Sandia National Laboratories

Exceptional service in the national interest

The Unfortunate Economics of Highly Specialized Chiplets

Ben Feinberg

SOS 27 Workshop

March 18, 2025

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administrationunder contract DE-NA0003525.

U.S. DEPARTMENT OF ENERGY

SAND2025-04747C

NNSA

Why Are We Talking About Chiplets?

- **The end of Dennard Scaling**
  - We cannot rely on pure technology improvements to provide major year-over-year performance improvements
- **The scale of hyperscalers and AI compute needs**
  - Scientific workload needs are less economically important to major vendors

Microsoft is deploying the equivalent of five 561 petaflops supercomputers every month

Amazon 2025 capex to reach $100bn, AWS 2024 revenue hit $100bn

Google expects 2025 capex to surge to $75bn on AI data center buildout

Meta plans $60-65bn capex on AI data center boom, will bring ~1GW of compute online this year

And re-announces its 2GW+ data center in Louisiana

OpenAI and Oracle to deploy 64,000 GB200 GPUs at Stargate Abilene data center by 2026 – report

Chips will be installed in phases, with 16,000 expected to be in place by end of summer 2024

Why Are We Talking About Chiplets?

- **The end of Dennard Scaling**
  - We cannot rely on pure technology improvements to provide major year-over-year performance improvements
- **The scale of hyperscalers and AI compute needs**
  - Scientific workload needs are less economically important to major vendors

How can we continue scaling performance?

Microsoft is deploying the equivalent of five 561 petaflops supercomputers every month

Amazon 2025 capex to reach $100bn, AWS 2024 revenue hit $100bn

Google expects 2025 capex to surge to $75bn on AI data center buildout

Meta plans $60-65bn capex on AI data center boom, will bring ~1GW of compute online this year

And re-announces its 2GW+ data center in Louisiana

OpenAI and Oracle to deploy 64,000 GB200 GPUs at Stargate Abilene data center by 2026 – report

Chips will be installed in phases, with 16,000 expected to be in place by end of summer 2024

Specialization to the Rescue?

SPARK: Sparsity Aware, Low Area, Energy-Efficient, Near-memory Architecture for Accelerating Linear Programming Problems

I-DGNN: A Graph Dissimilarity-based Framework for Designing Scalable and Efficient DGNN Accelerators

Uni-Render: A Unified Accelerator for Real-Time Rendering Across Diverse Neural Renderers

NOVA: A Novel Vertex Management Architecture for Scalable Graph Processing

EXION: Exploiting Inter- and Intra-Iteration Output Sparsity for Diffusion Models

Piccolo: Large-Scale Graph Processing with Fine-Grained In-Memory Scatter-Gather

Lincoln: Real-Time 50∼100B LLM Inference on Consumer Devices with LPDDR-Interfaced, Compute-Enabled Flash Memory

Cambricon-DG: An Accelerator for Redundant-Free Dynamic Graph Neural Networks Based on Nonlinear Isolation

FIGLUT: An Energy-Efficient Accelerator Design for FP-INT GEMM Using Look-Up Tables

EFFACT: A Highly Efficient Full-Stack FHE Acceleration Platform

Sampling of accelerator papers from HPCA 2025

Specialization to the Rescue?

15x Performance Over CPU
152x Performance Over GPU

2.4-3.5x Performance Over Precious Accelerators

0.7-119X Performance Over Commercial Renderers

2.35x Performance Over Previous Accelerator

3.2-379x Performance Over GPU

1.6-2.8 Performance Over Previous Accelerators

11-13x Prefill Performance Over SSD
158-254x Token Performance over SSD

7-49x Performance Over Previous Accelerators

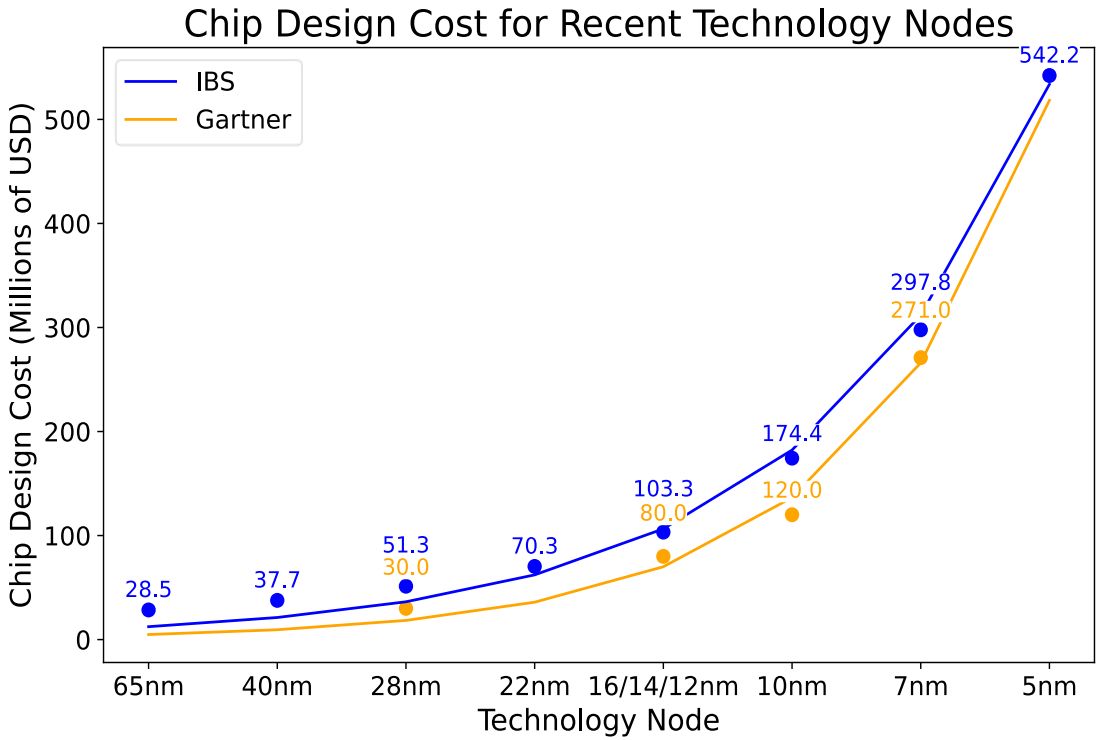
2x Performance over Previous Accelerators

0.57-6.16x Performance Over Previous Accelerators

Sampling of accelerator papers from HPCA 2025

Scale and Non-Recurring Engineering

- Chip design and fabrication have substantial non-recurring costs (NRE) which must be amortized with scale
- NRE costs are rising rapidly with new nodes due to increased design complexity
- Developing robust software stacks for new hardware adds substantial NRE

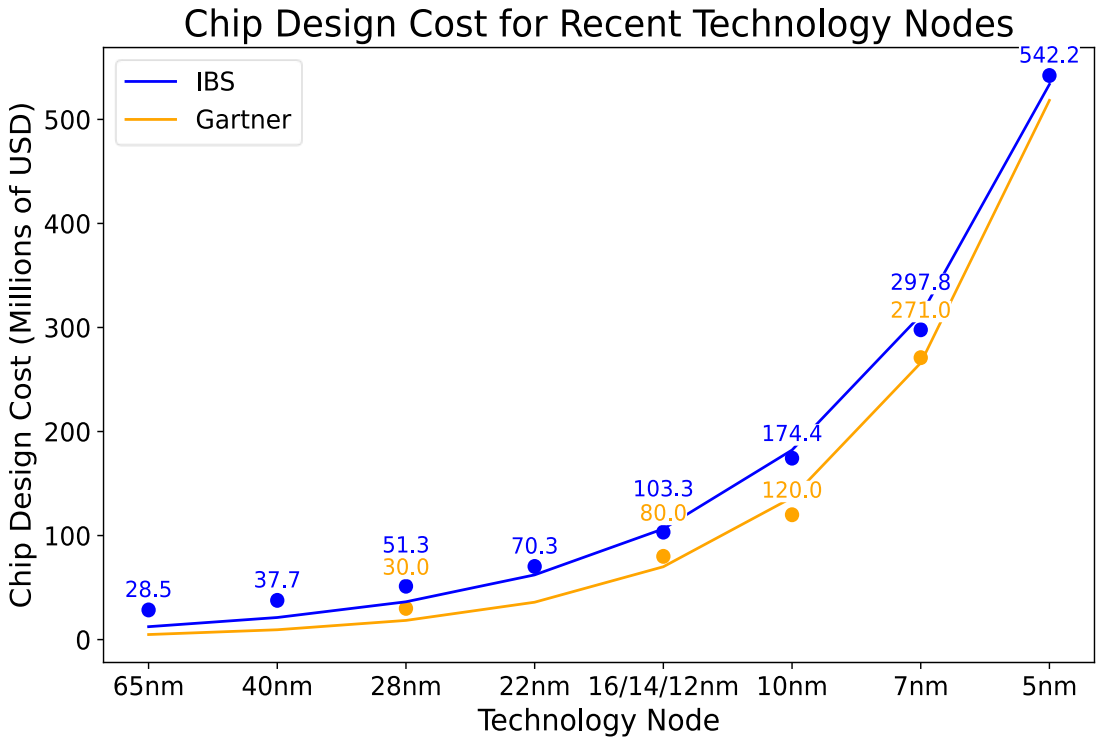


Data from: Khan and Mann, "AI Chips: What They Are and Why They Matter," April 2020.

Scale and Non-Recurring Engineering

- Chip design and fabrication have substantial non-recurring costs (NRE) which must be amortized with scale
- NRE costs are rising rapidly with new nodes due to increased design complexity
- Developing robust software stacks for new hardware adds substantial NRE

Low volume accelerators will struggle to effectively amortize NRE



Data from: Khan and Mann, "AI Chips: What They Are and Why They Matter," April 2020.

The Argument for Highly Specialized Chiplets

A chiplet designed to accelerate a substantial portion of a specific application

- Focused accelerator design reduces chip area and by extension NRE and per-chip cost
- Can rely on existing software stacks for functions not specific to the accelerator
- Large granularity offload reduces overheads associated with using accelerators
- Chiplets can provide greater bandwidth and lower latency than PCB-based interconnects

The Argument for Highly Specialized Chiplets

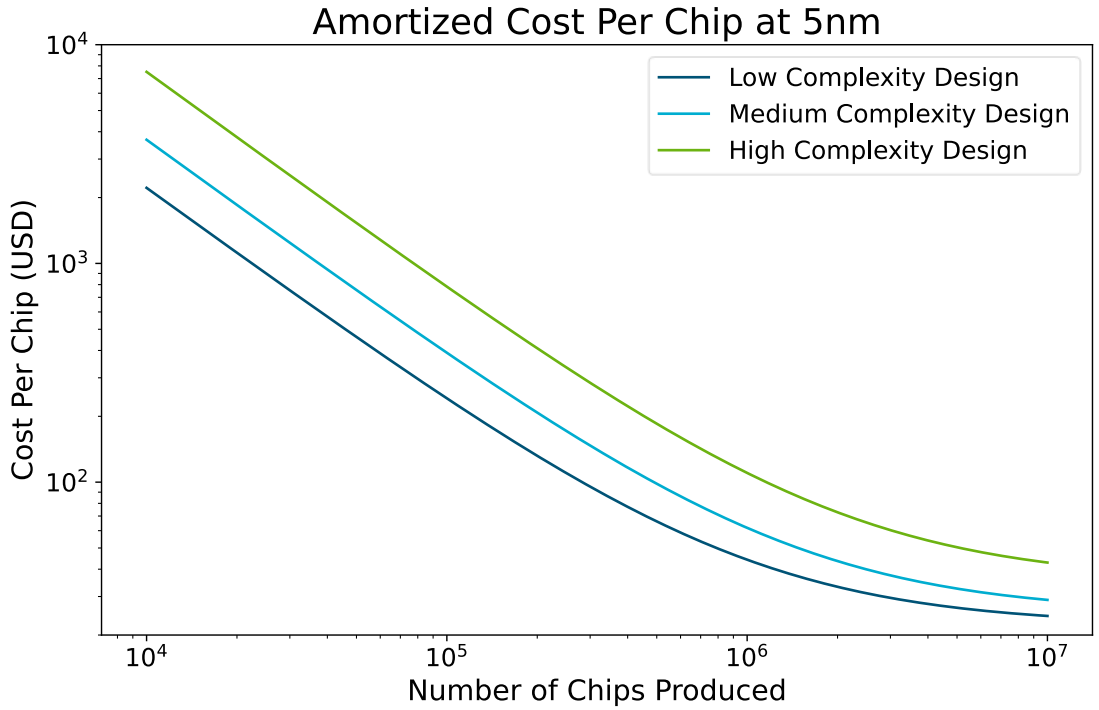
A chiplet designed to accelerate a substantial portion of a specific application

- Focused accelerator design reduces chip area and by extension NRE and per-chip cost
- Can rely on existing software stacks for functions not specific to the accelerator
- Large granularity offload reduces overheads associated with using accelerators
- Chiplets can provide greater bandwidth and lower latency than PCB-based interconnects

Thesis: Chiplets make the development of specialized accelerators for applications of interest feasible

Amortization with Highly Specialized Chiplets

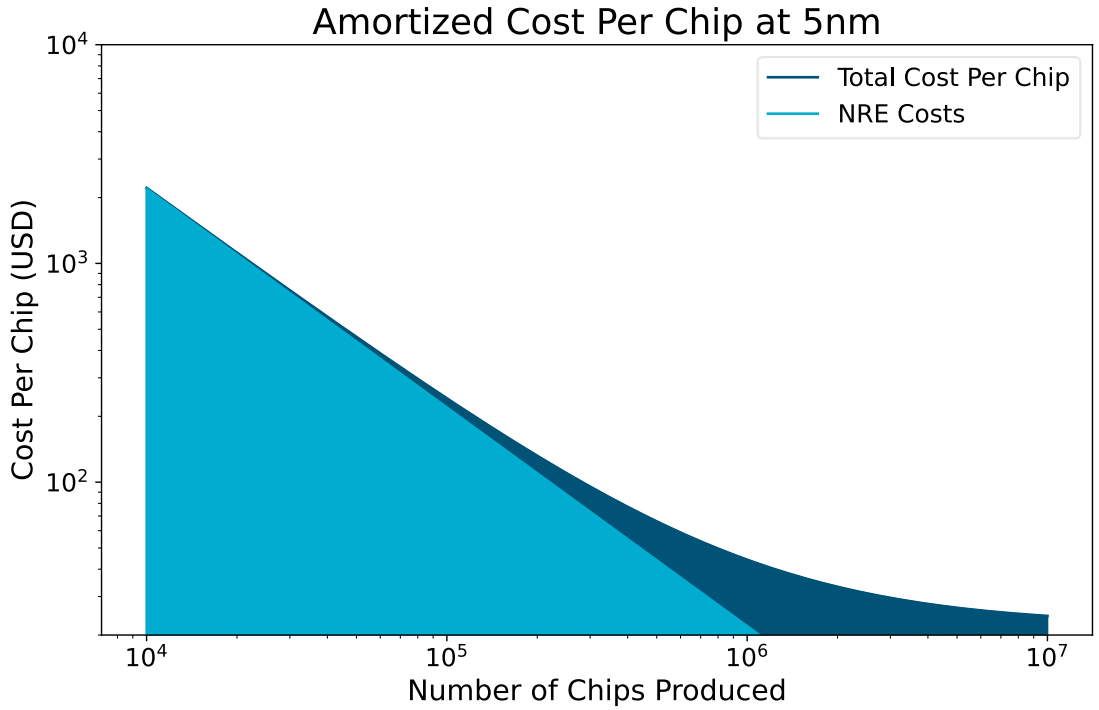
- Smaller chips do reduce costs, but on advanced nodes costs are still substantial
- Even for relatively small chiplets the cost for advanced nodes is unlikely to be effectively amortized at deployment scale



Based on model from: Ning, Tziantzioulis, and Wentzlaff, "Supply Chain Aware Computer Architecture," ISCA 2023.

Amortization with Highly Specialized Chiplets

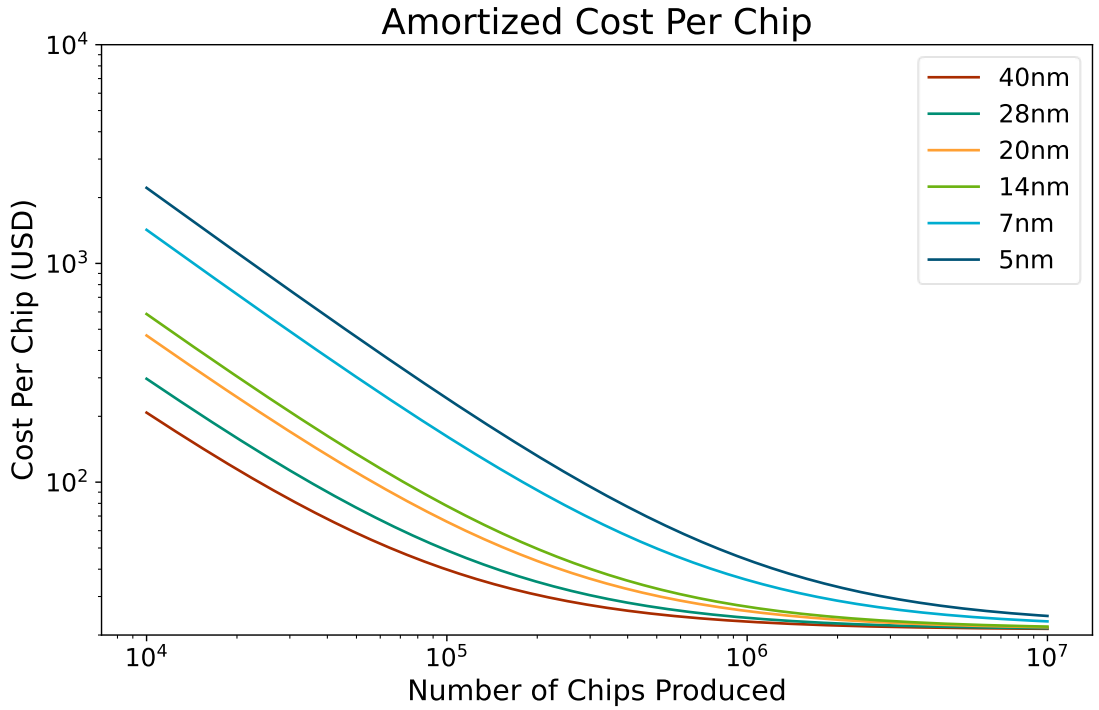
- Smaller chips do reduce costs, but on advanced nodes costs are still substantial
- Even for relatively small chiplets the cost for advanced nodes is unlikely to be effectively amortized at deployment scale



Based on model from: Ning, Tziantzioulis, and Wentzlaff, "Supply Chain Aware Computer Architecture," ISCA 2023.

Amortization with Highly Specialized Chiplets

- Smaller chips do reduce costs, but on advanced nodes costs are still substantial
- Even for relatively small chiplets the cost for advanced nodes is unlikely to be effectively amortized at deployment scale

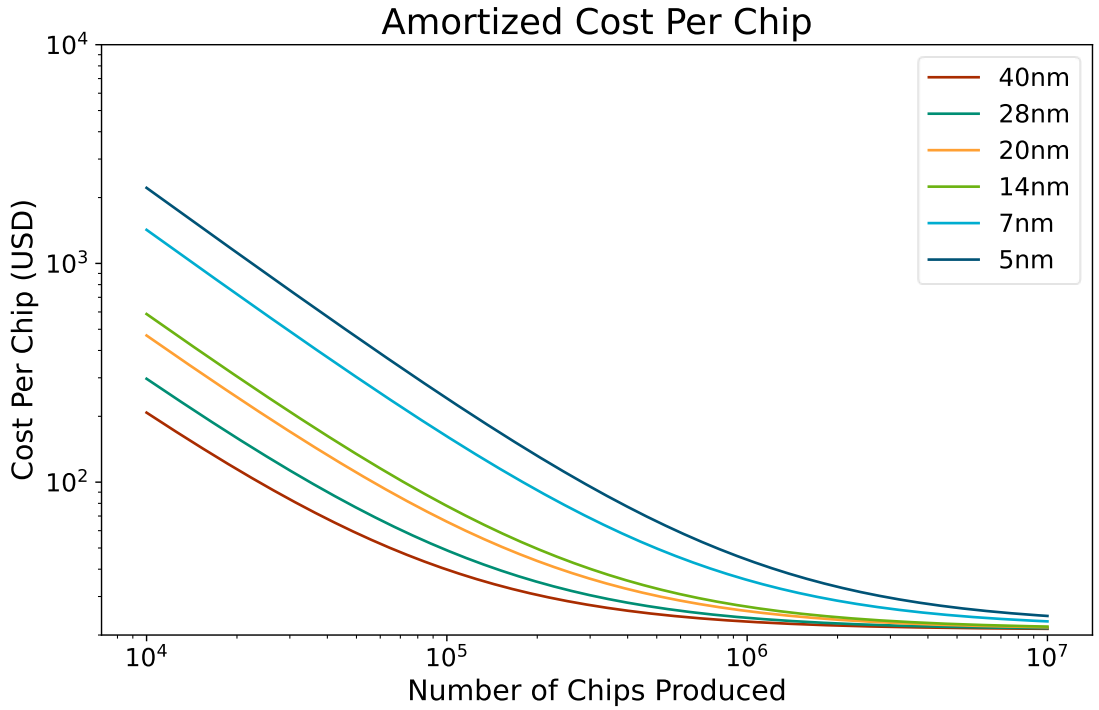


Based on model from: Ning, Tziantzioulis, and Wentzlaff, "Supply Chain Aware Computer Architecture," ISCA 2023.

Amortization with Highly Specialized Chiplets

- Smaller chips do reduce costs, but on advanced nodes costs are still substantial
- Even for relatively small chiplets the cost for advanced nodes is unlikely to be effectively amortized at deployment scale

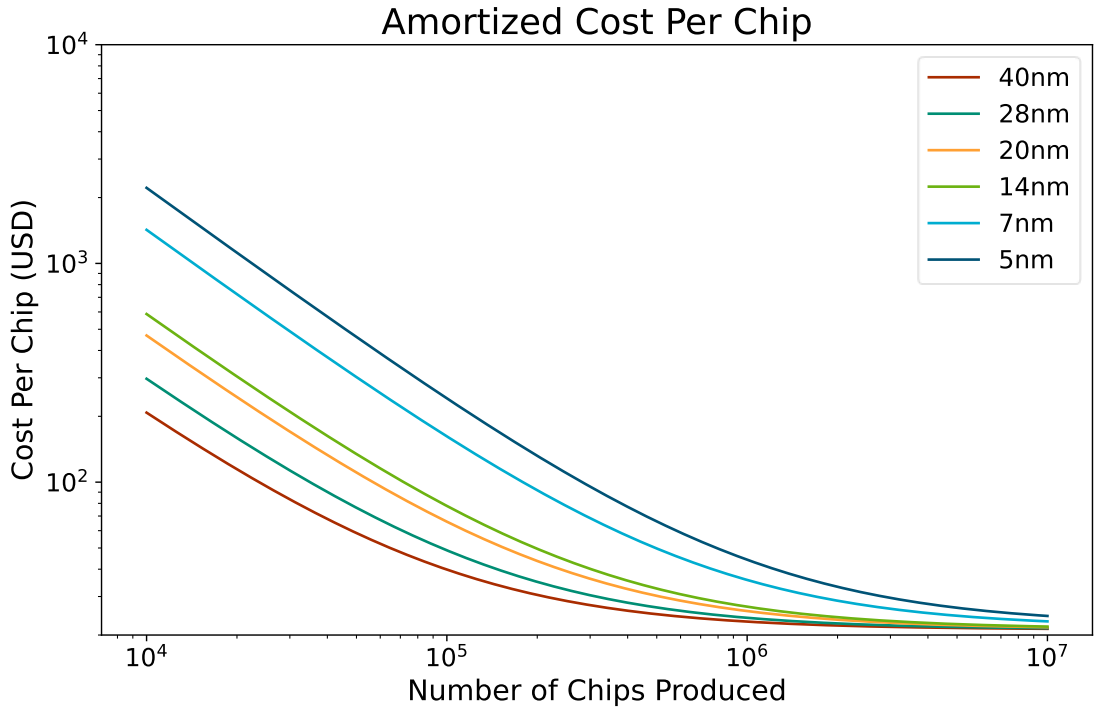
Highly specialized chiplets should consider less advanced nodes to deal with reduced scale



Based on model from: Ning, Tziantzioulis, and Wentzlaff, "Supply Chain Aware Computer Architecture," ISCA 2023.

Chiplets Enable Process Heterogeneity

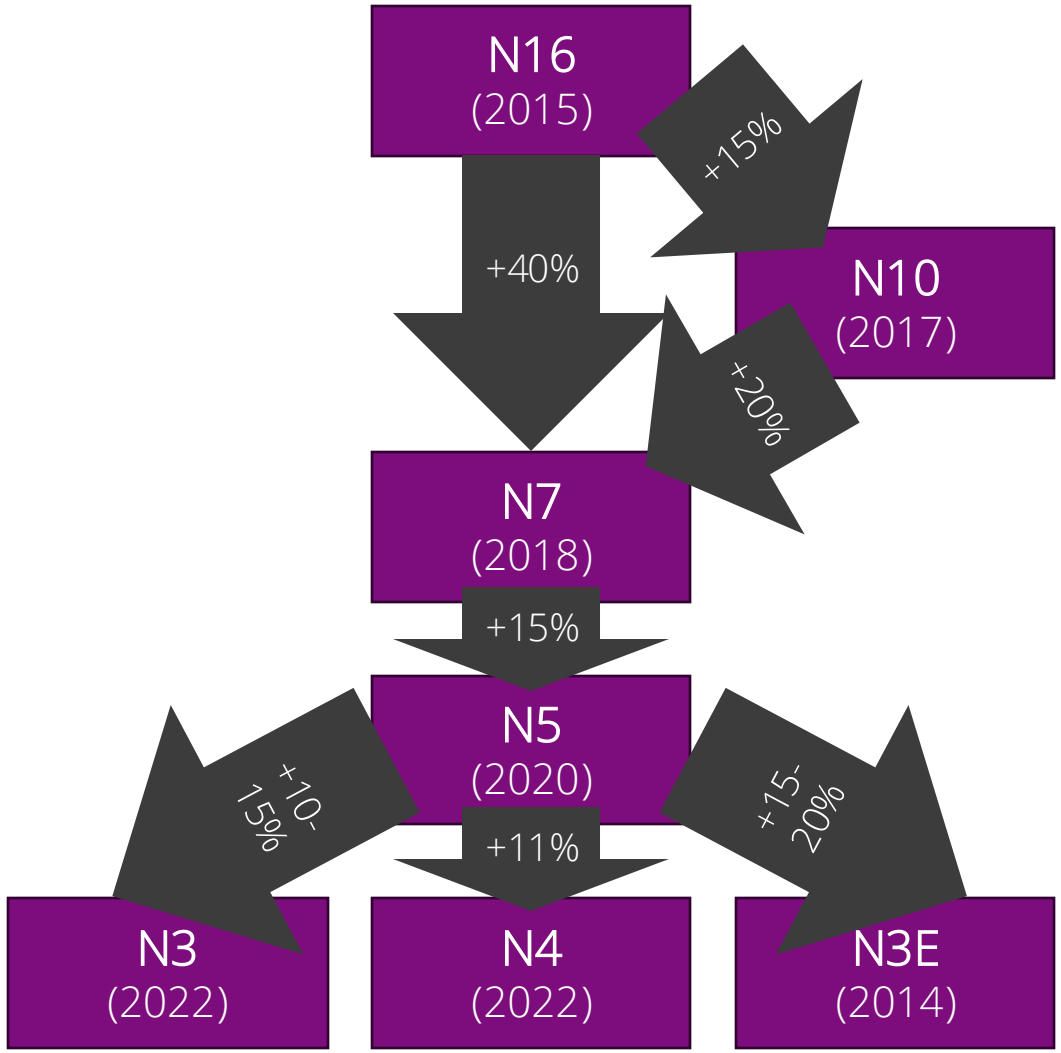
- Chiplets do not need to be on the same process node and can use the node that best fits the chiplet's needs
  - "Rather than halving the die size from going to 7nm, we would only achieve approximately a 28% reduction (i.e., 56%/2 + 44%=72%). When considering the relative cost increase of transitioning from 14nm to 7nm, this 28% die size reduction might not be sufficient to even reach a cost break-even point compared to the original 14nm "Zeppelin" chiplet."
    [*Naffziger et al,* ***ISCA 2021***]



Based on model from: Ning, Tziantzioulis, and Wentzlaff, "Supply Chain Aware Computer Architecture," ISCA 2023.

Specialization vs Technology Nodes

- Despite the end of Dennard Scaling, nodes are still improving year over year albeit at a reduced rate
- Transistor density scaling has also slowed, but still scaling achieving substantial node-to-node improvements

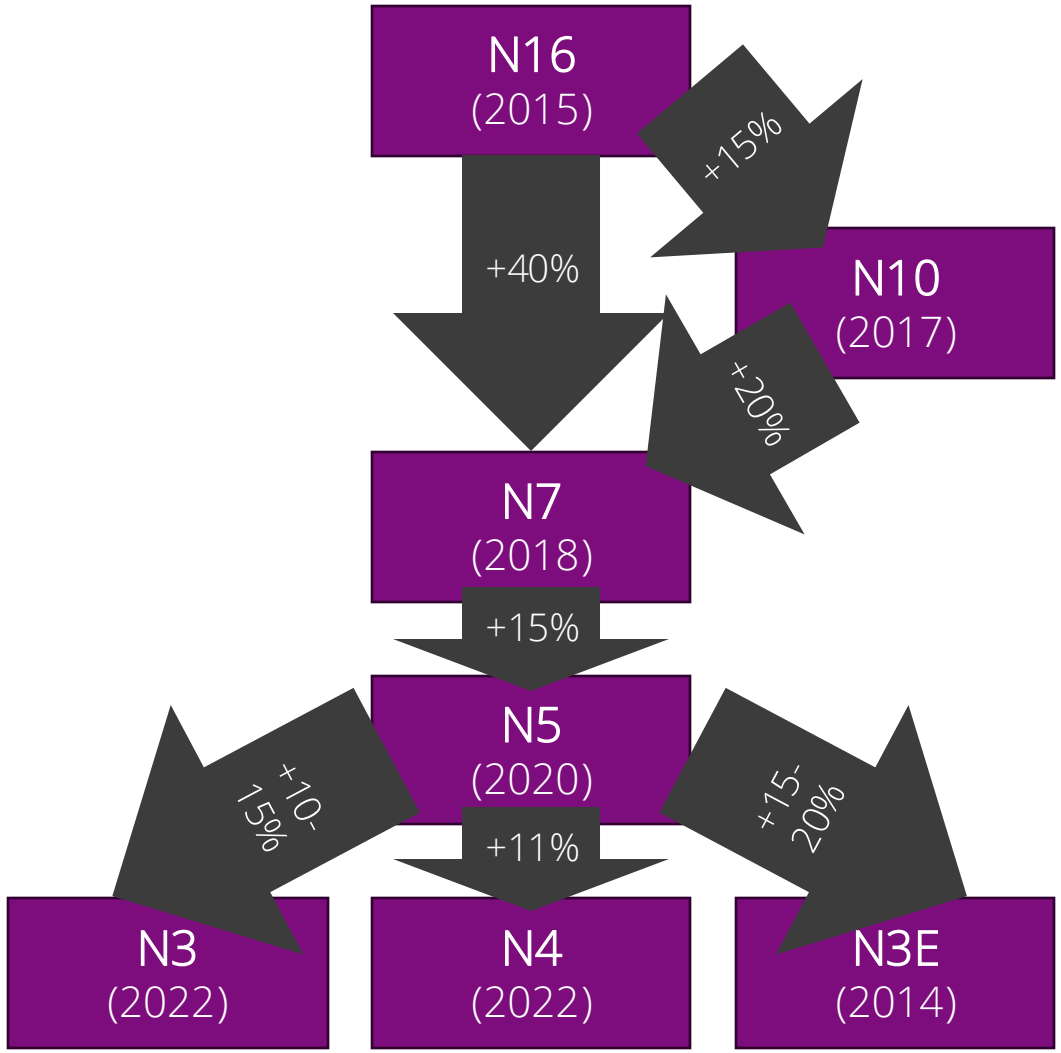


TSMC claimed node-over-node iso-power performance improvements

Specialization vs Technology Nodes

- Despite the end of Dennard Scaling, nodes are still improving year over year albeit at a reduced rate
- Transistor density scaling has also slowed, but still scaling achieving substantial node-to-node improvements

If we have to use less advanced nodes, can specialization still provide an edge?

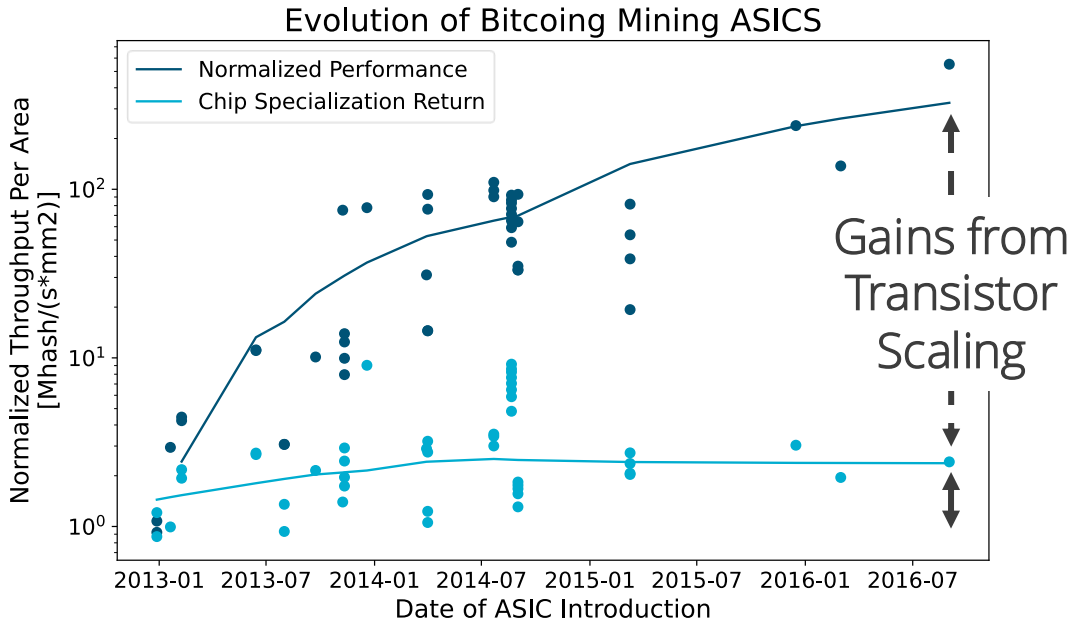


TSMC claimed node-over-node iso-power performance improvements

The Benefits of Specialization?

- Benefits of specialization can be difficult to predict without significant design effort
- As low-hanging optimizations are applied Amdahl's Law becomes a bottleneck

- **"[F]or mature computation domains…specialization returns either plateau or drop for high performing chips."**
  [Fuchs and Wentzlaff, HPCA'19]

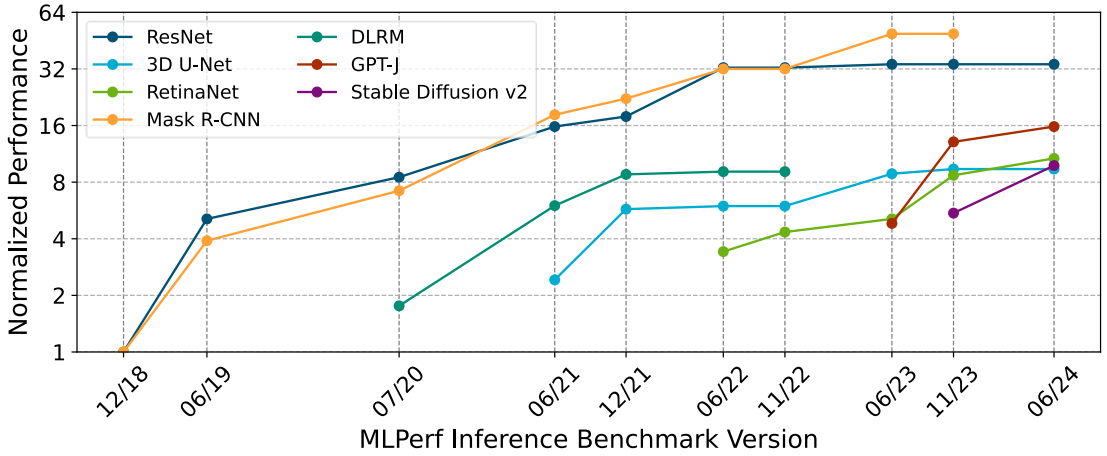


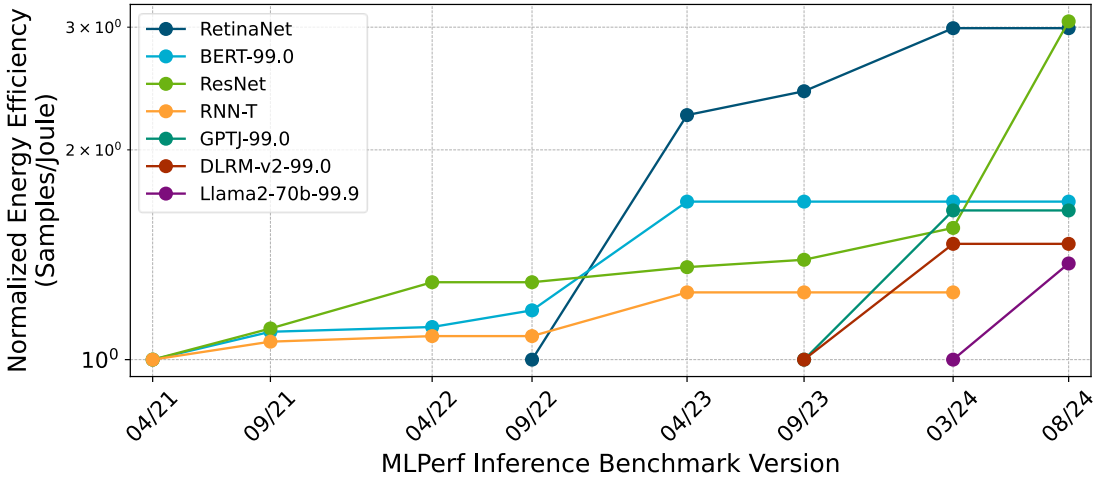
Data from: Fuchs and Wentzlaff, "The Accelerator Wall: Limits of Chip Specialization," HPCA 2019.

The Benefits of Specialization?

- Benefits of specialization can be difficult to predict without significant design effort
- As low-hanging optimizations are applied Amdahl's Law becomes a bottleneck

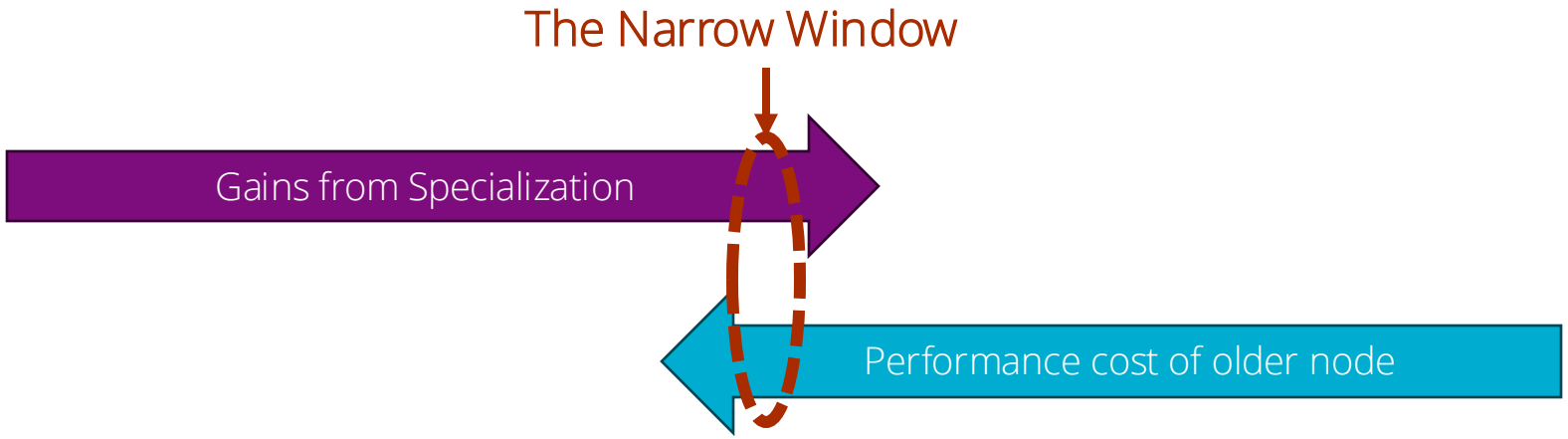
- **"[F]or mature computation domains…specialization returns either plateau or drop for high performing chips."**
  [Fuchs and Wentzlaff, HPCA'19]





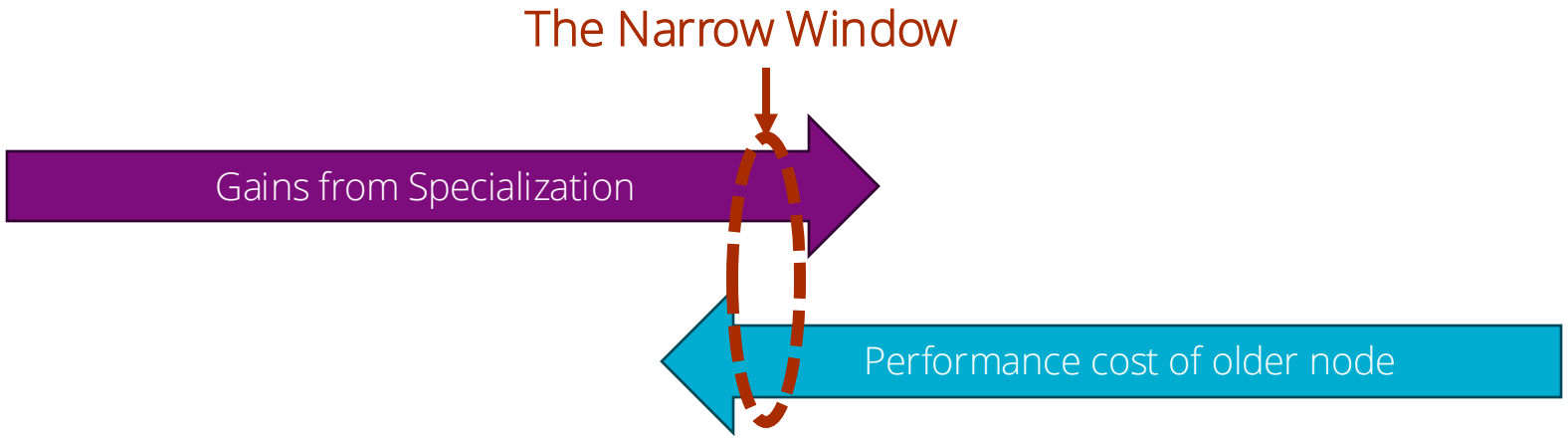
Data from: Tschand and Rajan et al, "MLPerf Power: Benchmarking the Energy Efficiency of Machine Learning Systems from Microwatts to Megawatts for Sustainable AI," HPCA 2024.

The Narrow Window for Highly Specialized Chiplets



Highly specialized chiplets will only provide significant improvements in narrow cases with high specialization gain or low node-to-node improvements

The Unfortunate Economics



How can use chiplets in a way that avoids the narrow window problem?

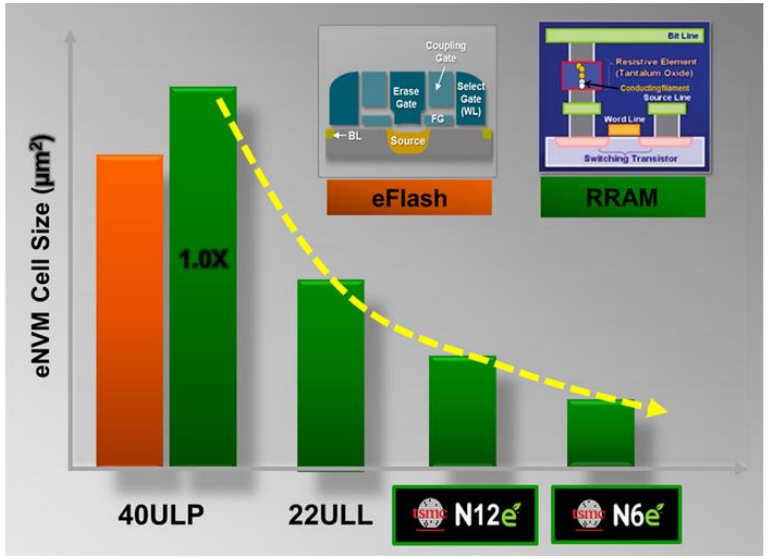
Another Narrow Window: Emerging Memory Technologies

- Slowdowns in SRAM and DRAM scaling create an opportunity for new memory technologies
- Extensive research on multiple technologies, MRAM, PCM, RRAM/ReRAM with some productization
- A common impression is large-scale deployment of emerging memories have been “a few years away” for the past decade

TSMC Logic-Compatible RRAM Supports Firmware, Data Storage and Security Memory

TSMC’s industry-leading Resistive Random Access Memory (RRAM) CMOS process provides good scalability, power reduction and logic migration. The non-volatile memory RRAM cell, formed between backend metal layers, is an excellent eFlash replacement for general micro-controlling units (MCUs) and Internet of Things (IoT) applications to support firmware, data storage, and security memory.

TSMC’s 40RRAM and 22RRAM with optimized power and form factor are moving into production. A 12RRAM is under development for next-generation AIoT products.



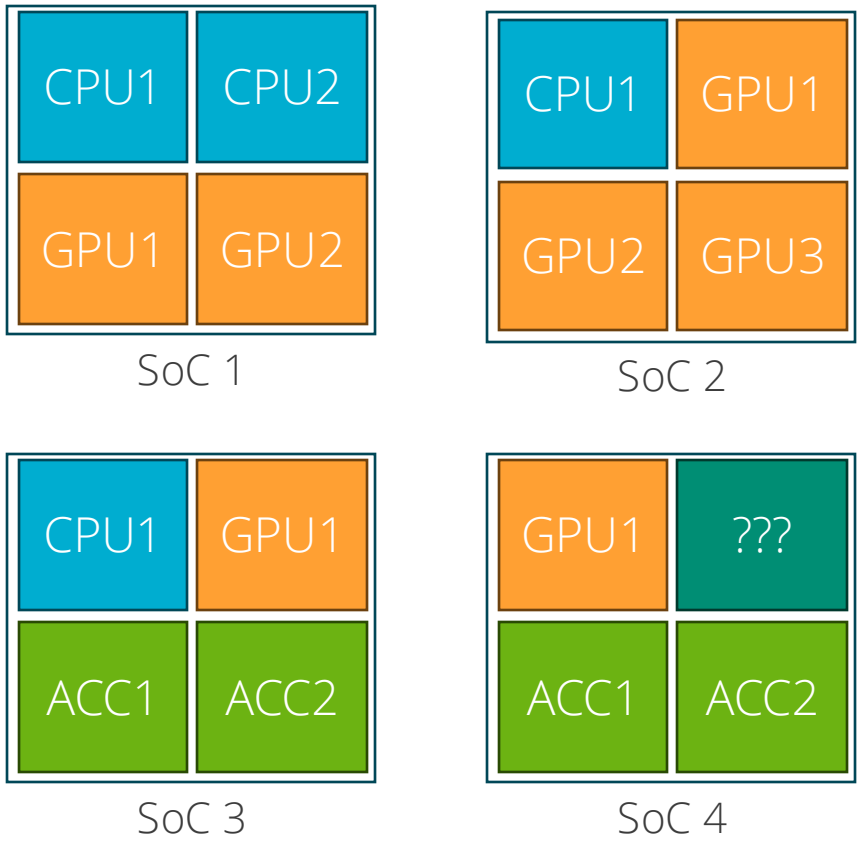
https://www.tsmc.com/english/dedicatedFoundry/technology/platform_IoT_tech_NVM

Chiplets for Semi Specialized Accelerators

- Focus on more generally applicable functions that are still performance critical
  - Scatter/Gather acceleration
  - Augmented FP64 matrix or vector engines
  - High density caches with novel memory technologies
- More potential users ⇒ greater amortization ⇒ wider window for success
- Higher bandwidth and lower latency can help mitigate the “narrow straw” problem of dedicated compute accelerators
- Process heterogeneity creates new opportunities for integrating new technologies

Chiplets for Specialized SoCs

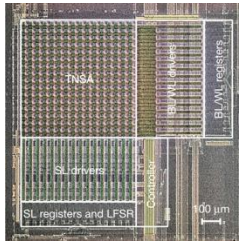
- With *a robust chiplet ecosystem* and *open standards* there are reduced barriers to customized combinations of compute chiplets
- Lower packaging costs compared to chip fabrication reduce the break-even point for custom SoCs
- Specialized SoCs provide an on-ramp for semi specialized accelerators
- Still substantial challenges around standardization and physical design



Chiplets for Novel Technologies (e.g., Analog MVM)

ReRAM
256×256 arrays
(Stanford)

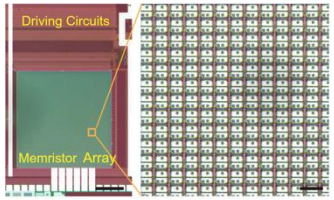
Wan et al, *Nature*, 2022 (130 nm)



Accuracy, performance, and energy demonstrator

ReRAM
256×256 arrays
(USC, Tetramem)

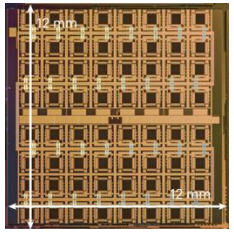
Song et al, *Science*, 2024 (65 nm)



Accuracy demonstrator

PCM
256×256 arrays
(IBM)

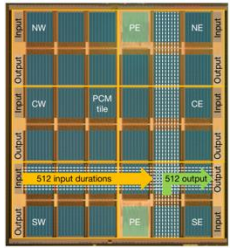
Le Gallo et al, *Nature Electronics*, 2023 (14 nm)



Accuracy, performance, and energy demonstrator

PCM
512×2048 arrays
(IBM)

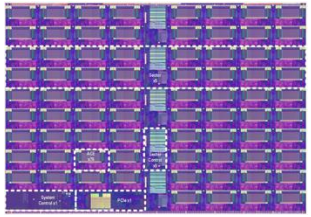
Ambrogio et al, *Nature*, 2023 (14 nm)



Accuracy, performance, and energy demonstrator

NOR flash
1024×2048 arrays
(Mythic)

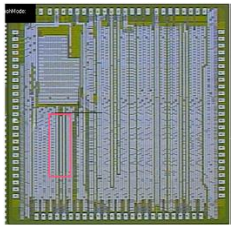
Fick et al, *ISSCC*, 2022 (40 nm)



Accuracy, performance, and energy demonstrator

SONOS flash
1024×1024 arrays
(Infineon)

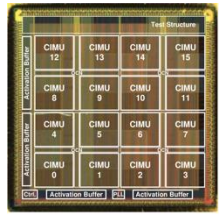
Agrawal, Xiao, Feinberg, et al, *IEDM*, 2022 (40 nm)



Accuracy demonstrator

SRAM
1152×256 arrays
(Princeton, Encharge AI)

Jia et al, *ISSCC*, 2023 (16 nm)

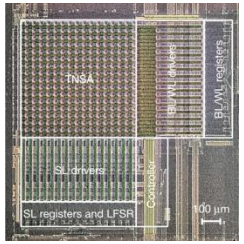


Accuracy, performance, and energy demonstrator

Chiplets for Novel Technologies (e.g., Analog MVM)

ReRAM
256×256 arrays
(Stanford)

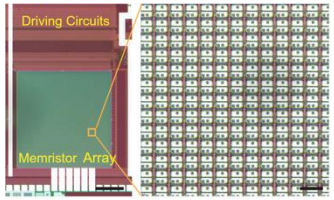
Wan et al, *Nature*, 2022 (130 nm)



Accuracy, performance, and energy demonstrator

ReRAM
256×256 arrays
(USC, Tetramem)

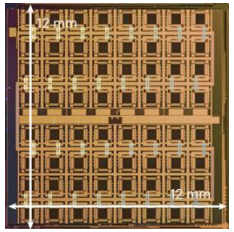
Song et al, *Science*, 2024 (65 nm)



Accuracy demonstrator

PCM
256×256 arrays
(IBM)

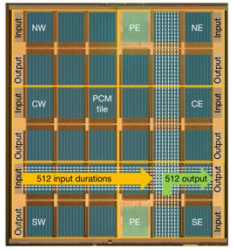
Le Gallo et al, *Nature Electronics*, 2023 (14 nm)



Accuracy, performance, and energy demonstrator

PCM
512×2048 arrays
(IBM)

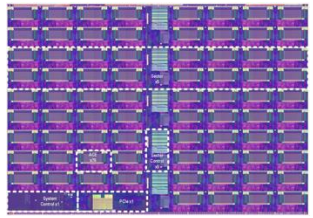
Ambrogio et al, *Nature*, 2023 (14 nm)



Accuracy, performance, and energy demonstrator

NOR flash
1024×2048 arrays
(Mythic)

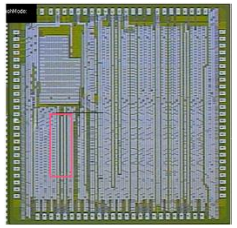
Fick et al, *ISSCC*, 2022 (40 nm)



Accuracy, performance, and energy demonstrator

SONOS flash
1024×1024 arrays
(Infineon)

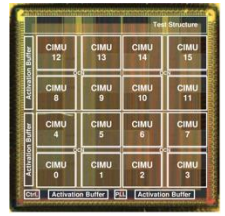
Agrawal, Xiao, Feinberg, et al, *IEDM*, 2022 (40 nm)



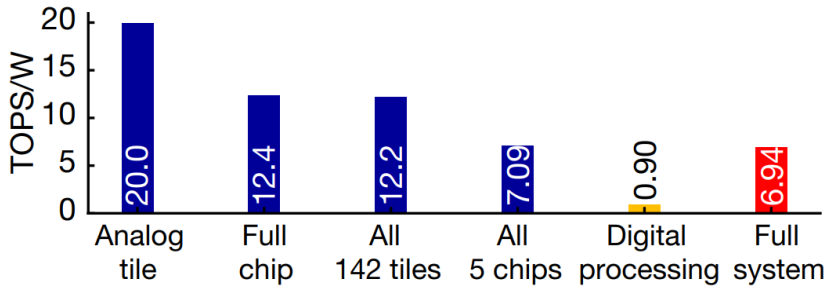
Accuracy demonstrator

SRAM
1152×256 arrays
(Princeton, Encharge AI)

Jia et al, *ISSCC*, 2023 (16 nm)



Accuracy, performance, and energy demonstrator



Benchmarked on MLPerf's RNNT speech transcription model (45M weights)

Conclusions

- Chiplets can lower the barrier of entry for customization and unlock new capabilities for architects with customized accelerators and SoCs

BUT

Conclusions

- Chiplets can lower the barrier of entry for customization and unlock new capabilities for architects with customized accelerators and SoCs

BUT

- Chiplets must be deployed judiciously to the avoid the narrow window problem from over-specialization

Conclusions

- Chiplets can lower the barrier of entry for customization and unlock new capabilities for architects with customized accelerators and SoCs

BUT

- Chiplets must be deployed judiciously to the avoid the narrow window problem from over-specialization

- Numerous open questions around standardization and effective use of chiplets that will require collaboration between academic/laboratory researchers and vendors