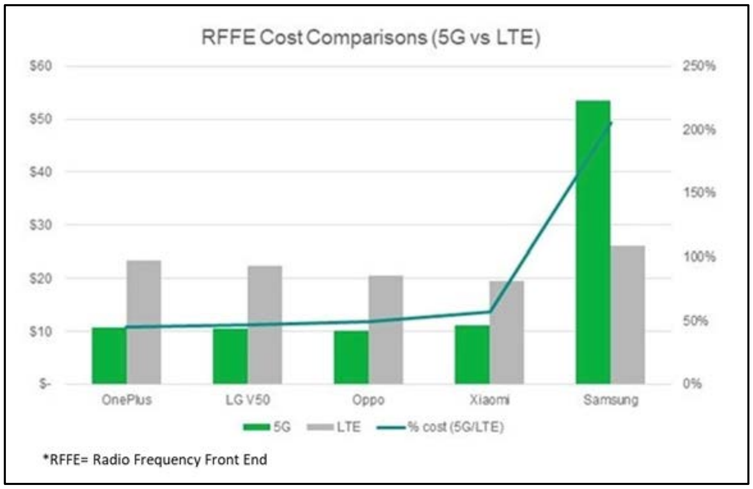
Source: IMEC

Figure ES18 Progress in stacking CMOS for logic applications

17. In 2005 the ITRS identified a new concept on how functionality could further increase. “More than Moore” (MtM) refers to the incorporation into devices of functionalities that do not necessarily scale according to “Moore's Law”, but provide additional value in different ways. The “More-than-Moore” approach allows for the non-digital functionalities (e.g., RF communication, power control, passive components, sensors, actuators) to migrate from the system board-level into the package (SiP) or onto the chip (SoC). (See the 2020 IRDS MtM white paper).

18. It was anticipated that the integration of CMOS-based system on chip (SoC) and non-CMOS based system in package (SiP) technologies within a single package would become increasingly important. Combination of More than Moore and Packaging Integration were defined as the foundations of future heterogeneous integration by ITRS in 2012. The first chapter of Heterogeneous Integration was published with the 2015 ITRS 2.0. The 2020 IRDS is finally resuming this subject by presenting the two new white papers: More than Moore and Packaging Integration.

19. The first 5G phones are in the market but get ready for the price shock (Figure ES19).



*RFFE= Radio Frequency Front End

Figure ES19 Introductory cost of 5G phones may challenge the consumers[3]

20. Due to the Covid-19 virus, video conferencing exploded in 2020 (e.g., 10 million Zoom meetings were held in Dec 2019 vs. 300 million Zoom meetings held in April 2020)

21. For the first time in years the personal computer industry saw a surge in demand in 1Q 2020, due to millions of people suddenly forced to work away from their offices because of Covid-19. Microprocessors revenue for PCs and notebooks surged to +14% in 1Q 2020 (year to year (YtY)) but due to problems with the supply chain shipments of PCs and notebooks sales decreased about 10% from 1Q 2019!

22. For the past 5 years the number of data centers has kept on increasing (Figure ES20). Traffic within a data center occupies the largest portion of time (Figure ES21). In the next 10 years the number of data centers will continue to increase but most likely in a very different way. At the 2020 VLSI Symposium it was stated that the proliferation of the **5G networks** (do not think only about the cell phones identified by this name) will facilitate access to storage and elaboration of data at the "edge" creating a surge in the demand for smaller but more efficient distributed data centers. See Figure ES22.

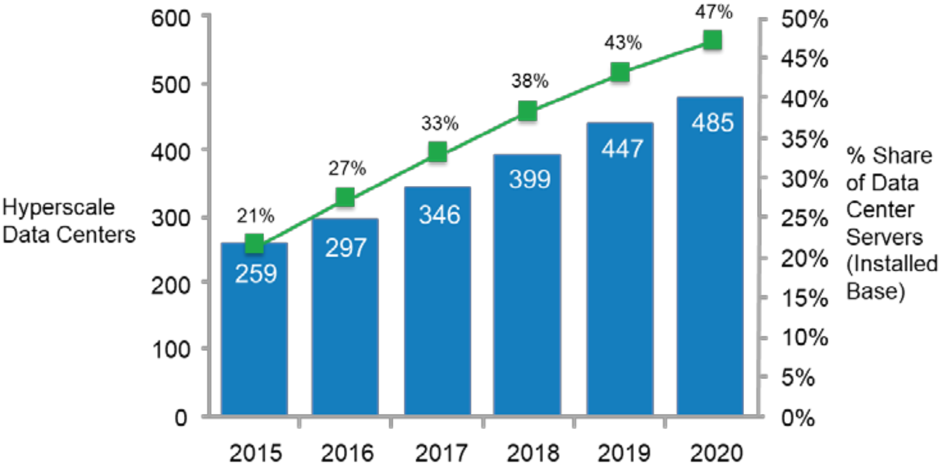


Figure ES20 Data center growth continues to accelerate

[3] *https://technology.informa.com/616863/in-5g-smartphone-designs-rf-front-end-graduates-from-traditional-supporting-role-to-co-star-with-modem*

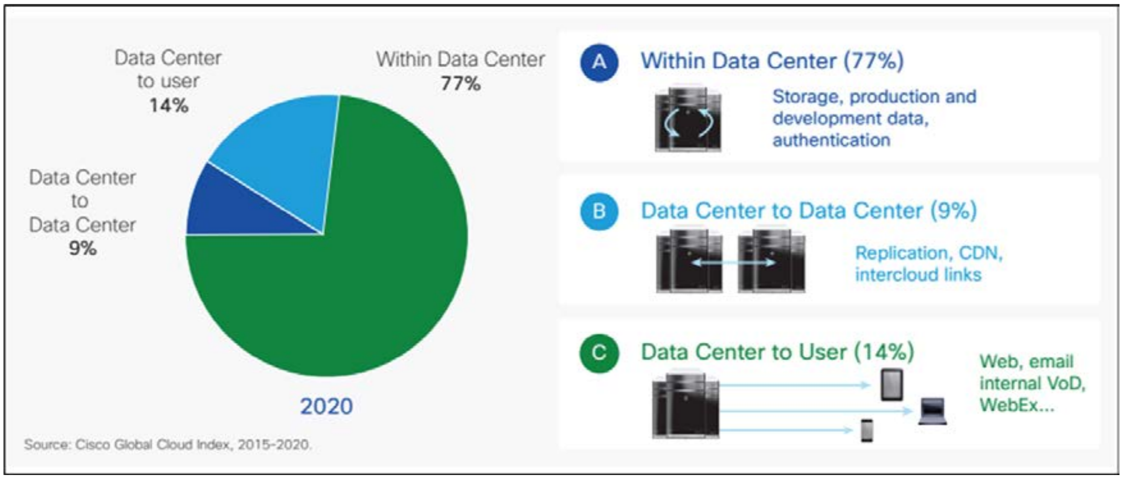
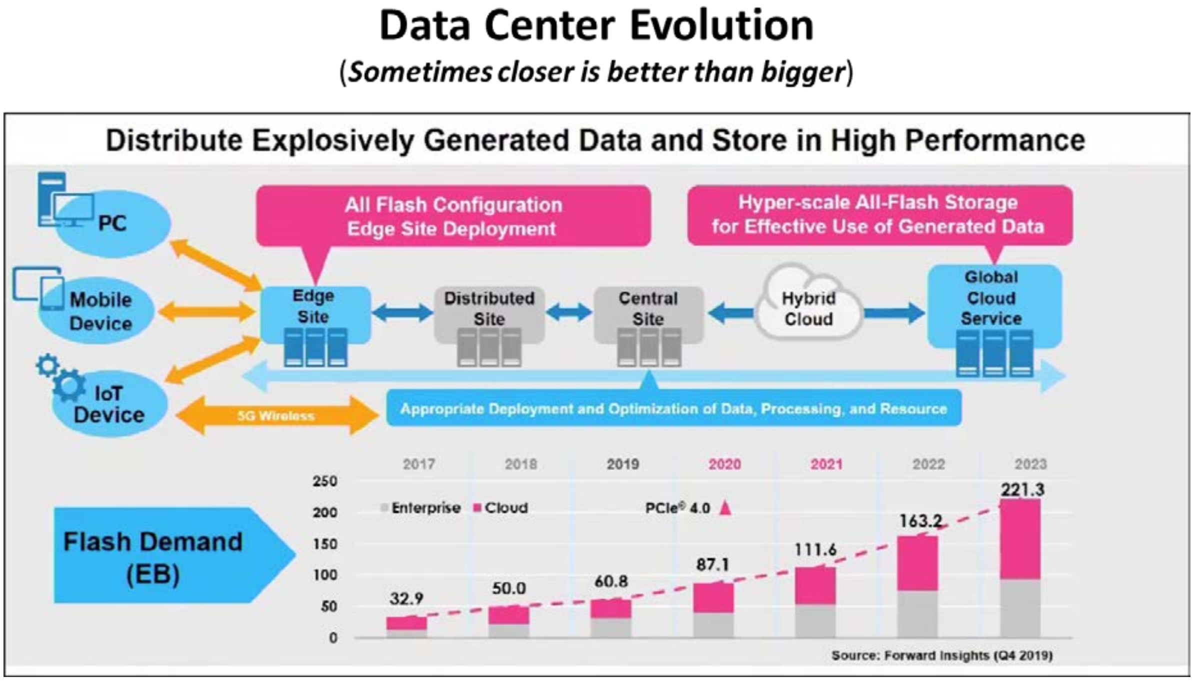


Figure ES21 Traffic within data center is main limiting factor



Source: Kloxia, 2020 VLSI Symposium

Figure ES22 Sensors data surge at the edge will require proximity of data centers

23. Most of the more promising implementations of Quantum Computing (QC) operate at temperatures as low as -273°C. Due to this constraint consumers will not be able to carry QC mobile devices in their pockets and will have to rely on powerful and far-reaching 5G/6G networks to remotely access Data Centers for information and Quantum Centers for computing; this vision for the year 2030 and beyond is supported by IRDS and the IEEE International Network Generations Roadmap (INGR). See Figure ES23.

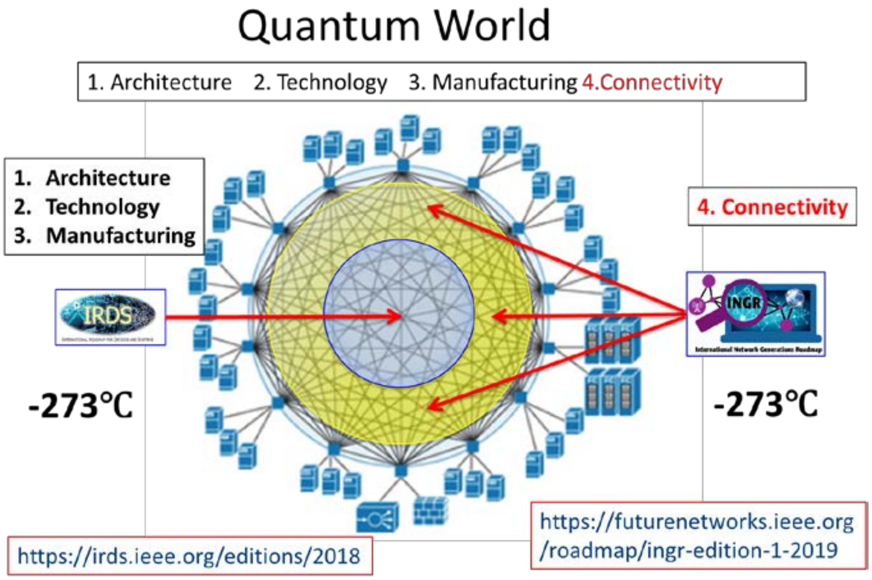

Quantum World
1. Architecture 2. Technology 3. Manufacturing 4.Connectivity
1. Architecture
2. Technology
3. Manufacturing
4. Connectivity
IRDS
INGR
International Network Generations Roadmap
-273°C
-273°C
https://irds.ieee.org/editions/2018
https://futurenetworks.ieee.org
/roadmap/ingr-edition-1-2019

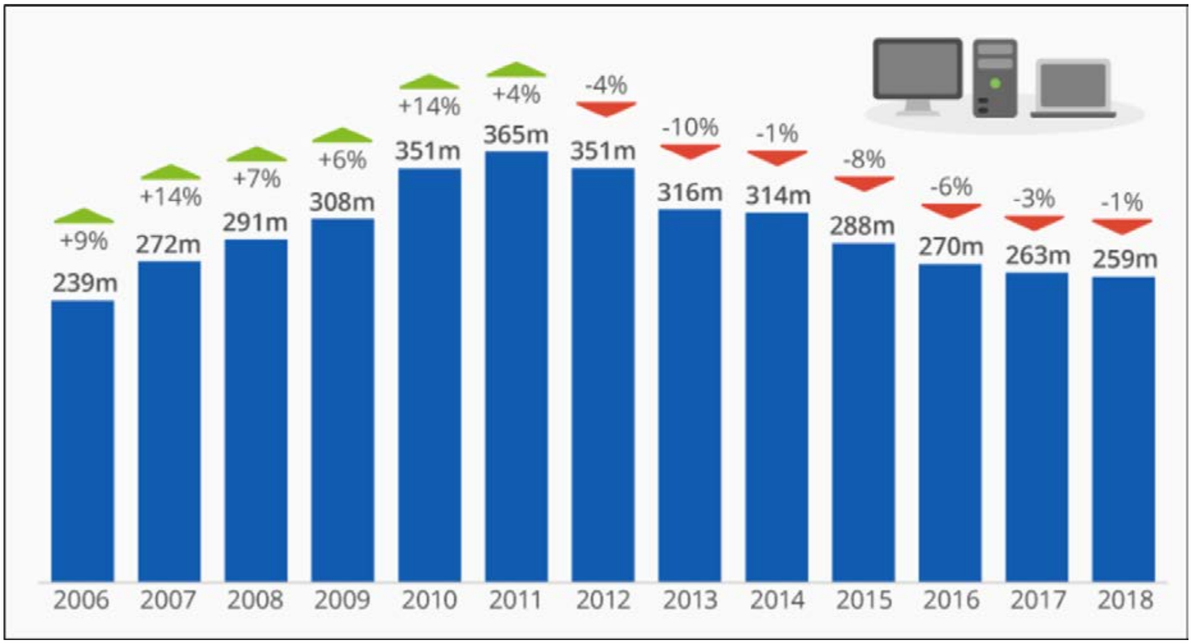
Figure ES23 Quantum centers require fast network for ubiquitous access

1.2. The New Ecosystem of the Electronics' Industry

1.2.1. The Big Picture

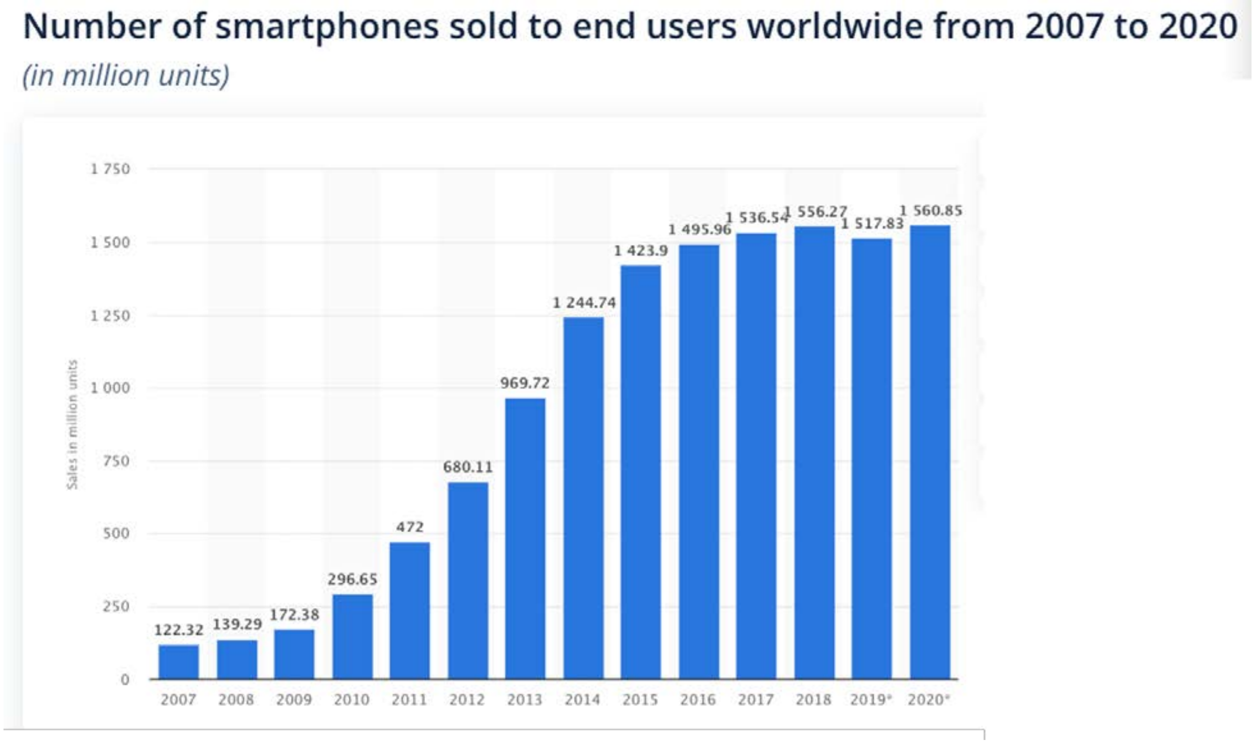
The **computer industry** and the semiconductor industry have been driving hand in hand the growth of the electronics industry for over 60 years. This combination has produced faster, smaller and cheaper components that have enabled progressively larger, faster and more powerful computing machines. Close cooperation with leaders in the software industry simplified and solidified the association between hardware and software. The introduction of the personal computer (PC) to corporations in the 80s was closely followed by the diffusion of PCs and notebooks to consumers in the 90s; these appliances were capable of doing computation and were also capable of connecting to the Internet by means of somewhat awkward cables. Introduction of Wi-Fi in 1998 eliminated this problem and increased employees' productivity by making the connection to the Internet completely ubiquitous and almost instantaneous inside corporations. On the other hand, by then the consumers were already dependent on cell phones and were ready for more.

Cell phones and Wi-Fi notebooks had already created the working attitude of "business on the go" especially in the business community but consumers were also closely following in the adoption of mobile devices. Meanwhile, the size of the electronics inside a cell phone kept on reducing driven by Moore's Law and from the 2" by 5" size of typical cell phones in the 90s new cell phones had reduced to a 1" by 2" format in early 2000. It may seem obvious looking back now but this implied that there was a lot of space potentially available to fill up the early cell phone format with a lot of electronics and in 2007 the iPhone did just that. Now a single device could provide everything that consumers had become used to do with PCs, notebooks and a cell phones and even more (books, songs, pictures and more). While sales of PCs and notebooks each hovered in the 300 million range at their peak (Figure ES24) the new generation of smart phones skyrocketed from about 122 million in 2007 (year of introduction) to 1.5 billion units in 2017 and maintained this unit sale level in subsequent years (Figure ES25).



Source: Statista

Figure ES24 Rise and decline of PC unit sales

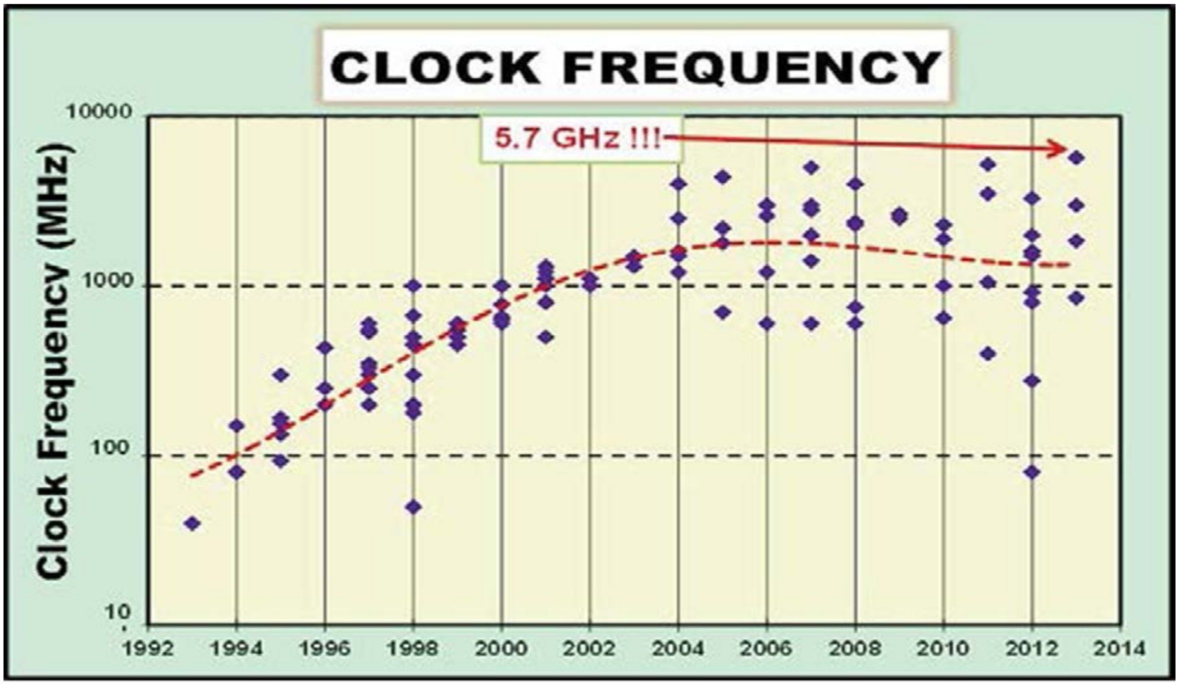


Source: Statista

Figure ES25 Cell phones climb and plateau at 1.5 billion units

On the **technology side** it is worth mentioning that major limitations were already identified by the NTRS/ITRS in the 90s giving enough time to the research community to provide alternative solutions to avoid major stumbling blocks later on. As a result, the transistor structure and component materials migrated from the planar topology to the vertical FinFET structure while a new high-κ material encased in a metal upper electrode replaced the silicon dioxide/polysilicon gate.

In addition, the power dissipation of microprocessors reached the 120-130 W limits by the middle of the previous decade but changing the architecture from a single core to multiple ones operating in parallel alleviated the problem. However, despite this architectural transformation the maximum frequency of operation had to be limited to a maximum value in the 5-6 GHz (Figure ES26) to avoid operational problem due to excessive power dissipation (See Section 5 for more details).



IEEE, ISSCC: Transistor's 60th year commemorative supplement

Figure ES26 Clock frequency limited to 4-6 GHz by power wall

For the next few years, some level of improvement in computing performance was achieved by implementing the aforementioned **multicore architecture** but intrinsically a microprocessor operates by alternating between serial and parallel operations. As a results performance could no longer double every two years following historical trends. Since it was becoming evident that it was impossible to completely adapt microprocessor operation to a 100% parallel approach it was then time to turn the table around and ask the question of whether it was possible to find an application that was intrinsically 100% parallel in nature. This concept was quickly reconnected with a type of architecture proposed in 1980 called neuromorphic computing. This architecture proposed the use of very-large-scale integration (VLSI) systems containing electronic analog circuits to mimic neuro-biological architectures present in the nervous system. A key aspect of neuromorphic engineering is understanding how the morphology of individual neurons, circuits, applications, and overall architectures creates desirable computations; affects how information is represented; influences robustness to damage; incorporates learning and development; how it adapts to local change (plasticity), and facilitates evolutionary changes. As an example oxide-based memristors, spintronics memories, threshold switches, and transistors can easily realize the hardware implementation of neuromorphic computing. It was now a time to answer the question of which application could be best matched to this architecture.

Finally, the right example arrived. In 2007 the ability of systems to compare and identify a generic subject to a database was rather limited since the typical error in object recognition was around 28% as compared to typical human error rate of about 5%. However, multiple neuromorphic algorithms progressively decreased this error to the 2.3% level by 2017, well below the human error rate!

Clever circuit architecture, feature-down scaling and high yielding larger die were and remain at the foundation of Moore's Law since the mid-60s. In order to image progressively smaller features on a wafer it had been necessary in time to introduce lithography exposure tools with lenses of progressively higher numerical apertures (NAs) and illumination sources with smaller wavelengths. However, the relentless pace of introducing new technologies every 2 years strained the equipment community and the suppliers ran out of technology for new tools capable of using exposure wavelengths smaller than 193 nm. As far back as 1997 research activities on development of exposure tools operating at 13.5 nm illumination were classified under the category of extreme UV (EUV); development had been initiated by the EUV LLC then but afterwards it had progressed at a very slow pace. The main limiting factor consisted in weakness of power sources capable only of operating in the few watts range as opposed to the hundreds of watts required to expose wafers fast enough to reach economically viable levels. Semiconductor manufacturers faced with these problems were compelled to expose several layers in the complex manufacturing process more than one time by using interlaced patterns to produce the required lines and spaces.

This workaround was quite viable to manufacture wafers but it implied continually increasing manufacturing costs, as more and more layers required multiple exposures; also, this solution implied acquiring more tools and manufacturing costs soon became inconsistent with cost targets of NAND memory producers. Under these conditions, memory manufacturers had to develop a workaround by stacking multiple memory cells on top of each other to reach density goals while avoiding using the most advanced and expensive lithography. Stacks of more than 100 memory cells have been successfully demonstrated but logic producers have not yet embraced this device stacking solution into manufacturing even though multiple research papers were presented in last 2-3 years on this subject.

A viable EUV source capable of delivering more than one hundred watts was introduced into several pilot lines in 2018. Last year about 30 EUV tools were delivered to high-volume manufacturing lines. In 2020, these tools are able to resolve lines and spaces with a 36 nm pitch on a single exposure. This means that this year (2020) the semiconductor industry is in the 18 nm generation using the correct definition of technology node. The industry misnamed this technology generation as the “5 nm” generation (See Section 1.2.2 for more detailed explanation).

1.2.2. Bringing the Node Nomenclature Back to Normal

From 1992 to present the methodology followed the NTRS, ITRS, and IRDS to name technology nodes was associated with the dimension of the smallest pitch typically utilized by the densest metal layer to be found in any integrated circuit. Figure ES27 illustrates the very first original definition of node: half pitch of the tightest metal layer. In the 70s, 80s and the most of the 90s the dimensions of the gate length and of the half-pitch of the tightest metal lines were essentially the same (Figure ES27); therefore this value was chosen as the node name since it conveyed with a single number the concept of density (i.e., half pitch of densest metal layer) and performance (i.e., shorter polysilicon gate implied faster transistors). Typically with the introduction of a new technology generation these dimensions were reduced to a value equal to 70% of the corresponding dimensions of the previous generation.

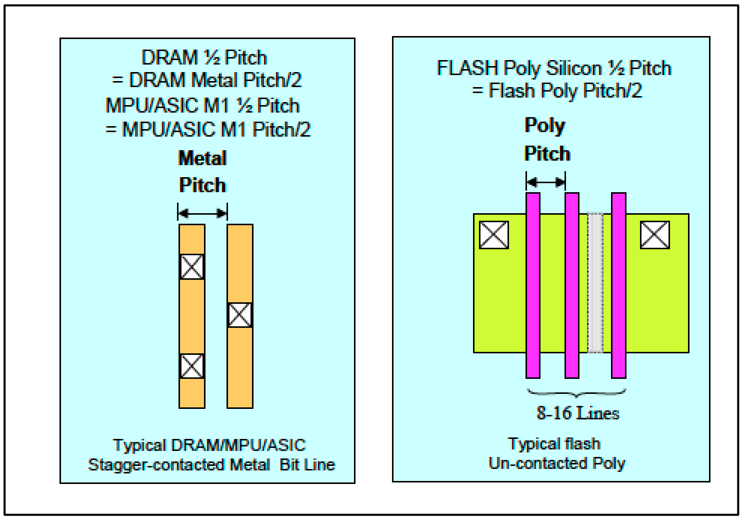
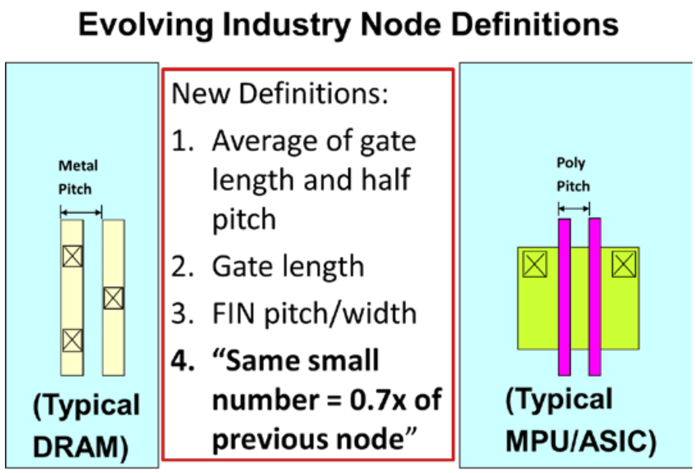


Figure ES27 Original technology node definition

In the second half of the 90s the adoption of the PC by the consumers imposed higher expectations of faster delivery and higher performance with any new product introduction. As a result, in the 90s the introduction of microprocessor technologies accelerated from a 3-4 year cycle to a 2-year cycle to respond to these consumers' demands. Furthermore, the length of the gate in any new technology was systematically reduced to 60% of the previous generation in order to produce faster transistors operating at higher frequencies. During this highly competitive time few companies began averaging the half-pitch dimension with the gate length dimension to define the name of the node associated with their newest technologies in order to attract the attention on more aggressive specifications than historical. Later on, some companies decided to use only the gate dimension to define the name of the technology node. Finally, the technology node definition became 70% of whatever the name of the node of the previous generation was! (See Figure ES28.)



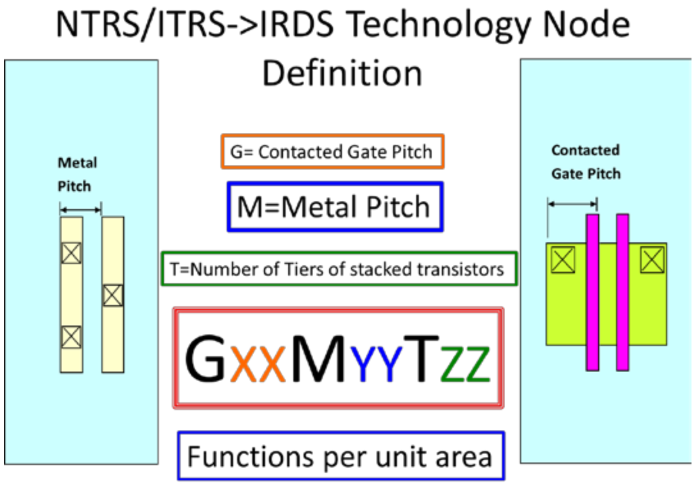
Evolving Industry Node Definitions

New Definitions:

1. Average of gate length and half pitch
2. Gate length
3. FIN pitch/width
4. **"Same small number = 0.7x of previous node"**

Figure ES28 Industry "adaptation" of technology node definition

This nomenclature has led to a complete detachment between IC features and technology nodes' names. In fact, there are companies nowadays announcing the introduction in the non-too-distant future of technologies below 1 nm in this decade!? It is therefore beneficial to go back to basics and revitalize the node definition to represent reality more closely. The IRDS has adopted a broader definition of node but still related to the NTRS and ITRS historical definition. This is illustrated in Figures ES29 and ES30.



NTRS/ITRS->IRDS Technology Node Definition

G= Contacted Gate Pitch

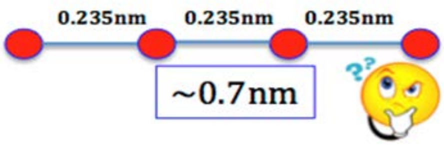
M=Metal Pitch

T=Number of Tiers of stacked transistors

GxxMyyTzz

Functions per unit area

Figure ES29 IRDS comprehensive technology node definition



YEAR OF PRODUCTION	2020	2022	2025	2028	2031	2034
	G48M36	G45M24	G42M20	G40M16	G38M16T2	G38M16T4
Logic industry "Node Range" labeling (nm)	"5"	"3"	"2.1"	"1.5"	"1.0 eq"	"0.7 eq"
IDM-Foundry node labeling	i7-f5	i5-f3	i3-f2.1	i2.1-f1.5	i1.5e-f1.0e	i1.0e-f0.7e
Logic device structure options	FinFET	FinFET LGAA	LGAA	LGAA	LGAA-3D	LGAA-3D
Mainstream device for logic	FinFET	FinFET	LGAA	LGAA	LGAA-3D	LGAA-3D
LOGIC TECHNOLOGY ANCHORS						
Patterning technology inflection for Mx interconnect	193i, EUV DP	193i, EUV DP	193i, EUV DP	193i, High-NA EUV	193i, High-NA EUV	193i, High-NA EUV
Beyond CMOS as complimentary to mainstream CMOS	-	-	-	2D Device, FeFET	2D Device, FeFET	2D Device, FeFET
Channel material technology inflection	SiGe25%	SiGe50%	SiGe50%	Ge, 2D Mat	Ge, 2D Mat	Ge, 2D Mat
Process, technology inflection	Conformal doping, Contact	Channel RMG	Lateral/ Atomic Etch	Non-Cu Mx	3D VLSI	3D VLSI
Stacking generation inflection	2D	3D stacking: W2W, D2W Mem-on-Logic	3D stacking: W2W, D2W Mem-on-Logic	3D stacking, Fine-pitch stacking, P-over-N, Mem-on-Logic	3D stacking, 3D VLSI: Mem-on-Logic with Interconnect	3D stacking, 3D VLSI: Logic-on-Logic

Notes: Mx—Tight-pitch routing metal interconnect IDM—independent device manufacturer FinFET—fin field-effect transistor LGAA—lateral gate all around EUV—extreme ultraviolet NA—numerical aperture Ge—germanium SiGe—silicon germanium RMG—replacement metal gate VLSI—very large scale integration W2W—wafer to wafer D2W—die to wafer Mem-on-Logic—memory on logic

Figure ES30 Technology node definition: IRDS vs. industry labeling

The top line of Figure ES30 refers to the year in which the technology is introduced into manufacturing. The second line indicates the key node attributes. G indicates the dimension of the contacted gated pitch while M indicates the dimension of the tightest metal pitch. ***The third line indicates the "industry labeling" of nodes that clearly appears completely devoted of any connection to reality***.

It should be noted once again that reduction of gate length dimension by itself has no longer a dominant influence on the performance of logic circuits. This is due to the severe power limitations that emerged in the middle of the previous decade. In order to limit power values to the 120-130 W range (i.e., no wafer cooling needed) it was no longer possible to increase operational frequency beyond the 5-6 GHz and therefore faster transistors no longer translated into to a higher operating frequency as shown earlier in Figure ES26. For this reason, semiconductor companies concentrated the transistor design effort on reducing power consumption instead of maximizing transistor speed since IC power dissipation had become such a major design constraint. However, reduction in contacted gate pitch is essential to increasing transistor density.

In few words, transistors are still getting faster generation-to-generation but not at the same rate than what used to be achieved in the 90s, since the major emphasis in transistor design has now shifted from speed to limiting power consumption.

It is however possible to fully reconnect with the historical definition of technology node defined by NTRS/ITRS if the value of the metal half pitch is once again used (Figure ES31) to identify a technology generation. It is clear from this picture that the semiconductor industry is not in the 7 nm or 10 nm generation but barely in the 18 nm generation!

No wonder Moore's Law will be still valid for the next 10 years once the real numbers of IC features are used for node definition; there is still a lot of room to run for scaling as a contributor to increasing transistor density!

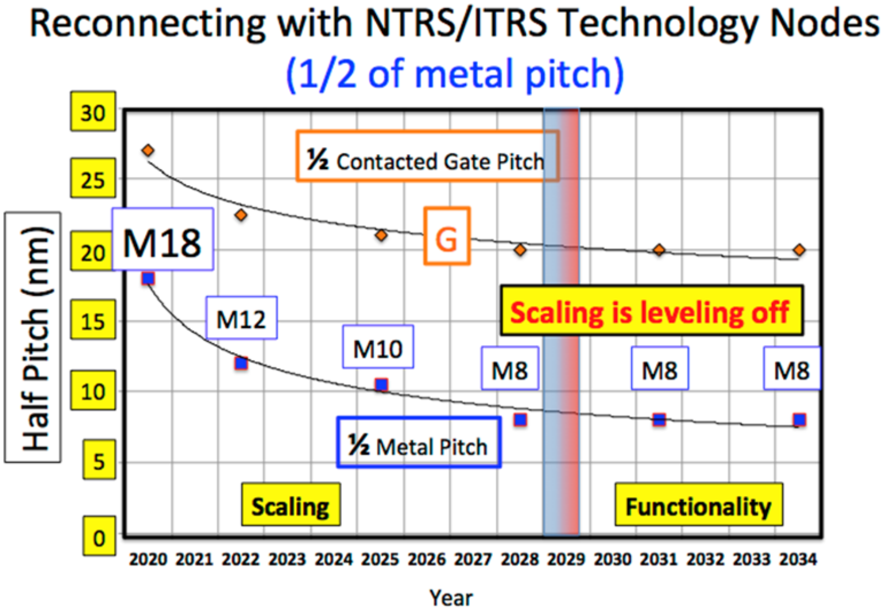


Figure ES31 Correct historical technology node definition and trend"

Figure ES31 shows also that feature scaling will reach fundamental limits of around 7-8 nm at the end of this decade. This prediction is consistent with forecasts of equipment technology leaders.[4] However, by early 2030 it is expected that quantum-computing technologies will begin to make real contributions to the advancement of the electronics industry. (See Section 1.4.)

1.2.3. 5G and Beyond Roadmap has become the "International Network Generations Roadmap"!

Cell phones began operation in the 90s using frequencies in the 800-900 MHz ranges in accordance with specifications of the Global System for Mobile Communications (GSM). These operational frequencies utilized by cell phones have increased multiple times and have now reached the present revision named 4G and LTE that operate in the 2,500-2,700 MHz ranges.

The adoption of a more powerful communication infrastructure under the name of 5G has been under discussion for the past few years. In 5G the utilization of frequencies ranging from 3 to 28 GHz and beyond had been under discussion for some time. In 2017 it became clear that 5G expectations and therefore its definition was quickly becoming by far much more complex that any of the previous transitions and therefore IEEE decided to launch a new network roadmap aimed at 5G.

In 2018 the working groups engaged in this IEEE network roadmap effort realized that the transition to 5G was no longer limited to the deployment of a communication system that introduced a new frequency spanning from 3.7 GHZ to 4.2 GHz for cell phones but it was much more than that.

5G defined a very broad new platform covering multiple aspects of communications. For instance, multiple bands operating in the 20−40 GHz and ~60 GHZ ranges were also proposed as additional elements of 5G. Therefore, the IEEE network

[4] *Martin van den Brink/President and CTO. ASML Keynote. "Continued scaling in semiconductor manufacturing enabled by advances in lithography". San Francisco, IEDM 2019.*

roadmapping effort was renamed the International Network Generations Roadmap (INGR) to encompass a much broader range of frequencies and novel network solutions. Close cooperation between IRDS and INGR has continued throughout the past two years.

As stated before, operation in these frequency ranges is still well within the capabilities of ICs. In the past 10 years, cell phones, portable PCs and many types of mobile appliances have become a viable means of accessing the Internet. Cell phone power consumption is typically below 5 watts, so this value is well within the thermal limits of IC operations. Most recently, access to the Internet via 4G-LTE or Wi-Fi has been continuously increasing since the areas of wireless coverage are continuously extending due to proliferation of hundreds of thousands cell towers and therefore mobile appliances have become the most convenient means of communication and access to any source of information anywhere at any time.

However, even though Wi-Fi and 4G-LTE have been developed with completely different market models and applications (i.e., wireless access to the Internet and cell phone, respectively) it is quite clear that both technologies are now interchangeably used and they are both contending for access to the same range of frequencies. In addition, distribution of TV programs, which early on relied exclusively on "wireless communications" is also contending for new network solutions. So cell phone companies, cable distributors and content providers are all contending for consumer attention. Is this a recipe for some type of unification and/or consolidation among all these business models? The question remains how will it be possible to reconcile different business models that support different applications.

1.2.4. Data Centers

The insatiable demand for information has led to the creation of gigantic clusters of servers and memory banks named "data centers" (Figure ES32). In this environment performance is still fundamental and using complex cooling systems can mitigate power issues. Power consumption of data centers is rapidly escalating into the hundreds of megawatts range. Communications within the data centers and for long distances is handled via fiber optics because of their stellar low rate of attenuation but traffic within data centers remains a bottleneck (Figure ES33).

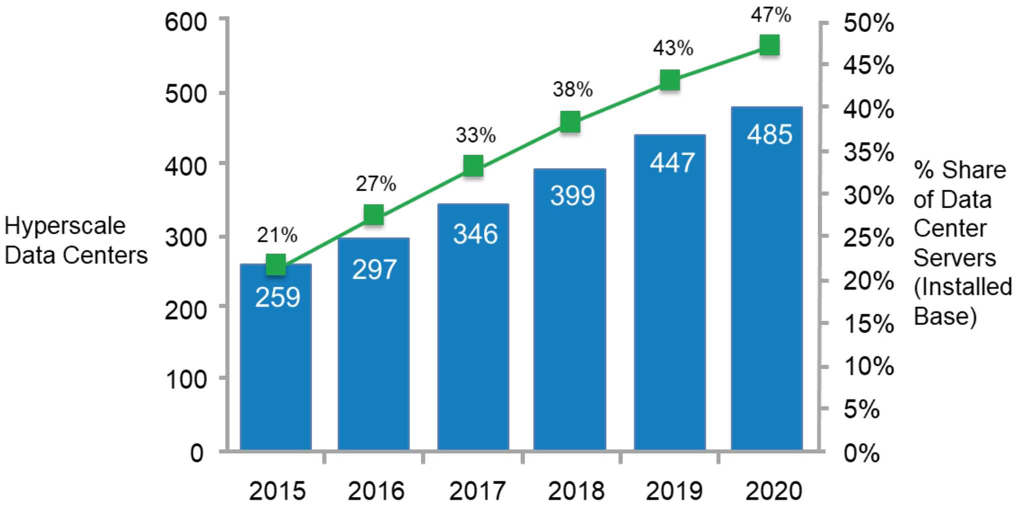


Figure ES32 Data center growth continues to accelerate

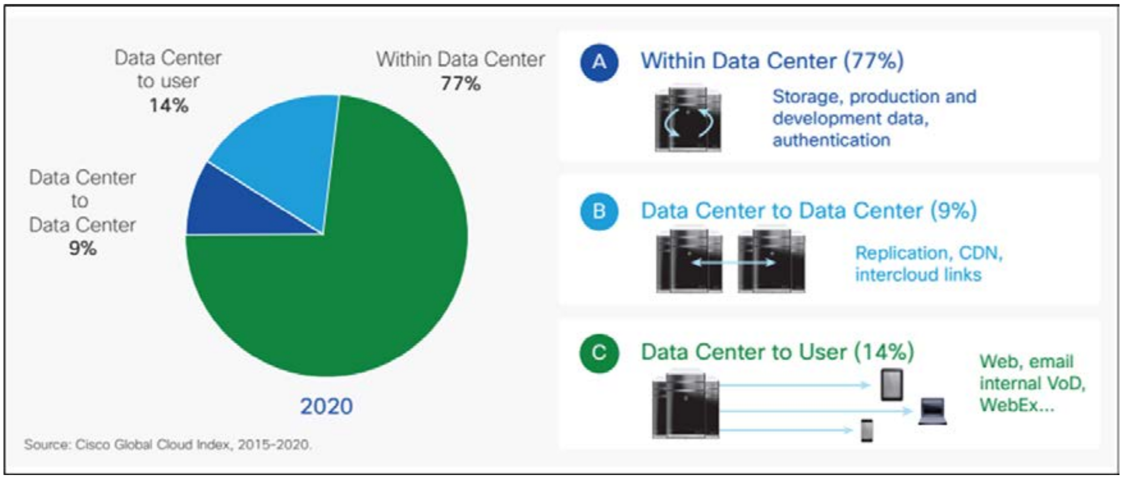


Figure ES33 Traffic within data center is main limiting factor

Adoption of multicore processors has also found the perfect application in servers used in data centers. In the past a separate bank of processors and memory aimed at a specific application had to be separately installed on a specific rack because the operating system required by the application was different from the operating system used for other applications. Under these conditions utilization of processors and memory devices was very inefficient. The advent of multicore processors, however, offered the opportunity to "host" different operating systems in each of the cores residing within the same microprocessor and therefore led to a dramatic increase in efficiency.

Since multiple applications were now residing within the same processor it followed that the rate at which the input of data could be handled by a single server was drastically increased; this led to higher requirements of the optical networks operating within a data center. Data center networks have gone from 1 Gigabit Ethernet (GbE) links out of a server rack just under a decade ago to 100 GbE today and 400 GbE already in the near term. To satisfy this requirement single mode fibers have been implemented in data centers.

1.2.5. PRODUCT CONFLUENCE AND TECHNOLOGY FUSION

In the past ten years the CMOS technology has evolved and continuously delivered generation-to-generation, reduced transistor size while burning less power/transistor. CMOS technology derivatives can be found from cell phone limited to 5-6 W power dissipation to data centers and large computers where power dissipation of few hundred watts can be managed. Will CMOS remain the only technology of choice for the foreseeable future? Different types of microprocessors operating at few gigahertz can be found in cell phones, in Wi-Fi and in large computers. Will 5G become the all-pervasive wireless technology of choice for cell phones, Internet devices and everything else? Similarly, the basic Von Neumann architecture where bits are moved back and forth between logic and memory still represent the architecture of choice.

The research community has been studying new logic and memory devices operating on completely new physical principles since the middle of the past decade. Similarly, new architectures have been explored for the past 10 years. Will tunnel transistors (TFET) and neuromorphic computing be the choice of the future? It is in any case clear that devices and systems can no longer be developed independently. New and different products are nowadays driving the growth of the electronics industry and therefore the ITRS, which had a technology-driven bottom up approach, had to evolve into the IRDS where application-driven, top-down requirements and bottom-up technology challenges are conjunctly harmonized.

1.2.6. SYSTEM INTEGRATION

It is generally impossible to find a single component that can provide all the required functionalities. This leads to the adoption of multiple subsystems that in the aggregate can provide the required functionalities. However, great care must be exercised in making all the subsystems work together. In general, system complexity increases very quickly as more subsystems are integrated, with cost and reliability requirements increasing as well. These considerations impose a severe limitation on which capabilities can be practically integrated in a single system at any point in time. However, as time goes by new technological capabilities become economically available so that higher functionality levels can be accomplished at affordable costs.

The main lesson out of these considerations consists in the fact that system integration is a dynamic process continually evolving. It would be therefore wrong to make absolute statements of what can or cannot be integrated in a system, and all these statements must be made in close reference to a very specific timeframe. What seemed impossible a few years ago may appear trivial nowadays.

In general it can be surmised that multi-components integration in packaging form is easier to realize but it is more costly to implement than monolithic or even heterogeneous integration whenever this technology can be realized in a cost-effective way. This observation explains why historically the two approaches continuously alternated to increase system functionality.

Nobody said it better than Gordon Moore in his 1965 foundational publication:[5]

No limits to increasing transistor count

> *"While to those of us in the semiconductor industry it sometime seems difficult to realize, there is no fundamental reason why the device yields are limited below 100%".*

But this is not the only approach.

> *"It may prove to be more economical to build large systems out of smaller functions, which are separately packaged and interconnected. The availability of large functions, combined with functional design and construction, should allow the manufacturer of large systems to design and construct a considerable variety of equipment both rapidly and economically."*

Many people have used these statements to suggest that multi-chip packaging solutions will become the ONLY way forward….. **but at this point the editor decided to remove the next sentence for space or other reasons and by so doing the whole meaning of the two previous sentences got completely distorted.**

> *"As far as the technologies for achieving large functions is concerned, several possibilities exist, any one of which is capable of being developed to the point that these arrays are feasible. It is not clear if one of these will dominate or if **a combination will be employed.**"*

From these statements it is clear that **either monolithic homogeneous or heterogeneous integration at the die level or heterogeneous integration at the package** level are just tools to be selected as appropriate on a case-by-case to optimize system performance and minimize cost.

Historically, the first practical cases of heterogeneous integration occurred early on at the board level. A printed circuit board (PCB) mechanically supports and electrically connects different electronic components using conductive tracks, pads and other features etched from copper sheets laminated onto a non-conductive substrate. Components (e.g., capacitors, resistors and several active devices) are generally soldered on the PCB. Advanced PCBs may contain also components embedded in the substrate.

PCB boards consist of non-conductive glass reinforced epoxy laminated sheets covered by a maze of conductive tracks. Special receptacles are provided to insert ICs and other components onto the PCB. Of course, multiple varieties of board and packaging technology exist.

Recently flip chip (FC) technology associated with 2.5D tight integration, and TSVs' 3D stack high bandwidth memory (HBM) has become available. All together these packaging technologies have allowed breaking the "Memory Wall" by doubling throughput and halving energy requirement like it has been demonstrated by Google tensor processing units (TPUs) before and after enabling use of HBM.

The key limiting factor in building any computing or communication system remains the speed of communication between logic and memory chips. It is well known that most of the total computing time is spent in memory. In addition, any time the signal travels from one chip to the other additional delays are introduced in the overall computing time. Therefore, reducing the delays introduced by the logic-memory connection is of paramount importance. Recently, new ways of connecting these die have been introduced by taking advantage of some of those very same advanced flip chip techniques mentioned before in conjunction with the more recently developed technology of through silicon vias (TSVs). In this case vias are dry etched into silicon wafers and then filled by electroplating with Cu to stack memory chips on top of logic chips into 3D stacks. The chips have a central array of up to 5,000 TSVs and are stacked by attaching the memory on top of the logic and connecting them to one another by means of an array of micro-pillars with a 50-micron pitch. A redistribution layer (RDL) is used to reroute connections to desired locations. For example, a bump array located in the center of a chip

[5] *https://www.chiphistory.org/20-moore-s-law-original-draft-1965*

can be redistributed to positions near the chip edge. The ability to redistribute points can enable higher contact density and enable subsequent packaging steps.

1.2.7. More than Moore and Packaging Integration = Heterogeneous Integration

Over the years, the scope of the ITRS was enlarged to include not only the CMOS-based digital domain for memory and microprocessor devices (driven by miniaturization as described by Moore's Law), but also heterogeneous integration of multi-functional analog and mixed-signal technologies for smart system applications ("More than Moore"). At the same time, the perspective of the roadmap shifted from being mostly technology driven to being increasingly determined by application requirements. In line with this, the ITRS evolved the International Roadmap for Devices and Systems.

The roadmapping effort has given rise to new insights in innovation methodology and strategy. This is in particular the case for "More than Moore", which requires a highly multidisciplinary R&D environment. It has become clear that progress in highly complex technology fields can only be achieved by cooperation along the complete innovation chain, which implies that multiple fields of expertise can be combined for the development of generic technology modules, which can be made available on open technology platforms. This trend is clearly demonstrated in the present developments in, e.g., the automotive industry and the medical domain.

The concept of "More than Moore" was introduced in the 2005 edition of the ITRS:

> "More than Moore" refers to the incorporation into devices of functionalities that do not necessarily scale according to Moore's Law, but provide additional value in different ways. The More-than-Moore approach allows for the *non-digital* functionalities (e.g., RF communication, power control, passive components, sensors, actuators) to migrate from the system board-level into the package (SiP) or onto the chip (SoC).

In addition, the increasingly intimate integration of complex embedded software into SoCs and SiPs means that software might also need to become a fabric under consideration that directly affects performance scaling.

The objective of More than Moore is to extend the use of the silicon-based technology developed in the microelectronics industry to provide new, non-digital functionalities. It often leverages the scaling capabilities derived from the More Moore developments to incorporate digital and non-digital functionality into compact systems and, eventually, system-of-systems.

The Packaging Contribution

The simplest way of integrating multiple dissimilar functions and technologies has historically occurred via packaging.

Even though multiple die located in the same package cavity have historically been connected via wire bonding due to the simplicity and low cost of this technology, it has been the use of FC that has really revolutionize (once again) the packaging industry in the past 10 years.

Flip chip is essentially a vertical and area array interconnect system using various fusible metallurgies (called "bumps") deposited on the bond pads of IC when they are still at the wafer level during what is called the back end of line (BEOL) processing. Separated and tested good chips with the bumps are *flipped* over a high density circuit board (called a *substrate*) and the bumps on the chip are aligned to corresponding bond pads on the substrate using high-precision/high-speed robots with machine vision. The metallurgies of the aligned bond pad-interconnect are then fused (*reflowed in a reducing atmosphere*) and alloyed to form an area array of robust joints.

Several advancements in FC technologies have contributed to making this approach more viable and cost effective than in the past. Earlier versions of flip chip technology had been developed in the 1960s and used to build multi-chip modules (MCMs) for mainframes

FC technology now has very broad applications ranging from original microprocessors as well as the latest SoCs, low cost GaAs power amplifiers (which enable fast data transfer by RF for Internet and video access from handhelds) to newer applications such as back-illuminated sensors (BIS) imaging chips, flat panel displays, lasers, LEDs, 3D stacked chips including logic chip and HBM memories.

All of the above systems specifications dictate the requirements for the semiconductor industry. (See Figure ES1 in the Introduction.)

1.3. Beyond Dimensional Scaling

As outlined in the previous paragraph the combination of 3D scaling both monolithically (SoC) and heterogeneously (SiP) will enable compacting more and more devices in a well-defined area/volume according to Moore's Law for another couple of decades. However, it is inevitable that physical limits of individual devices will eventually be reached due to some topological or electrically related limitations by the end of this decade.

How will it be possible then to continue to increase performance of either logic or memory operations when the number of physical devices can no longer be increased? It is clear that when overall 3D topological limitations are reached it will not be possible to increase performance by increasing the number of devices in an IC; but could performance be increased by any other means? Is it possible to increase performance without changing the physical size or the number of the devices that constitute an IC?

Once again, memory products are paving the way towards a new approach beyond 3D by providing invaluable clues to practical ways of solving this problem. For over 20 years producers of Flash memories have been studying how to store **more than one bit in a single memory cell**. In the past decade they were finally successful in finding a viable solution, and nowadays-Flash memory cells storing up to 4 bits have been successfully demonstrated. This memory cell solution points in the direction of storing and manipulating more than one bit in a single physical location as a possible way on continuously improving performance. This solution provides the additional benefit that device features can actually be substantially relaxed, alleviating the need for development of new expensive lithography tools capable of higher resolution. (Lithographic tools are the most expensive contributor to the total cost of the IC tool kit, demanding close to 40% of the total equipment cost of any leading-edge production line.) At present, it is not yet clear how more than one bit at a time could be simultaneously manipulated in a single logic device, but here is where the research community needs to rise up to the challenge. However, at least one possible solution using a multi-bit approach in a single physical element has been already demonstrated by a completely new way of performing logic operations: Quantum Information Processing!

1.4. Quantum Information Processing (QIP)

Quantum computing takes a very different approach to computation, relying on quantum bits, or qubits. In addition to representing a 0 or 1 as a conventional bit does, a qubit possesses two completely new features: first of all, a qubit can be in a quantum-mechanical superposition of 0 and 1 at the same time and secondly multiple qubits can be correlated through quantum entanglement. These novel features allow the use of massive quantum parallelism on a single quantum core. Although quantum mechanics is typically relevant only when describing behavior at the atomic level, it can apply to the behavior of superconducting circuits at extremely low temperatures, typically below 0.1 K.

It is important to realize that quantum computing does not represent a practical solution to all computational problems but offers the potential to carry out exponentially more efficient algorithms for several important problem classes. Devices for quantum computing are very different from conventional devices and fine-tuning their characteristics to avoid decoherence while organizing them effectively into scalable architectures has, to date, proved to be a formidable but not impossible engineering challenge.

Another approach utilizing coexistence of more than one bit in a physical location consists in quantum annealing. This technique, related to adiabatic quantum computing, is a computing approach in which binary variables are represented with qubits, each of which is initialized into a superposition of 0 and 1. The algorithm works by gradually adjusting an arrangement of programmable qubit fields and qubit-to-qubit interactions in a way that encodes an optimization problem defined by a cost function. Qubit states that correspond to a minimization of the cost function are most likely to emerge at the end of the algorithm, at which point the result can be read. However, hybrid cooperation with a classical computer is highly recommended.

Another approach to quantum computing, the quantum gate model, uses quantum logic gates to achieve a general-purpose quantum computer, essentially creating a quantum von Neumann architecture using quantum gates instead of classical gates. Potential applications include classically challenging computational problems, such as factoring large numbers. Recent advancements include a theoretical proof that the number of steps needed to solve certain linear algebra problems using parallel quantum circuits is independent of the problem size, whereas the number of steps grows logarithmically with problem size for classical circuits. Further applications include database search, portfolio optimization, machine learning, and combinatorial optimization. This so-called quantum advantage comes from the quantum correlations present in quantum circuits, but not present in classical circuits. At a scale of 50 sufficiently coherent qubits, known classical supercomputers are no longer able to simulate quantum computers.

It should however be pointed out that quantum computing will not replace classical computing, but it will work in conjunction with it since most of the results of quantum computing calculations will often need additional data manipulation to become practically useable. In addition, classical computers can accomplish several tasks and calculations in a much better way than any quantum computer. In fact, in multiple cases it is not clear how a quantum computer will ever equate the performance of classical computers. It is expected that this symbiotic relation will continue in the foreseeable future. For more information refer to the 2020 IRDS CEQIP Chapter.

2. Roadmap Process and Structure

2.1. Roadmap Process

The IRDS process is an evolution of the NTRS, ITRS, and ITRS 2.0 roadmapping process. The most relevant change consists in the fact that in the past device requirements were determined by the readily available technologies, and system integrators were left with very few options on how to assemble their products. However, with the advent of the fabless/foundry ecosystem the system integrators regained the leadership position in establishing device requirements. To adjust to this new environment the IRDS process has been strengthen by a close association with the IEEE Rebooting Computing Initiative and by adding Applications Benchmarking and Systems and Architectures requirements to the 2017 IRDS roadmap process.

The IRDS is kept current with a continuous review by the IFTs. Every other year reflects a full revision to the previous year's work. These "full revision" years, (usually odd numbered years (2017, 2019, etc.)) are followed by "update" years where additions and/or changes to IFT chapter table values are included to keep the roadmap current or to respond to industry input.

The 2020 IRDS is a revision to many of the 2017 or 2018 IFT chapters. All 2017 chapters were reviewed by the teams and many reflect revisions to either chapter narrative or table values, or both. The Beyond CMOS chapter now includes Emerging Research Materials information. Also, the updated 2020 More Moore chapter and tables' values reflect the most current trends of the industry. It should be noted that other 2020 IFT chapters do not reflect these recent updates as aligned to the More Moore IFT chapter. Alignment throughout the IRDS IFT chapters is the objective of the 2020 IRDS revision, a full revision year.

For the 2020 IRDS the Focus Teams are as follows:

1. Application Benchmarking (AB)
2. Systems and Architectures (SA)
3. Outside System Connectivity (OSC)
4. More Moore (MM)
5. Beyond CMOS (BC)
6. Cryogenic Electronics and Quantum Information Processing (CEQIP)
7. Packaging Integration (PI) white paper
8. Factory Integration (FI)
9. Lithography (L)
10. Yield Enhancement (YE)
11. Metrology (M)
12. More than Moore (MtM) white paper
13. Environment, Safety, Health, and Sustainability (ESH/S)

Additional roadmap documentation first introduced in 2018 includes market drivers. The 2018 IRDS included an update to Medical Devices Drivers, and added Automotive Drivers.

2.2. IRDS International Focus Teams (IFTs)

2.2.1. 2.2.1 Application Benchmarking (AB)

The mission of the Applications Benchmarking (AB) IFT in the 2020 IRDS is to update and identify key application drivers, and to track and roadmap the performance of these applications for the next 15 years. The output of the AB market drivers in conjunction with the drivers of the Systems and Architectures (SA) IFT, generates a cross-matrix map showing which application(s) are important or critical (gating) for each market.

Historically, applications drive much of the nanoelectronics industry. For example, 10 years ago the PC industry put pressure on semiconductor manufacturers to advance to the next node in the roadmap. Today, as applications shift to the mobile market, it is again system manufacturers that are applying pressure for new technologies. The market for *Internet* *of Things edge devices* (IoT-e) has its own set of requirements and needs, including low cost and low energy consumption. Most of this is discussed in the Systems and Architectures (SA) 2020 roadmap chapter. It is the function of the AB chapter to step back from the current markets and their needs, and to consider the current and near-future application needs in each of these markets.

2.2.2. 2.2.2 Systems and Architectures (SA)

The mission of the System Architecture (SA) chapter in the 2020 IRDS is to establish a top-down, system-driven roadmapping framework for key market drivers of the semiconductor industry in the 2019-2034 period. The SA chapter is proposing roadmaps of relevant system metrics for mobile applications, datacenter, IoT, and cyber-physical systems (CPS). The Systems and Architectures (SA) roadmap chapter of the 2020 IRDS roadmap constitutes a bridge between application benchmarks and component technologies. The systems analyzed in this chapter cover a broad range of applications of computing, electronics, and photonics. By studying each of these systems in detail, we can identify requirements for the semiconductor and photonics technologies that make these systems and applications possible.

The SA chapter considers four different types of systems: IoT edge (IoTe) devices provide sensing/actuation, computation, security, storage, and wireless communication. They are connected to physical systems and operate in wireless networks to gather, analyze, and react to events in the physical world. Cyber-physical systems provide real-time control for physical plants. Vehicles and industrial systems are examples of CPS. Mobile devices such as smartphones provide communication, interactive computation, storage, and security. For many people, smartphones provide their primary or only computing system. Cloud systems power data centers to perform transactions, provide multimedia, and analyze data. Cloud systems represent a trend towards common design principles and methodologies between traditional enterprise, high-performance scientific, and web-native computing.

2.2.3. Outside System Connectivity (OSC)

The mission of the OSC IFT in the 2020 IRDS consists in identifying and assessing capabilities needed to connect most elements of the Internet of Everything (IoE) and highlight technology needs and gaps. This includes supporting interconnection of a broad range of sensors, devices, and products to support information communication, processing and analysis for many applications including automobiles, aerospace, and a wide range of IoT applications for personal use, home, transportation, factory, and warehouse. Communication of data over fiber optic circuits to data centers and fiber optic communication within data centers is in scope for this chapter.

The 2020 IRDS OSC chapter is a review and update to these topics, including updates to several tables, such as for wavelength performance, data centers' requirements and potential solutions, optical interconnects for telecom, office and factory LAN, fiber to X, mobile device performance, and technology needs for compound semiconductor FET and bipolar transistors.

2.2.4. More Moore (MM)

The More Moore (MM) IFT of 2020 IRDS provided 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 (IoT and server) applications and forecasted logic and memory technologies (15 years) in main-stream/high-volume manufacturing (HVM). The 2013 ITRS already anticipated that fundamental limits of 2D scaling were going to be reached for all product lines between 2015 and 2021. Flash products have been the technology leaders in pitch scaling since the mid-70's and they have already overcome the 2D limitations by aggressively implementing 3D memory cell strictures—already 72-96 layers of Flash memory cells have been demonstrated. It is anticipated that logic technologies will transition to 3D approaches in the next few years. These technological solutions will assure continuation of Moore's Law for an additional 10−15 years.

The updated 2020 More Moore chapter and tables' values reflect these current trends of the industry. It should be noted that other 2020 IFT chapters do not reflect these recent updates. Alignment throughout the IRDS IFT chapters is the objective of the 2021 IRDS revision.

2.2.5. Beyond CMOS (BC)

The goal of the Beyond CMOS (BC) IFT of the 2020 IRDS is to survey, assess, and catalog viable new information processing devices and system architectures due to their relevance on technological choices. It is also important to identify the scientific/technological challenges gating their acceptance by the semiconductor industry as having acceptable risk for further development. Another goal is to pursue long-term alternative solutions to technologies addressed in More-than-Moore (MtM) entries. This is accomplished by addressing two technology-defining domains: 1) extending the functionality of the CMOS platform via heterogeneous integration of new technologies, and 2) stimulating invention of new information processing paradigms. It is important to notice that many new memory devices identified by BC in past roadmaps have already been successfully demonstrated and are making their way into manufacturing by means of heterogeneous monolithic integration. Refer to Figure ES49.

2.2.6. Cryogenic Electronics and Quantum Information Processing (CE&QIP)

The goal of this chapter for the 2020 IRDS is to survey, catalog, and assess the status of technologies in the areas of cryogenic electronics and quantum information processing. Application drivers are identified for sufficiently developed technologies and application needs are mapped as a function of time against projected capabilities to identify challenges requiring research and development effort.

Cryogenic electronics (also referred to as low-temperature electronics or cold electronics) is defined by operation at cryogenic temperatures (below –150 C or 123.15 K) and includes devices and circuits made from a variety of materials including insulators, conductors, semiconductors, superconductors, or topological materials. Existing and emerging applications are driving development of novel cryogenic electronic technologies.

Information processing refers to the input, transmission, storage, manipulation or processing, and output of data. Information processing systems to accomplish a specific function, in general, require several different interactive layers of technology. A top-down list of these layers begins with the required application or system function, leading to system architecture, micro- or nano-architecture, circuits, devices, and materials. A fundamental unit of information (e.g., a bit) is represented by a computational state variable, for example, the position of a bead in the ancient abacus calculator or the voltage (or charge) state of a node capacitance in CMOS logic. A binary computational state variable serves as the foundation for von Neumann computational system architectures that dominated conventional computing.

Quantum information processing is different in that it uses qubits, two-state quantum-mechanical systems that can be in coherent superpositions of both states at the same time, which can have computational advantages. Measurement of a qubit causes it to collapse to either one state or the other.

Technology categories covered in this report include:

- Superconductor electronics (SCE)
- Cryogenic semiconductor electronics (Cryo-Semi)
- Quantum information processing (QIP)

2.2.7. Packaging Integration (PI)

The Packaging Integration (PI) focus of the 2020 IRDS will be concentrated on packaging roadmap requirements and the introduction of many new requirements and potential solutions to meet market needs in the longer term. Packaging integration is the final manufacturing process transforming semiconductor devices into functional products for the end user. Packaging provides electrical connections for signal transmission, power input, and voltage control. It also provides for thermal dissipation and the physical protection required for reliability. In the past packaging was considered a limiting factor in both cost and performance for electronic systems. On the contrary packaging is now stimulating an acceleration of innovation. Heterogeneous integration of multiple technologies has become a dominant factor in the past 10 years enabling a variety of new products, especially in the mobile category. Design concepts, packaging architectures, materials, manufacturing processes, and systems integration technologies are all changing rapidly. This accelerated pace of innovation has resulted in development of several new technologies and the expansion and acceleration of others introduced in prior years. Most of the new technologies are an evolution of flip chip (FP) technologies originally developed more than 30 years ago by IDM. AT that time FC was too expensive for assembly only companies but nowadays Foundries have re-energize3d this FC technology with 2.5D and 3D combinations for instance to accelerate introduction of products when the IC technology is not fully ready yet. Wireless and mixed-signal devices, biochips, optoelectronics, and MEMS are placing new requirements on packaging and assembly as these diverse components are introduced as elements for System-in-Package (SiP) architectures.

2.2.8. Factory Integration (FI)

The Factory Integration (FI) focus area of 2020 IRDS is dedicated to ensuring that the semiconductor-manufacturing infrastructure contains the necessary components to produce items at affordable cost and high volume. Realizing the potential of Moore's Law requires taking full advantage of device feature size reductions, new materials, yield improvement to near 100%, wafer size increases, and other manufacturing productivity improvements. This in turn requires a factory system that can fully integrate additional factory components and utilize these components collectively to deliver items that meet specifications determined by other IRDS focus areas as well as cost, volume, and yield targets.

For the 2020 IRDS chapter, there is new content on several emerging future factory elements, such as smart manufacturing, big data, and security requirements.

2.2.9. Lithography (L)

The Lithography (L) focus area of patterning technology has been high-performance logic chips, DRAM memory, and Flash memory. High performance logic chips are now the drivers for better resolution features, Extreme Ultraviolet Lithography (EUVL) is now being implemented into manufacturing for leading edge logic due to the development cycle time, the manufacturing cycle time, the increased numbers of patterning levels and the overall complexity of extending multiple patterning to higher multiples. DRAM memory is trailing high performance logic in critical dimensions and in using EUVL. Flash memory has switched from scaling horizontally to stacking vertically and its patterning challenges relate to cost and to finding processes that reduce process steps. Consequently, nanoimprint patterning is being qualified now for potential Flash manufacturing in 2019. Multi electron beam direct write and directed self-assembly are patterning techniques that are under development but farther from manufacturing use than EUVL or nanoimprint. In the short term, that is for the next 8 years of so, logic critical dimensions will keep shrinking and DRAM dimension will continue to shrink but will trail logic devices in minimum resolution. Resolution is available for all the projected line and space patterns, assuming EUVL double patterning. Line and space patterns will have challenges of defects, of inspection overlay and particularly of managing stochastic issues. Stochastic issues arise from the random placement of molecules, from the random placement of chemical reactions in the photoresist and from the random arrival of photons during the imaging process. They cause CD variation, feature roughness and critical pattern defects. Critical dimensions of features printed with EUVL are small enough now that stochastic effects are a substantial fraction of tolerance budgets.

Stochastics will become a bigger issue as dimensions continue to shrink. Contact holes and other hole like patterns have all the issues lines and spaces do and, in addition, have resolution challenges. After two more logic nodes, their dimensions will require something better than EUVL double patterning. In the long term, when logic scaling is done by stacking vertical instead of by shrinking dimension, the challenges will relate to yield, process step consolidation, etch and deposition and possibly process cost and overlay over topography.

2.2.10. Yield Enhancement (YE)

The Yield Enhancement (YE) focus area is dedicated to activities ensuring that semiconductor manufacturing is optimized for production of the maximum number of functional units. Identifying, reducing, and avoiding relevant defects and contamination that can adversely affect and reduce overall product output are necessary to accomplish this goal. Yield in most industries has been defined as the number of functional and sealable products made divided by the number of products that can be potentially made. In the semiconductor industry, yield is represented by the functionality and reliability of integrated circuits produced on the wafer surfaces. During the manufacturing of ICs yield loss is caused, for example, by defects, faults, process variations, and design. The relationship of defects and yield, and an appropriate yield to defect correlation, is critical for yield enhancement. The Yield chapter of 2020 IRDS presents the current advanced and next generation future requirements for high-yielding manufacturing of More Moore as well as More than Moore products separated in “critical process groups" including MEMS, and back-end processes (e.g., packaging). Consequently, an inclusion of material specifications for Si, SiC, GaN, etc. is considered.

The 2020 IRDS chapter includes two white papers, entitled as follows: “Proactive Particle Control in Ultrapure Water (UPW) in Silicon Wafer Cleaning Process” and “Metal Contamination of Image sensors by Ultrapure Water in Silicon Wafer Cleaning Process.” Additionally, general updates for the chapter and the Table YE3 for Technology Requirements for Wafer Environmental Contamination Control is included.

2.2.11. Metrology (M)

The Metrology Chapter (M) of the 2020 IRDS identifies emerging measurement challenges from devices, systems, and integration in the semiconductor industry and describes research and development pathways for overcoming them. The 2020 Edition for the Metrology Chapter includes, but is not limited to, measurement needs for extending CMOS, beyond CMOS technologies, novel communication devices, sensors and transducers, materials characterization and structure/function relationships. This also includes metrology required in research and development, and techniques providing process control in manufacturing, yield, and failure analysis. With device feature sizes projected to decrease to less than 5 nanometers within the next 10 years, scaling as we know it is expected to soon reach its physical limits or get to a point where cost and reliability issues far outweigh the benefits. Already transistors for chips at the 7-5 nm nodes have already been demonstrated. The adoption of complex 3D structures fabricated using new materials and processes with ever decreasing dimensions are also projected to make their way into manufacturing within the next few years. The metrology roadmap addresses some of the measurement science challenges caused by these new developments and aims to provide a long-term view of the challenges, potentials solutions, technology, tools, and infrastructure needed to characterize new devices and materials for process control, and manufacturability.

2.2.12. Environment, Safety, Health and Sustainability (ESH/S)

The Environmental, Safety, Health, and Sustainability (ESH/S) chapter was not updated for 2020.That chapter serves to provide a long-range framework and process for all key stakeholders in the semiconductor and microelectronics industry, to develop proactive technical solutions to address critical ESH/S challenges up front, without gating industry R&D, mitigating cost, ensuring business continuity, and identifying key new markets and opportunities. The 2017 ESH/S chapter reflects that *transitional nature of the technology roadmap itself*, from the previous ITRS to the new scope and vision of the IRDS. Reflecting this fundamental shift, the 2017 edition of the ESH/S chapter is primarily grounded on the work done by the transitional team in 2016, to form a basis for a substantial rewrite of the next roadmap update in 2021. The 2017 ESH/S formally added the area of ‘sustainability,’ given the increasing constraints posed by natural resources (water, energy), and materials usage. It also includes the topic of governance, in the context of how processes and systems are managed and reported. Note that broader sustainability topics typically included in standard reporting (such as fair labor practices and social responsibility that are not directly related to technology and operations), are considered out of the scope of this roadmap chapter.

2.2.13. More than Moore (MtM)

The objective of the present More than Moore white paper is to provide an overview of a number of technology/application areas that are representative for the More than Moore domain in the sense that they require multifunctional heterogeneous system solutions, rather than miniaturization of devices only. These are: Smart sensors, Smart energy, Energy harvesting and Wearable, flexible and printed electronics.

The IRDS International Focus Team (IFT) on More than Moore, as listed in the acknowledgments section has generated the content of this chapter. Extensive use has been made of the NEREID NanoElectronics Roadmap for Europe.[6]

“More than Moore” refers to the incorporation into devices of functionalities that do not necessarily scale according to Moore's Law, but provide additional value in different ways. The More-than-Moore approach allows for the *non-digital* functionalities (e.g., RF communication, power control, passive components, sensors, actuators) to migrate from the system board-level into the package (SiP) or onto the chip (SoC).

In addition, the increasingly intimate integration of complex embedded software into SoCs and SiPs means that software might also need to become a fabric under consideration that directly affects performance scaling. The objective of More-than-Moore is to extend the use of the silicon-based technology developed in the microelectronics industry to provide new, non-digital functionalities. It often leverages the scaling capabilities derived from the More Moore developments to incorporate digital and non-digital functionality into compact systems and, eventually, system-of-systems.

[6] *NEREID NanoElectronics Roadmap for Europe (2018), https://www.nereid-h2020.eu/roadmap.*

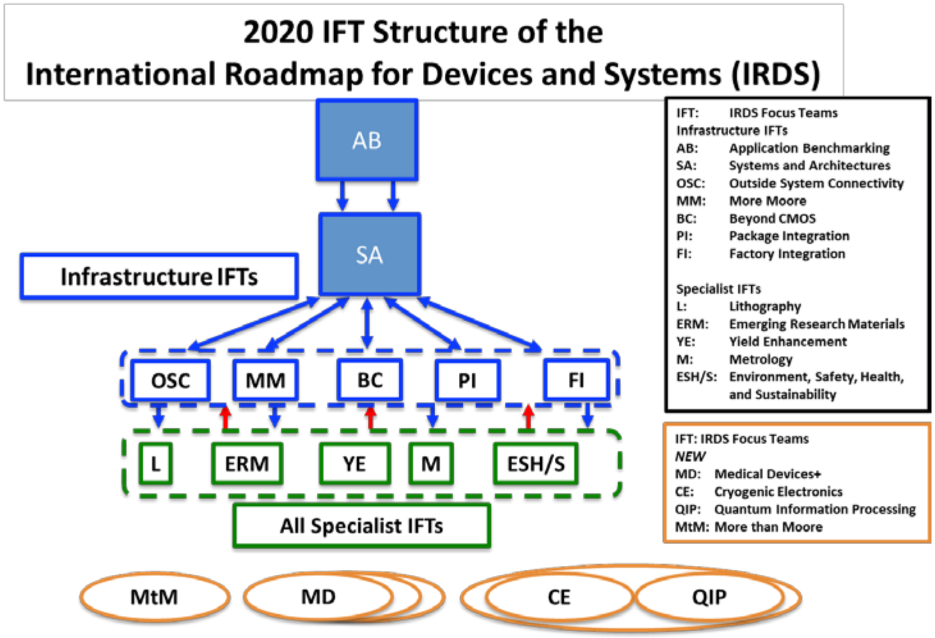


Figure ES34 IFT structure of the IRDS

3. Overall Roadmap Drivers—ORSC and ORTC

3.1. System Performance Considerations

System performance is determined by the harmonious confluence of hardware, architecture, and software algorithms. In the PC era any hardware upgrades easily "fit in" the well-established system architecture and dramatically improved system performance. The NTRS and the ITRS followed the impact of any new technology generation "up the food chain" to report continuous system performance improvements. However, the imbalance between the speeds at which processor and memory devices operated compelled the migration of ever-larger amount of cache memory on the processor chip.

However, this solution was not good enough, and further worsening the situation was the fact that superscalar microarchitectures were enabling higher frequencies though deeper pipelines. This meant more instructions needed to be "in flight" than was possible by waiting for branch instructions to execute. This led to *speculative execution*: predicting what path a program would take and then doing that work ahead of time, in parallel. Thus, higher frequencies meant deeper pipelines, which in turn required more and more speculatively executing instruction. But no prediction is 100% accurate. Invariably, these microprocessors did a lot of extra, wasted work by mis-speculation. The deeper the pipeline, the more power was wasted on these phantom instructions.

In the middle of the past decade, microprocessor's power dissipation reached fundamental limits and operational frequency could no longer be increased even though transistor performance could have easily allowed circuits to operate in the tens of GHz and above. These power limitations compelled a dramatic change in processor architecture to multicore in an attempt to partition data fetch and computation tasks among several cores that could then operate almost independently. Nowadays multiple new architectures are being explored and especially tailored for specific applications.

The restructuring of the roadmap process to ITRS 2.0 was initiated in 2014 and published early in 2016 (www.itrs2.net). In the same year, the roadmap was further restructured under IEEE and this process led to the 2017 IRDS and now to the 2020 IRDS. This IRDS is an attempt to formulate and harmonize system and device requirements in a comprehensive and synergistic way. This explains why new and more powerful system indicators needed to be generate in parallel to logic and memory indicators.

3.1.1. *Application Benchmarking and Systems and Architectures*

The new requirements outlined above led to the formation of the Application Benchmarking (AB) IFT and the expansion of the System Integration IFT to also include architectural elements and thus became the Systems and Architectures (SA)

IFT. Much more information on these subjects can be found in the AB and SA chapters but it is worth including one example that epitomizes the new comprehensive approach of the IRDS.

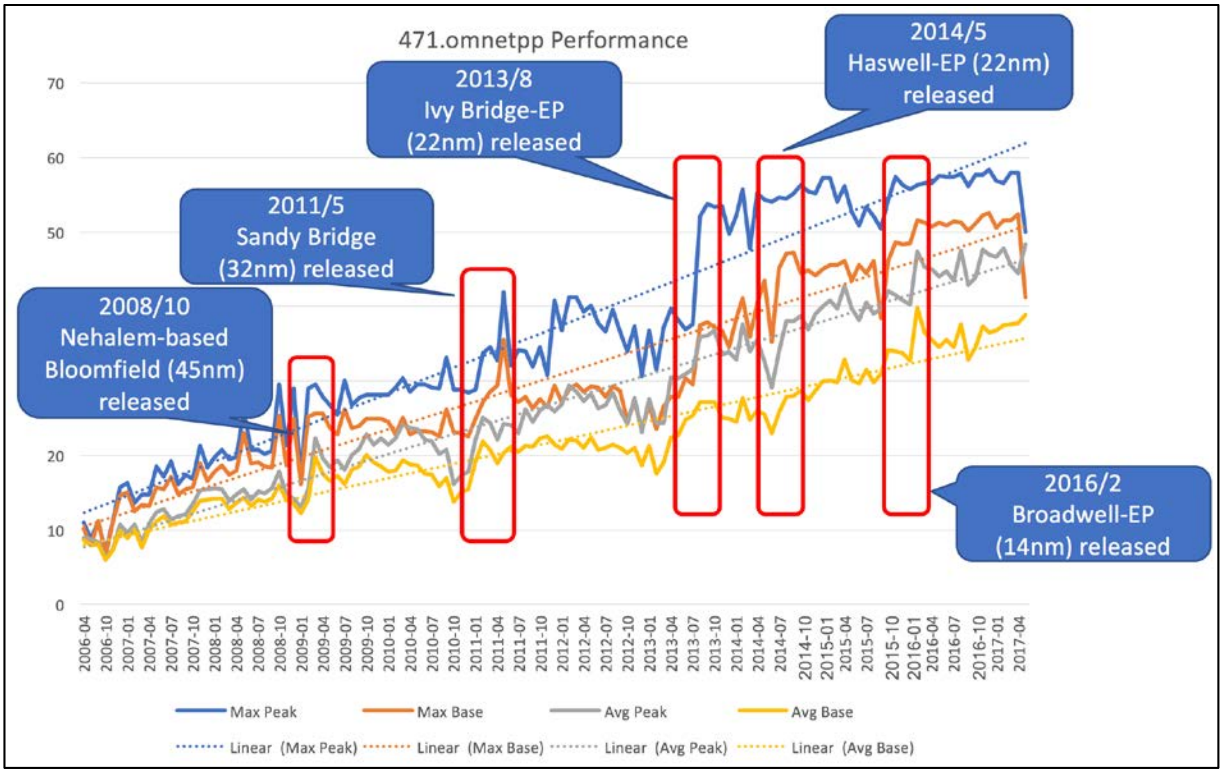


Figure AB-10 Historical performance over time of 471.omnetpp benchmark

Figure AB-10 from the Application Benchmarking chapter shows the performance of 471.omnetpp over time. Discrete Event Simulation (DES) models the operation of a system as a discrete sequence of events in time. The plot indicates both base and peak performance with max and average scores per monthly bins, represented by the four solid lines. The linear regression of each metric is also represented by the four dotted lines. In general, performance appears to be improving over time, but it should be noted that there are occasional steep jumps in performance. These generally align well with major CPU releases from Intel as depicted in the figure. One thing to note is that the data uploaded to SPEC only has the test date associated with the data, not the system release date. While, in theory, it is possible to look up all the system release dates for all 8,600 data points, we assumed that, in general, newer systems would be more popular to benchmark at any given time, and, thus, the test dates would align well with the system release dates.

The Ivy Bridge-EP processor released around this timeframe is the first processor to have a 25MB last-level cache, which is enough to fit the entire working set of 471.omnetpp. This could also explain why the max peak performance of the benchmark has remained somewhat flat after that point. Of course, the whole domain of DES should not be restricted to this single benchmark with a limited size input set. However, being memory bound and cache sensitive would be a general characteristic to describe DES workloads.

In addition, benchmarking and the trend prediction on “artificial intelligence (AI)/machine learning (ML)” is one of the important topics, which is discussed in the Application Benchmarking chapter. While AI/ML algorithms have been around for many years, the specific class of deep neural networks have found commercial application on a wide variety of workloads in the past decade, including image classification, object detection, speech recognition, machine translation, recommendation systems, sentiment analysis/classification of text, language modeling, text-to-speech, face identification, image segmentation, and image enhancement. Compute-in-memory (CIM) is a paradigm that achieves the parallel multiply-accumulate capabilities of the “Analog AI” techniques. Instead, digital memory devices such as SRAM or a compact NVM such as RRAM are used to encode multi-bit synapses in a customized memory array, which is then designed to perform some or all of the multiply-accumulate operations needed for AI/ML inference (and training) during the actual memory read. The advantages of SRAM-based CIM macros are that the high-endurance of SRAM allows small macros to be reused across many different weight values, so that macro size does not need to match the total size of the AI/ML model; the advantages of RRAM-based CIM macros are, at least eventually, in higher density (e.g., lower area-per-bit).

Based on some publicly available data on raw computational capabilities of various systems either released or announced for training and inference of DNNs, Figure ES35 shows how TOPS (data points) and the resulting TOPS/W (assessed against the green dotted lines) vary as a function of reported system power in (a) the training and (b) the inference. In inference, the performances are mostly on the trend of the equivalent performance vs. power consumption as of 1 Top/W.

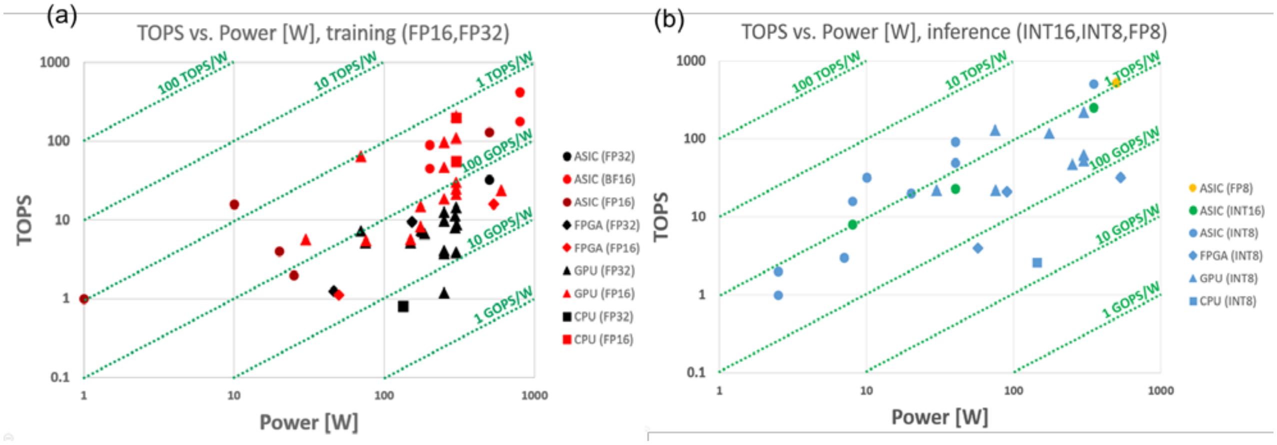


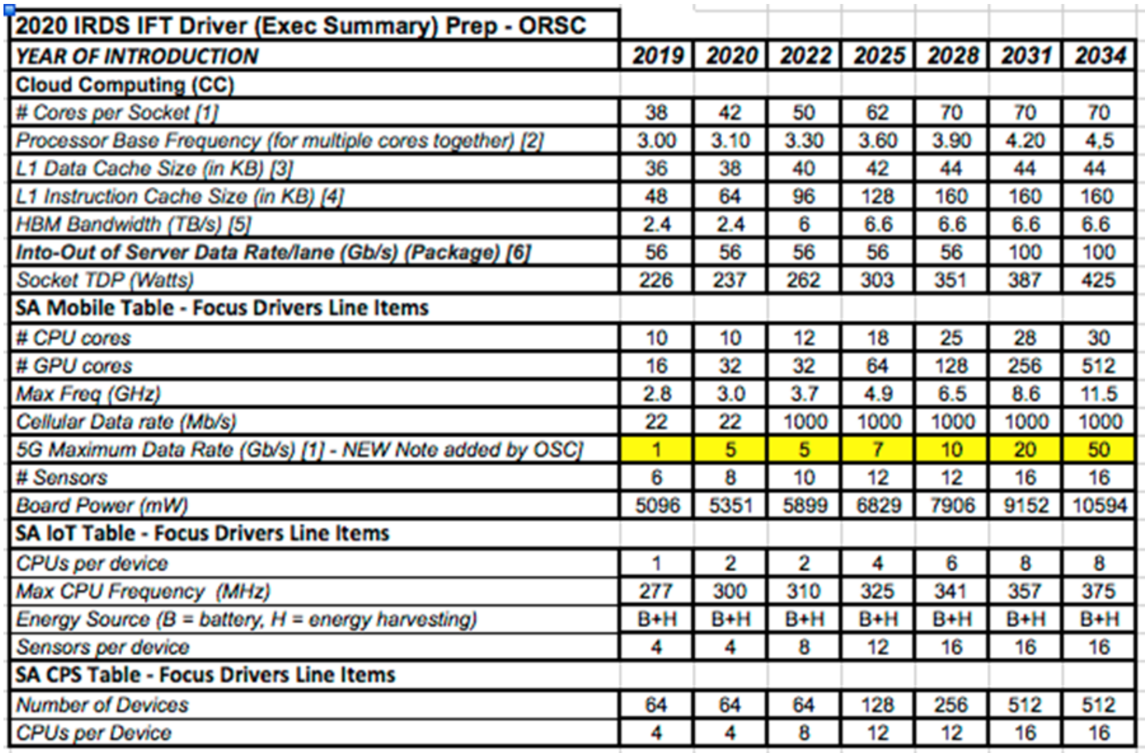
Figure ES35 TOPS (data points) and the resulting TOPS/W (assessed against the green dotted lines) as a function of reported system power in (a) the training and (b) the inference.

3.2. Overall Roadmap Systems and Technology Characteristics (ORSC and ORTC)

Providing a simple top-down view of the new very complex ecosystem it is not an easy task and it is therefore necessary to have a simplified depiction of what is under study including some basic definitions. The 2018 IRDS developed several summary tables to capture all these elements and the reader will be able to find detailed information in the chapters addressing these subjects. This is reflected in the 2020 Edition.

The executive summary presents a succinct overview of the above elements progressing from system requirement all the way to device specifications.

Table ES1 Overall Roadmap System Characteristics



2020 IRDS IFT Driver (Exec Summary) Prep - ORSC							
YEAR OF INTRODUCTION	***2019***	***2020***	***2022***	***2025***	***2028***	***2031***	***2034***
Cloud Computing (CC)							
# Cores per Socket [1]	38	42	50	62	70	70	70
Processor Base Frequency (for multiple cores together) [2]	3.00	3.10	3.30	3.60	3.90	4.20	4,5
L1 Data Cache Size (in KB) [3]	36	38	40	42	44	44	44
L1 Instruction Cache Size (in KB) [4]	48	64	96	128	160	160	160
HBM Bandwidth (TB/s) [5]	2.4	2.4	6	6.6	6.6	6.6	6.6
Into-Out of Server Data Rate/lane (Gb/s) (Package) [6]	56	56	56	56	56	100	100
Socket TDP (Watts)	226	237	262	303	351	387	425
SA Mobile Table - Focus Drivers Line Items							
# CPU cores	10	10	12	18	25	28	30
# GPU cores	16	32	32	64	128	256	512
Max Freq (GHz)	2.8	3.0	3.7	4.9	6.5	8.6	11.5
Cellular Data rate (Mb/s)	22	22	1000	1000	1000	1000	1000
5G Maximum Data Rate (Gb/s) [1] - NEW Note added by OSC]	1	5	5	7	10	20	50
# Sensors	6	8	10	12	12	16	16
Board Power (mW)	5096	5351	5899	6829	7906	9152	10594
SA IoT Table - Focus Drivers Line Items							
CPUs per device	1	2	2	4	6	8	8
Max CPU Frequency (MHz)	277	300	310	325	341	357	375
Energy Source (B = battery, H = energy harvesting)	B+H	B+H	B+H	B+H	B+H	B+H	B+H
Sensors per device	4	4	8	12	16	16	16
SA CPS Table - Focus Drivers Line Items							
Number of Devices	64	64	64	128	256	512	512
CPUs per Device	4	4	8	12	12	16	16

Notes for Table ES1:

[1] Required for specintrate scaling

[2] Frequency increase slowing down, but increases because of better cooling (allowing higher TDP))

[3] Load-to-use latency dictates constant or limited growth in L1 data cache size

[4] Instruction footprint for cloud apps going up (refer Google data warehouse paper)

[5] HBM: 128GB/s per port, in sockets 2015, HBM2: 256GB/s per port, can be in sockets 2017, HBM3: 512GB/s, can be in sockets 2019

[6] A lane is defined as:

a) Single wavelength and polarization that is received by a single detector;

b) Pulse Amplitude Modulation with 4 amplitude levels (PAM4) would a single lane;

c) Input (IQ) and Output (OQ) can be independent lanes;

d) Multiple wavelengths in a fiber would be multiple lanes:

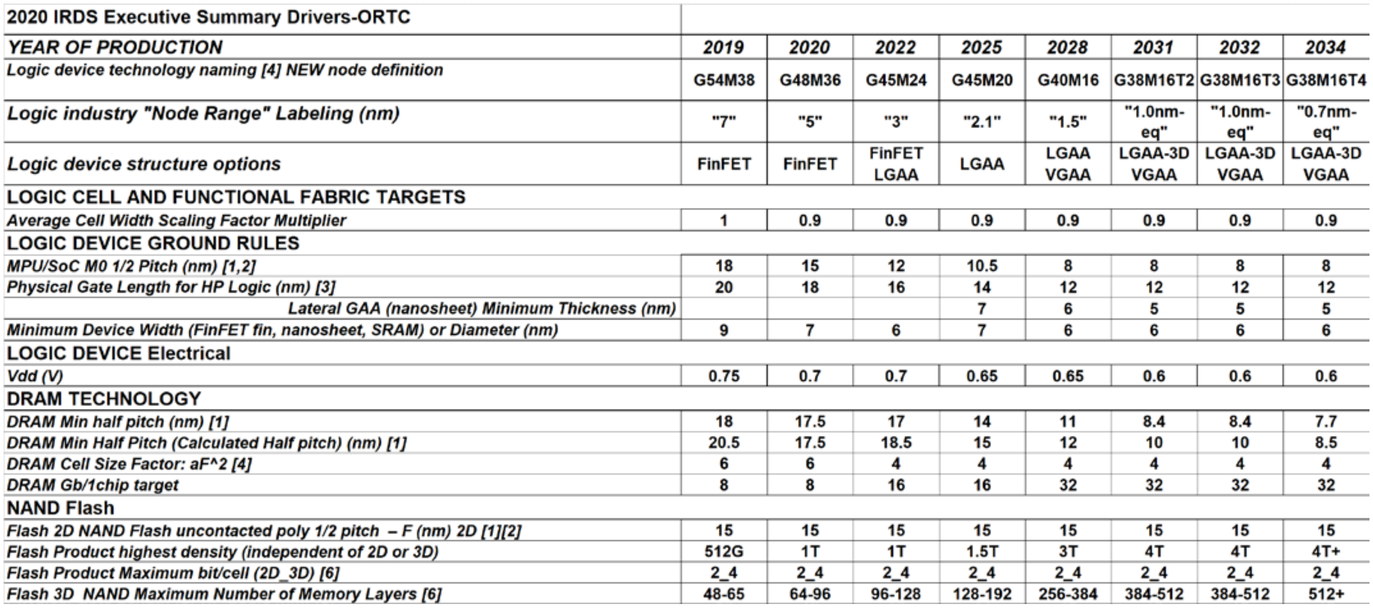
e) If multiple lanes are required to support the required data rate, circuitry will be required to separate the data stream into multiple lanes and recombine in the receiver.

This Table ES1 summarizes some of the major system characteristics in cloud computing, mobile, IoT, and cyber-physical systems.

It can be noticed that in all cases the number of CPU or GPU cores continues to increase throughout the time horizon. In all cases it is expected that the amount of data elaboration will continue to increase and actually it does not seem that there is any limit on these requirements. Frequency of operation will continue to increase as well but only at a moderate rate except for mobile systems where the strong demand for wider bandwidth can only be satisfied if the operational frequency is increased. It may be observed however that this increase in frequency can only be tolerated if power dissipation is kept at very low levels by careful power management.

Table ES2 summarizes the major technology characteristics of logic and memory devices. For convenience, the traditional industry "Node Range" Labeling is indicated even though in the past it only closely tracked the half-pitch of nonvolatile memory (NVM) products. The NTRS/ITRS definition of technology naming was based on half-pitch of gate (traditionally polysilicon based) and metal half pitch (Figure ES27). This definition can be easily connected with the IRDS definition using metal pitch (Figure ES29).

Table ES2 Overall Roadmap Technology Characteristics



2020 IRDS Executive Summary Drivers-ORTC								
YEAR OF PRODUCTION	*2019*	*2020*	*2022*	*2025*	*2028*	*2031*	*2032*	*2034*
Logic device technology naming [4] NEW node definition	G54M38	G48M36	G45M24	G45M20	G40M16	G38M16T2	G38M16T3	G38M16T4
Logic industry "Node Range" Labeling (nm)	"7"	"5"	"3"	"2.1"	"1.5"	"1.0nm-eq"	"1.0nm-eq"	"0.7nm-eq"
Logic device structure options	FinFET	FinFET	FinFET LGAA	LGAA	LGAA VGAA	LGAA-3D VGAA	LGAA-3D VGAA	LGAA-3D VGAA
LOGIC CELL AND FUNCTIONAL FABRIC TARGETS								
Average Cell Width Scaling Factor Multiplier	1	0.9	0.9	0.9	0.9	0.9	0.9	0.9
LOGIC DEVICE GROUND RULES								
MPU/SoC M0 1/2 Pitch (nm) [1,2]	18	15	12	10.5	8	8	8	8
Physical Gate Length for HP Logic (nm) [3]	20	18	16	14	12	12	12	12
Lateral GAA (nanosheet) Minimum Thickness (nm)				7	6	5	5	5
Minimum Device Width (FinFET fin, nanosheet, SRAM) or Diameter (nm)	9	7	6	7	6	6	6	6
LOGIC DEVICE Electrical								
Vdd (V)	0.75	0.7	0.7	0.65	0.65	0.6	0.6	0.6
DRAM TECHNOLOGY								
DRAM Min half pitch (nm) [1]	18	17.5	17	14	11	8.4	8.4	7.7
DRAM Min Half Pitch (Calculated Half pitch) (nm) [1]	20.5	17.5	18.5	15	12	10	10	8.5
DRAM Cell Size Factor: aF^2 [4]	6	6	4	4	4	4	4	4
DRAM Gb/1chip target	8	8	16	16	32	32	32	32
NAND Flash								
Flash 2D NAND Flash uncontacted poly 1/2 pitch – F (nm) 2D [1][2]	15	15	15	15	15	15	15	15
Flash Product highest density (independent of 2D or 3D)	512G	1T	1T	1.5T	3T	4T	4T	4T+
Flash Product Maximum bit/cell (2D_3D) [6]	2_4	2_4	2_4	2_4	2_4	2_4	2_4	2_4
Flash 3D NAND Maximum Number of Memory Layers [6]	48-65	64-96	96-128	128-192	256-384	384-512	384-512	512+

Notes for Table ES2:

Logic Notes

[1] Industry naming convention (x0.7 with respect to earlier node) showing a full PPA (performance-power-area) gain from node cycle to another where SoC area typically scales with a 0.55-0.70x factor.

[2]Horizontal local interconnect (M0) pitch. This is in most of cases the tightest metal pitch in a SoC enabling the scaling of standard cell height. M0 pitch follows a scaling factor of 0.7-0.85x. When this scaling factor combined with the other scaling factors coming from the contacted poly pitch and design area scaling provide a full PPA (performance-power-area) gain to the next node where SoC area typically scales with a factor of 0.55-0.70x.

[3] Defined as distance between metallurgical source/drain junctions

[4] GxxMxxTx notation refers to Gxx: contacted gate pitch, Mxx: tightest metal pitch in nm, Tx: number of tiers. This notation illustrates the technology pitch scaling capability. On top of pitch scaling there are other elements such as cell height, fin depopulation, DTCO constructs, 3D integration, etc. that define the target area scaling (gates/mm2).

DRAM Notes

[1]The definition of DRAM Half pitch has been changed from this edition. Because of 6F2 DRAM cell, BL pitch is no more critical dimension. Calculated half pitch is use the following equation "Calculated half pitch=(Cell Area/ Call size factor)^0.5." Critical dimension for process development, the Minimum half pitch is also introduced. Currently Active area (long rectangle island shape) half pitch is the critical dimension of 6F2 DRAM.

Critical dimension for process development, the Minimum half pitch is also introduced. Currently Active area (long rectangle island shape) half pitch is the critical dimension of 6F2 DRAM.

Calculated half pitch is use the following equation "Calculated half pitch=(Cell Area/ Call size factor)^0.5." Critical dimension for process development, the Minimum half pitch is also introduced.

Currently Active area (long rectangle island shape) half pitch is the critical dimension of 6F2 DRAM.

[4] Cell size factor = a = (DRAM cell size/F2), where F is the DRAM ½ pitch. The current values of a are 6 from 2009. And a=4 will be predicted in 2021.

NAND Flash Notes

[1] 2D NAND strings consist of closely packed polysilicon control gates (the Word Lines) that separate the source and drain of devices with no internal contact within the cell. Up to now this uncontacted word line pitch is still the tightest in all technologies.

[2] In order to gain high data bandwidth 3D NAND packs two bit lines within each cell x-y print space. These metal bit lines contact the array vertical channels in a complex and staggered fashion, resulting in tight metal (usually metal-2) pitch. Currently this tight contacted pitch is smaller than DRAM and Logic contacted metal pitches. But since 3D NAND x-y footprint is likely stay constant, other technologies may surpass 3D NAND in metal pitch in the future.

[6] The number of 3D layers spans a range since the same density product may be achieved by using smaller 1/2 pitch, bits per cell, and fewer layers, or larger 1/2 pitch, bits per cell, and more layers. The number of 3D layers is not a unique function, depending on the cell 1/2 pitch, number of bits per cell, and 3D NAND technology architecture chosen. Lower number of 3D layers generally has lower bit cost, but other factors such as decoding method, speed performance, easier or harder to get yield, also need to be considered.

4. Grand Challenges

4.1. In the Near-term

4.1.1. Application Benchmarking

4.1.1.1. Big Data Analytics

The gains in graph processing performance come from three main sources: 1) improvements in algorithms, 2) memory bandwidth, and 3) bandwidth. Processor performance does not have a first-order impact.

Some of the gains in traversed edges per second (TEPS) over the years have been due to improvements in the algorithm used. While we expect this factor to continue to have an impact, the improvements are likely to provide diminishing returns.

The most critical resources today are bandwidth and latency of the memory and the global network.

- Graph problems have a high ratio of communication to computation
- Very little locality

Memory bandwidth needs: the current top machines have an aggregate memory bandwidth of 5 petabytes per second.

- The next generation machines will attain about 20 petabytes per second; thanks to in-package DRAMs like high-bandwidth memory (HBM).

4.1.1.2. Feature Recognition

For near-term digital hardware, the critical hardware needs are the design of systolic computing units that align well with the deep neural network (DNN) algorithms; and, ability to receive and send data to large amounts of memory at high-bandwidth yet reasonable power.

Further improvements can be obtained for distributed training by highly efficient use of network bandwidth, allowing the multiple “workers” to collaborate with a minimal amount of information exchange.

4.1.2. Systems and Architectures

4.1.2.1. IoT Edge Devices

IoT edge devices must satisfy several stringent requirements. They must consume small amounts of energy for sensing, computation, security, and communication. They must be designed to operate with strong limits on their available bandwidth to the cloud.

Many IoT devices will include artificial intelligence (AI) capabilities, which may or may not include online supervision or unsupervised learning. These AI capabilities must be provided at very low-energy levels. A variety of AI-enabled products have been introduced. Several AI technologies may contribute to the growth of AI in IoT edge devices: convolutional neural networks; neuromorphic learning, and stochastic computing.

IoT edge devices must be designed to be secure, safe, and provide privacy for their operations.

4.1.2.2. Mobile Systems

Mobile systems present several challenges for system designers. Multimedia viewing, such as movies and live TV, have driven the specifications of mobile systems for many years. We have now reached many of the resolution limits of human perception, so increases in requirements on display resolution and other parameters will be limited in the future based on multimedia needs. However, augmented reality will motivate the need for advanced specifications for both input and output (I/O) in mobile devices. Mobile device buyers demand frequent, yearly product updates. This fast refresh rate influences design methodologies to provide rapid silicon design cycles; it can also suggest the use of programmability to provide a broad range of models on a given platform. Financial transactions are now performed using mobile devices. We expect this trend to grow, particularly in developing nations, where financial technology will leapfrog. Security and privacy are key concerns for mobile devices, particularly for financial transactions.

4.1.2.3. Cloud Applications and Systems

Cloud applications present several challenges for system designers. Data centers are starting to take advantage of heterogeneous core types, much as embedded systems have done for many years. System architects need to balance the performance improvements for chosen applications provided by specialized accelerators against the utilization of these specialized cores. The huge scale of problems in social networking and AI means that algorithms run at memory speed and that multiple processors are required to compute. The radius of useful locality—the distance over which programmers can use data as effectively local—is an important metric. We expect optical networking to greatly enlarge useful locality radius over the next few years. Memory bandwidth is a constraint on both core performance and number of cores per socket. Stacked memories, which are starting to come into commercial use, provide higher bandwidth memory connections. Thermal power dissipation continues to be an important limit.

Cloud systems present significant challenges. Heterogeneous architectures can provide more efficient computation of key functions. Novel memory systems, including stacked memories, offer high performance and lower power consumption. Advances in internal interconnect may create tipping points in system architecture. The term hyperconvergence is used to describe the point at which I/O speeds approach internal interconnect speeds.

4.1.3. Outside System Connectivity

4.1.3.1. RF Analog Technology

The key challenges for radio frequency (RF) are to achieve high-performance, energy-efficient RF analog technology compatible with CMOS processing and delivering capabilities to support a broad range of applications for IoT devices. To achieve high-performance RF with high-energy efficiency, CMOS gate resistance must be reduced with technologies that are compatible with CMOS processing. Furthermore, SiGe and III-V performance needs increased ft and fmax while being integrated with CMOS. Passive devices need to be integrated on CMOS with higher performance.

To enable systems and components to meet 5G performance requirements, we need devices to support <6 GHz massive multiple input multiple output (MIMO) and 28 GHz communication with low power and cost effectively, increasing energy efficiency of amplifiers while increasing operating frequency; systems to deliver high density communication without interference and with low noise, antennas to support multiple band communication in compact mobile devices, and low-cost high-efficiency directional antennas to support mmWave and massive MIMO 5G.

4.1.4. More Moore

4.1.4.1. Logic Device Scaling

Beyond 2022 a transition from FinFET 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.

FinFET and lateral GAA devices enable a higher drive at unit footprint if fin pitch can be aggressively scaled while increasing the fin height. 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. This trend in reducing the number of fins while balancing the drive with increased fin height is defined as fin depopulation strategy, which also simultaneously reduces the standard cell height, therefore the overall chip area.

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.

4.1.4.2. DRAM and 3D NAND Flash Memory

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 metal-insulator-metal (MIM) capacitors have been adopted using high-κ ($ZrO_2/Al_2O/ZrO_2$) as the capacitor of 40–30 nm half-pitch DRAM. This material evolution and improvement will continue until 20 nm high-performance (HP) and ultra-high-κ (perovskite $\kappa > 50 \sim 100$) materials 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 3D structure will be changed from cylinder to pillar shape.

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 ~20 nm 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. Also, the depositing of many layers of dielectric and/or polysilicon, as well as metrology for multilayer films and deep holes, all constitute challenge in unfamiliar territories. These all translate to large investment in new equipment and floor space and new challenges for wafer flow and yield.

4.1.5. Emerging Research Materials

4.1.5.1. Materials for Logic Device Scaling

Materials and processes that achieve performance and power scaling of lateral fin- and nanowire FETs (Si, SiGe, Ge, III-V) are as follows:

- Integrated high-κ dielectrics with equivalent oxide thickness (EOT) <0.5 nm and low leakage
- Integrated contact structures that have ultralow contact resistivity
- Achieving high-hole mobility in III-V materials in FET structures
- Achieving high electron mobility in Ge with low-contact resistivity in FET structures
- Processes for achieving low dislocations and anti-phase boundary generating interface between Ge/III-V channel materials and Si
- Dopant placement and activation, i.e., deterministic doping with desired number at precise location for V_{th} control and source/drain (S/D) formation in Si as well as alternate materials.

4.1.5.2. Materials for Copper Interconnect

Materials and processes that improve copper interconnect resistance and reliability are as follows:

- 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.
- 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.
- 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.
- Reduction of the κ value of intermetal dielectrics. Reduction of the interlevel dielectric (ILD) κ value is slowing down because of problems with manufacturability. The poor mechanical strength and adhesion properties of low-κ materials are obstructing their incorporation.

4.1.6. Beyond CMOS

4.1.6.1. Emerging Memories and Logic Devices

One of the grand challenges is to realize scaled volatile and nonvolatile memory technologies to replace SRAM and NAND flash memory in appropriate applications. The key components of such emerging memories are novel memory devices and selector devices.

Another grand challenge is to extend ultimately scaled CMOS as a platform technology into new domains of applications. The emerging logic and information processing devices will be extended CMOS devices and/or beyond CMOS devices.

4.1.7. Lithography

4.1.7.1. EUV Lithography for 7 nm Node Logic and Beyond

With the successful implementation of EUV in manufacturing, patterning challenges for logic and DRAM have shifted from resolution to noise, defects, overlay and edge placement. For flash memory the challenges are cost and demonstrating nanoimprint lithography with sufficiently low defects and cost.

Some of the defect challenges relate to keeping masks clean. Although pellicles are available, their transmission is low, thereby reducing exposure tools throughput significantly. Actinic patterned mask inspection tools have recently been introduced, addressing a problem that has long been without a satisfactory solution. Other defects are due to what are called stochastics, which are random variations in light exposure and in resist chemistry.

Stochastic defects come from random variations in the number of photons in a discrete exposure of a small area and come from the random placement, reaction and dissolution of the various molecular components that make up photoresist. These defects can take the form of bridging between lines, missing contact holes, line opens or merged contact holes. Recent work has shown that they are actually more common than simple extrapolation of CD variation assuming a normal distribution would predict. These sorts of defects currently limit the usable resolution of EUV tools. There are fewer such defects with slower resist, so EUV users typically use slower resists than they would like. A slower resist is one that requires a higher exposure dose to define the desired pattern. This results in decreased exposure tool throughput and more expensive exposures. Long term, the need for slower resists is expected to drive the development higher power light sources and/or a more efficient optical train in exposure tools.

4.1.8. Packaging Integration

4.1.8.1. Packaging Technology

The challenges in 3D and 2.5D packaging are as follows:

- Materials and process for through silicon vias (TSVs) compatible with silicon
- Improve process for die stack compatible with future shrinks
- Head extraction form die stack
- Dense planer (2.5D) bridge to fill in the “interconnect gap”
- Establish hard (simulation and/or measurement based) universal performance metrics for package selection.

4.1.9. Metrology

4.1.9.1. Measurement of Complex Three-Dimensional (3D) Structures

The 3D structures, such as FinFETs, place increased need for inline metrology for dimensional, compositional, and doping measurements. The materials’ properties of block co-polymers for directed self-assembly (DSA) result in new challenges for lithography metrology. The increased use of multi-patterning techniques introduces the need to independently solve a large set of metrics to fully characterize a multi-patterning process.

The 3D interconnect comprises several different approaches and new process control needs are not yet established. For example, 3D (critical dimension and depth) measurements will be required for trench structures including capacitors, devices, and contacts.

4.1.9.2. Measurement of Complex Material Stacks and Interfacial Properties

Reference materials and standard measurement methodology for new high-κ gate and capacitor dielectrics with engineered thin films and interface layers as well as interconnect barrier and low-dielectric layers, and other process needs are required.

Optical measurement of gate and capacitor dielectric averages over too large an area and needs to characterize interfacial layers. Carrier mobility characterization will be needed for stacks with strained silicon and silicon on insulator (SOI), III-V, GeOI, and other substrates, or for measurement of barrier layers. Metal gate work function characterization is another pressing need.

4.1.10. Factory Integration

4.1.10.1. Responding to Business Requirements

In responding to rapidly changing and complex business requirements, the grand challenges are as follows:

- Increased expectations by customers for faster delivery of new and volume products
- Rapid and frequent factory plan changes driven by changing business needs
- Ability to load the fab within manageable range under changeable market demand, e.g., predicting planning and scheduling in real-time
- Enhancement in customer visibility for quality assurance of high reliability products: tie-in of supply chain and customer to factory information and control systems (FICS) operations
- Addressing the big data issues, thereby creating an opportunity to uncover patterns and situations that can help prevent or predict unforeseeable problems difficult to identify such as current equipment processing/health tracking and analytical tools
- To strengthen information security: maintaining data confidentiality (the restriction of access to data and services to specific machines/human users) and integrity (accuracy/completeness of data and correct operation of services), while improving availability (a means of measuring a system’s ability to perform a function in a particular time) contradictive to needs of data availability.

4.1.10.2. Re-emergence of 200 mm Production Line

The increased heterogeneity and variety of devices combined with market pressures such as those associated with IoT solutions have given rise to 200 mm production as an important component of microelectronics ecosystem. While basic tenants of FI challenges and potential solutions associated with 300 mm translate well to 200 mm, there are specific FI challenges such as connectivity, variability, and availability of replacement components that must be addressed so that 200 mm can remain as a viable production capability in the ecosystem.

4.1.11. Yield Enhancement

4.1.11.1. Detection of Multiple Killer Defects

One of the important key challenges will be the detection of multiple killer defects and the signal-to-noise ratio. It is a challenge to detect multiple killer defects and to differentiate them simultaneously at high capture rates, low cost of ownership, and high throughput. Furthermore, it is difficult to identify yield relevant defects under a vast amount of nuisance and false defects.

- Existing techniques trade off throughput for sensitivity, but at expected defect levels, both throughput and sensitivity are necessary for statistical validity.
- Reduction of inspection costs and increase of throughput is crucial in view of cost of ownership (CoO).
- Detection of line roughness due to process variation.
- Electrical and physical failure analysis for killer defects at high capture rate, high throughput and high precision.
- Reduction of background noise from detection units and samples to improve the sensitivity of systems.
- Improvement of signal to noise ratio to delineate defect from process variation.
- Where does process variation stop and defect start?

4.1.12. Environment, Safety, Health, and Sustainability

4.1.12.1. Material Challenges in Environmental, Safety, Health, and Sustainability

Material challenges in Environmental, Safety, Health, and Sustainability are as follows:

- Emerging/novel materials, i.e., III-V (GaN, InP, InGaP, etc.), nano and energetic materials (assessing their ESH/S impacts, along with social responsibility implications)
- Utilization challenge (materials efficiency of incoming fab materials is <2%)
- Treatment and abatement solutions to meet current and future regulatory requirements can gate development ramp, and costs increasing
- Restrictions on recycling, repurposing, and reuse are significant due to technology and regulatory hurdles
- There is not universally accepted or applicable alternative assessment (frameworks, methods, and tools) strategy, nor are there clear guidelines or standards for how these should be applied for picking less ESH/S impactful materials.

4.2. In the Long-term

4.2.1. Application Benchmarking

4.2.1.1. Big Data Analytics

The gains in graph processing performance came mainly from improvements in algorithms, memory bandwidth, and bandwidth. Processor performance does not have a first-order impact.

In the long term, large memory bandwidth and lower latency and higher bandwidth of the global network, which could use optical links, will be needed.

4.2.1.2. Feature Recognition

The main long-term challenges are DNN hardware based on either in-memory digital or analog computation, with the following critical technology need: analog memory devices, low-power, high-bandwidth, moderate-precision and extremely area-efficient A/D converters.

4.2.2. Systems and Architectures

4.2.2.1. IoT

The development of IoT will face significant long-term challenges. Low cost-efficient energy harvesting using multiple sources in order to develop autonomous systems, energy storage and management, low power sensing, computing and communication, automatic network configuration, and security will be needed.

4.2.2.2. Mobile

Video will drive demand for both bandwidth and display, and augmented reality applications will require further increases in communication, computation, capture, and display. An important long-term challenge is also the substantial reduction of power consumption and increase of battery capacity to meet the demand of very active users.

4.2.2.3. Cloud

Three-fourths of all data important to organizations will never be in the data center. High bandwidth memory and large socket thermal power dissipation using improved packaging and cooling will be needed.

4.2.2.4. Cyber-Physical Systems

Long-term challenges are associated with hardware and software reliability, security, exponential increase in the number of bytes generated and need of local analysis, and substantial advances in storage technologies.

4.2.3. Outside Systems Connectivity

Long-term grand challenges are the following: development of circuits for cancellation of 5G mmWave noise; reconfigurable and high-efficiency directional MIMO antenna with circuits to reconfigure and synchronize signals; agreement on optical technology standards; development of technologies for communication between systems with different wavelengths; polarization and modulations; reduction of latency of communication between CPUs and memory in data centers specifically due to routing, and conversion of electrical and photonic signals.

4.2.4. More Moore

Power scaling is a major long-term challenge, which should use steep subthreshold slope devices but there is a lack of manufacturable candidates up to now. Diversification with novel architectures like vertical GAA (VGAA) devices, 3D stacking and possible co-integration of CMOS and beyond CMOS will be required for improving performance. These will need good management of thermal challenges, yield, and cost, together with introduction of alternatives to Cu-interconnects with low resistance and good reliability.

4.2.5. Lithography

Lithography limits will be reached by 2030 but resolution may not be any longer a major limiter by then due to the likely introduction of several types of 3D device structures into manufacturing The new potential patterning challenges will thus be related to cost, yield, defectivity, and optimization of complex 3D structures. Etch and deposition of sub 10 nm structures will also become major challenges. Another potential challenge might be implementing patterning on 450 mm wafers. However, as EUV becomes mainstream patterning method providing substantial cost reduction it could limit the financial benefit of switching to 450 mm wafers. Little work is currently being done on extending any other patterning methodology (multiple patterning, nanoimprint, maskless, directed self-assembly (DSA) to larger wafer sizes than 300 mm.

4.2.6. Factory Integration

Important long-term challenges are the flexibility, extendibility, and scalability needs of a cost-effective, leading-edge factory, tackling environmental issues like material recycling and substitution (scarce, toxic) and future global regulations, and management of the uncertainty of novel device types replacing conventional CMOS and the impact of their manufacturing requirements on factory design.

4.2.7. Yield Enhancement

The next generation inspection is a significant challenge. We need to explore new alternative technologies that can meet inspection requirements to discriminate defects of interest, like high speed scanning probe microscopy, near-field scanning optical microscopy, interferometry, scanning capacitance microscopy, and e-beam. In-line defect characterization and analysis for smaller defect sizes and feature characterization will be required alternatively to optical systems and energy dispersive X-ray spectroscopy.

4.2.8. Beyond CMOS

The beyond CMOS era is facing major research challenges. Nanoscale volatile and nonvolatile memory technologies to replace traditional SRAM, DRAM and Flash in appropriate applications are needed, for instance by using resistive memories (phase-change RAM (PCRAM), Resistive RAM (ReRAM), magnetic RAM (MRAM)). The scaling of information processing technology substantially beyond that attainable by ultimately scaled CMOS will require new computing paradigms like neuromorphic or quantum computing, novel architectures, device technology breakthroughs using charges (e.g., small slope switches) or in the longer term alternative state/hybrid state variables (e.g., spin, magnon, phonon, photon, electron-phonon, photon-superconducting qubit, photon-magnon), the states being be digital, multilevel, analog, or entangled.

4.2.9. PACKAGING INTEGRATION

In the field of packaging, the significant long-term challenges are reliable interconnects and substrates for wearable electronics (bendable, washable); bio compatible systems for miniaturized implants; efficient integration of electronic and optical components; and integration of cooling systems for quantum computing.

4.2.10. METROLOGY

Nondestructive wafer and mask-level metrology with better precision for novel device architectures and 3D structures are needed. Complementary and hybrid metrology combined with state-of-the-art statistical analyses will be required to reduce the measurement uncertainty due to statistical limits of sub-7 nm process control. Materials characterization and metrology methods are also needed for control of interfacial layers, dopant positions, defects, size, location, alignment and atomic concentrations relative to device dimensions and for direct self-assembling processes.

4.2.11. ENVIRONMENT, SAFETY, HEALTH, AND SUSTAINABILITY

We are facing substantial challenges with the possible impact of health and environment of emerging materials (e.g., III-V materials, perfluorooctanoic acid (PFOA)) and potential biological interactions with e.g., mmWave (28-330 GHz). Driving green chemistry and engineering concepts will become a very important asset for future technologies, considering their impact on sustainability and future regulations in this domain.

4.2.12. MORE THAN MOORE

As many application fields, such as automotive, healthcare, communication and energy management, increasingly require the development of multifunctional smart systems, multidisciplinary cooperation along the complete value chain becomes more important every day. In the More than Moore domain, which is characterized by the heterogeneous integration of digital and non-digital components, this leads to the development of generic technology modules and open technology platforms. This makes it possible to explore novel opportunities for innovation, e.g., organ-on-chip for human disease modeling and the development of personalized medicine. Some of these trends—in particular on the needs regarding smart sensors, smart energy, energy harvesting, and wearable, flexible and printed electronics—are described in a new More than Moore white paper.

5. HISTORICAL EVOLUTION OF THE ROADMAP METHODOLOGY

From NTRS, to ITRS, and finally to IRDS

5.1. MOORE'S LAW

For over 50 years the semiconductor industry has marched at the pace of Moore's Law. Transistor scaling associated with doubling the number of transistors every two years on the average has been and continues to be the unique feature of the semiconductor industry. As a consequence, as transistors became smaller, they could also be switched from the off to the on state at faster rates while simultaneously they became cheaper to manufacture. System integrators assembled new products utilizing the building blocks provided by the semiconductor industry, but they were barely able to complete assembling a new system when a yet new more powerful IC was becoming available. Any new technology generation enabled multiple new products with better performance than the previous ones. Integrated device manufacturers (IDM) in conjunction with software companies providing operating systems and applications were in full control of the pace at which the whole electronics industry ecosystem was moving forward. Therefore, past editions of technology roadmaps (i.e., National Technology Roadmap for Semiconductors (NTRS) and the International Technology Roadmap for Semiconductors (ITRS)) concentrated on forecasting the rate of transistor scaling, the technical impediments to be overcome, and how transistor density and performance affected the evolution of integrated circuits.

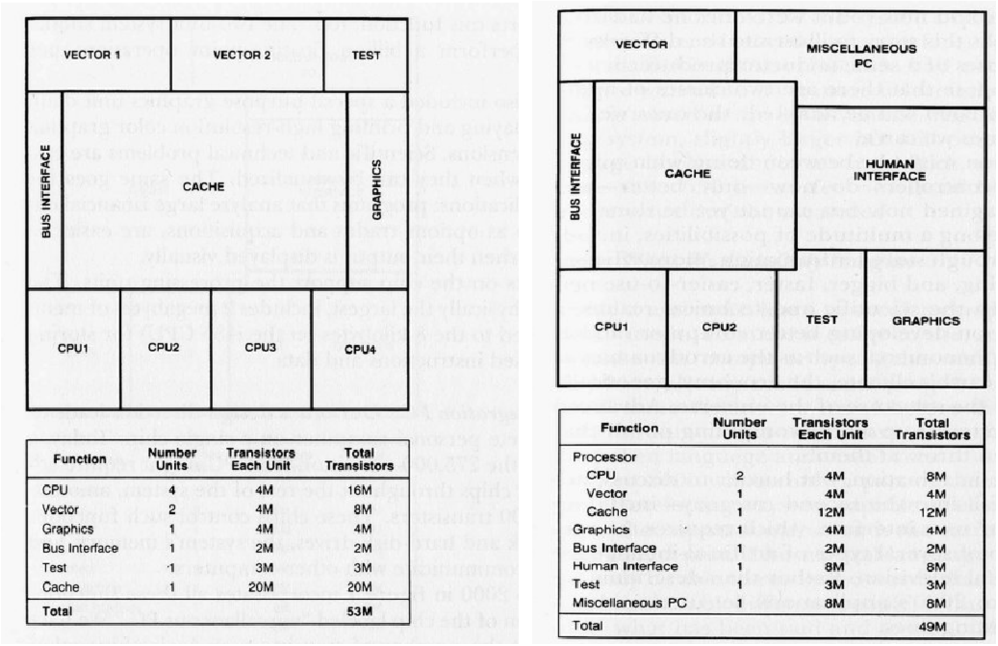
During the past 15 years the advent of fabless design houses and foundries has revolutionized the way in which business is done in the new semiconductor industry, and because of this change system integrators have regained full control of the business model. This implies that system requirements are set at the beginning of any new product design cycle and step-by-step-related requirements percolate down through the manufacturing production chain to the semiconductor manufacturers. No longer does a faster microprocessor trigger the design of a new PC but on the contrary the design of a new smart phone generates the requirements for new ICs and other related components. In addition, fast approaching fundamental 2D topological limits have been threatening the ability of the semiconductor industry to continue scaling at historical rates. 2016 has seen the introduction in major conferences of new very creative 3D transistors, memory cells, and overall 3D IC structures that through the next 10 years will revolutionize the way ICs are designed and produced.

5.2. The Dawn of the New Computer Industry

The first arithmetic mechanical machine was invented by Blaise Pascal as far back as 1642 but the first machine encapsulating most of the elements of modern computers was introduced by Charles Babbage in 1837, whereas machine language was first pioneered by Ada Lovelace in 1843. ENIAC was introduced in 1946, being the first fully electronic computer powered by vacuum tubes. By the early 60’s IBM had established itself as a leader in transistorized computers in four product lines aimed at multiple applications, but to merge these different lines in a more productive way, in 1964 it introduced the first general-purpose machine. The Model 360 could perform up to 34,500 instructions per second, with memory from 8 to 64 KB. Bipolar transistors were by far faster and more reliable than any MOS device and for the subsequent 30 years it was bipolar technology that powered the evolution of computers.

The personal computer industry began as a “hobby past-time” in the mid 70’s. Apple was the most significant company shipping personal computers until IBM decided to enter this business. The IBM PC was introduced in 1981 with the support of Intel and Microsoft. As time went by personal computers became more powerful while large computers pushed the performance of bipolar transistors beyond their power limits despite the use of very sophisticated cooling techniques. By the mid 90’s both branches of the industry relied on CMOS technology for both logic and memory products. However, power limits were again reached by the middle of the first decade of the new century imposing severe limitations on performance. (Refer to Sections 1.2.5 and 1.2.6.)

It was proposed as far back as 1989 that in order to continue providing increased computational performance while remaining within severe power limitations could be achieved by means of operating multiple cores processing information in parallel (Figure ES36).



Function	Number Units	Transistors Each Unit	Total Transistors
CPU	4	4M	16M
Vector	2	4M	8M
Graphics	1	4M	4M
Bus Interface	1	2M	2M
Test	1	3M	3M
Cache	1	20M	20M
Total			53M

Function	Number Units	Transistors Each Unit	Total Transistors
Processor			
CPU	2	4M	8M
Vector	1	4M	4M
Cache	1	12M	12M
Graphics	1	4M	4M
Bus Interface	1	2M	2M
Human Interface	1	8M	8M
Test	1	3M	3M
Miscellaneous PC	1	8M	8M
Total			49M

Figure ES36 Prediction of MPU migration multicores outlined in 1989

Under these conditions the value of the operating frequency could be approximately scaled down inversely proportionally to the number of cores. Under this approach the complete output result would be constituted by a combination of the results of the individual cores. This approach became the standard by the middle of the previous decade when fundamental power limits were finally reached and indeed improved performance was demonstrated while operating within allowable power limits. However, the resulting rate of improvement was much slower than historical rates, as shown in Figure ES37. Despite the use of multicores, the system architecture remained solidly anchored to the Van Neumann concept.

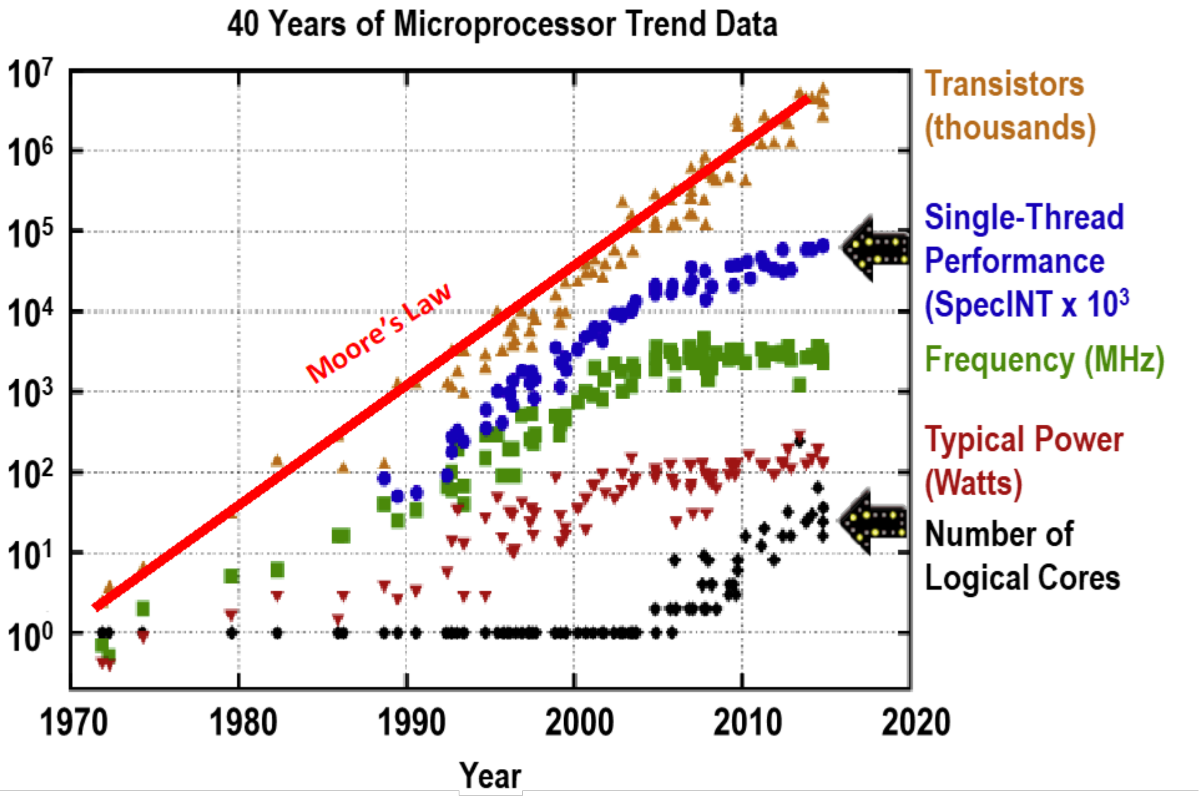
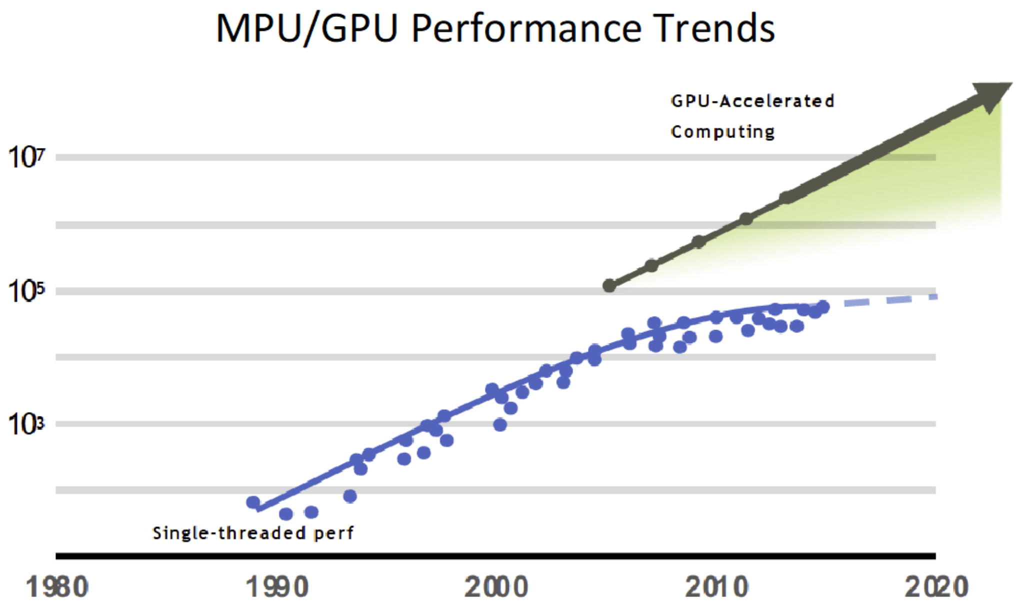


Figure ES37 Adoption of multicore architecture allowed moderate performance increase in CPU design within power limits

It slowly became clear that it was necessary to substantially or completely depart from past architectures and devise new solutions tailored to solving specific problems in order to achieve the rate of improvement in performance comparable to the past. In essence the system designers came to the realization that in order to make progress it was no longer possible to seek for a universal solution, but it was necessary to depart from the past approach and develop many new architectures each capable of delivering better solutions to specific problems (Figure ES38).

New ways of performing a variety of computing functions have been demonstrated and more yet are under development. Among these, neuromorphic computing, approximate computing, and perhaps most importantly , quantum computing appear as the most promising architectures for some special applications.



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

Source: NVIDIA

Figure ES38 Performance resuming and exceeding historical trends by using GPU-accelerated computing

5.3. SoC and SiP

In the past 15 years the advent of the Internet; the extensive deployment of Wi-Fi base stations; consumer acceptance of a broad variety of cell phones and wireless mobile appliances, plus the successful combination of fabless companies working in conjunction with foundries has completely changed the electronics industry. System integrators are nowadays able to conceive, design, and realize any integrated circuit they wish without having to refer to integrated device manufacturers. System integrators can integrate multiple functionalities in a single chip called System on Chip (SoC) or by means of integrating multiple dice in a single package called System in Package (SiP) as opposed to connecting multiple standard specialized ICs on a board (PCB). It is clear that these methods of integration are more efficient and less costly than acquiring several separate ICs (e.g., , microprocessor, graphic processor, multiple memory types, USB, etc.) and assembling them on a board. In addition, the limited space available in mobile products has further accelerated the integration of multiple capabilities in a very confined environment. Usually packaging solutions are the first to be implemented since they can expeditiously allow integration of multiple heterogeneous ICs into a single package with little impact on the underlying technology of the ICs components. Heterogeneous monolithic integration requires more development effort and typically follows with a delay of one to two technology generations from SiP demonstrations. Even under these conditions integration of a wide variety of different technologies into a single chip has some limitations. Therefore, it is expected that a combination of SoC and SiP will become the dominant technological approach to increase system performance by cost effectively augmenting the integration complexity of novel ICs and packaging solutions. System integrators are by and large setting the pace of innovation for the electronics industry. The IC industry has also contributed to provide valuable technology building blocks to other industries that either did not exist or were in their infancy before, and by adopting well-established technologies and brand-new devices like micro-electromechanical systems (MEMS), flat panel displays, multiple sensors and so on have been realized. All these somewhat dissimilar technologies have been readily included into mobile appliances by means of heterogeneous integration. This trend, where continued miniaturization of the digital components was combined with the integration of new, often analog functionalities, was forecasted as far back as 2006 by the ITRS under the name of More than Moore (MtM).

5.4. POWER CHALLENGE

Each new technology generation has continued to produce smaller and better transistors that could switch faster than those produced with any previous technology generation. In the past this electrical feature of transistors (e.g., intrinsic transistor delay) enabled microprocessors to operate in each new generation at higher frequencies and therefore computer performance, as measured by industry benchmarks, (e.g., measured by millions of instructions executed per second (MIPS), continued to improve at a very fast rates (almost doubling with each new technology generation) without any major change in computer architecture. In fact, the basic computer architecture has not changed much through the years since Von Neumann introduced his concept on how to perform computing in 1945. However, power consumption of integrated circuits kept on increasing until the beginning of the past decade when fundamental thermal limits were finally reached by some ICs. It became clear then that it had become practically impossible to sustain concurrently increasing both the frequency of operation and the number of transistors: one of the two features (i.e., either the frequency or the number of transistors) had to level off in order to make the ICs capable to operating under practical thermal conditions. Frequency was selected as the sacrificial victim and indeed it has stalled in the range of a few GHz since the middle of the previous decade. New transistor designs and new architectures have been aimed at alleviating this problem especially in the past 5-10 years. See the following paragraph for a roadmap towards a practical solution.

5.5. CONSEQUENCES OF FREQUENCY LIMITATIONS

These limitations on maximum useable frequency have impacted the rate of progress of the computer industry that has been compelled to develop such methods as complex software algorithms and clever instruction management to improve performance to partially compensate for the aforementioned conditions. The architecture of the microprocessors has changed from single core to multi-core. With this partitioning of the process architecture (very easy from a technology and layout point of view) each core can run in the lower GHz range while the output rate is increased multifold by combining the output of multiple cores to produce the output signal. By extension, the concept of parallel processing has already found specific applications that go beyond classical von Neumann computing. A typical example is the case of face (or any other object) recognition. In this case, as a first step, the image of an unknown object is subdivided in multiple pixels. A graphics' processor capable of simultaneously processing thousands of parallel streams of pixel then compares this information with an extensive memory library of objects of the similar type. By means of this process a subset of objects matching a large number of pixels of the object under evaluation are identified. Based on this comparison evaluation a feedback is sent back towards the object under evolution. In the meantime, variable electrical weights located in the physical nodes of each line of the paralleled processing array are recalibrated to further refine the evaluation; this process (back and forth) is repeated multiple times until an accurate match is found. Object recognition exemplifies one of the possible implementations of what has been defined as neuromorphic architecture. In summary, this approach allows the network to compare any object under investigation with a vast library of objects of the same family by means of a comparative and iterative process where the array nodes are readjusted multiple times until an essentially perfect match is found. A very successful demonstration of the power of parallel computing has shown that performance in line with the historical improvement rate of the 90s can be achieved (Figure ES38).

Unfortunately, this parallel type of solution cannot (yet?) be used in all computational cases since some problems, or part of them, can only be solved in a serial way, but it does open the way for new innovative architectures.

However, frequency or power limitations have not impacted the development and expansion of either cell phones or devices using Wi-Fi to access the Internet. Cell phone and mobile devices in general operate below 3 GHz due to the way operating frequencies are allocated in each country by very specific rules. In addition, power consumption of portable devices has been limited to 5 watts to allow multiple hours of operation without a recharge. Historically, consumers began accessing the Internet via desktop appliances (e.g., personal computers) and then progressively became accustomed to accessing it via mobile multipurpose appliances. Reaching any source of information via the Internet takes tens of milliseconds due to the speed at which signals can travel on any interconnect lines, so microprocessors operating in the low GHz frequency range in mobile appliances are more than adequate to handle the communication traffic.

For these reasons neither power nor frequency limitations have affected the mobile world so far. This situation may change to some degree with the advent of 5G where frequency and power considerations will substantially change. (Refer to Section 1.2.3.)

5.6. INTERNET OF THINGS, INTERNET OF EVERYTHING (IOT, IOE)

The US Department of Defense awarded contracts as early as the 1960s for packet network systems, including the development of the ARPANET (which would become the first network to use the Internet Protocol).

Access to the ARPANET was expanded in 1981 when the National Science Foundation (NSF) funded the Computer Science Network (CSNET). Since the mid-1990s, the Internet has had a revolutionary impact on culture and commerce, including the rise of near-instant communication by electronic mail, instant messaging, voice over Internet Protocol (VoIP) telephone calls, two-way interactive video calls, and the World Wide Web. This worldwide connectivity has created new phenomena like social networking and online shopping and banking. Increasing amounts of data are transmitted at higher and higher speeds over fiber optic networks operating at 1-Gbit/s, 10-Gbit/s, and beyond 40-Gbit/s. The Internet's takeover of the global communication landscape occurred almost instantly in historical terms: it only communicated 1% of the information flowing through two-way telecommunications networks in 1993, increasing to 51% by 2000, and more than 97% of the telecommunicated information by 2007! Today the Internet of Things continues to grow, driven by ever-greater amounts of online information, commerce, entertainment, and social networking, just to mention a few. Access to the Internet was originally done via hardwired desktop computers but the introduction of wireless technology (Wi-Fi), smart phones in 2007 and tablets in 2010 has revolutionized the way people interact via the Internet. The world of communications has truly become a wireless, ubiquitous, and continuously interconnected world.

With this all said and done it is important to remember that the ever-expanding Internet of Everything (IoE) could not have happened if semiconductors had not powered a variety of communication devices, data centers, routers, and sensors. The advent of foundries and fabless companies enabled the customization of semiconductor products that now can cover all the aspects of IoE and it would be a mistake to assume that the semiconductor industry is by now a mature industry and it has not much more to offer. The new fabless/foundry ecosystem has opened the door to an endless flow of innovation available at very reasonable and affordable costs. The advent of the third phase of device integration (i.e., 3D power scaling) plus the many new capabilities associated with the introduction of revolutionary materials in the semiconductor industry will revolutionize how computers are constructed. New computers built with revolutionary architectures enabled by new devices will be surrounded from top to bottom with a variety of new sensorial capabilities and communications capabilities that will offer new and exciting options to system designers (see the Rebooting Computing website for more details).

5.7. The 3 Eras of Scaling

The foundations of the IC industry were laid out with the invention of the self-aligned silicon gate planar process in the late 1960's. Moore's predictions of the doubling of the number of transistors per die on an annual and then bi-annual pace formulated in 1965 and in 1975, respectively, in conjunction with Dennard's scaling guidelines led to the growth of the semiconductor industry until the beginning of the last decade.

Geometrical scaling characterized the 70s, 80s and 90s. This was the first generation of transistor scaling. The National Technology Roadmap for Semiconductors (NTRS) was initiated in the U.S. with a workshop held in 1991 and subsequent publications occurred in 1992, 1994, and 1997, respectively. The electronics industry was primarily "bottom up" driven during this period since any new technology generation provided transistors operating with continuously better performance that could power new memory and processors that easily fitted in the existing system architecture. System integrators could barely keep up with the rate at which new memory and new processors products were introduced since the industry changed in the 90's from a 3-4-year cycle to a 2-year technology cycle. However, major upcoming material and structural limitations were identified by the NTRS between 1994 and 1997. These problems were so fundamental that it was deemed necessary to engage the whole international semiconductor community to successfully identifying possible solutions in a timely fashion. A proposal to extend the NTRS to European, Japanese, Korean, and Taiwanese technical communities was presented in April 1998 to the World Semiconductor Council (WSC). The proposal was accepted and the ITRS was formed. The international research community met in San Francisco on July 1998 and at that meeting the research activities necessary to completely restructure the MOS transistor and the necessary methodology were approved and launched worldwide [Figure ES39].

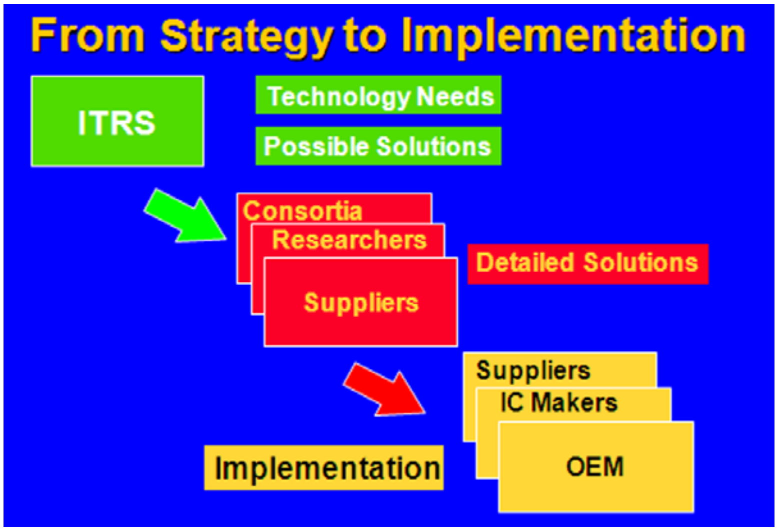


Figure ES39 1998 ITRS program: from strategy to implementation

This new approach to restructuring the transistor was named "equivalent scaling". The goal of this program consisted of reducing the historical time of ~25 years between major transistor innovations to less than half in order to save the semiconductor industry from reaching a major crisis. Strained silicon, high-κ/metal gate, FinFET, and use of other semiconductor materials (e.g., germanium) represented the main features of this scaling approach (Figure ES40). By 2011 all these new process modules were successfully introduced into high volume manufacturing (Figure ES41).

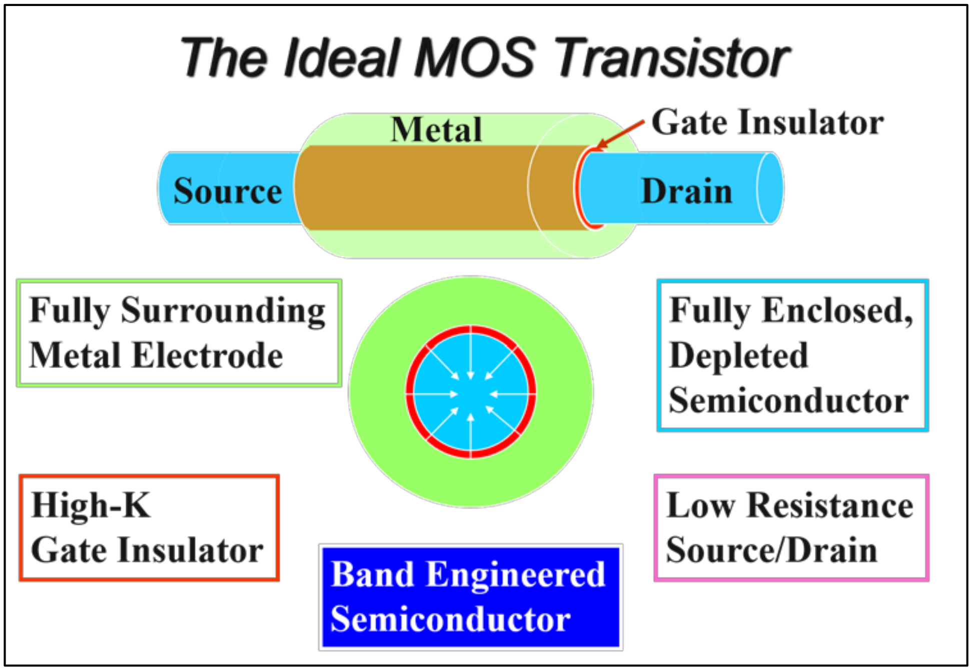
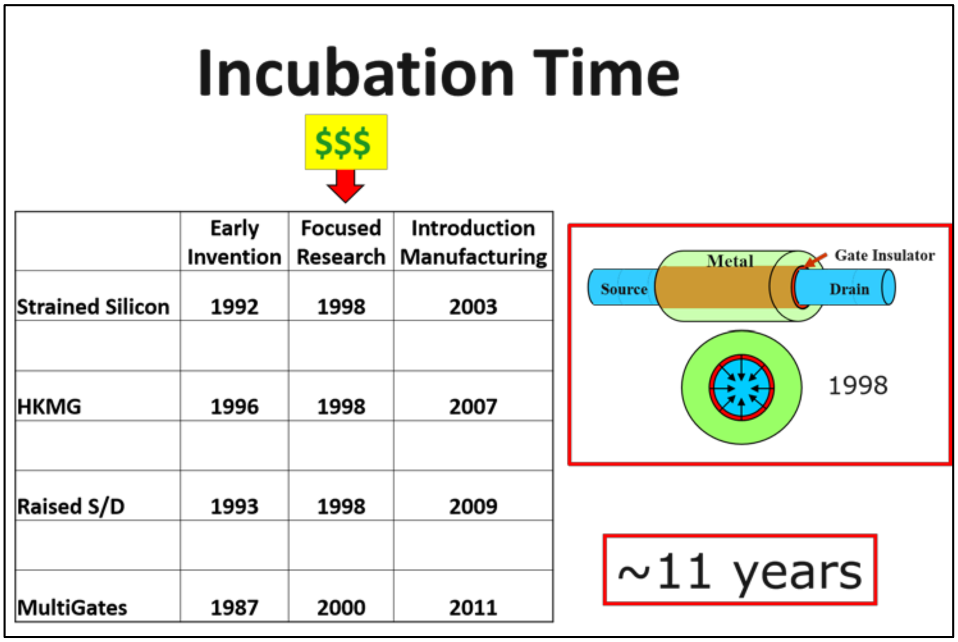

The Ideal MOS Transistor
Metal
Gate Insulator
Source
Drain
Fully Surrounding Metal Electrode
Fully Enclosed, Depleted Semiconductor
High-K Gate Insulator
Low Resistance Source/Drain
Band Engineered Semiconductor

Figure ES40 Vision of the completely refurbished MOS transistor



Incubation Time

$$$

	Early Invention	Focused Research	Introduction Manufacturing
Strained Silicon	1992	1998	2003
HKMG	1996	1998	2007
Raised S/D	1993	1998	2009
MultiGates	1987	2000	2011

Metal
Gate Insulator
Source
Drain
1998

~11 years

Figure ES41 From strategy to implementation in high-volume manufacturing in record time

The advent and success of the combination of fabless design houses and foundries about 10 years ago revolutionized the way in which business was done and heralded the coming of the new semiconductor industry. Because of this environmental change system integrators finally regained full control of the business model. This implied that system requirements were going to be set at the beginning of any new product design cycle and step by step corresponding device requirements percolated down through the design/development/manufacturing production chain to the semiconductor manufactures. No longer was a faster microprocessor triggering the design of a new PC but on the contrary the design of a new smart phone generated the requirements for new ICs and other related components. Under these conditions it became clear in 2012 that the ITRS needed to adapt and morph to the new ecosystem (Figure ES42). It was anticipated then that this transformation process would take some time and it was decided that the 2013 ITRS was going to be the last of its kind. Next, 2014 and 2015 were going to be dedicated to the construction of a new intermediate roadmap that was named ITRS 2.0.

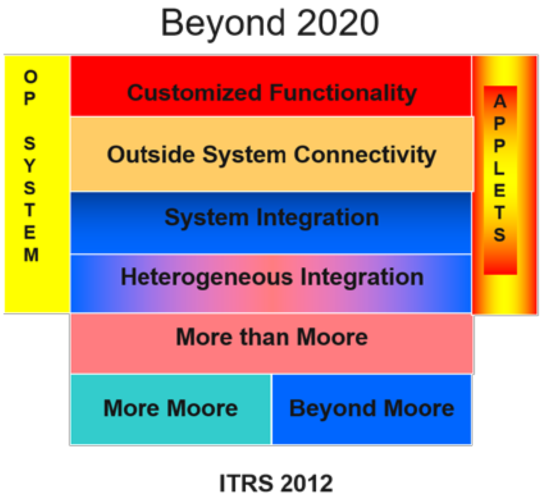


Figure ES42 The new ecosystem of the electronics industry

During the preparation of the 2013 ITRS it was also assessed that horizontal (2D) features were going to be approaching the range of a few nanometers shortly beyond 2020 (Figure ES43) and so it became clear that the semiconductor industry was going to be running out of horizontal space by then! The question was: “Which products were going to be reaching these 2D limits first?” Memory products have always been the leaders in transistor density (i.e., smallest feature pitch) and so it should not have been surprising to realize that the solution to this problem was to come first from companies producing Flash memories. In fact, multiple companies announced in 2014 that future products were going to fully utilize the vertical dimension (Figure ES44). This is not too dissimilar from the approach taken in Manhattan, Tokyo, Hong Kong, or similarly highly crowded places to deal with space limitations: skyscrapers have become the standard approach to maximize “packing density”. In addition, the rapid increase in the number of transistors (i.e., 2×/2-years) and the comparably rapid increase in operating frequency throughout the 80s and 90s drove the power dissipation of microprocessors way beyond the 100 W by the 2003−2005 timeframe. This implied that number of transistors and frequency could no longer simultaneously increase. Under these conditions the electronics industry decided to convert to a multicore architecture, and continued to increase the number of transistors at historical rate but limited the operating frequency to few gigahertz. All the above considerations indicated that the structure of integrated circuits needed to evolve from 2D to 3D structures and transistor design needed to be aimed at reduced power consumption as opposed to be optimized for maximum operating frequency.

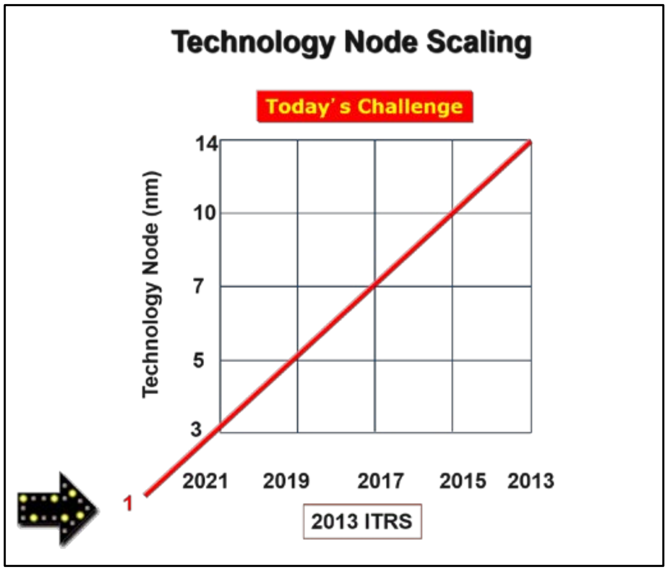


Figure ES43 2D scaling will reach fundamental limits beyond 2020

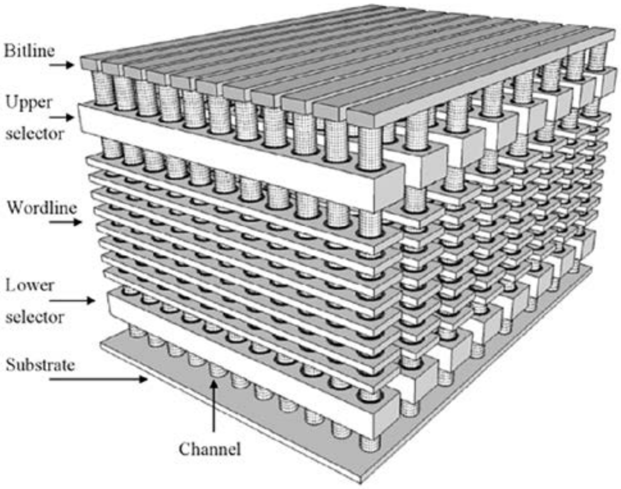
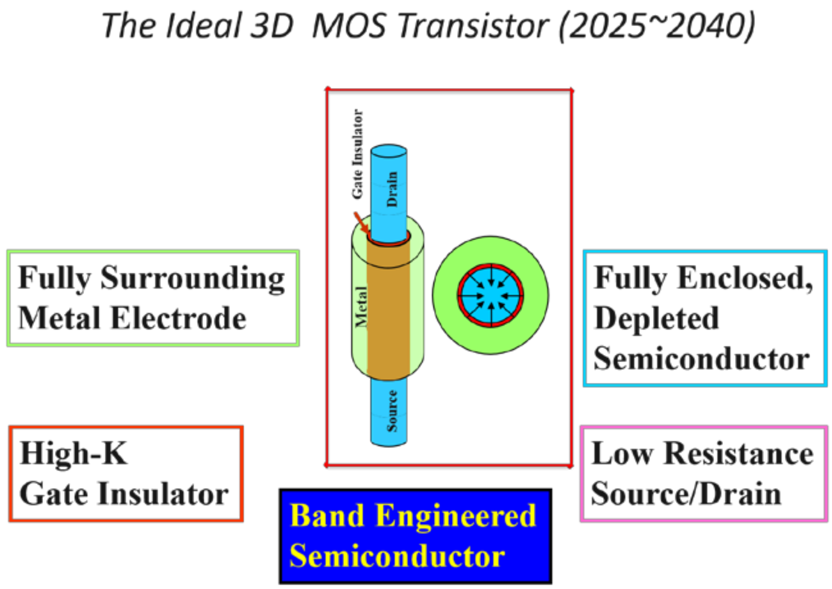


Figure ES44 Flash memory aggressively adopts 3D scaling in 2014

For the reasons described above, the new scaling method was named “3D Power Scaling” by the IRDS to symbolically include in a very succinct way all the challenges facing the semiconductor and electronics industries in the next 15 years (Figure ES45).



The Ideal 3D MOS Transistor (2025~2040)

Fully Surrounding Metal Electrode

Fully Enclosed, Depleted Semiconductor

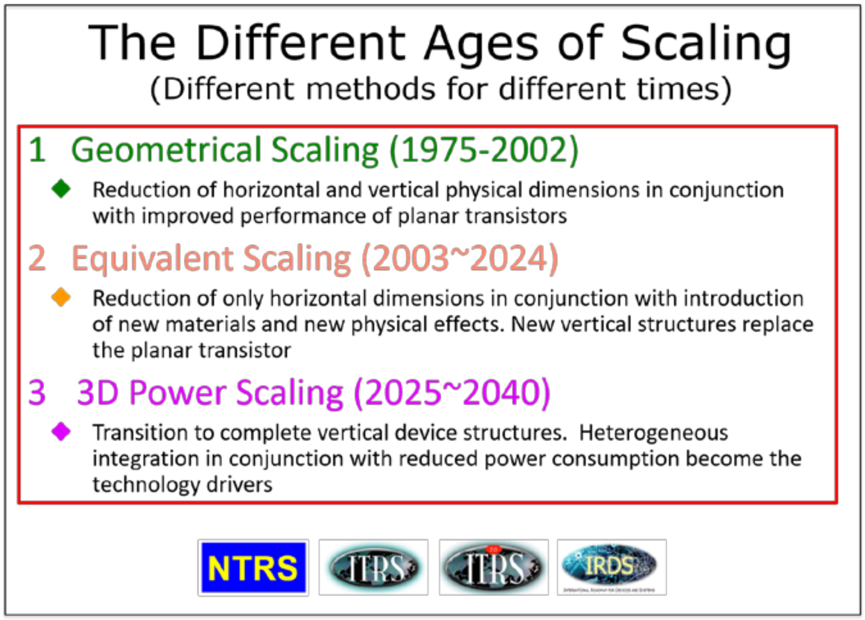
High-K Gate Insulator

Band Engineered Semiconductor

Low Resistance Source/Drain

Figure ES45 The ideal 3D transistor

A summary of the 3 eras of transistor scaling is shown in Figure ES46.



The Different Ages of Scaling

(Different methods for different times)

1 Geometrical Scaling (1975-2002)

- Reduction of horizontal and vertical physical dimensions in conjunction with improved performance of planar transistors

2 Equivalent Scaling (2003~2024)

- Reduction of only horizontal dimensions in conjunction with introduction of new materials and new physical effects. New vertical structures replace the planar transistor

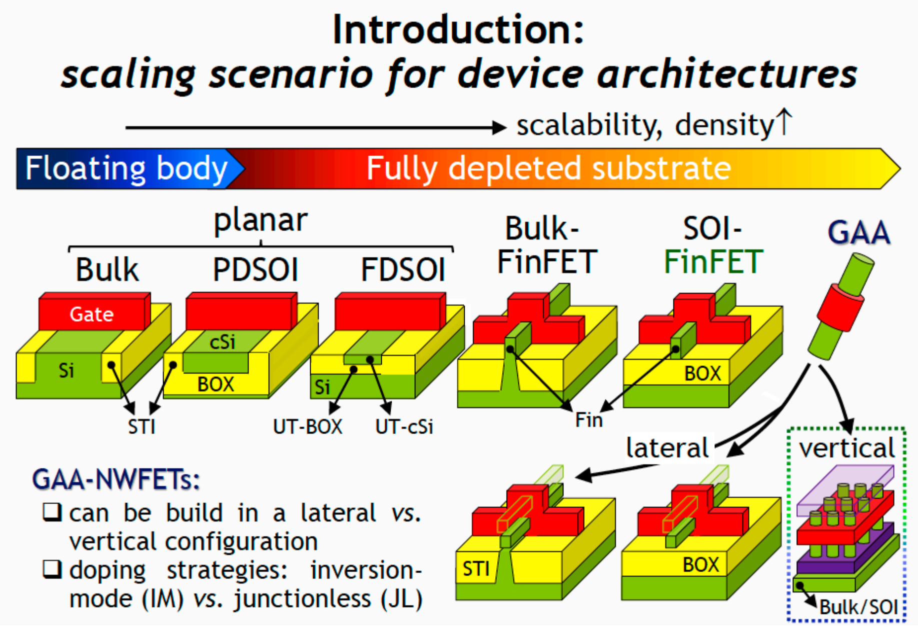
3 3D Power Scaling (2025~2040)

- Transition to complete vertical device structures. Heterogeneous integration in conjunction with reduced power consumption become the technology drivers

Figure ES46 The 3 eras of scaling heralded by NTRS, ITRS, ITRS 2.0, and IRDS

6. PRACTICAL CONSIDERATIONS

It is the goal of the IRDS to outline the most aggressive and likely solutions that will eventually become mainstream in the long term (i.e., 10-15 years). However, practical implementation of new technologies always occurs in steps to make sure that the highest manufacturability results are maintained or even exceeded. Therefore, 3D implementation will also happen in steps and the most relevant transistor milestones are shown in Figure ES47. These new transistor architectures will introduce benefits related to controlling electrostatic effects but will introduce several new stringent requirements on how any new system is designed. The MOSFET device architecture has been changing from the planer 2D through 2.5D of FinFET to GAA either of nanowire or nanosheet structures. These GAA MOSFETs will be stacked monolithically to be 3D VLSIs. See Figure ES48.

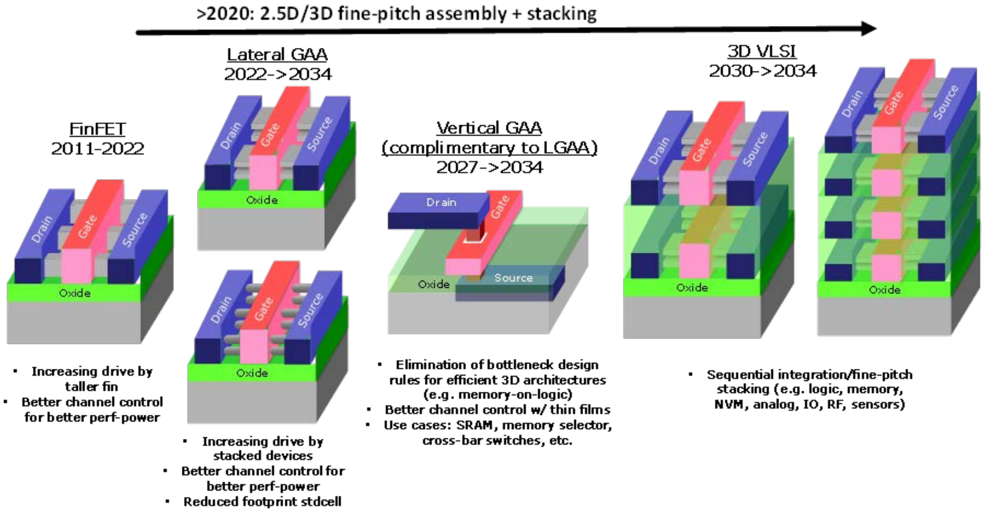


Introduction:
scaling scenario for device architectures

GAA-NWFETs:

- ❑ can be build in a lateral *vs.* vertical configuration
- ❑ doping strategies: inversion-mode (IM) *vs.* junctionless (JL)

Figure ES47 Practical migration of transistor structure from FinFET to GAA to fully vertical



- Increasing drive by taller fin
- Better channel control for better perf-power

- Increasing drive by stacked devices
- Better channel control for better perf-power
- Reduced footprint stdcell

- Elimination of bottleneck design rules for efficient 3D architectures (e.g. memory-on-logic)
- Better channel control w/ thin films
- Use cases: SRAM, memory selector, cross-bar switches, etc.

- Sequential integration/fine-pitch stacking (e.g. logic, memory, NVM, analog, IO, RF, sensors)

Figure ES48 Change in the MOSFET device architecture from the 2D planar through 2.5D FinFets to 3D monolithic VLSI with GAA

In the 2020 IRDS Beyond CMOS roadmap chapter the reader will find multiple new and exciting devices that, after 10 years of research, have already become or are soon becoming key players in the next decade. Integration of several memory circuits on top of logic circuits has been successfully demonstrated with consequent improvement in performance (Figure ES49). This monolithic heterogeneous integration is made possible by the fact that these memory circuits can be fabricated at temperatures below 400°C. These temperatures are comparable to temperatures utilized nowadays to produce multiple interconnect lines on top of microprocessors.

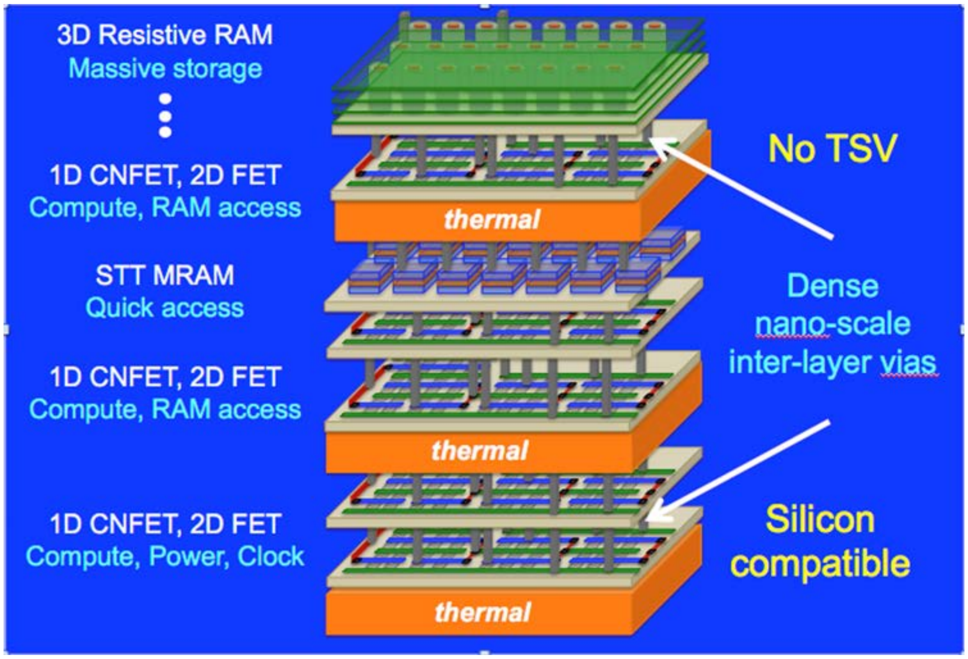


Figure ES49 Planning for the advent of monolithic heterogeneous integration

7. Acronyms/Abbreviations

Term	Definition
4G	Fourth generation
5G	Fifth generation
AB	Application Benchmarking
AI	Artificial intelligence
ARPANET	Advanced Research Projects Agency Network
ASIC	Application-specific integrated circuit
BC	Beyond CMOS
BEOL	Back end of line
BIS	Back illuminated sensors
CAGR	Compound annual growth rate
CASS	IEEE Circuits and Systems Society
CEDA	IEEE Council on Electronic Design Automation
CEQIP	Cryogenic Electronics and Quantum Information Processing
CIM	Compute in memory
CMOS	Complementary metal oxide semiconductor
CS	IEEE Computer Society
CSC	IEEE Council on Superconductivity
D2W	Die to wafer
DNN	Deep neural network
DRAM	Dynamic random access memory
DSA	Directed self-assembly
EDS	IEEE Electron Devices Society
EPS	Electronics Packaging Society
EOT	Equivalent oxide thickness
ESH/S	Environment, Safety, Health and Sustainability
ESI	European SiNANO Institute
EUV	Extreme ultraviolet
FC	Flip chip
FET	Field effect transistor
FI	Factory Integration
FinFET	Fin field-effect transistor
GAA	Gate all around
GbE	Gigabit ethernet
Ge	Germanium
GHz	Gigahertz
GSM	Global System for Mobile Communications
HBM	High bandwidth memory
HDD	Hard disk drive
I/O	Input/Output
IC	Integration circuit
IDM	Independent device manufacturer
IEDM	International Electron Devices Meeting
IEEE	Institute of Electrical and Electronic Engineers
IEEE-SA	IEEE Standards Association
IFT	International focus team
IMEC	Interuniversity Microelectronics Centre
iNEMI	International Electronics Manufacturing Initiative
INGR	International Networks Generation Roadmap
IoE	Internet of everything
IoT	Internet of things
IRC	International roadmap committee
IRDS	International Roadmap for Devices and Systems
ISSCC	International Solid-State Circuits Conference

ITRS	International Technology Roadmap for Semiconductors
LED	Light emitting diode,
LGAA	Lateral gate all around
LTE	Long-term evolution
MAG	IEEE Magnetics Society
MET	Metrology IFT
MHz	Megahertz
MIMO	Multiple input multiple output
MIPS	Millions of instructions per second
ML	Machine learning
MM	More Moore
mmWave	Millimeter wave
MOS	Metal oxide semiconductor
MOSFET	Metal oxide semiconductor field effect transistor
MPU	Microprocessor unit
MtM	More than Moore
Mx	Tight-pitch routing metal interconnect
NA	Numerical aperture
NAND	A logic gate (NOT-AND) that produces an output that is false only if all its inputs are true
NSF	National Science Foundation
NTC	IEEE Nanotechnology Council
NTRS	National Technology Roadmap for Semiconductors
ORSC	Overall roadmap system characteristics
ORTC	Overall technology system characteristics
OSC	Outside System Connectivity
PC	Personal computer
PCB	Printed circuit board
PFOA	perfluorooctanoic acid
RAM	Random-access memory
RCI	IEEE Rebooting Computing Initiative
RDL	Redistribution layer
RMG	Replacement metal gate
RRAM	Resistive random-access memory
RS	IEEE Reliability Society
S/D	Source/drain
SA	Systems and Architectures
SDRJ	Systems and Devices Roadmap of Japan
SIA	Semiconductor Industry Association
SiGe	Silicon germanium
SiNANO	European Academic and Scientific Association for Nanoelectronics
SiP	System in package
SoC	System on chip
SRAM	Static random-access memory
SRC	Semiconductor Research Corporation
SSCS	IEEE Solid State Circuits Society
SSD	Solid state drive
TEPS	Traversed edges per second
TFET	Tunnel field-effect transistor
TPU	Tensor processing units
TSV	Through silicon via
VLSI	Very large scale integration
W2W	Wafer to wafer
WSC	World Semiconductor Council
YE	Yield Enhancement
YoY	Year over year
YtY	Year to year

8. APPENDIX

8.1. APPENDIX A—IFT CHAPTER FILES LINKS

- Application Benchmarking (AB)
- Systems and Architectures (SA)
- Outside System Connectivity (OSC)
- More Moore (MM)
- Beyond CMOS (BC)
- Cryogenic Electronics and Quantum Information Processing (CEQIP)
- Packaging Integration (PI) white paper
- Factory Integration (FI)
- Lithography (L)
- Yield Enhancement (YE)
- Metrology (M)
- Environment, Safety, Health, and Sustainability (ESH/S)
- More than Moore (MtM)
- Medical Devices Market Drivers
- Automotive Market Drivers

8.2. APPENDIX B—OVERALL ROADMAP CHARACTERISTICS (ORSC AND ORTC) SOURCE INFORMATION LINKS

- Systems and Architectures Tables
- More Moore Tables