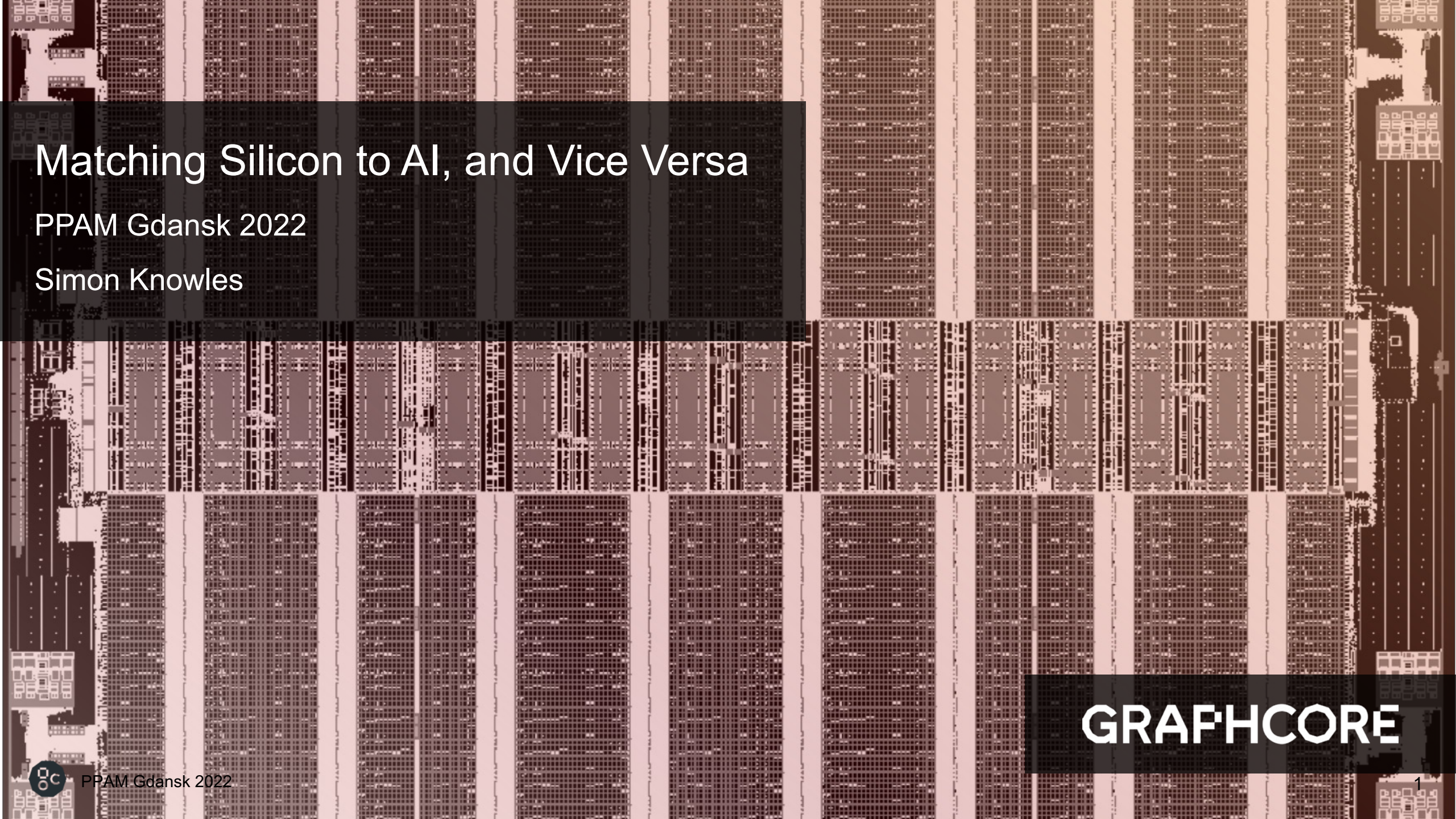


Matching Silicon to AI, and Vice Versa

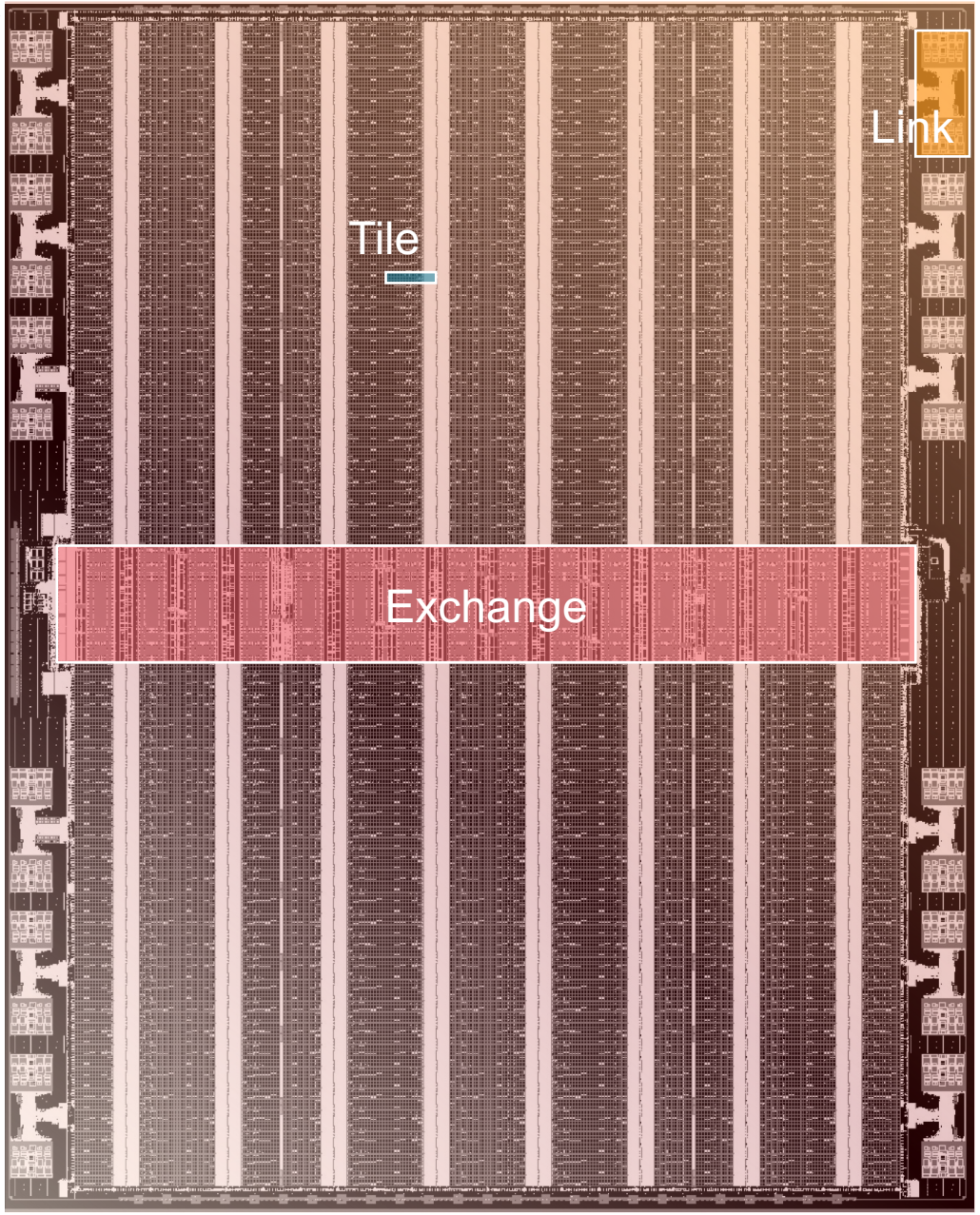
PPAM Gdansk 2022

Simon Knowles

GRAPHCORE

Graphcore Colossus Mk2 IPU

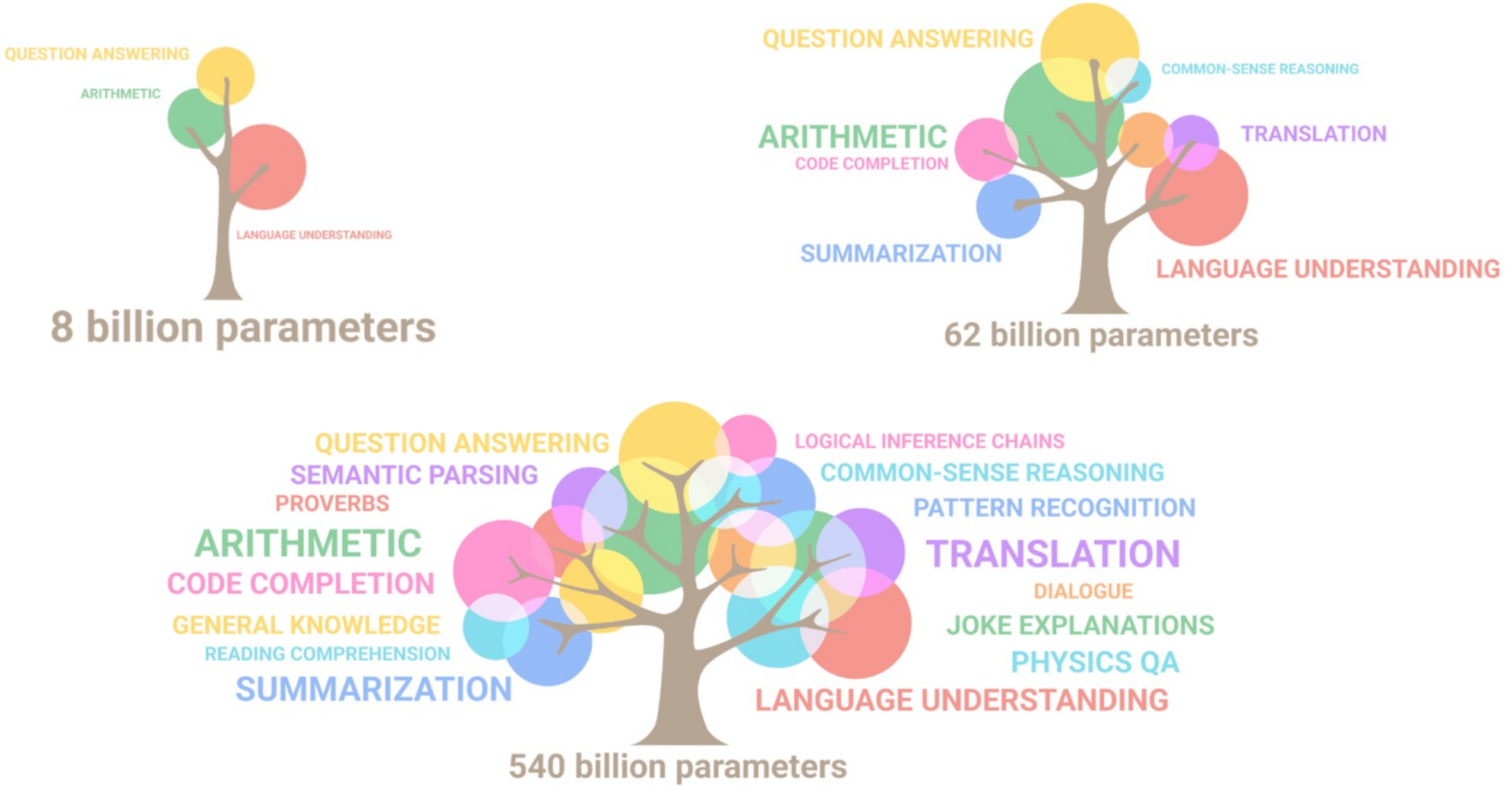
- Full-reticle N7, 14 metals, 59bn transistors
- 1472 processors (MIMD)
- 350Tflop/s fp16, 87Tflop/s fp32
- 897MiB distributed SRAM @ 65TB/s
- 11TB/s crossbar inter-tile interconnect
- 10x 16GB/s inter-chip links, 2x 16GB/s PCIe



AI Algorithms - State of Play

- Dense neural networks are pervasive.
- Useful, efficient models of O(100m - 100bn) weights.
- Training compute ~ O(100 • weights^2) … Eflop - Yflop
- Inference compute ~ O(100 • weights) … Gflops - Tflops
- All signs point to bigger models being more capable

Intelligent Capabilities Emerge with Bigger Models



Chowdhery et al, “PaLM: Scaling Language Models with Pathways”, arXiv:2204.02311

"*The survival of man depends on the early construction of an ultra-intelligent machine*.

… *defined as a machine that can far surpass all the intellectual activities of any man however clever*."

Isadore Jacob Gudak / Irving John Good, 1962.

Parametric Scale of a Human

- Human brains have 100-1000 trillion trainable synaptic weights[(1)], probably highly redundant.
- Hippocampal synapses have a weight resolution of ~4.5 bits[(2)].
- Artificial neural nets can reuse learned weights across structure; brains cannot, so perhaps NNs need fewer weights.
- AI can specialize to “intellectual activities” more than a human.

=> Ultra-intelligence might require less than 100TB of learned state?

(1) Wikipedia.org/wiki/Neuron

(2) Bartol et al, 2015, “Hippocampal spine head sizes are highly precise”, bioRxiv

Dense Neural Network Training Energy

~3pJ/flop SoTA training transformers on “infrastructure class” AI machines:

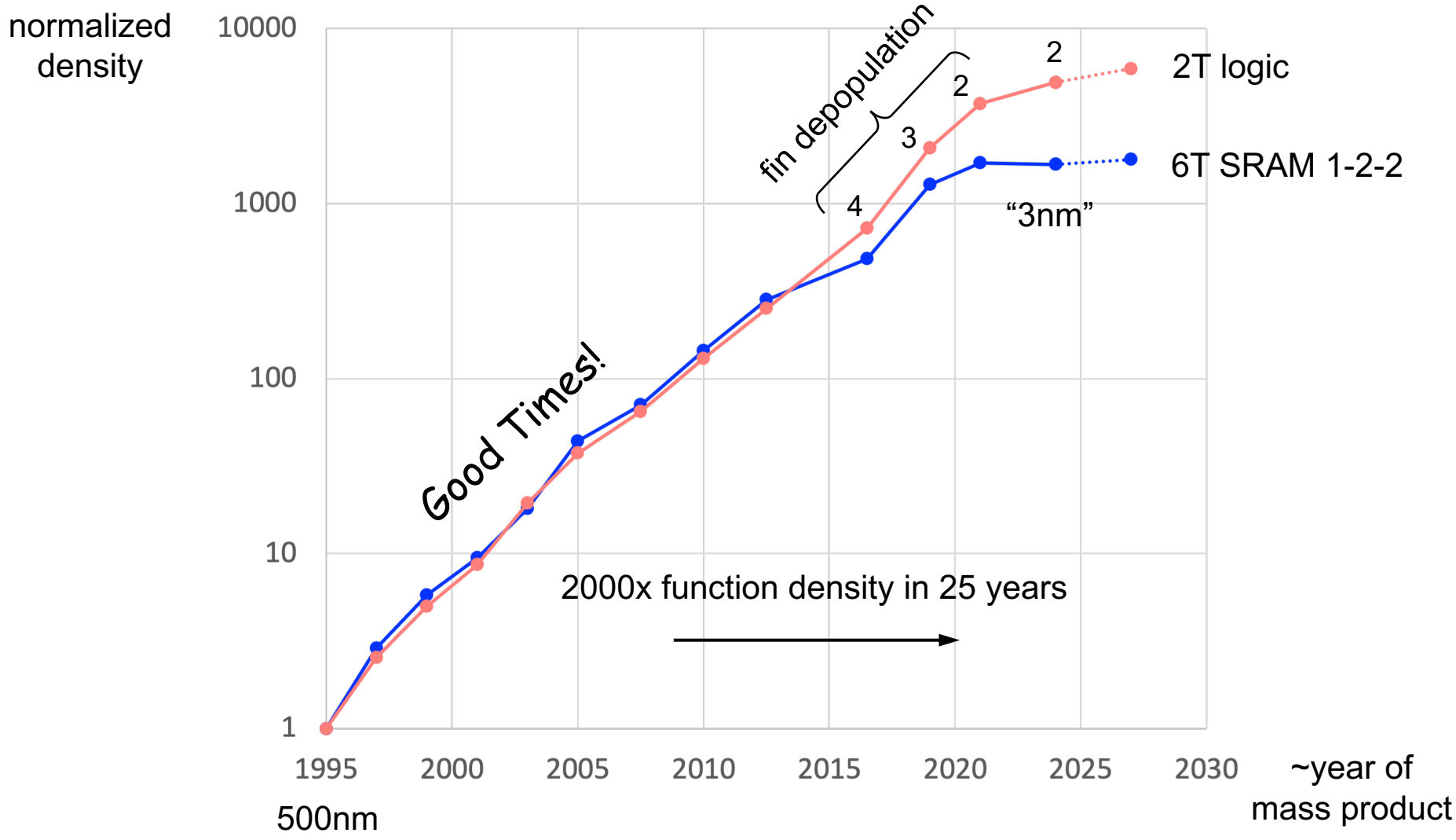
- 1 billion parameters, 20 billion tokens* … 250 chips (100kW) for 1 hour.
- 1 trillion parameters, 20 trillion tokens* … 25,000 chips (10MW) for 1 year.

(*) Guided by Hoffman et al, “Training Compute-Optimal Large Language Models”, arXiv:2203.15556.

US DoE ACE 2022 targets for an AI computer in 2030:

- ~1 Zflop/s … 140x Frontier HPL-AI
- ~60 fJ/flop … 1/50th Frontier HPL-AI

Transistor Density



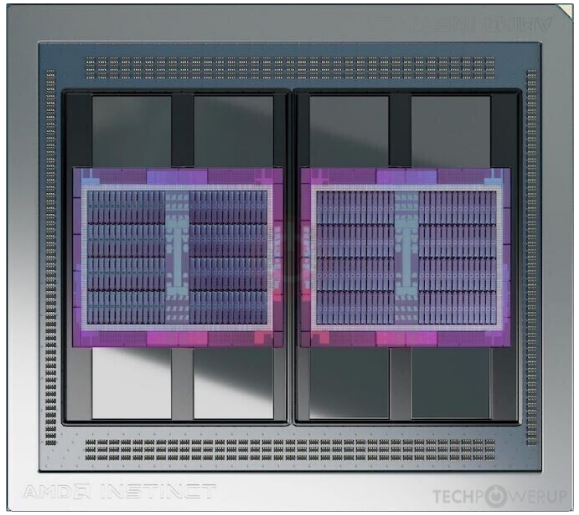
Die density engineering continues, but slowly.

IMEC roadmap:

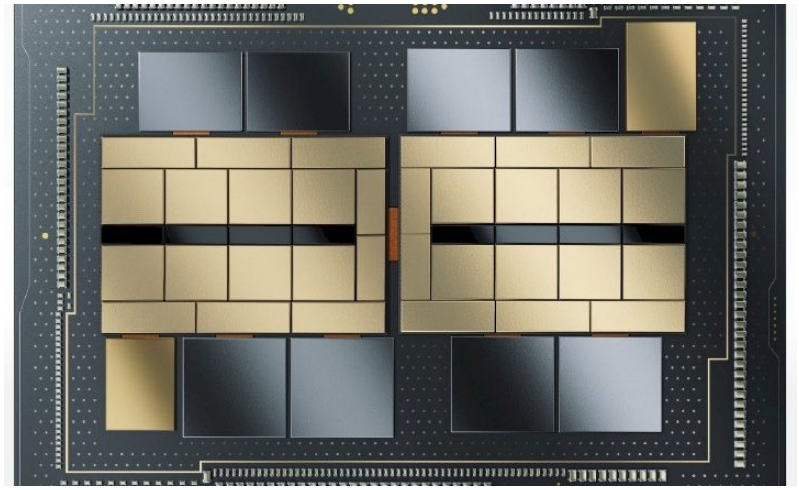
- GAA/forksheet expected at N2 … minimal density effect.
- Buried power rails (BPR) … in XPUs 2026?
- Vertical P/N stack (CFET) … in XPUs 2033?

Multi-die integration is replacing die density scaling

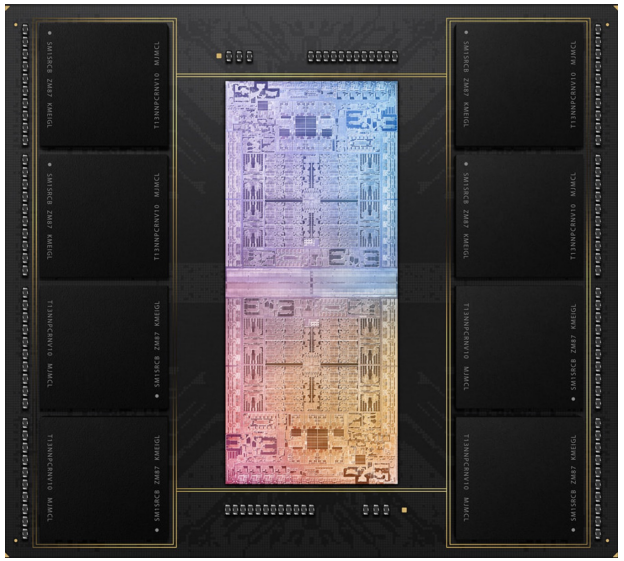
AMD MI250X: inter-CoWoS buried bridge



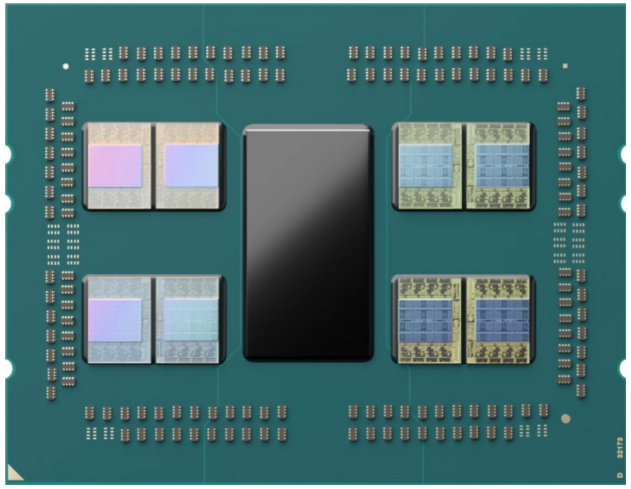
Intel Ponte Vecchio: 42-die on 2 interposers



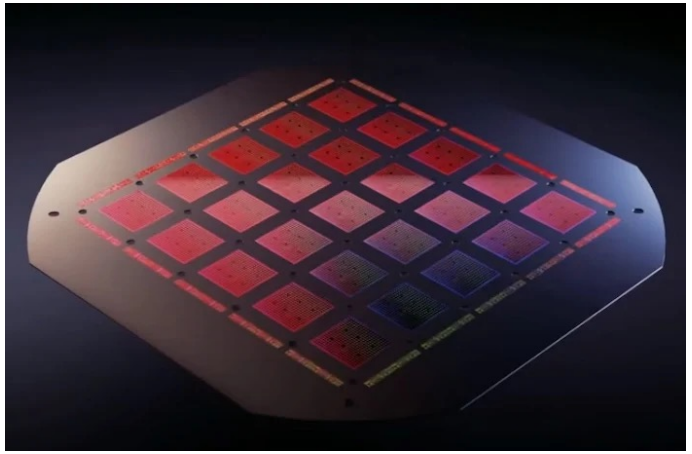
Apple M1-Ultra: buried silicon bridge, LPDDR5 on substrate



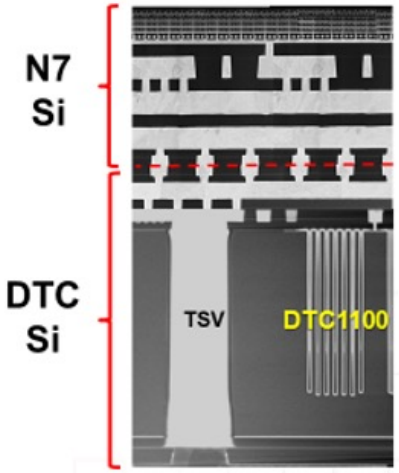
AMD Milan-X: Chip-on-Wafer caches



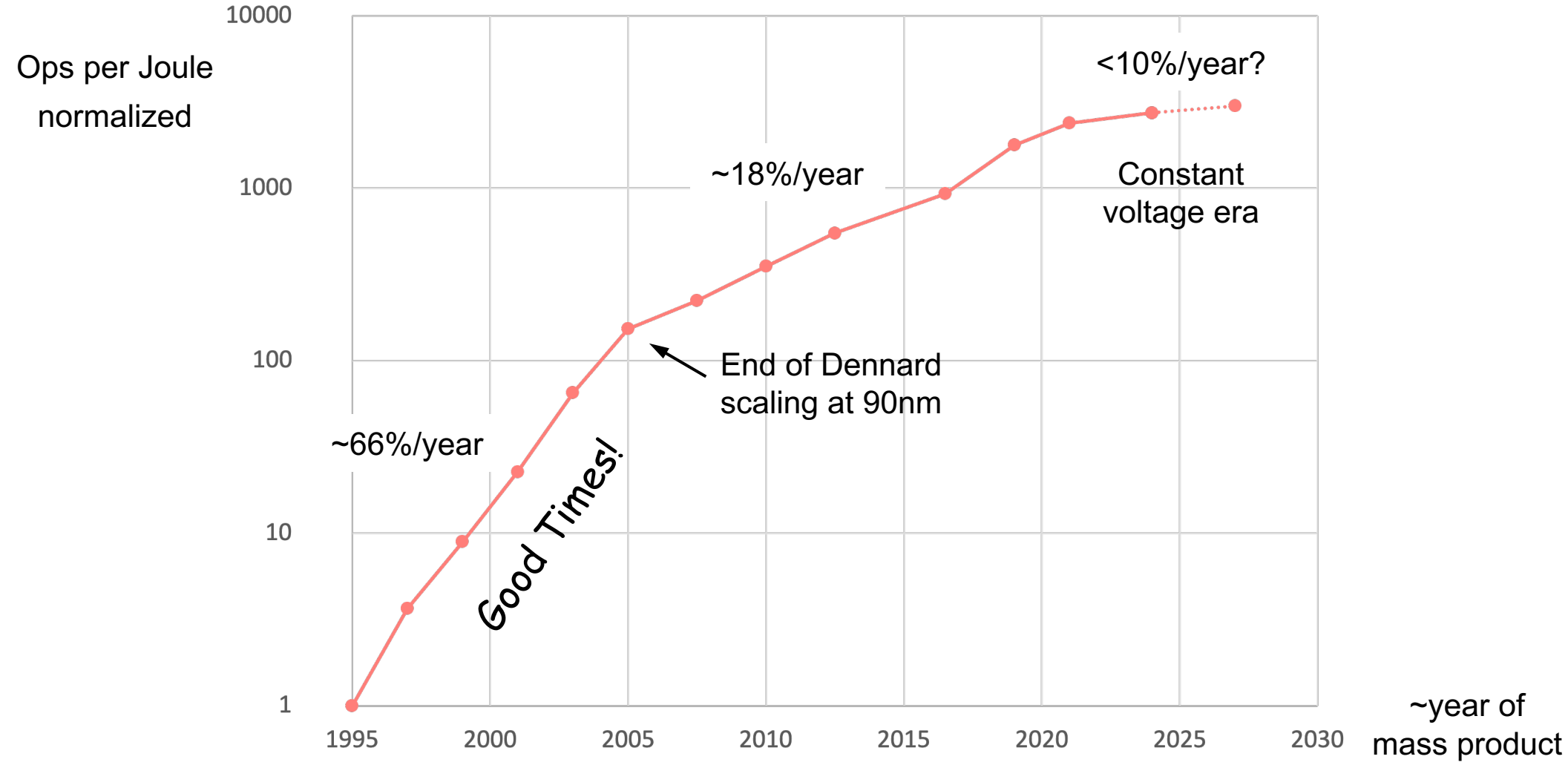
Tesla D100 wafer-scale InFO



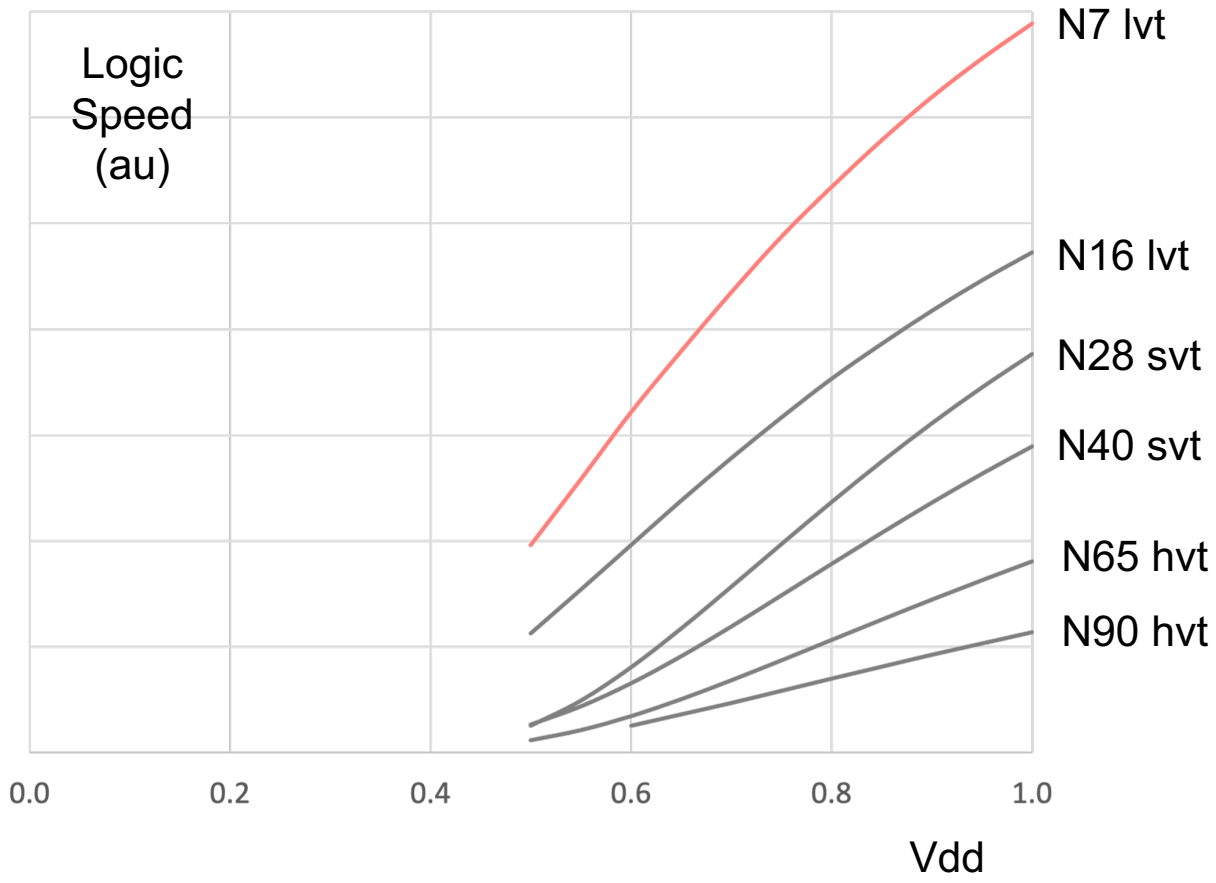
Graphcore: Wafer-on-Wafer decoupler



Performance per Watt



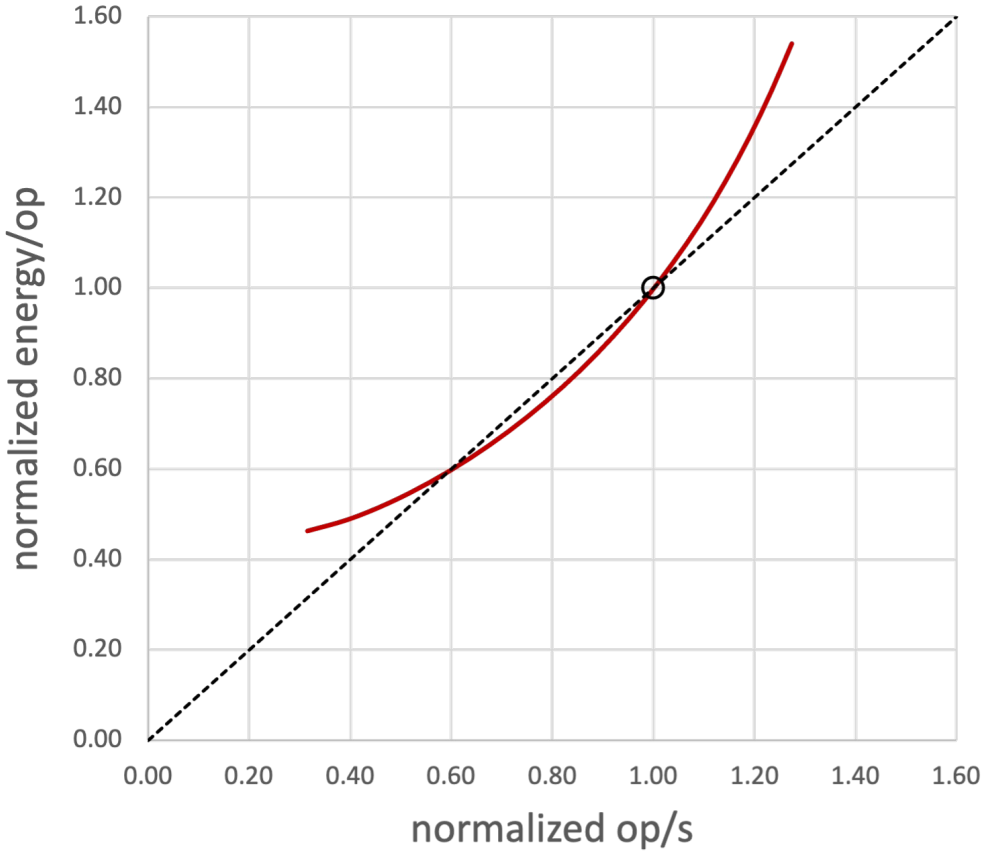
Why do all the XPUs operate at 1-2GHz?



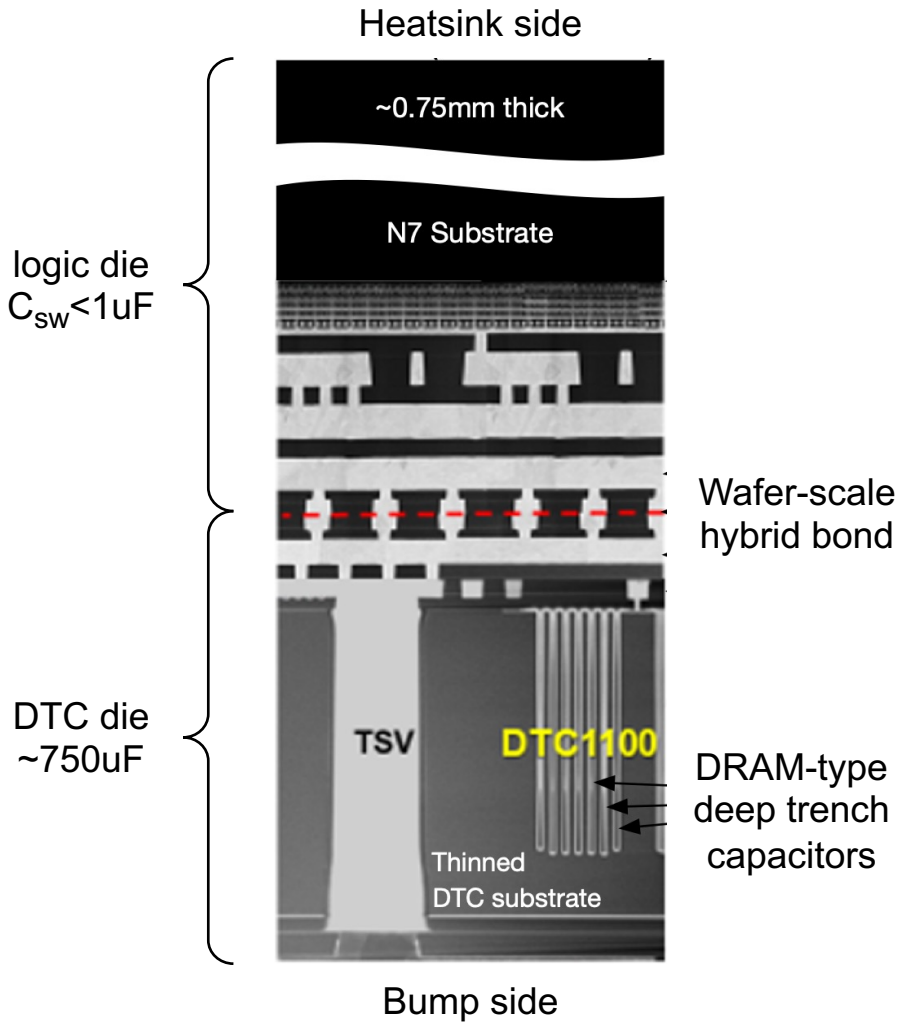
Energy Cost of Speed by Modulating Voltage

XPU design consensus is ~1.85GHz @ ~800mV

- 20% faster would cost ~40% more energy
- 40% slower would save ~40% energy

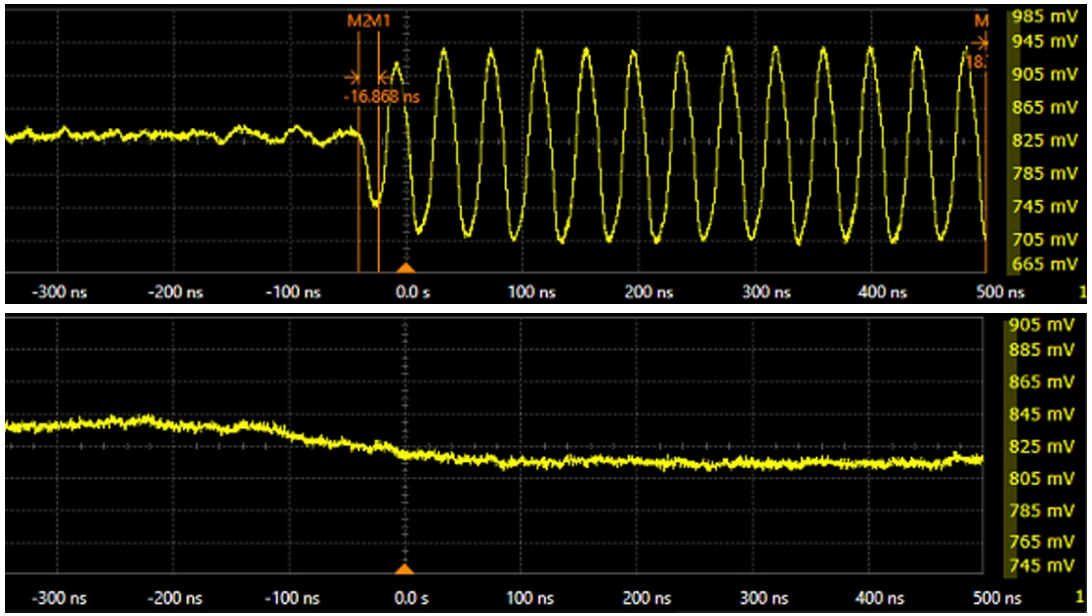


Taming Vdd

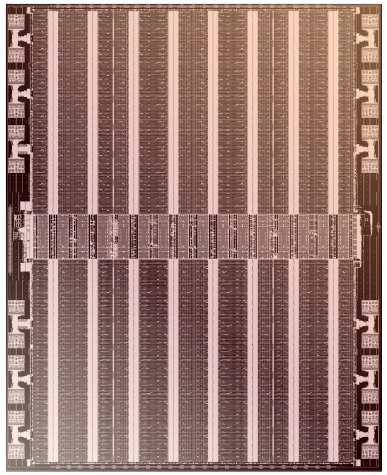


Graphcore Colossus Mk2w:

- First Wafer-on-Wafer (WoW) 3D logic chip
- 1.4x speed at same energy/op



Vdd at die without/with DTC; 25MHz 50/50 min/max activity virus

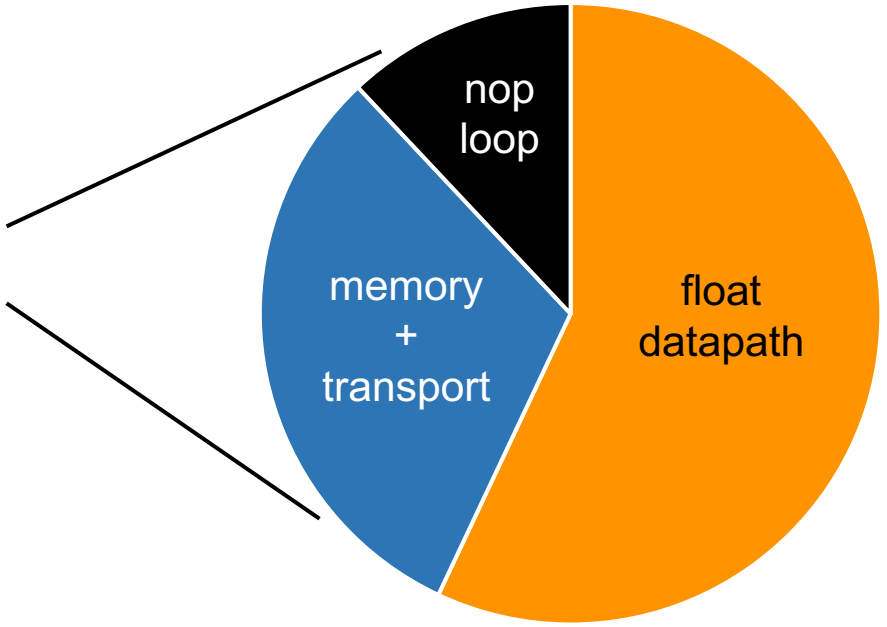


Kernel Power

Convolution dynamic power measured at the die with virus data

- Real application data is typically 1/3~1/2 less energetic.
- Power-optimized Mk2 die with wafer-on-wafer DTC decoupler.

Multiply	Accumulate		pJ/flop
	Datapath	Memory	
f16	f32	f16	1.0
	f32	f32	1.3
f32	f32	f32	2.5



float8 IEEE Standards WG P3109

- ~50% of fp16 energy/flop
- Works for training with managed scaling
- More accurate for inference than int8

1.4.3 “af8” : 4b exponent, 4b precision

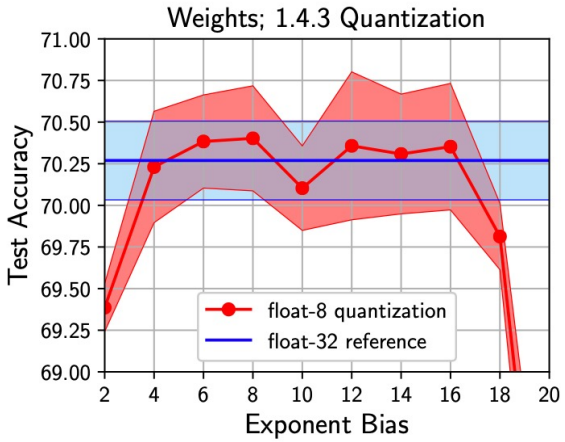
For weights and activations; best accuracy

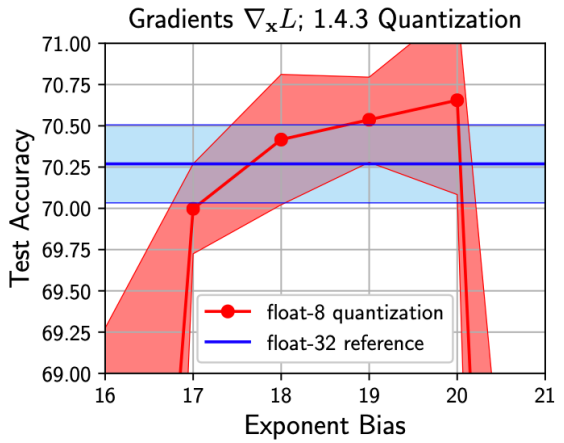
1.5.2 “bf8” : 5b exponent, 3b precision

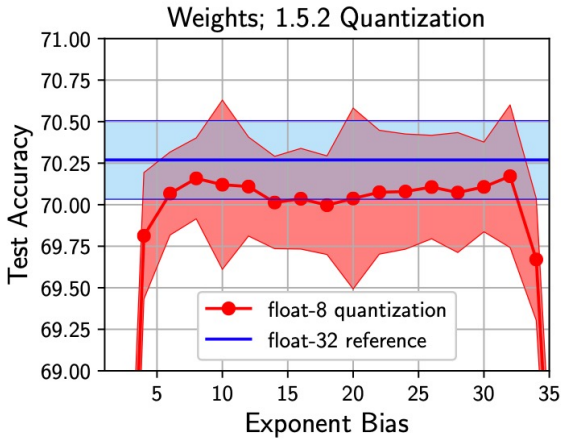
For gradients; best stability

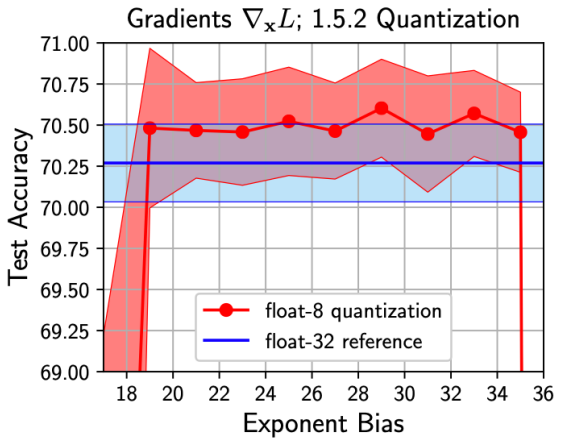
weights
(activations similar)

gradA
(gradW similar)





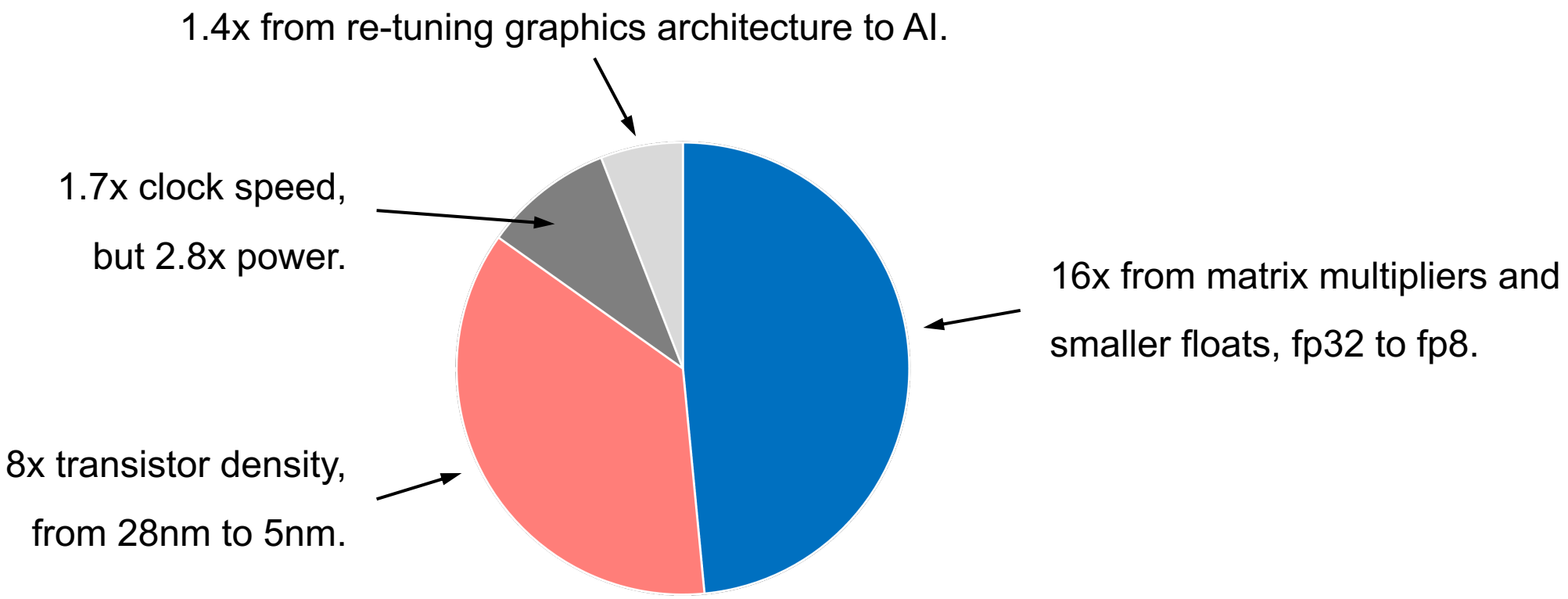




Noune et al. 2022, “8-bit Numerical Formats for Deep Neural Networks”, arXiv:2206.02915

~300x peak GPU arithmetic in the first AI Decade

NVIDIA Maxwell 6.6Tflop32/s in 2014 to Hopper 2000Tflop8/s in 2023



The Next Decade?

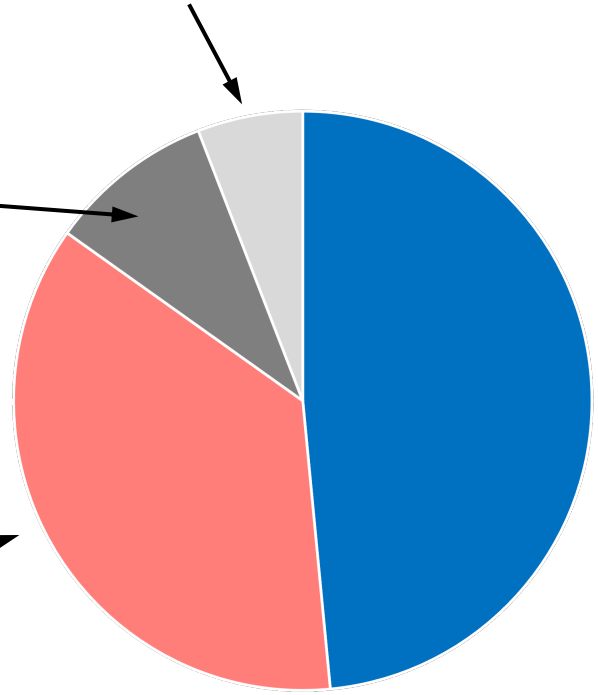
More from AI-specific architectures

1.4x from re-tuning graphics architecture to AI.

1.7x clock speed,

but 2.8x power.

Another 2x, at 3x power?



16x from matrix multipliers and

smaller floats, fp32 to fp8.

Done?

8x transistor density,

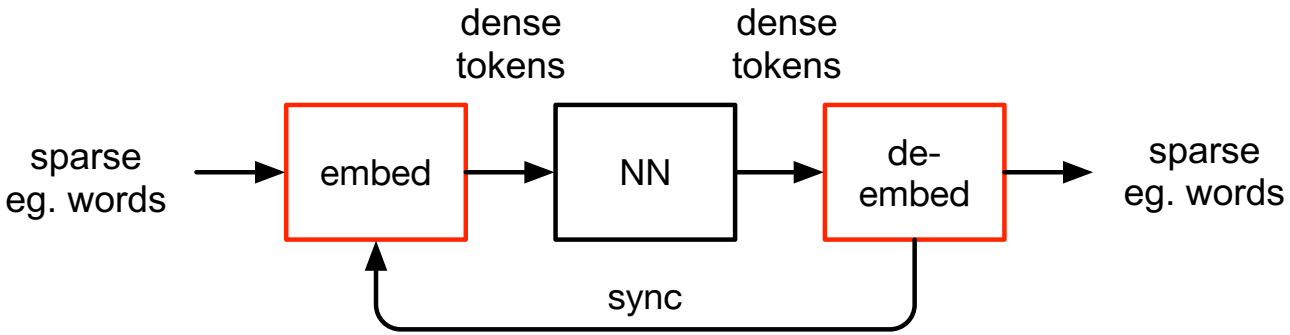
from 28nm to 5nm.

Another 2-3x?

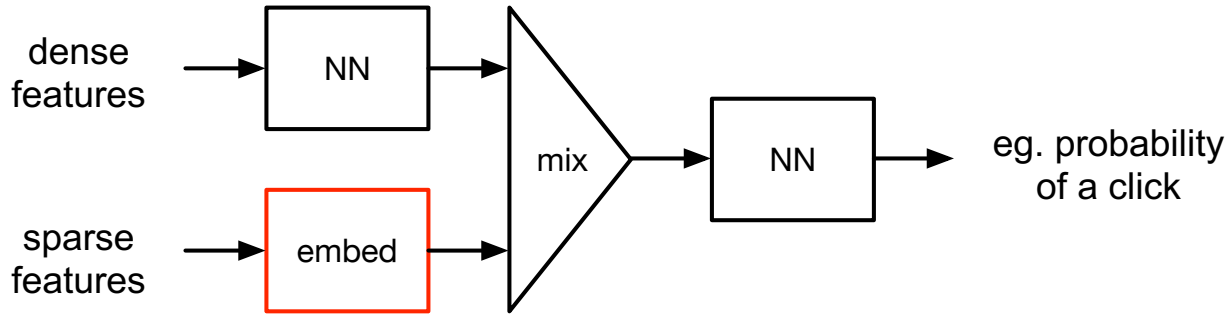
Not All AI is Flop-Dominated

Gradient backprop requires continuous functions. Embeddings map sparse categorical data to a dense vector representation. Embeddings are learnable, with arithmetic intensity = 1.

One machine architecture may not be best at both dense NNs and sparse embeddings.



Language Model



DL Recommender

Algorithm Imperative: Use Fewer Flops, More Information

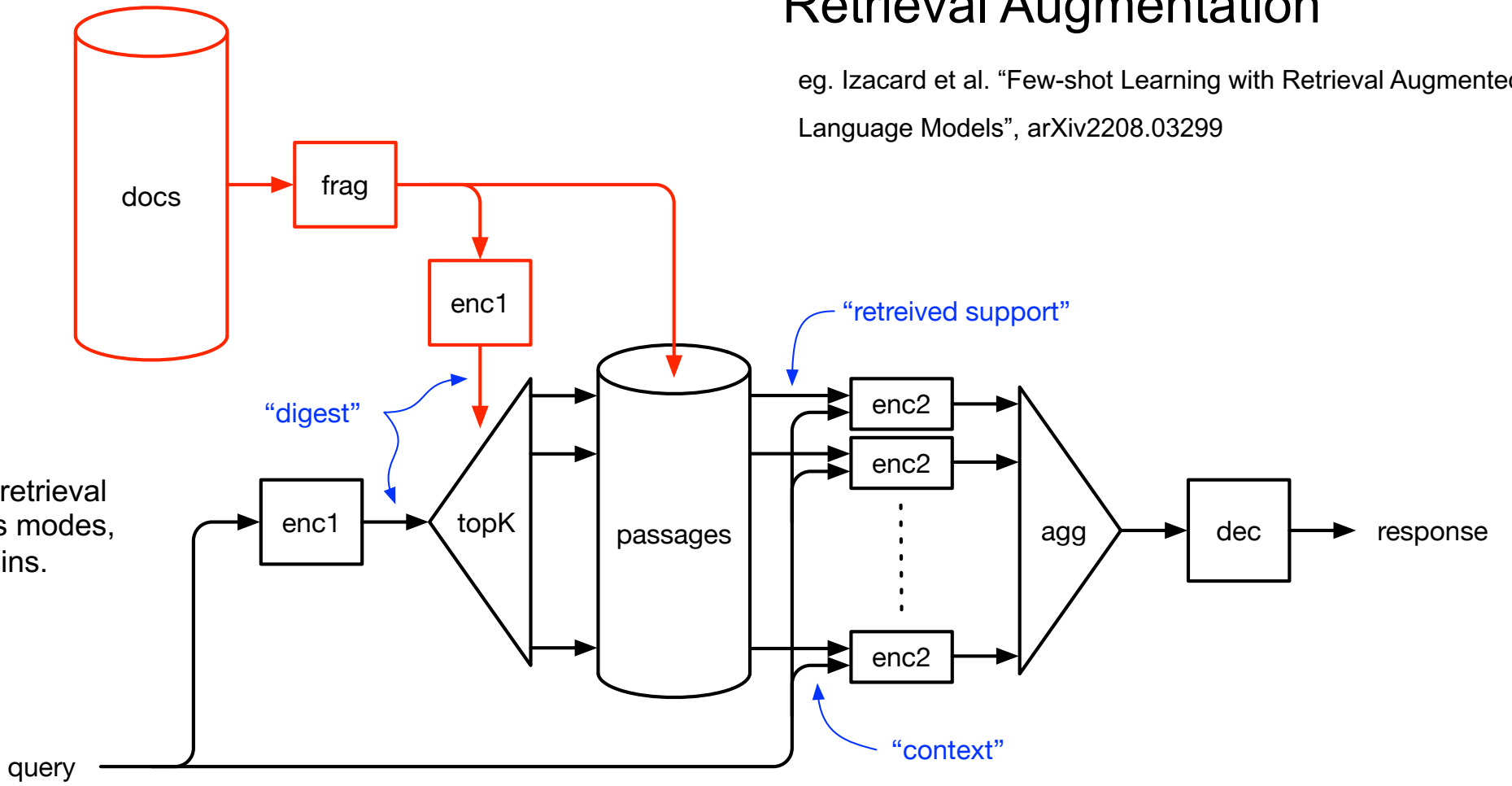
- Routing *aka* Conditional Sparsity … like a brain
- Retrieval Augmentation … like a human + www

Routing Networks

- In a dense neural network, every datum interacts with every weight.
- Brains don’t fire all their neurons in response to every stimulus.
- Efficient multi-task, multi-domain, multi-modal AI must obviously access its “knowledge” selectively.

Retrieval Augmentation

eg. Izacard et al. “Few-shot Learning with Retrieval Augmented Language Models”, arXiv2208.03299



Encoder1 allows retrieval of material across modes, languages, domains.

Here retrieval is by approximate matching of digests to index a set of passages as the knowledge base.

Enrich this support system by encoding a Knowledge Graph to allow path-based retrieval.

Model Scale Breakthroughs

Cost of data

O(100M) params

Deep learning

1000x

Cost of compute

O(100G) params

Unsupervised learning

1000x

Cost of memory

O(100T) params?

Routing networks

Retrieval augmentation

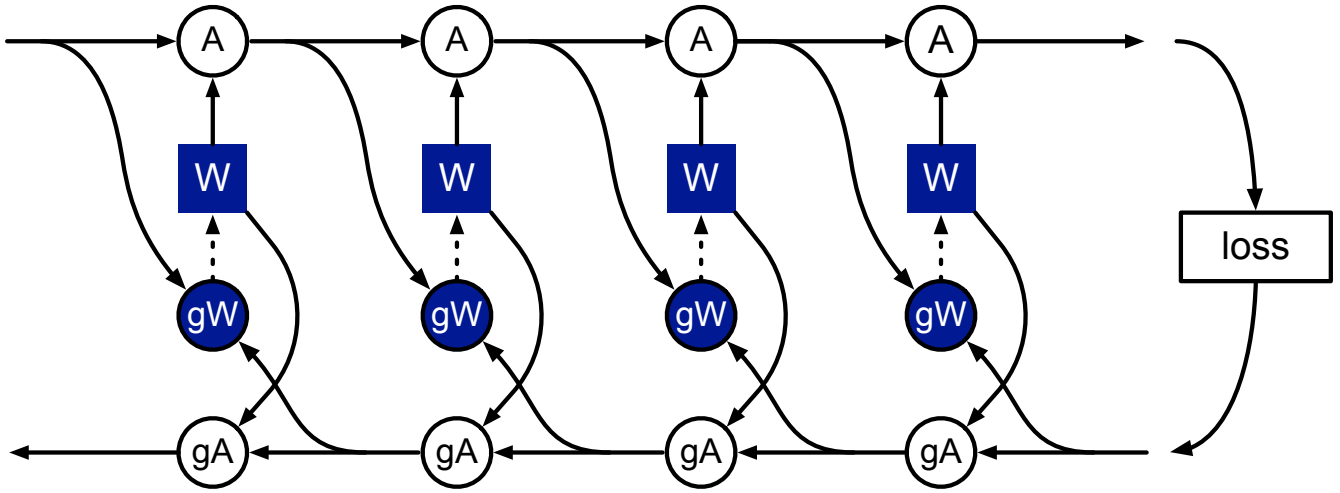
AI accelerators will need more (cheap, fast) memory

SoTA per reticle-sized logic die:

	GB	GB/s	pJ/B	normalized $cost/B
SRAM over 50% die	1	>> 50,000	<< 1	1
6x HBM3-4800 on silicon substrate	96	3000	40	4
★ 16x LPDDR5-6400 on organic substrate	512	800	50	1
12x DDR5-5600 on 128GB DIMMs	1536	500	300	1.25

…all +0.5pJ/B/mm on die (max 30pJ/B for half-perimeter)

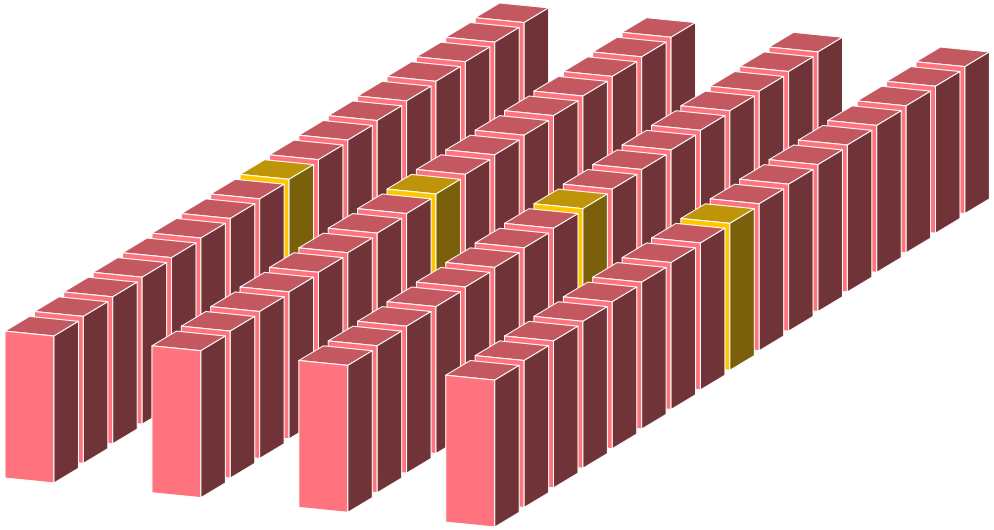
Memory you can’t update in ~1s isn’t very useful



SGD requires O(1 million) training iterations:

- each iteration reads and writes *all* model and optimizer state
- 1s per iteration => 2 weeks to train
- SSD would wear out, DRAM requires GB/s ~ GB

Brain-Scale Computing



GRAPHCORE **Good Computer** [mid-size]

- 1 PB DRAM ~ 200TB model (100T params)
- 2 PB/s ~ 1s model iteration
- 2k Mk3 IPUs ~ 1 Eflop16/s real
- ~2.5MWatts, 68 datacentre racks, 100m^2

Remember this:

- Silicon is approaching “constant energy per op”.
- Information capacity ultimately determines the potency of an AI, given sufficient training.
- AI computing will be limited by power and memory.
- “Human scale” AI is feasible.
- AI algorithm innovation needs to focus more on memory, less on flops.