SC22
Dallas, TX | hpc accelerates.

Hybrid AI/HPC Approaches for Next Generation Multi-Trillion-Parameter Models

ScalAH22: 13th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Heterogeneous Systems

Phil Brown, VP Scaled Systems

GRAPHCORE

“*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*.”

Irving John Good, 1962.

Valuable AI Computers

Capacity machines : cheaper than human, per unit of work.

Capability machines : super-human, at least in specific domains.

Capability is determined by:

- *quantity and quality of training data*
- *model scale*
- *effectiveness of representations*
- *effectiveness of training and inference processes*

$\left.\begin{array}{l}\text{quantity and quality of training data}\\ \text{model scale}\end{array}\right\}$ Educational effort

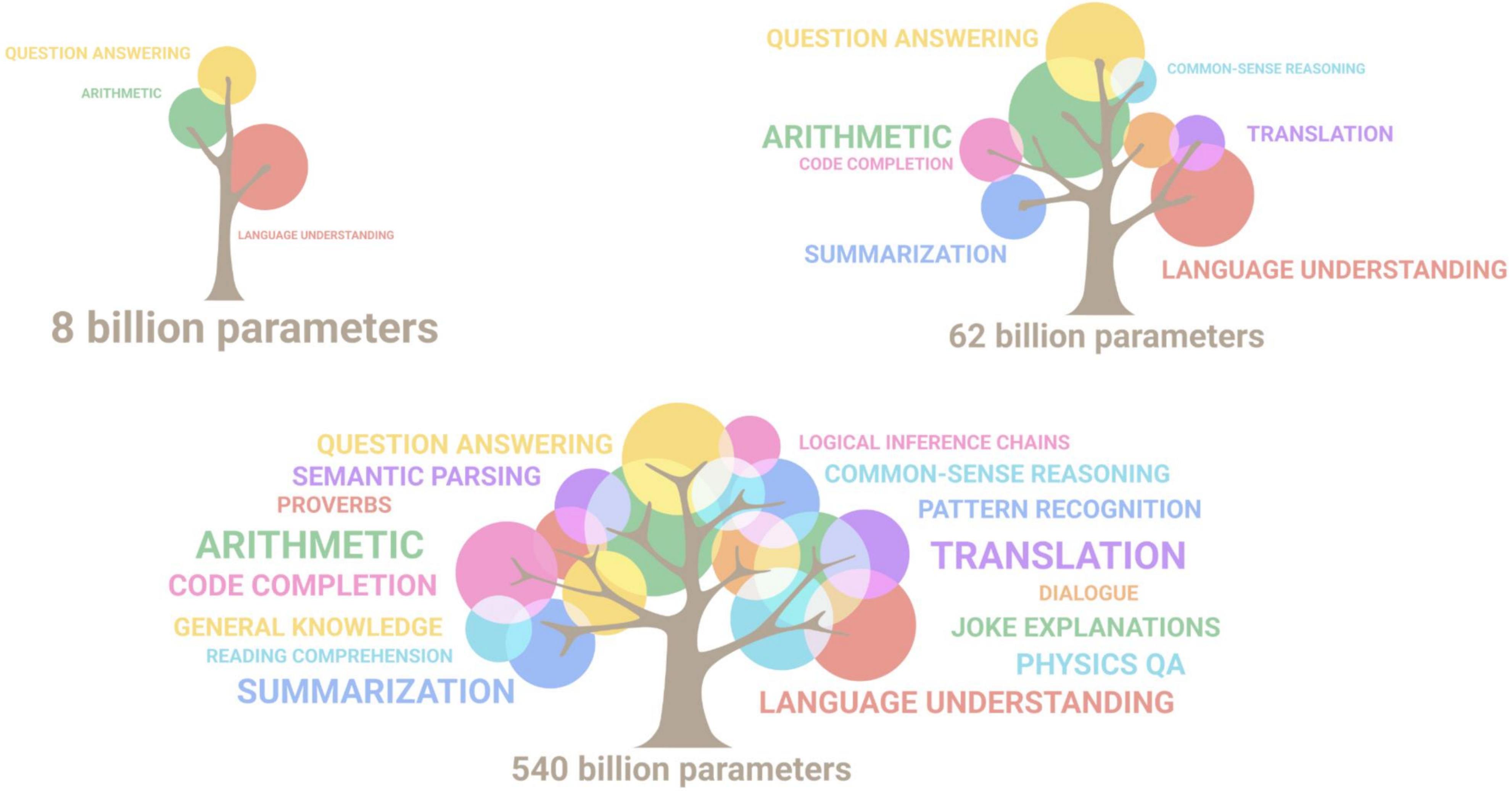
$\left.\begin{array}{l}\text{effectiveness of representations}\\ \text{effectiveness of training and inference processes}\end{array}\right\}$ AI design

Bigger Models => More Potent AI

Stored information capacity (#parameters) limits what can be computed by an AI, given sufficient training.

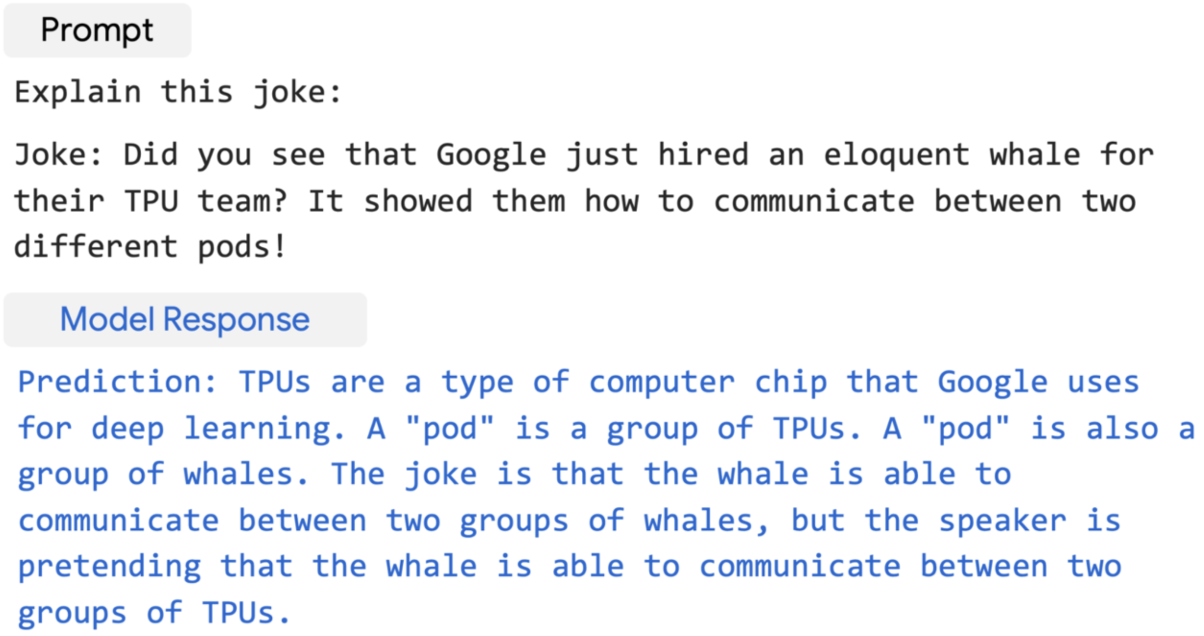
The computation rate just determines how quickly it can learn and infer.

Capabilities Emerge with Model Scale



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

Sense of Humour



Prompt

Explain this joke:

Joke: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

Model Response

Prediction: TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

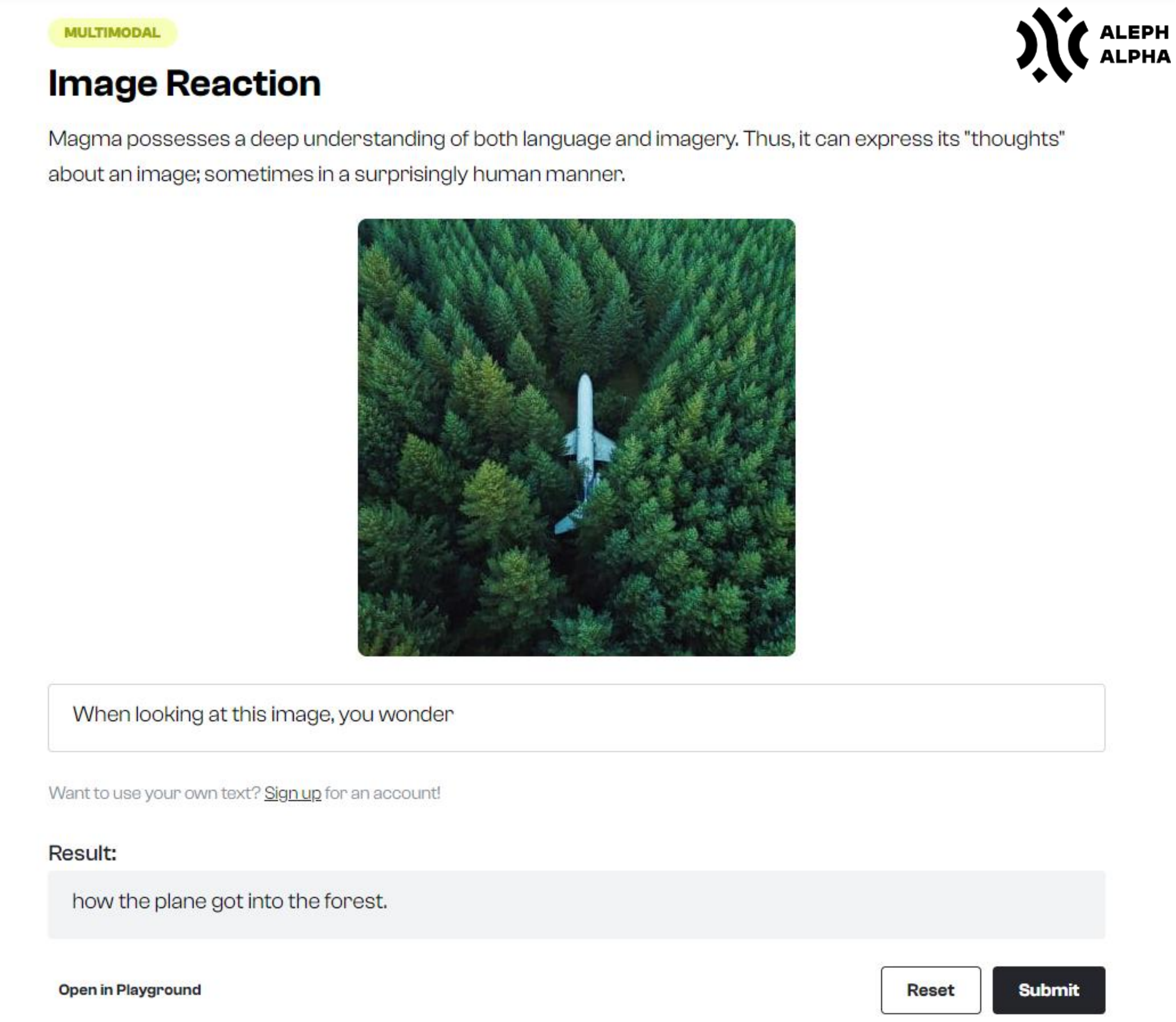


Image Reaction

Magma possesses a deep understanding of both language and imagery. Thus, it can express its "thoughts" about an image; sometimes in a surprisingly human manner.

When looking at this image, you wonder

Want to use your own text? Sign up for an account!

Result:

how the plane got into the forest.

Parametric Scale of a Brain

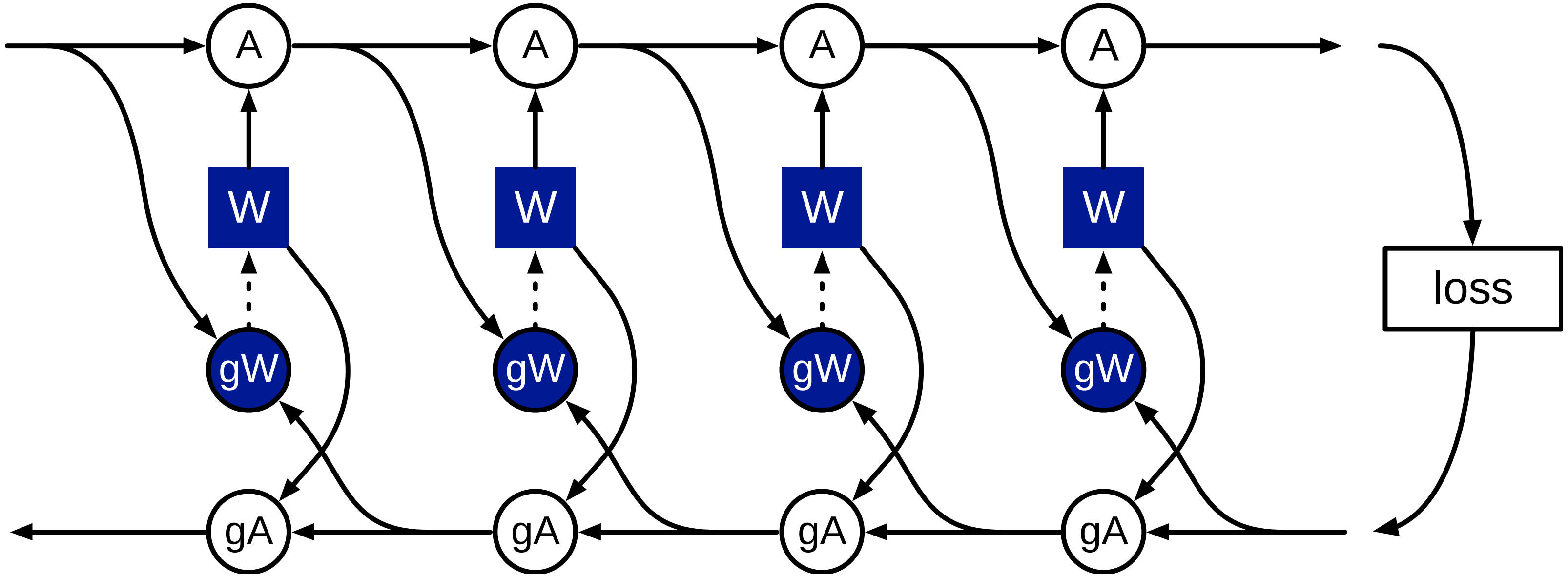
- Human brains have 100-1000 trillion trainable synaptic weights(1), probably highly redundant.
- Hippocampal synapses have a weight resolution of ~4.5 bits(2).
- Artificial neural nets can reuse learned weights convolutionally, and can specialize to “intellectual activities” more than a human.
- So 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

Master Learning Algorithm

First-order stochastic gradient descent (SGD) by back-propagation



Implications of the Master Learning Algorithm

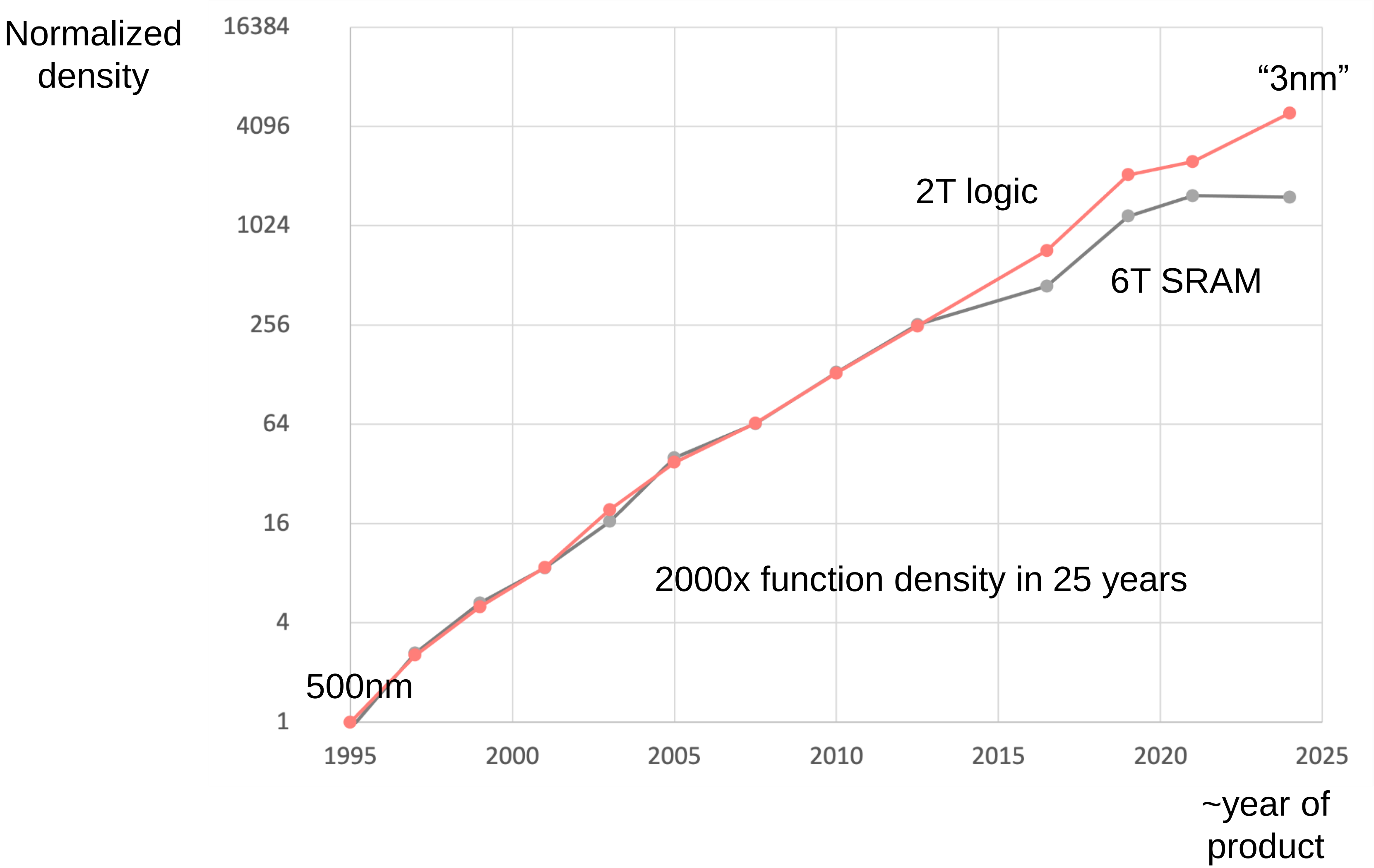
1) End-to-end sequential dependencies => no natural pipelining.

2) Major data flow is explicit; little value in automatic caches.

3) O(1 million) SGD iterations, each reading and writing all model and optimizer state:
   - SSD endurance is only a few thousand cycles, so all SRAM/DRAM.
   - 1 second /iteration ~ 12 days to train; TBytes must be matched by TB/s.

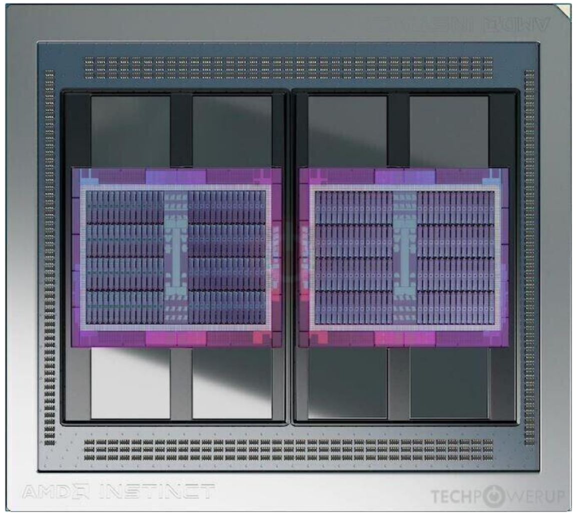
Meanwhile, in Silicon…

Silicon Density Scaling

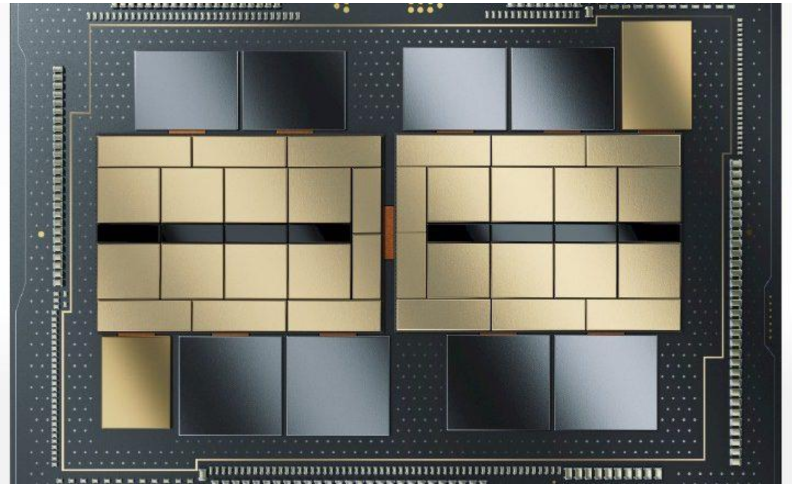


Multi-die integration replacing “Moore’s Law” scaling

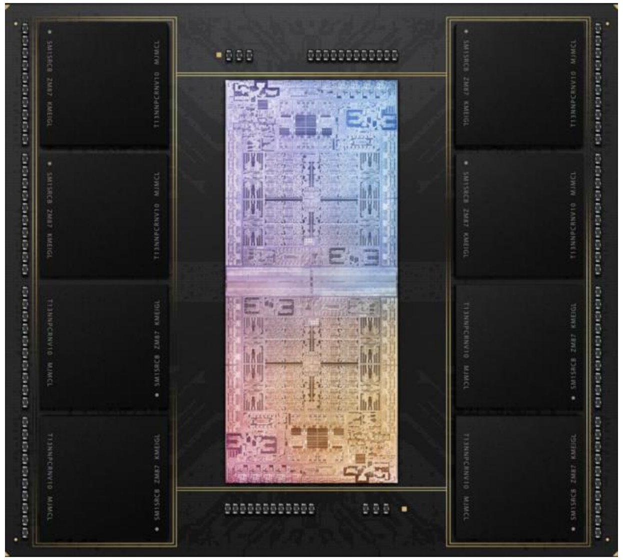
AMD MI250X: inter-CoWoS buried bridge



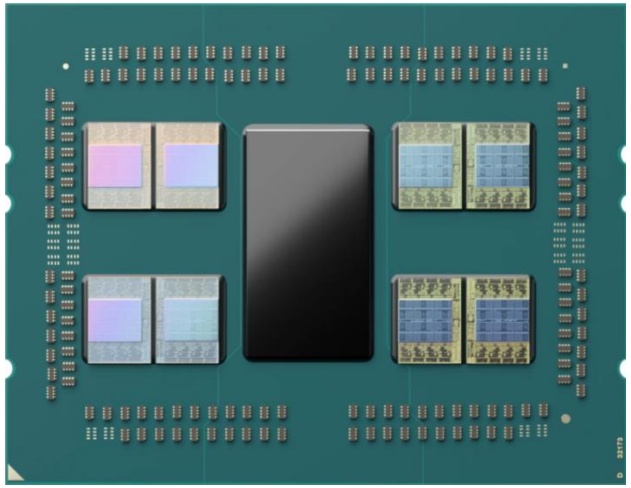
Intel Ponte Vecchio: 42-die on 2 interposers



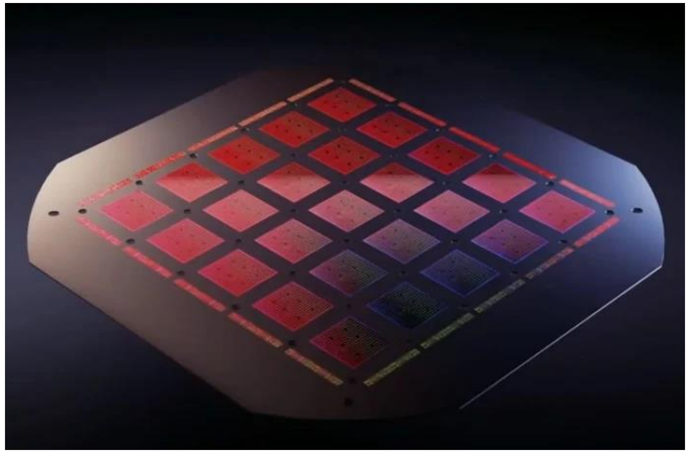
Apple M1-Ultra: buried silicon bridge, LPDDR5 on substrate



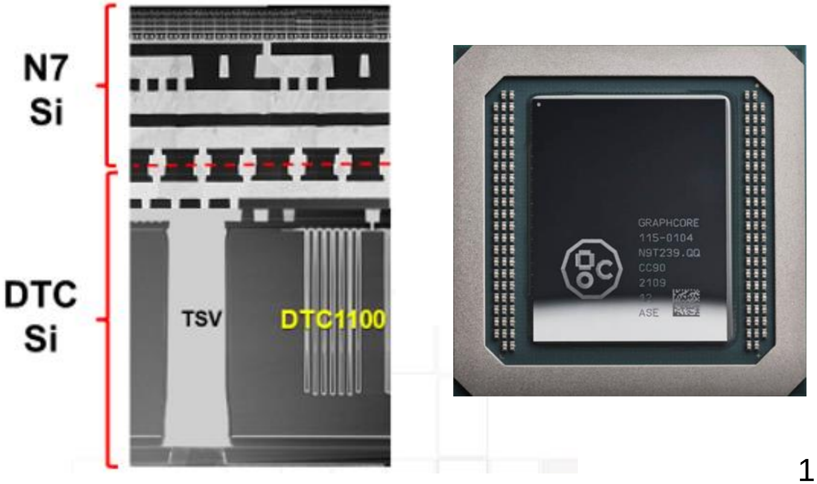
AMD Milan-X: Chip-on-Wafer caches



Tesla D100 wafer-scale InFO

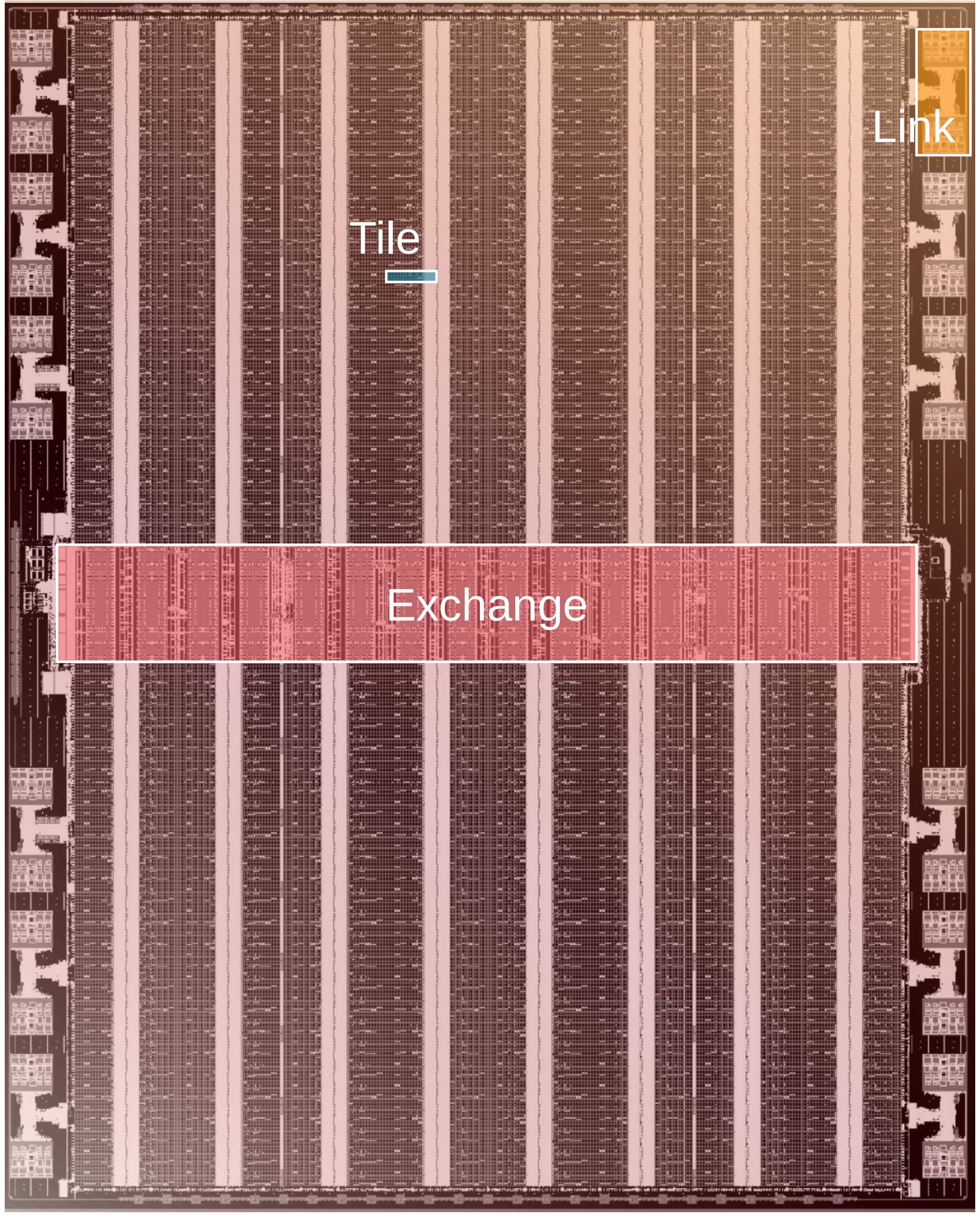


Graphcore: Wafer-on-Wafer decoupler

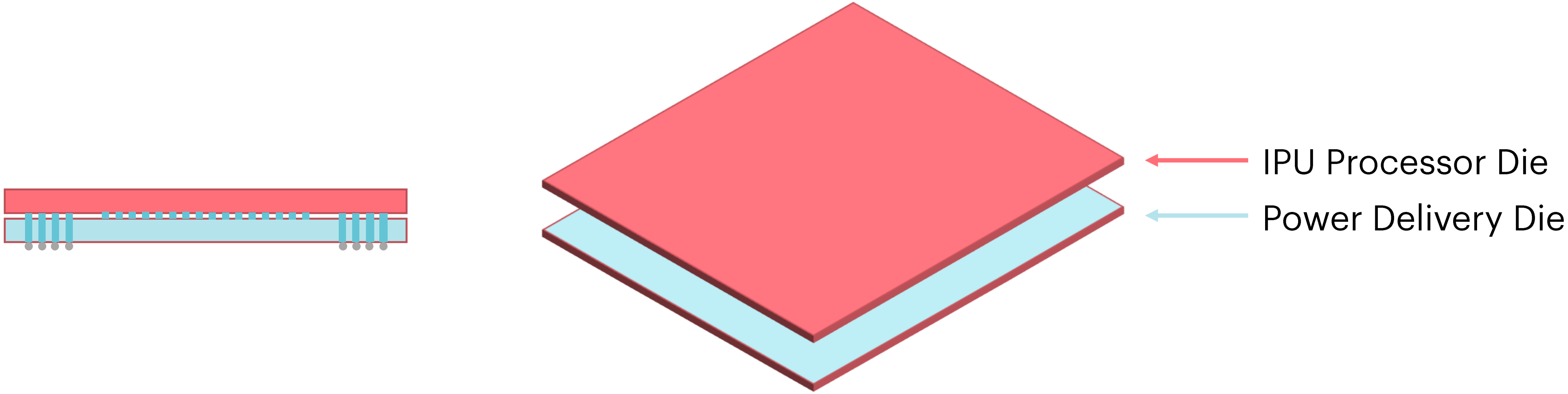


Graphcore Colossus Mk2 IPU

- 59,334,610,787 active transistors
- 7nm process, 14 metals, 86 masks, full reticle 8cm^2
- 1472 independent processors = 350Tflop16/s
- 900MiB distributed SRAM ~ 50% die area
- 11TB/s non-blocking inter-tile interconnect
- 1.85GHz mesochronous clock



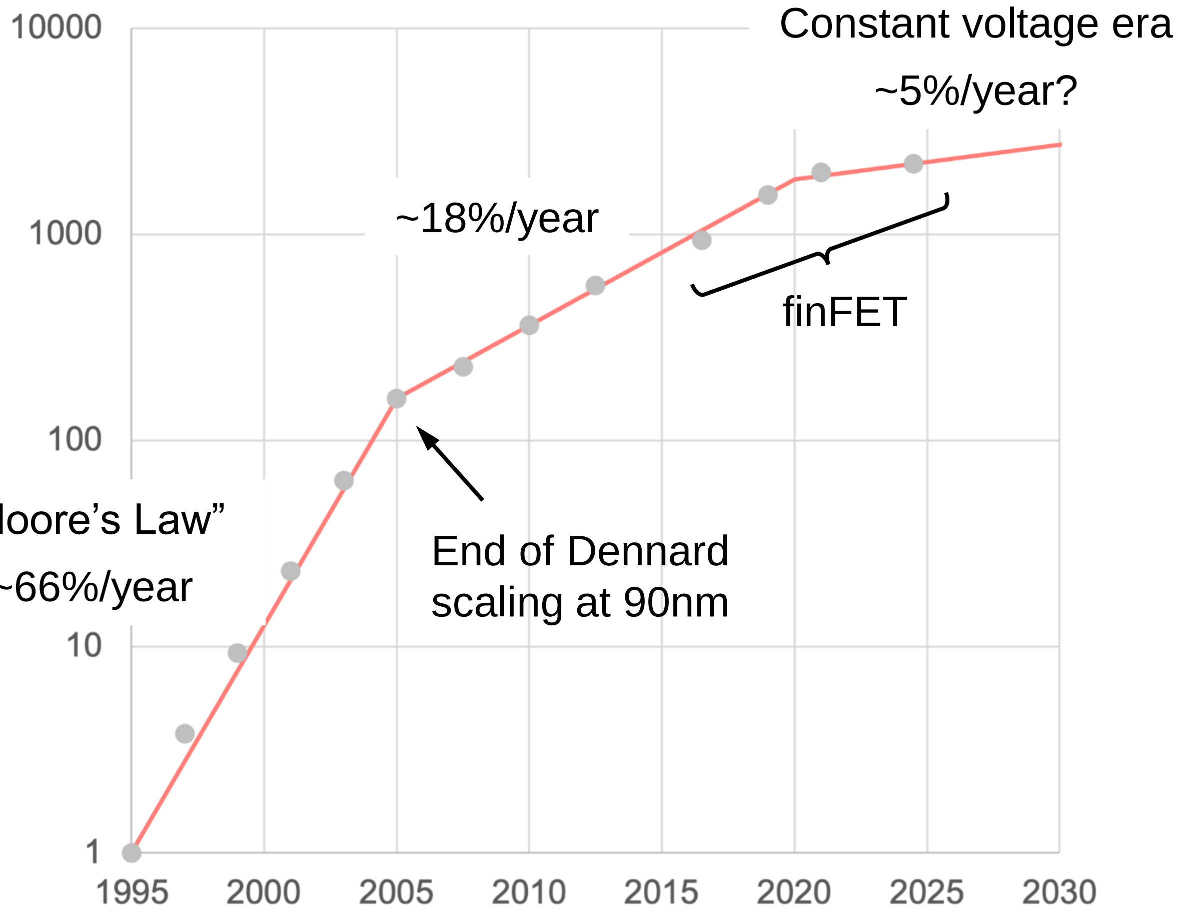
First stacked 3D wafer-on-wafer chip



Logic Wafer + Power Delivery Wafer

Silicon Energy Scaling

Performance per Watt
from a fixed silicon area
(normalized)



Constant voltage era

~5%/year?

~18%/year

finFET

"Moore's Law"

~66%/year

End of Dennard
scaling at 90nm

~year of
product

Large AI computers will need >1pJ per flop for the foreseeable future.

10 Years of GPU Adaptation to AI

~300x in GPU peak arithmetic over the first AI decade:

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

- 16x from matrix multipliers using smaller numbers, fp32 to fp8.
- 8x transistor density, 28nm to 5nm.
- 1.7x clock speed, but 2.8x power, 250W to 700W.
- 1.4x from shifting architecture from graphics to AI.

What Next?

~300x in GPU peak arithmetic over the first AI decade:

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

- 16x from matrix multipliers and smaller numbers, fp32 to fp8. ⬅ Tapped out?
- 8x transistor density, 28nm to 5nm. ⬅ Another 2x, over 5+ years?
- 1.7x clock speed, but 2.8x power, 250W to 700W. ⬅ Another 2x, at 3x power?
- 1.4x from shifting architecture from graphics to AI. ⬅ Room to be more radical?

Total Energy for AI Training

SoTA for “Infrastructure class” AI is ~3pJ/flop, full system, dense isotropic network:

- 1 billion parameters, 20 billion tokens* … 250 chips (100kW) for 1 hour.
- 100 billion parameters, 2 trillion tokens* … 2,500 chips (1MW) for 4 months.
- 1 trillion parameters, 20 trillion tokens* … 25,000 chips (10MW) for 1 year.

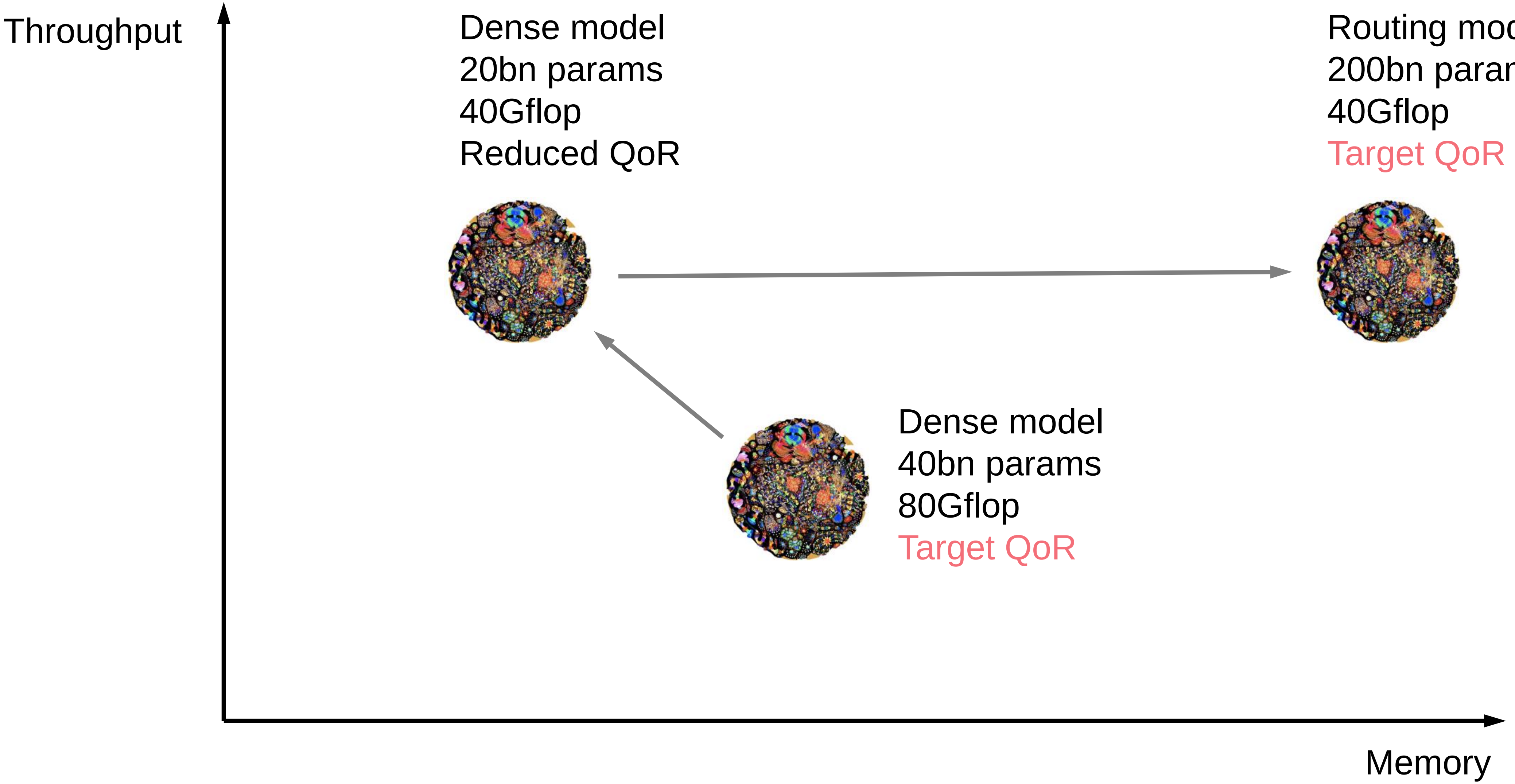
Further progress will require algorithmic innovation

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

Brains do Routing

- In a dense isotropic neural network, such as GPT3, every token interacts with every weight.
- Brains don’t fire all their neurons in response to every stimulus.
- An efficient AI capable of more that one task, or responsive to more than one type of data, must obviously access its “knowledge” selectively.

Trading Memory for Throughput



Throughput

Dense model
20bn params
40Gflop
Reduced QoR

Routing model
200bn params
40Gflop
Target QoR

Dense model
40bn params
80Gflop
Target QoR

Memory

Deep Neural Network Scale Evolution

Cost of data

O(100M) params

2014

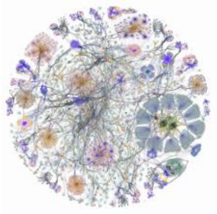
Deep learning

eg. ResNet

1000x

Cost of compute

O(100G) params



2018

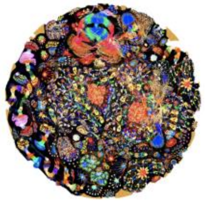
Unsupervised learning

eg. BERT, GPT

1000x

Cost of memory

O(100T) params?

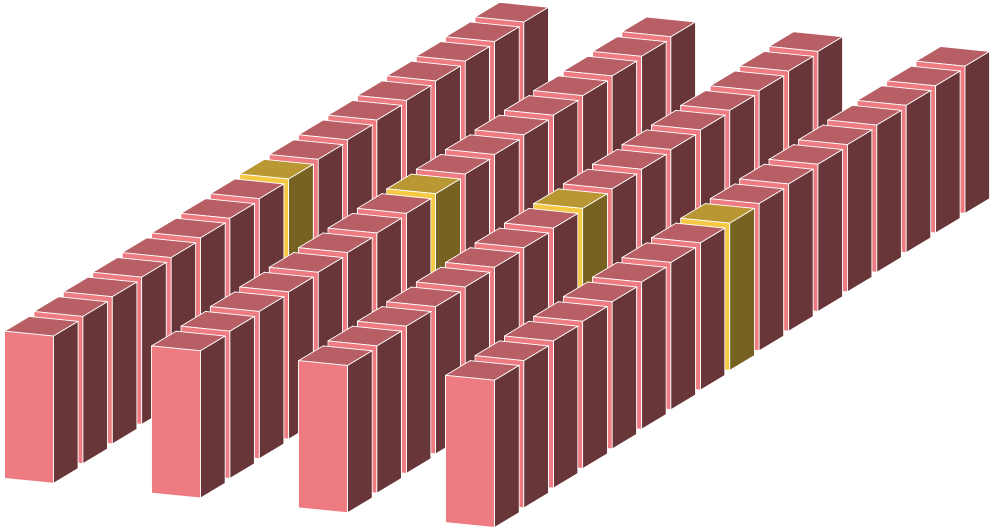


2022

Routing networks

eg. SGMoE, Switch

Practical ”Brain-Scale” Computing



GRAPHCORE Good Computer [mid-size]

- 2048 Mk3 IPUs ~ 1 real Eflop_{16}/s
- 1PB DRAM at > 2 real PB/s
- ~$50m, 2.5MWatts, 68 standard racks, 100m^2

Take-Aways

- Silicon scaling is almost done, especially energy per op.
- Brain-scale dense neural networks are infeasible; larger models must be sparse.
- Extremely sparse routed brain-scale neural networks will be feasible soon.
- The arrival of AI at this ending of “Moore’s Law” demands a new era of algorithm and architecture co-innovation.