CPU Trends

HEPiX Techwatch Working Group
July 10, 2024

Shigeki Misawa
Scientific Data and Computing Center
Brookhaven National Laboratory

Trends Affecting CPUs

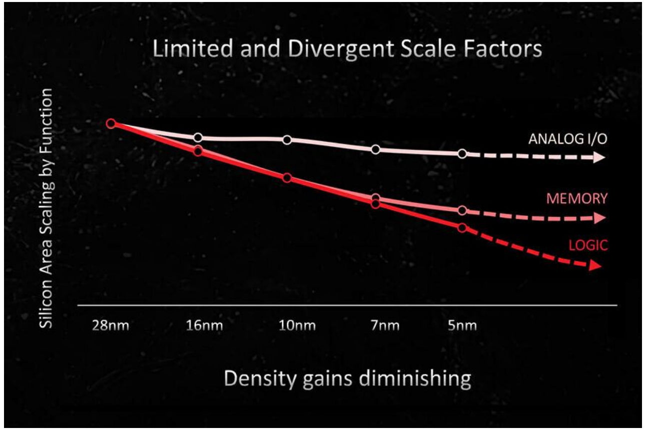
- End of Dennard scaling
- Unequal scaling
  - Static RAM cell vs logic sizes
  - I/O vs logic sizes
- Reduced lithography reticle size
- Advances in packaging
- Explosive growth in the use of AI/ML technologies
- Changing relationships among semiconductor foundries
- Increased competition in the CPU market
- Changing dynamics between CPU producers and consumers

End of Dennard Scaling

- Dennard scaling - Scale transistor size by 1/K then:
  a. Transistor area decreases by $1/K^2$
  b. Delay decreases by 1/K ➡ Max frequency increase by K
  c. Transistor power consumption decreases by $1/K^2$
  d. a and b combined ➡ Power consumption per unit area remains the same
- Scaling failure - Dennard scaling ignores leakage current
  a. Power density no longer constant as logic density increases
     - For a given die size, power consumption increases (roughly $\propto K^2$) <u>if only transistor dimensions are scaled</u>
     - CPU frequency has effectively stalled at ~4GHz

SRAM Scaling

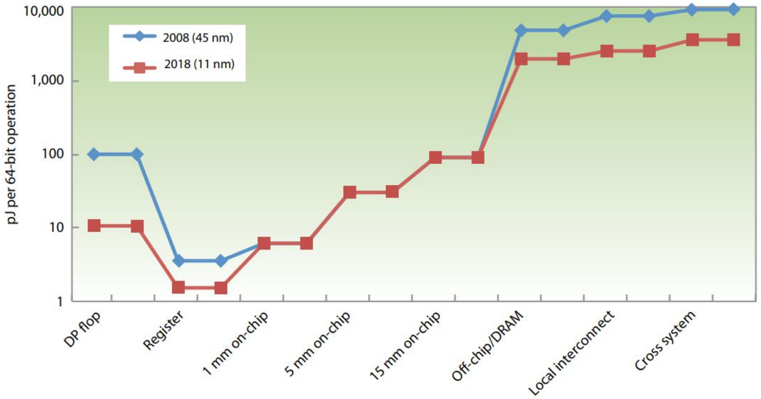
- For a fixed SRAM capacity and logic gate count, SRAM die area remains roughly constant as logic area shrinks in newer processes
  - Trade off between core count and on die SRAM cache capacity per core when moving to more advanced nodes.
  - SRAM costs higher with more advanced processes



TechPowerup.com, “AMD Explains the Economics Behind Chiplets for GPUs”, Nov 14, 2022, https://www.techpowerup.com/301071/amd-explains-the-economics-behind-chiplets-for-gpus

I/O Scaling

- Scaling of communication circuitry has not matched logic scaling
  - I/O energy consumption per bit roughly constant over time
  - I/O energy per bit increases with distance
  - I/O circuitry does not benefit from process feature size shrink
    - Size of I/O circuitry per lane remains roughly constant
- To minimize power and area while increase I/O bandwidth :
  - Increase # lanes and reduce baud rate per lane
  - Reduce signal propagation distance
  - Increase bits per symbol
    - e.g NRZ vs PAM3 (GDDR7) and PAM4 (PCIe-Gen 6+)



Exascale Computing Trends: Adjusting to the “New Normal” for Computer Architecture, IEEE Computing Society, Computing in Science & Engineering, Nov/Dec 2013.

Drive to Chiplets

- Process optimized fabrication
        - Older, lower cost processes for I/O and SRAM
        - Newer, higher cost process with smaller feature sizes for logic
- Reduced reticle area in newer processes - Limits size of monolithic die.
    - i193 (DUV) and EUV limit ~ 853mm^2
    - High-NA EUV limit ~ 450mm^2.
    - As reference, single Intel Emerald Rapids XCC die ~760mm^2 on Intel 10 process
- Die yield increases with smaller die size
    - Reduced on die device variation, increases yield at higher performance
    - Loss due to defects is reduced
- Allows for “Lego-like” CPU designs by interconnecting desired chiplets
    - Standardized chiplet interfaces like Universal Chiplet Interconnect Express (UCIe) may simplify this task in the future.

Explosive Growth in AI/ML Applications

- AI/ML accelerators embedded in CPU’s
  - On die accelerators (monolithic CPU/GPU)
  - Off die, in package chiplet GPU (AMD MI300)
- Driving need for higher memory bandwidth
  - HBM3/3e/4 memory - Higher performance than DDR5, but reduced capacity
  - LPDDR5 - Similar performance to DDR5, but lower power and overall capacity
- Driving need for tighter coupling between CPU’s and external AI/ML/GPU accelerators
  - Communication link with coherent memory access
    - CXL CPU-GPU link
    - NVLink Nvidia CPU-GPU link
  - MCM CPU/GPU Packaging - Nvidia Grace-Hopper

Advanced Packaging

- High performance die to die connectivity required for chiplet based CPUs
  - Best communication performance (BW, latency, power) still achieved with monolithic die
- Higher performance, lower power connectivity to external subsystems like memory and accelerators requires reducing their distance from the CPU
- Variety of 2D, 2.5D, and 3D interconnects are available or in development
  - Differ in tradeoff among cost, performance, complexity, area, thermal constraints, interconnect density and yield
  - Simplest are 2D interconnect like multichip modules (MCM) with ceramic or PCB substrates
  - 3D interconnects offer higher signal density and higher performance at a cost
  - Development of more advanced interconnects with interposers and bridges, through silicon vias (TSVs) and stacked die is an industry priority

Power/Performance/Area (PPA)

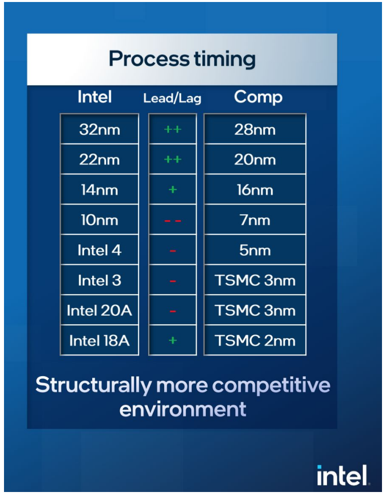
- CPU designs target specific trade offs in power consumption, performance and die area to satisfy specific classes of applications
- Example 1: Mobile CPUs (laptop/phone)
  - Emphasize low power consumption over power and area
  - Mix of performance (“Big”) and low power (“Little”) cores
  - Power management
  - On die application specific hardware (Crypto, GPU, AI) processors
    - More efficient than CPUs for the application
  - Close coupled memory (LPDDR5) for lower power
- Example 2: “Embedded” on die accelerators
  - Instruction set extensions (effectively embedded application specific processors)
    - AVX-512, VNNI, AES-NI, AMX

Data Center CPUs

- Target performance and power over area
- Product differentiation
  - Performance oriented
    - Emphasize single thread performance over power consumption
    - Use core complexity and clock speed, i.e. area and power, to increase performance
  - Throughput oriented
    - Emphasize performance/watt and performance/area
    - Use a larger number of smaller “efficient” cores that sacrifice some single threaded performance to reduce power consumption, but increase total throughput give area and power constraints
  - Specialized (Target specific use cases)
    - Large cache/fast local memory
    - Tightly coupled CPU/GPU and APUs

Changing Landscape in the Foundry Market

- Intel late with multiple process nodes over the past decade and is working to leapfrog TSMC
    - Intel late to transition to EUV lithography (Intel 4)
    - First with Gate All Around (GAA/Ribbon) FET and backside power (Intel 20A)
        - Improves transistor performance
        - Reduced power consumption (leakage current)
- TSMC first to move EUV lithography into production (N7 process).
    - Trails Intel to GAAFET
    - Backside power not on roadmap

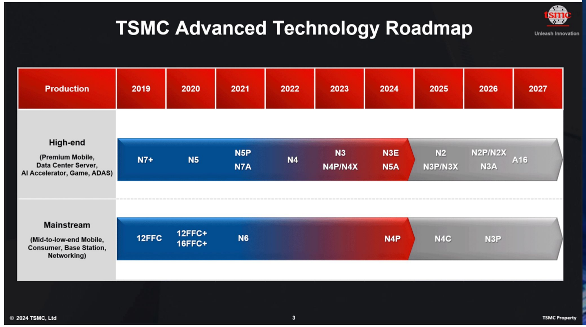


Process timing

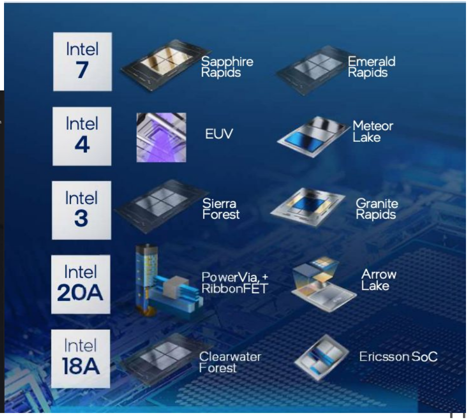
Intel	Lead/Lag	Comp
32nm	++	28nm
22nm	++	20nm
14nm	+	16nm
10nm	- -	7nm
Intel 4	-	5nm
Intel 3	-	TSMC 3nm
Intel 20A	-	TSMC 3nm
Intel 18A	+	TSMC 2nm

Structurally more competitive environment

https://d1io3yog0oux5.cloudfront.net/_fd22122fc9f6cbb0ceaf96bf8341310b/intel/db/861/9069/pdf/INTC+DB+Fireside+Chat+8-31-23.pdf
https://d1io3yog0oux5.cloudfront.net/_fd22122fc9f6cbb0ceaf96bf8341310b/intel/db/861/9068/pdf/IAO+Investor+Webinar+Slides+to+post+on+our+INTC+website+PDF.pdf



TSMC Advanced Technology Roadmap



https://www.anandtech.com/show/21370/tsmc-2nm-update-n2-in-2025-n2p-loses-bspdn-nanoflex-optimizations

Increasingly Competitive CPU Market

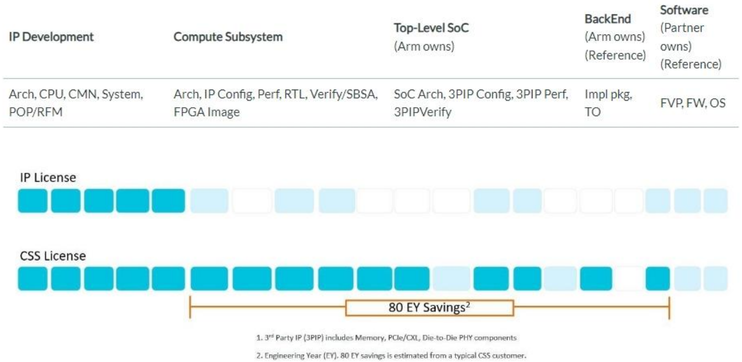
- Intel competitors making inroads
  - In 1Q24, AMD captured 23.6% of unit sales in the x86 server market [1]
  - ARM ISA based CPUs constituted 10% of the total market for server CPUs [2]
    - It is estimated that 20% of Amazon AWS CPU instances in 2022 were ARM based[2]
  - ARM ISA and AMD CPUs offer competitive performance relative to Intel CPUs
- Intel splits foundry operations from its CPU division
  - Intel utilized TSMC for some CPU's, providing Intel access to more advanced processes
- ARM Neoverse and Neoverse CSS reduces barrier to entry for competitors to AMD and Intel

[1] https://www.theregister.com/2024/05/10/amd_gains_on_intel/
[2] https://www.theregister.com/2023/08/08/amazon_arm_servers/

ARM and the Changing Producer/Consumer Relationship

- ARM is a supplier of CPU IP that entered the server market in 2018
    - ARM does not sell CPU chips
    - ARM sells designs for the major components of a complete CPU; cores, MMU, interconnect fabric, memory controller, etc.
    - Designs are either “soft” or “hard” IP i.e., logical designs (RTL models) or physical implementations from foundry partners (e.g. TSMC)
- Three generations of ARM Neoverse cores, E*n,* N*n*, and V*n*, where n=generation (1,2,3). Core types target different environments
    - E*n* - Low power (energy efficiency)
    - N*n* - “Balanced” power and performance
    - V*n* - High performance
- ARM IP significantly reduces the expertise and effort required to develop a complete CPU
    - Costs well within the budget of the large public cloud providers
    - Ability to create CPUs tailored to specific tasks presents a value proposition not provided by AMD or Intel

ARM IP Market Disruption



IP Development	Compute Subsystem	Top-Level SoC (Arm owns)	BackEnd (Arm owns) (Reference)	Software (Partner owns) (Reference)
Arch, CPU, CMN, System, POP/RFM	Arch, IP Config, Perf, RTL, Verify/SBSA, FPGA Image	SoC Arch, 3PIP Config, 3PIP Perf, 3PIPVerify	Impl pkg, TO	FVP, FW, OS

1. 3rd Party IP (3PIP) includes Memory, PCIe/CXL, Die-to-Die PHY components

2. Engineering Year (EY). 80 EY savings is estimated from a typical CSS customer.

Neoverse CSS

- Neoverse CSS IP
  - Preconfigured, mostly complete SoC with tunables (e.g., #cores, cache size)
  - Chiplet support via UCIe or other interface
  - External accelerator support via PCIe-5/CXL1.1
  - Significantly reduces effort and expertise required to design from components IP
- Open question for non-captive ARM developers:
  - What is the value proposition?
  - Is there enough demand to support a custom core or Neoverse derived ARM CPU in the open market? (Is there a sustainable business model?)

Intel Xeon 6

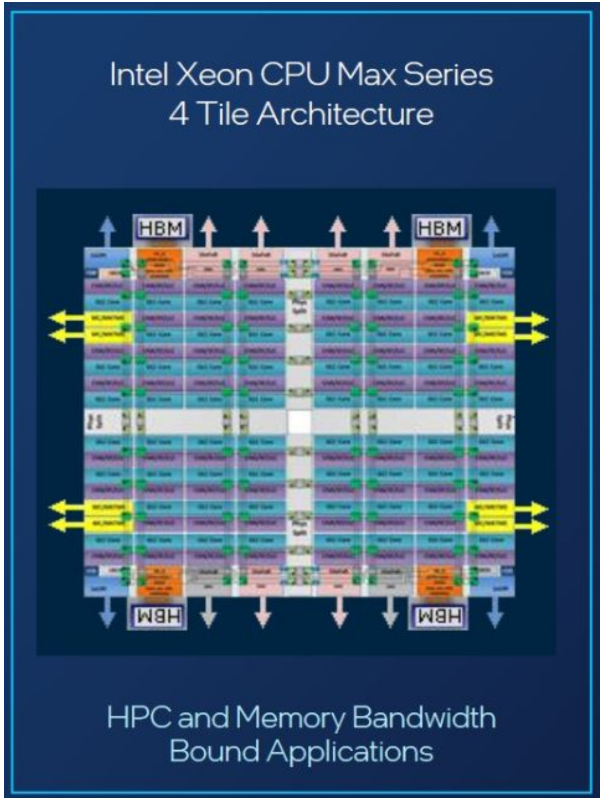
- Granite Rapids - SP and AP versions
  - Performance optimized
  - Up 86 (SP) or 128 (AP) performance optimized, hyperthreaded cores (P cores) per CPU
  - One or more compute chiplet with a mesh of P core “tiles”
  - I/O chiplet shared with Sierra Forest. Intel 7 process
- Sierra Forest - SP and AP versions
  - Throughput optimized
  - 144 (SP) or 288 (AP) single threaded efficiency cores (E cores) per CPU
  - Compute chiplet with mesh of E core tiles on Intel 3 (EUV enabled) process[2]
    - 2 or 4 cores per tile with shared L2 cache and L3 shared with all cores on chiplet
    - 12 tiles per chiplet, 3 chiplets for the 144 core Sierra Forest

---

[2] Intel Gets Its Chiplets In Order With 6th Gen Xeon SPs (nextplatform.com)

Intel Xeon Max

- Sapphire Rapids Max (Xeon 4th Gen)
    - Sapphire Rapids CPU augmented with 64GB of HBM2e high bandwidth memory
        - XCC die package with 4 EMIB bridge connected compute tiles
        - 4 HBM stacks, one per compute tile
        - Up to 56 Cores
    - Reconfigurable memory :
        - HBM only
        - HBM as cache to main DDR memory
        - HBM as part of main memory


Intel Xeon CPU Max Series
4 Tile Architecture

HPC and Memory Bandwidth
Bound Applications

Intel via ServeTheHome.com
[4th Gen Intel Xeon Scalable Sapphire Rapids Leaps Forward - Page 7 of 13 (servethehome.com)](https://www.servethehome.com)

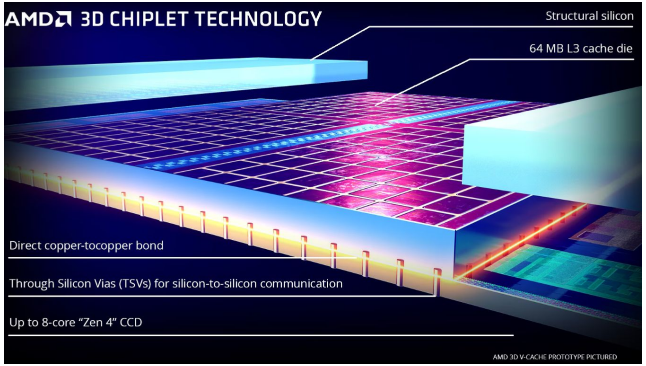
AMD Genoa and Bergamo

- Genoa - Up to 96 cores/192 threads (12 CCDs)
    - Zen 4 CCD - Chiplet with eight performance optimized, SMT enabled, Zen 4 cores with L1/L2/L3 cache. Built using TSMC N5 process. 66.3mm^2 die size
    - Separate I/O die fabricated with TSMC N6 process
- Bergamo - 128 cores/256 threads (8 CCDs)
    - Zen4c CCD - Chiplet with sixteen area optimized, SMT enabled, Zen 4 cores with L1, L2,L3 cache. TSMC N5 process. 72.7mm^2 die size
    - Lower base and boost clock and 33% more cores at the same TDP as comparable Genoa
    - Same L1/L2 cache size per core as Genoa, half size L3 cache per core compared to Genoa
    - Same I/O die as Genoa
- Genoa targets maximum single thread performance, while Bergamo targets maximum throughput per watt.

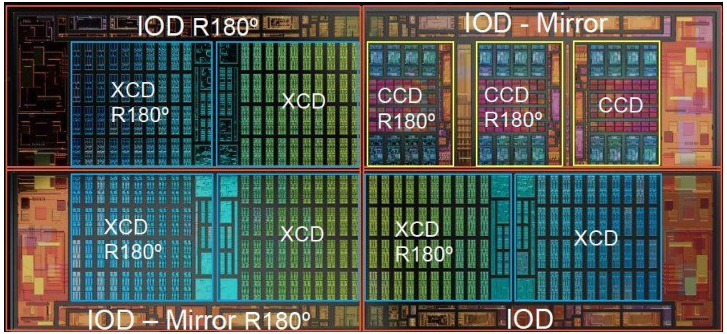
[1] source: https://www.semianalysis.com/p/zen-4c-amds-response-to-hyperscale

AMD Genoa X and MI300A

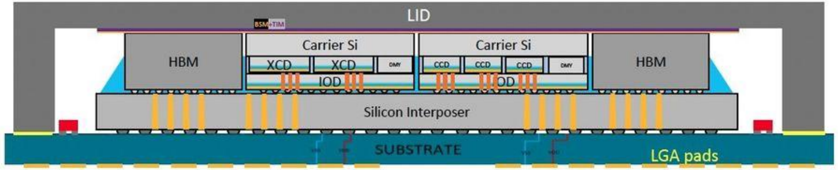
- Genoa X - Up to 96 cores/192 threads
  - Genoa CPU with CCDs stacked with L3 SRAM chiplet
  - Up to 1.1GB of L3 cache, 3x the L3 cache of similar core count Genoa CPUs
- MI300A APU
  - 3 Zen 4 CCDs - Total of 24 Zen 4 cores
  - 128 GB of unified memory
    - Eight 16GB HBM3 memory stacks
  - 6 GPU “XCD” chiplets - Total of 228 CDNA 3 GPU com units



https://www.amd.com/en/products/processors/technologies/3d-v-cache.html



AMD via IEEE Spectrum https://spectrum.ieee.org/amd-mi300



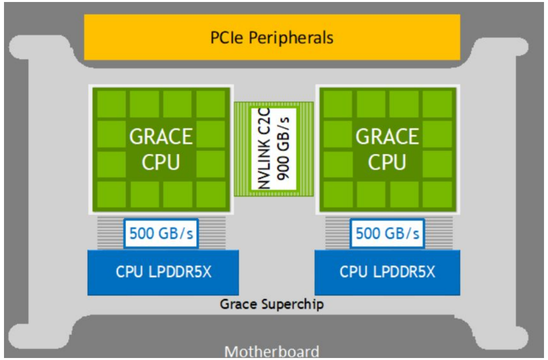
ARM Based Data Center CPUs

- Neoverse V2 derived CPUs
  - Amazon Graviton4
  - Nvidia Grace
  - Google Axion
- Neoverse CSS N2 derived CPUs
  - Microsoft Cobalt 100
- Neoverse N1 derived CPUs
  - Ampere Altra / Altra Max
- Custom (non Neoverse) derived CPUs
  - Ampere AmpereOne

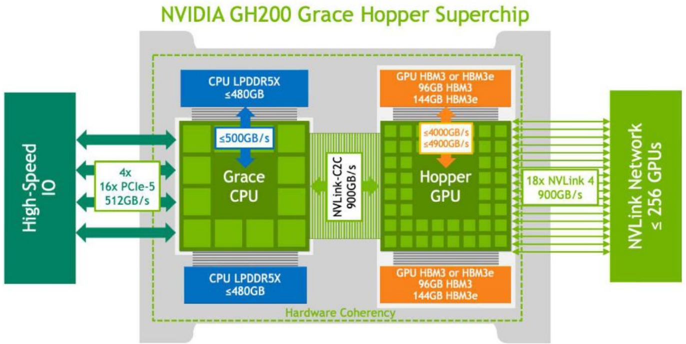
- Note that all recent data center ARM ISA CPUs and ARM Neoverse cores do not support simultaneous multi-threading (SMT)
- All CPUs typically compared to Intel Sierra Forest and AMD Bergamo.

Nvidia Grace and Grace Hopper

- PCB based MCM
  - NVLink-C2C chip interconnect
  - Grace Superchip
    - Two Grace 72 core CPUs
      - Up to 480 GB per CPU
      - 32 channel LPDDR5X memory
  - Grace Hopper Super Chip
    - One Grace CPU
    - One Hopper GPU
      - 144 GB of HBM3e memory
      - 6 stacks HBM3/HBM3e memory



https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-cpu-superchip#page=1?ncid=no-ncid



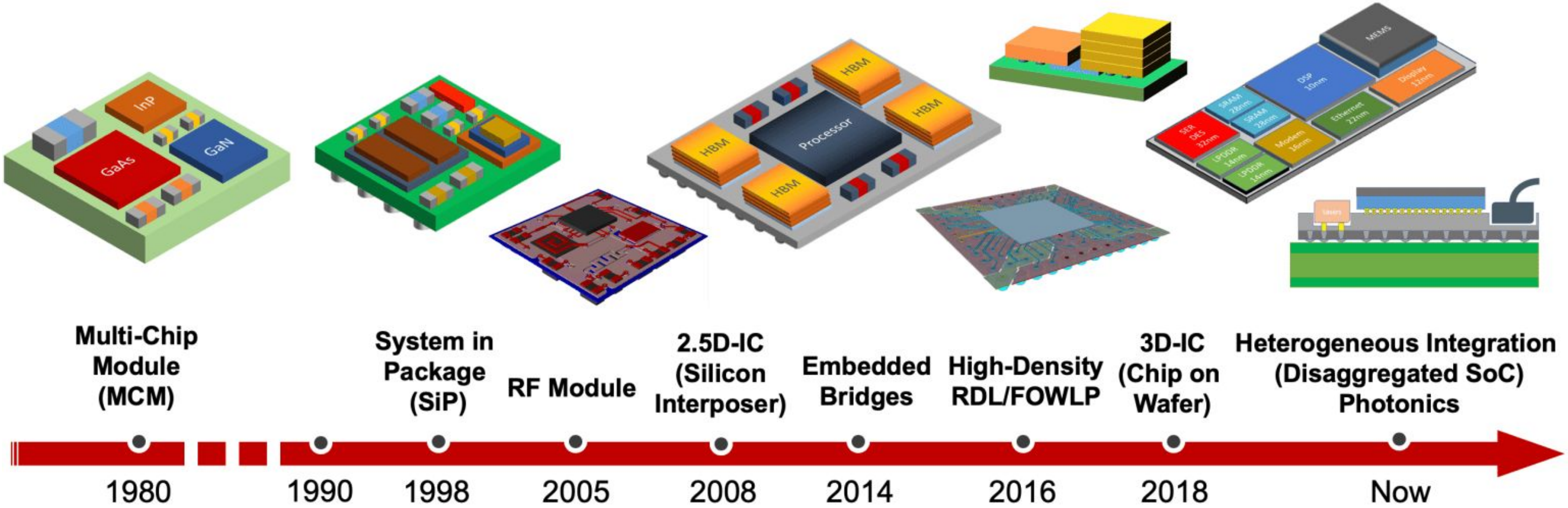
https://resources.nvidia.com/en-us-grace-cpu/grace-hopper-superchip?ncid=no-ncid

Summary

- Current generation data center CPUs exhibit more explicit differentiation compared to previous generations
  - At the early stages of utilizing chiplets
- Realistic benchmarks are needed determine if HEP applications are sensitive to these differences.
  - Quantitative understanding of the workload mix required if performance differences between different CPUs varies substantially by application
- CPU and server power consumption is becoming an issue
  - Limits of air cooling. Is direct chip cooling in the future ?

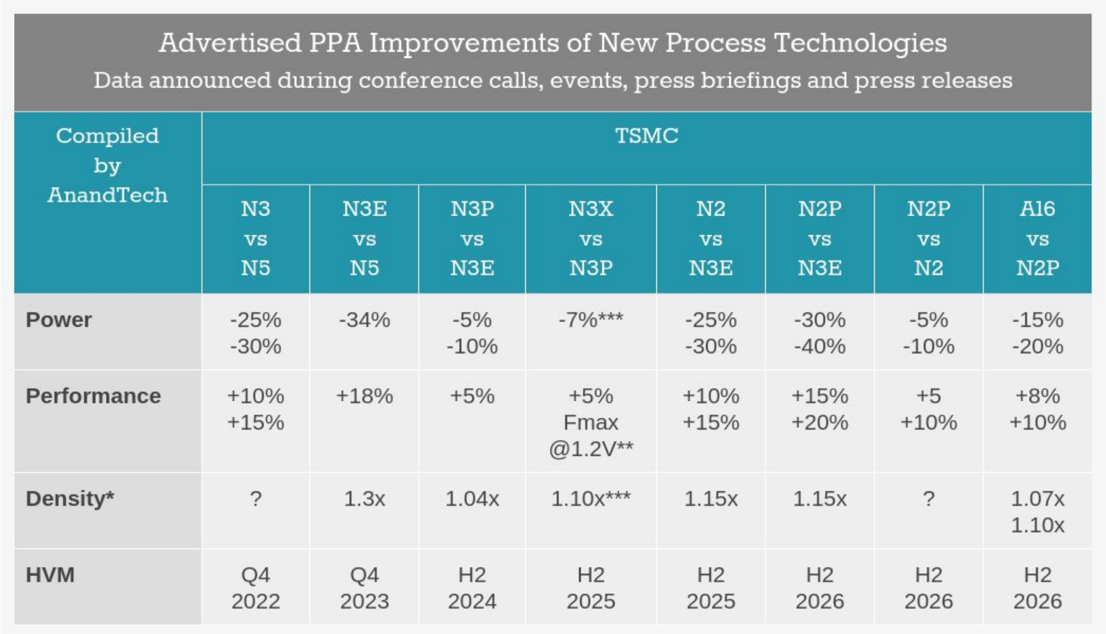
Backup Slides

2D 2.5D 3D Evolution



Cadence via SemiEngineering.com
https://semiengineering.com/eda-on-board-with-advanced-packaging/

TSMC Process Node Stats



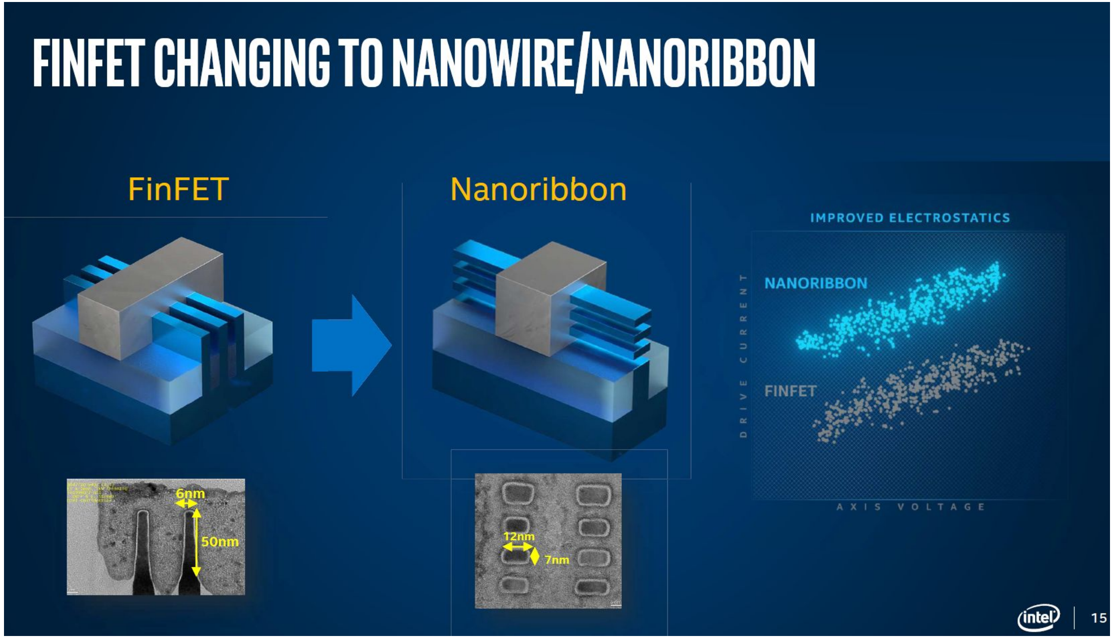
Advertised PPA Improvements of New Process Technologies

Data announced during conference calls, events, press briefings and press releases

Compiled by AnandTech	TSMC							
	N3 vs N5	N3E vs N5	N3P vs N3E	N3X vs N3P	N2 vs N3E	N2P vs N3E	N2P vs N2	A16 vs N2P
Power	-25% -30%	-34%	-5% -10%	-7%***	-25% -30%	-30% -40%	-5% -10%	-15% -20%
Performance	+10% +15%	+18%	+5%	+5% Fmax @1.2V**	+10% +15%	+15% +20%	+5 +10%	+8% +10%
Density*	?	1.3x	1.04x	1.10x***	1.15x	1.15x	?	1.07x 1.10x
HVM	Q4 2022	Q4 2023	H2 2024	H2 2025	H2 2025	H2 2026	H2 2026	H2 2026

https://www.anandtech.com/show/21408/tsmc-roadmap-at-a-glance-n3x-n2p-a16-2025-2026

GAA FET (aka Ribbon/Nanowire FET)



FINFET CHANGING TO NANOWIRE/NANORIBBON

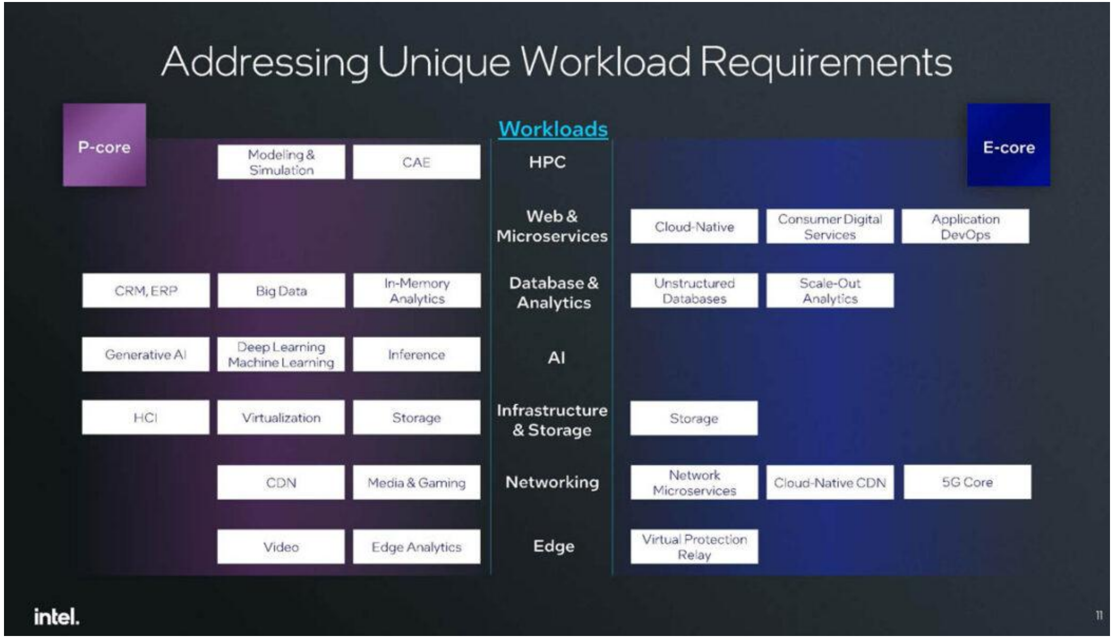
FinFET

Nanoribbon

IMPROVED ELECTROSTATICS

Intel via anandtech.com
https://www.anandtech.com/show/16823/intel-accelerated-offensive-process-roadmap-updates-to-10nm-7nm-4nm-3nm-20a-18a-packaging-foundry-emib-foveros/3

Workload Differentiation



Addressing Unique Workload Requirements

P-core			Workloads	E-core		
	Modeling & Simulation	CAE	HPC			
			Web & Microservices	Cloud-Native	Consumer Digital Services	Application DevOps
CRM, ERP	Big Data	In-Memory Analytics	Database & Analytics	Unstructured Databases	Scale-Out Analytics	
Generative AI	Deep Learning Machine Learning	Inference	AI			
HCI	Virtualization	Storage	Infrastructure & Storage	Storage		
	CDN	Media & Gaming	Networking	Network Microservices	Cloud-Native CDN	5G Core
	Video	Edge Analytics	Edge	Virtual Protection Relay		

Intel via ServetheHome.com
https://www.servethehome.com/intel-xeon-6-6700e-sierra-forest-shatters-xeon-expectations/