UQ Guest Seminar

2-Mar-2026

Next-Generation System Integration and R&D

Technology for Sustaining AI/HPC Growth

Josh Fryman, PhD

Intel Fellow

CTO Intel Government Technologies

Director of IGT R&D

An interlude . . . innovation is . . .

Definitions ..

- *Innovation*
  1. the action or process of innovating.
     "innovation is crucial to the continuing success of any organization"
  2. a new method, idea, product, etc.
     "technological innovations designed to save energy"

- *Learned Helplessness*
  1. a condition in which a person suffers from a sense of powerlessness, arising from a traumatic event or persistent failure to succeed. It is thought to be one of the underlying causes of depression.

"an elephant never forgets.." ← what?!

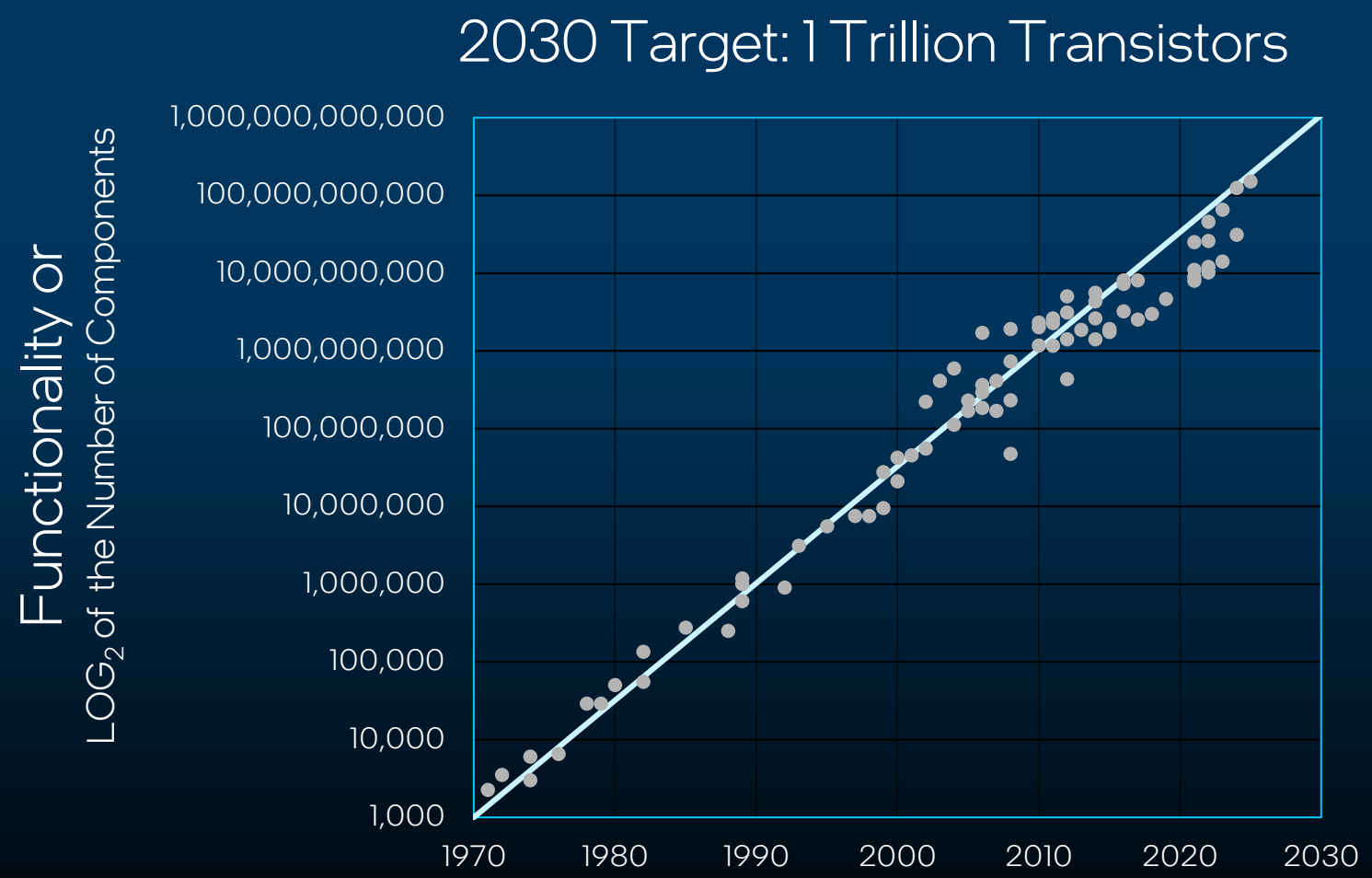
Innovation and (sat on by) an Elephant?

- *Education: reward or punish*
- *Title: manager knows best*
- *Seniority: must get permission*
- *Territory: I must own on my idea*
- *Praise: staying within the box*

- *Don't ask, don't tell, be quiet*
- *Just do what I'm told*
- *Giving up entirely*
- *Individuals don't scale*
- *Stop looking outside the box*

. . . and back to our show

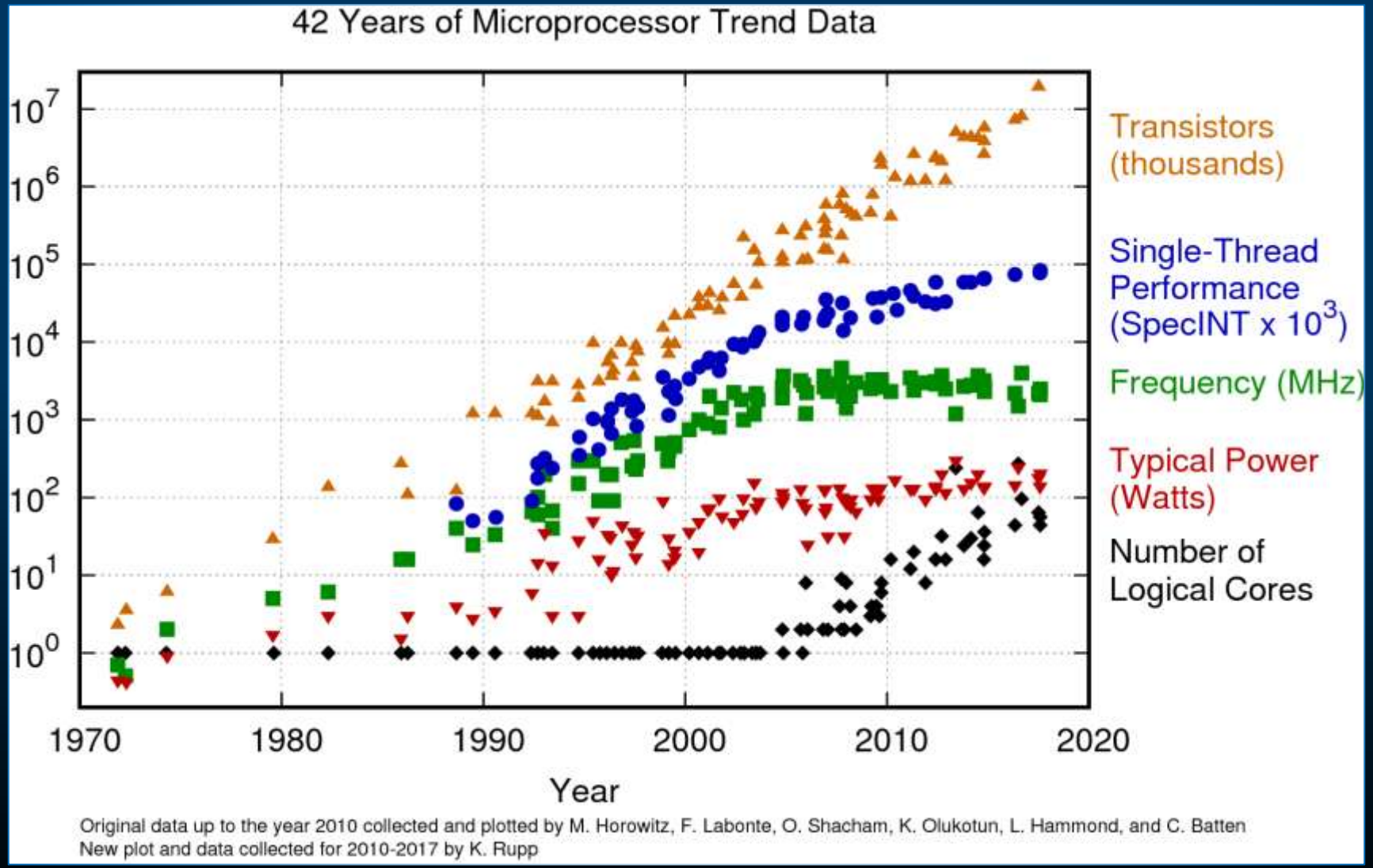
What is the industry doing? Pushing on **hard** limits



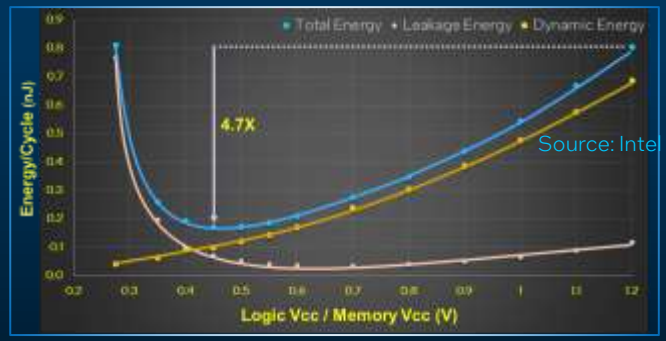
- **Enable multi-core CPU & XPU**
- **Enable Accelerators**
- **Lower Latency & Energy**
- **Increase Memory Capacity & Bandwidth**

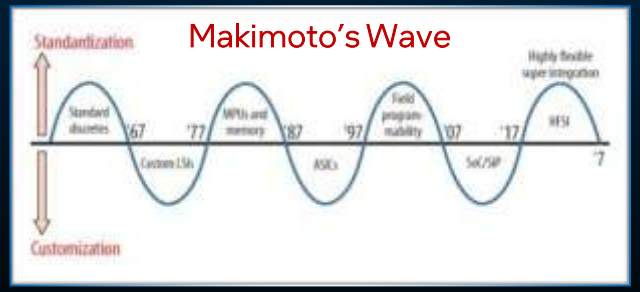
Image credit: Intel Corporation

But compute performance hit other limits



https://www.nextplatform.com/2019/06/18/dennard-scaling-demise-puts-permanent-dent-in-supercomputing/





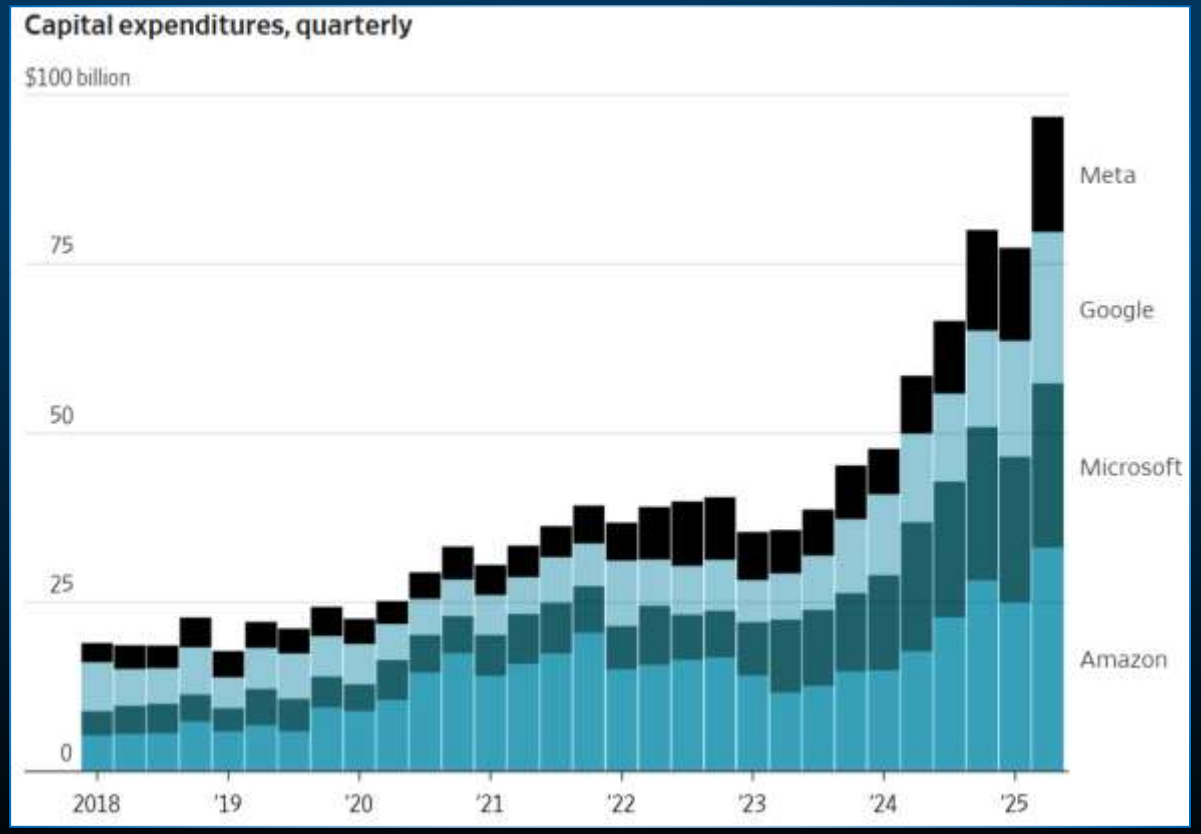
https://semiengineering.com/knowledge_centers/standards-laws/laws/makimotos-wave/

“Planet-scale AI infrastructure” shifted the ecosystem

Silicon Valley’s New Strategy: Move Slow and Build Things

Big tech companies are becoming infrastructure companies—just like the steel and railroad giants of old

https://www.wsj.com/tech/ai/silicon-valley-ai-infrastructure-capex



Budgets for building	%
HPC only	3.8%
AI only	10.2%
Blended	86%

HPC-only segment continues a steady decline YoY

Source: Intersect360 Research

TECH STOCKS

‘It’s existential’: How Big Tech found itself in a $650 billion

For hyperscalers like Amazon, how much capex is too much?

Microsoft Shares Dive as Data-Center Spending Overshadows Earnings Surge

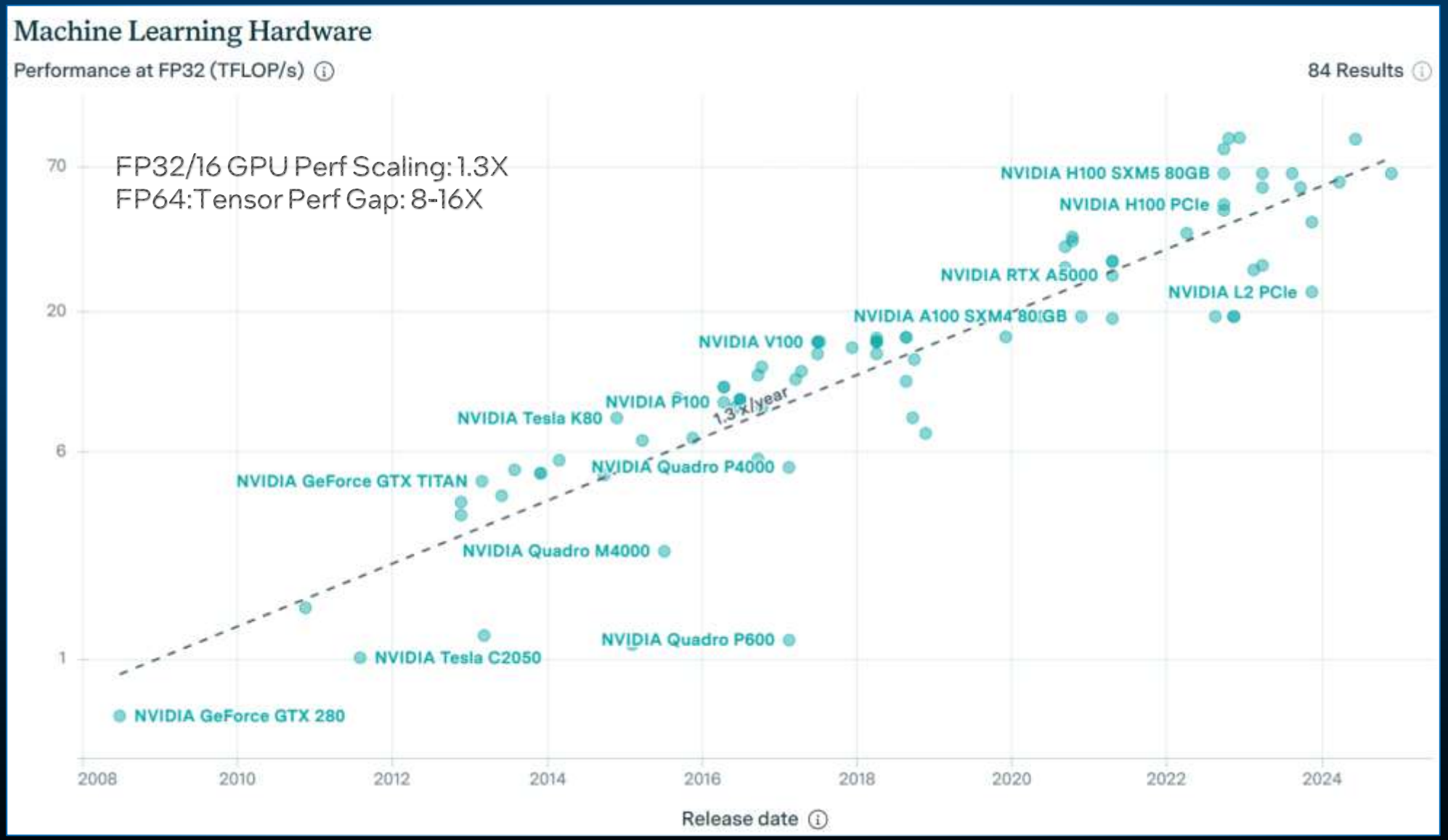
Tech giant reports higher-than-anticipated capital spending and slower growth in cloud computing

https://www.wsj.com/tech/ai/microsofts-earnings-surge-elevated-by-cloud-business-251829c2

By Sebastian Herrera Follow

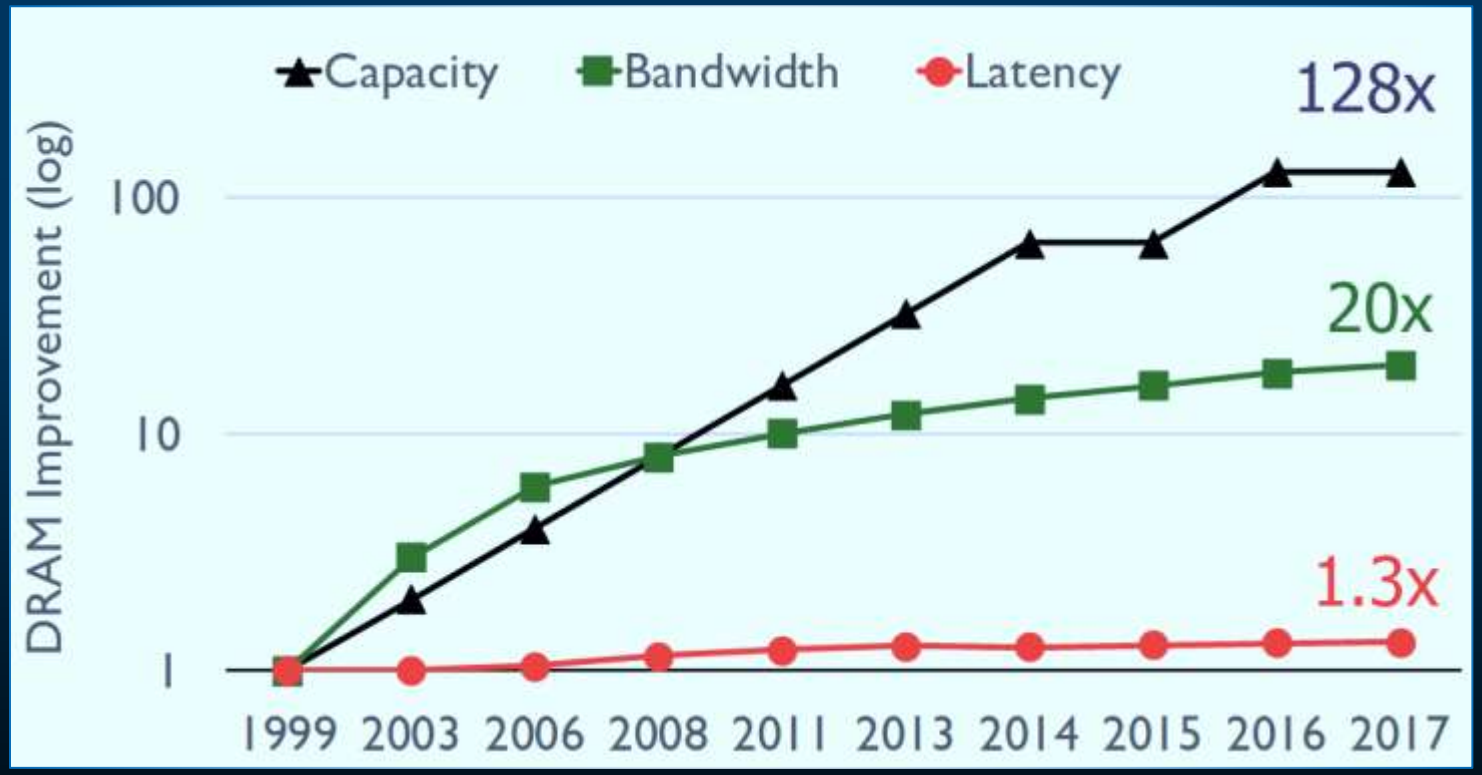
Updated Jan. 29, 2026 9:59 am ET

Let's look closer at a fundamental bottleneck



https://epoch.ai/data/machine-learning-hardware

Memory Wall: Why there's really a cliff



Source: "A modern primer on processing in memory," by Onur Mutlu et al, arxiv.org, Dec 2020.

Historical Prices

DDR5 memory ~ $5/GB

LPDDR5 ~ $3/GB

HBM memory ~ $10-25/GB

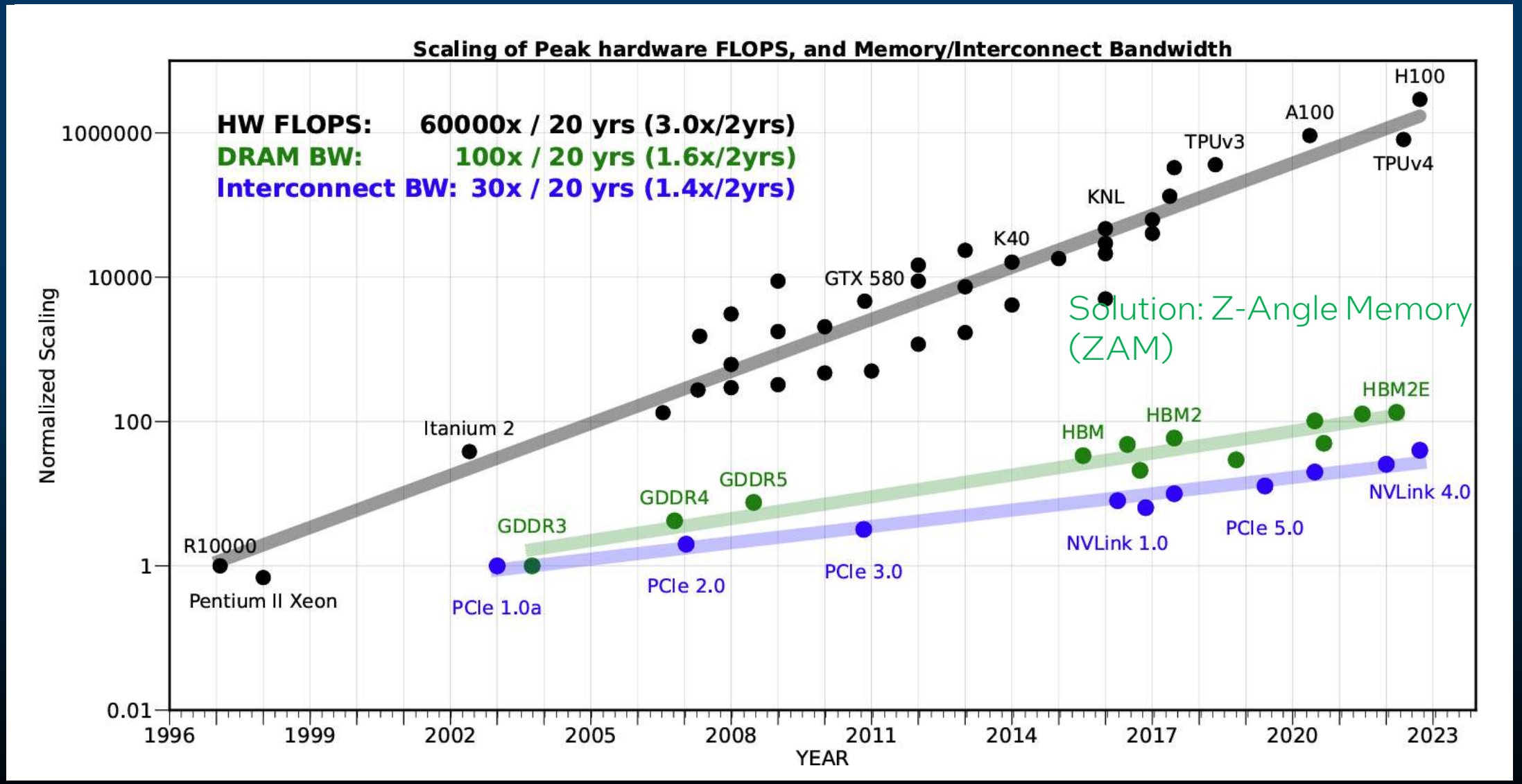
Capacity:BW scaling is off by >6x

× ~~2~~ ~~3~~ ~~4~~ 5?

Customers buy for bandwidth, but pay for (unused) capacity – creating a TCO problem at a global scale

Volume customers drive market solutions, dragging everyone along for the ride – whether or not they have the same problems

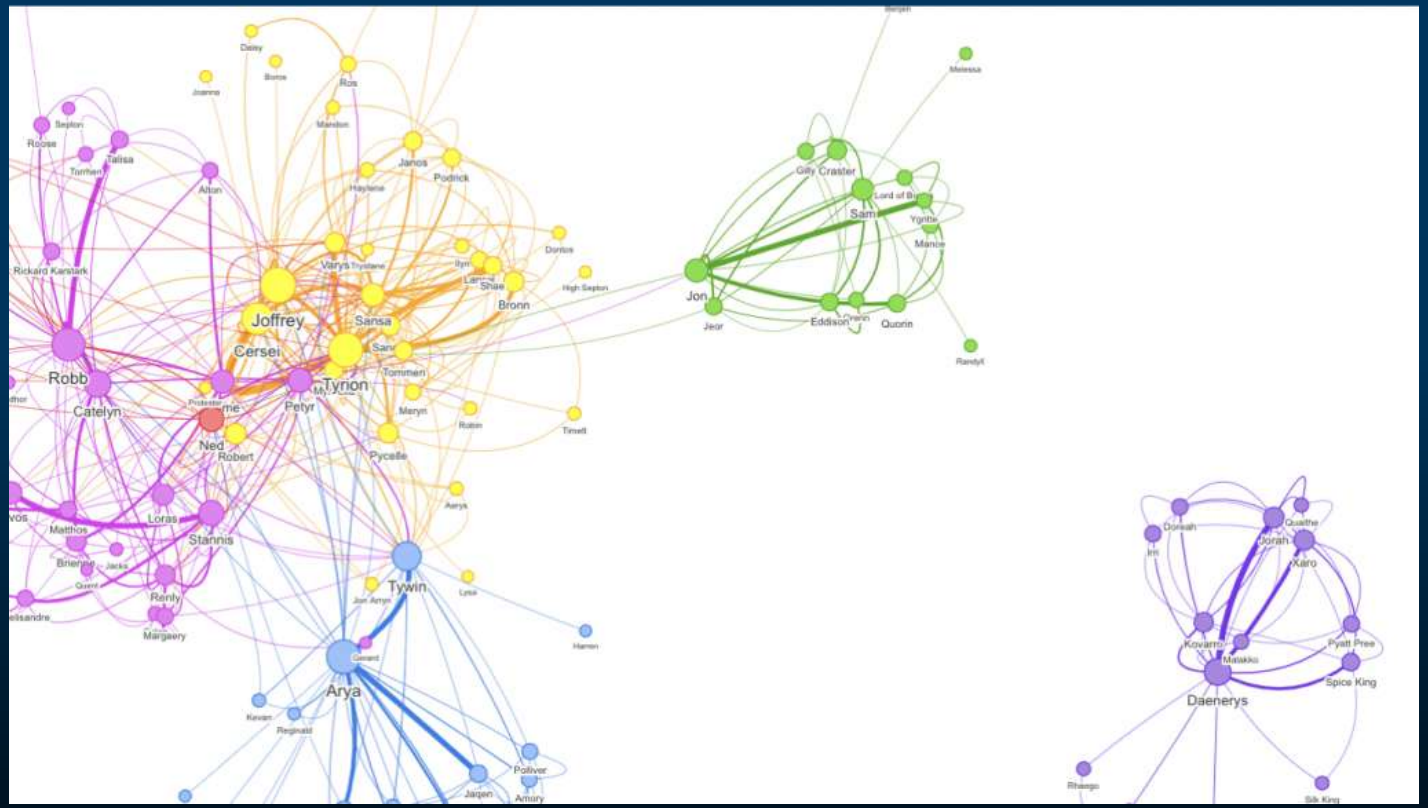
Behind the Memory Wall . . . looking closer



An interlude . . . DRAM R&D . . .

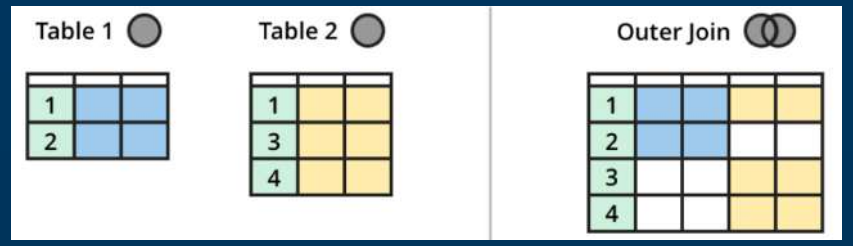
Step 1: Analyze the *real* problem

Community Detection in *Game of Thrones*



https://opendatascience.com/why-we-need-graph-analytics-for-real-world-predictions/

Why Graph Analytics? (Recognize the Relationship)

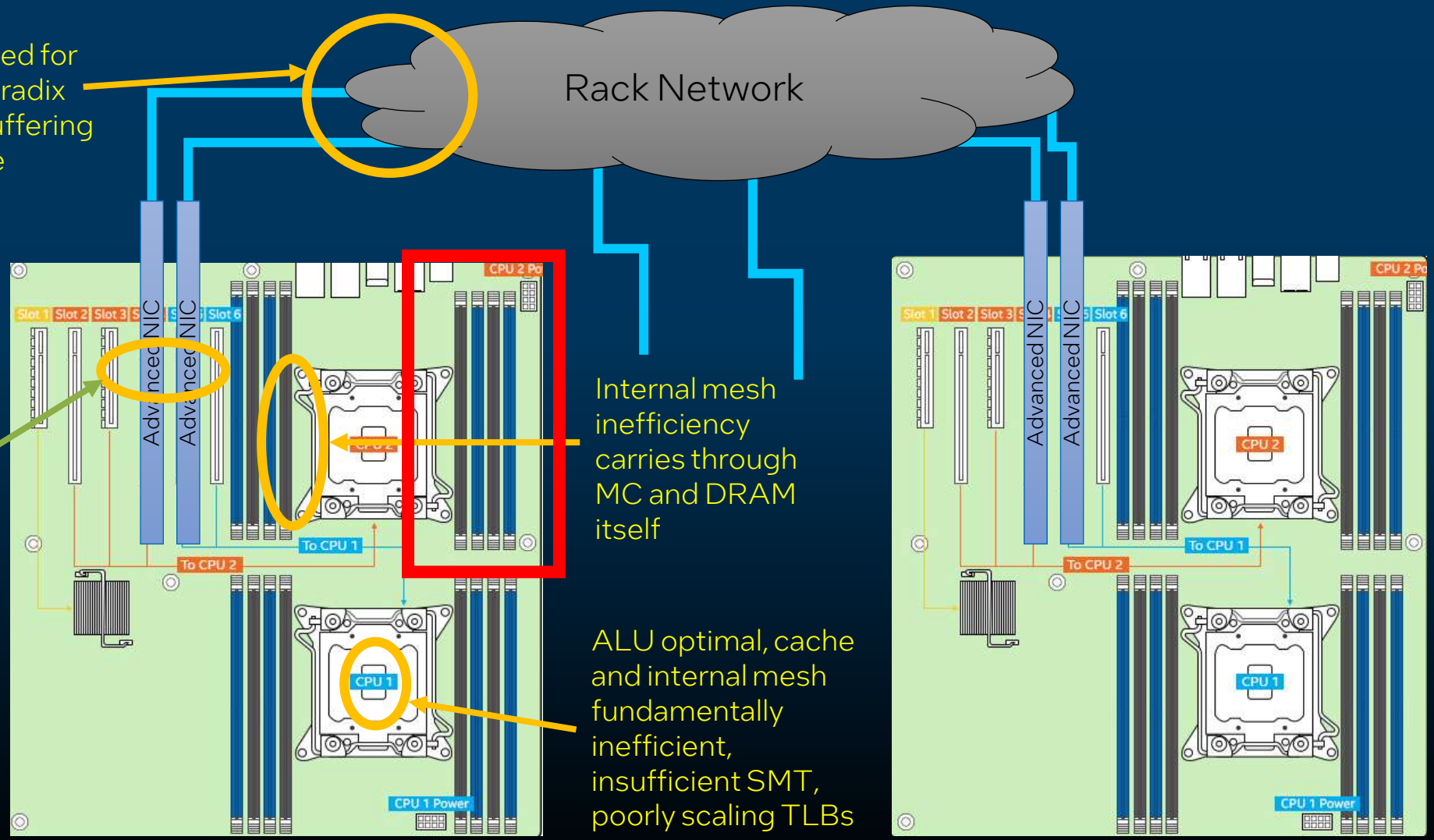


Data Traversal is the first class citizen

- T-TEPS (Traversed Edges Per Second) as key
- *Compute:* Modest to Low
- *Memory & Cache:* Large & Granular
- *Network & IO:* Extreme
- *Access Pattern:* Arbitrary

Step 2: Analyze Limits

Infrastructure designed for big messages on low radix topology, protocol buffering and hierarchy throttle performance

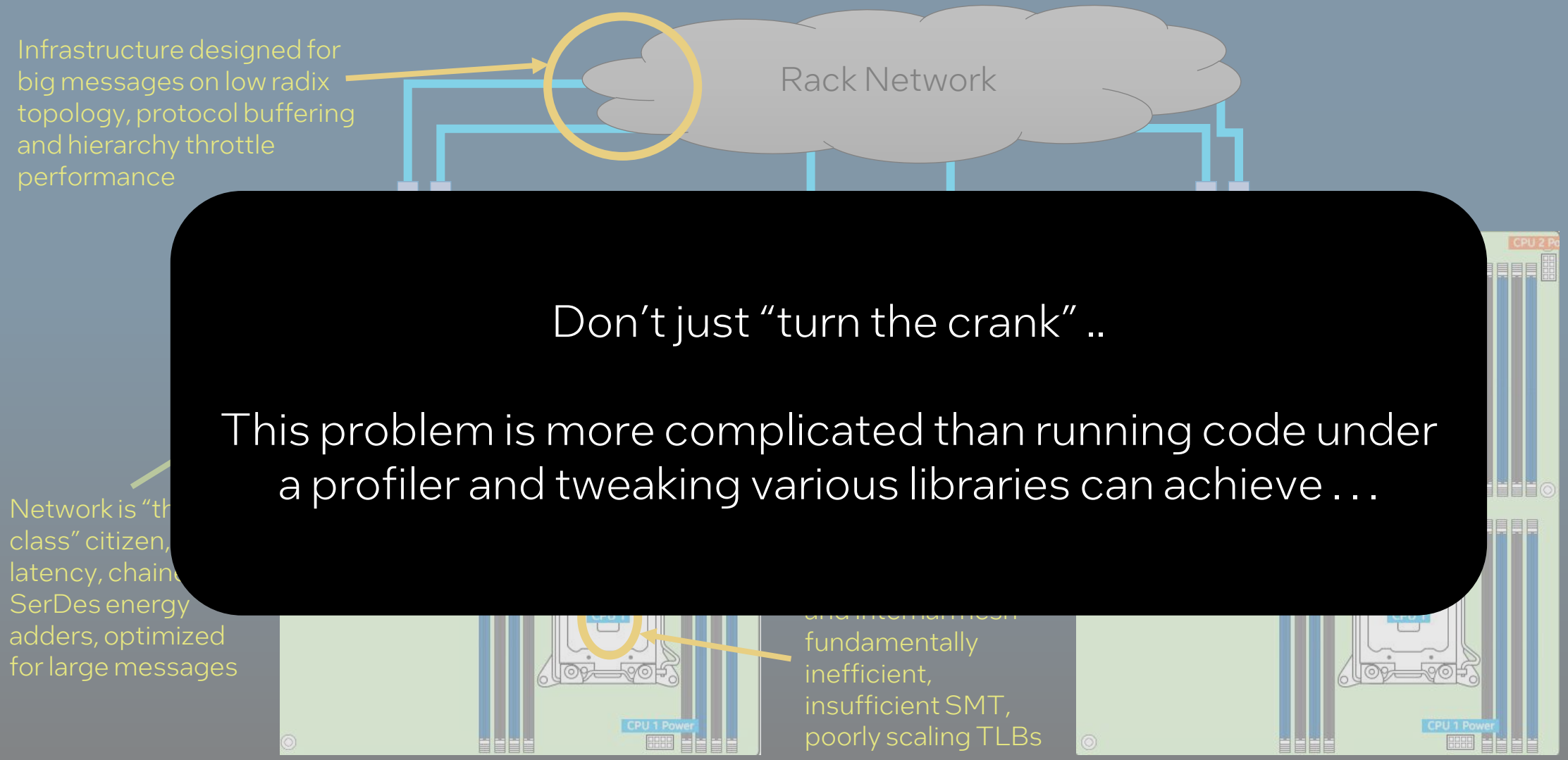


Network is "third class" citizen, high latency, chained SerDes energy adders, optimized for large messages

Internal mesh inefficiency carries through MC and DRAM itself

ALU optimal, cache and internal mesh fundamentally inefficient, insufficient SMT, poorly scaling TLBs

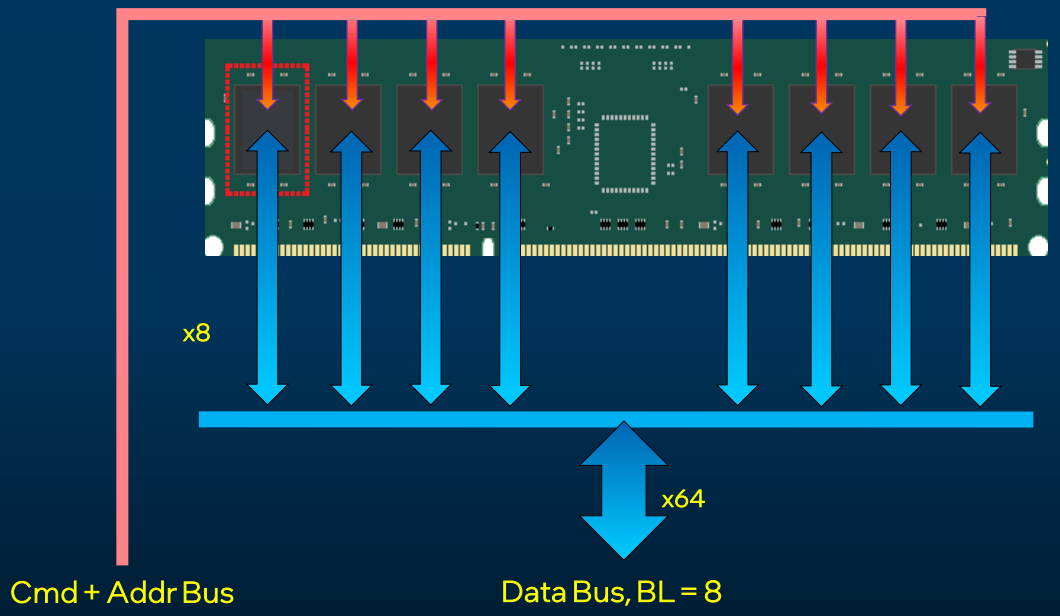
Step 2: Analyze Limits



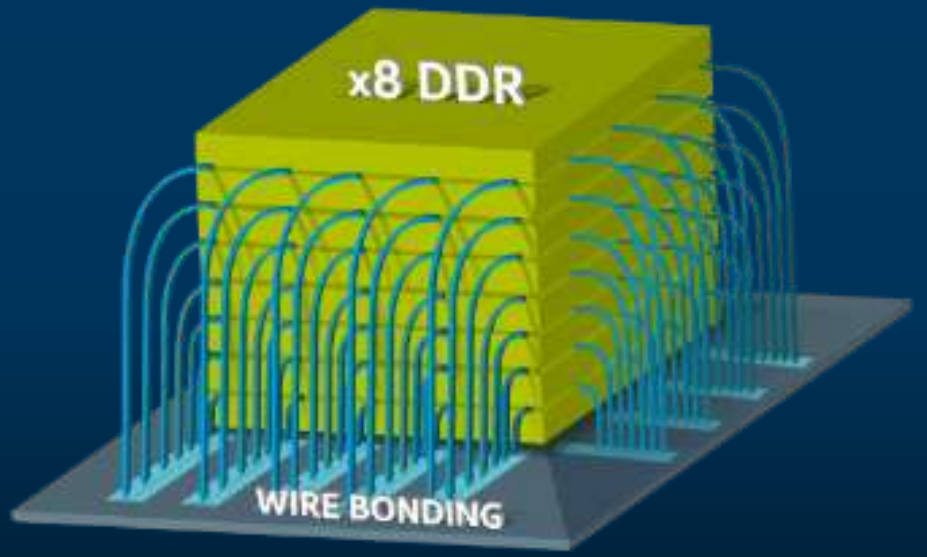
Don't just "turn the crank" ..

This problem is more complicated than running code under a profiler and tweaking various libraries can achieve . . .

Step 3: Novelty Awareness



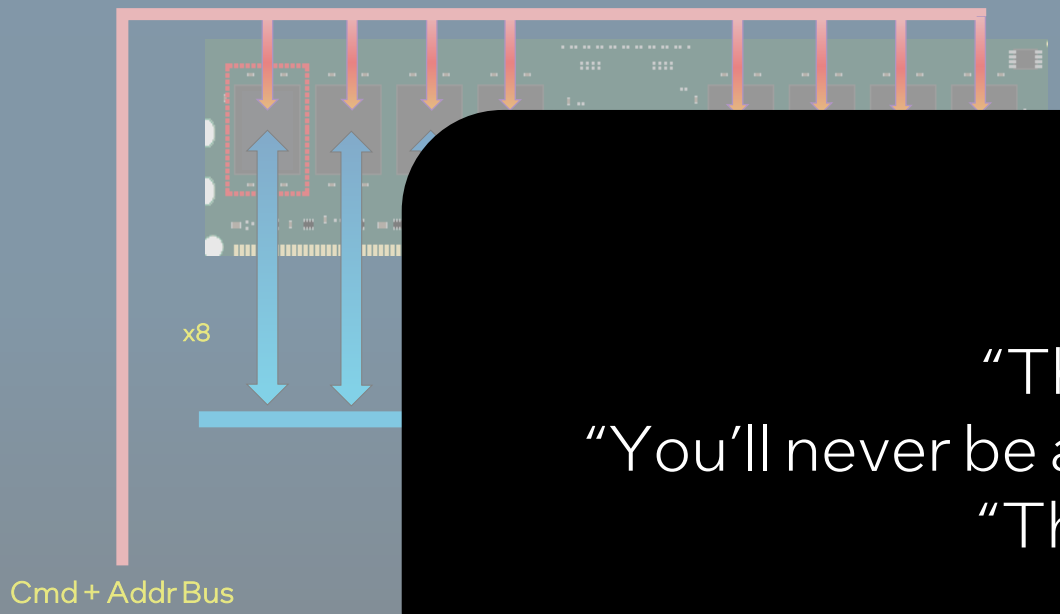
12.5% useful BW
(87.5% wasted power)

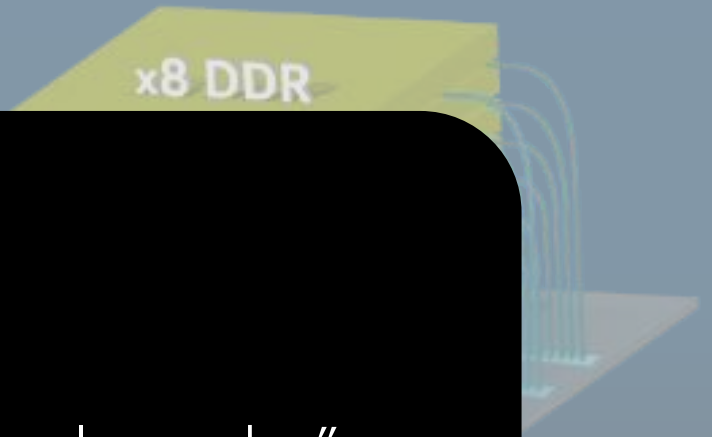


New commodity DRAM package, individual die addressing:

- Optimal for 8B access efficiency
- Power optimal
- 8x parallelism, same aggregate BW

Step 3: Novelty Awareness





"You're crazy.."
"This will never work.."
"You'll never be able to make the thermals work .."
"This will never yield .."

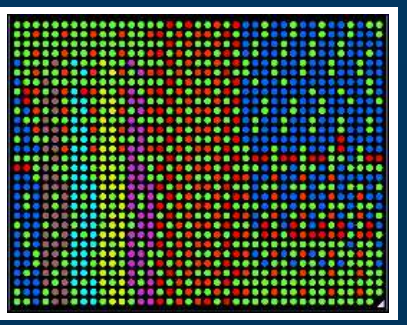
12.5% useful BW
(87.5% wasted power)

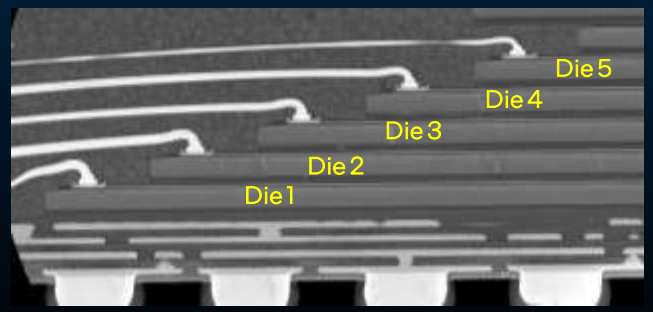
kage,

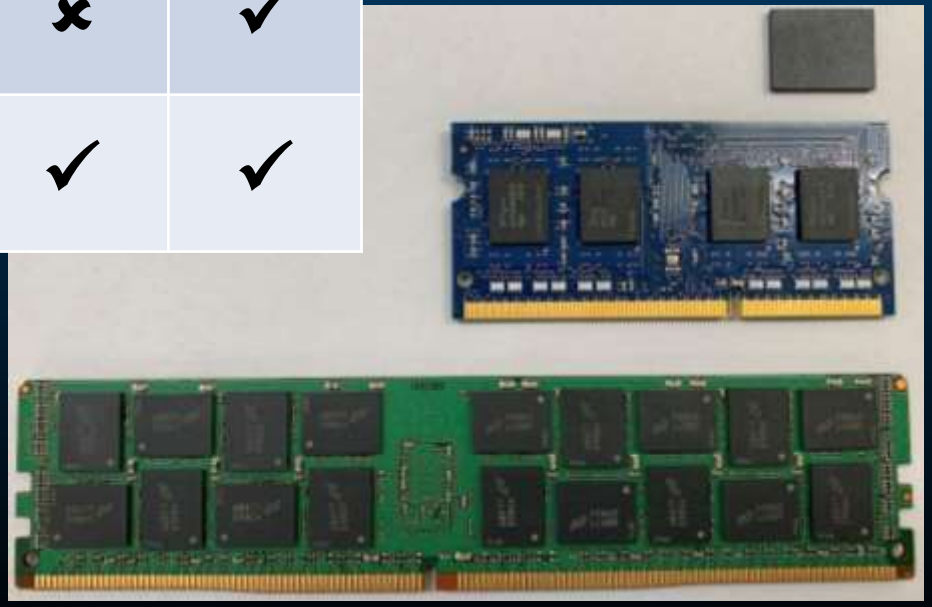
- Optimal for 8B access efficiency
- Power optimal
- 8x parallelism, same aggregate BW

Step 4: Do it! And prove it!

Tech	GB	$/GB	Raw GB/s	8B Read		64B Read		IPM Ok?	Ext Ok?
				pJ/b	Eff. BW	pJ/b	Eff BW		
HBM2	8	15	256	31	32	5.8	256	✓	✗
DDR4 DIMM	64	4	25.6	99	3.2	10.3	25.6	✗	✓
Stacked DDR4	20	<5	32	10.3	32	5.8	32	✓	✓

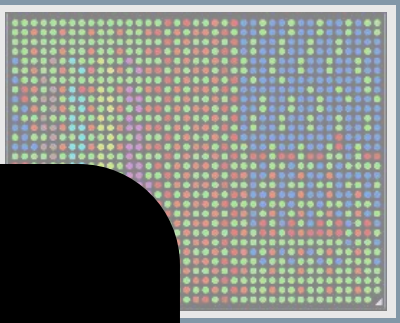






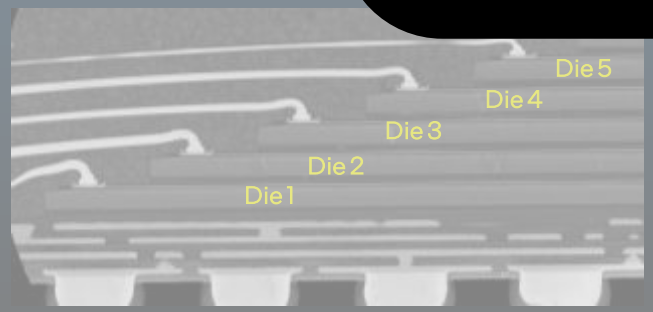
Step 4: Do it! And prove it!

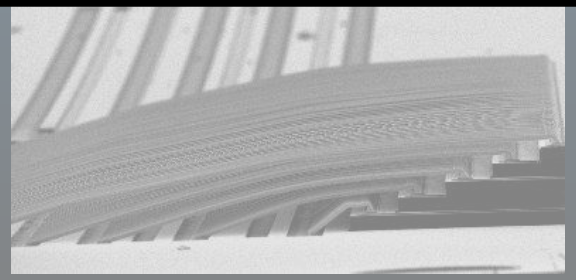
Tech	GB	$/GB	Raw GB/s	8B Read		64B Read		IPM Ok?	Ext Ok?
					Eff.		Eff		
HBM2	8								
DDR4 DIMM	64								
Stacked DDR4	20								

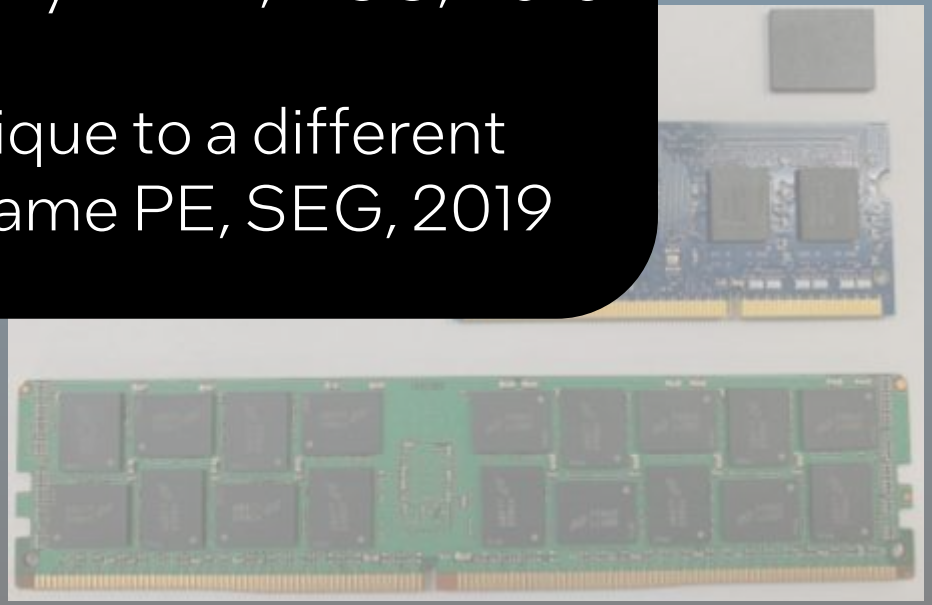


"If I had to consider this based on the data sheets, complexity, and analysis before even thinking of a simulation, I would have killed it instantly." – PE, DCG, 2016

"Can you help me apply this technique to a different memory technology right now?" – same PE, SEG, 2019

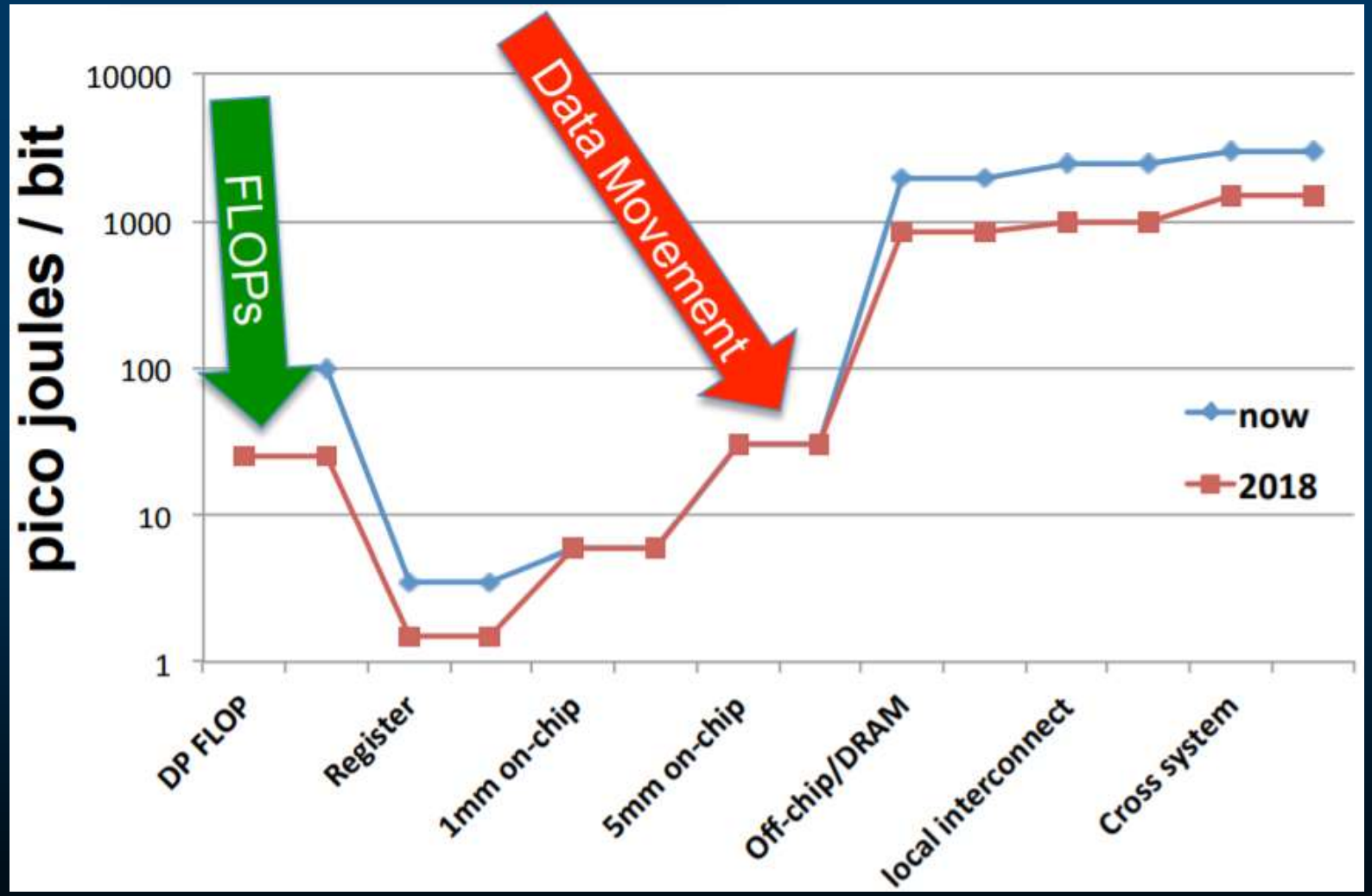






. . . and back to our show

The Energy Cost of Data movement



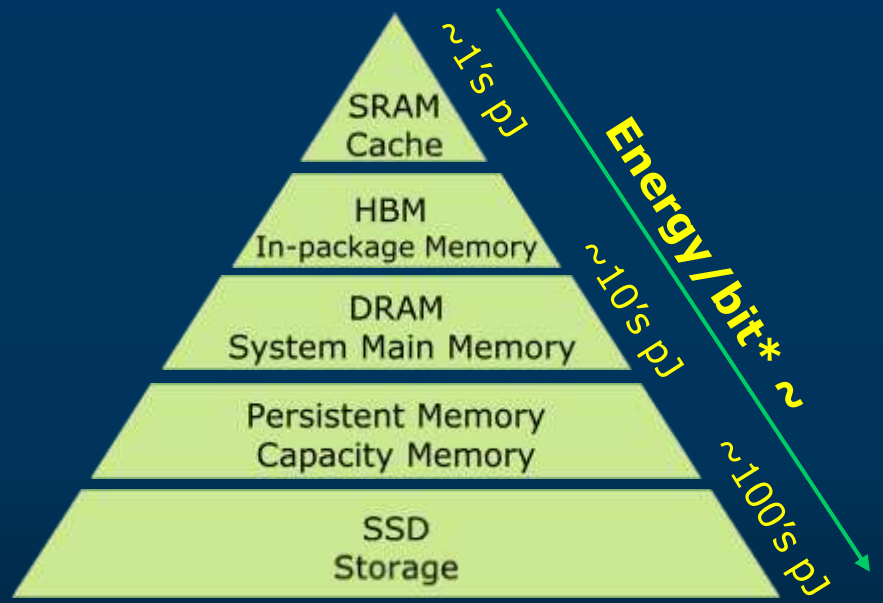
2014 projection of 2018 energy per bit moved*
(by Horst Simon, adapted from John Shalf)
* H. D. Simon, "Why we need Exascale and why we won't get there by 2020," 2014.

"Ten Lessons from Three Generations Shaped Google's TPUv4i"

Operation		Picojoules per Operation		
		45 nm	7 nm	45 / 7
+	Int 8	0.03	0.007	4.3
	Int 32	0.1	0.03	3.3
	BFloat 16	--	0.11	--
	IEEE FP 16	0.4	0.16	2.5
	IEEE FP 32	0.9	0.38	2.4
×	Int 8	0.2	0.07	2.9
	Int 32	3.1	1.48	2.1
	BFloat 16	--	0.21	--
	IEEE FP 16	1.1	0.34	3.2
	IEEE FP 32	3.7	1.31	2.8
SRAM	8 KB SRAM	10	7.5	1.3
	32 KB SRAM	20	8.5	2.4
	1 MB SRAM[1]	100	14	7.1
GeoMean[1]		--	--	2.6
DRAM		Circa 45 nm	Circa 7 nm	
	DDR3/4	1300[2]	1300[2]	1.0
	HBM2	--	250-450[2]	--
	GDDR6	--	350-480[2]	--

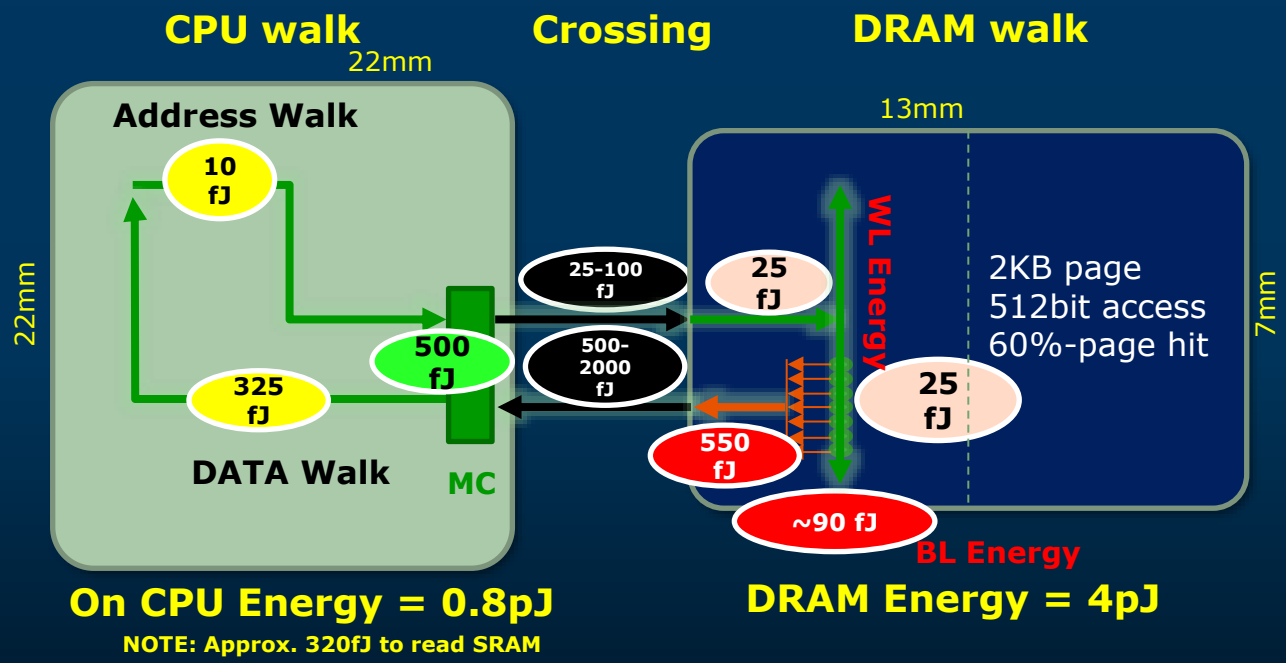
Jouppi et al, ISCA 2021

Decreased Energy – Fundamental



Data movement energy*

- Power limits compute rate
- Not moving data saves energy
- Fundamental opportunity for NMC



Energy consumption**, DRAM example

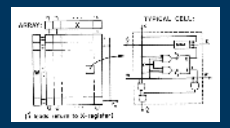
* Intel & Shen, Meng, "Silicon Photonics for Extreme Scale Systems", Journal of Lightwave Technology, 2019
** Borkar, Fryman, "HPC Memory Subsystem – Beyond Myths & Hype", ISC, 2016

Compute 'In' Memory: Long History of Research

Data-bit with Cryotron flip-flop

1960

Seeber Self-sorting memory



Cellular sorting array

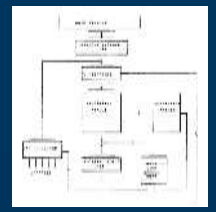
1969

Kautz Cellular logic-in-Memory Arrays

Cache logic-in-memory with sector add

1970

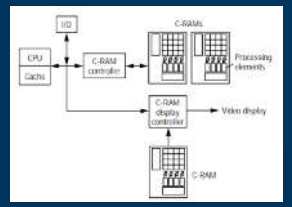
Stone Logic-in-Memory computer



BLIMP config

1972

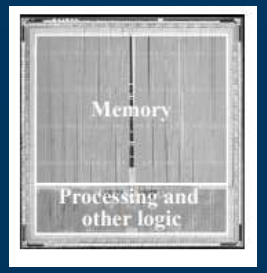
Brookhaven Logic-In-Memory Processor (BLIMP)



Add logic within DRAM

1981-1997

NON-VON, Computational RAM, EXECUBE, Terasys, IRAM



Substrates that Tightly integrate logic within DRAM

1998-2002

Active Pages, FlexRAM, Smart Memories, DIVA

HBM-PIM

Persistent Memory

Network Cube (HMC) Vault

Tesseract architecture

2002-now

3D-stacked memory with TSVs, HBM, HMC, PCM, MRAM, RRAM, byte addressable NVM, persistent memory

* S.Ghose, et.al., "Processing-in-Memory: A workload-driven perspective", IBM Journal of Research and Development, Nov.-Dec. 2019

ISSCC 2022 - Forum 1: We've rethought our commute; can we rethink our data's commute?
Credit: Frank Hady, Intel

It always comes down to .. software

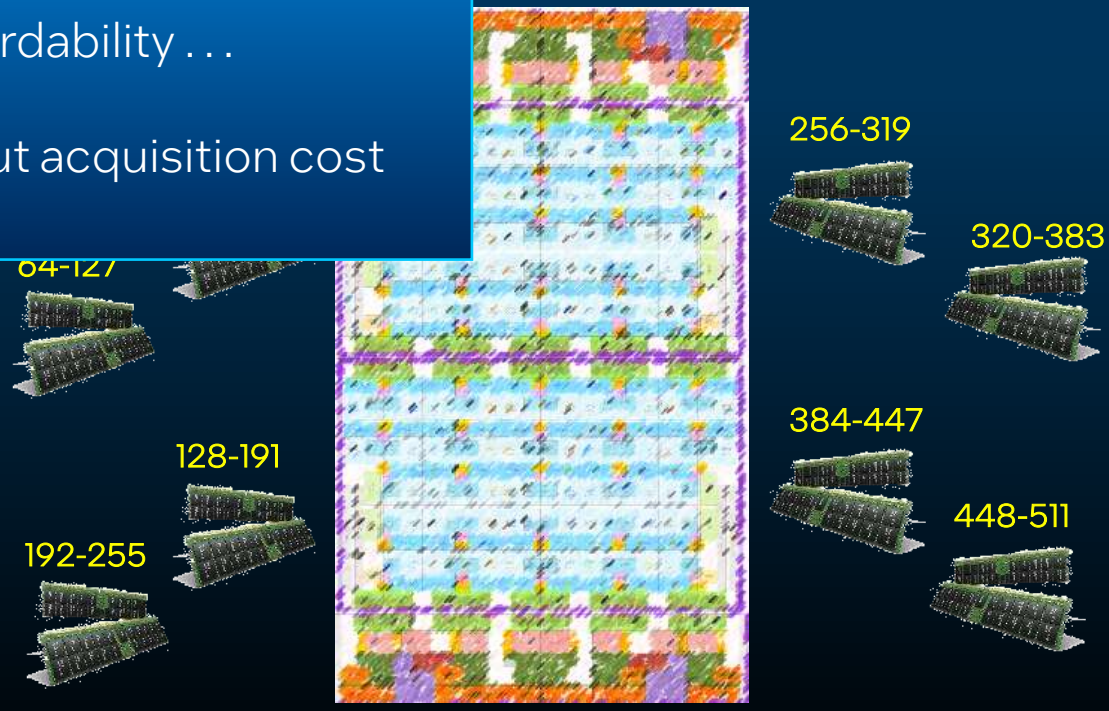
```
struct data_element {
    struct data_element * prev;
    struct data_element * next;
    uint64_t data_size;
    void * data_ptr;
    struct semaphore_t lock;
};

struct queue {
    struct queue_element * hea
    struct queue_element * tai
    uint64_t entries;
    struct semaphore_t lock;
};
```

40+ bytes

A story of affordability . . .

. . . is not actually about acquisition cost

Address Interleaving
Address Translation
Active Messaging (MC peer)
atastructure Layout

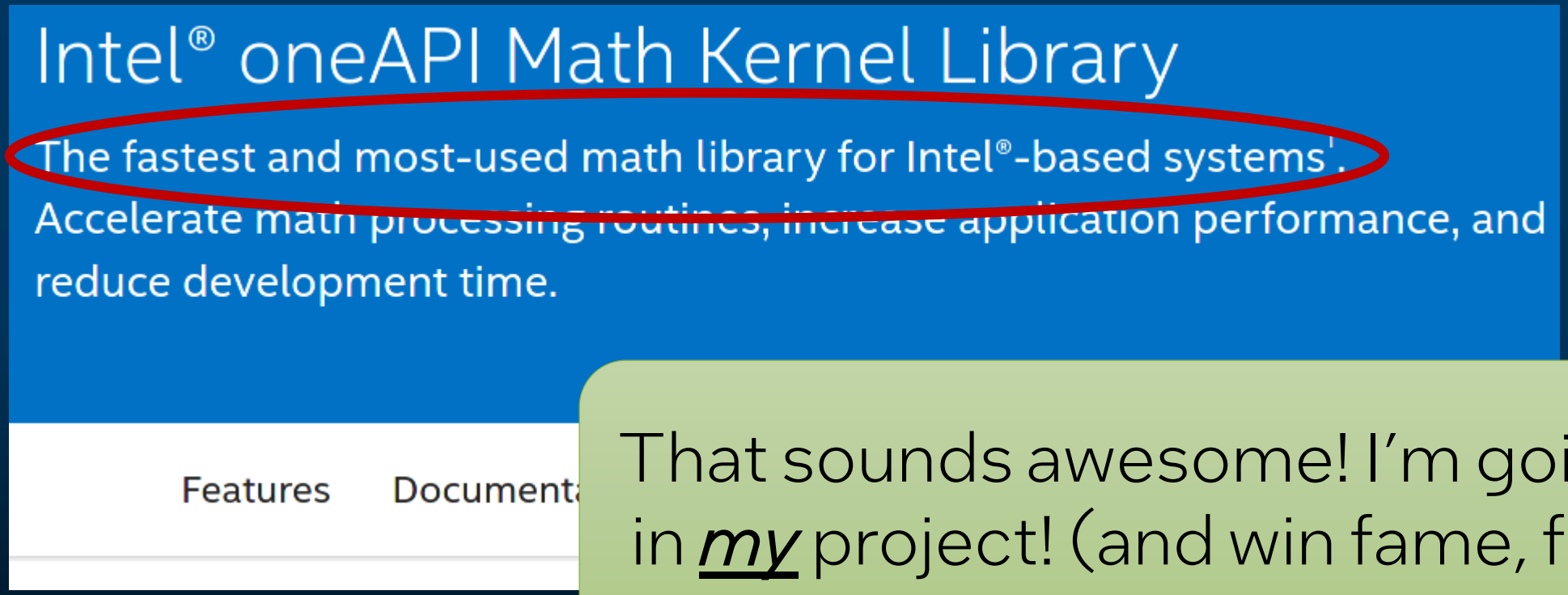


Clever programmers optimize their datastructures and code to account for concurrent programming, access alignment, cache structure, and memory organization . . .

Source: Intel Corporation

An interlude . . . Software R&D . . .

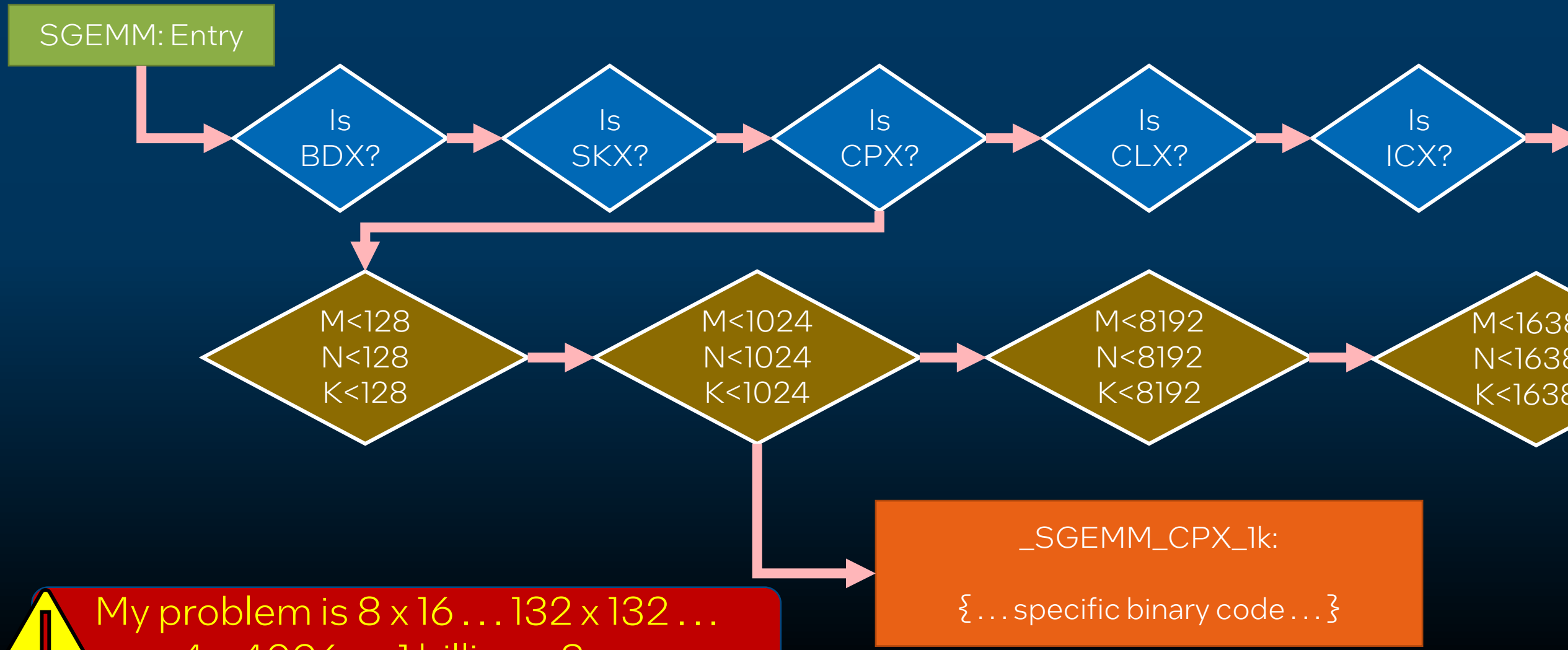
Step 1: Recognize *a* problem



That sounds awesome! I'm going to use it in ***my*** project! (and win fame, fortune, . . .)

Uh-oh . . . why is MKL *slower* than a naïve textbook GEMM implementation on ***my*** dataset?

Step 2: Analyze Limits (simplified example)



My problem is 8 x 16 . . . 132 x 132 . . .
4 x 4096 . . . 1 billion x 8 . . .

Step 3: Novelty Awareness

Dynamic pathing ("fat binaries") is a clever idea but starts to have complex drawbacks in overhead for excessive conditionals

To reduce overheads, granular chunks based on historical workloads divide segments (M, N sizing) and algorithms

Can this be done better via a DSL and JIT at run-time with a dynamic binary cache? This is another well-known technique in instrumentation tools!

```
switch (architecture) {
    case BDX: _sgemm_bdx( argc, argv );
              break;
    case SKX: _sgemm_skx( argc, argv );
              break;
    . . .
}

if ((M <= 128) && (N <= 128) && (K <= 128)) {
   _sgemm_bdx_128( argc, argv );
} else
if ((M <= 1024) && (N <= 1024) && (K <= 1024)) {
   _sgemm_bdx_1k( argc, argv );
} else
. . .

int my_func( int argc, void ** argv )
{
   int err = 0;
   . . .
   err =
   . . .
}
```

Patch: "call 0xCDEAF000"

Step 4: Do it! And prove

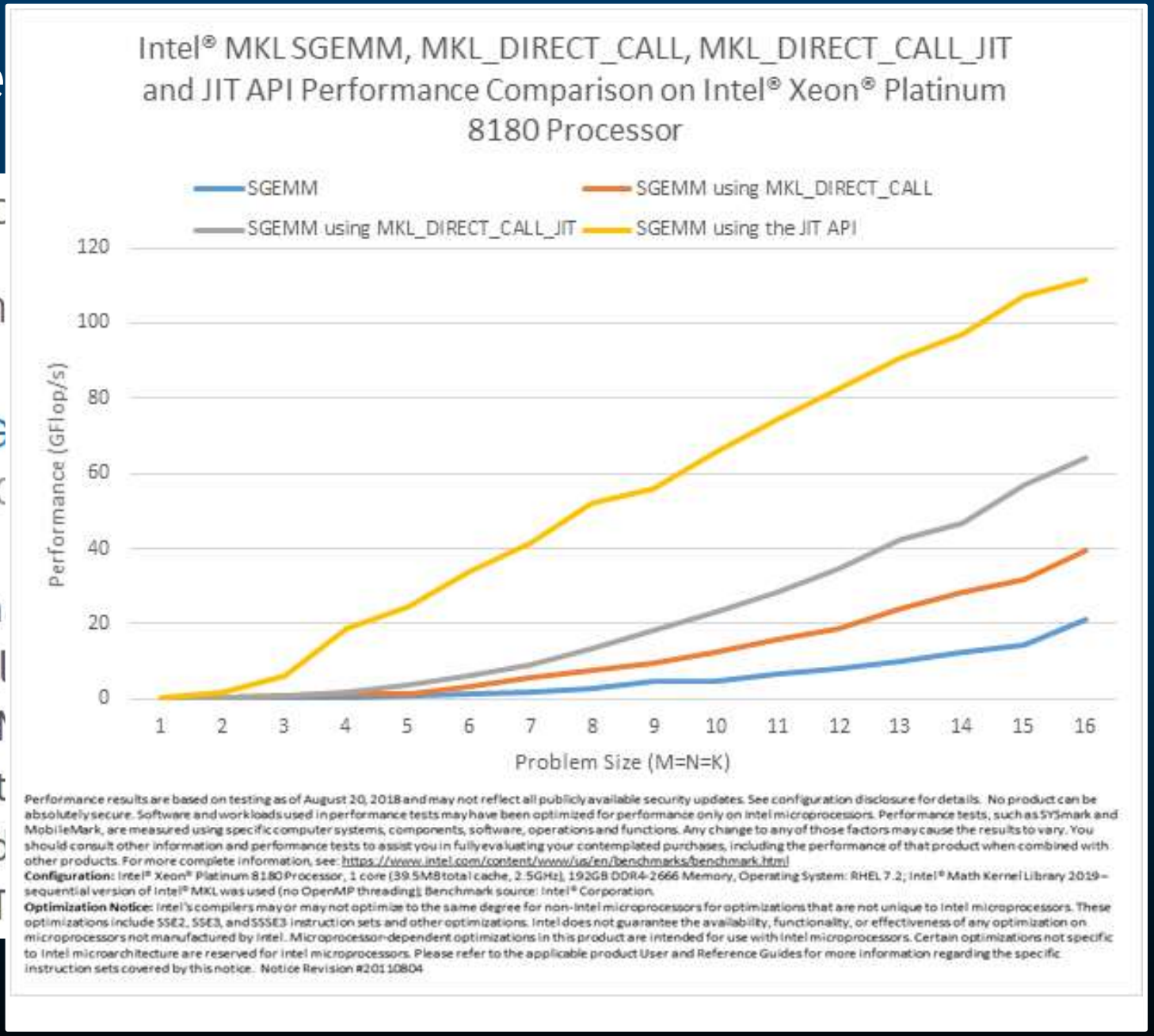
Intel® Math Kernel Library Improvec

Code Generation

By G

Published:09/(

The most commonly used and perform
functions are the general matrix multipl
optimizations for small problem sizes (M
Just-In-Time (JIT) code generation for t
Extensions 2 (Intel® AVX2) and Intel® Ac
architectures. The new just-in-Time (JIT



. . . and back to our show

Tackling the root cause of the memory wall cliff

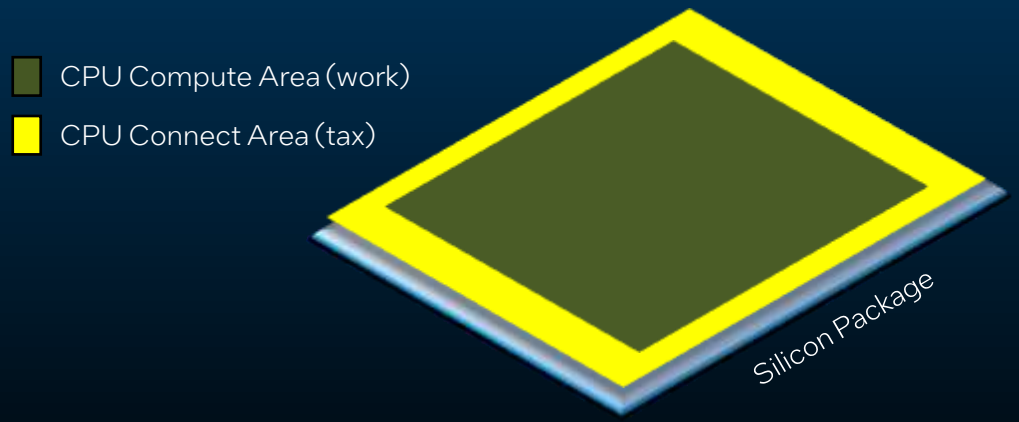
Uncle Sam needs novel memory for nuke sims. So why did it choose Intel?

Didn't the x86 giant just blow up its data storage biz?

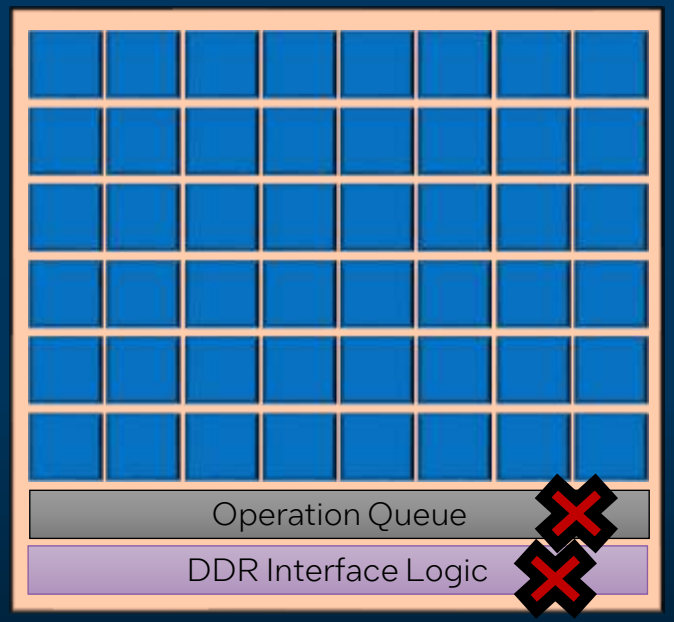
Tobias Mann Tue 13 Dec 2022 | 11:30 UTC

The US Department of Energy's Sandia National Labs believes that novel memory tech may be the secret to faster, more accurate nuclear weapon simulations.

https://www.theregister.com/2022/12/13/intel_doe_nukes_mem/



Impact to area for "work" by 2030?



Breaking the perimeter-area problem means delivering O(1) TB/s of bandwidth into O(10) GB of capacity with both dense and sparse access performance

Source: Intel Corporation

Feb 3 in Tokyo, Japan: Intel Connection

tom's HARDWARE premium

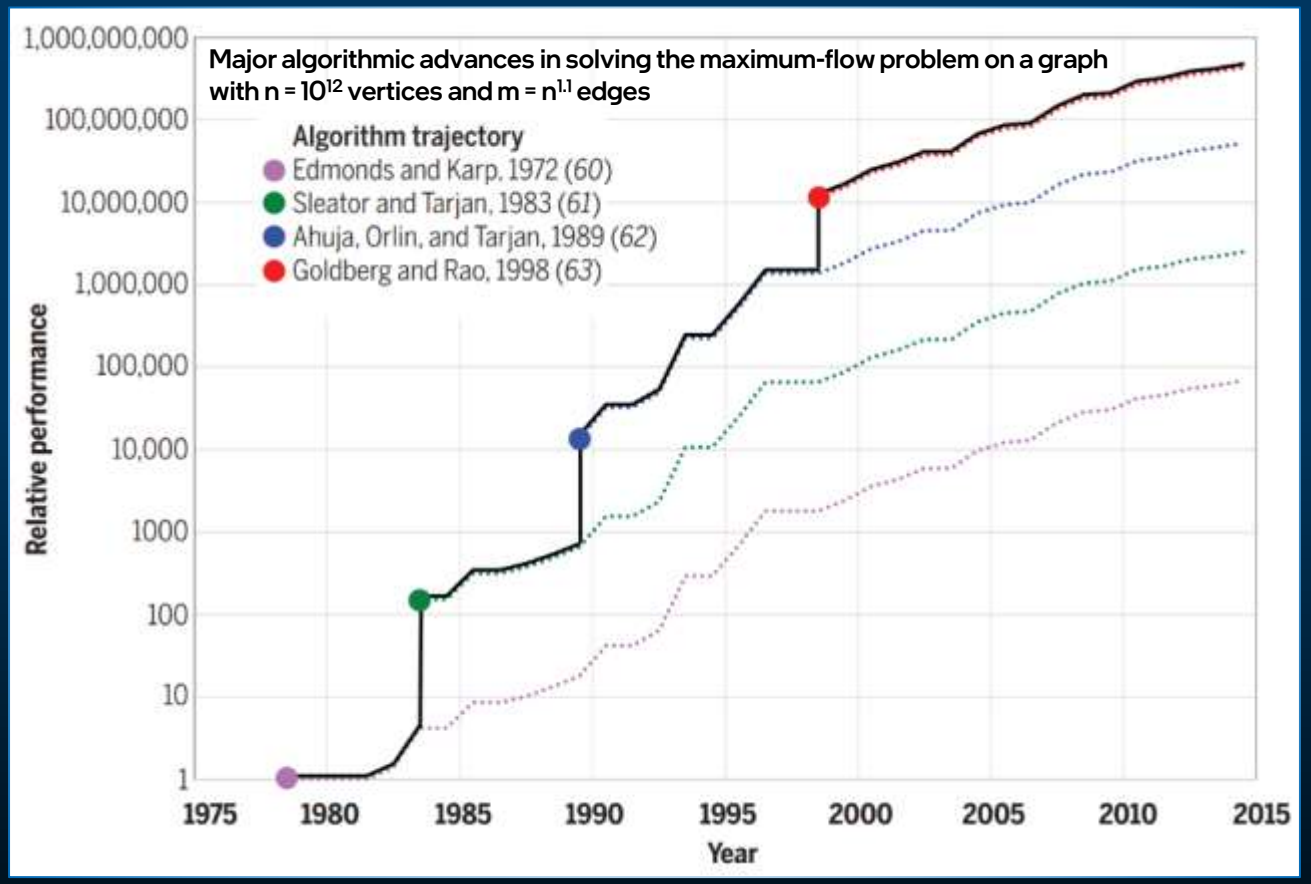
Tech Industry > Artificial Intelligence

Intel is co-developing new Z-Angle Memory to compete with HBM used in AI data centers — vertically-stacked memory touts 2 to 3x more capacity, greater bandwidth, and half the power consumption

News Analysis By Jon Martindale published February 3, 2026

With the first prototypes expected in 2027.

"A New Golden Age of System Architecture"



Credit: There's Plenty of Room at the Top, Leiserson et al, *Science, June 2020, Vol 368*

Ozaki Scheme II: A GEMM-oriented emulation of floating-point matrix multiplication using an integer modular technique

Katsuhisa Ozaki[1] Yuki Uchino[2]

and Toshiyuki Imamura[2]

[1] Department of Mathematical Sciences, Shibaura Institute of Technology, Japan

[2] RIKEN Center for Computational Science, Japan

ozaki@sic.shibaura-it.ac.jp

HPC Wire, April 2025

Modern Computing Challenge: Big-O needs a Do-Over

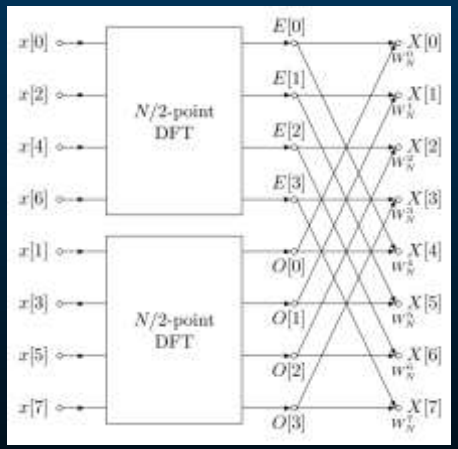
Fourier transform

$$\widehat{f}\,(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi\xi x}\, dx, \quad \forall \xi \in \mathbb{R}. \qquad \textbf{(Eq.1)}$$

DFT Sequence O(n^2)

$$X_k = \sum_{m=0}^{n-1} x_m e^{-i2\pi km/n} \qquad k = 0, \ldots, n-1$$

Partial FFT Sequence



O(n log n)

. . . but O(-) analysis is based on precious compute and free communications
. . . sometimes with memory capacity but not bandwidth factor
. . . and no sensitivity to latency or pressure on any IO interface
. . . O(right) in 1894-1909 is O(wrong) by 2020

Image credit: Wikipedia

. . . and One More Thing

AIRSPEEDER

AIRSPEEDER TEAM

AUS

2026

AIRSPEEDER

AUSTRALIA LEADING THE WORLD IN THE SKY

Airspeeder is a global competition between nations, where each team carries its flag into a new era of flight. It's the return of international rivalry to aviation, staged as sport, but powered by breakthroughs that can reshape how the world moves.

Australia intends to lead in the sky the way it's always led on the edge: with tough engineering, straight talk, and a willingness to test it for real. From workshops and wind-swept airfields where performance is earned, not posed, Team Australia brings a national instinct for turning speed into capability and competition into a proving ground.

Property of Airspeeder

AIRSPEEDER

AI, HPC, Materials Science, and more .. volunteer help wanted!
Ilaria Angeli <ilaria.angeli@alauda.aero>

THE AIRSPEEDER

A Living Testbed for Autonomy, Electrification, and Safety


intel