ICOS
International Cooperation
On Semiconductors

IRDS More Moore Roadmap for edge and cloud computing

Mustafa Badaroglu

IEEE-IRDS More Moore Roadmap Chair

Cloud and edge computing w/ intelligent connectivity

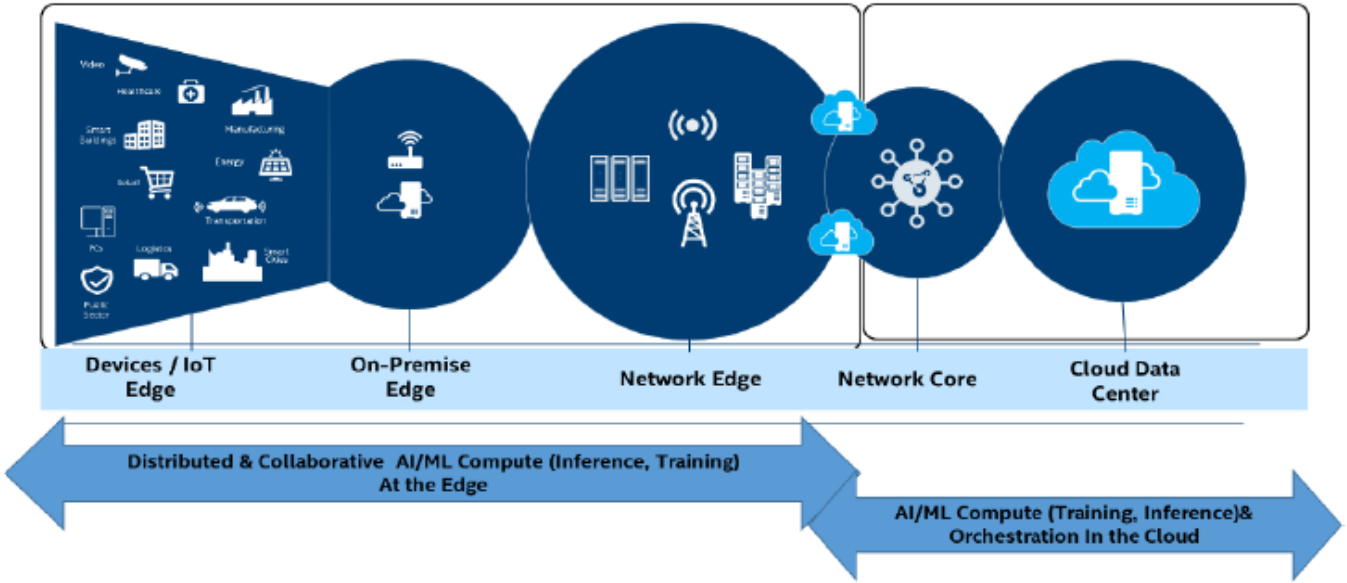
- Device-interconnect-memory technologies for mobile and cloud+HPC computing
  - Mobile computing - additional functionality, biometrics, and display/camera/sensing for increased consumer value
  - Cloud+HPC computing - 2.5D/3D integration for data-abundant computing
- Convergence of edge and cloud computing by 5G/6G and distributed AI

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**

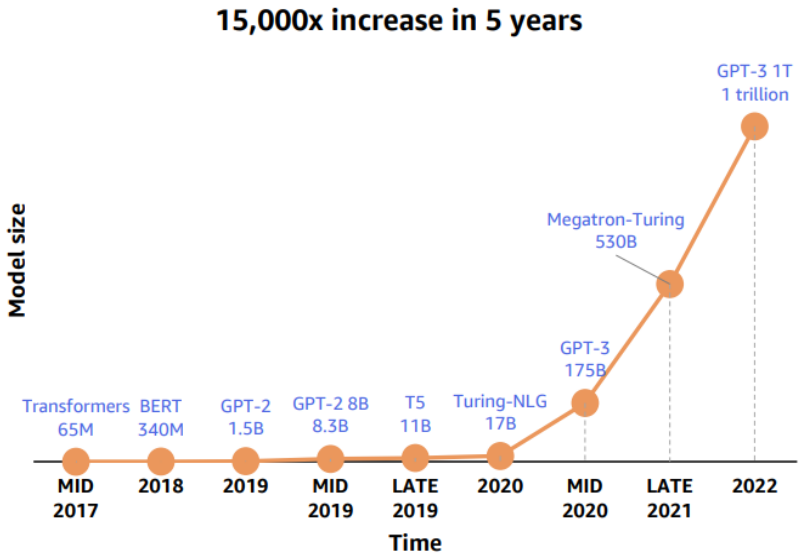
Mobile computing
AI at edge requires fast and high-density memories



Source: 5G At the Edge, 5G Americas, 2019

New directions on system scaling – More Moore

- Semiconductor content in applications
  - Increasing die size due to memory and machine learning – era of parallel computing
  - Leading edge nodes in multiple platforms: consumer, cloud, data center, industrial, automotive
  - Large LLM (e.g. ChatGPT) - next-gen DRAM adoption
  - System reliability - always-on nature (24/7) and RF/power/logic convergence
- Time-to-market
  - Sophisticated integration - e.g. AR glass, fog-cloud-edge fusion, form factor
  - Need to deploy product fast
- Process complexity/heterogeneity
  - High-aspect ratio stacked transistors
  - New materials (e.g. resists, di-electric, metallization, 3D) for processing
  - New unit processes – direct metal etch, backside process, GAA inner spacer, CFET, …
  - More process steps – advanced device, interconnect, more materials, 3D integration



Source: AWS@2022

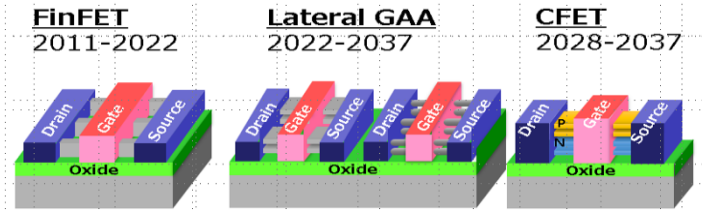
More Moore technology market drivers

- Mobile applications
  - Heterogenous integration of NPU, CPU, GPU, and domain-specific ASICs
  - Extreme reality (VR/AR) for enriched media
  - Real-time NLP and LLM
  - Edge-AI (mobile phone, smart cameras/speakers, sensor fusion)
  - Seamless office
- HPC applications
  - AI accelerators in enterprise/cloud
  - Codec ASICs - 24/7/365 continuous run of video and audio (codec), 5 years minimum time
  - Networking – Always-on, 500W power envelope
  - ADAS chips – Autonomous driving
  - Memory and IO solutions for AI, graphics, HPC
- Sustainable compute platforms with novel fabrics
  - Neural processing unit
  - Fine-pitch 3D stacking
  - Reconfigurable compute fabrics
  - Smart 2.5D interposers

More Moore technology targets

- Mission – Roadmap of unified technology platform providing concurrent physical, electrical and reliability enablements for logic and memory technologies to sustain PPAC (power, performance, area, cost) scaling and system energy/area-efficient performance for edge and cloud applications
- Node-to-node PPACTS scaling targets – Logic Datapath
  - (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
  - (S)chedule: 1y product cycle for consumer-mobile, 3y product cycle for data center
- Node-to-node System system scaling targets – Memory and CPU combined
  - TOPS: Throughput
  - TOPS/W: Energy efficiency
  - TOPSxTOPS/W/Area: Energy-efficient performance (aka 1/EDP: inverse of Energy Delay Product) at unit area
  - 2.4x scaling of TOPSxTOPS/W/Area every logic generation – (1.15x1.35/0.65~2.40)
  - Per Frame (GPU), Per inference (edge), Per training (cloud), Per Pocket (network)

2 complementing routes for More Moore scaling



2D scaling for energy-efficient computing at scaled area

Classical 2D scaling

2022 — Nanosheet

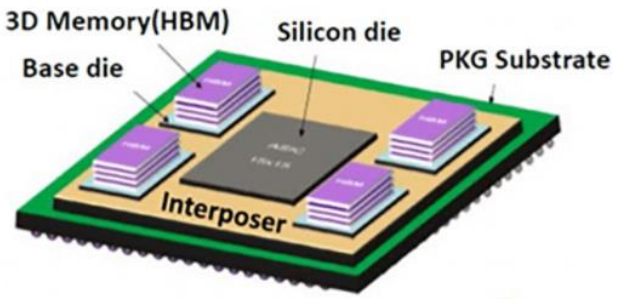
2028 — CFET SRAM

2030+ — Logic-on-Logic (Sequential-3D) 3D-SRAM

Source: IRDS 2022 More Moore roadmap

2020

System-technology-co-optimization
e.g. System Infrastructure to boost scaling



2022 — Memory (SRAM) On Logic

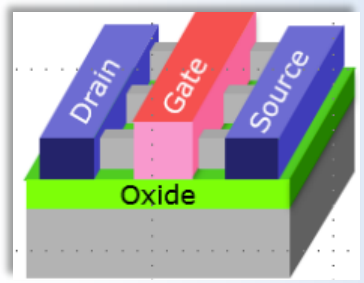
2025 — 3D-Memory on Logic Backside PDN

2028 — Backside Devices Logic on Logic

2030+ — Domain-specific chiplets on active interposer (e.g. decap, memory, NOC, IO, ESD, VR)

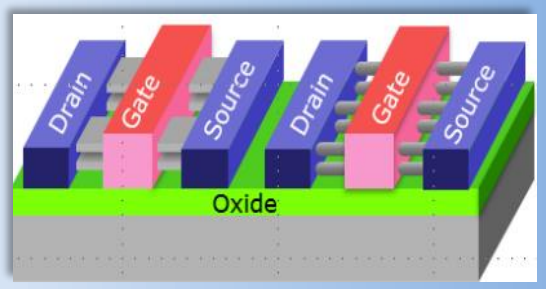
Evolution of Device Architectures

FinFET
2011-2022



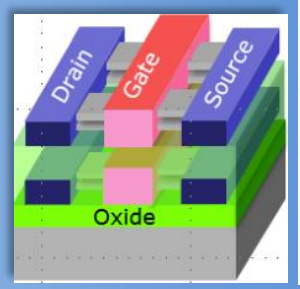
- Increasing drive by taller fin
- Better channel control for better perf-power

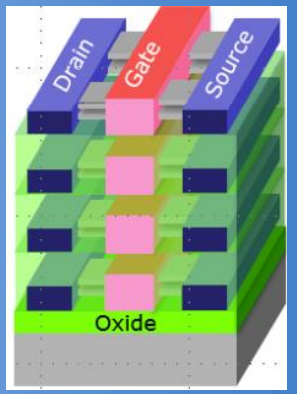
Lateral GAA (=> Backside-PDN, Direct Metal Etch, Seq/Mono-3D)
2022-2034



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

3D VLSI
2031 - 2034

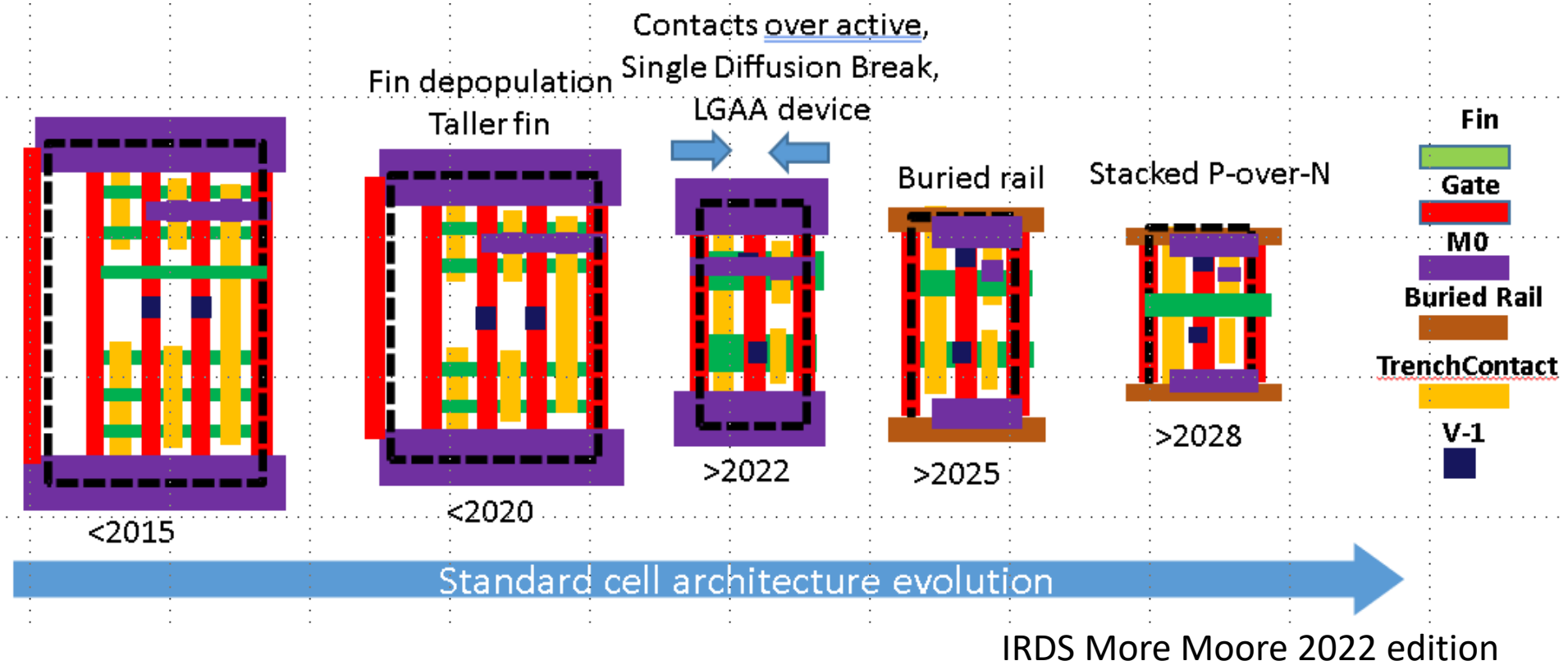




- Sequential/Monolithic heterogenous 3D integration/fine-pitch stacking (e.g. memory, IO)

Source: Modified from IEEE IRDS More Moore 2022 edition

Buried power rail by >2025 and CFET by >2028



IRDS More Moore 2022 edition

- Front-side and back-side wafer processing
- High-aspect ratio vias and metallization (e.g. buried rail) below the device

Logic device roadmap 2022-2037

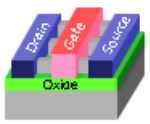
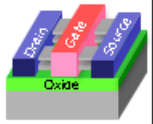
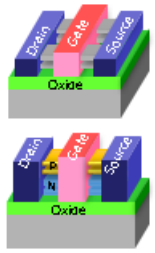
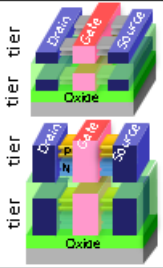
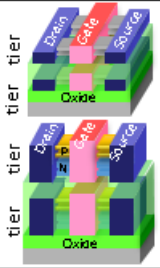
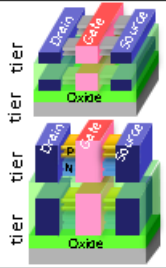
YEAR OF PRODUCTION	2022	2025	2028	2031	2034	2037
	G48M24	G45M20	G42M16	G40M16/T2	G38M16/T4	G38M16/T6
Logic industry "Node Range" Labeling	"3nm"	"2nm"	"1.5nm"	"1.0nm eq"	"0.7nm eq"	"0.5nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI
Logic device structure options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

Source: IRDS 2022 More Moore roadmap

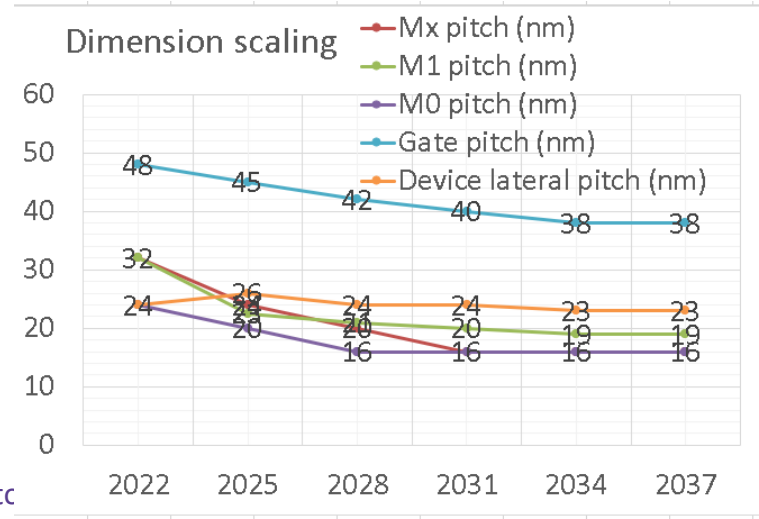
G: Gate Pitch (nm)
M: Metal Pitch (nm)
T: Tiers
eq: equivalent
e: equivalent
LGAA: Lateral Gate All Around

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

Logic ground rules – IRDS More Moore Logic

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

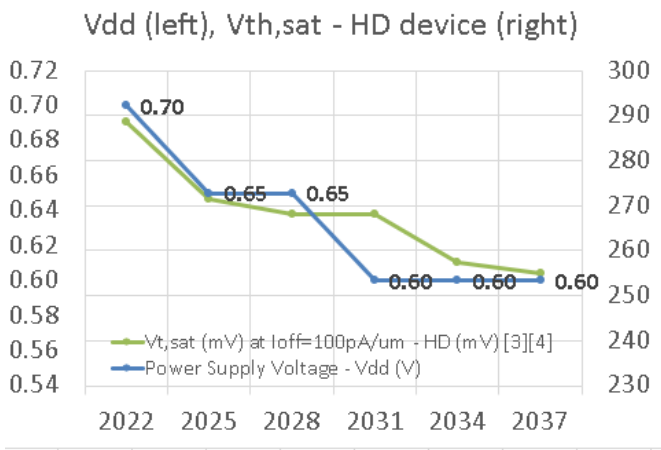
- M0 pitch critical to sustain cell area scaling
- Lg scaling saturating around 10-12nm
- 3 discrete GAA widths in SoC: HP, HD, SRAM but choosing which width to assign still flexible in GAA
- Device height, device width, and device vertical pitch critical for overall PPA scaling



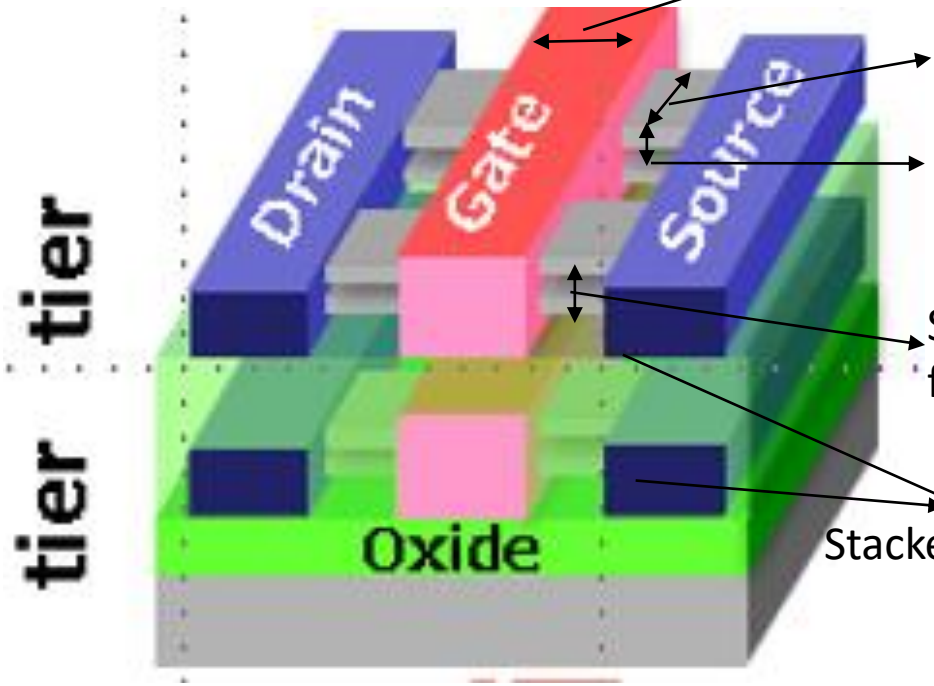
Roadmap specifications from ground rules, device, integration schemes to system

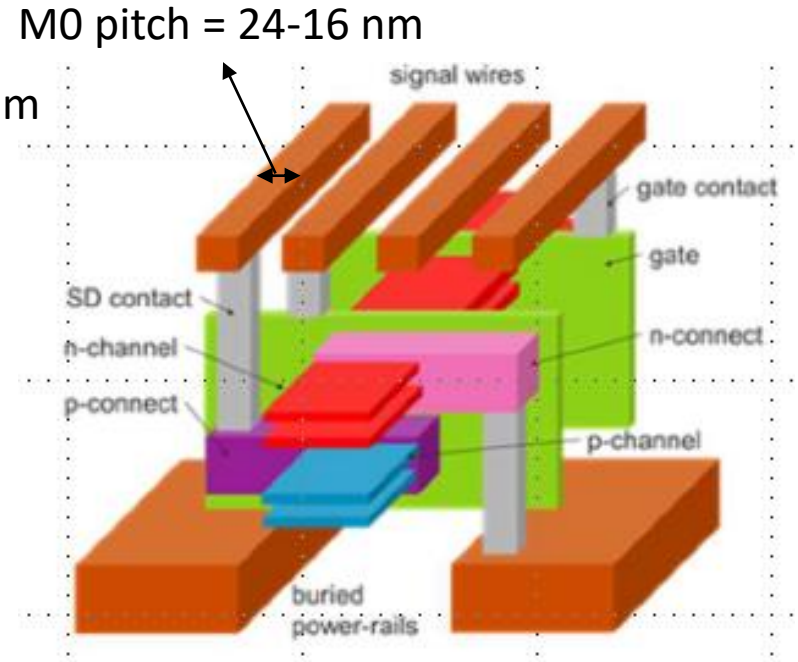
- Logic technology anchors
- 3D stacking ground rules
- Logic technology integration capacity
- Power and performance scaling
- Defectivity and yield targets
- Logic device ground rules
- Patterning
- Logic device electrical specs
- Analog device electrical FoMs
- Interconnect technology
- Logic and bitcell architecture
- Logic cell electrical FoMs

YEAR OF PRODUCTION	2022	2025	2028	2031	2034	2037
	G48M24	G45M20	G42M16	G40M16/T2	G38M16/T4	G38M16/T6
Logic industry "Node Range" Labeling	"3nm"	"2nm"	"1.5nm"	"1.0nm eq"	"0.7nm eq"	"0.5nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI
Logic device structure options	finFET LGAA	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM
Platform device for logic	finFET	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D
LOGIC DEVICE ELECTRICAL SPECS						
Power Supply Voltage - Vdd (V)	0.70	0.65	0.65	0.60	0.60	0.60
Subthreshold slope (mV/dec) - HP (mV/dec)	82	72	70	70	70	70
Subthreshold slope (mV/dec) - HD (mV/dec)	75	67	67	65	65	65
Capacitive equivalent thickness (CET) (nm) [2]	1.00	1.00	0.90	0.90	0.90	0.90
Vt,sat at Ioff=10nA/um - HP (mV)	156	165	165	164	156	154
Vt,sat (mV) at Ioff=100pA/um - HD (mV) [3][4]	288	271	268	268	258	255
Effective mobility (cm2/V.s)	125	100	80	60	40	40
Rsd (Ohms.um) [5]	271	257	245	232	221	210
Ballisticity.Injection velocity (cm/s)	9.00E+06	9.00E+06	9.00E+06	9.00E+06	9.00E+06	9.00E+06
Vdsat (V) - HP	0.092	0.101	0.108	0.144	0.216	0.216
Vdsat (V) - HD	0.104	0.101	0.108	0.144	0.216	0.216
Ion (uA/um) at Ioff=10nA/um - HP [6]	874	787	851	753	737	753
Ion (uA/device) at Ioff=10nA/um - HP [7]	88	170	184	157	118	115
Ion (uA/um) at Ioff=100pA/um - HD [8]	644	602	656	551	532	547
Ion (uA/device) at Ioff=100pA/um - HD [9]	65	130	142	115	85	83
Cch,total (fF/um2) - HP/HD [8]	34.52	34.52	38.35	38.35	38.35	38.35
Gate height over fin (nm)	20	15	10	10	10	10
Cch (fF/um) - HP [8]	0.44	0.39	0.37	0.37	0.37	0.37
Cch (fF/um) - HD [8]	0.50	0.39	0.37	0.37	0.37	0.37
CV/I (ps) - FO3 load, HP [9]	1.06	0.96	0.84	0.88	0.90	0.88
I/(CV) (1/ps) - FO3 load, HP [10]	0.94	1.04	1.18	1.14	1.11	1.14
Energy per switching [CV2] (fJ/switch) - FO3 load, HP	0.65	0.49	0.47	0.40	0.40	0.40

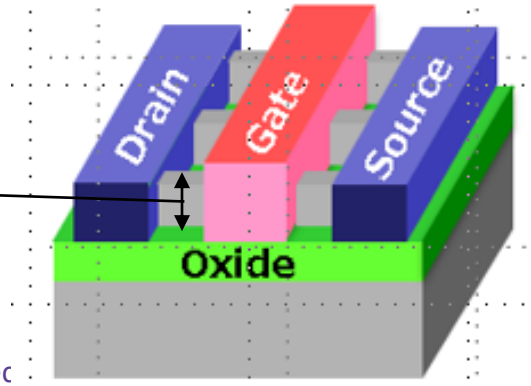


Critical dimensions across 2022-2037



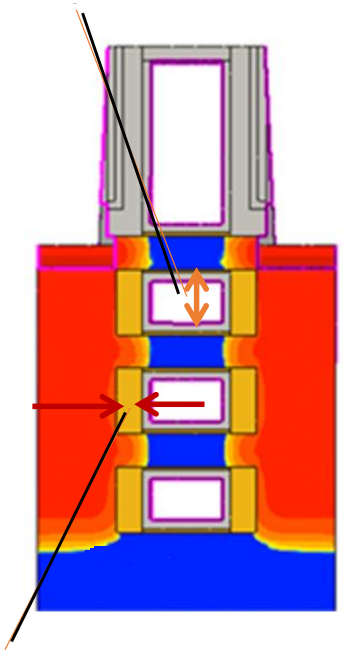


Source: TEL@IEDM2021

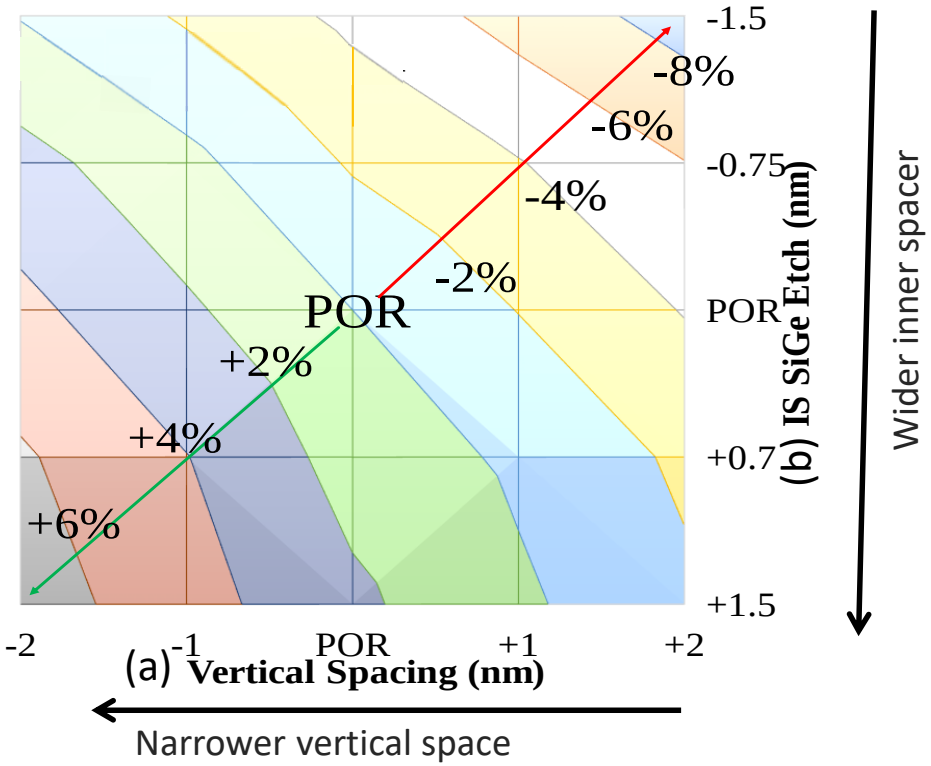


Two key geometry knobs for NSFET device

(a) Vertical space btw NS

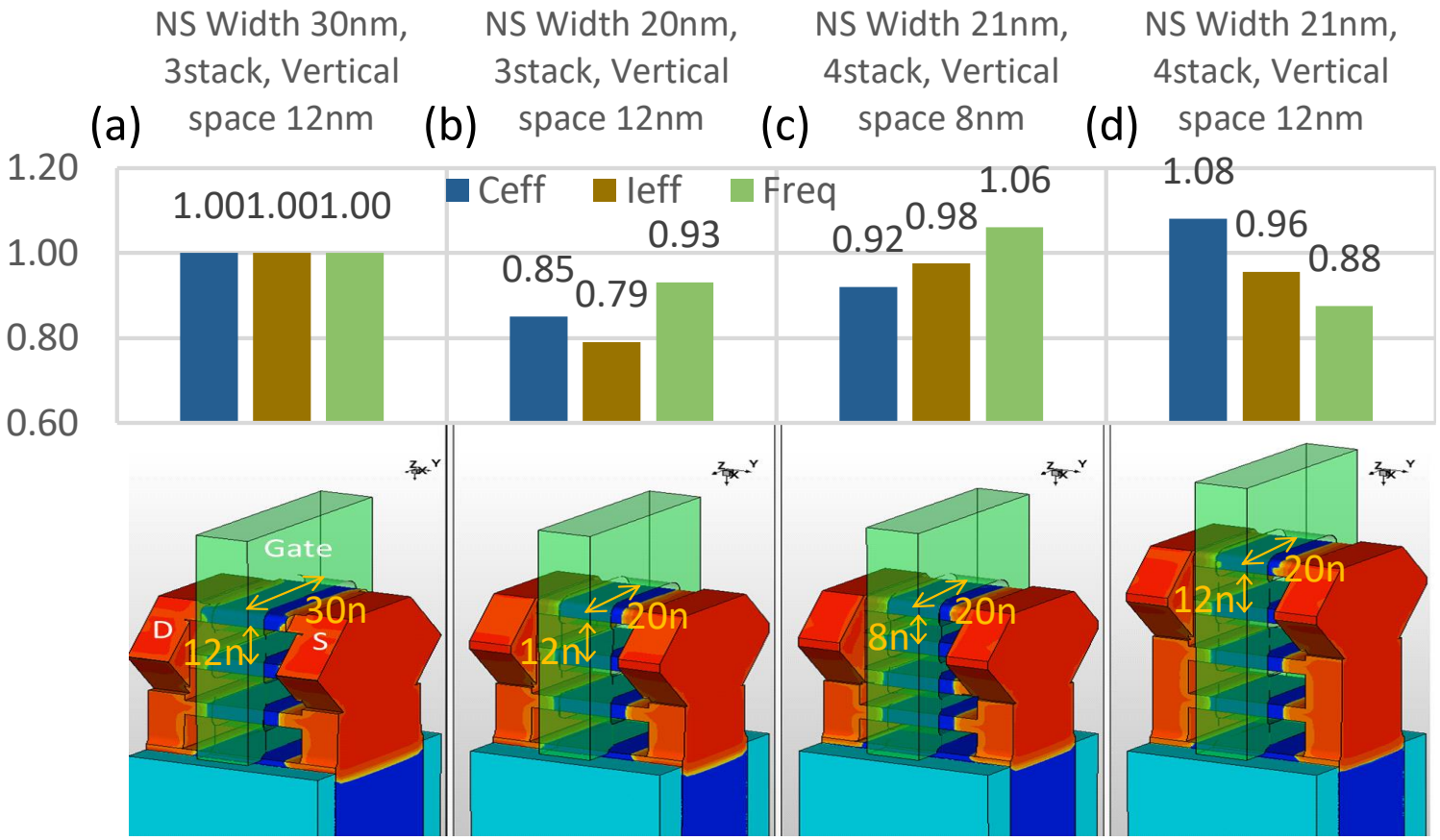


(b) Inner spacer width



Two key geometry knobs for NSFET device optimization.
(a)vertical space between NS
(b)inner spacer width.

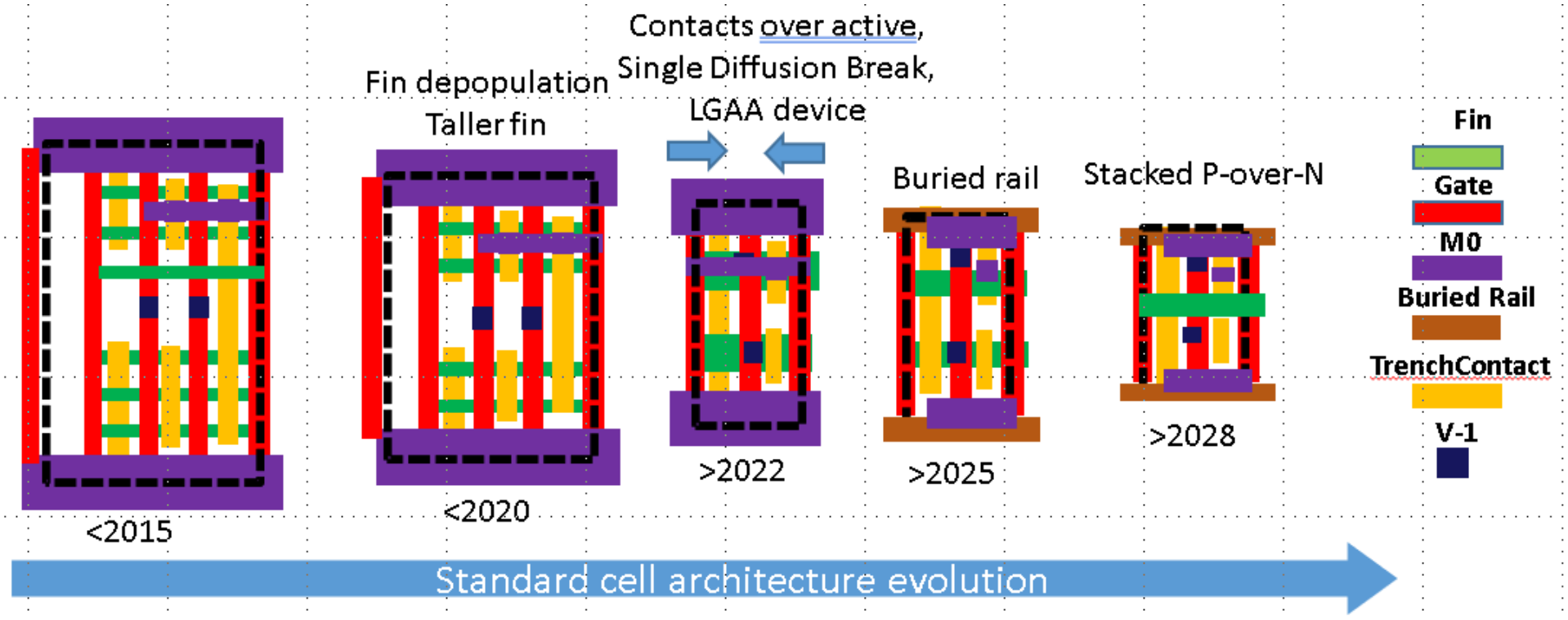
Ultra stack NSFET optimization



- (a) Three stacked NSFET with 30nm NS width
- (b) Scaling NS width from 30nm to 20nm at same vertical space.
  - Cell level perf drops as I_{eff} drops more than C_{eff} gain.
- (c) When one more NS is added at tighter vertical space,
  - perf improves due to significant I_{eff} gain much more than C_{eff} penalty.
- (d) When vertical space is kept same as control,
  - +1 NS stack degrades cell perf due to significant C_{eff} penalty.

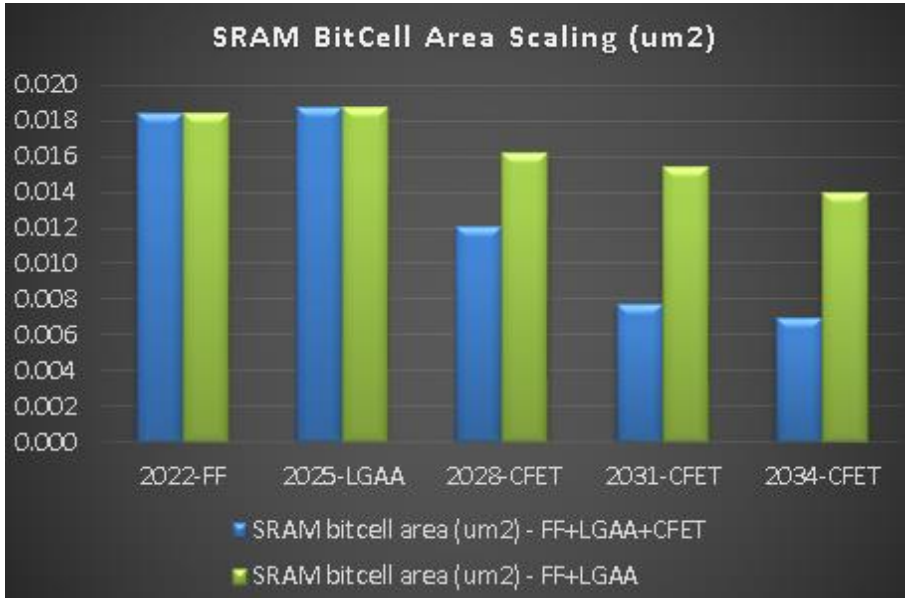
Standard cell evolution with DTCO

- Front-side and back-side wafer processing
- High-aspect ratio vias, direct metal etch, backside metallization



Source: IRDS More Moore 2022 edition

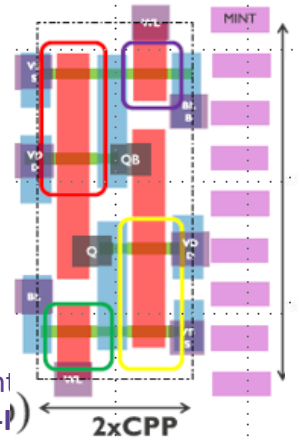
Bitcell and cell height saling



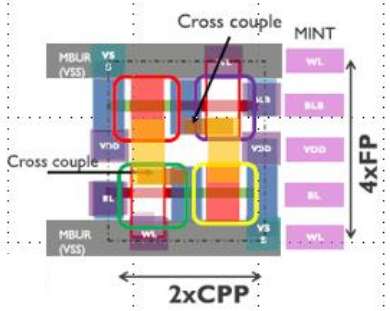
Standard cell employing finFET

Standard cell
Employing GAA + buried rail + backside metallization

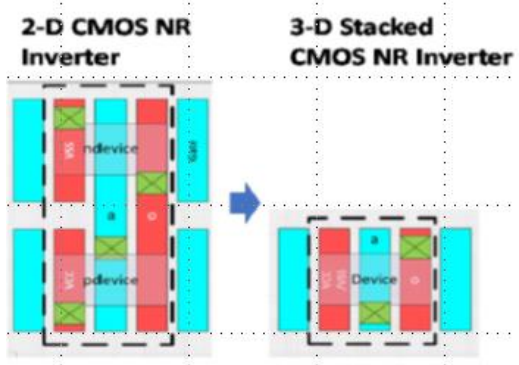
- CFET helping HD bitcell area scaling by 2028
- Buried rail enabling ~110nm cell height by 2025
- Cell height expected to reduce down to 90nm by 2028 and to 80nm by 2031



SRAM

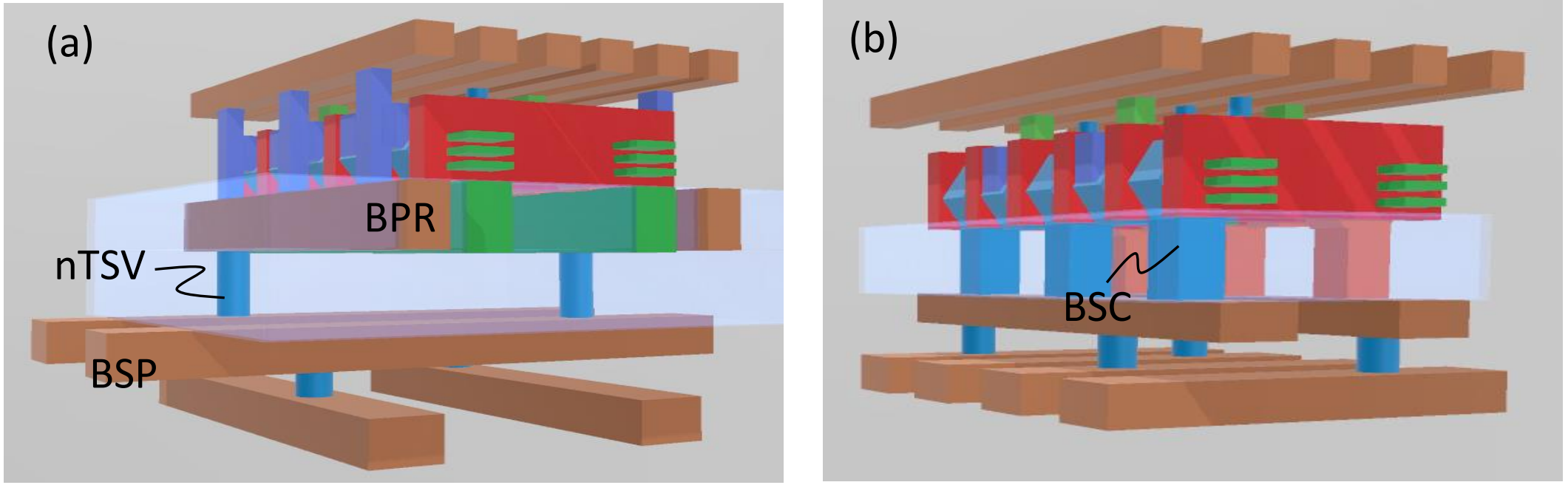


Source: IMEC



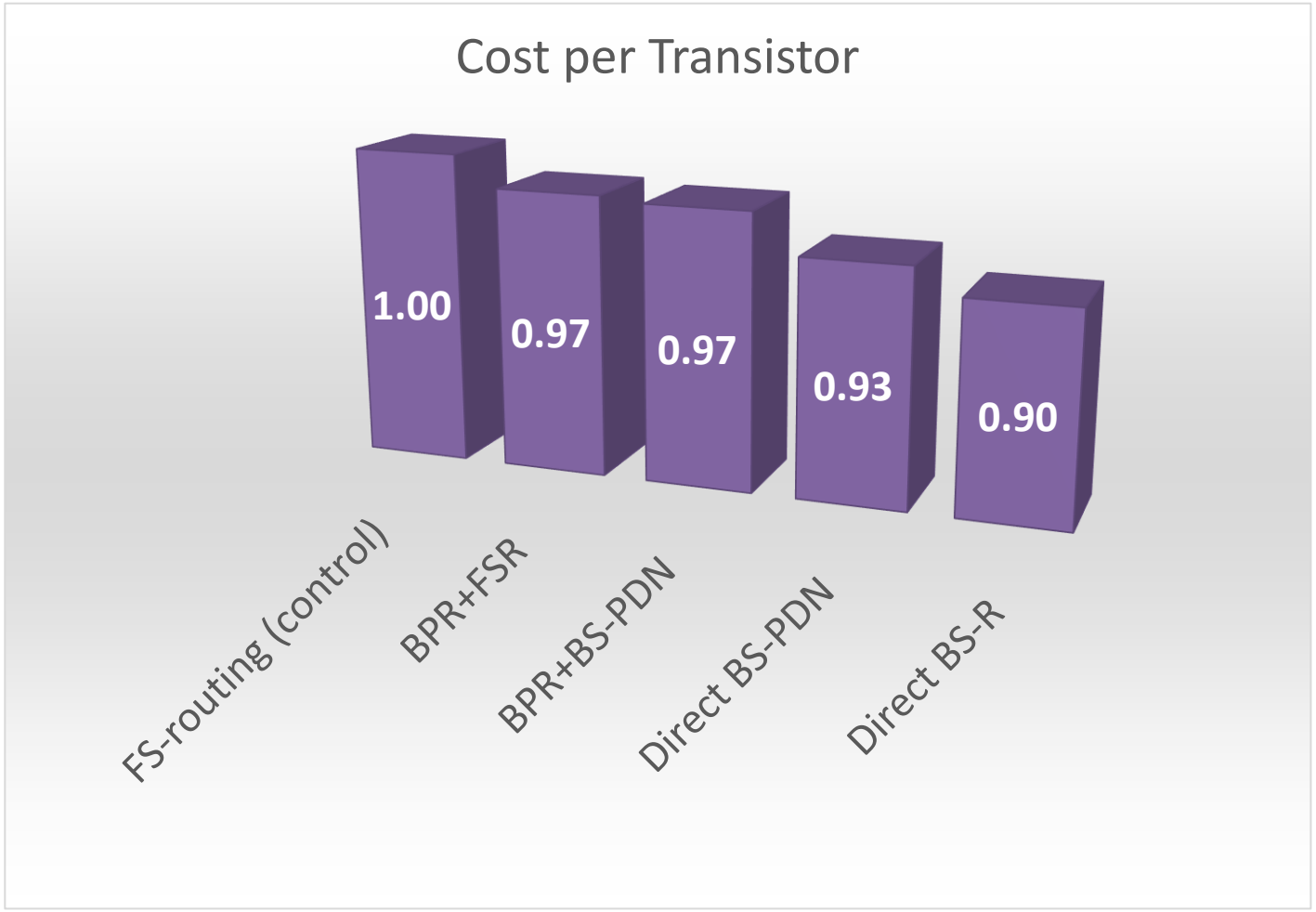
Source: Intel

Backside power connection schemes



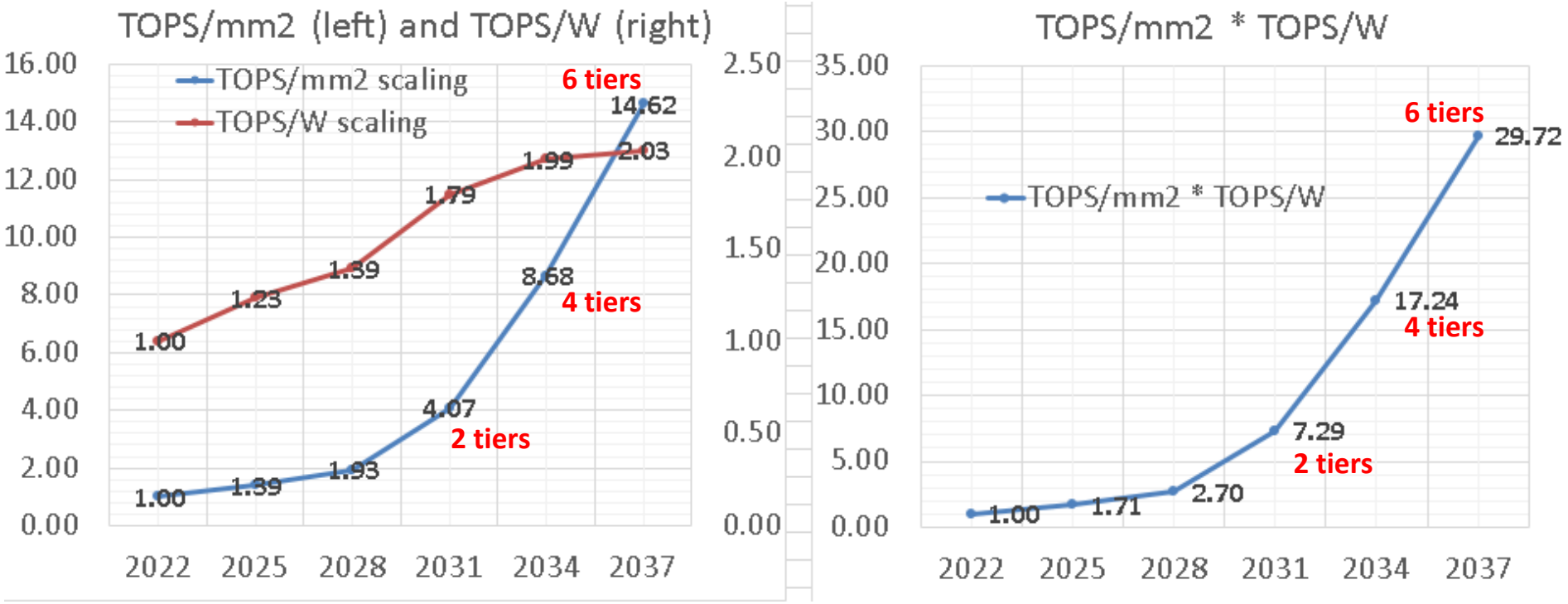
Two different Back Side Power (BSP) scheme.
(a) BPR+nTSV, (b) BSC (Back Side Contact)

BPR, BS-PDN, BS-Routing Cost



- ~10% cost per transistor scaling potential using backside routing.
- Significant area saving in logic area offset additional process cost.
  - Backside thinning process
  - Back side BEOL process
- Simpler Frontside process further offset backside process cost.
- Due to less IR drop in PDN, perf improves

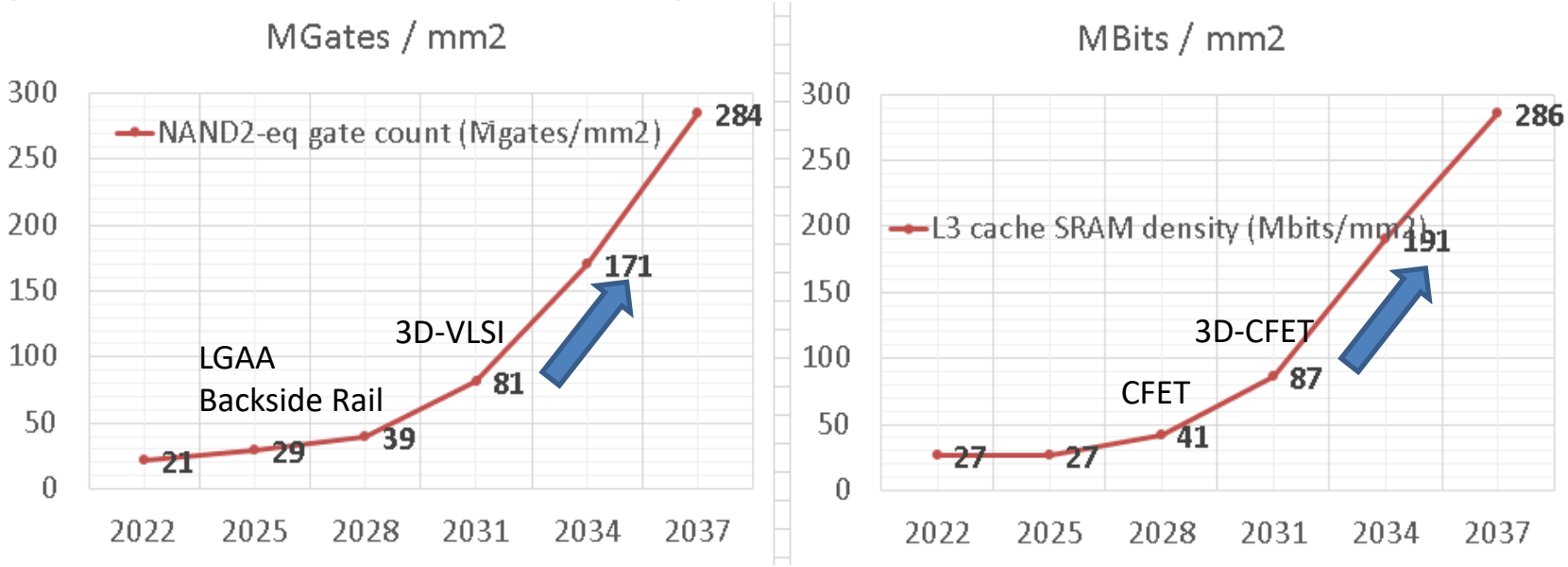
IRDS MM projection for TOPS/mm2 and TOPS/W scaling



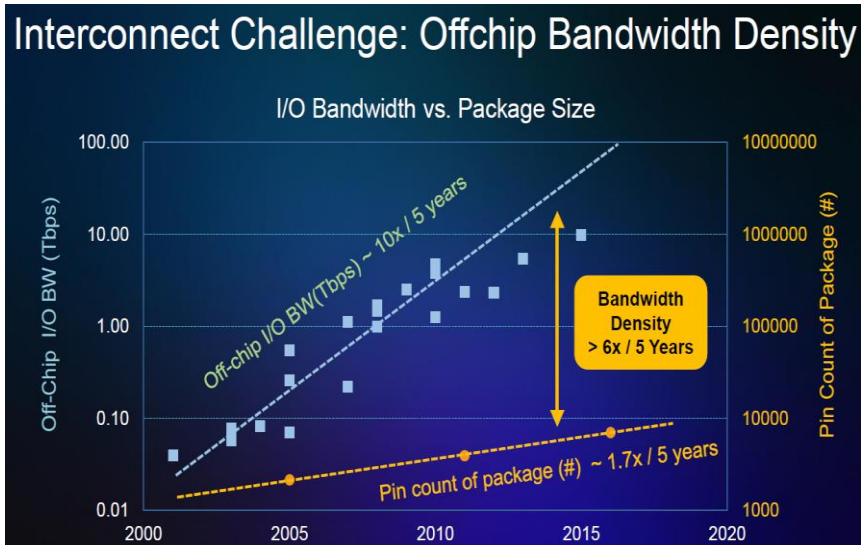
Source; IRDS 2022 More Moore roadmap

- Energy efficiency slowing down – techniques focusing on efficient power delivery more important
- TOPS/mm2 scaling boosted by 3DVLSI ensuring both compute and memory capacity/bandwidth

Integration capacity



Source: IRDS More Moore 2022 edition

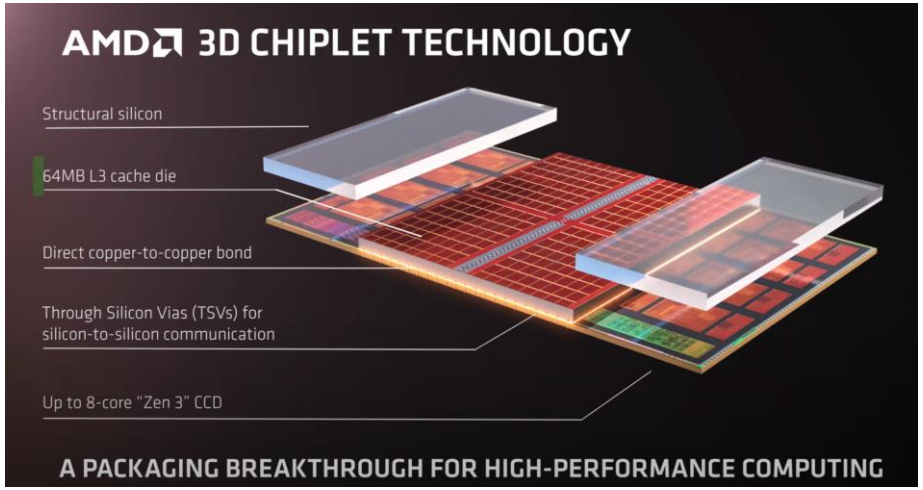


Source: ISSCC 2020

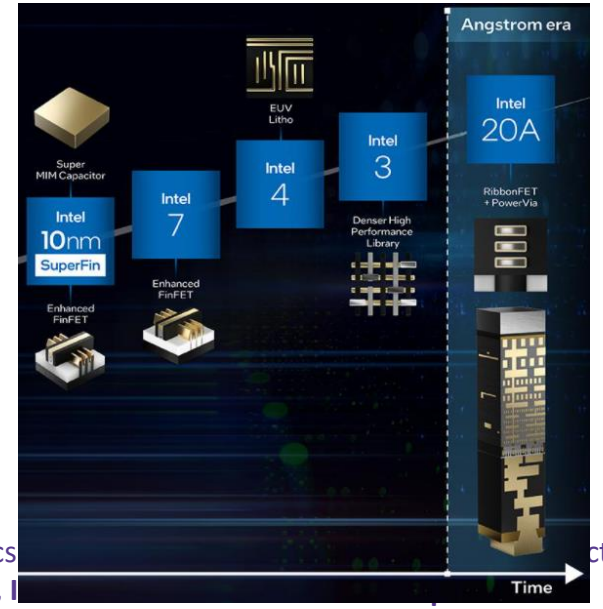
- Today: >100B transistors per chip, In 2034: >1T transistors per chip
- STCO comes into play to maintain the scaling
  - Larger memory bandwidth/capacity
  - **High-BW and low-energy IO** signaling for AI workload in memory
  - Better (e.g. backside) power delivery to improve voltage headroom
  - Die split (e.g. 3D/2.5D split of memory and logic) for better yield
  - Improving variability and better process centering to improve voltage headroom and frequency

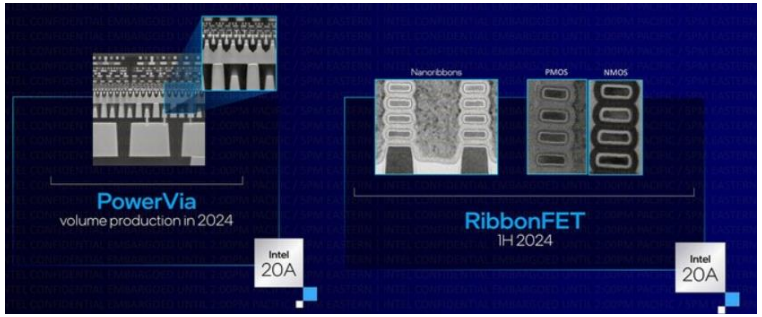
STCO elements continue system scaling

- Increasing cache size
  - Using mature technology for top and bottom tiers
  - Increasing L3 cache from 32MB to 96MB per chiplet with 3D hybrid bonding
  - L3 cache bandwidth increases to > 2 TB/sec.
  - Close to full-node performance gain - +12% FPS, +15% gaming perf with potential energy reduction with less cache misses
- Concurrent Transistor and Chip infrastructure development
  - Transition from finFET to Nanosheet
  - MIMCAP to improve PDN – better voltage headroom
  - PowerVia (backside power rail) to decouple the chip from PDN



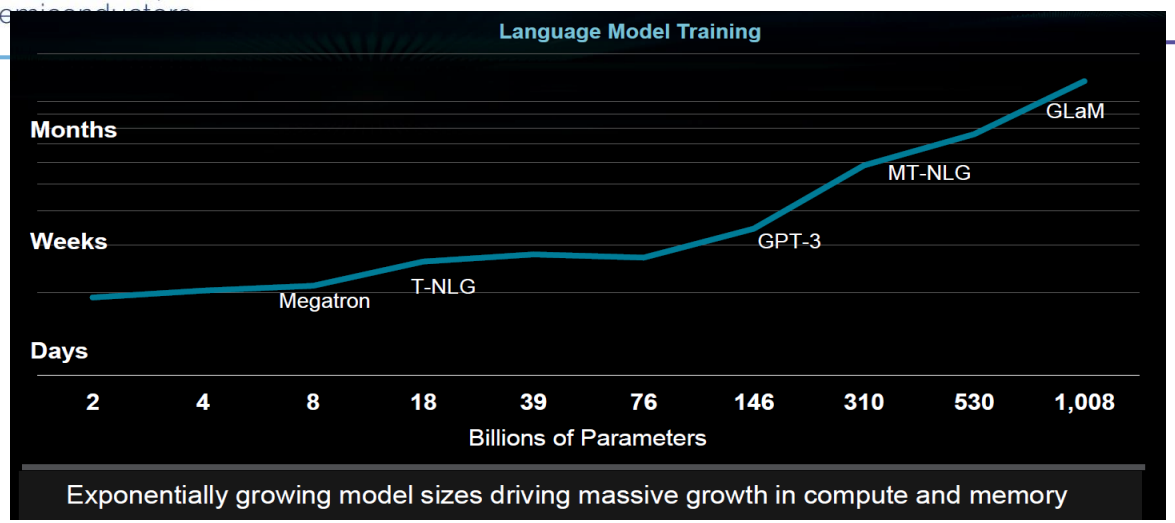
Source: AMD

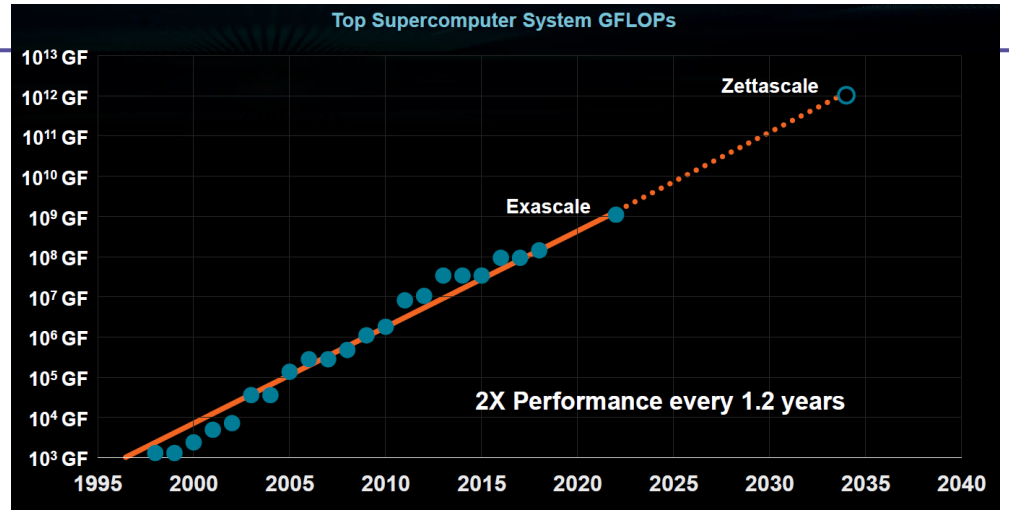


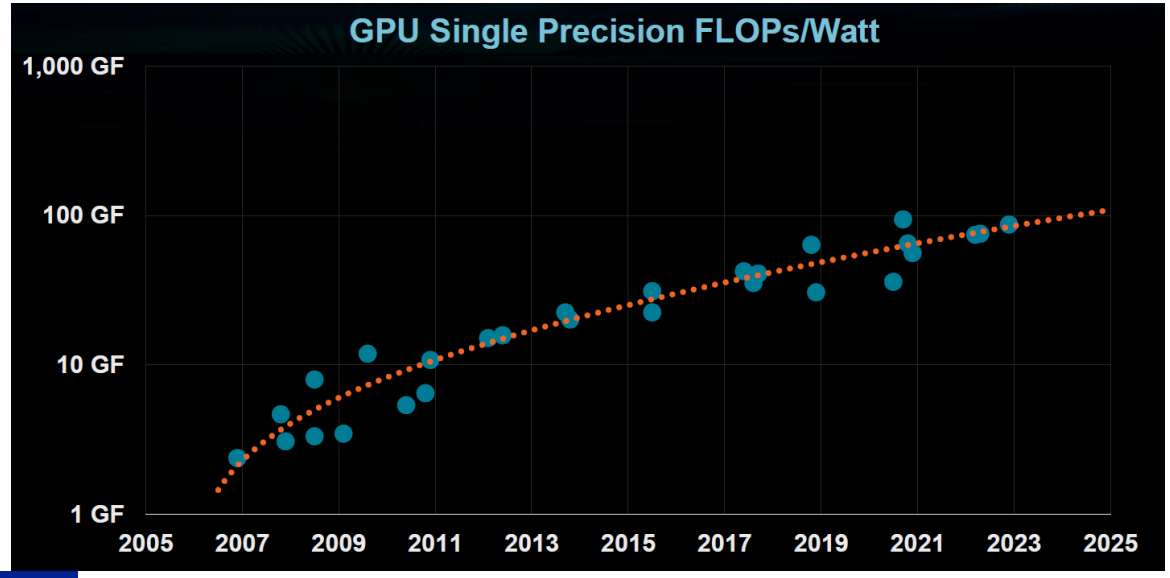


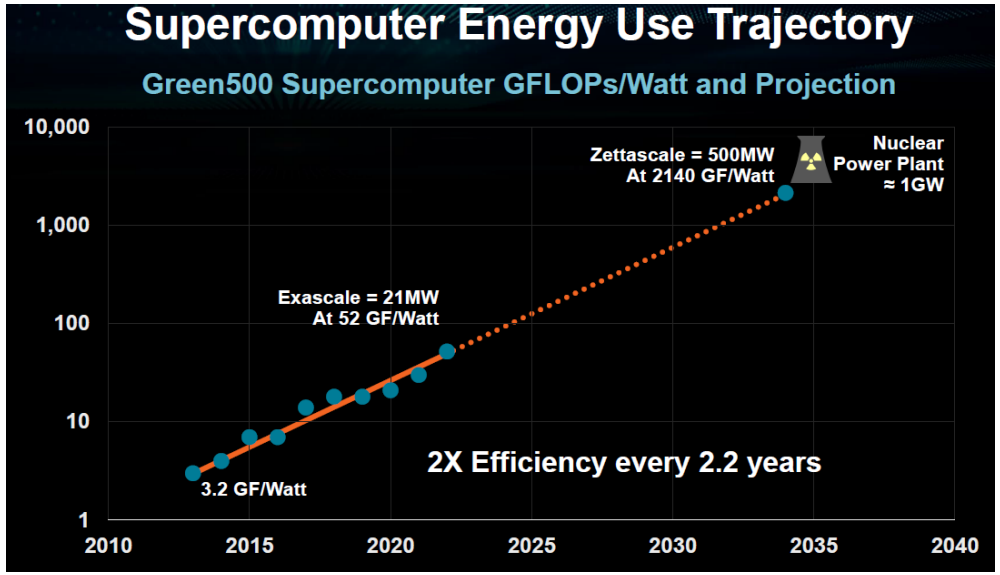
Source: Intel

AMD Plenary@ISSCC2023 – Energy efficiency bottleneck

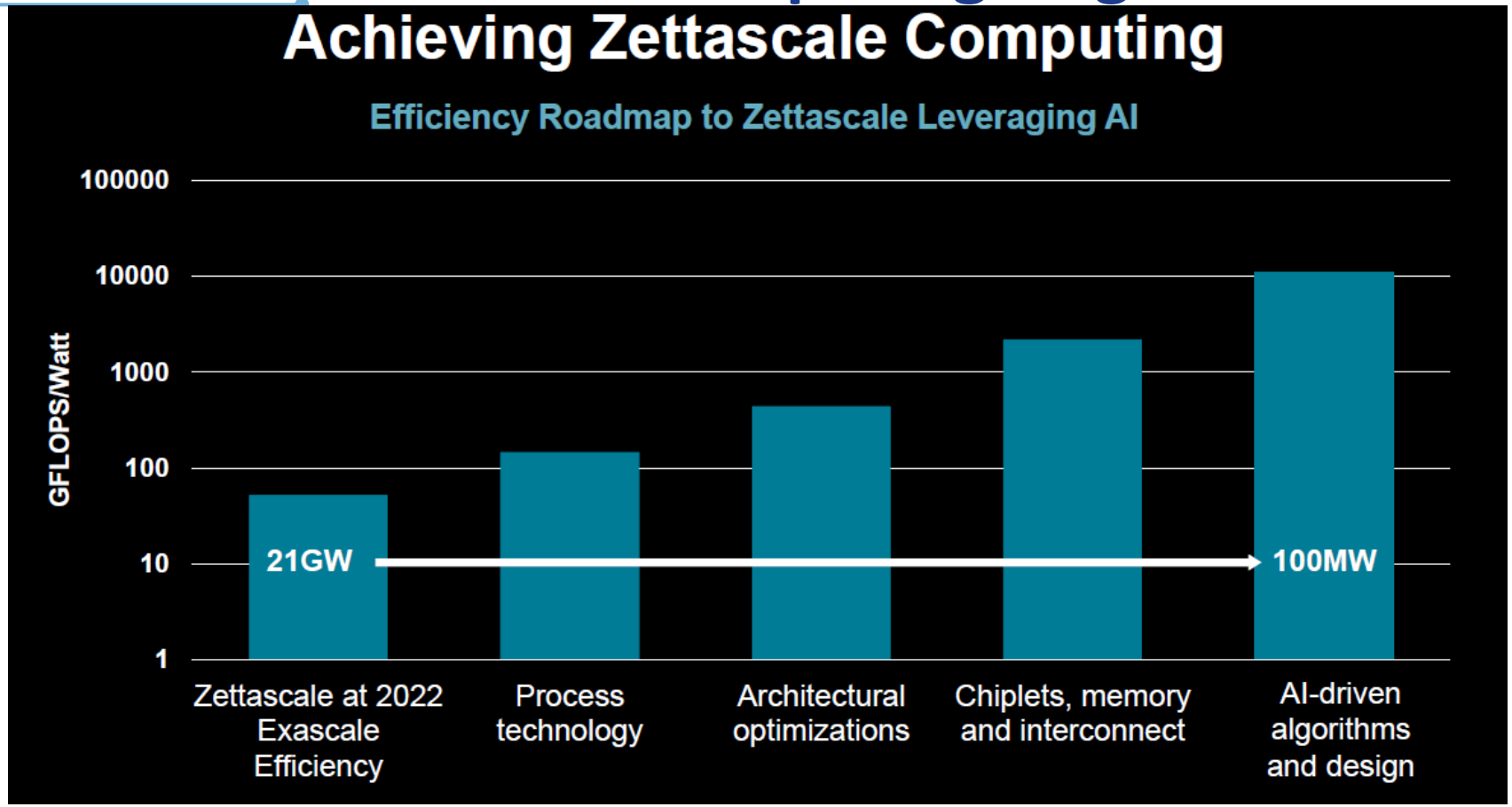








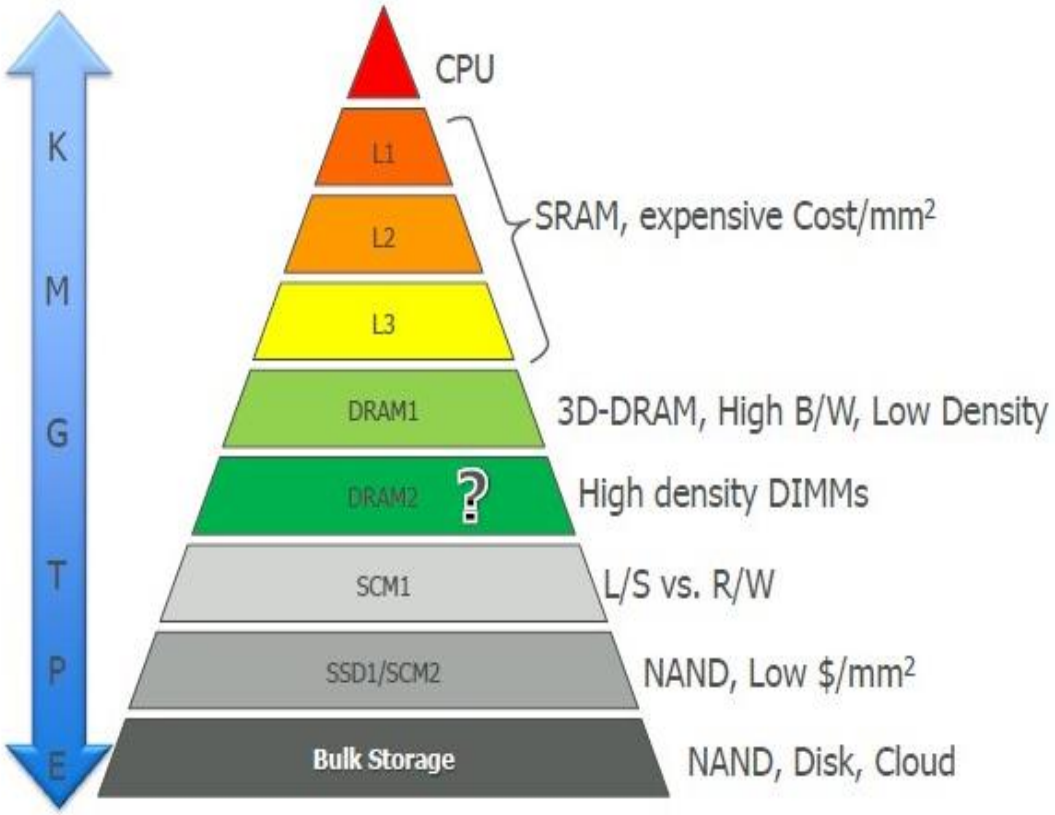
AMD Plenary@ISSCC2023 - How to enable zettascale computing target



- Target: >10,000 GF/Watt to enable <100MW Zettascale
- Process technology allowing pathway for architecture and algorithm optimizations

Memory options and fusion with compute

COMPUTE MEMORY SUBSYSTEM



Compute memory hierarchy

Today	Trending	Long-Term /Emerging
SRAM	SRAM	CFET-SRAM (monolithic/sequential-3D)
SRAM	3D-SRAM LLC/LLW-DRAM	3D-SRAM 3D-DRAM Emerging Memories (Oxide,FeFET,MRAM)
DRAM (LP,PC,GP) HBM-DRAM	DRAM (LP,PC,GP) HBM-DRAM	DRAM (LP,PC,GP) HBM-DRAM
PCM (3D-XPoint) NAND-FLASH	NAND-FLASH	NAND-FLASH Emerging Memories (RRAM,MRAM,FeFET)
SSD Magnetic	SSD	SSD DNA Storage

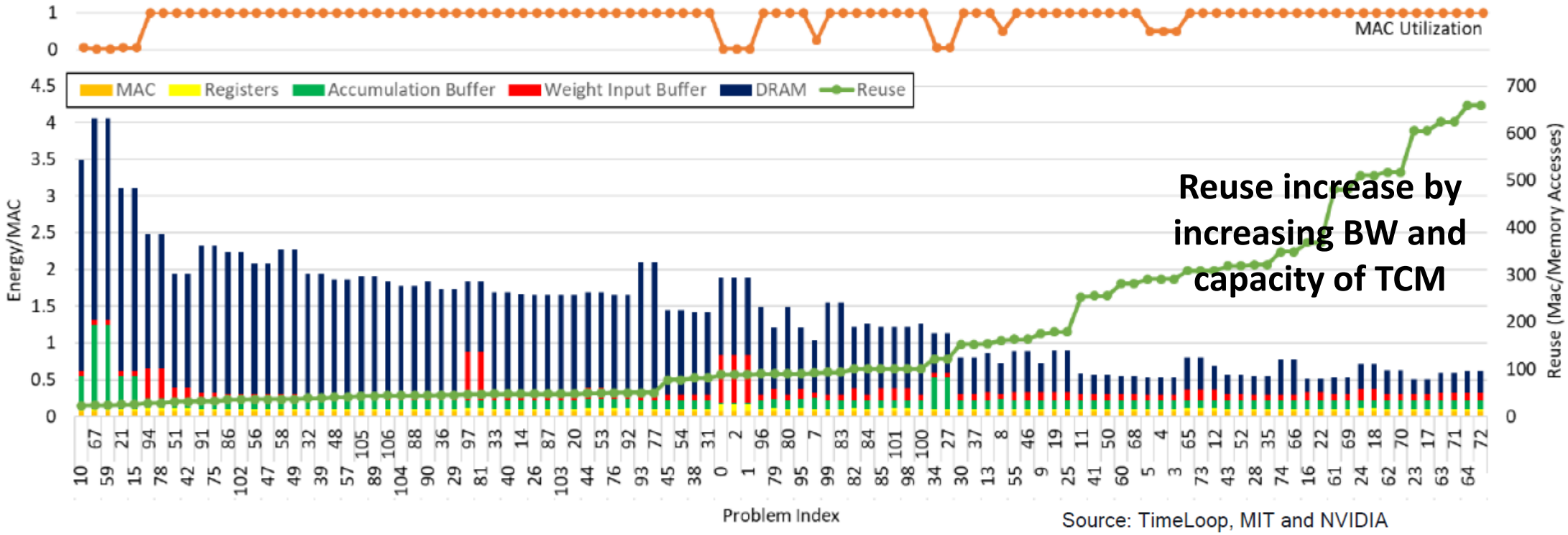
Other memory applications

AI buffer memory, Code memory for IoT, CIM, Frame buffer for GPU, LUT for FPGA

Source sample text

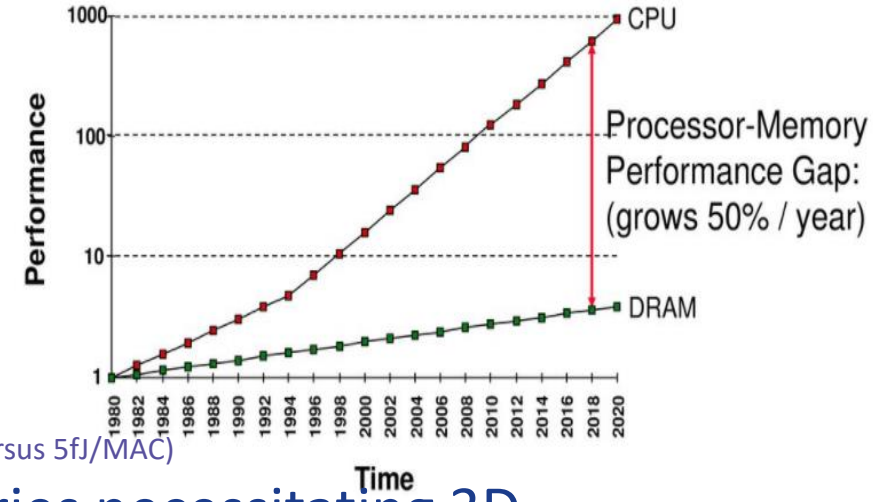
Significant energy reduction at system-level

- More than 70% of energy devoted to DRAM access – in LLM models this is >99%
- Energy reduction achieved by re-use level increase if higher buffer capacity and BW are utilized
- Capacity and BW increase enablers for system-energy reduction

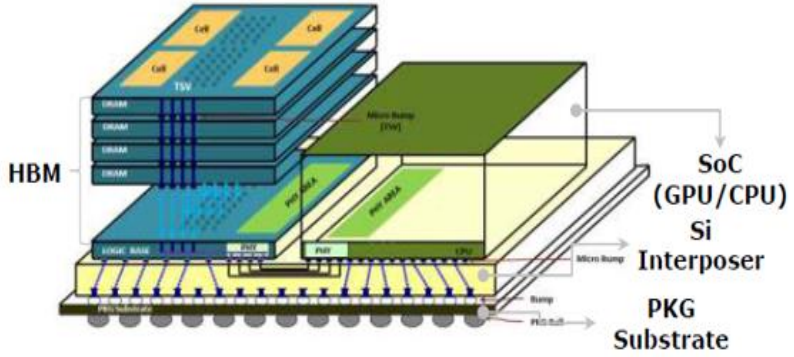


Memory wall problem

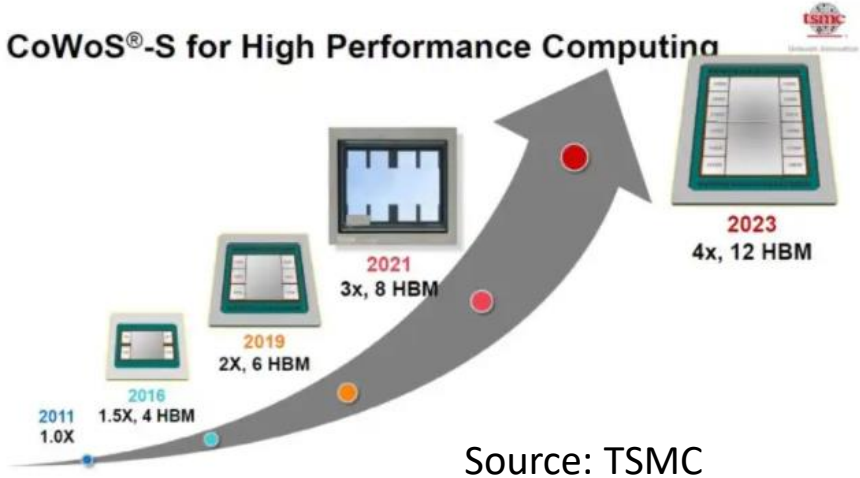
- Typical AI chip running at >10/500 TOPS (e.g. 5K/250K parallel MACs at 1GHz)
  - Challenge is how to feed data to CPU
- On-chip memory cache access
  - 2048 TBytes/sec for L0 (4K) – very wide BW, difficult data re-use at inner loop
  - 200 TBytes/sec for L1 (256KB) – difficult data re-use between kernels
  - 20 TBytes/sec for L2 (8MB) – difficult data re-use between kernels
- 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 (~3.5pJ/bit in LPDDR5X versus 5fJ/MAC)
- 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 supply data demand by HBM is not enough
  - Need for 2-5 TBytes/sec bandwidth access to high capacity data (>256MB)



Memory wall addressed by system-level interconnect: High-capacity Memory on SiP, I/O, and interconnect



Source: SK Hynix



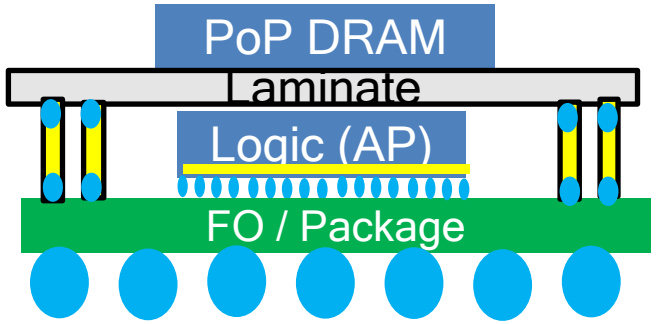
Source: TSMC

IO speed and bandwidth scaling much faster than CPU speed
Demand for large interposers

Source: Rambus

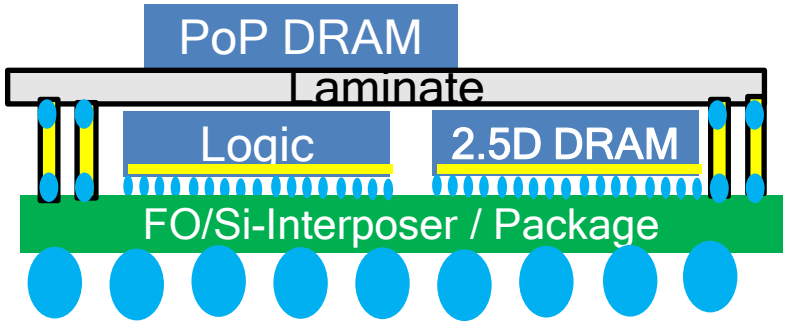
State-of-the-art – memory+compute integration schemes

Current integration for a mobile processor - FO-PoP



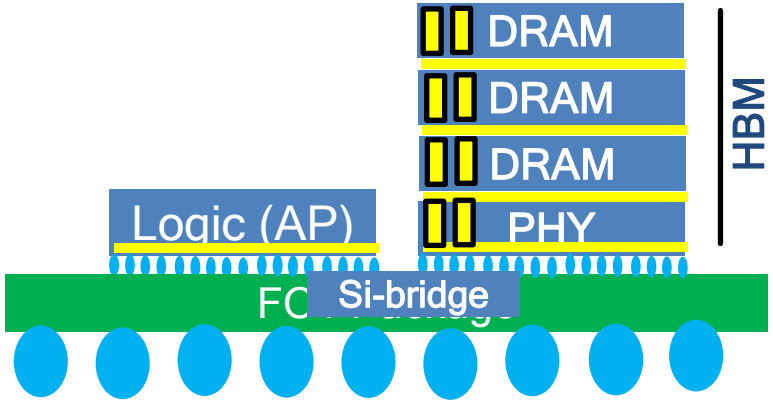
- Good xyz factor
- All computation takes place in logic (AP: application processor) die
- Limited cache (SRAM) in logic requiring more accesses to PoP DRAM
- Low-BW from PoP DRAM
- High energy/bit DRAM access

Trending integration for mobile/compute processors - FO-PoP w/ 2.5D DRAM



- DRAM (e.g. last-level cache) side-by-side (2.5D) with logic providing higher-BW access but at limited capacity (1GB)
- Moderate energy/bit
- xy form factor increase
- 2.5D routing complicating IO fanout
- Complex memory controller

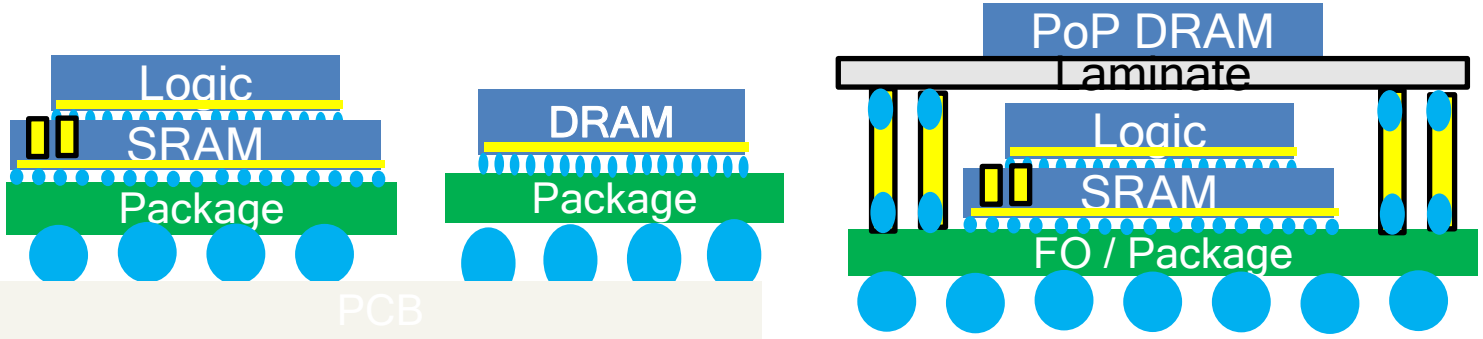
Current integration for high-performance GPU/AI/Server applications

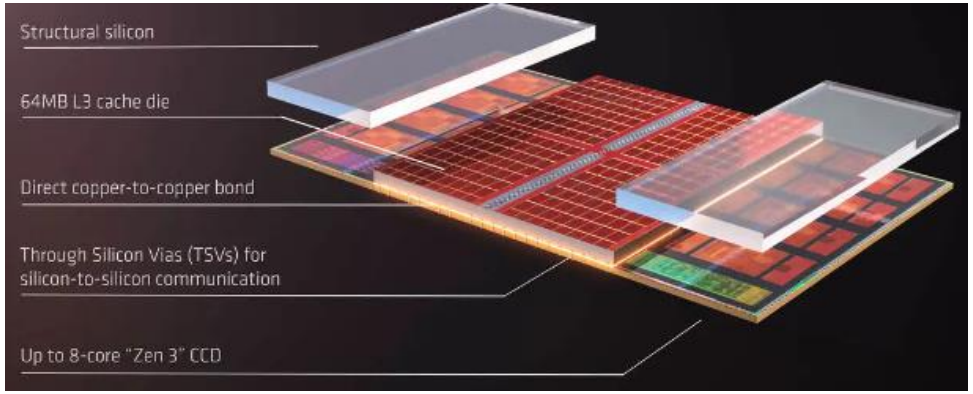


- 3D stacked DRAM (HBM) next to logic providing high-BW and high-capacity
- High energy/bit
- Xyz form factor penalty
- Need for TSV process
- 2.5D routing complicating IO fanout
- Complex memory controller
- Expensive interposer needing embedded Silicon bridge to route signals between HBM and logic

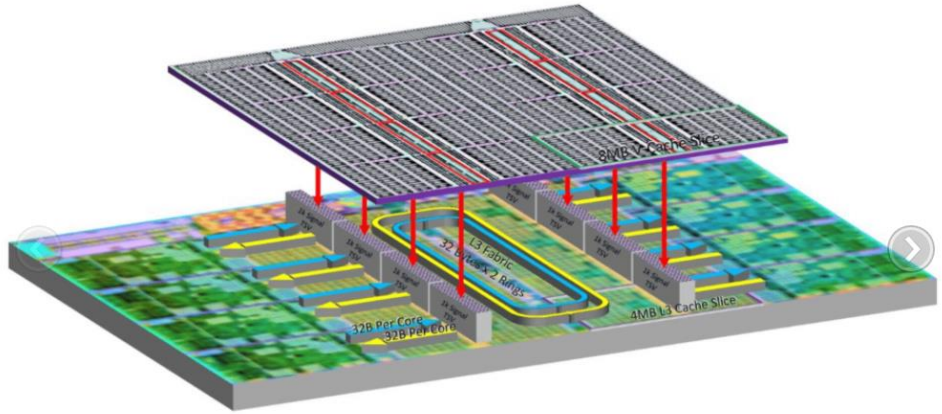
State-of-the-art – memory+compute integration schemes-2

Trending integration for a compute processor+3D SRAM – FO-PoP compatible





- All computation takes place in logic
- PoP DRAM compatible
- Increased SRAM capacity with 3D SRAM but not as high as DRAM - slightly reducing DRAM accesses in certain workloads but not effective in large AI models
- Need for TSV process



Example from AMD

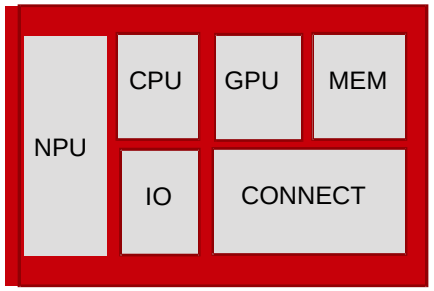
Source sample text

2D to fine-3D Transition – 3DVLSI era

- Memory limited
- Limited Performance
- Overhead in clocking
- Increasing parasitics
- Yield and cost issues
- Increasing dark Si

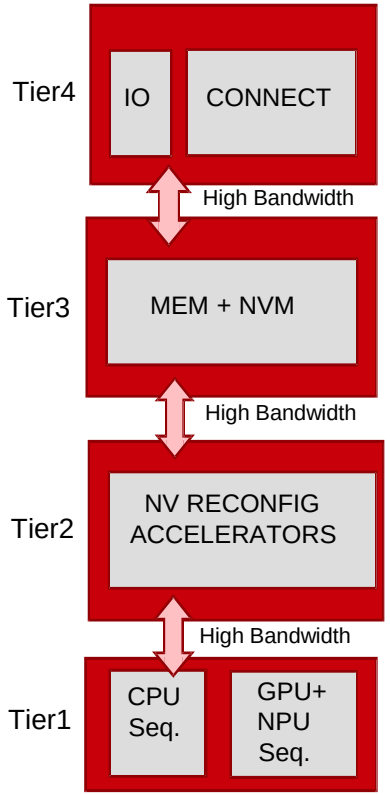
2D Era

Max Frequency at Low Power



3D Era

Max Throughput at Low Power Using Reconfig. Accelerators/Fabric



- Memory dominant
- High throughput

- Less overhead in clock
- Less parasitics
- Better yield, low-cost
- Less dark Si
- Parallelism and NVL/M
- Drivers: Big Data, Machine Learning, VR/AR

TECHNOLOGIES

SENSORS
- GRYSCOPE
- ACCELEROMETERS
- PIEZO TF TRANSDUCER
- OPTICAL SENSING
- MEMS MIC
- SOLID-STATE COMPASS
- WIRELESS SENSING
- ENV-HEALTH SENSING

DISPLAY
- OLED
- uLED
- uPROJECTOR
- LASER

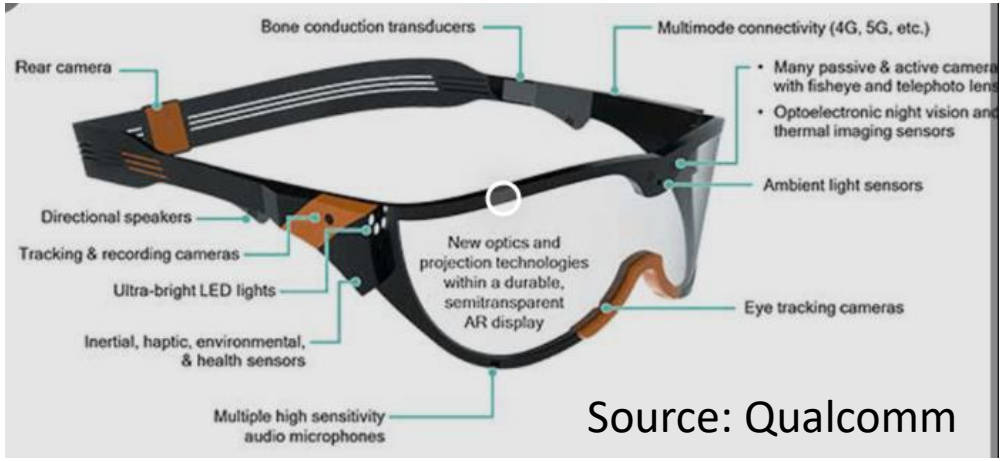
COMPUTE
- COMPUTER VISION
- AI
- NPU / CPU / NVM
- GPU

COMMUNICATIONS
- mmWave
- 5G / 6G
- UWB
- ANTENNA
- NETWORK

ENERGY
- BATTERY
- WIRELESS CHARGING
- SOLID-STATE STORAGE

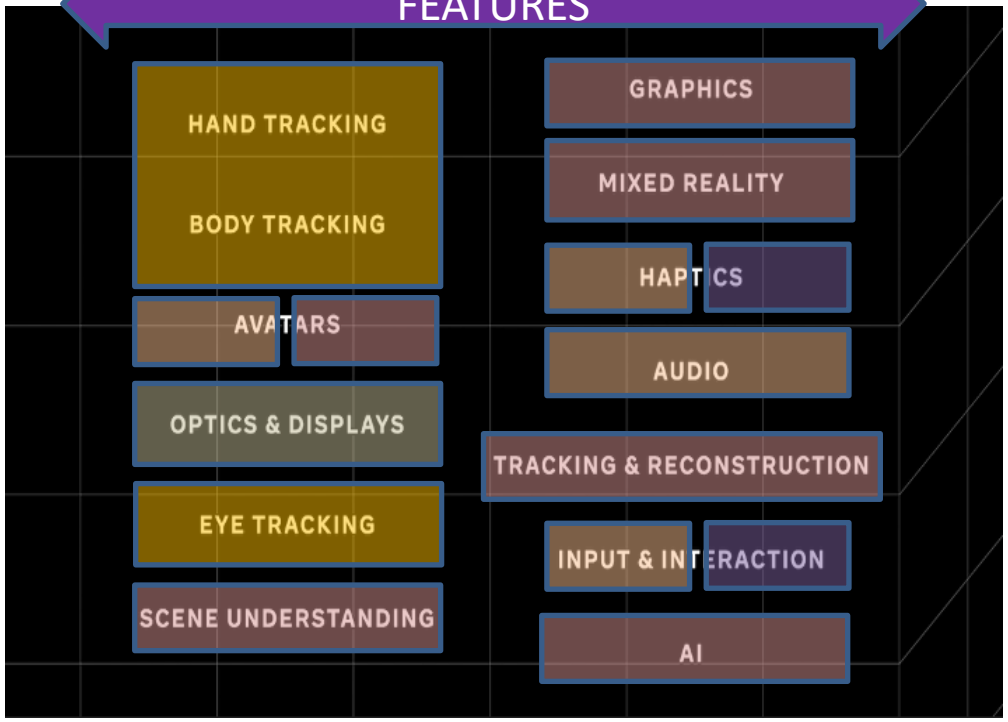
ACTUATORS
- ROBOTICS
- TRANSDUCERS
- MEMS OPTICS

ASSEMBLY
- 2.5D
- 3D
- THERMAL



Source: Qualcomm

FEATURES



ASSEMBLY

ENERGY

COMMUNICATIONS

More Moore difficult challenges

- Near Term (2022-2028)
  - Cost reduction and yield (at large die sizes)
  - Power scaling
  - Parasitics scaling
  - Integration enablement for SRAM-cache/AI applications
  - Local interconnect and power delivery
- Long Term (2028-2037)
  - Power scaling
  - 2D-Mat channel growth and device contact resistance hindering their manufacturability
  - Thermal issue due to increased power density
  - Cost reduction and lack of architectures supporting logic-on-logic stacking
  - Tight-pitch interconnects for low-RC and meeting EM/TDDB

Potential solutions

- Near Term (2022-2028)
  - Stacked GAA devices for PPA improvement and use DTCO to scale area
  - Reduce contact and rail resistance through new materials and backside processing
  - Low-k spacer to reduce Ceff and volumeless Vt tuning to reduce vertical pitch
  - High-NA EUV for single exposure patterning
  - 3D stacked SRAM and custom DRAM to scale capacity and bandwidth
- Long Term (2028-2037)
  - Reduce wirelength through 3D stacking by exploiting the vertical routing much more
  - Employ vertically stacked memory-logic subsystems with highly-parallel compute schemes
  - CFET for SRAM bitcell scaling
  - Fine-grain power gating and regulation
  - 3D architectures allowing minimum overhead and fault-tolerance in cross-tier interconnects

Take-aways

- Mobile computing and Cloud/Edge driving More Moore scaling
- IRDS More Moore roadmap connecting the HVM technology capability and integration schemes to the system FoMs
- GAA, buried rail with backside metal, and 3D stacked SRAM maintaining PPA scaling
  - GAA allows selection of widths with better drive than finFET
  - GAA vertical spacing reduction key for sustaining scaling with GAA
  - Buried rail with backside metal reducing cell height and resources for frontside routing
  - 3D SRAM and/or DRAM integration options help increasing memory capacity and bandwidth
- Memory bandwidth and capacity critical enablers for scaling
- Form factor reduction and heterogenous integration are enabled by 3D

THANK YOU

This project has received funding from the European Union's Horizon Europe research and innovation programme under GA Nº 101092562