SNIA DEVELOPER CONFERENCE

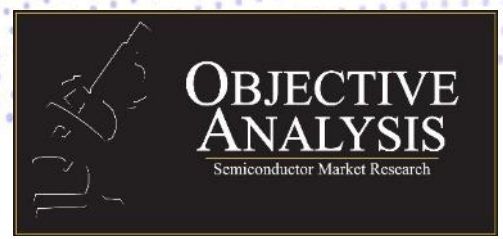
SDC 25

By Developers FOR Developers

Hyatt Regency Santa Clara, CA
September 15-17, 2025

The Processor Chip of the Future!

Chiplets, UCIe, Persistent Memory, and Heterogeneous Integration

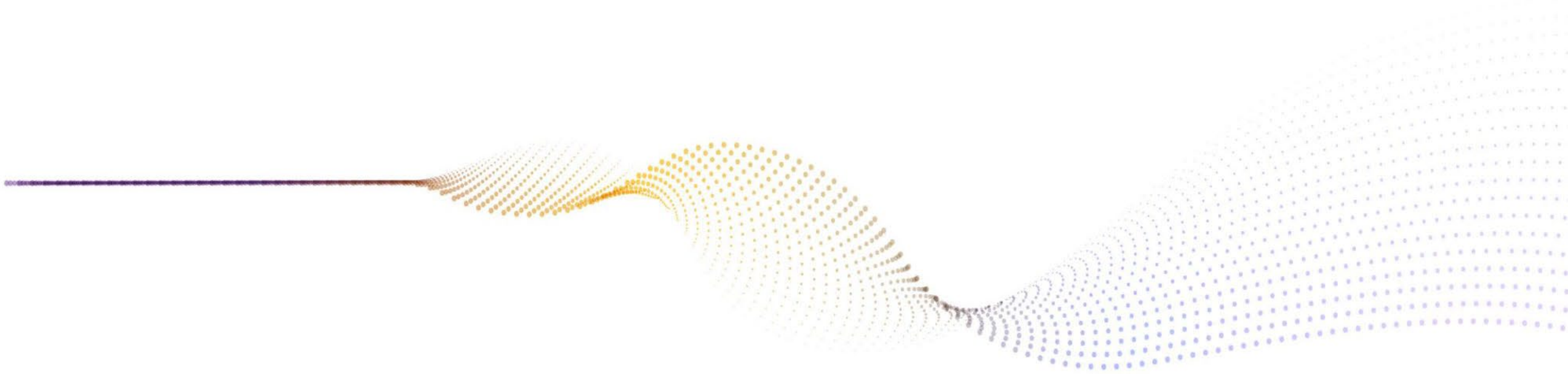


Coughlin Associates

www.sniadeveloper.org

Outline

- AI: The chiplet success story
- New memories and chiplets
- Changing the story on energy
- Boosting cost/performance

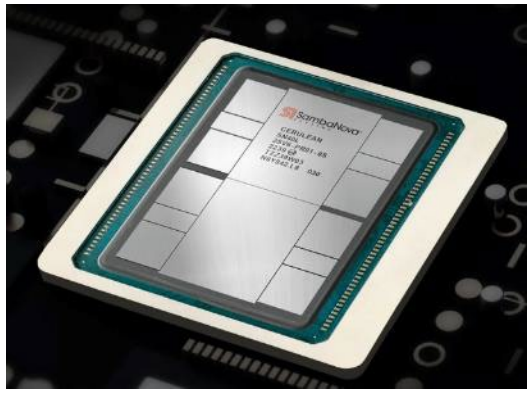


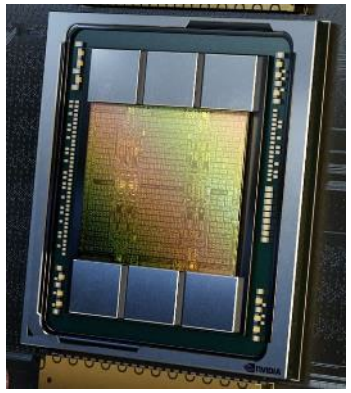
AI and Chiplets

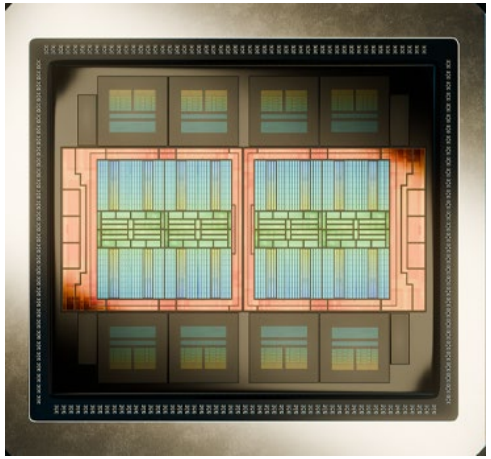
What *IS* a Chiplet???

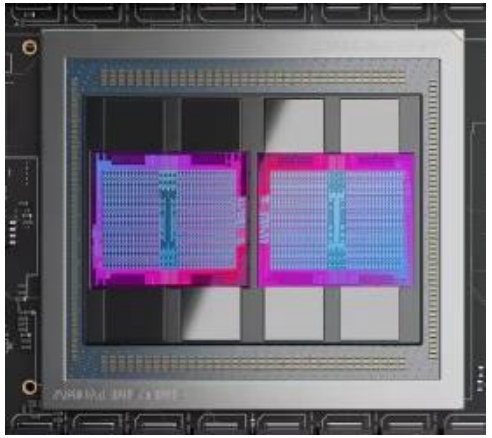
- Breaks one impossibly large chip into smaller ones
  - Can build a very large chip from smaller "chiplets"
- Moves us past Moore's Law limitations
  - Moore's law: Smaller transistors x larger chips + "Cleverness"
  - Lithography equipment can't make a chip larger than ~850mm²
    - The "Reticle Limit"
- Ideally, data would never leave the chip
  - Faster data communication
  - Lower power consumption
  - Chiplets allow this on a larger scale than before

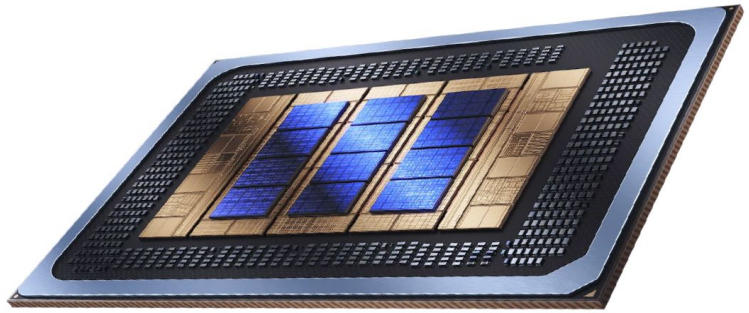
Chiplets Are Already Widely Used



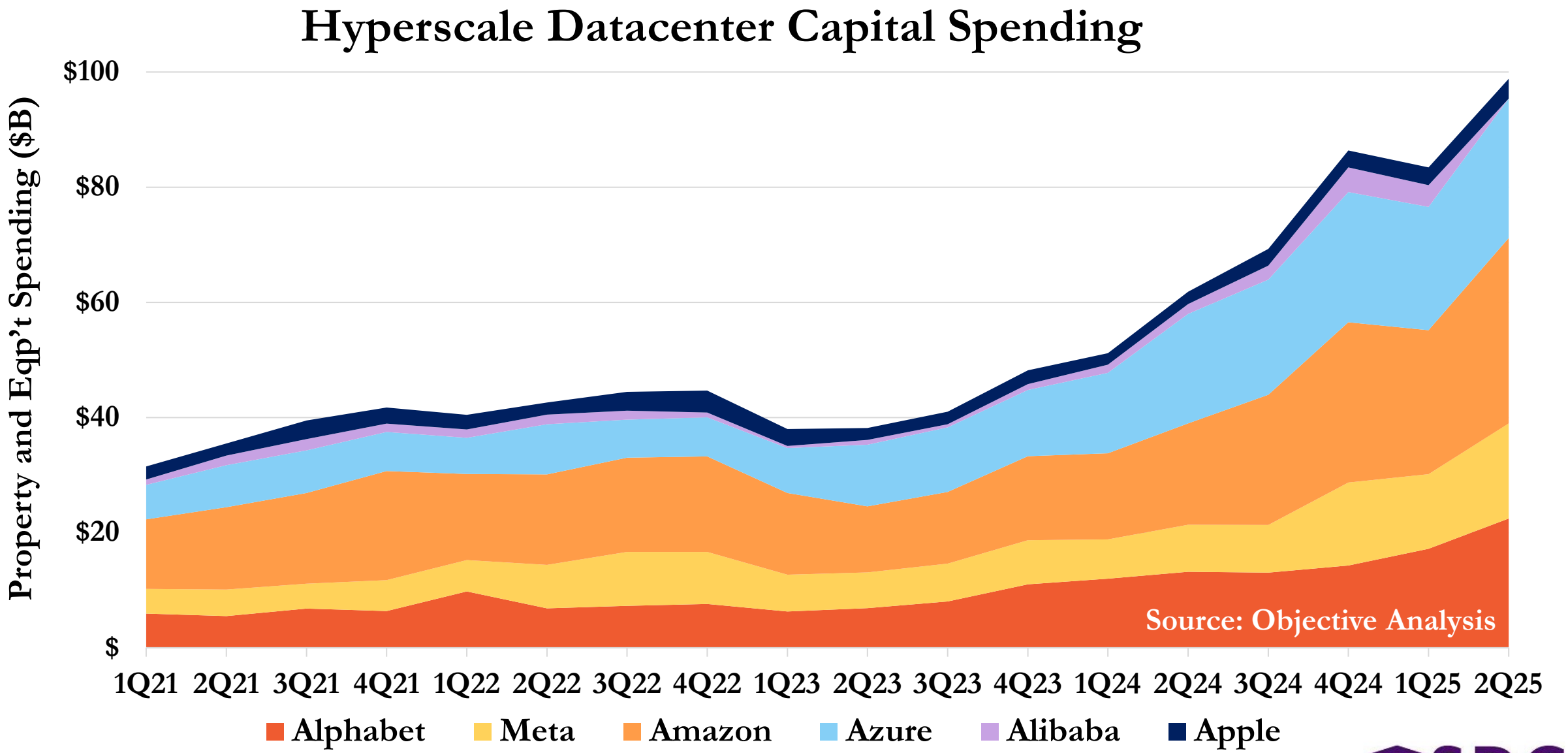




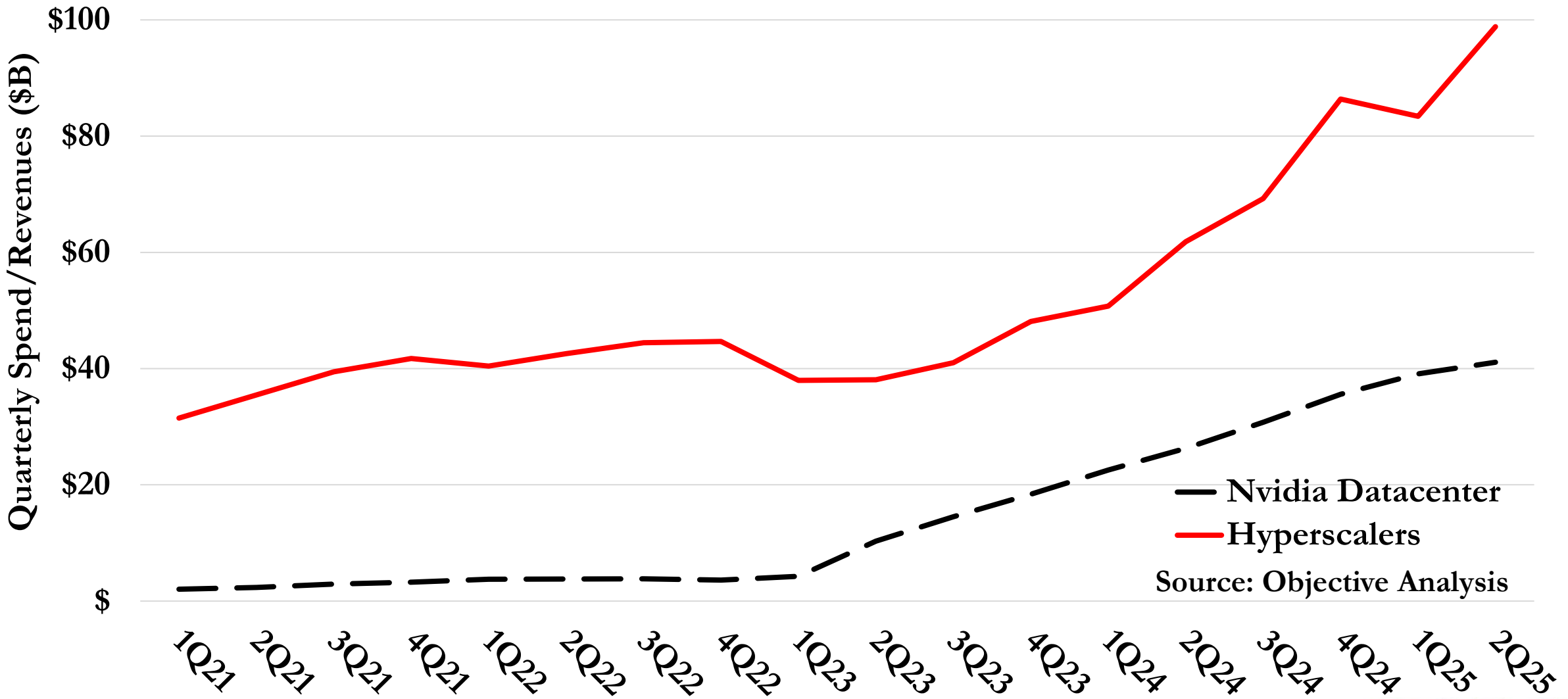




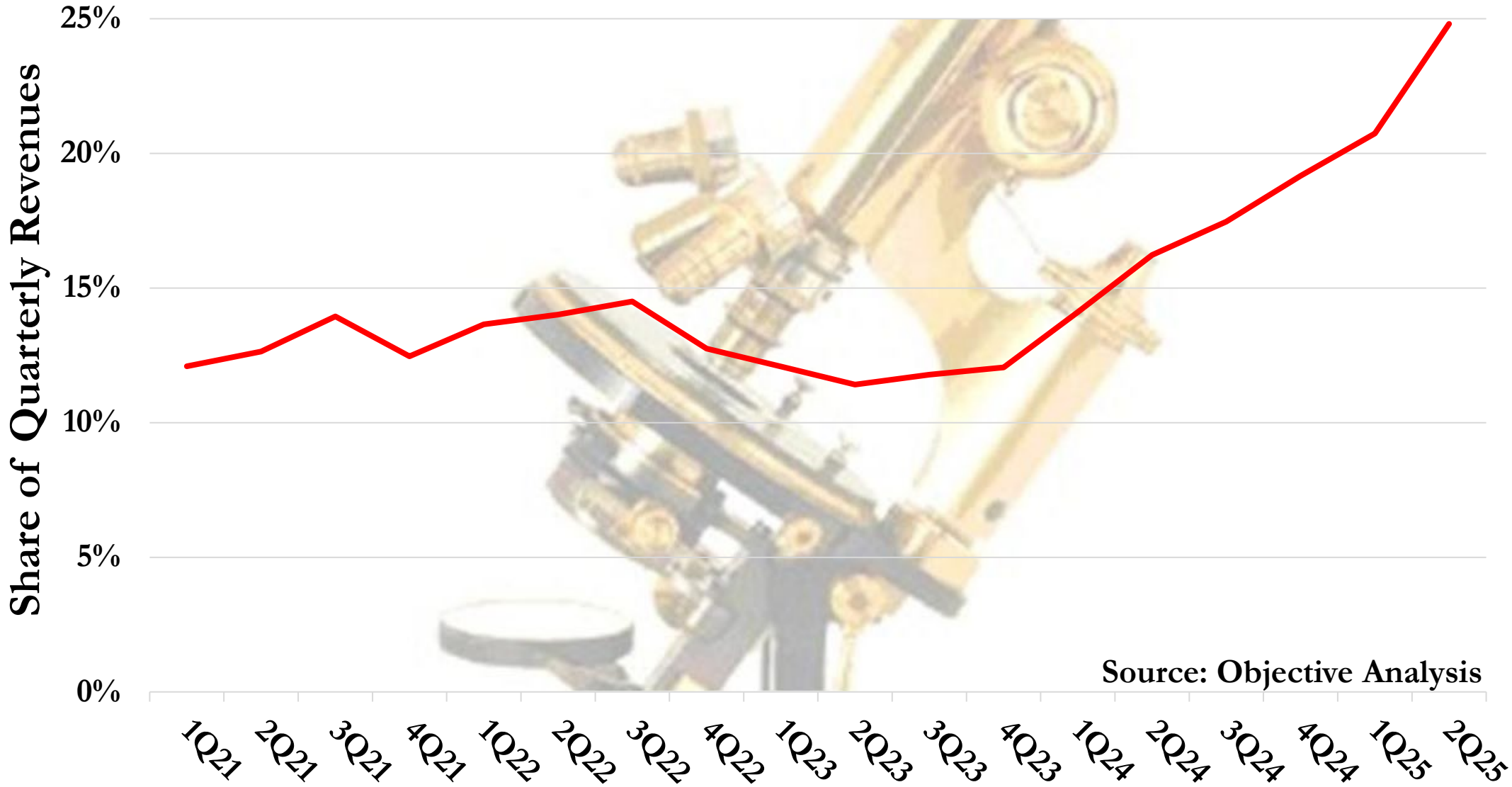
Hyperscaler CapEx Binge Continues



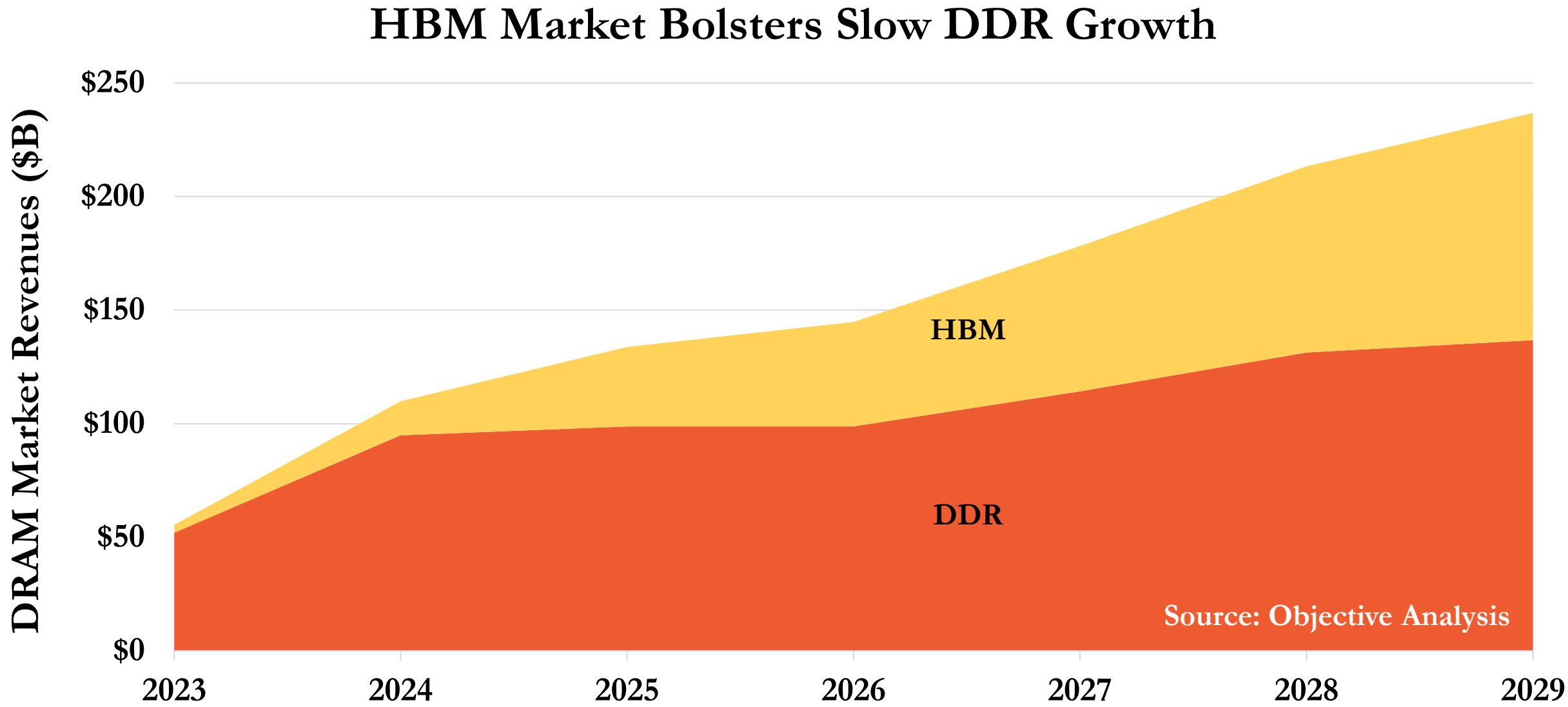
Nvidia Datacenter Growth is Steady



CapEx is a Growing Share of Hyperscaler Revenues



The DRAM Market Needs HBM

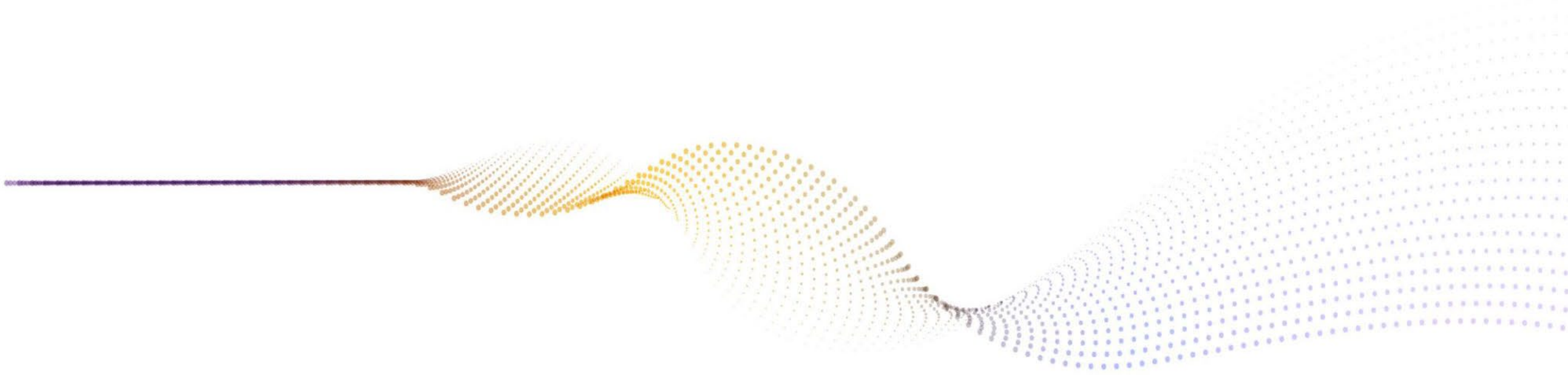


Will HBM Rescue the DRAM Market?

Upcoming report from Objective Analysis

- Covers all perspectives
  - What is HBM and why it's so costly
  - Supply Chain – who buys it and why
  - Price dynamics: HBM vs. DRAM
  - Forecast (Revenues, units, ASP)
- Coming soon for immediate download:

Objective-Analysis.com/reports



Chiplets Enable New Memory Types

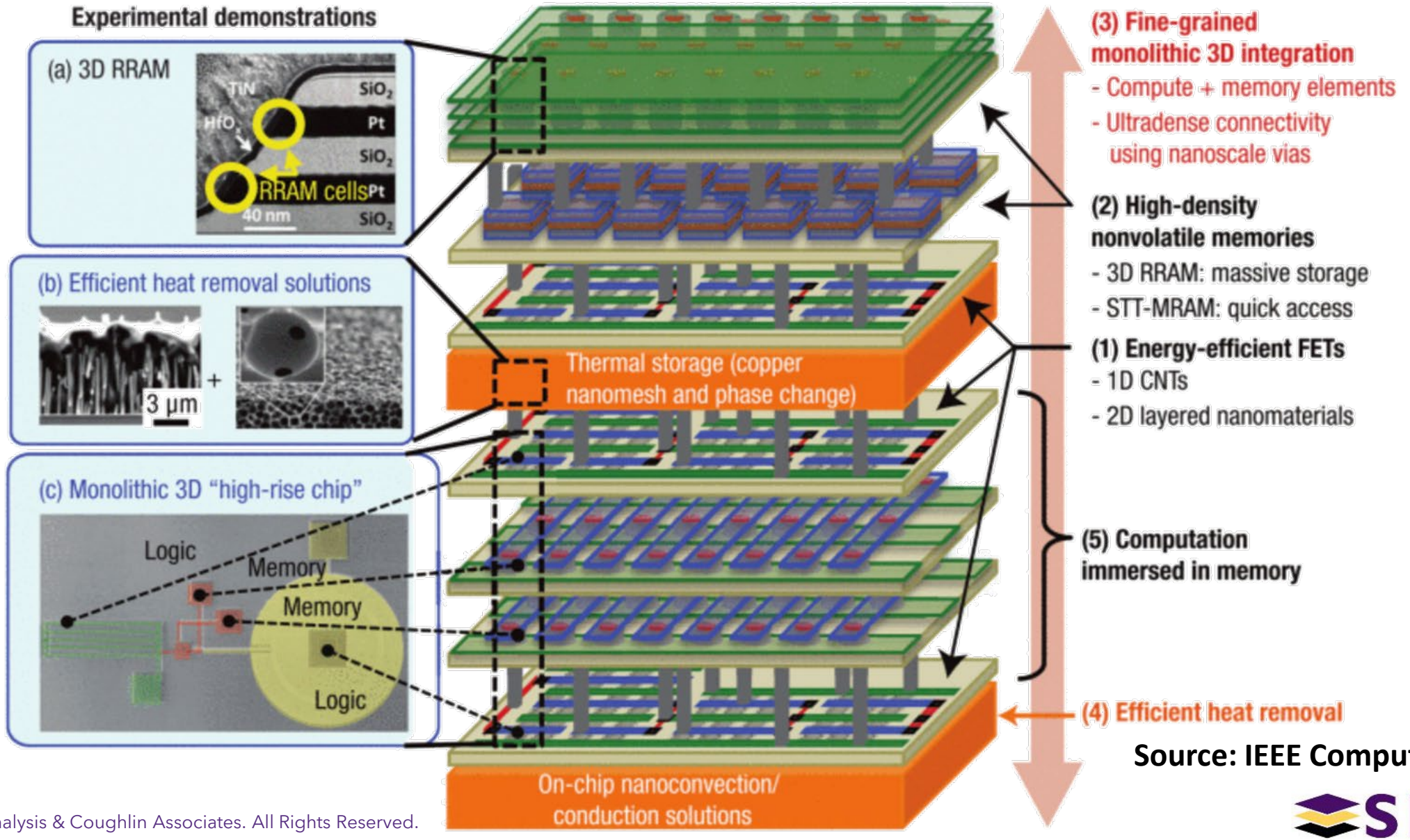
Chiplets Today

- Only for ultra-high-end applications
  - $10K+ package cost
- Very large die sizes
  - Two Blackwell GPUs: 814mm² each
  - Ninety-six DRAM chips: 120mm² each
    - Eight 12-high HBM stacks
  - Eight HBM base logic chips: 120mm² each
  - Total silicon area = 20% of a 300mm wafer
- Two basic processes: Logic and DRAM

Chiplets Tomorrow

- More chiplets in more applications
- Smaller die sizes
- Wider variety of semiconductor fabrication processes
  - Logic & DRAM, as before
  - NAND flash (HBF) and power management
  - Optical interconnect
  - MRAM, ReRAM, FRAM, and PCM
  - True heterogeneous integration
- Lower-cost end products
- More stacking too!

The Processor Complex of Tomorrow

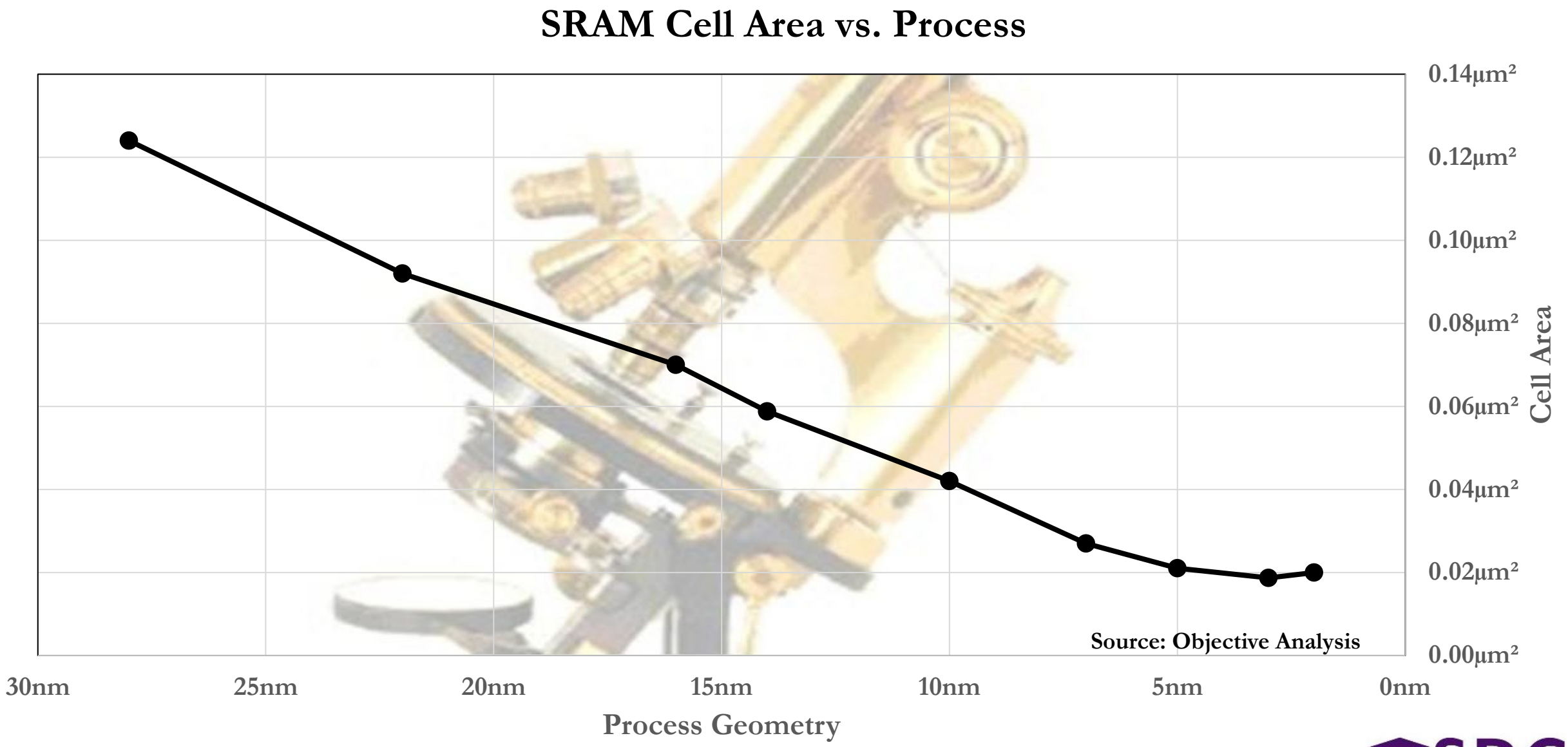


Source: IEEE Computer 2015

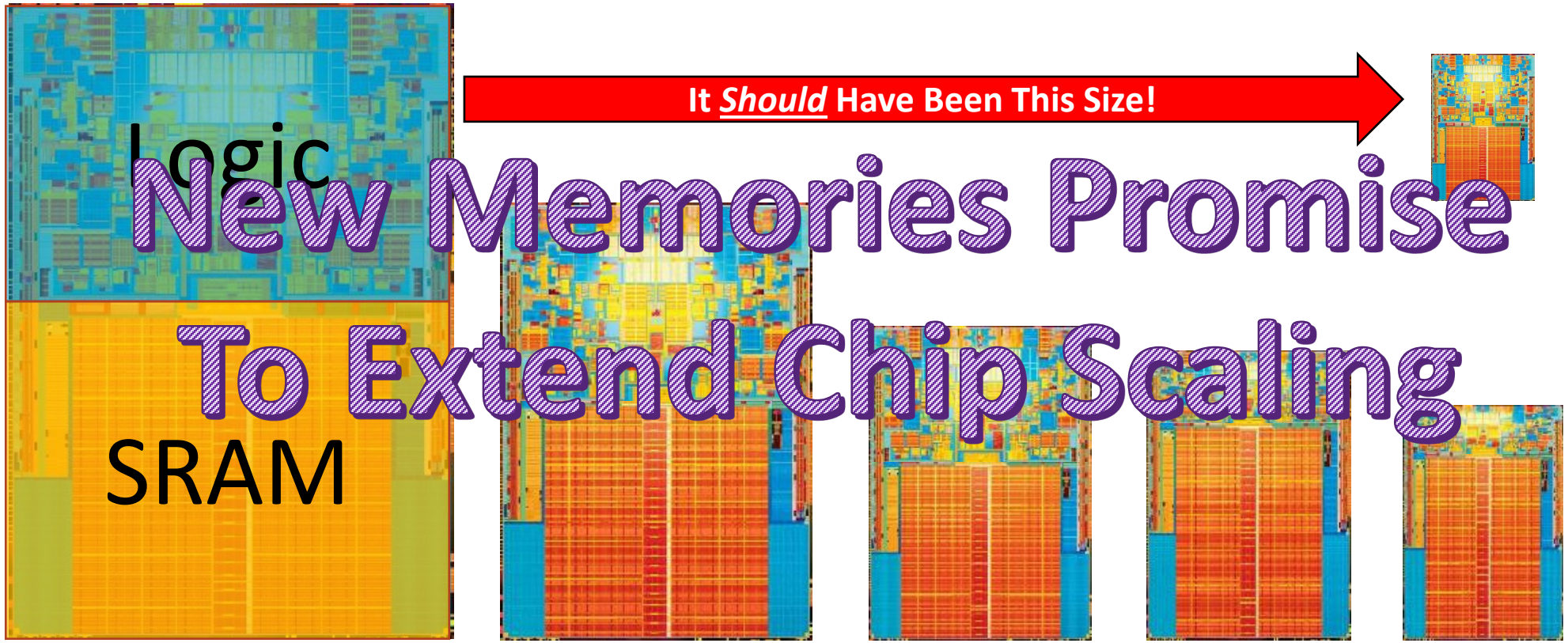
New Memories in Chiplets

- New memory processes and logic processes don't mix well
  - Some new memories use tricky new elements
- Even established memory & logic processes disagree
  - Memories make logic slower
  - Logic makes memory more costly
  - Mixed memory + logic = slow and costly
- More flexibility with chiplets than with embedded memories
  - Any memory can be paired with logic
  - Can scale like the memory options in cell phones: 256GB, 512GB, or 1TB

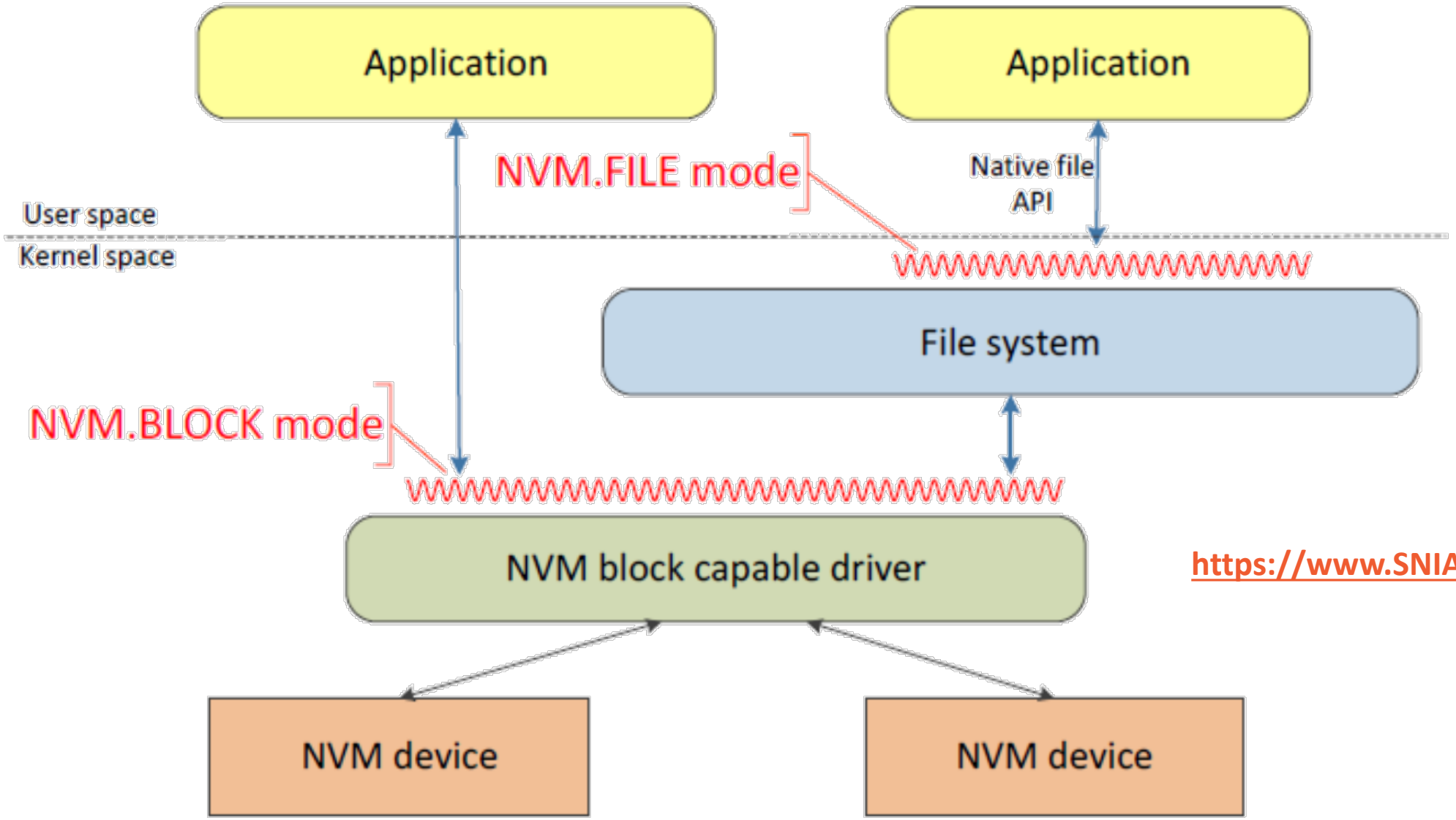
SRAM No Longer Scales with Process



SRAM Caches Limit Chip Scaling



SNIA's NVM Programming Model Works for Caches, Too!



https://www.SNIA.org/node/14641

Report: New Memories–Not Just for AI

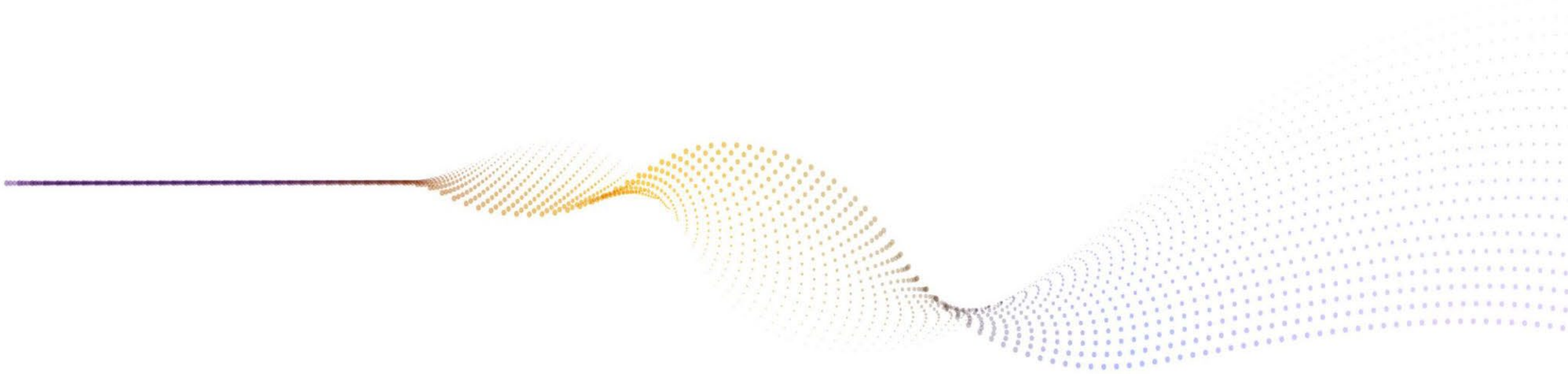
- Covers all new memory types
- Profiles over 175 companies
- Over 300 pages, 200 figures, & 34 tables
- Tech, market, & equipment forecasts
- Purchase for immediate download

Coughlin Associates

Objective Analysis
Semiconductor Market Research

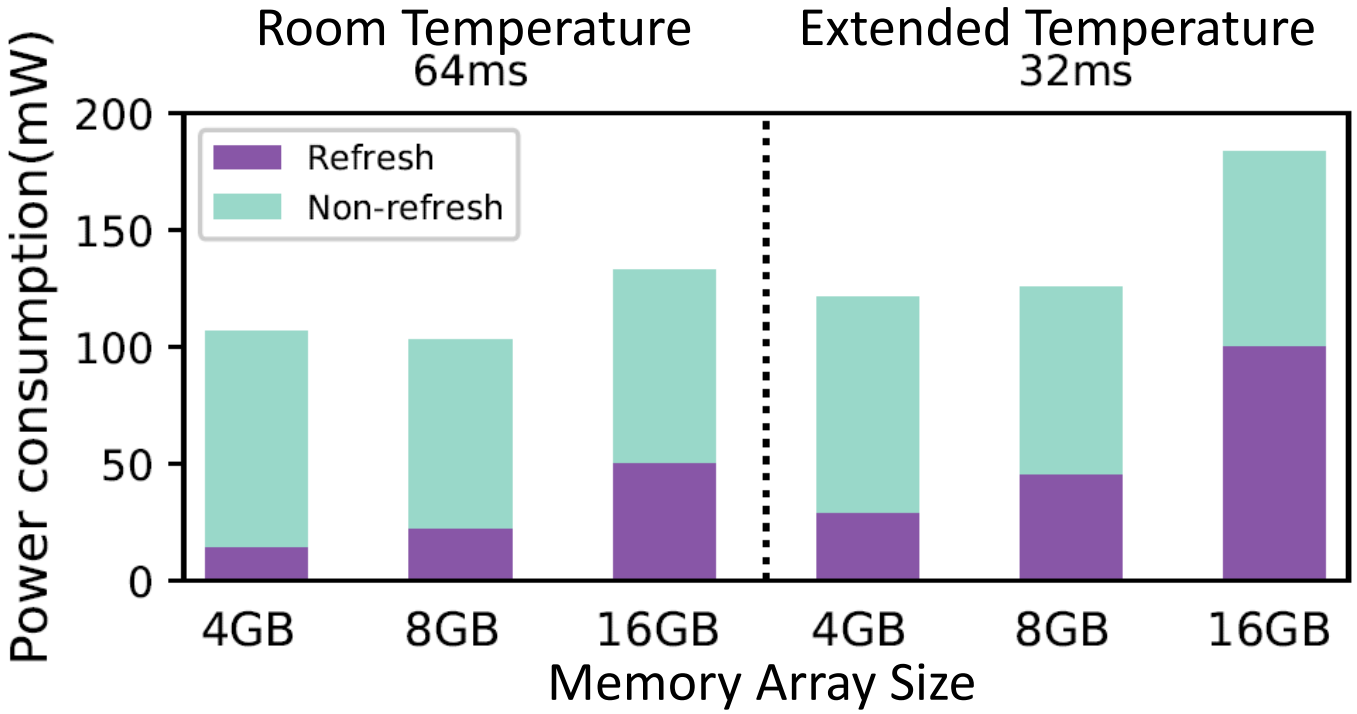
Coming in November!

https://Objective-Analysis.com/reports/#Emerging
http://www.TomCoughlin.com/techpapers.htm



Let’s Talk Energy

DRAM Refresh: A Major Culprit

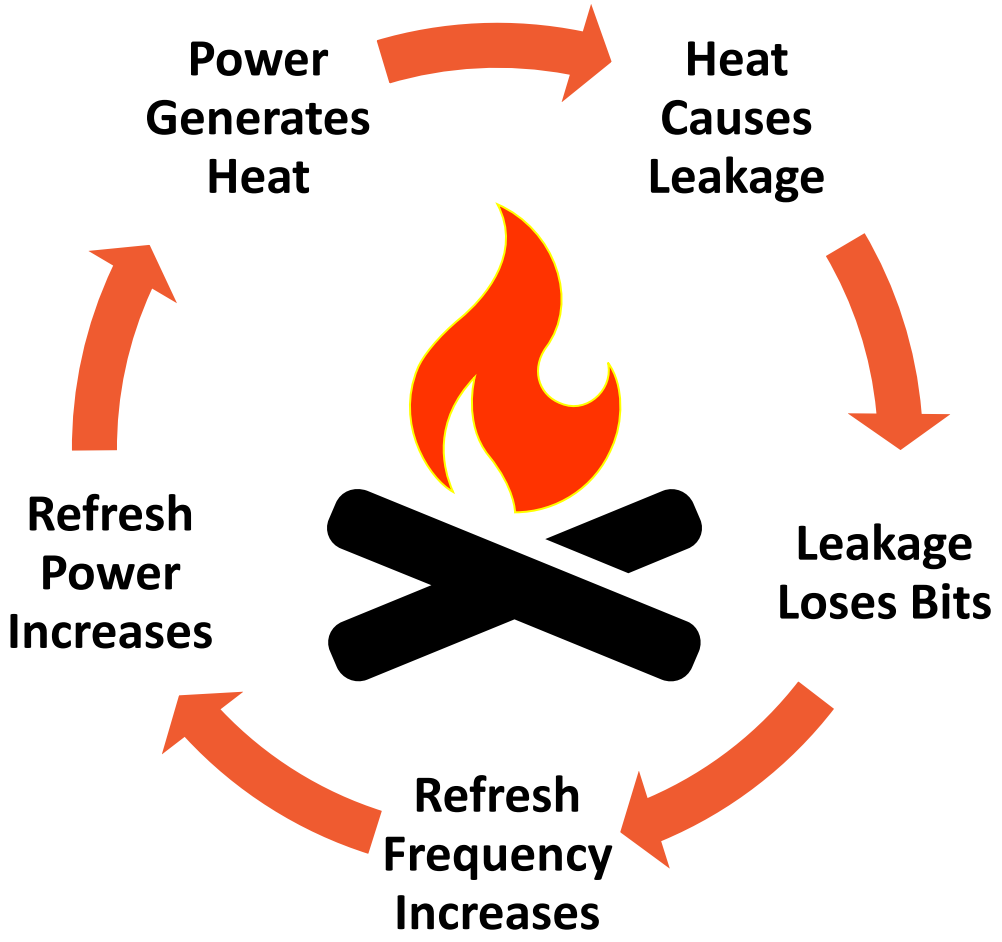


- DRAM requires more refreshing when hot
- Chart:15-55% of DRAM power goes to refresh
- DRAM accounts for up to 46% of power in a server*
- A large datacenter can consume 100MW of power
- DRAM refresh could be as much as 25MW

SOURCES:
Chart: Kim, S. *et al.* Charge-Aware DRAM Refresh Reduction with Value Transformation
https://jaehyuk-huh.github.io/papers/kim_charge_aware_hpca2020.pdf

* S. Ghose *et al.* 2018. What Your DRAM Power Models Are Not Telling You: Lessons from a Detailed Experimental Study.
In Proc. ACM Measurement and Analysis of Computing Systems
https://www.pdl.cmu.edu/PDL-FTP/associated/18sigmetrics_vampire.pdf

DRAM Refresh's Vicious Cycle



Moving Data to the Processor is Costly

One floating-point calculation

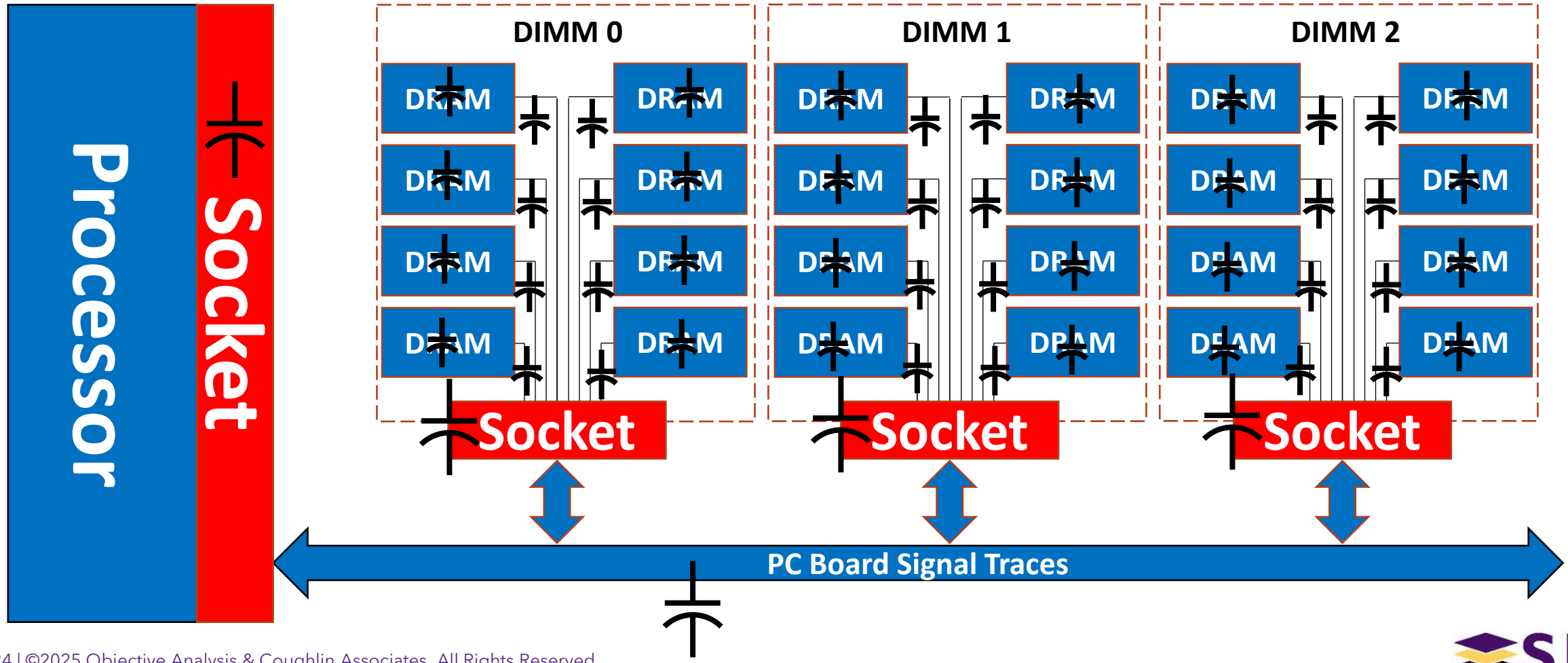
17 picojoules

Moving data from DRAM to CPU

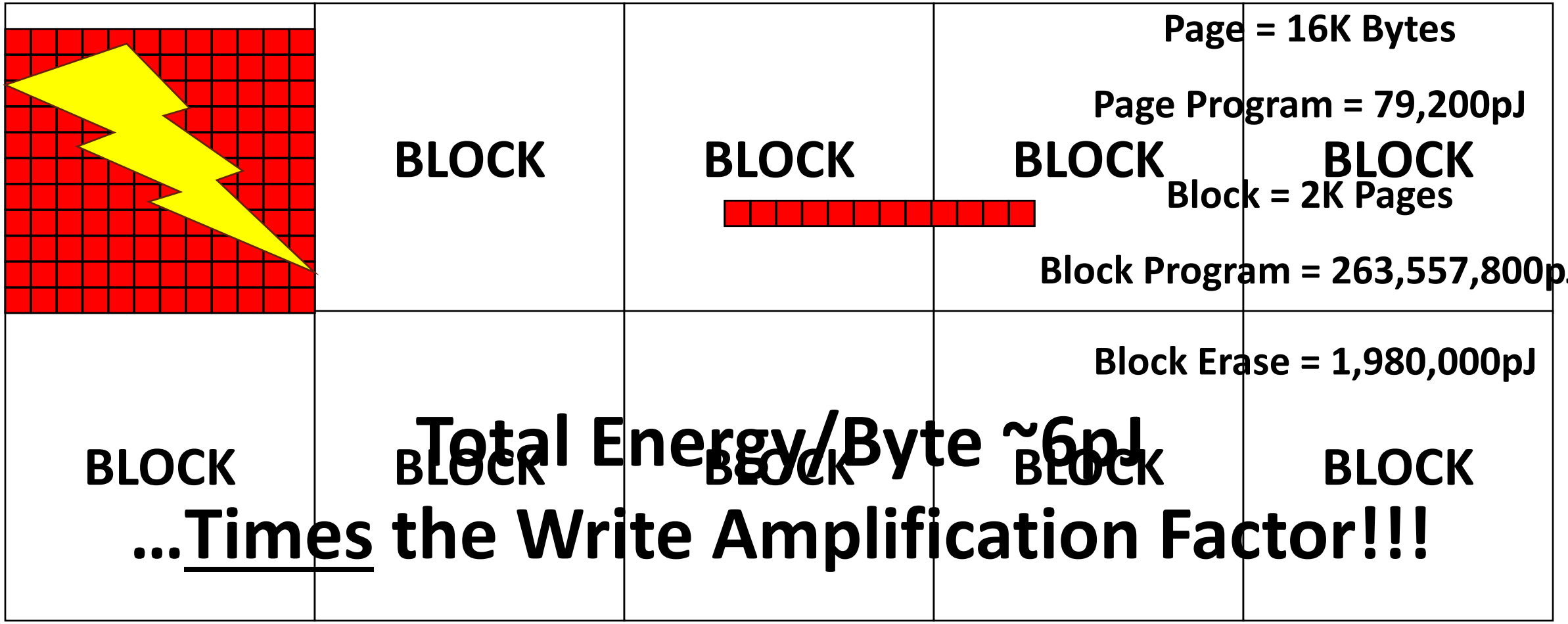
17,000 picojoules

Opportunities for **1000x improvement** are increasingly rare

To the Processor Everything Looks Like a Capacitor



Flash's Energy-Hungry Program/Erase



Some New Memories Have Lower Write Energy

Table 7. Summary of Emerging Memory Technologies

		MRAM			ReRAM	
	PCM	DRAM-like	SRAM-like	FRAM	OxRAM	Weebit
Source	STMicro	Everspin	GlobalFoundries Sony, Avalanche	Infineon/FMC Fujitsu/TI/ Panasonic	Panasonic	Weebit
Cell Type	2T2R	1T1J	2T2J	2T2C/1T1C	1T1R	1S1R/1T1R
Cell Size (f^2)	4-50	40-160	40-160	10-32	8-35	4
Stackable	No	No	No	No	No	Yes
MLC	Yes	Yes	No	No	Yes	TBD
Selector	Transistor	Transistor	Transistor	Transistor	Transistor	OTS
Materials	10+	Many	Many	2		10+
Layers	?	10+	10+	3		3-8
Masks	3-5	2-5	2-5	2	2	2-3
Current Process (nm)	20	40	22	130	130	40
Minimum (nm)	<10	<10	<10	TBD	28	<10
Status	Production	Production	Production	Production	Production	Prototypes
Impact of Scaling	Higher bit resistance	Data Retention-Endurance Ratio tightens			Worse sensing margin	
Read (pJ/bit)		1	1	0.37	66	3
Read (ns)	5-100	3-15	3-20	20-50	300	10
Write (pJ/bit)	10	<1	<0.5	0.37	150	~1
Write (ns)	300	1-15	10-100	50	850	100
Endurance	10^7	10^{15}	10^{15}	10^{15}	10^6	10^6
Retention (yr)	10	0.25	1	100	10	10
Max Temp (°C)	165	165	165	125	>150	>150

Source: Objective Analysis, 2024

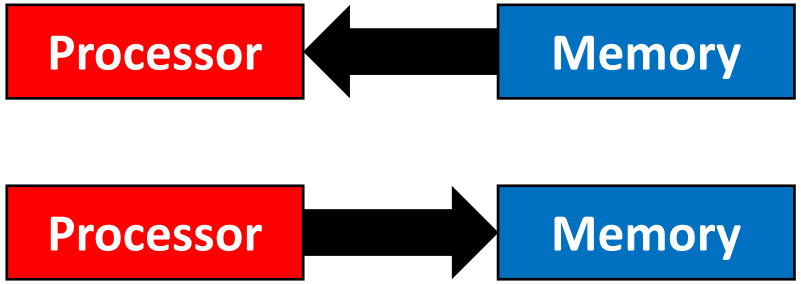
Write Energy/Byte

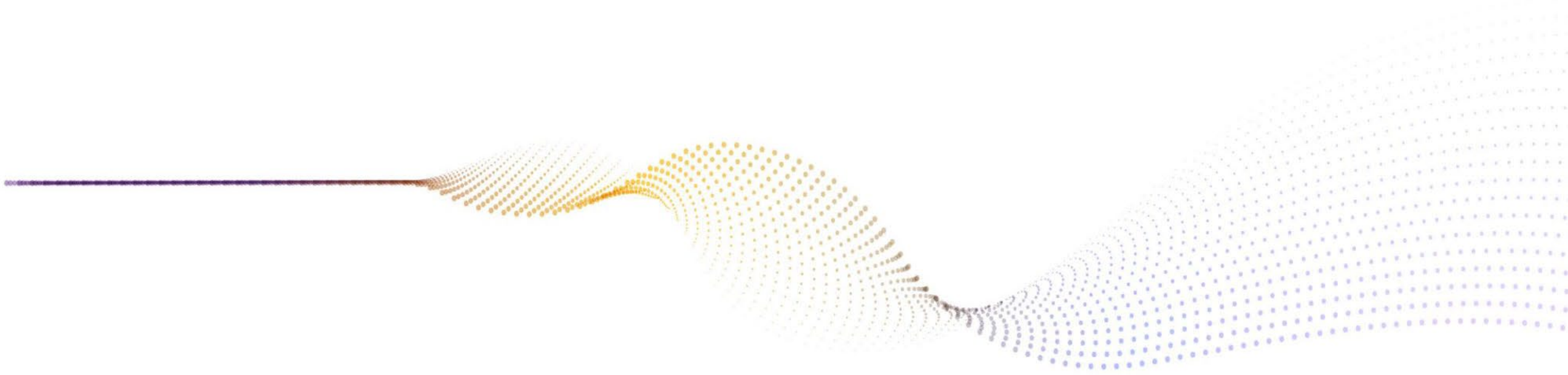
- PCM – 8[X]0pJ
- MRAM (DRAM-like) – <8pJ
- MRAM (SRAM-like) – <4pJ
- FRAM – 3pJ
- OxRAM – 1,2[X]0pJ
- Weebit – 8pJ

Write-in-Place Helps Enormously

Approaches Under Investigation

- Replace DRAM with a new memory technology
  - MRAM, ReRAM, FRAM, PCM
  - Intel's Optane proved this to be very difficult
    - But Sandisk plans to try this again
- Move memory into the processor
  - HBM is a small step in that direction
- Move processing into the memory
  - Digital compute-in-memory
  - Analog neural networks
    - New memories are good for this approach
- Support persistence in memory & cache





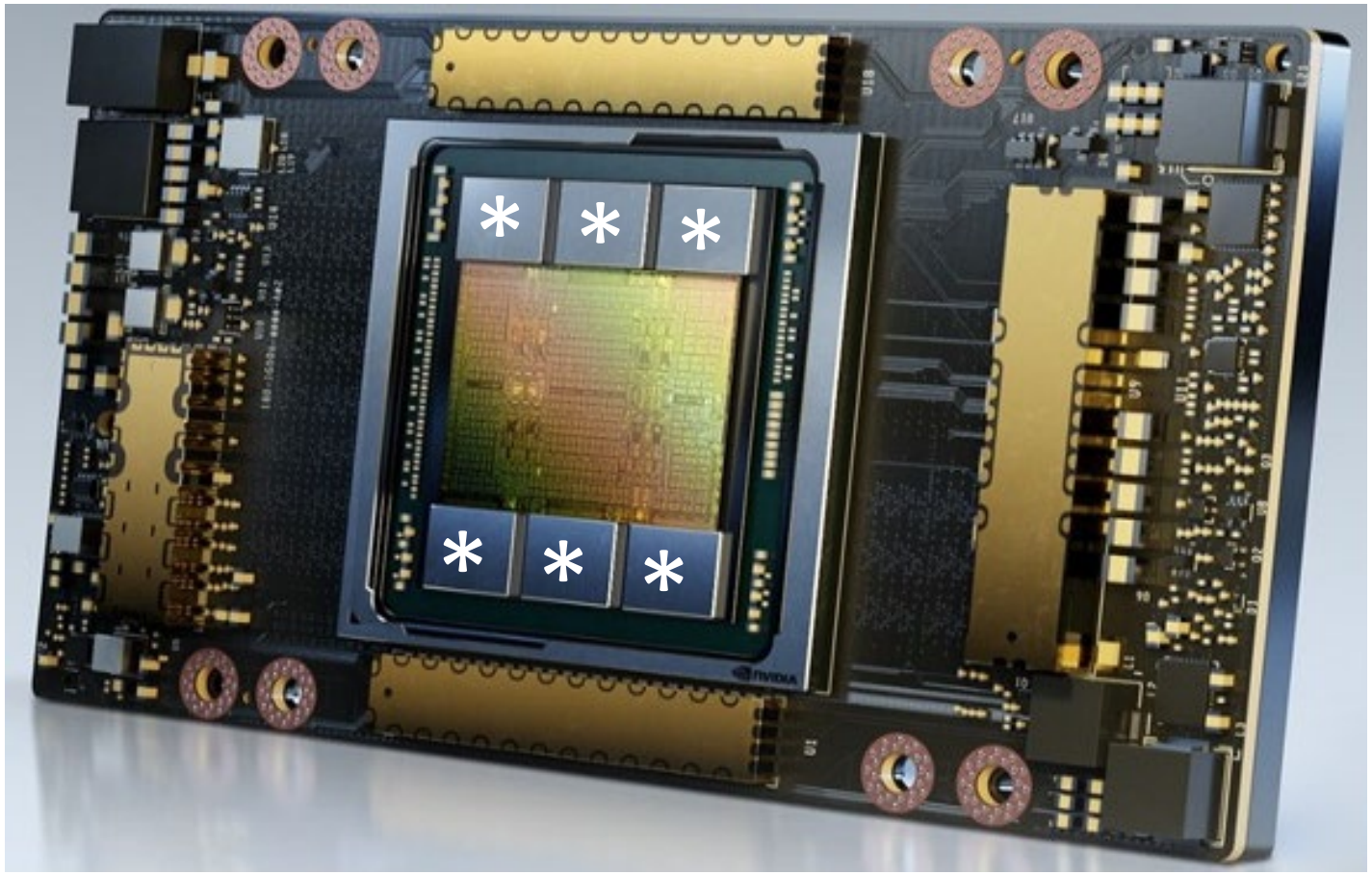
Improving Cost/Performance

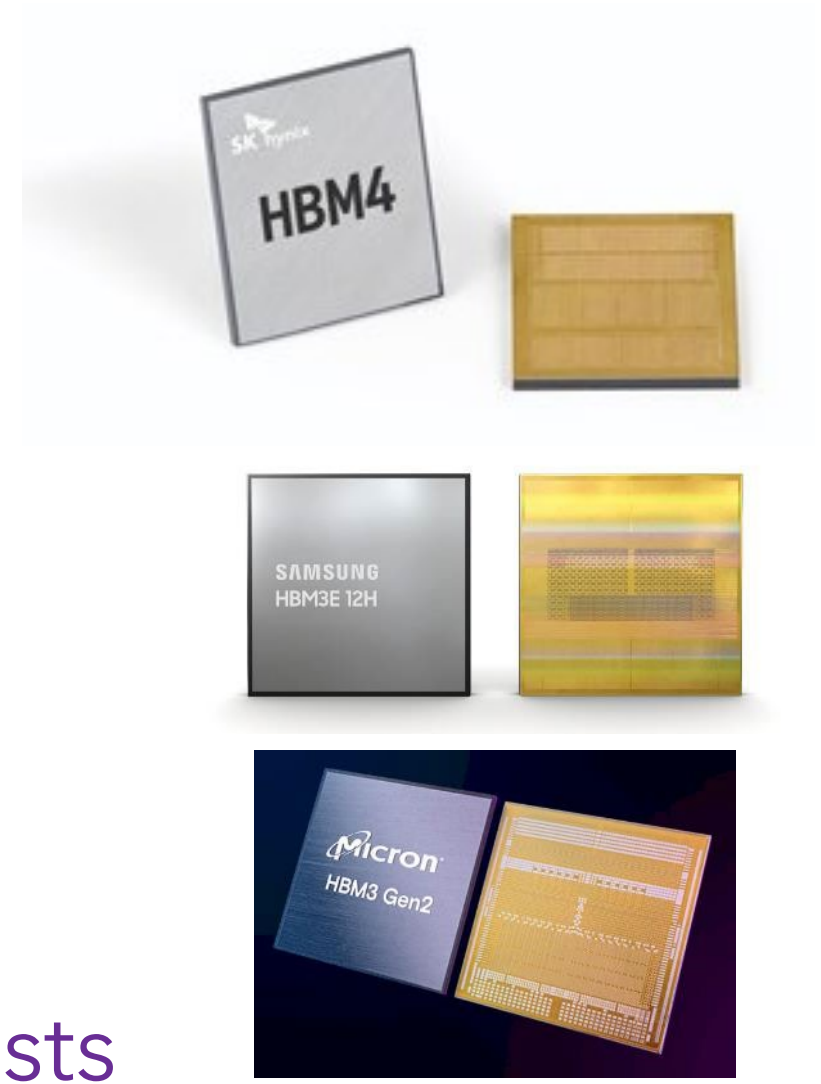
TCO is a Big Driver

- TCO = Capital Expense + Operating Costs
  - Capital Expense = Buying the system
  - Operating Costs = Energy, Cooling, Maintenance, etc.
- Energy is a large part of this equation
  - Remove the need for refresh
  - Reduce data transfer costs
  - Reduce high write energy
  - Reduce moving data to storage
- Capital expense follows die costs
  - Use new memories to improve chip scaling

Chiplets Can Bring New Memories Into the Package

HBM – An Open Standard for DRAM





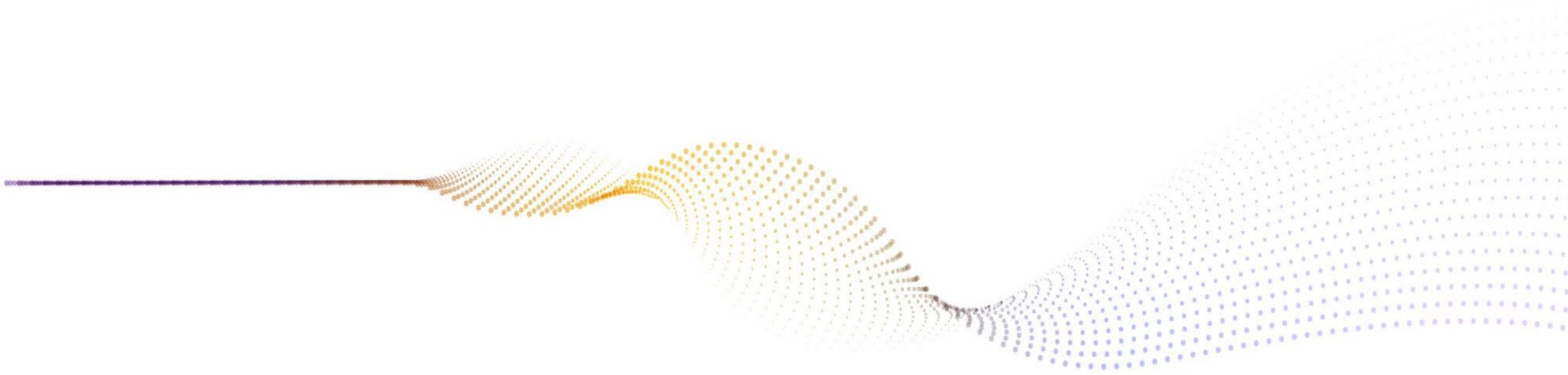
Competitive Sourcing = Lower Costs

UCIe Will Be a Useful Tool

- New open standard for chiplet interconnect
- UCIe is like CXL within the package
  - Flexible data movement
  - Hides differences in memory types
  - Allows commoditization of chiplets
- UCIe helps reduce costs
  - Logic on a logic-only process
  - Memory on a memory-only process (cheap!)
  - Optics on an optical process
  - Broadly-used commodity parts will cost much less

Summary

- Chiplets will move from large and expensive to small and cheap
- New memories will reduce energy while extending chip scaling
- New architectures will reduce data movement
- Cost/Performance improvements will continue



Thank you for attending!

Please remember to rate this session. You get access the presentations at
http://sniadeveloper.org/conference

