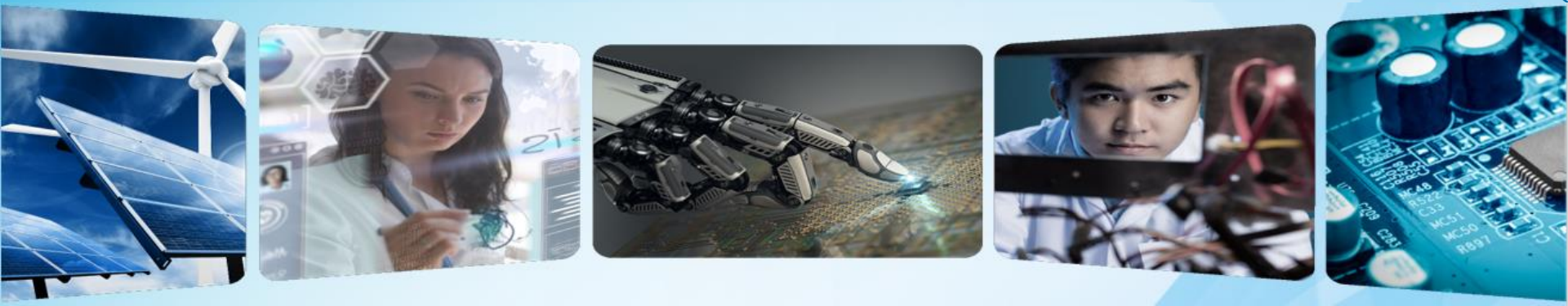


More Moore roadmap for high-volume manufacturing- IRDS view

Mustafa Badaroglu – IEEE IRDS More Moore Global Roadmap Chair

IWAPS 2020, November 5-6, 2020

IRDS
International Roadmap for Devices and Systems

IEEE

Cloud and edge computing drive More Moore

- Device-interconnect-memory technologies for mobile and cloud computing
- Edge computing - additional functionality, biometrics, and display/camera/sensing for increased consumer value
- 2.5D/3D integration to scale memory bandwidth / power and latency

Big Data and abundant computing power are pushing computing to the **Cloud**

Micro (data) servers
Increasing data on cloud requires high-density **memories**

Instant Data generated by sensors and users are pushing computing to the **Edge**

Mobile computing
AI at edge requires fast and high-density memories

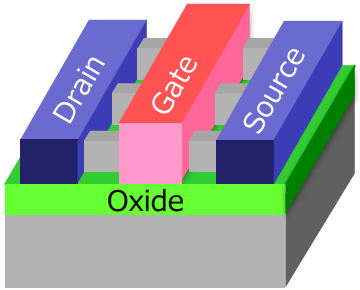
More Moore mission and targets

- Goal
  - Provide physical, electrical and reliability requirements for logic and memory technologies to sustain More Moore (PPAC: power, performance, area, cost) scaling for big data, mobility, and cloud (IoT and server) applications and
  - Forecast logic and memory technologies (15 years) in main-stream/high-volume manufacturing (HVM)
- Product Drivers
  - Mobile – 5G, hetero integration, edge computing, extreme reality (VR/AR)
  - Data/micro servers – cache integration, memory, IO
  - New compute fabrics
    - High-performance energy-efficient graphics/sensing
    - Ultra-low power computing exploiting non-volatility
    - Machine learning at edge and cloud
- PPAC
  - (P)erformance: >15% more performance at iso power
  - (P)ower: >25% less power at constant performance
  - (A)rea: >35% less area
  - (C)ost: <30% wafer cost and >10-20% less die cost for the same function
  - (T)emperature: <10% increase in power density

More Moore device structure evolution

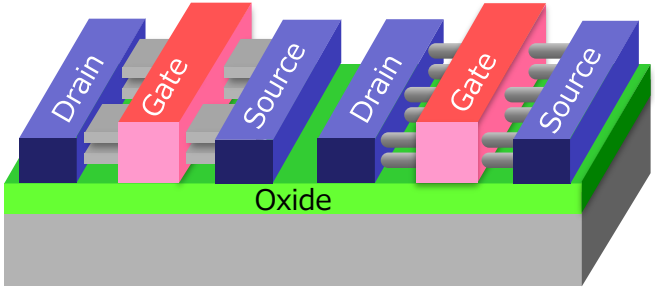
>2020: 2.5D/3D fine-pitch assembly + stacking

FinFET
2011-2022



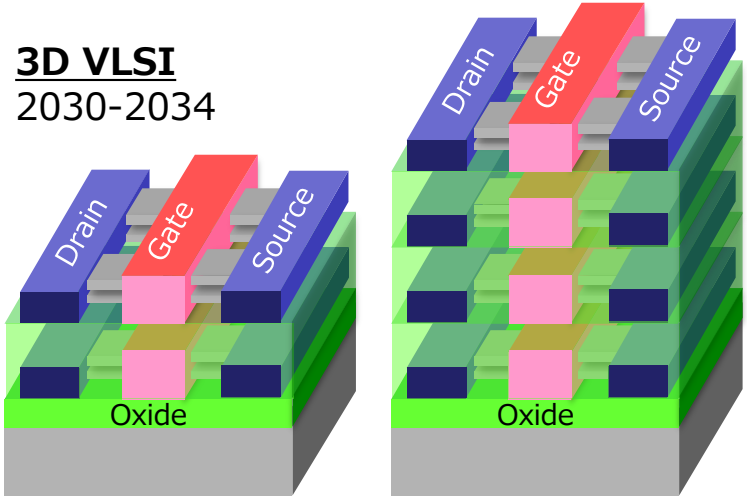
- Increasing drive by taller fin
- Better channel control for better perf-power

Lateral GAA
2022-2034



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

3D VLSI
2030-2034



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

IO: Input/Output
RF: Radio Frequency

Source; IRDS 2020 Edition, More Moore

Technology Naming Convention – IRDS definition

FEBRUARY 20, 2020

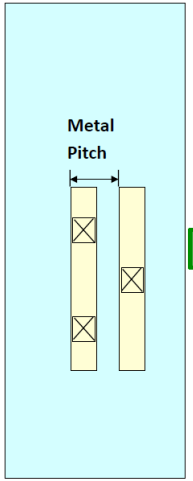
Logic/Foundry Process Roadmaps (for Volume Production)

	2015	2016	2017	2018	2019	2020	2021
Intel		14nm+	10nm (limited) 14nm++		10nm	10nm+	7nm EUV 10nm++
Samsung	28nm FDSOI	10nm		8nm	7nm EUV 6nm EUV	18nm FDSOI 5nm	4nm
TSMC	16nm+ finFET	10nm	7nm 12nm		7nm+ EUV	5nm 6nm	5nm+
GlobalFoundries	14nm finFET			22nm FDSOI 12nm finFET		12nm FDSOI	12nm+ finFET
SMIC	28nm				14nm finFET	12nm finFET	
UMC			14nm finFET			22nm planar	

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

NTRS/ITRS->IRDS Technology Node Definition



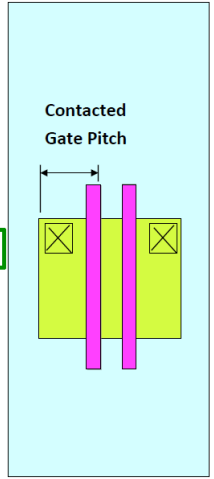
G= Contacted Gate Pitch

M=Metal Pitch

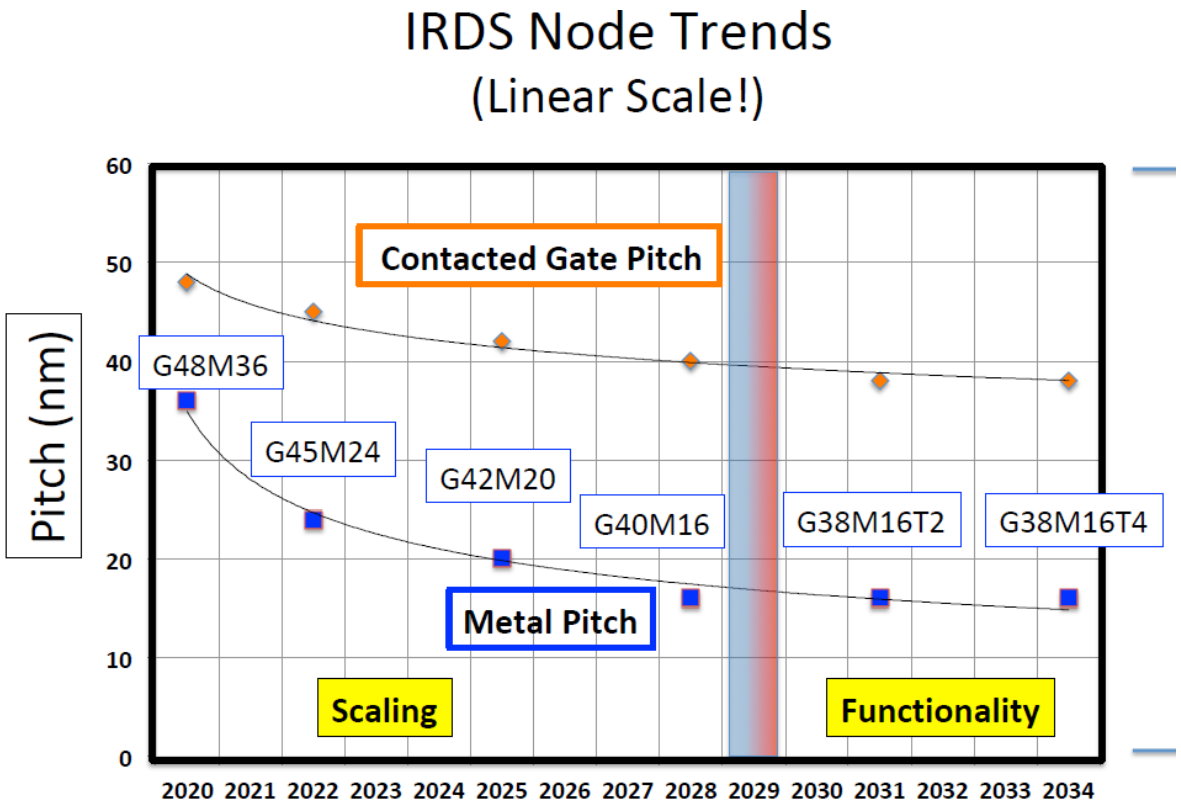
T=Number of Tiers of stacked transistors

GXXMYYTZZ

Functions per unit area



Technology node trends - IRDS



IRDS 2020 logic device roadmap 2020-2034

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 1	finFET LGAA 2	LGAA	LGAA	LGAA-3D 5	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 4	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	3DVLSI	3DVLSI
Stacking generation inflection	2D	3 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, 3DVLSI: Mem-on-Logic with Interconnect	3D-stacking, 3DVLSI: Logic-on-Logic

Source; IRDS 2020 Edition, More Moore

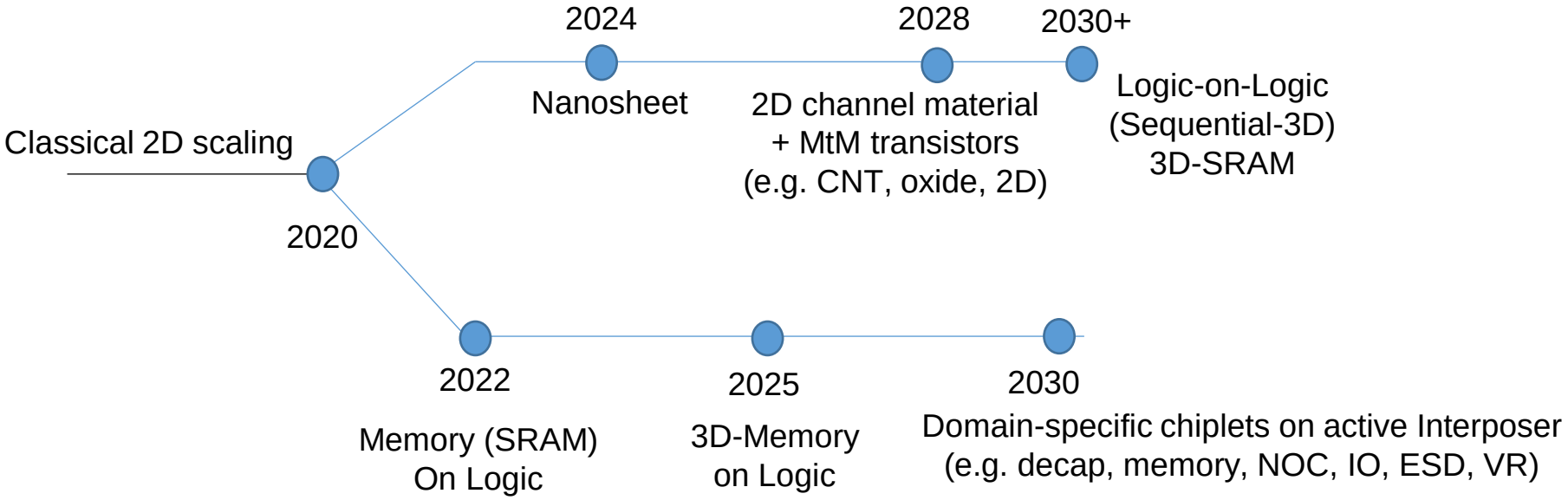
1. FinFET – leading device option until 2025
2. LGAA – potential entry around 2022
3. 3D fine-pitch stacking assembly projected to picking up 2022 for memory-on-logic applications
4. 2D materials as a potential channel material to further scale the gate length and Ceff reduction
5. Beyond 2031 – true 3D device stacking

G: Gate Pitch (nm)
M: Metal Pitch (nm)
T: Tiers
IDM: Integrated Device Manufacturer
eq: equivalent
e: equivalent
LGAA: Lateral Gate All Around

EUV: Extreme UltraViolet
DP: Double Patterning
NA: Numerical Aperture
Fe: Ferroelectric
RMG: Replacement Metal Gate
W2W: Wafer to Wafer
D2W: Die to Wafer

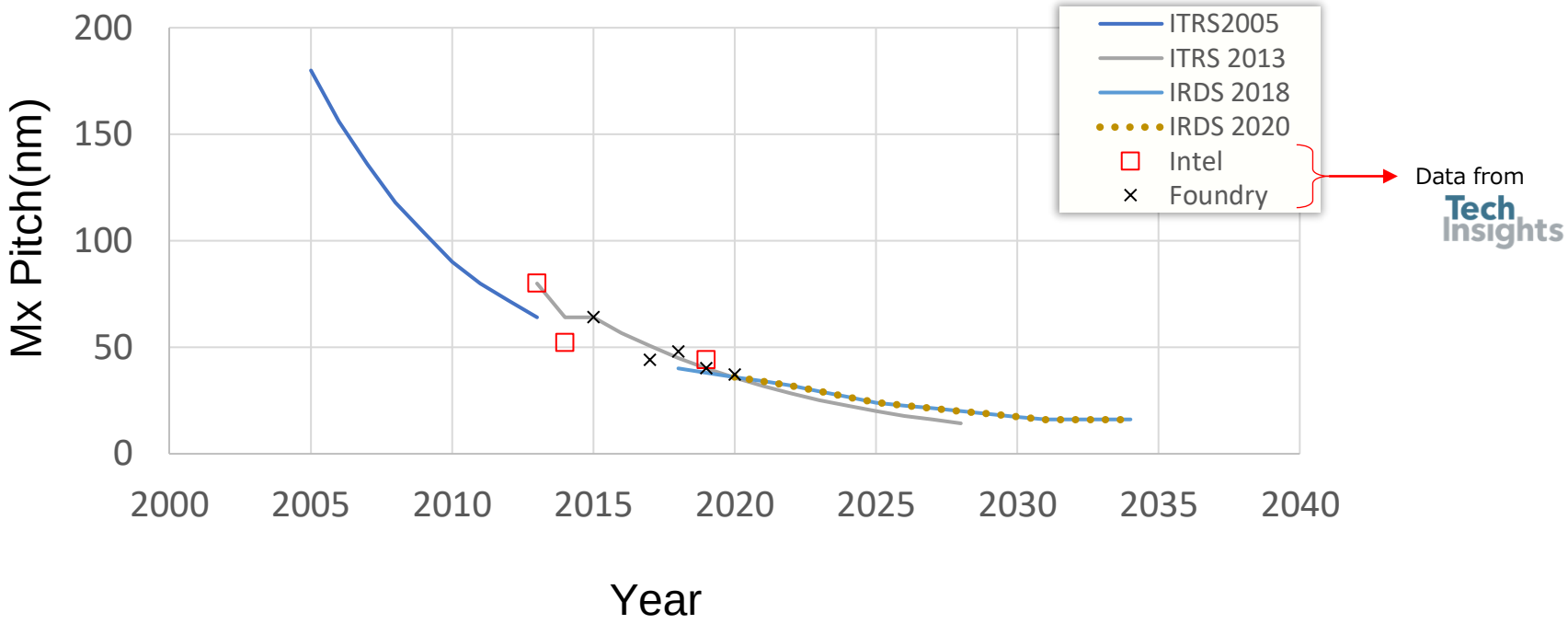
2D and 3D scaling routes should be hand-to-hand

2D scaling for energy-efficient computing – 3D compatible device

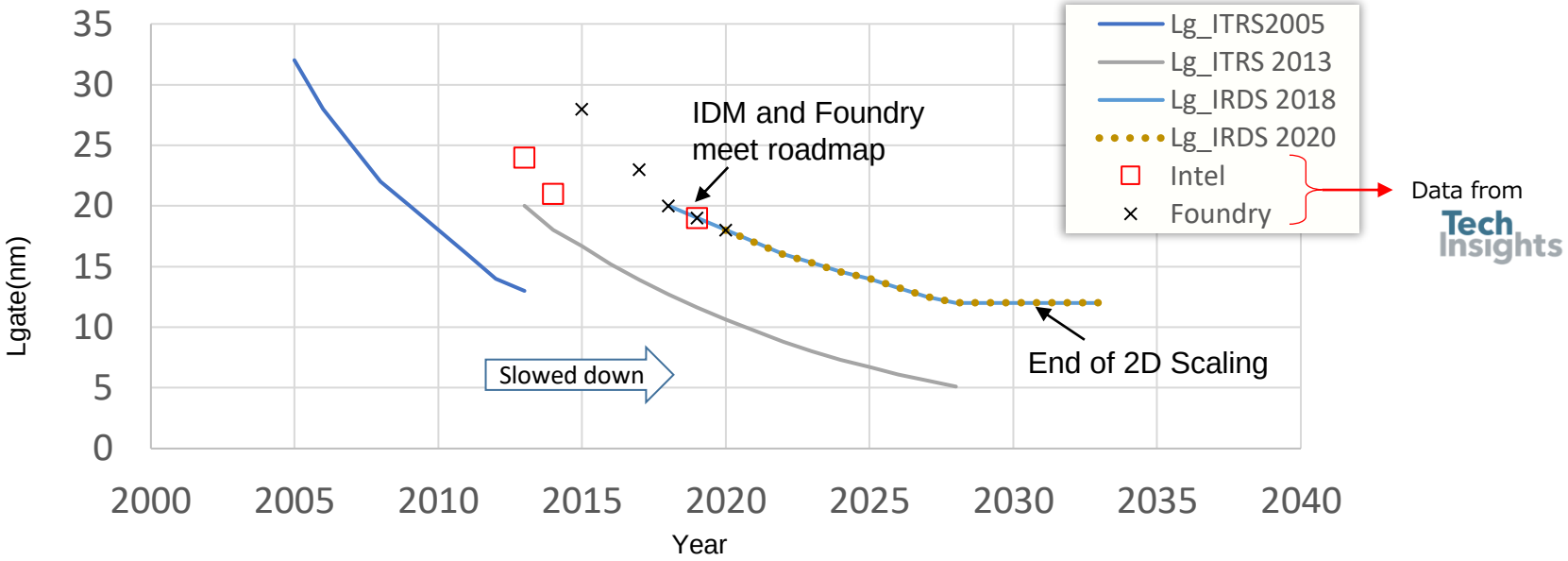


2.5D/3D+chiplet assembly high-speed/bandwidth bus + memory integration for increased system throughput – TOPS/Watt, TOPS/mm2

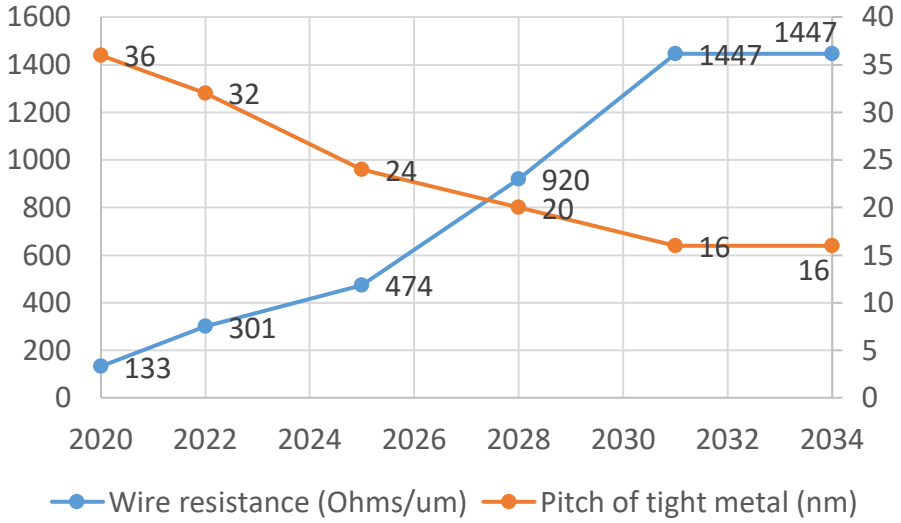
Mx Pitch Scaling Trend

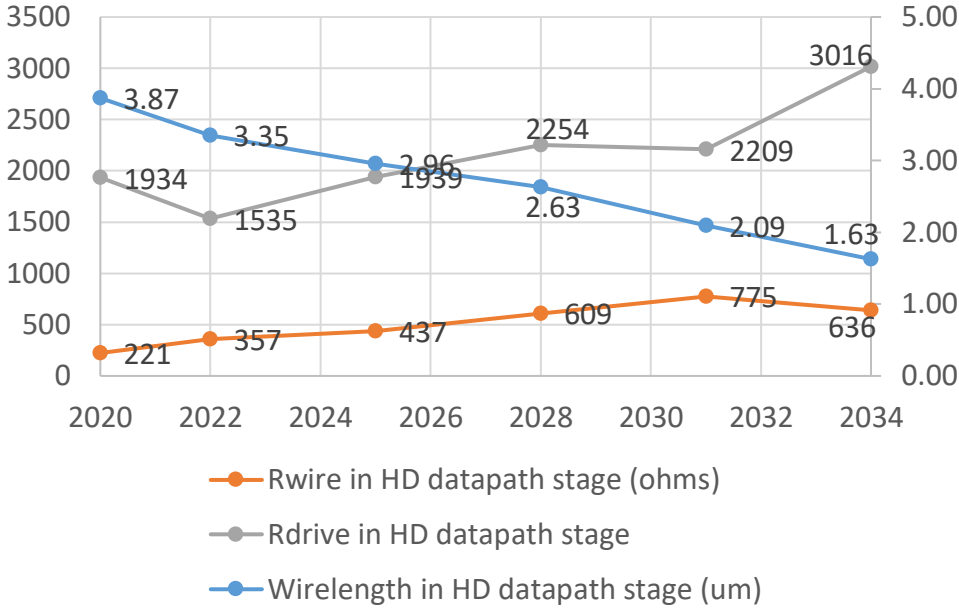


Lgate scaling trend



Rwire / um compensated by wiring hierarchy and area scaling

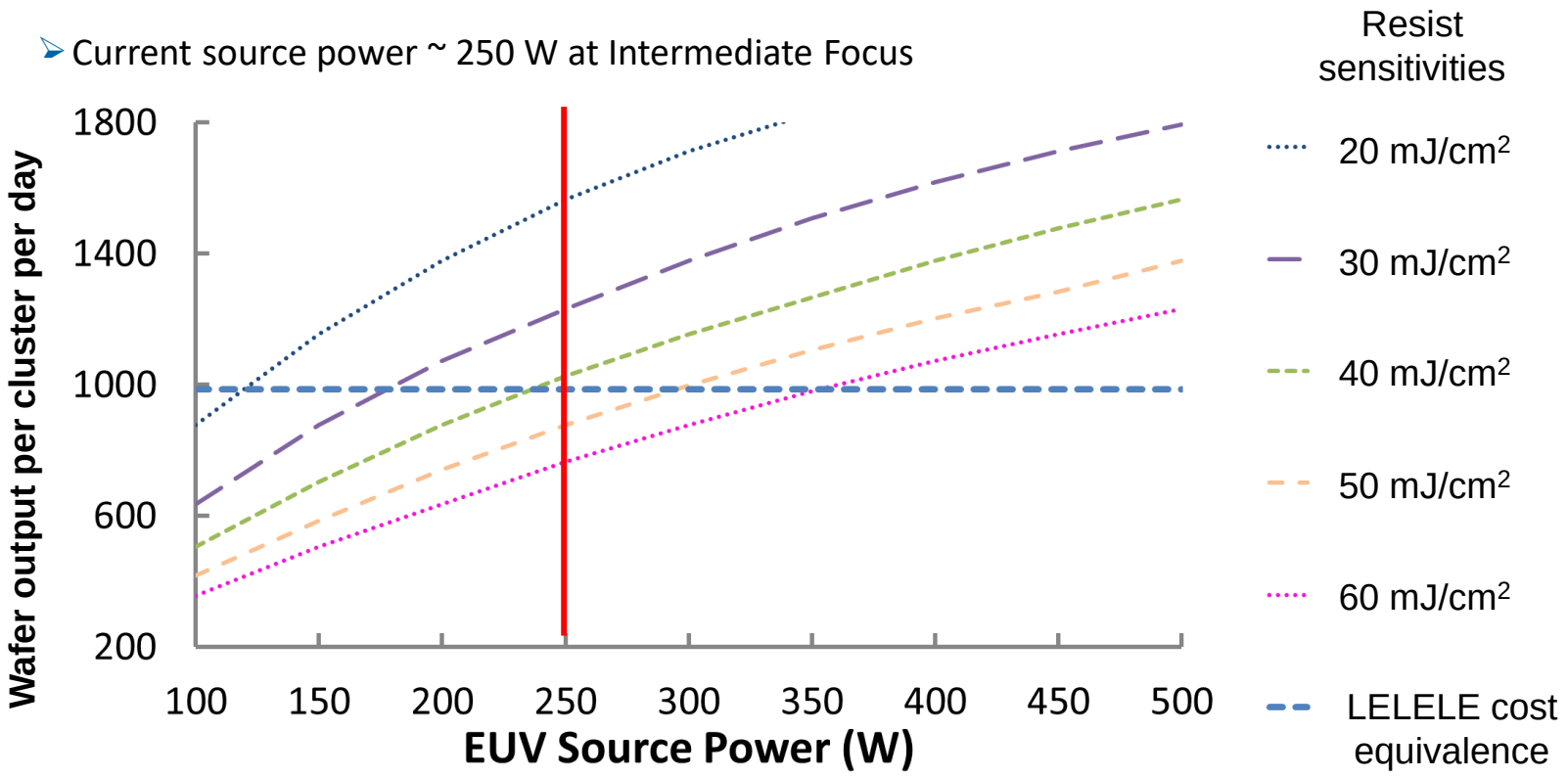




- Exponential increase of wire resistance per um length
- Compensated by increasing the wiring layers and reduced via resistance
- Reduced via resistance significantly contributor to the reduced IR drop

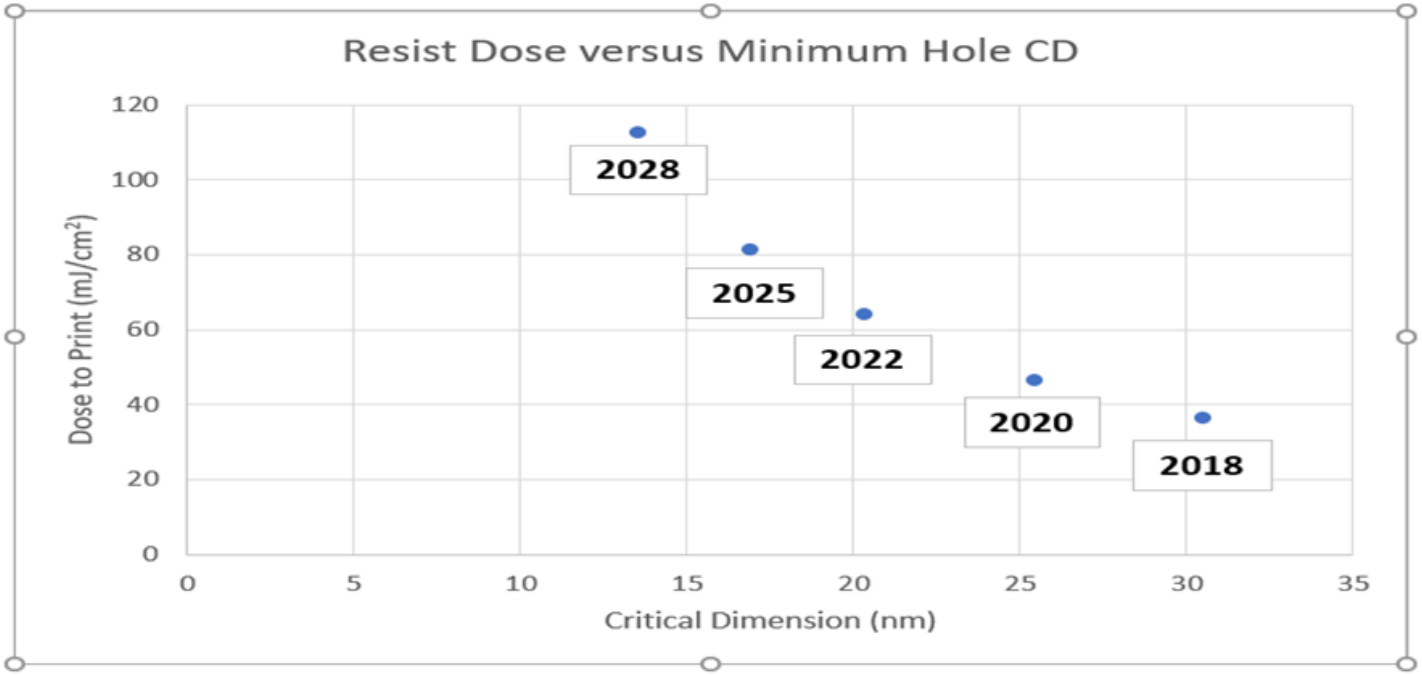
EUV Needs continual increases in Source Power

- Current source power ~ 250 W at Intermediate Focus



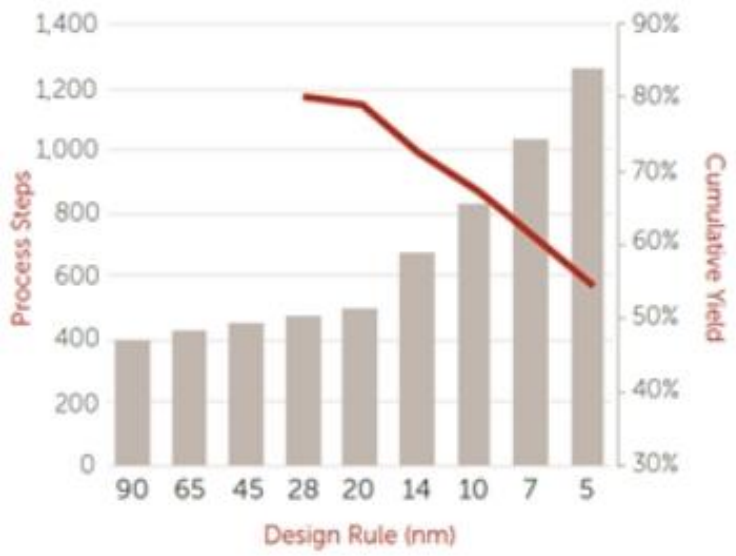
Courtesy: Lithography IFT chair, Mark Neisser

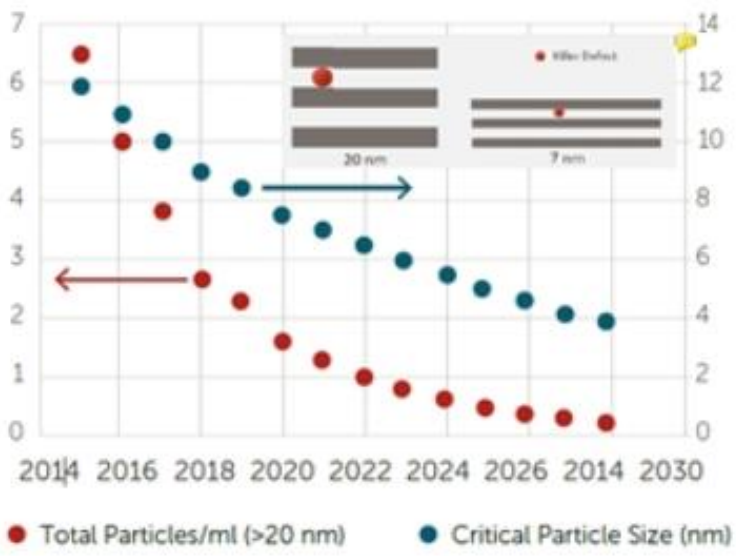
Imaging Noise will Force the use of Slower Photoresists



Courtesy: Lithography IFT chair, Mark Neisser

Critical defect size and process complexity w/ more steps in wiring

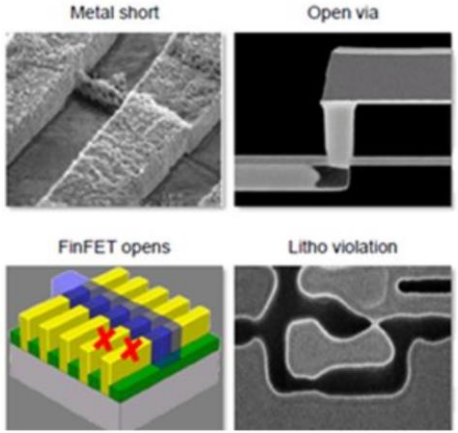




Source: Entegris @ Semiconductor Digest 2019

Defect ranking – no direct winner

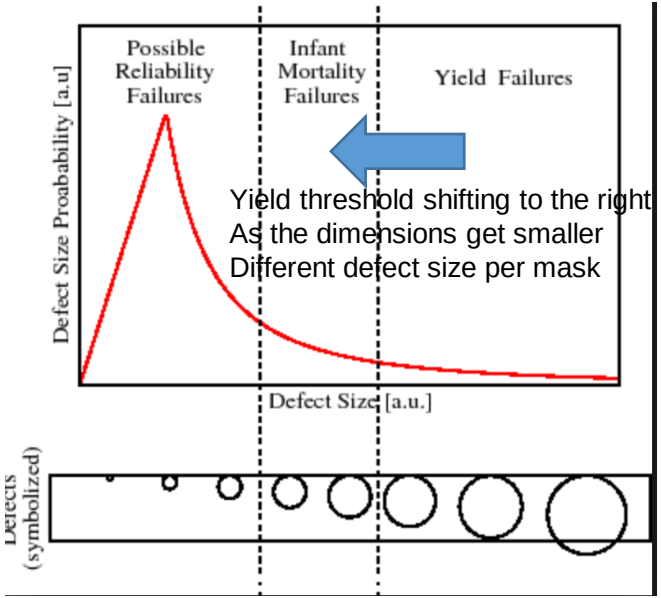
- Metal Density - ~20-40%
- Via density - 2-5%
- Fin density – 10-15%
- Gate density – 10-30%
- Criticality – Density x (1um/Defect Size)xk
  - k = Process Risk x (1/tolerance)
  - Critical Defect Size = CD/4
  - Metal = 0.4x(1um/12nmx4)=133xk
  - Via = 0.05x(1um/12nmx4)=17xk
  - Fin = 0.15x(1um/6nmx4)=100xk
  - Gate = 0.30x(1um/16nmx4)=75xk

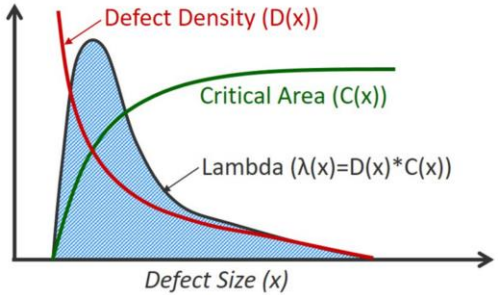


Source: eInfochips

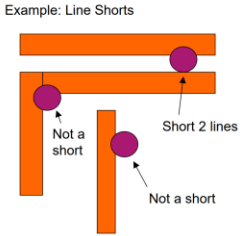
- No direct winner as some has “dirty” process and/or “less tolerating” failures
  - Metal lines could “tolerate” shorts –driver strength and noise margin threshold
  - Vias less likely tolerate defects – direct open could lose the connection
  - Fin process is relatively “clean” and “tolerate” defects due to multiple fins but fail in small-size defects (where defect density is high)

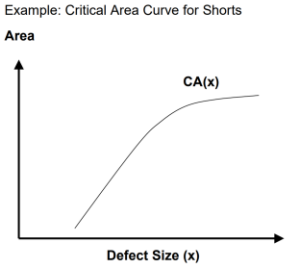
Defect size and criticality





Source: PDF Solutions





Critical Area Analysis (CAA) is a DFM technique that measures the susceptibility of a specific layout to random defects and indicates **areas** of the layout where design modifications can have the greatest positive impact on overall **yield**.

finFET (2022) - IRDS 3nm

2022	Steps/ layer	Layer Count	Step Count	Comp	Overall Complexity	Width (nm)	Defect Size (nm)	Density@Chip	Critical Area (cm2) in 80mm2	Dx/wafer	1/(A*Dxi)^n	1/(A*D0i)^n
Gate	2	1	2	1	0.75	18	9.0	32%	0.256	5.0	0.994	0.978
Fin	2	1	2	1	0.75	6	3.0	12%	0.092	136.1	0.980	0.978
Contact Via (VC)	2	1	2	2	1.50	18	9.0	3%	0.027	5.0	0.999	0.978
Contact Metal (MC)	3	1	3	2	2.25	18	9.0	37%	0.296	5.0	0.979	0.978
Local Via (V0)	1	1	1	2	0.75	12	6.0	2%	0.016	17.0	0.999	0.978
Local Metal (M1)	5	1	5	2	3.75	12	6.0	46%	0.370	17.0	0.906	0.978
Tight Pitch Via (Vx)	1	2	2	2	1.50	16	8.0	1%	0.010	7.2	0.999	0.978
Tight Pitch Metal (Mx)	3	2	6	2	4.50	16	8.0	46%	0.370	7.2	0.935	0.978
Routing Via (Vy)	1	14	14	0.5	2.63	40	20.0	1%	0.005	0.5	1.000	0.978
Routing Metal (My)	1	14	14	0.5	2.63	40	20.0	46%	0.370	0.5	0.994	0.978
	21	38	51	**SUM**	21.00			23%	0.181		0.801	0.800
				AVG	2.100							

Key numbers

D0 (def/cm2)		0.059
D0 (def/inch2)		0.381
Defect improvement factor		0.052
Block Area	13	0.750
Total complexity target	61	21.00
Maskcount target	59	86
Comlexity targeting factor		0.375

Average defect density

k-factor

Same-function chip area scaling factor wrt 2020

Process complexity ~ #masks, critical steps

$$Y_r = \frac{1}{(1 + AD_0)^n}$$

finFET (2022) vs GAA (2025) comparison

finFET (2022)

D0 (def/cm2)		0.059
D0 (def/inch2)		0.381
Defect improvement factor		0.052
Block Area	13	0.750
Total complexity target	61	21.00
Maskcount target	59	86
Comlexity targeting factor		0.375

GAA (2025)

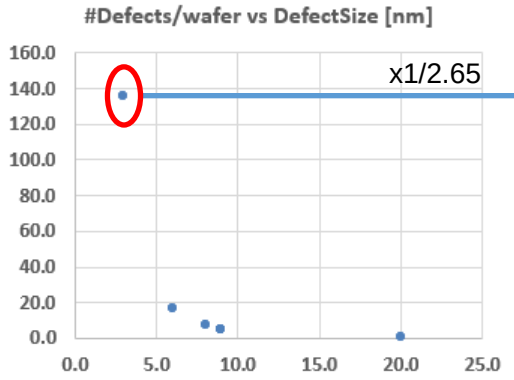
D0 (def/cm2)		0.053
D0 (def/inch2)		0.342
Defect improvement factor		0.032
Block Area	13	0.585
Total complexity target	61	21.00
Maskcount target	59	86
Comlexity targeting factor		0.375

-11%

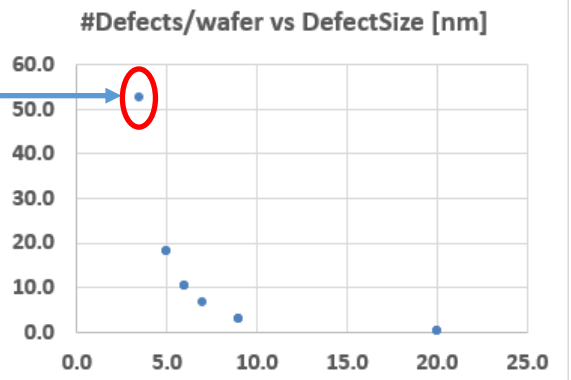
x1/1.6

x0.78

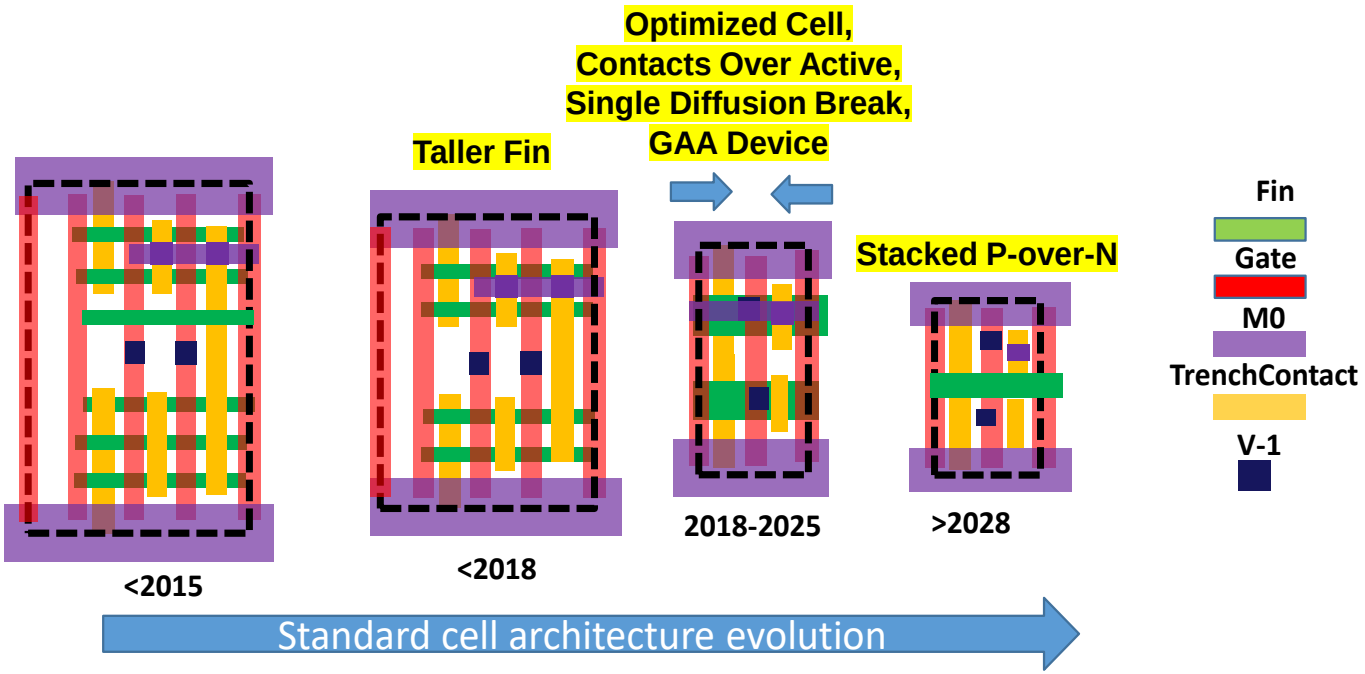
No change



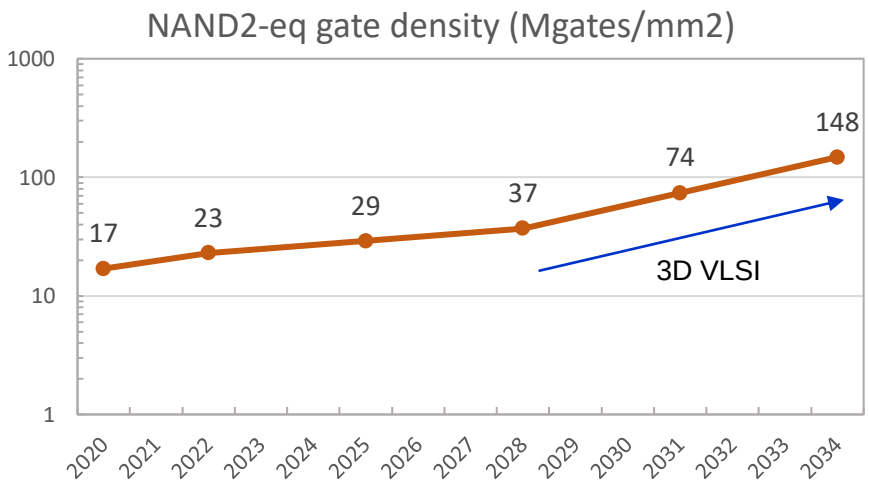
x1/2.65

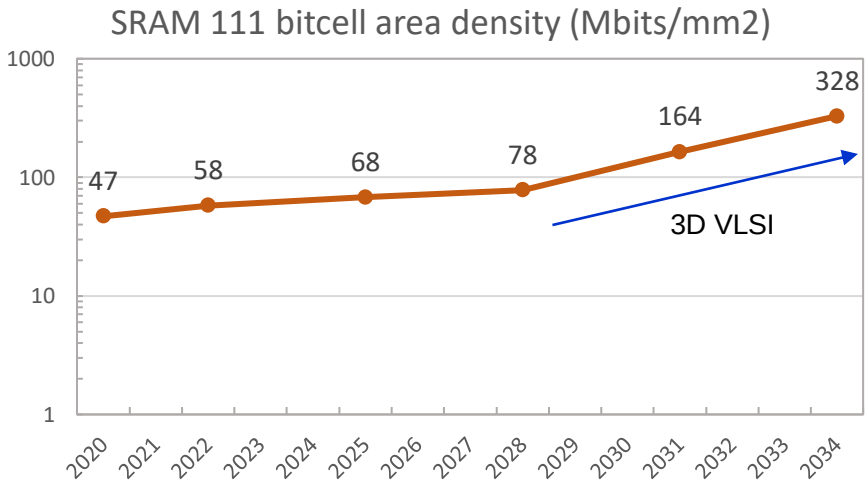


More than 50% of area scaling is tackled by DTCO in >5nm



Logic Standard Cell and SRAM scaling



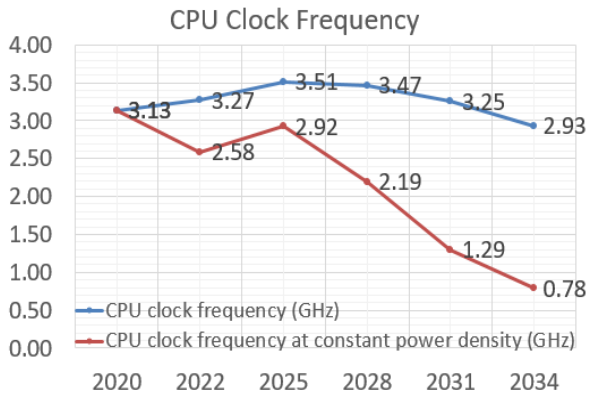


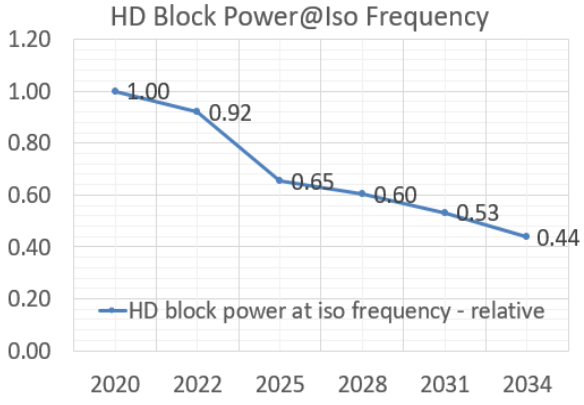
Original Source: IRDS 2020 Edition, More Moore

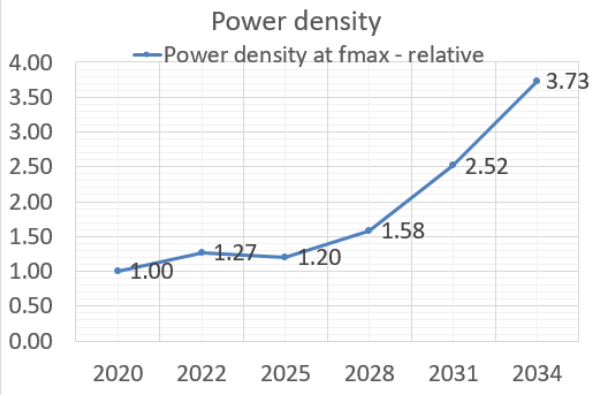
NPU: Neural Processing Unit

- Scaling - 2.2x in stdcell and x1.7 in bitcell in 3 generations (before 2025)
  - 2020: 68M transistors / mm2 in stdcell logic, 282M transistors/mm2 in bitcell
  - 2034: ~0.6B transistors / mm2 in stdcell logic, ~2B transistors/mm2 in bitcell
  - 2020: 8B transistors, 2034: >60B transistors in a 3D mobile processor (80mm2 die footprint)
- Tier-level stacking is necessary to increase #functions (after 2031)
- Gate and bitcell density improvements enable more GPU/NPU cores provided the fact that access bandwidth issues are solved

Frequency and Power Scaling





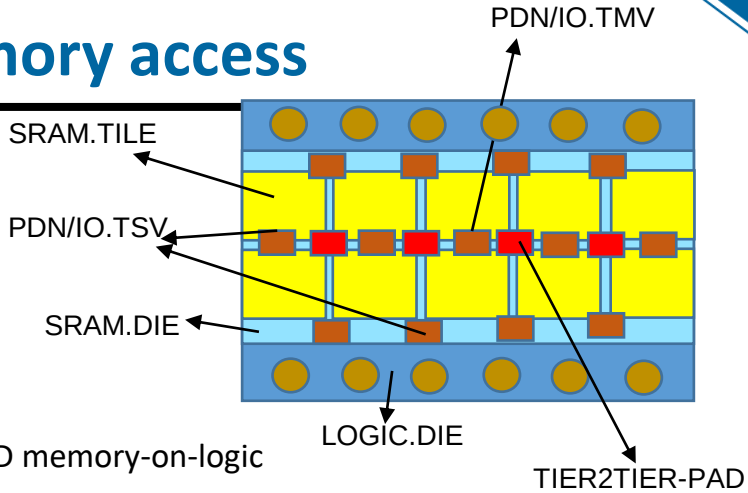


HP: High Performance
HD: High Density

Original Source: IRDS 2020 Edition, More Moore

- Frequency scaling limited because of parasitics and stalling after 2028
- Power reduction limited because of slow-down in Vdd and capacitance
- Thermal (increasing power density) reduces the average system frequency

Scaling needs a solution for the memory access



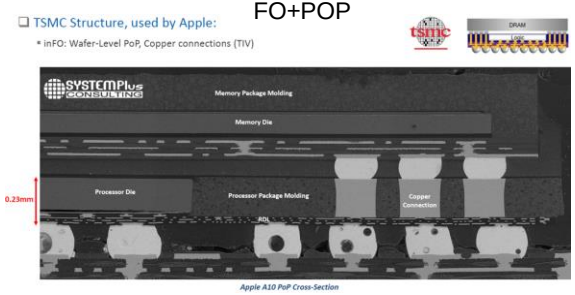
- For an execution engine running at 512 TOPS
  - On-chip memory cache access
    - 2048 TBytes/sec for L0 (4K) – difficult to feed data
    - 200 TBytes/sec for L1 (256KB)– difficult to do intra-kernel data re-use
    - 20 TBytes/sec for L2 (8MB) – difficult to do intra-kernel data re-use
  - On-package/off-chip memories
    - 1 TByte/sec for HBM3 (64GB) – size limitations of interposer size and RDL pitch
    - 512 GBytes/sec for GDDR6 (2GB)– severe bandwidth and energy constraints
- Large gap between on-chip and on/off-package memories necessitating 3D memory-on-logic
  - 2 orders of magnitude difference in memory
  - Spatial processing helps in AI to reduce latency constraints, but the demand for data in each time slot is huge. HBM evolution cannot meet the expectations.
  - Need for 2-5 TBytes/sec bandwidth access to high capacity (1GB)
- Integration challenges
  - TSV density impacting stacked SRAM cell efficiency – IO, PDN
  - Die enclosure constraints limiting the SRAM capacity (e.g. logic die on top needs to enclose the memory)
  - TSV process readiness with the advanced technology
  - Low-temperature assembly process for DRAM
- Tiled memory banks connected by Network-On-Chip reducing the overhead of pass-over TSV

3D Landscape – Manufacturing Status

Fine pitch D2W / W2W assembly	Market	Tech Characteristics	Scaling driver	Challenges
FO+POP	Mobile	chip-first/last, face-up/down, ~2-10um RDL pitch, 40um ubump pitch	Multi-stack FO, RDL pitch, #layers	Warpage, MC topography
W2W stacking w/ TSV + uBump	Memory	F2B, 40um TSV pitch, 10x50um TSV, 50um thick Si	Wafer thickness, uBump pitch	Size matching, TSV
W2W F2F di-electric/hybrid bonding	Memory, Imagers	F2F, <1um pad pitch, 1x5um TSV, 5um thick Si	Wafer thickness, uBump pitch	Size matching, TSV, pad density, alignment (~200nm) limits
Si interposer w/ TSV and w/o TSV	HPC, Network	<1um pitch RDL, 40um ubump pitch, 40um 10x100um TSV pitch, >1500mm2 Si area	Si size, #layers, decap/ESD co-integration	High cost (~$600 wafer cost adder), test
EMIB (Embedded Multi-die Interconnect Bridge) in laminate	HPC	<1um pitch interconnect	FO compatibility	EMIB Alignment to FO/laminate

Fine-pitch 3D stacking assembly technologies today

FO+POP

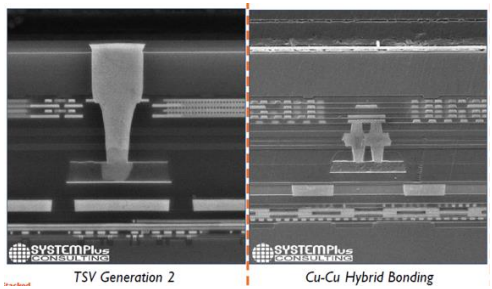


Apple A10 on FO+POP

Chiplets with coarse/fine interconnections (multi-sourcing of KGP)

W2W F2F di-electric bonding+TSV

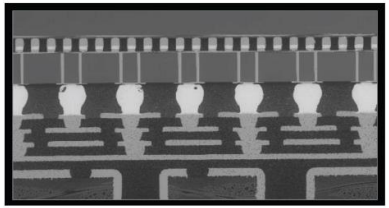
W2W F2F hybrid bonding w/o TSV



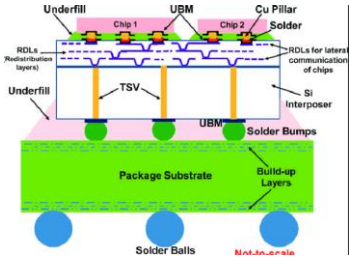
Sony: CMOS Image Sensors

Fine-pitch stacking (~1-2um pitch) (form factor and performance critical)

Si Interposer + TSV

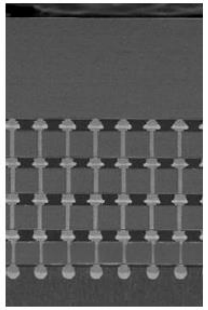


AMD Fiji GPU



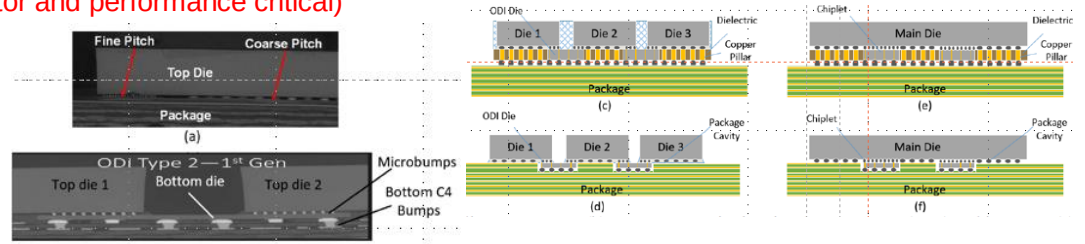
Hetero-integration fine-pitch interconnect in 2.5D (HBM2E and high-performance computing)

W2W stacking+ TSV+uBump



SK Hynix HBM memory

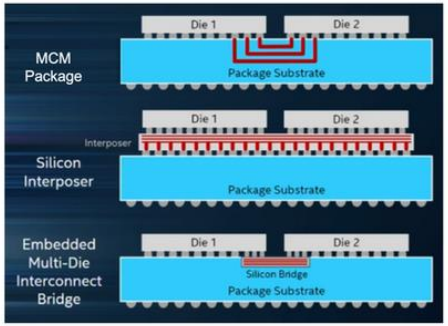
12 tiers in 3D, 40um pitch (HBM2E and AI)



Intel: Omni-directional interconnect

Chiplets with coarse/fine interconnections (HBM2E and AI)

Need for open interfaces to extend beyond Memory+Accelerator

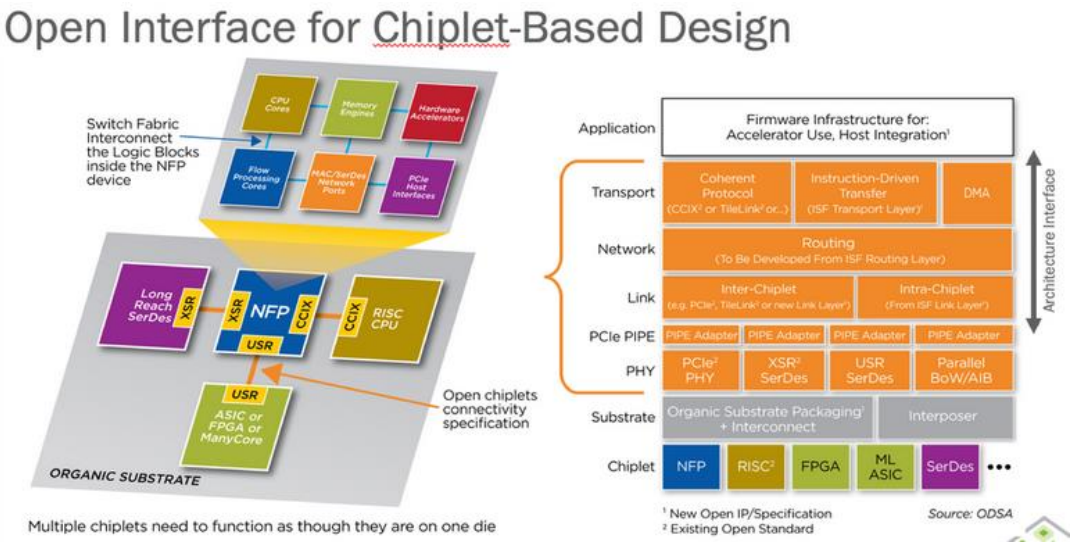


- Cheap, Mature, Reliable
- Screening for KGD (known-good die)
- Low density die-die traces
- High power using basic CMOS IO

- High density die-die traces (6-8x)
- Low interconnect power (<0.5pJ/bit)
- Poor screening for KGD
- High cost/Low yield of large Interposer

- High density die-die traces (6-8x)
- Low interconnect power (<0.5pJ/bit)
- Lower cost to Interposer
- Poor screening for KGD
- Limited routing length (<5mm)

Observation: An ultra-low-power die-die Interface with ~8x higher throughput/trace than that of Interposer/EMIB makes the MCM package the best SIP solution

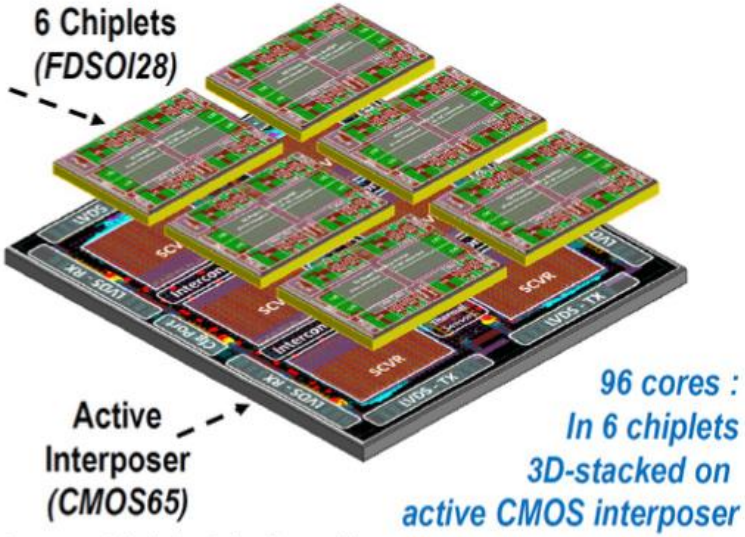


Source: ODSA (Open Domain-Specific Architecture)

Active Interposer – ISSCC 2020

6 Chiplets 3D-stacked on an Active Interposer

- Chiplet Overview
  - 4 cluster of 4 cores
  - Distributed L1$ + L2$ + L3$
  - Scalable Cache Coherency
- Active Interposer
  - Distributed flexible interconnects
  - Integrated SCVRs (1/chiplet)
  - Memory Controller & System IO's
  - SOC Infrastructure, DFT



➢ *2 technology nodes difference between chiplets & bottom die*

Source: Leti @ ISSCC2020

Conclusions

- Mobile computing and Cloud+AI+HPC driving More Moore scaling
- Slow-down in pitch scaling tackled with DTCO
  - Lateral-GAA - viable path for high PPA value
  - Monolithic/Sequential 3D integration needed beyond 2028-203
  - EUV source power scaling needs to be accompanied with the noise introduced by the resists to reach CoO targets
  - Main challenge is the defect-free clean of high-aspect ratio structures in lateral/vertical trenches
  - finFET transition to GAA requires a significant improvement in the defect density factor (x0.78) while (x1/2.65) for the reduction of minimum defect size
- Memory bandwidth and latency critical for many applications
  - Read bandwidth and capacity getting more critical for embedded NVM
  - Accompanying IO and bus solutions necessary to couple memory to logic
  - 2.5D/3D requiring various memory hierarchies to be integrated, such as stacked SRAM/MRAM/DRAM
- 3D stacking approaches
  - Motivating cases are memory-on-logic in large chips or product bundling/time-to-market needs
  - Active interposer are needed for interoperability and bussing needs of chiplets