Computing for Super-Human Cognition

TERATEC Forum 2022

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 Machines

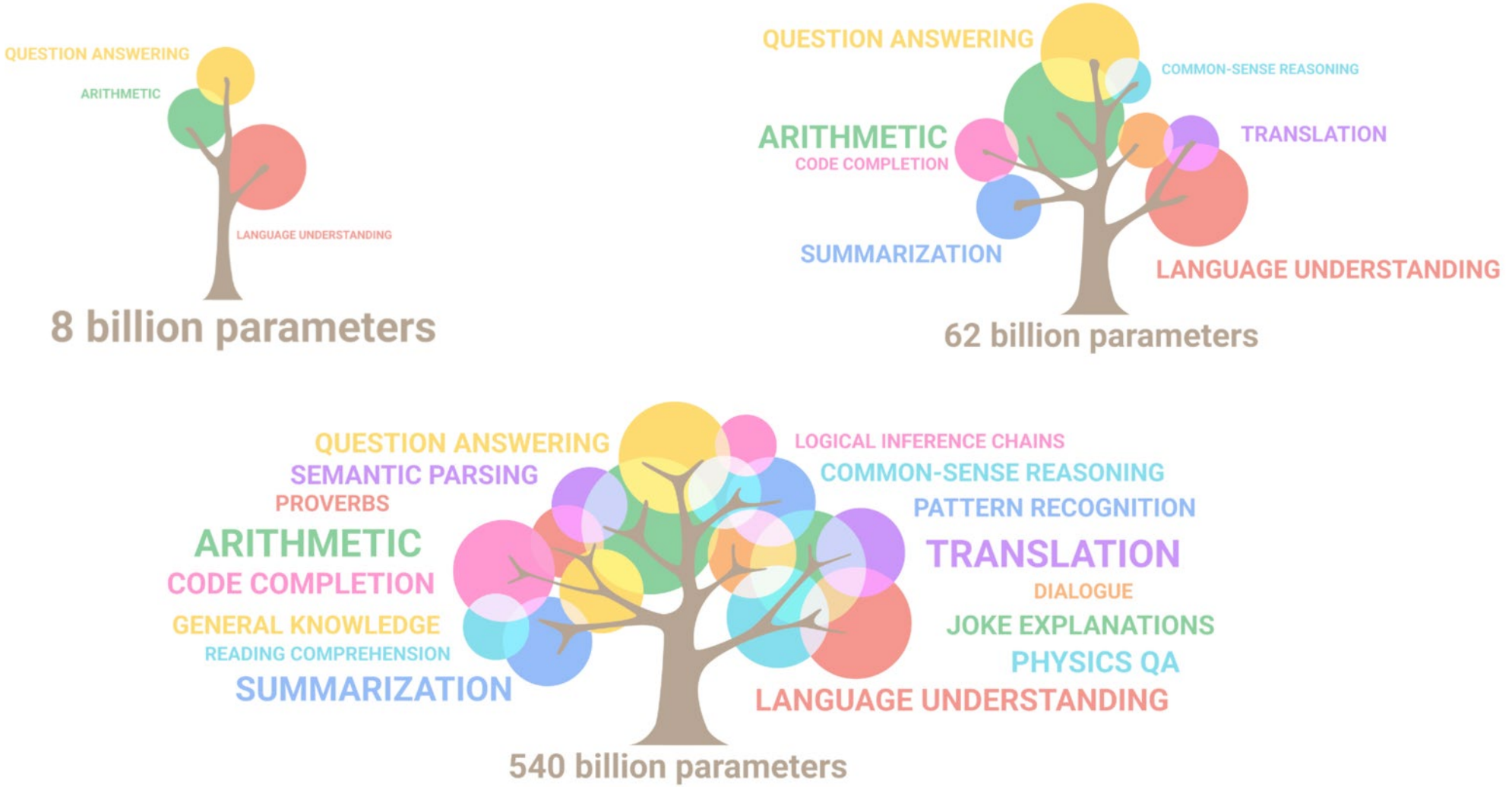
Capacity machines : cheaper than a human, per unit of work.

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

effectiveness of data representations
effectiveness of training and inference processes
} Brain design

quantity and quality of training data
model scale (#parameters)
} Educational effort

Intelligent Capabilities Emerge with Model Scale



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

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.
- AI can specialize to “intellectual activities” more than a human.

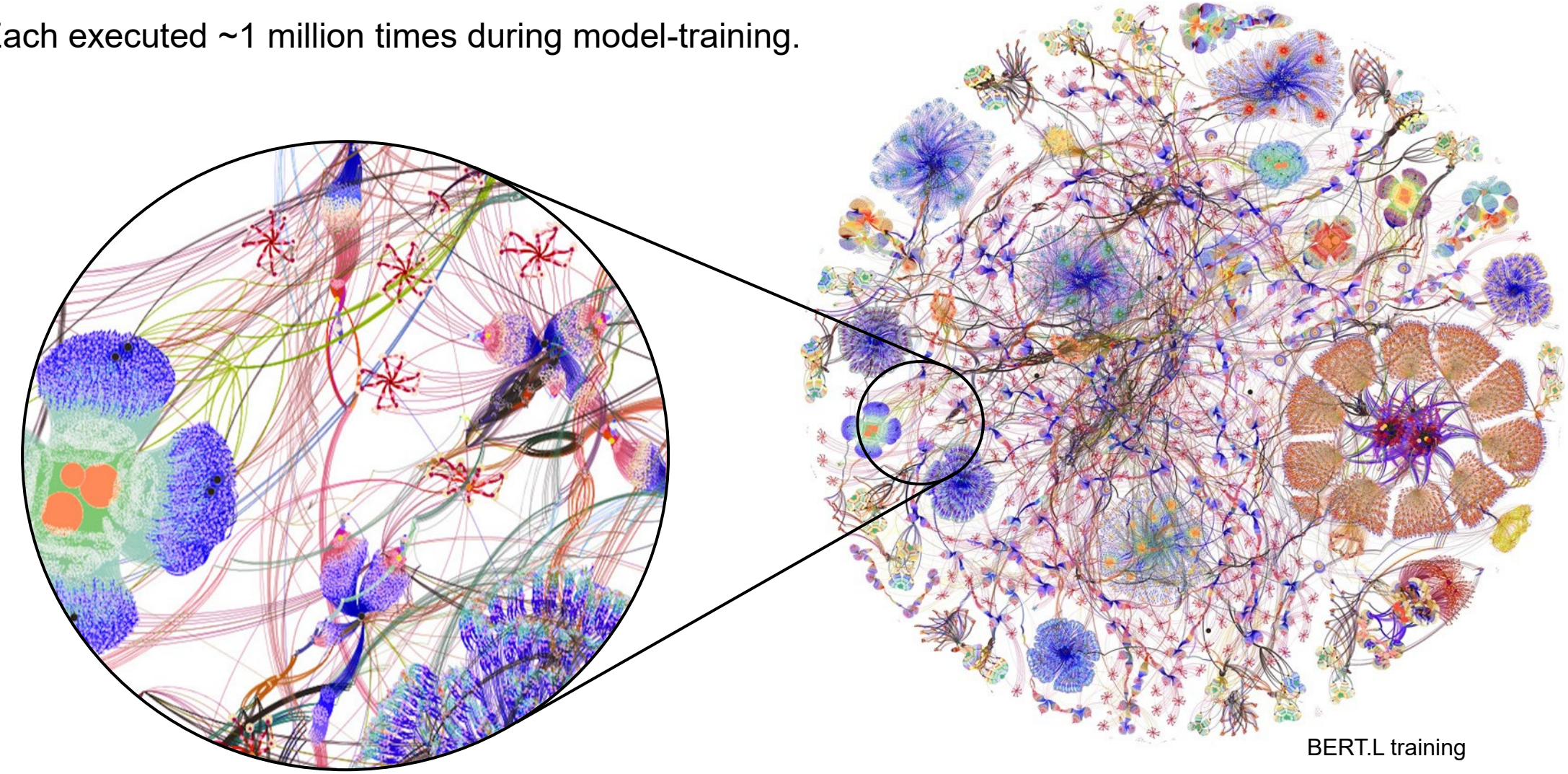
=> Ultra-intelligence might require less than 100TB of learned state?

(1) Wikipedia.org/wiki/Neuron

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

AI computing today, software view…

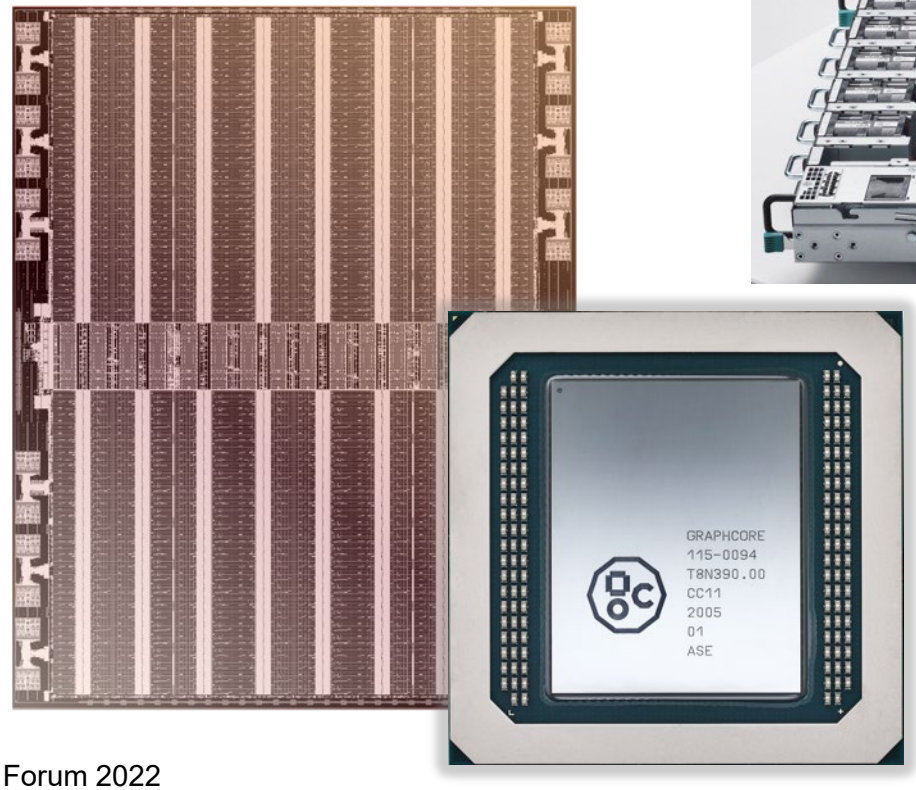
- Millions-to-billions of inter-communicating C++ programs.
- Each executed ~1 million times during model-training.

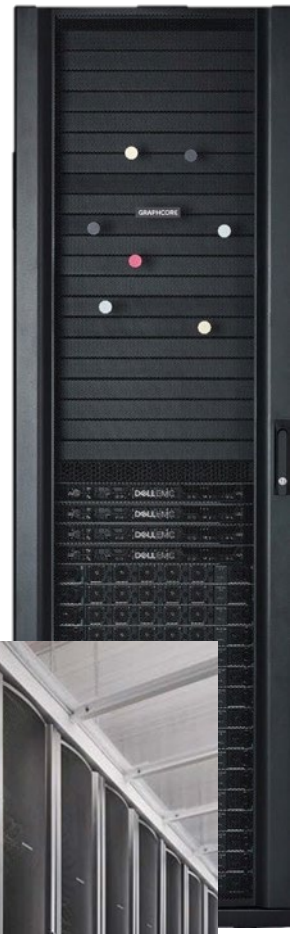


BERT.L training

AI computing today, hardware view…

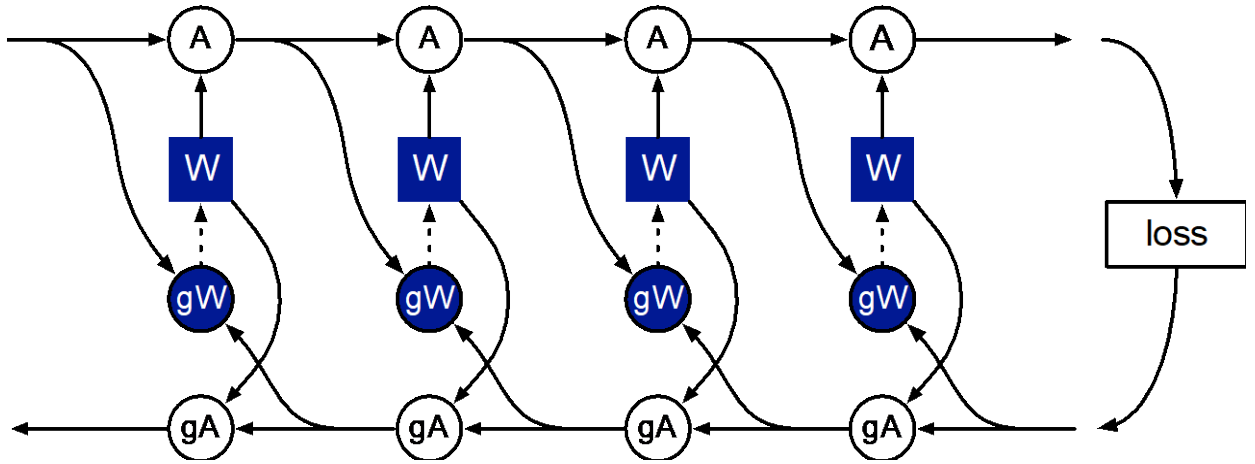
- Trillions of transistors applied to one training run.
- Tens-to-thousands of chips, thousand-to-millions of Watts.
- Hours-to-months of training program run time.





The Master Learning Algorithm of AI

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



- ~1PB memory required for “brain scale” 100 trillion parameters.
- ~2PB/s memory bandwidth required for 1 million SGD iterations in 2 weeks.

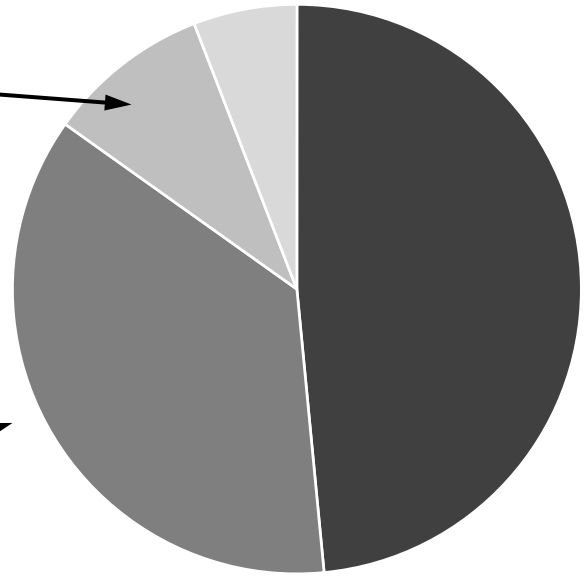
Neural Network Compute Scale-Up

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

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

1.4x from re-tuning graphics architecture to AI.

1.7x clock speed, but at
2.8x power, 250W to 700W.



16x from matrix multipliers and
smaller floats, fp32 to fp8.

8x transistor density,
from 28nm to 5nm.

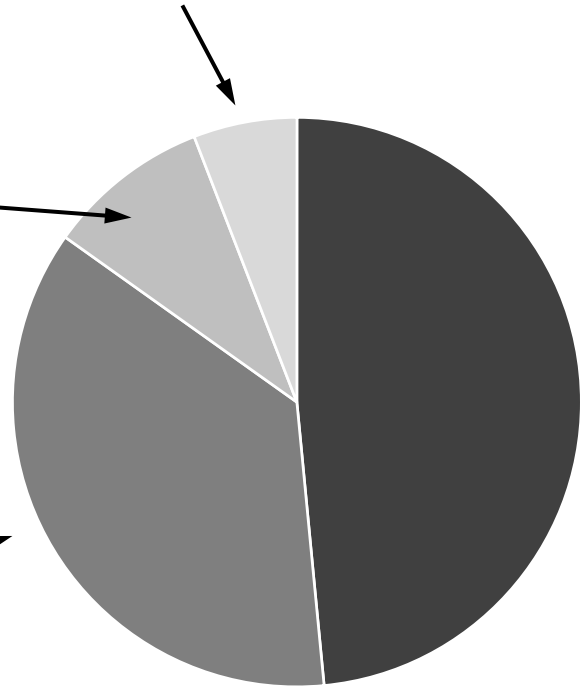
The Second AI Decade?

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

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

1.4x from re-tuning graphics architecture to AI.

More from ground-up AI architectures



1.7x clock speed, but at 2.8x power, 250W to 700W.

Another 2x, at 3x power?

16x from matrix multipliers and smaller floats, fp32 to fp8.

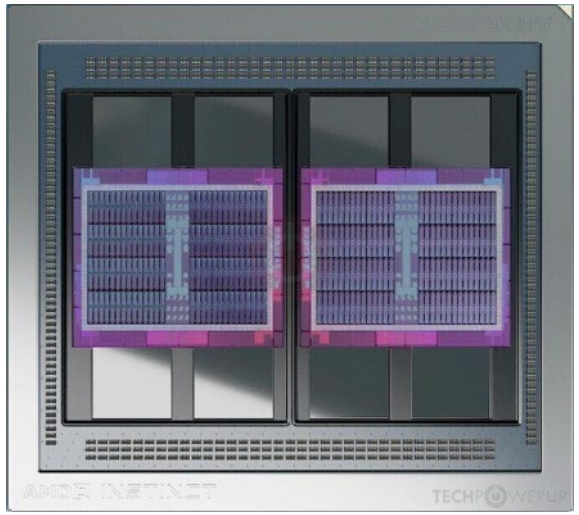
Done?

8x transistor density, from 28nm to 5nm.

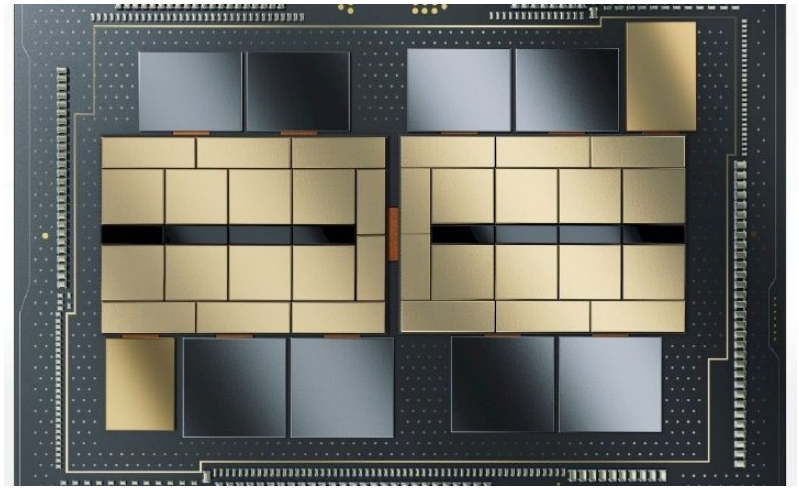
Another 2-3x?

Multi-die integration is replacing die density scaling

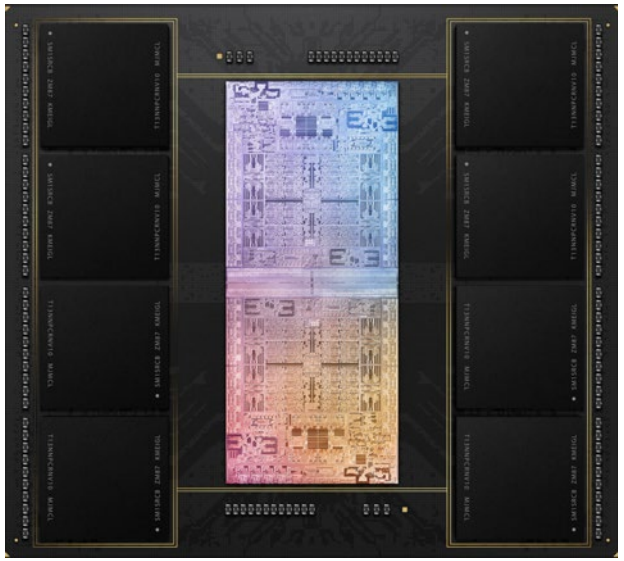
AMD MI250X: inter-CoWoS buried bridge



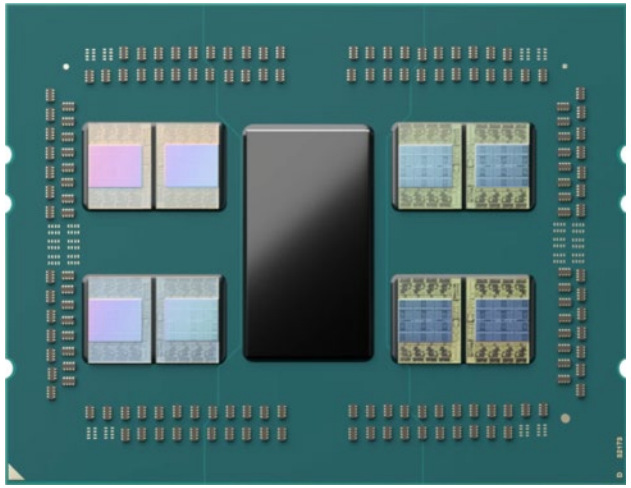
Intel Ponte Vecchio: 42-die on 2 interposers



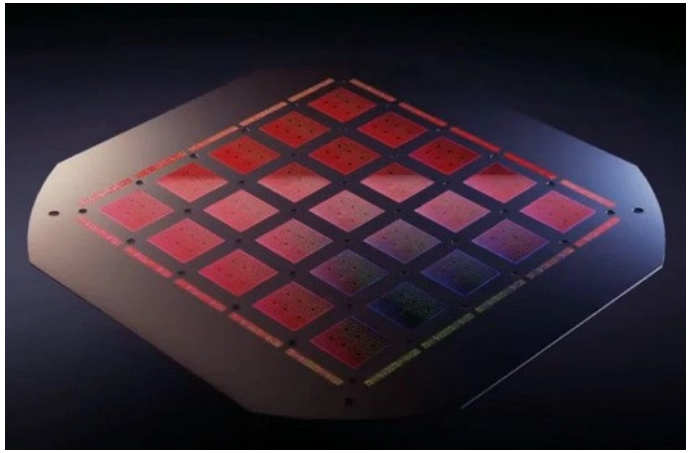
Apple M1-Ultra: LPDDR5 on substrate



AMD Milan-X: Chip-on-Wafer caches



Tesla D100 wafer-scale InFO



Graphcore: Wafer-on-Wafer

Dense Neural Network Energy Scaling

SoTA training for dense isotropic transformers using AI supercomputers:

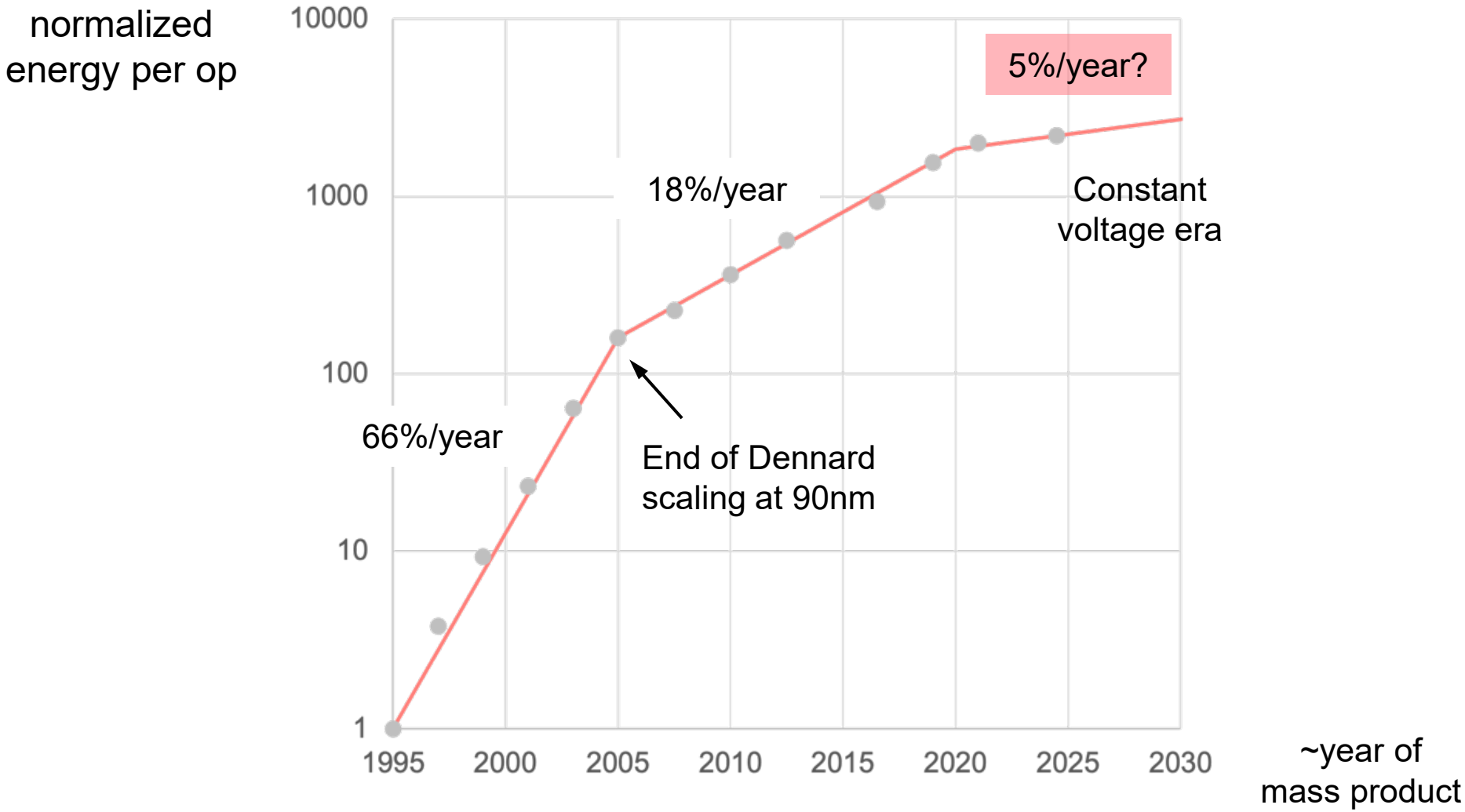
- 100 billion params ~ 10^{24} flops @ 3pJ/flop, 500W/chip … 4k chips, 2MW, 2.5 weeks.

Extrapolate to “brain scale” using next-gen chips:

- 100 trillion params ~ 10^{30} flops* @ 2pJ/flop, 1000W/chip … 64M chips, 64GW, 1 year.

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

Silicon Energy Scaling



…fortunately, Brains do Routing

- 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 stored “knowledge” selectively.

Model Scale Breakthroughs

Cost of data

O(100M) params

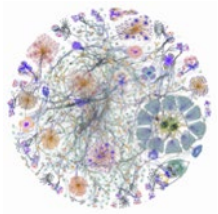
~2014

Deep learning

1000x

Cost of compute

O(100G) params



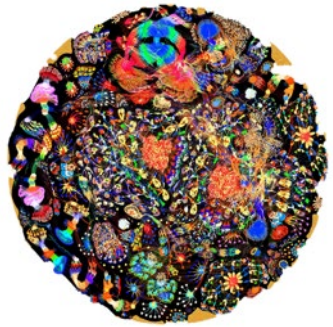
~2018

Unsupervised learning

1000x?

Cost of memory

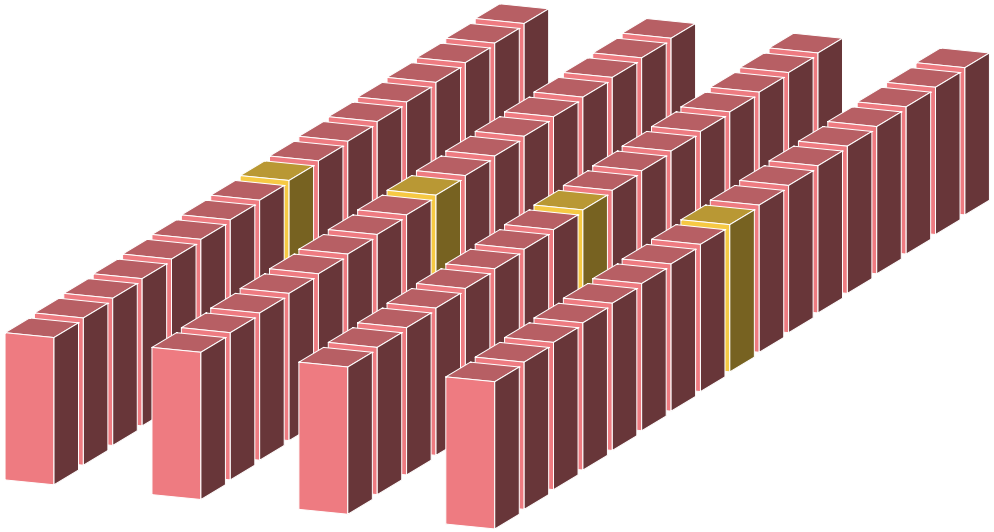
O(100T) params?



~2022

Routing networks

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; large AIs will all be sparse.
- Extremely-sparse routed brain-scale neural networks are feasible now.
- The arrival of AI at this ending of “Moore’s Law” demands a new era of algorithm and architecture co-innovation.