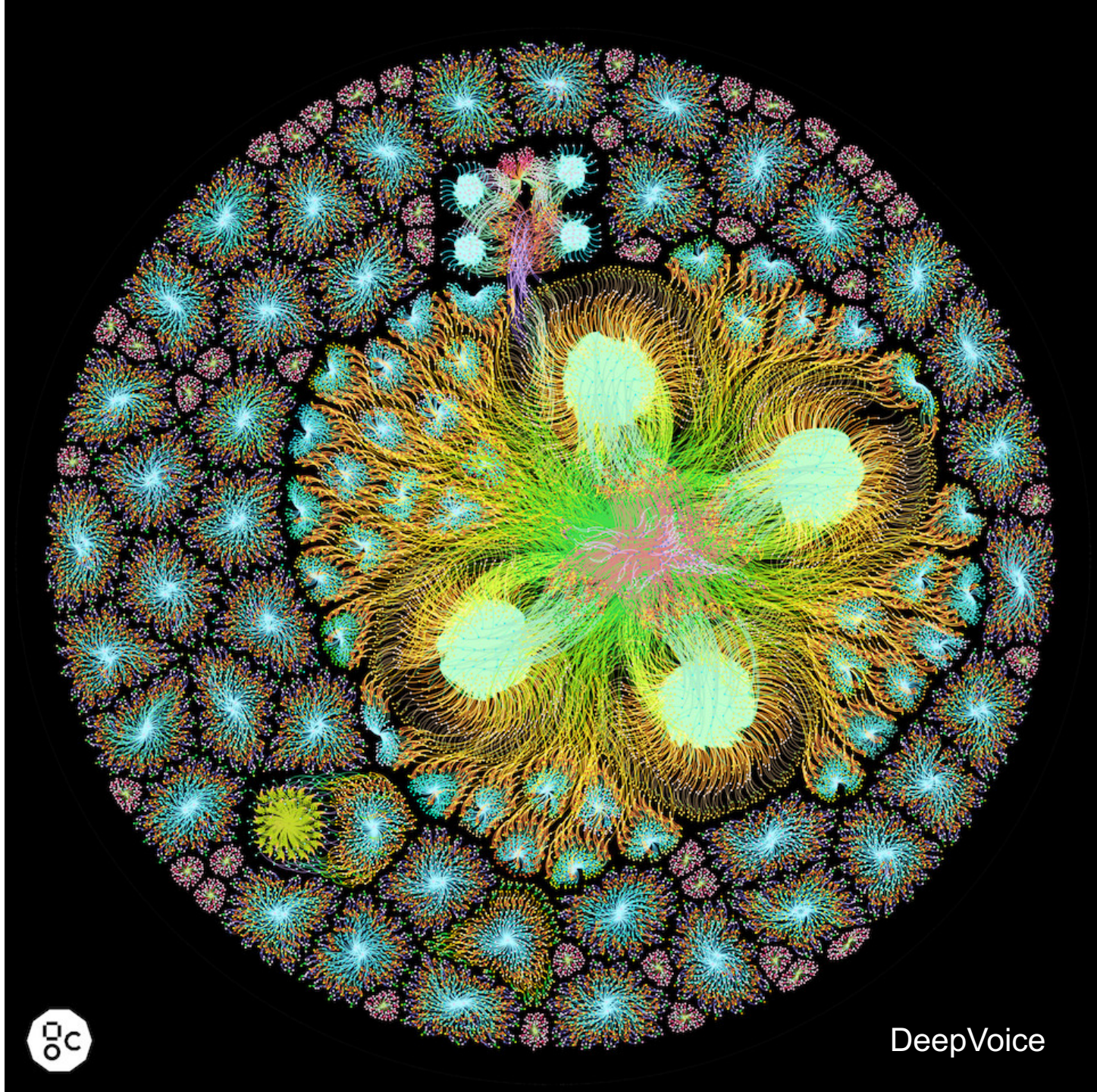
GRAPHCORE

Designing Processors and Systems for Intelligence

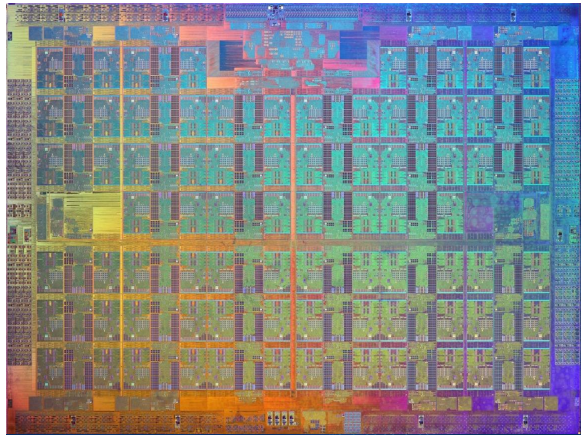
Ola Toerudbakken, SVP Rack Scale Systems

HiPINEB 2018

Vienna, 24th February 2018



INTELLIGENCE IS A NEW WORKLOAD

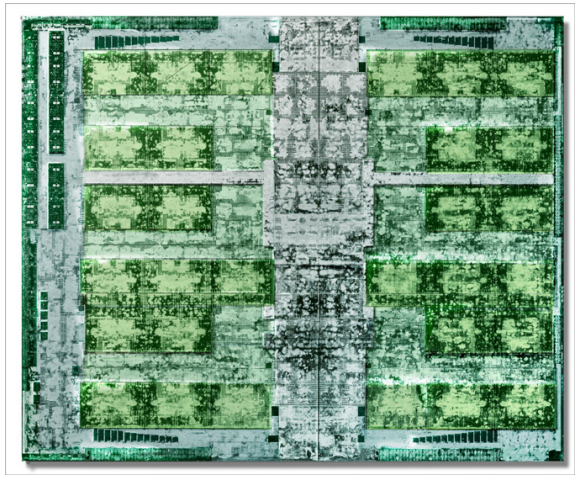


CPU

Scalar

Designed for office apps

Evolved for web servers

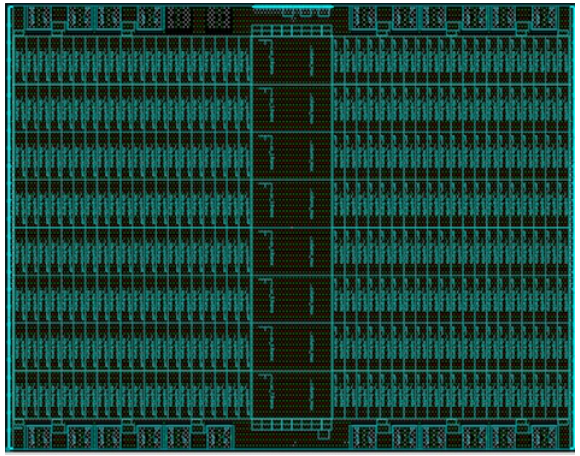


GPU

Vector

Designed for graphics

Evolved for HPC



IPU

Graph

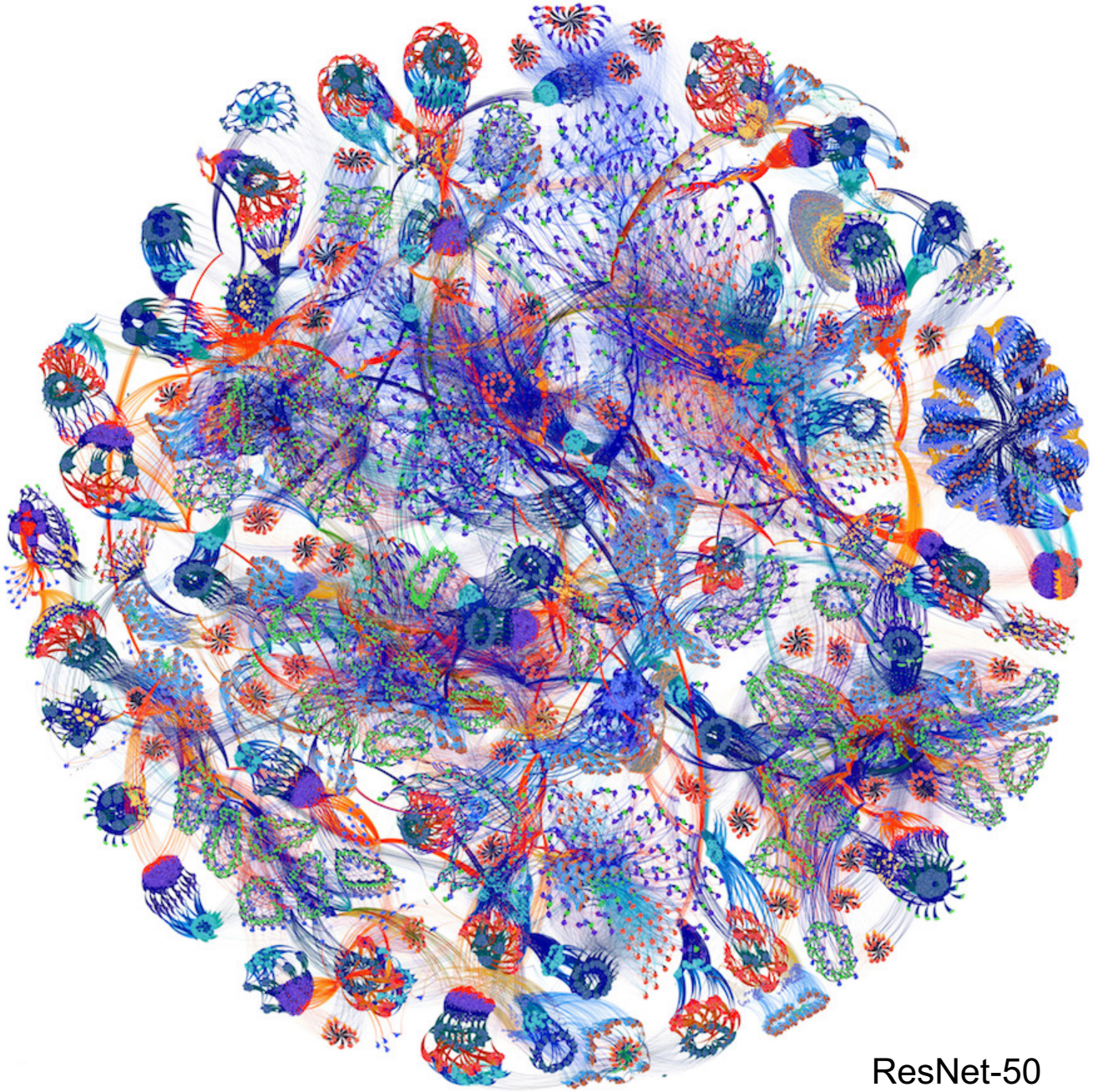
Designed for intelligence

Knowledge models and inference algorithms are naturally and jointly representable as graphs…

- Vertices contain parts code & state.
- Edges pass data.
- Graph structure is static.

Graphs expose a lot of parallelism...

O(1000) work items/processor
x O(1000) processors/chip
x O(1000) chips/system

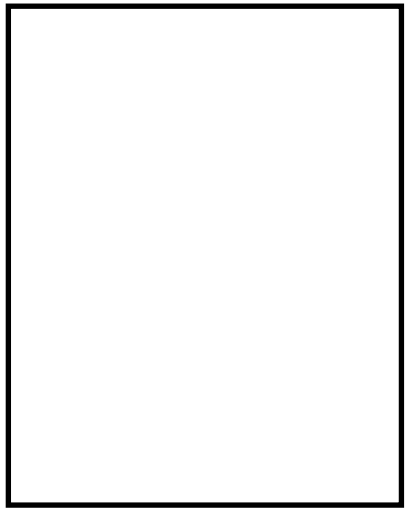


ResNet-50
training batch=4

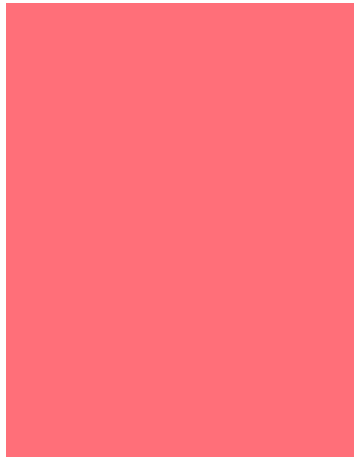
INTELLIGENCE MACHINE CHARACTERISTICS

- Computation on graphs
  - massive parallelism, sparse, high-dimensional models
  - distributed memory
- Low precision, wide dynamic range arithmetic
  - mixed-precision float 16.32 (and smaller?)
- Static graph structure
  - compiler can partition work, allocate memory, and schedule messages
  - bulk-synchronized, un-ordered, address-less communication
- Massive Communication
  - e.g. human brain ~¼ neurons and ~¾ synapses
- Entropy generative
  - noise in hardware (learning is training + exploring; compute & estimate gradients)

ALL LOGIC CHIPS ARE POWER LIMITED



Largest manufacturable die
~825mm^2

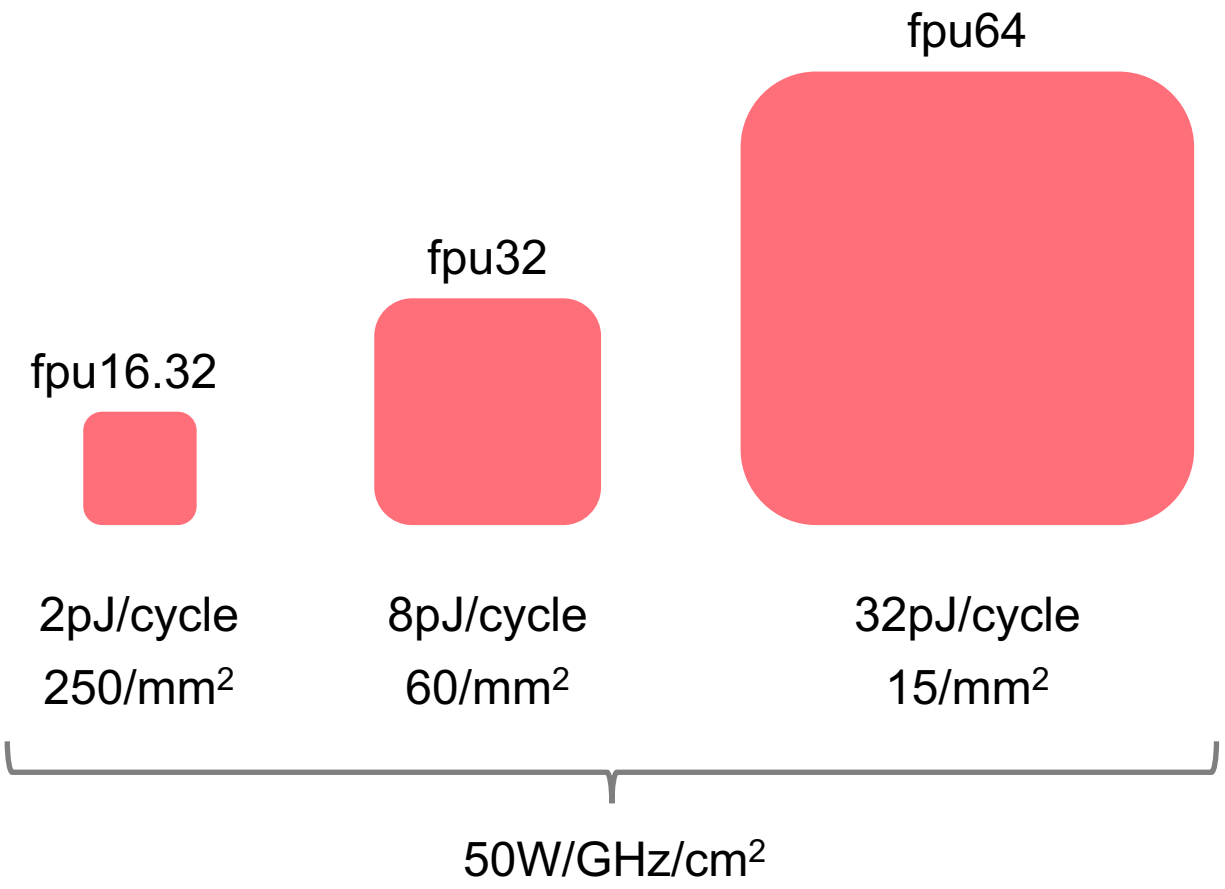


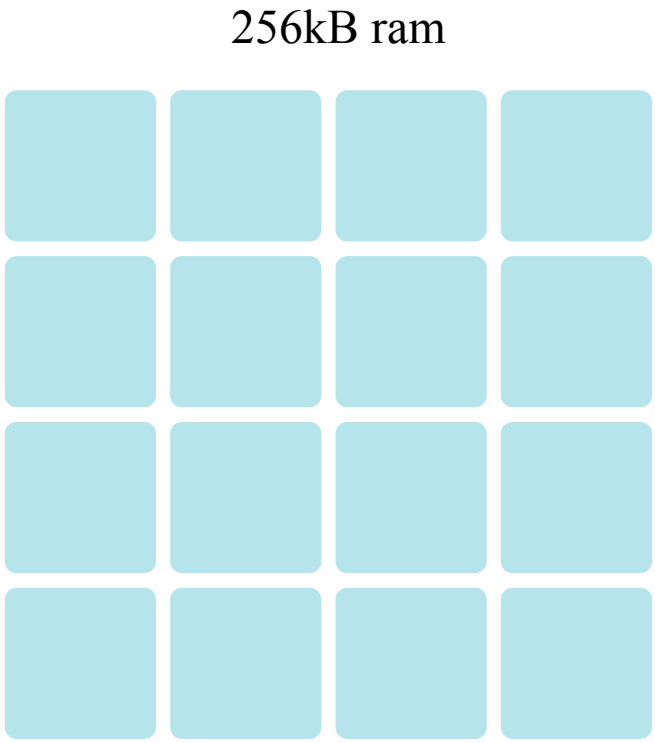
16nm 1Pflop16.32
<700mm^2

1000W

8cm^2 die @ 200W

POWER DENSITY





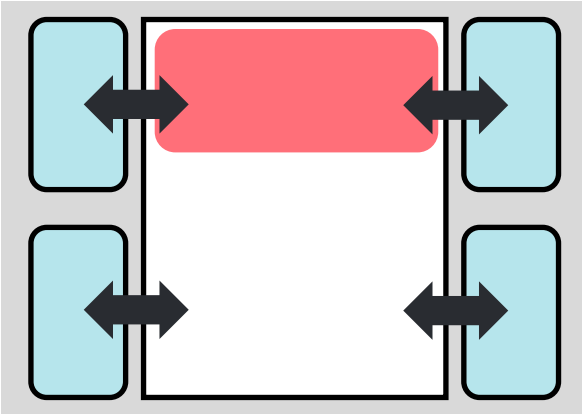
3 accesses x 8pJ/cycle

~1/5 power density of logic

MEMORY BANDWIDTH @ 240W

DRAM on interposer

180W GPU + 60W HBM2

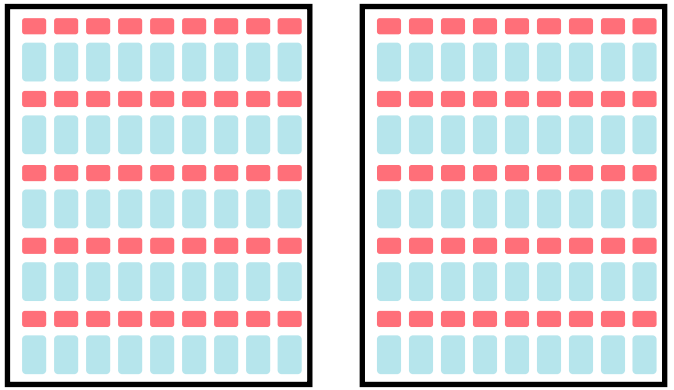


16GB @ 64pJ/B

900GB/s

Distributed SRAM on chip

2x IPU (75W logic + 45W ram)

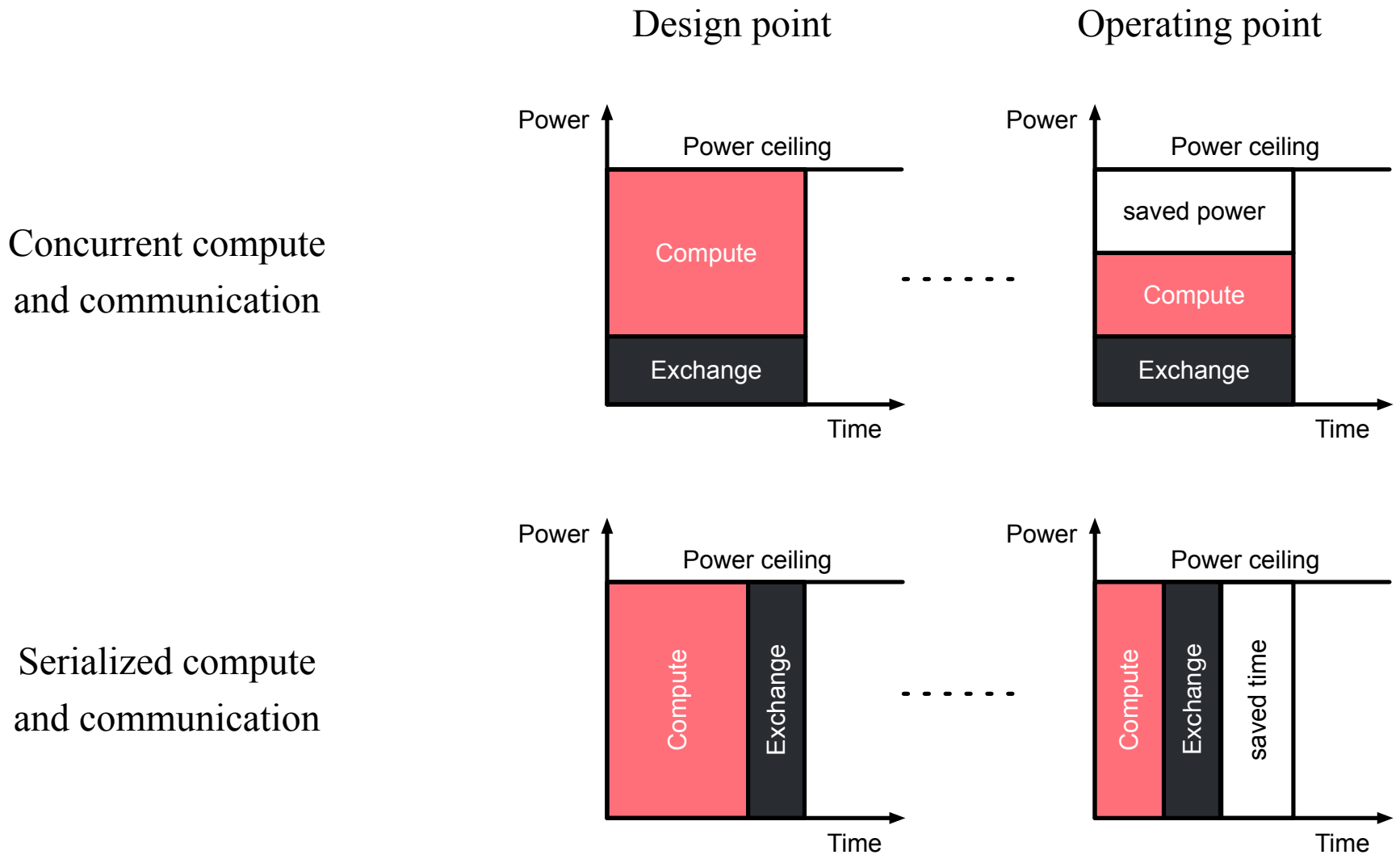


600MB @ 1pJ/B

90,000GB/s

100x

SERIALIZE COMPUTE AND COMMUNICATION



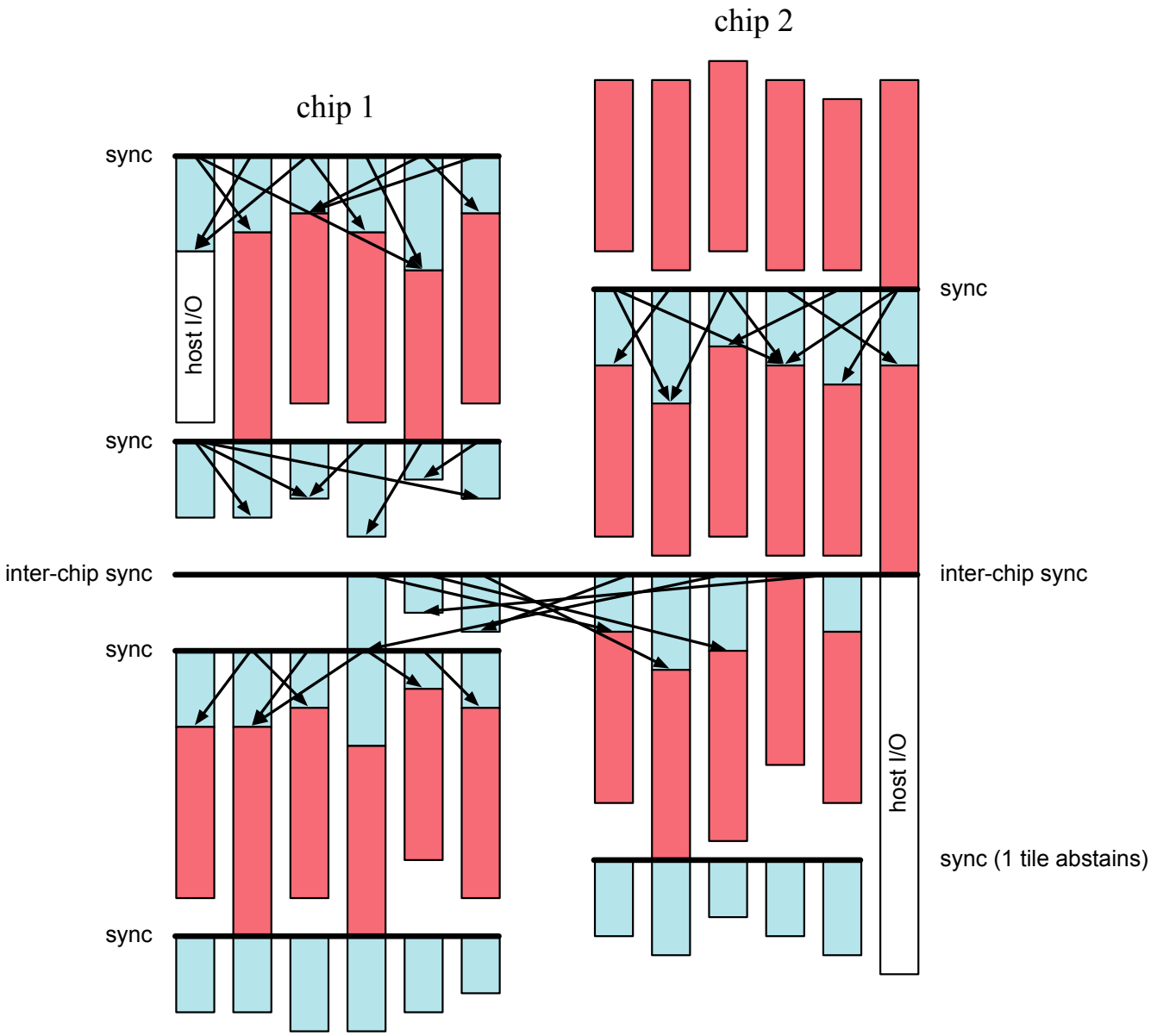
BULK SYNCHRONOUS PARALLEL

On-Chip & Inter-Chip.

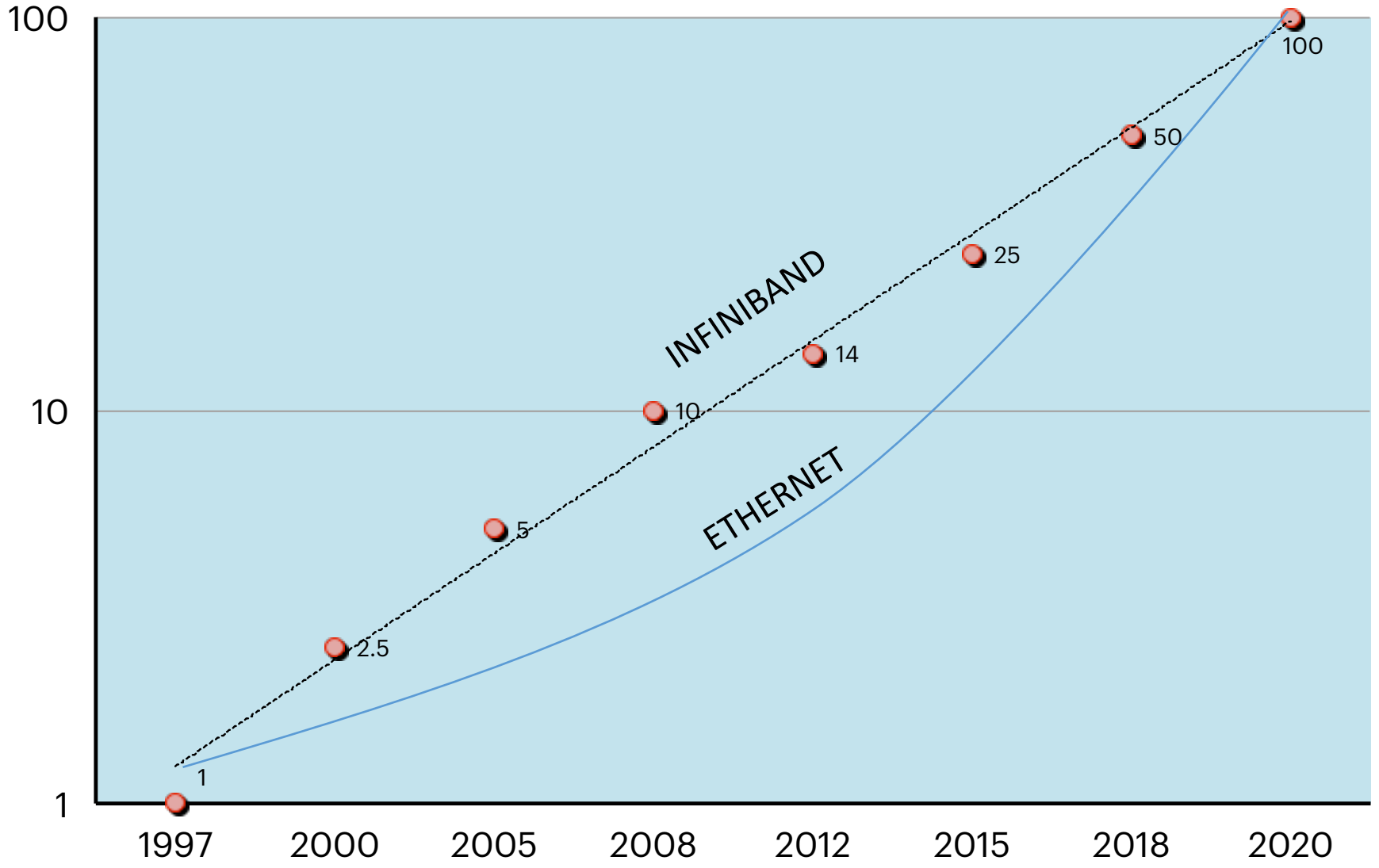
Massive parallelism with
no concurrency hazards.

compute phase

exchange phase



INTERCONNECT SPEED EVOLUTION

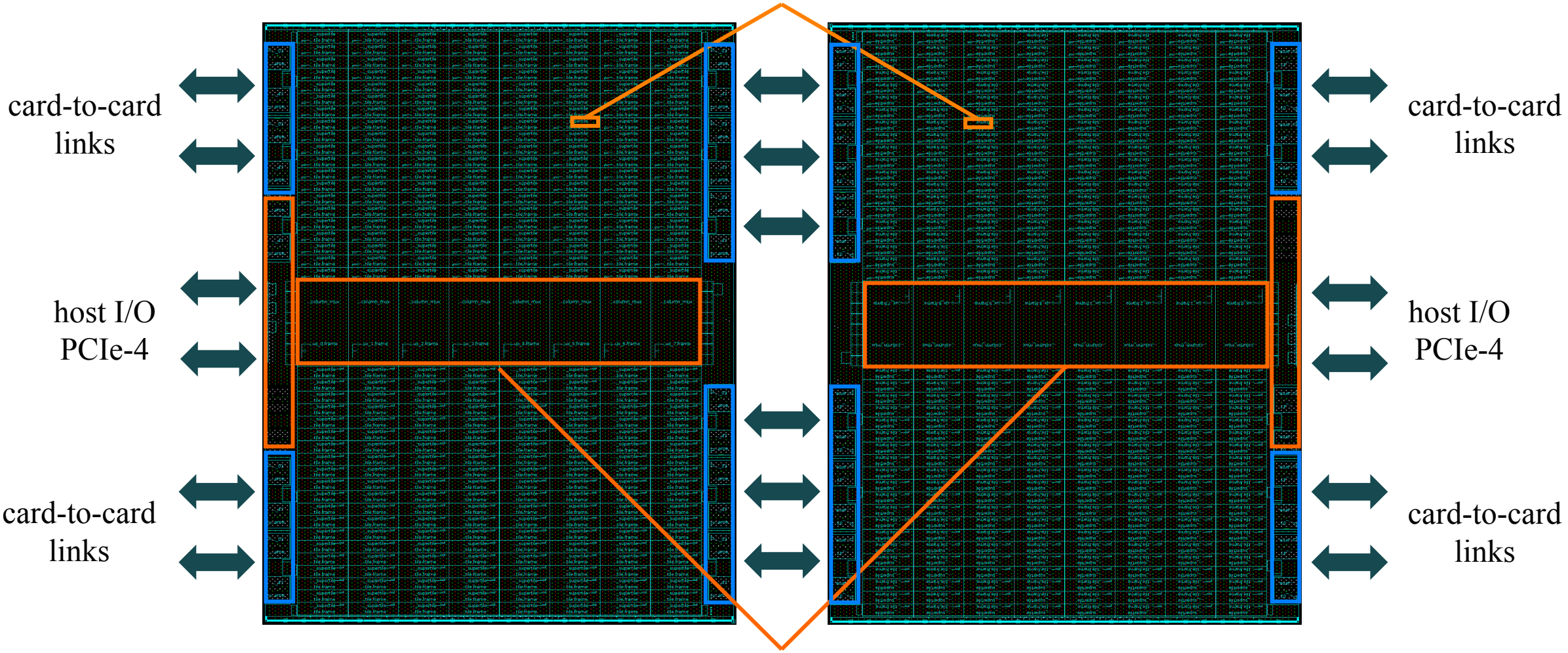


NETWORK FOR MI:

Latency
Collectives
Massive Connectivity
+
High BW Data Access

GRAPHCORE IPU

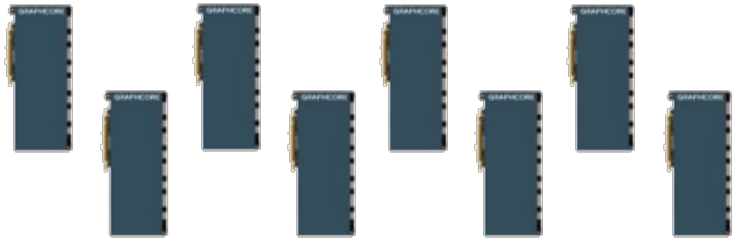
>2000 processor tiles >200Tflop ~600MB



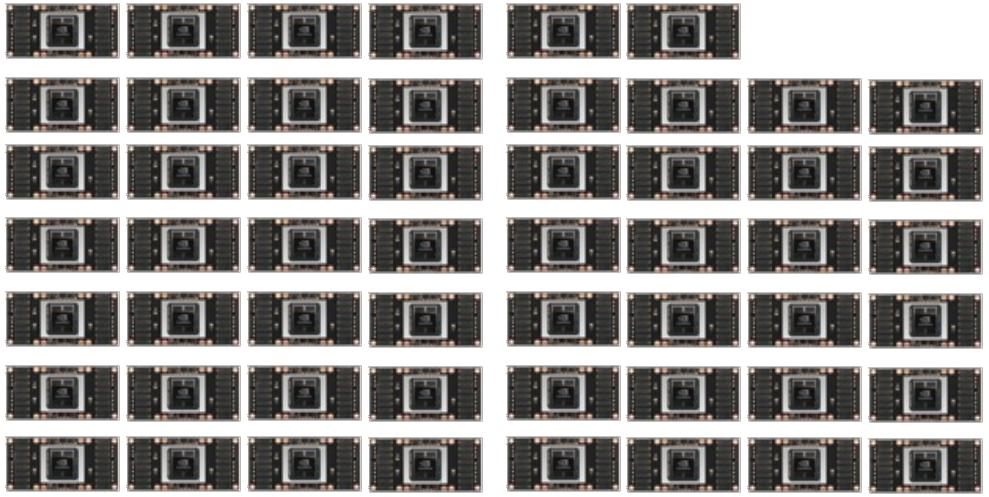
all-to-all exchange spines each ~8TBps

I/O bandwidth 384GBps/chip

Benchmark: ResNet-50 ImageNet Training at 16,000 Images/Sec



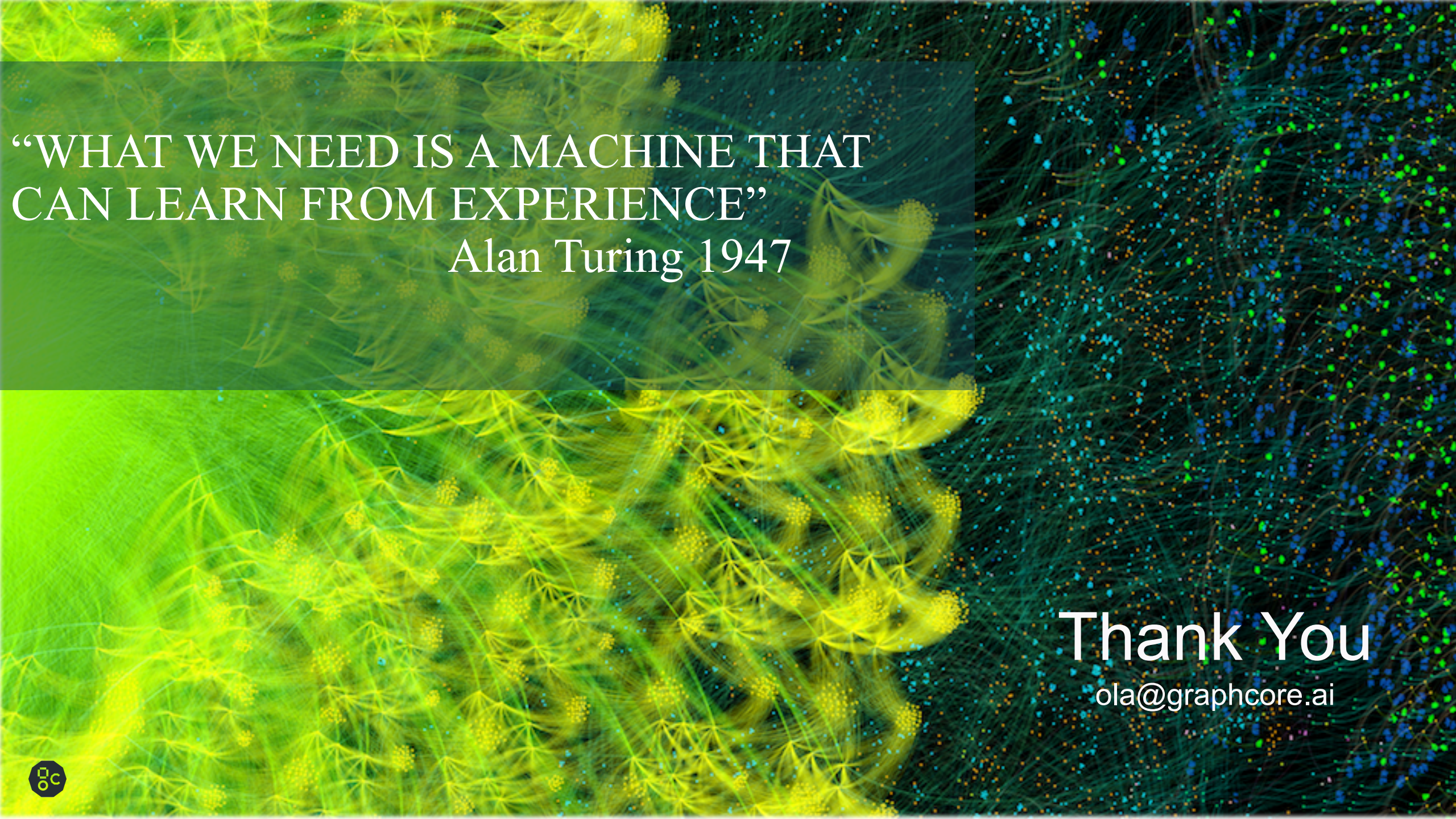
8x C2 IPU Accelerator PCIe cards



54x NVIDIA Volta V100

Scaled 1/2.4x from 128x Pascal P100

Source for P100: Facebook https://arxiv.org/abs/1706.02677
Source for P100 to V100 scaling: Nvidia


“WHAT WE NEED IS A MACHINE THAT CAN LEARN FROM EXPERIENCE”
Alan Turing 1947
Thank You
ola@graphcore.ai