Post-Moore AI Infrastructure

Babak Falsafi

Collaboration w/ Louis Coulon, Mario Drumond, Simla Harma, Martin Jaggi, Tao Lin, Yunho Oh, Canberk Sönmez and the EcoCloud Community

ecocloud

an EPFL research center

ecocloud.ch

EPFL

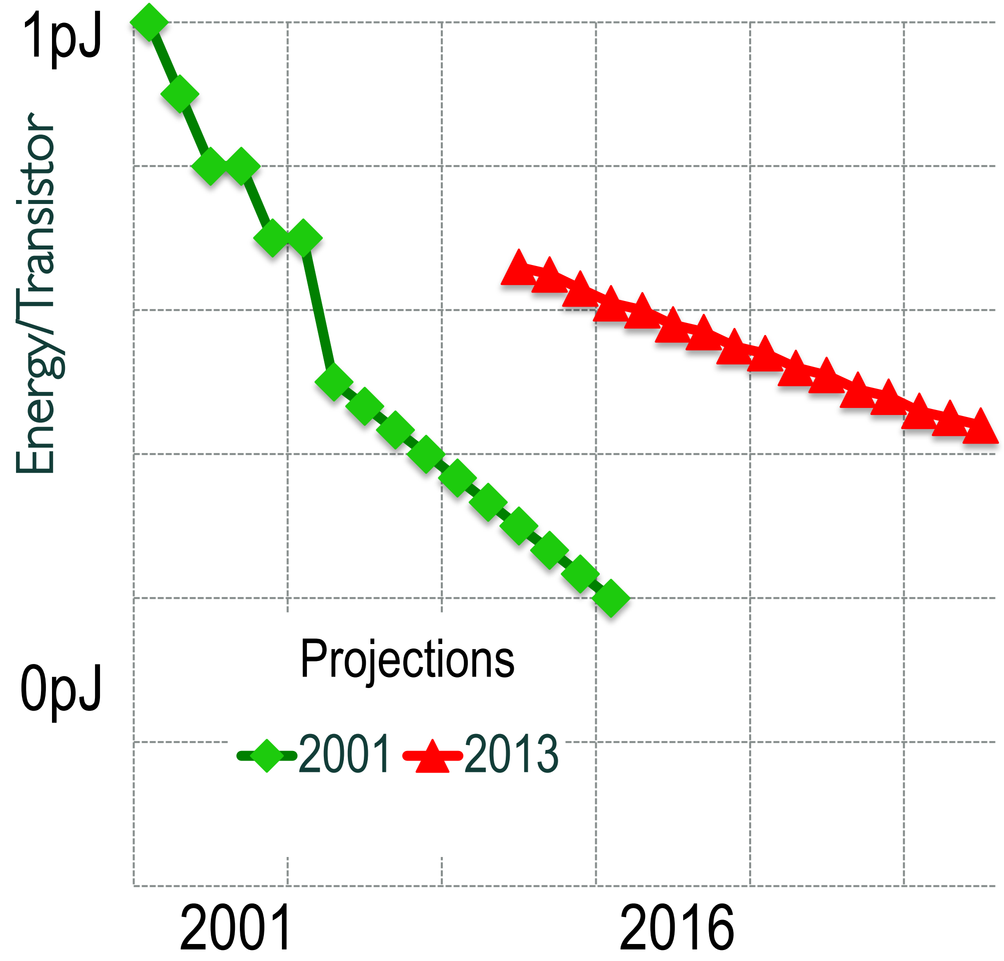
Modern Datacenters:
The Backbone of Cloud

- A million home-brewed servers
- Centralized to exploit economies of scale
- Network fabric w/ μ-second connectivity
- At physical limits
- Need sources for
  - Electricity
  - Network
  - Cooling

250MW, Bedford

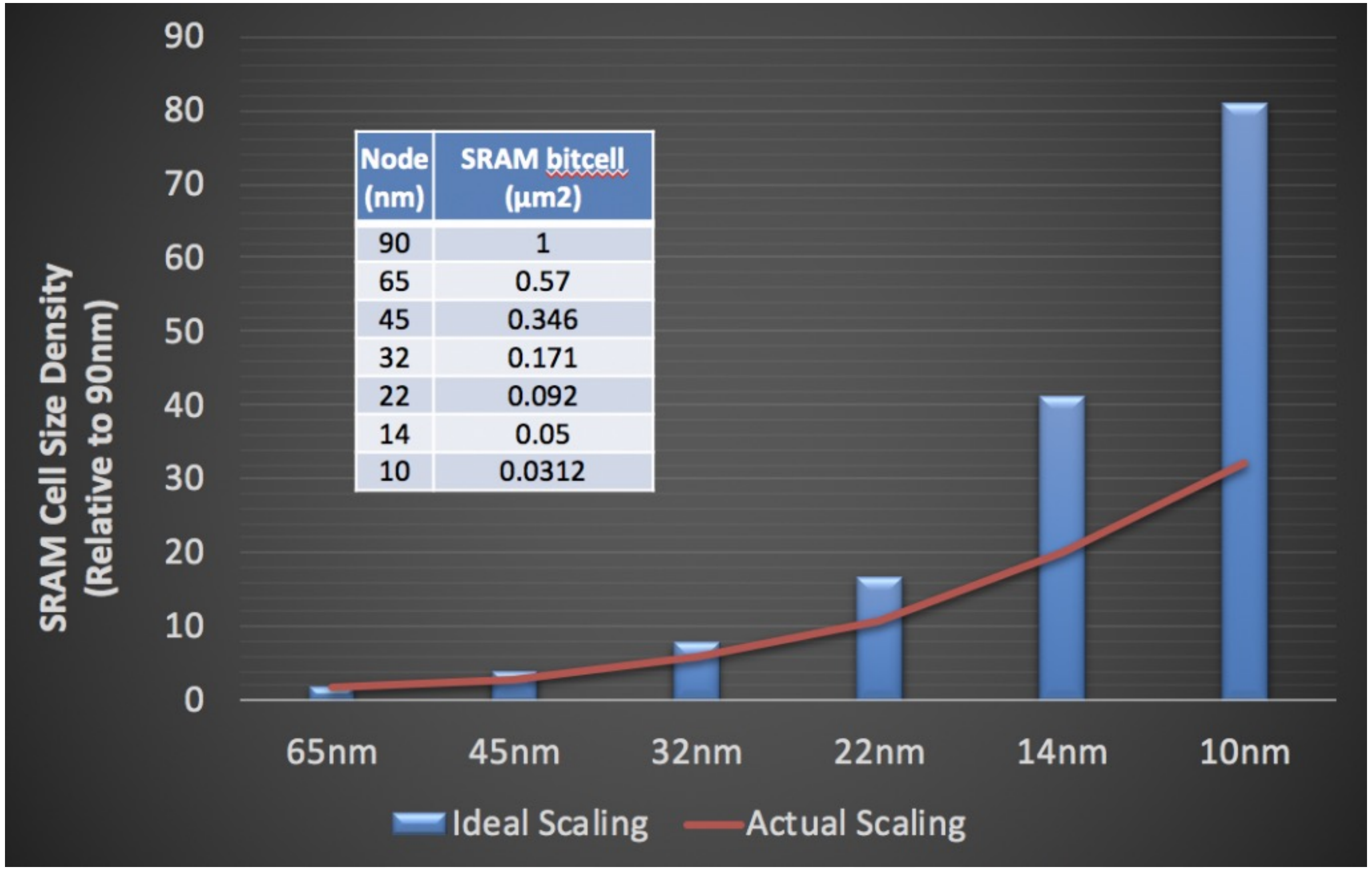
But, Silicon out of steam!

Silicon efficiency is dead (long live efficient silicon)



Moore's Law dying

[David Brooks, SIGARCH'18]



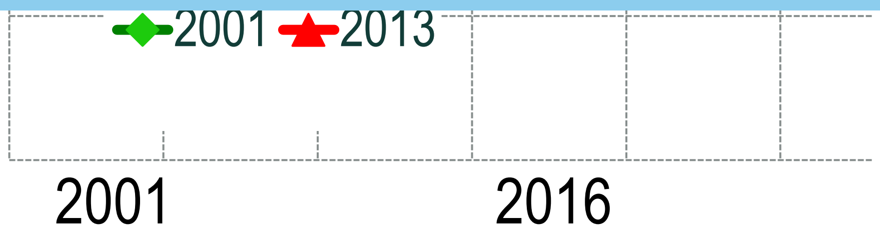
Node (nm)	SRAM bitcell (µm2)
90	1
65	0.57
45	0.346
32	0.171
22	0.092
14	0.05
10	0.0312

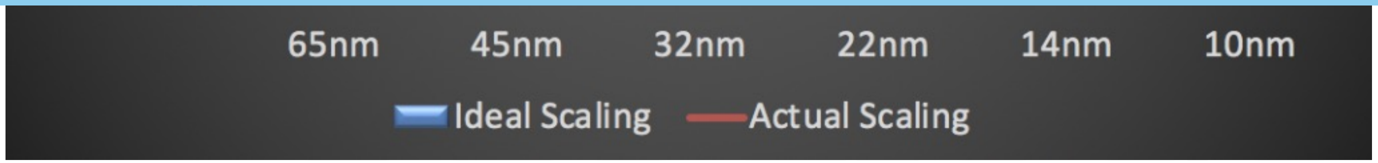
But, Silicon out of steam!

Silicon efficiency is dead
(long live efficient silicon)

Moore's Law dying
[David Brooks, CAT'18]

Conventional scaling 41%/year
Recent years 15%/year!
[David Brooks, Computer Architecture Today]



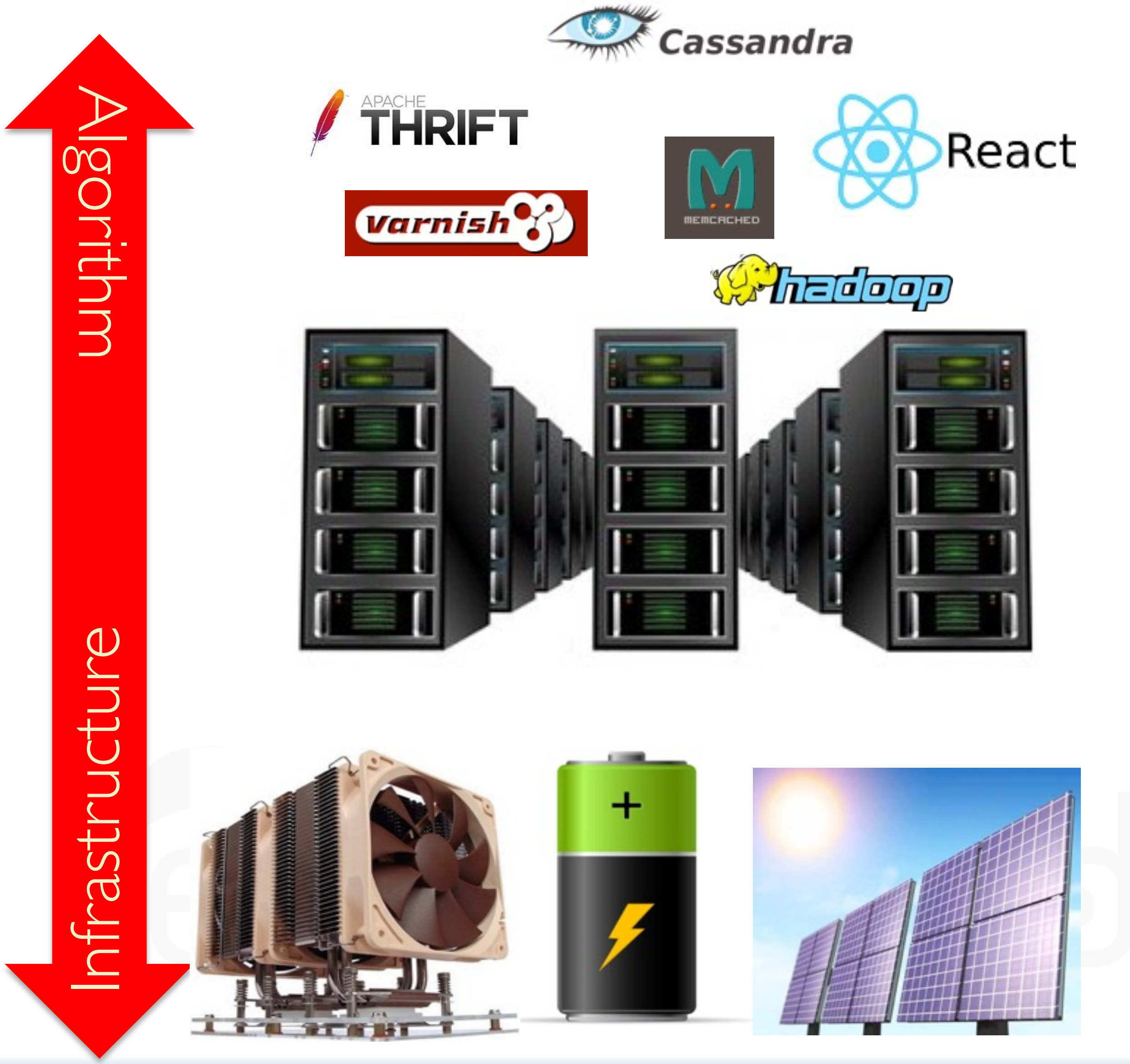


The future of Digital Platforms: Cross-Stack Optimization

ecocloud
an EPFL research center

ISA opportunities

- Integration
  - Move less frequently
  - Move less distance
- Specialization
  - Customize work
  - Less work/computation
- Approximation
  - Adjust precision



ecocloud

Decarbonizing datacenters:

- Center at EPFL founded in 2011
- 21 faculty, 100+ researchers

Holistic datacenter design:

- Minimizing electricity in IT services
- Post-Moore server design
- Integrated cooling, renewables
- From algorithms to infrastructure

Microsoft

HUAWEI

AXA

IMO Ventures

Hewlett Packard Enterprise

intel

IBM

facebook

ORACLE

EPFL

ecocloud.ch

DATACENTER EFFICIENCY LABEL

sdea.ch

IT INFRASTRUCTURE EFFICIENCY

- compute, storage, network and workloads

DC INFRASTRUCTURE EFFICIENCY

- electrical, cooling and heat recycling components

DC CARBON FOOTPRINT

- energy efficiency and sustainability of the electricity sources

DATA CENTER EFFICIENCY

Sign up for our newsletter

EPFL ecocloud Hewlett Packard Enterprise asut VIGISWISS suisse énergie Notre engagement : notre futur. Green IT Switzerland HOCHSCHULE LUZERN

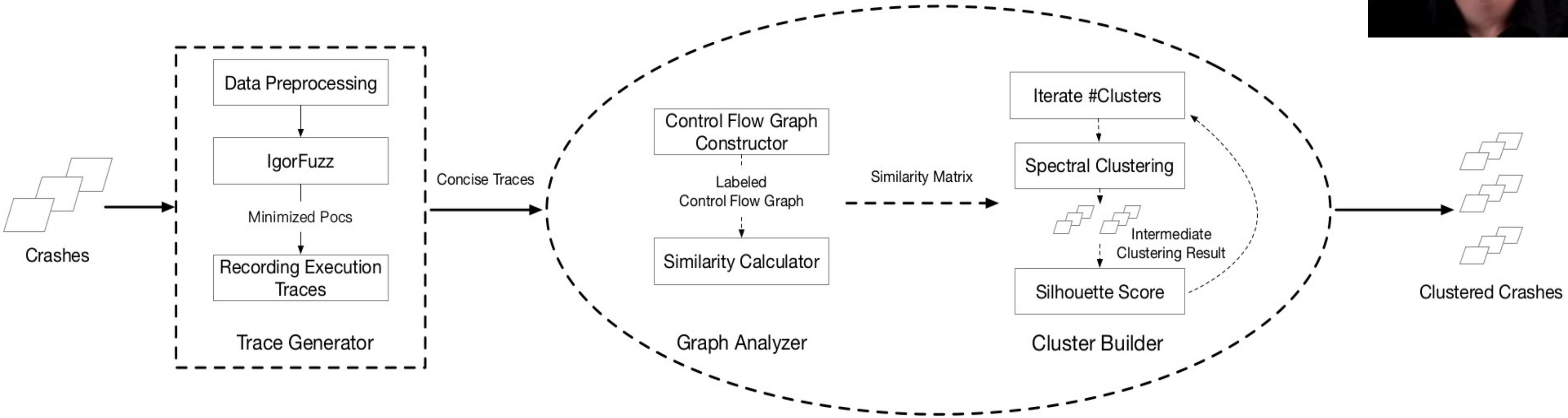
mlbench.github.io

Benchmark Suite:

- distributed machine learning
- public & reproducible collection of reference implementations
- algorithms, frameworks and systems
- PowerSGD is now integrated to FB's PyTorch

Enhanced Fuzzing Using AI [CCS'21]

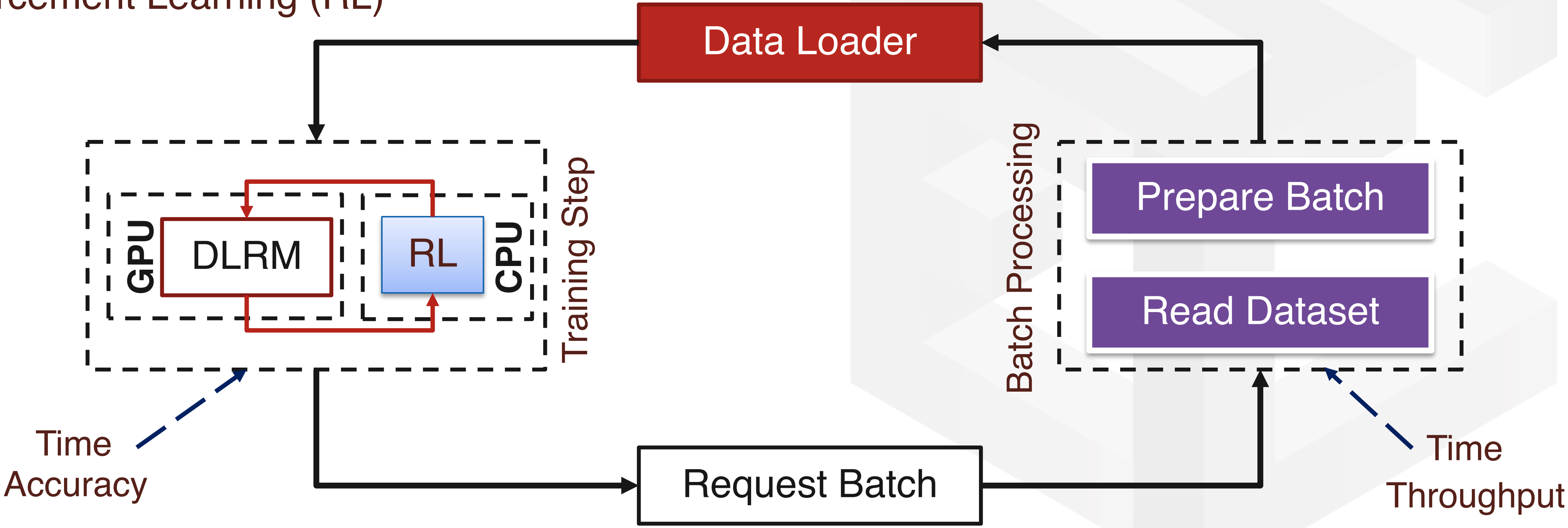
- Fuzzing produces thousands of crashes for tens of bugs
- Grouping reduces developer cost but risks missing bugs
- Our two-phased approach ***minimizes test cases*** and groups them using ***CFG-based clustering***



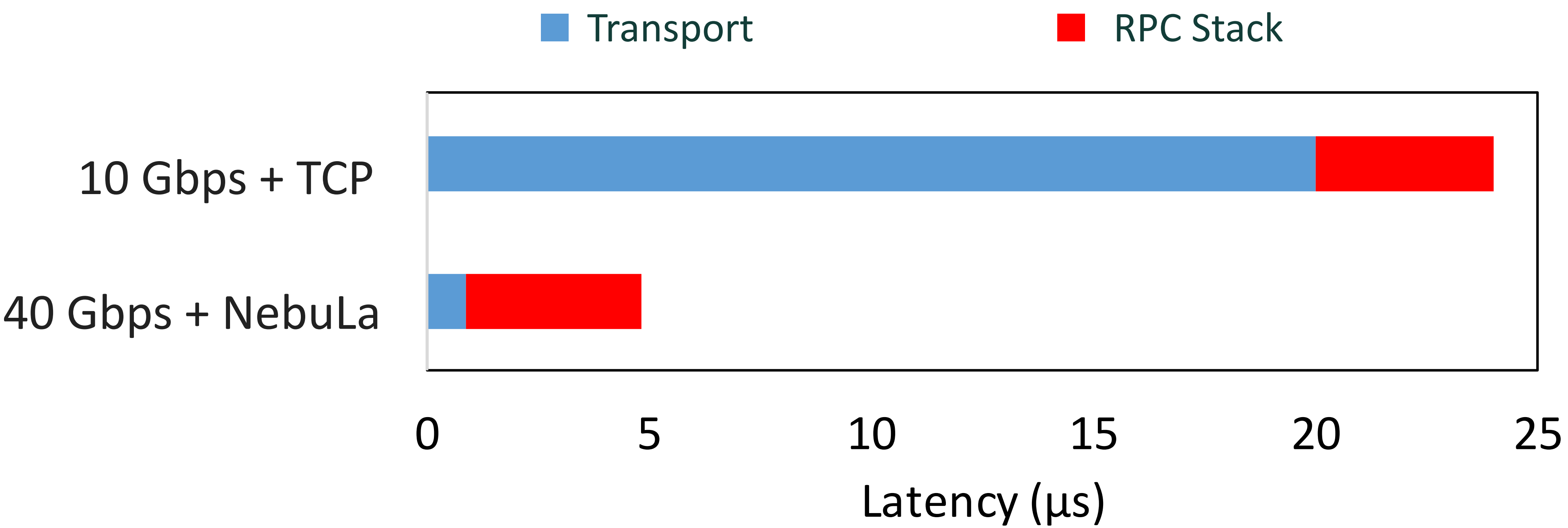
Training for Recommendation Models

1. **Memory optimization**
   - Dataset reading
   - Data Loader
   - CPU-GPU interactions
2. **Automatic neural architecture search**
   - Reinforcement Learning (RL)

Up to 29.5x and **23x** speedup for Terabyte datasets!

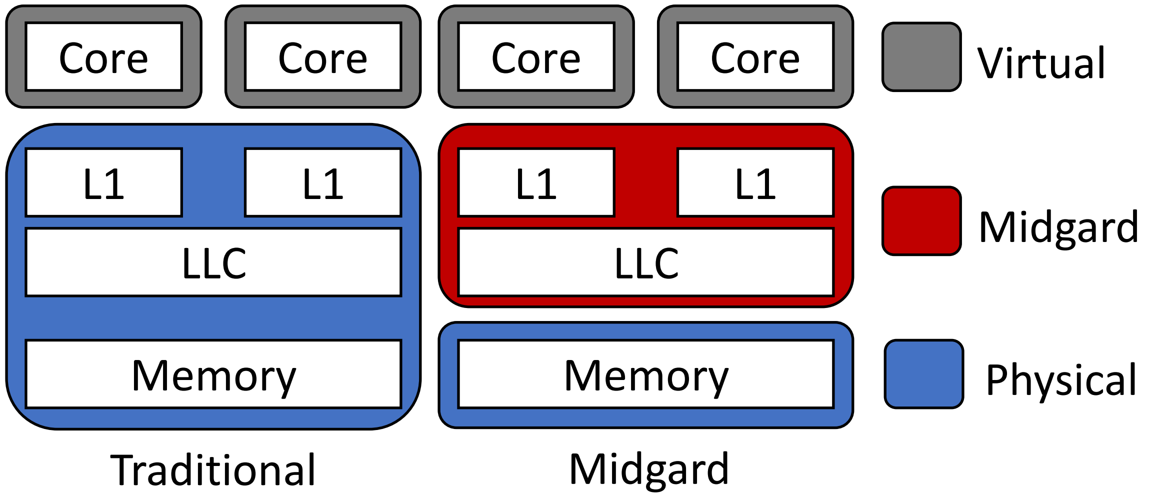


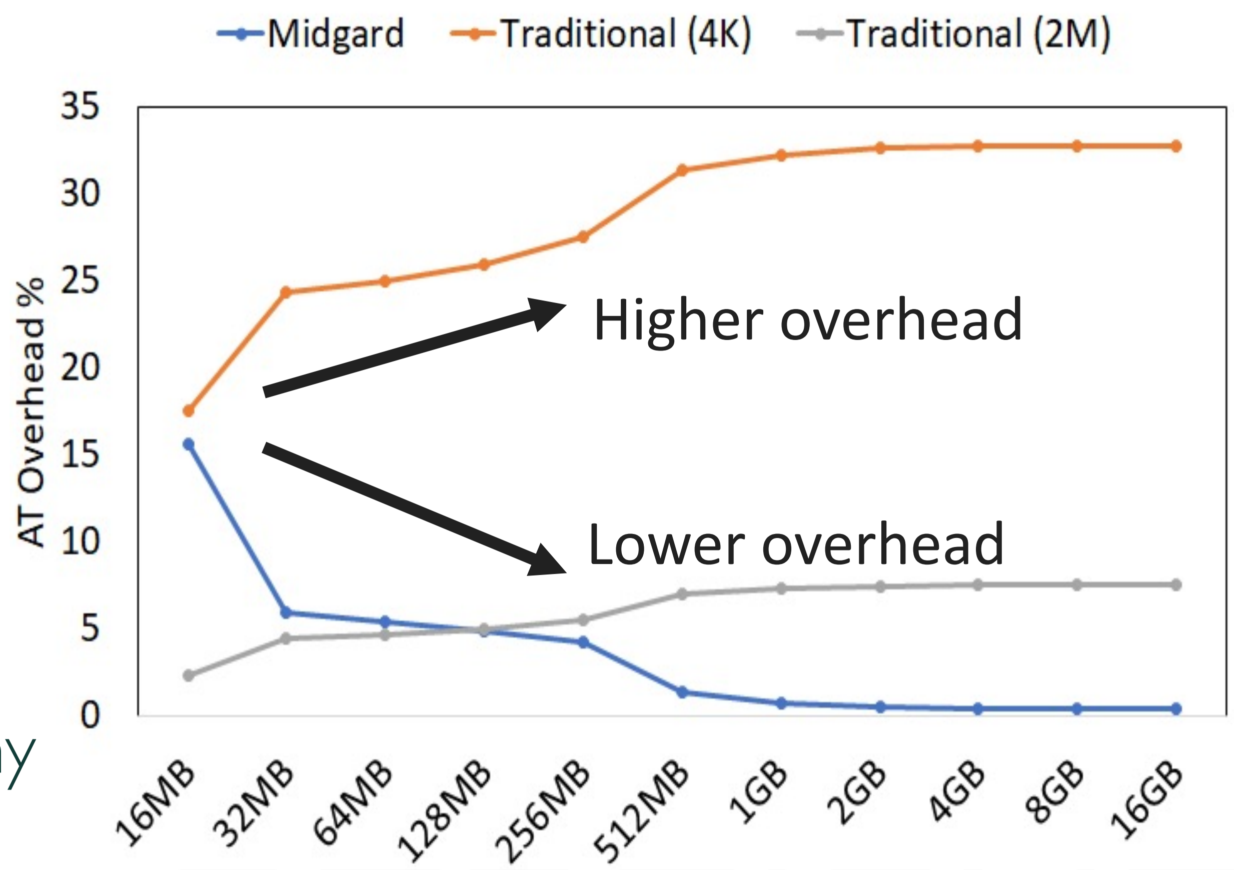
Cerebros: RPC Processor @ Line Rate



- Orchestration is 10x-100x slower than network line rate
- NebuLa: Hardware-terminated network protocol [ISCA'20]
- Cerebros: NIC-integrated RPC Processor [ASPLOS'20, MICRO'21]

Rebooting Virtual Memory with Midgard





Source: Midgard [ISCA'21]

- Decouple VMA's from pages
- Unique namespace for cache hierarchy
- Eliminates POSIX overhead
- Create/revoke access control instantly

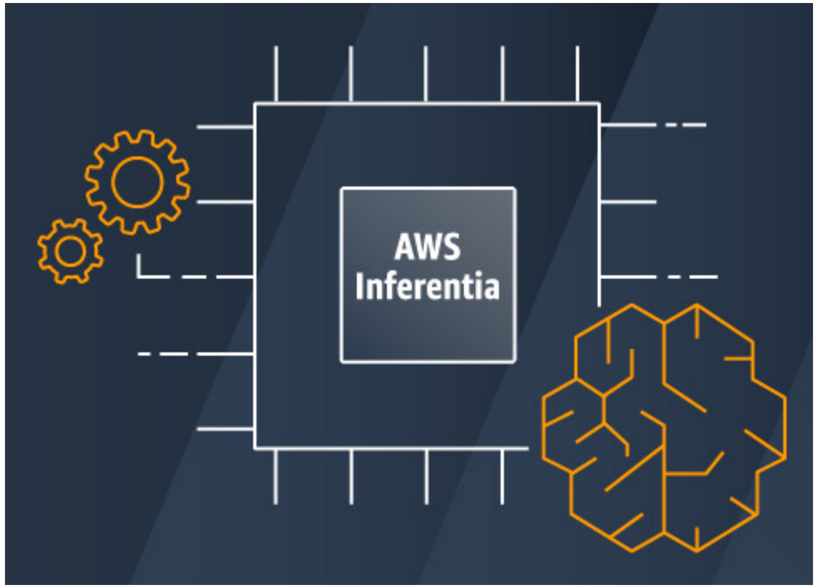
Outline

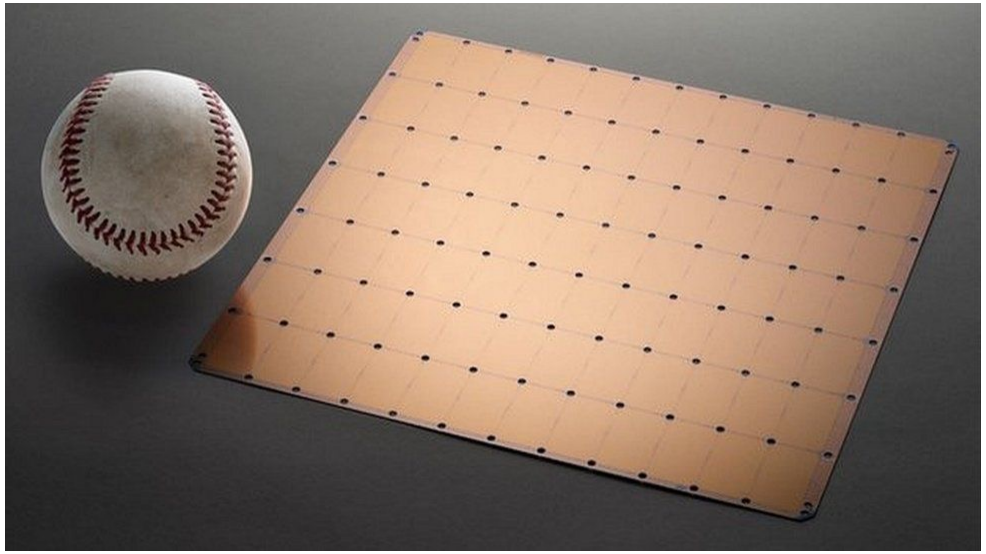
- Overview of EcoCloud
- DNN Accelerators
  - Inference/Training Divergence
  - Encoding
  - Inference Accelerator
- Summary

DNN's Platform Divergence

Inference platforms:

- Tight latency constraints
- Ubiquitous deployment
- Relies on fixed-point arithmetic

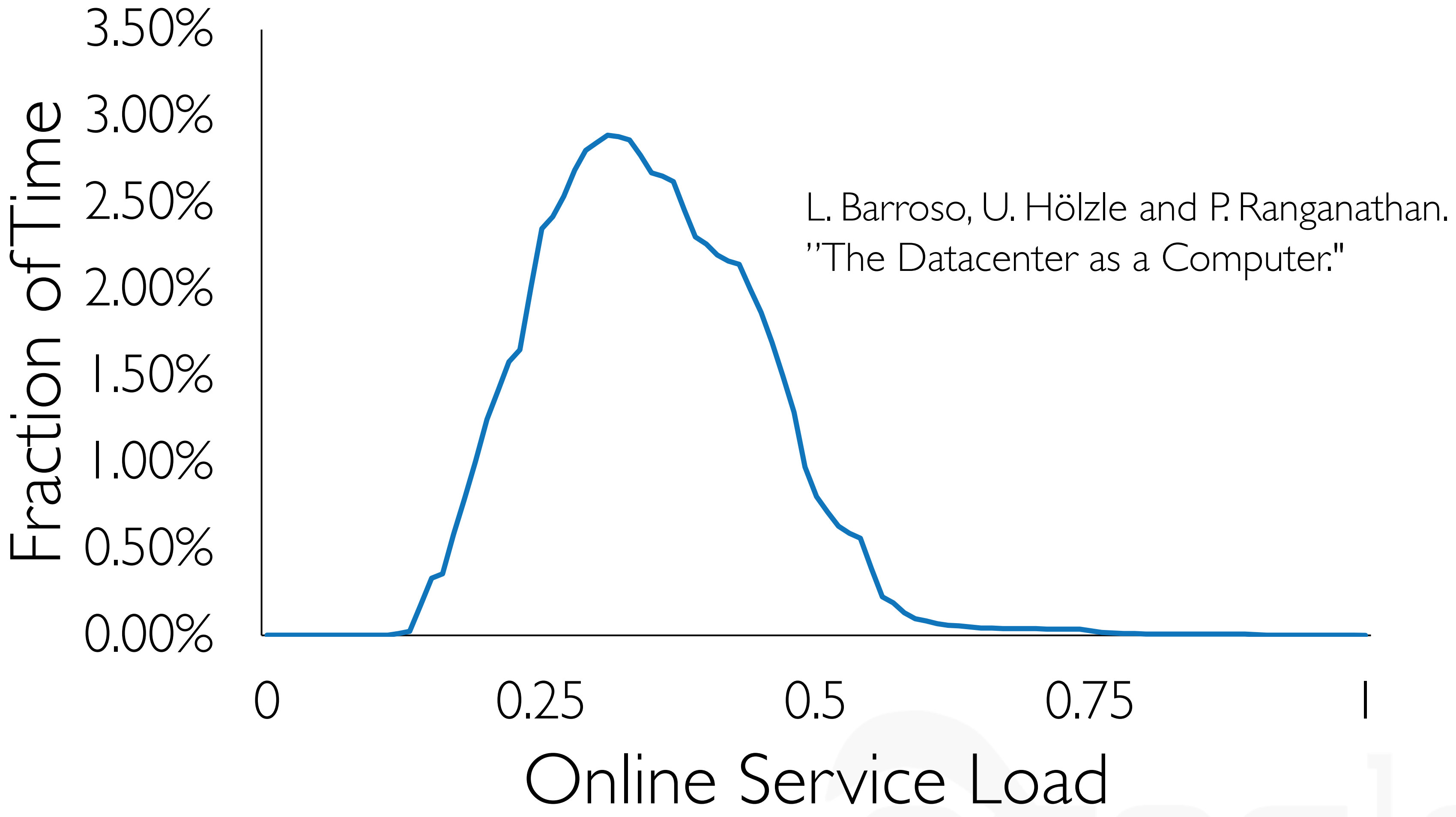




Training platforms:

- Throughput optimized
- Server deployment
- Requires floating-point arithmetic

Inference Services Idle



Can idle cycles be used otherwise (e.g., for training)?

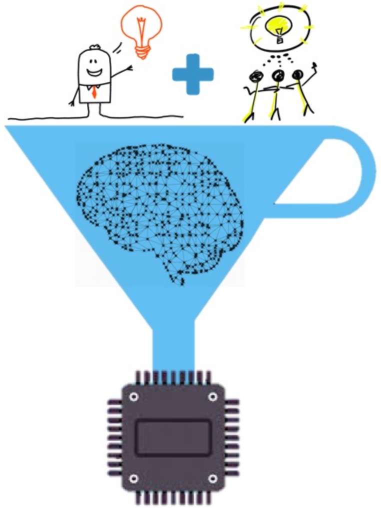
Contributions

Answer to questions, can we

- train with fixed-point w/o loss of accuracy?
  - how low can we push precision?
- build an inference accelerator that can also train?
  - how much latency do we sacrifice?

Sneak peek answers, yes!

(visit ColTraIn @ parsa.epfl.ch/coltrain/)

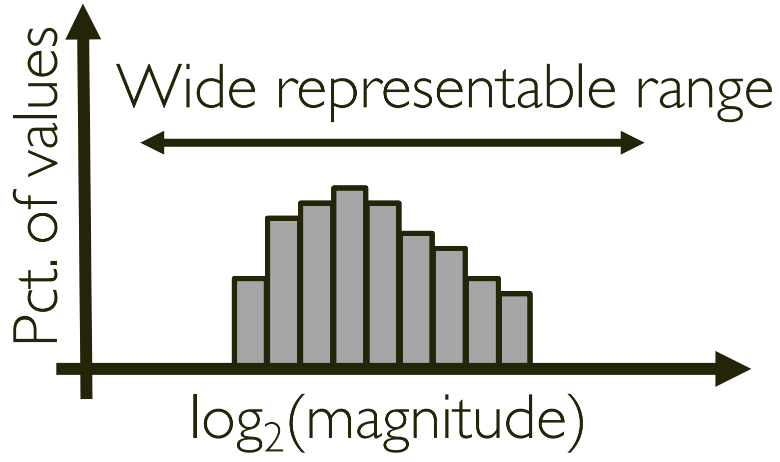


Floating vs. Fixed Point: Representable Range

Floating point

- Mantissa + exponent

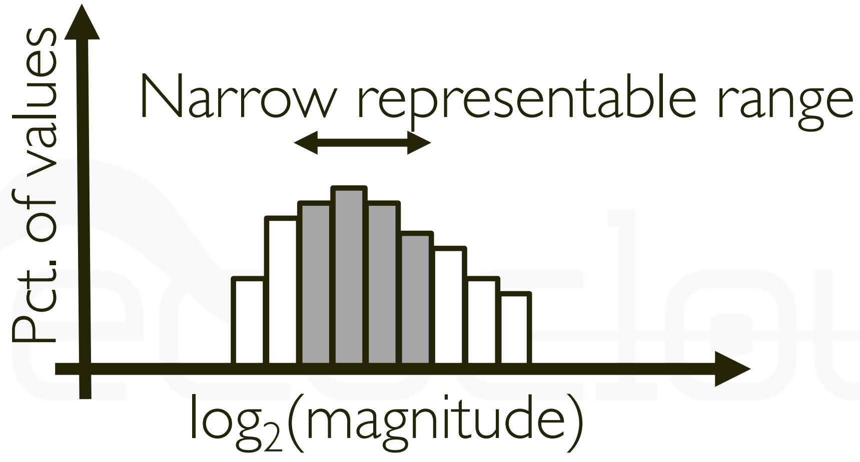
- Wide representable range
- Value has independent range



Fixed point

- Mantissa

- Narrow representable range
- Values range pre-determined

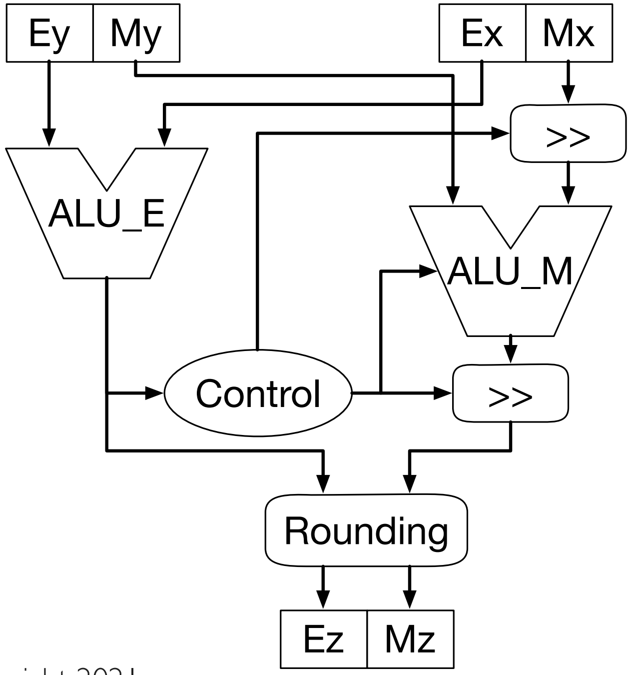


Floating vs. Fixed Point: Area and Power

▪ Floating point

- Mantissa + exponent

- Complex exponent management

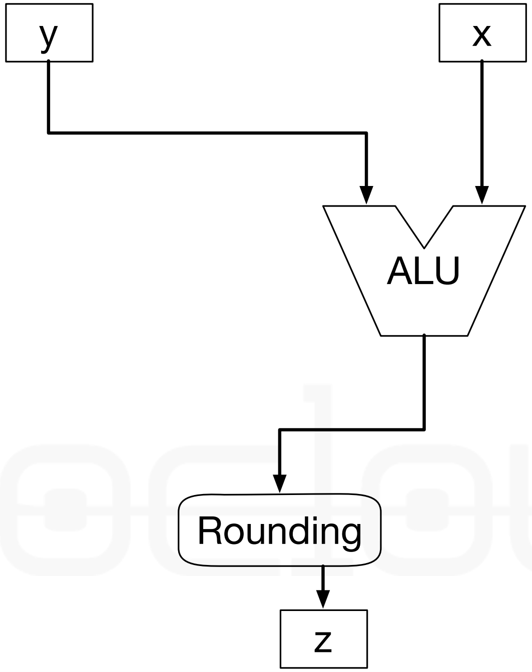


▪ Fixed point

- Mantissa

- No exponent management

ALU Hardware



Hybrid BFP-FP (HBFP) [NeurIPS'18]

Block floating point (BFP) shares exponents in blocks

- Proposed for DSP's to reduce silicon footprint
- > 90% of arithmetic with one exponent/tensor

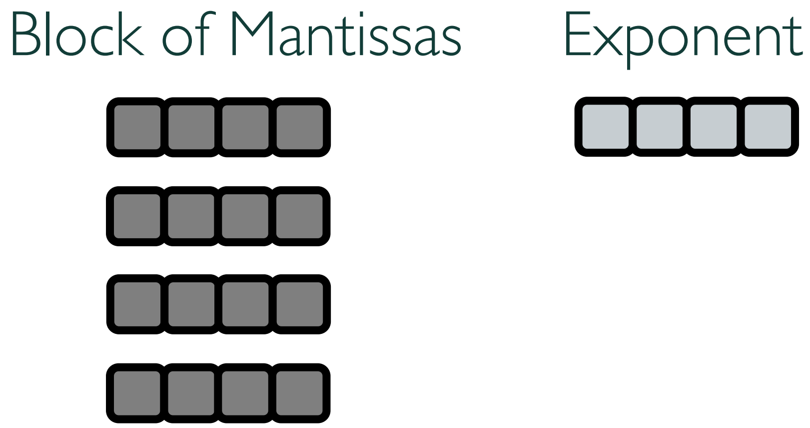
Use FP32 for all activations and other arithmetic

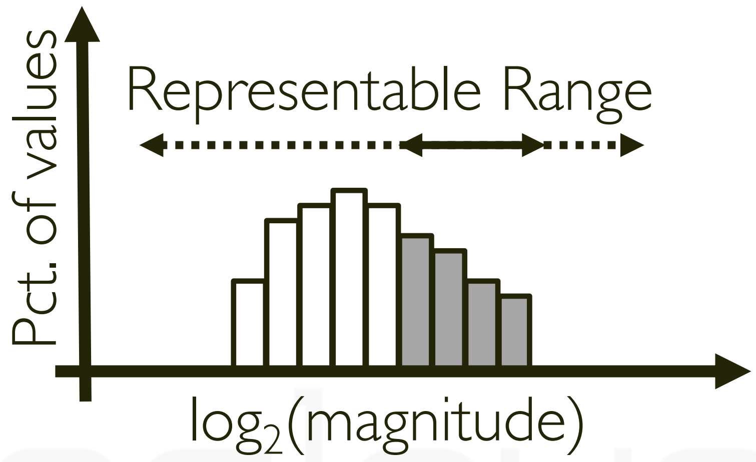
Co-Located Training & Inference (ColTraIn)

✓Uses fixed point logic (hbfp) for both training & inference
✓Piggyback training on a custom inference accelerator

Open-source emulator

github.com/parsa-epfl/HBFPEmulator





HBFP vs. FP32

Emulated HBFP dot products with PyTorch

- Saturated and rounded inputs/outputs of dot products
- Covered forward and backward passes
- Weight updates in fp32 but weights kept in BFP

Models & datasets:

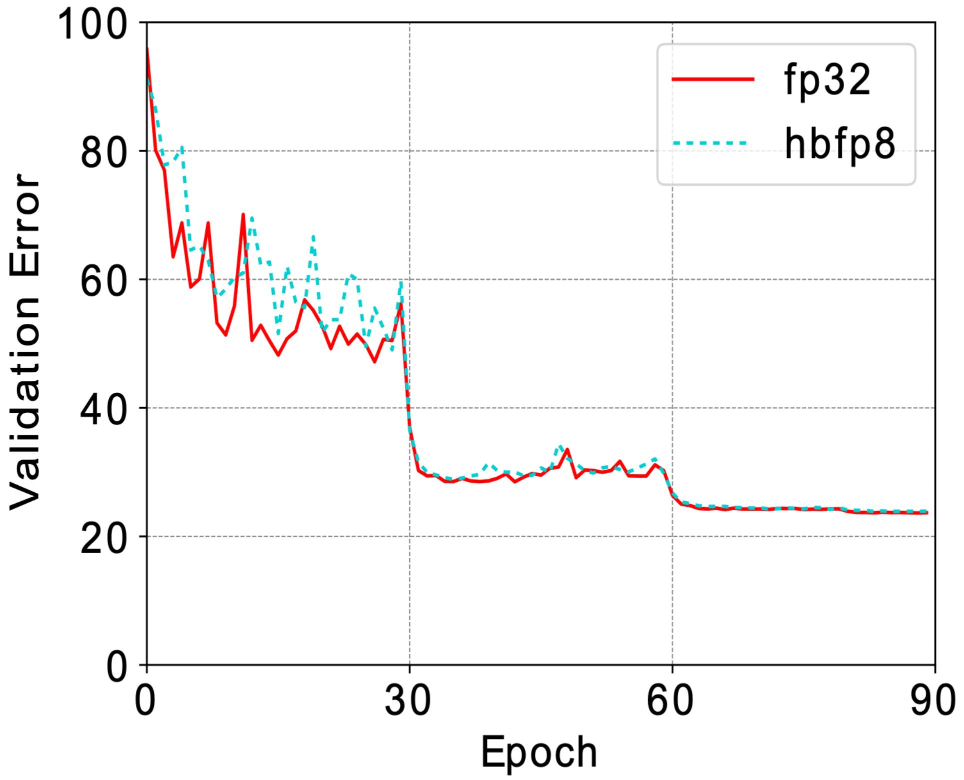
- ImageNet, CIFAR-100, SVHN, Penn Tree Bank and English Wikipedia
- ResNet, WideResNet, Densenet, LSTM and BERT

HBFP parameters:

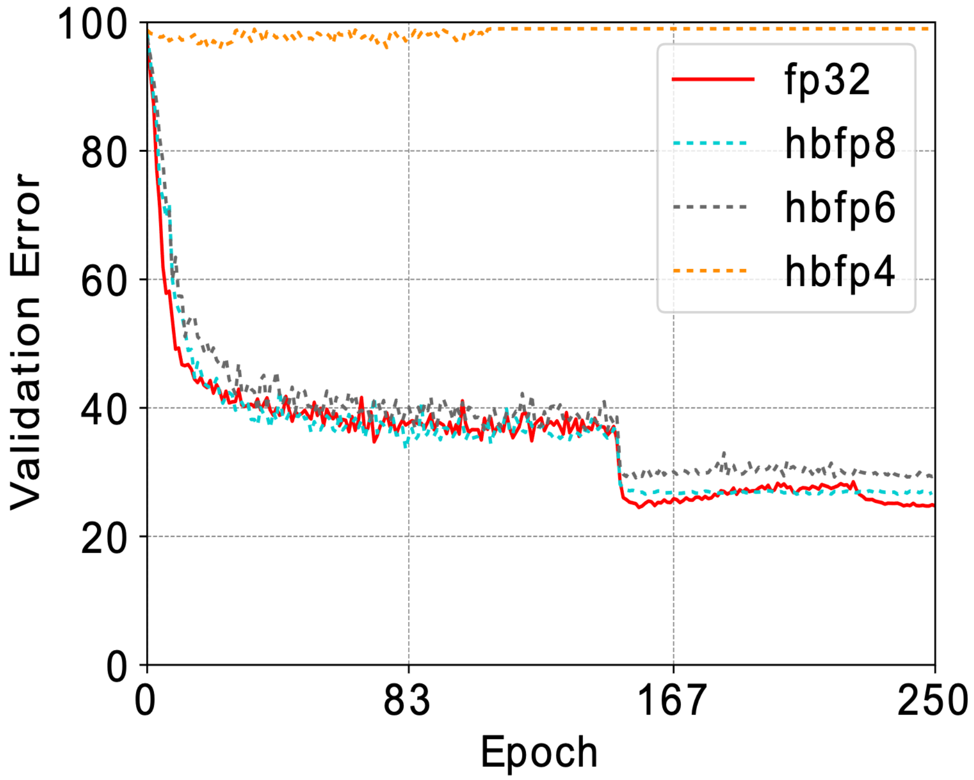
- 10-bit exponent (we vary the mantissas from 4 to 8 bits)
- Baseline block size: 24x24 tiles (576 mantissas sharing an exponent)
- All hyperparameters tuned using fp32

CNN

ResNet50 on ImageNet



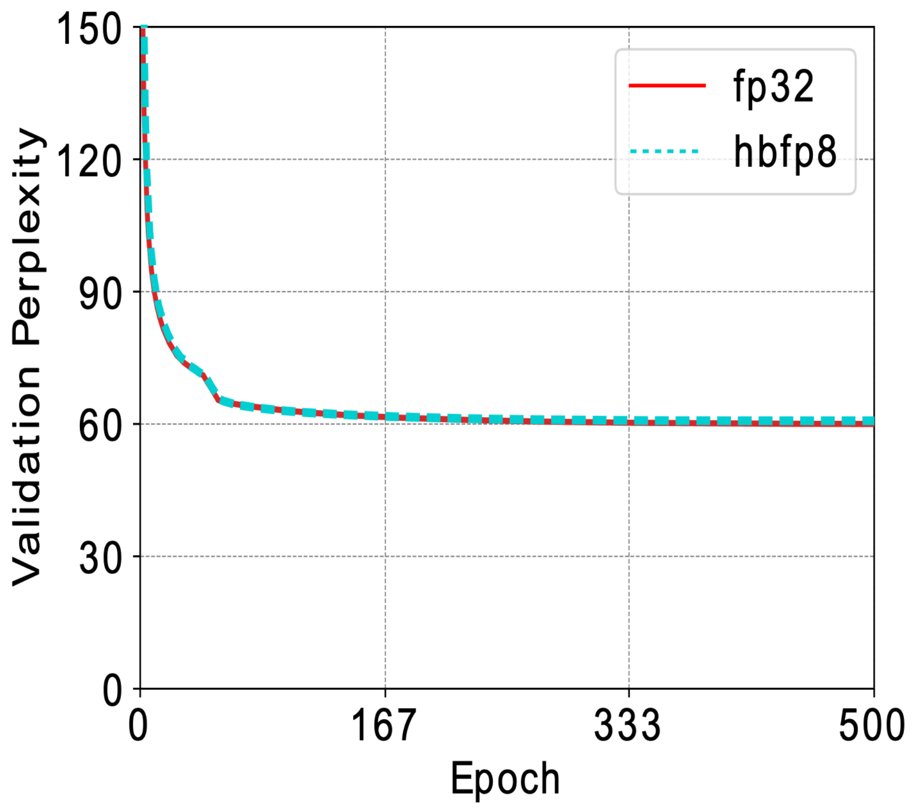
ResNet50 on CIFAR100



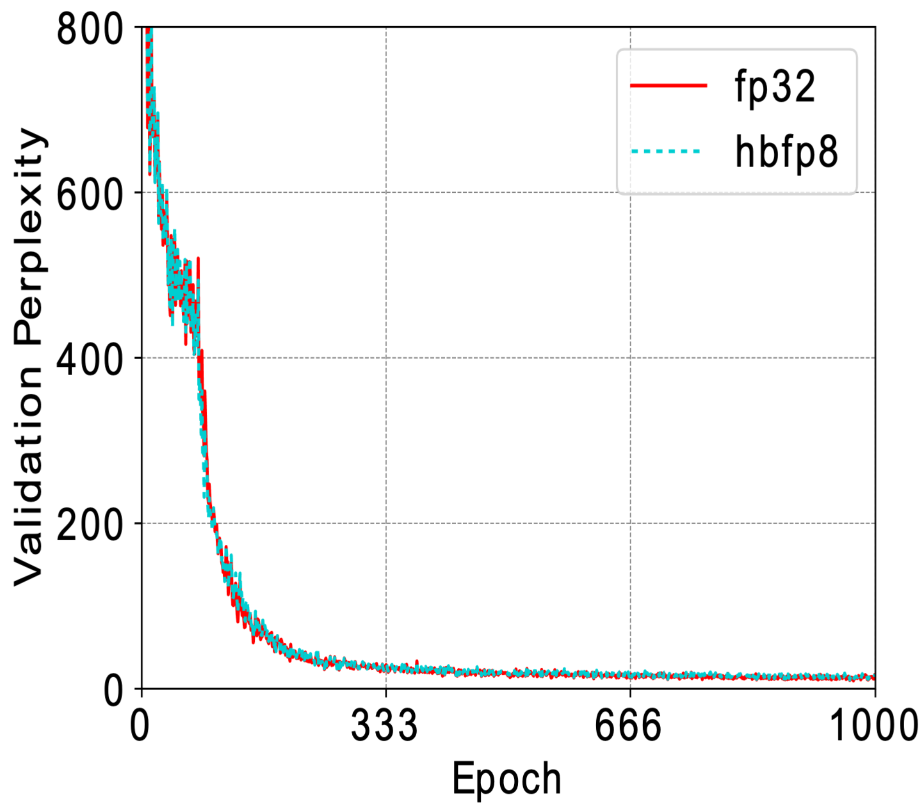
hbfp8 tracks fp32 with an 8-bit fixed encoding

NLP

LSTM on Penn Tree Bank

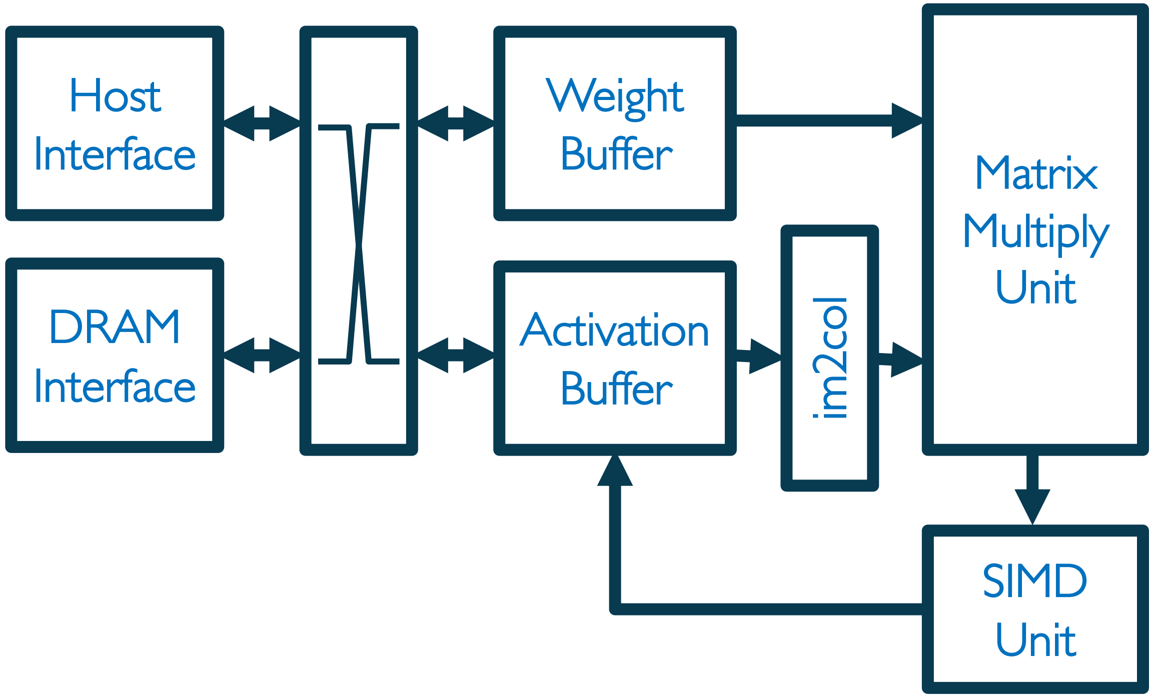


BERT on English Wikipedia

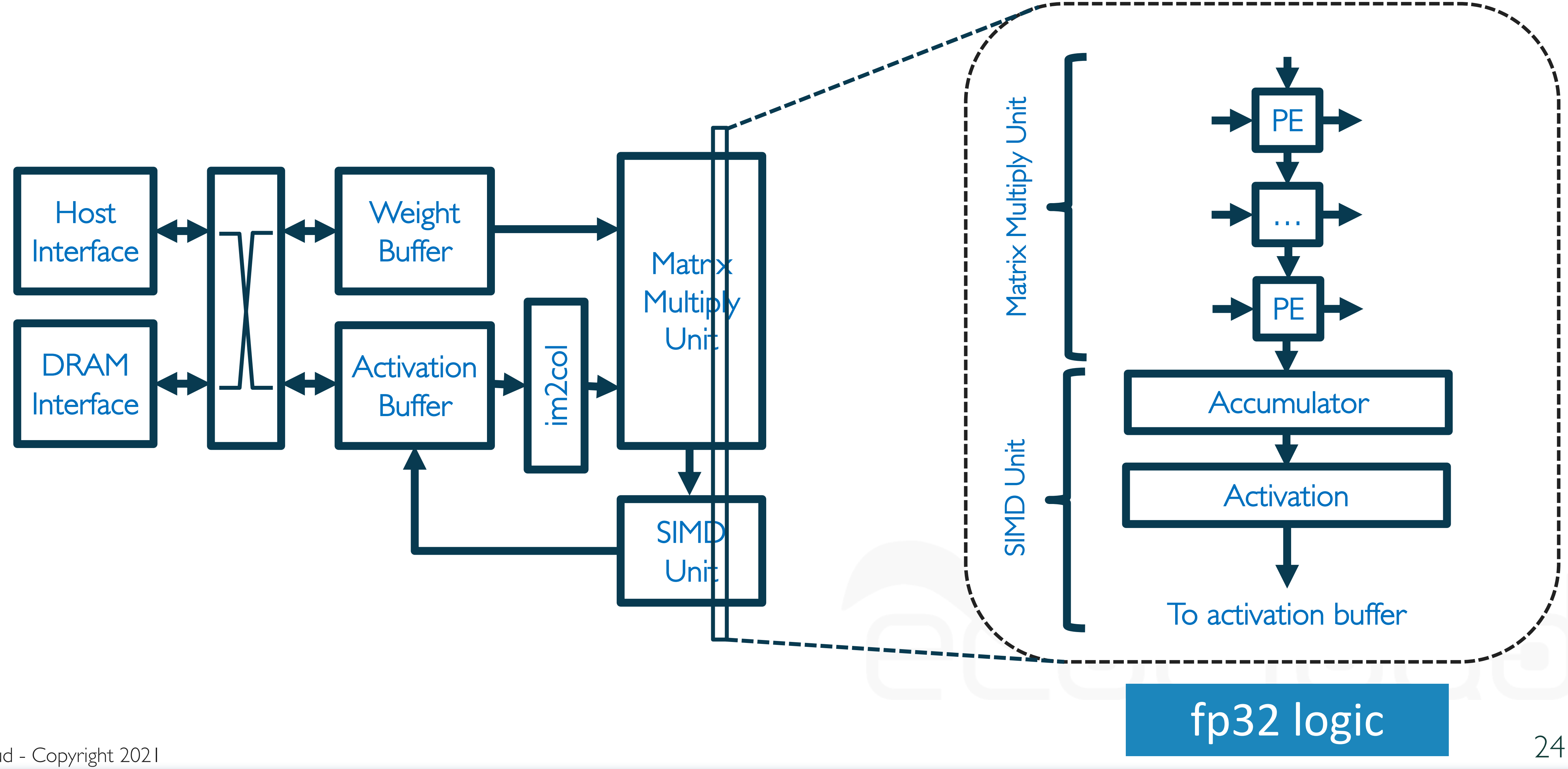


fp32 accuracy with an 8-bit logic

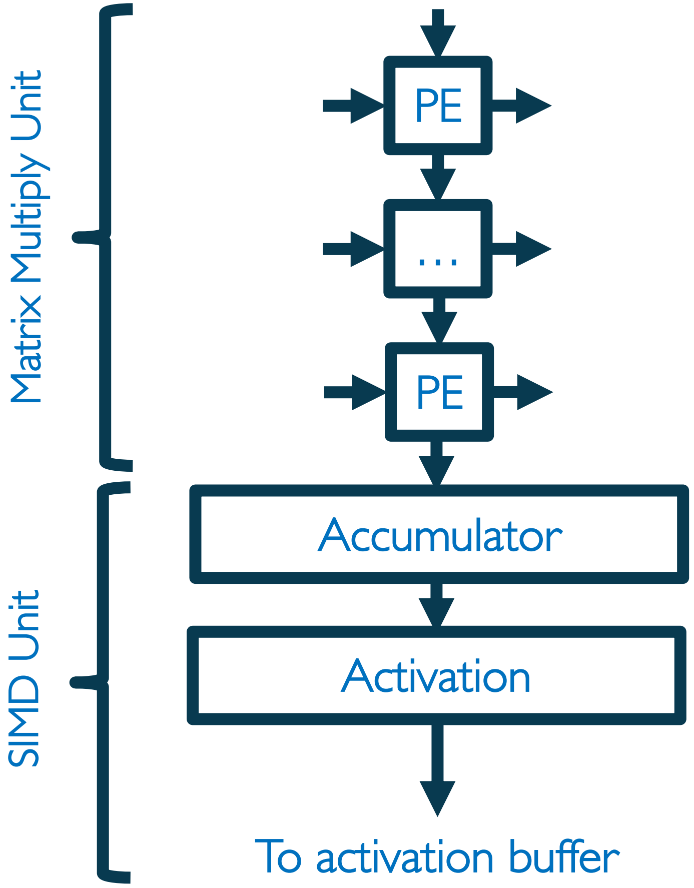
Equinox: Our Baseline Inference Accelerator



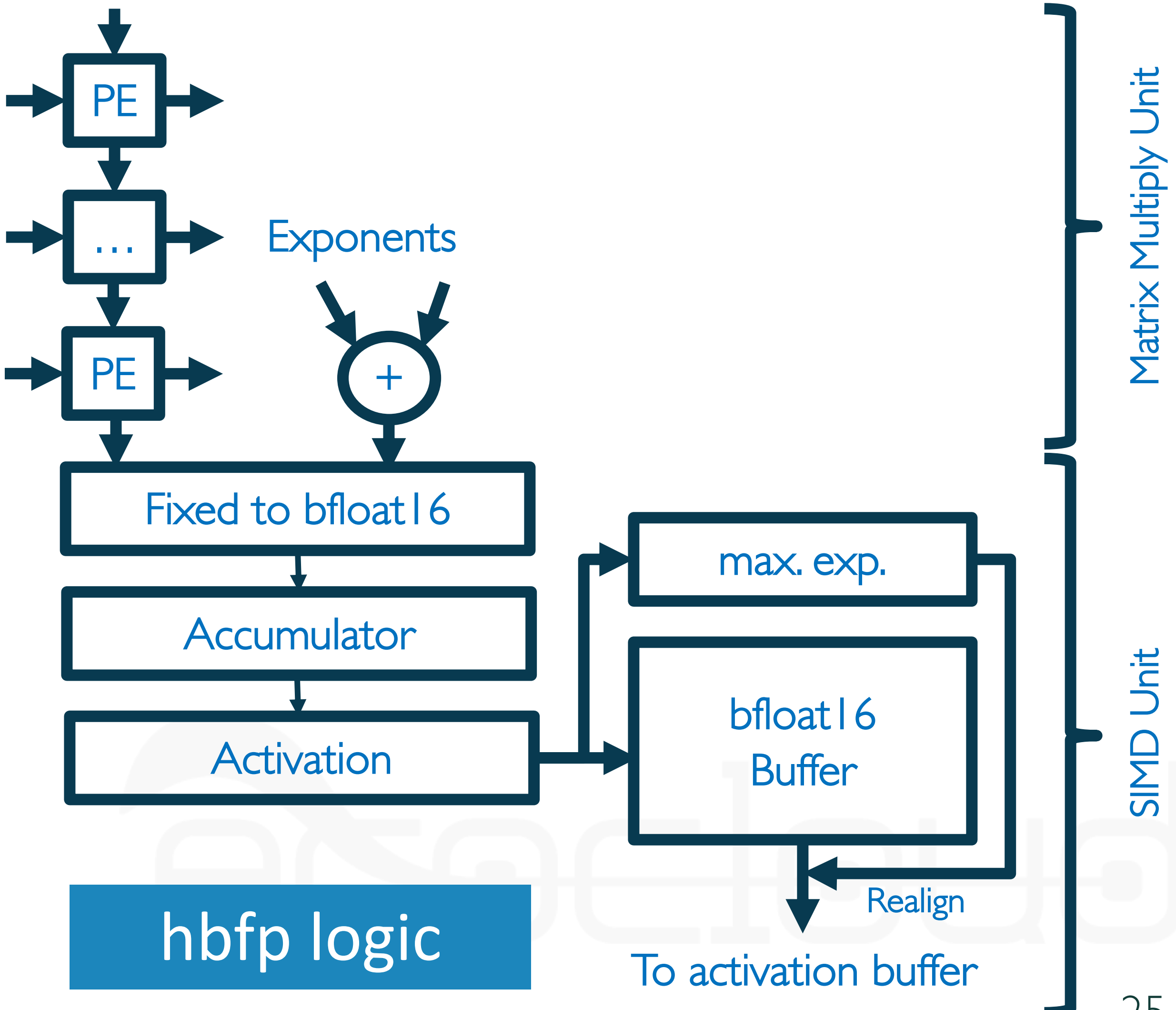
Datapath with FP32 Logic



FP32 vs. HBFP Datapath Logic

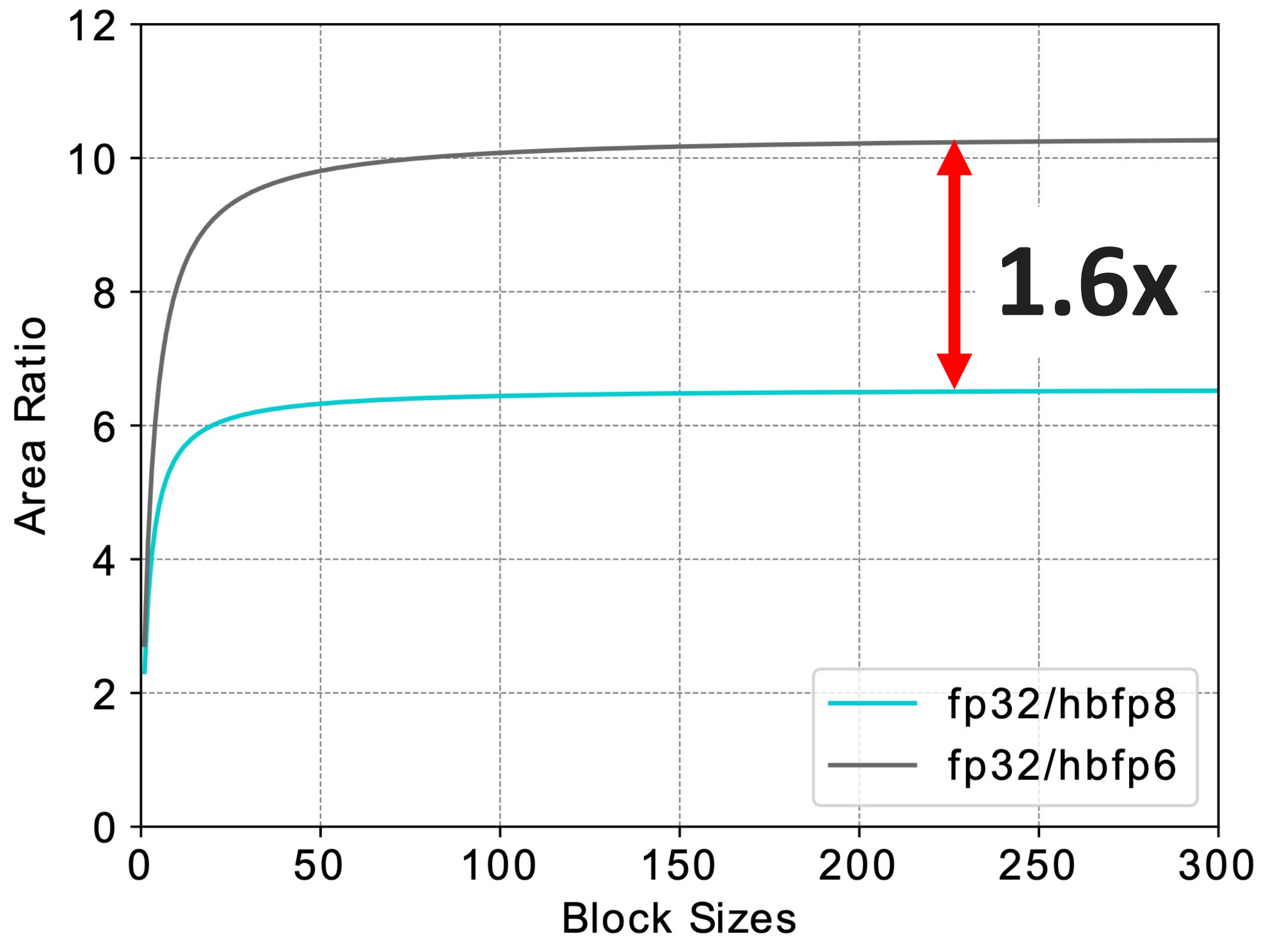


fp32 logic



hbfp logic

Logic Area Comparison



ResNet20 on CIFAR10

Encoding (w/ block size)	Validation Error (%)
fp32	7.9
hbfp8_576	8.4
hbfp8_256	8.3
hbfp6_256	26.7
hbfp6_225	8.8
hbfp4_1 (fp14)	11.4

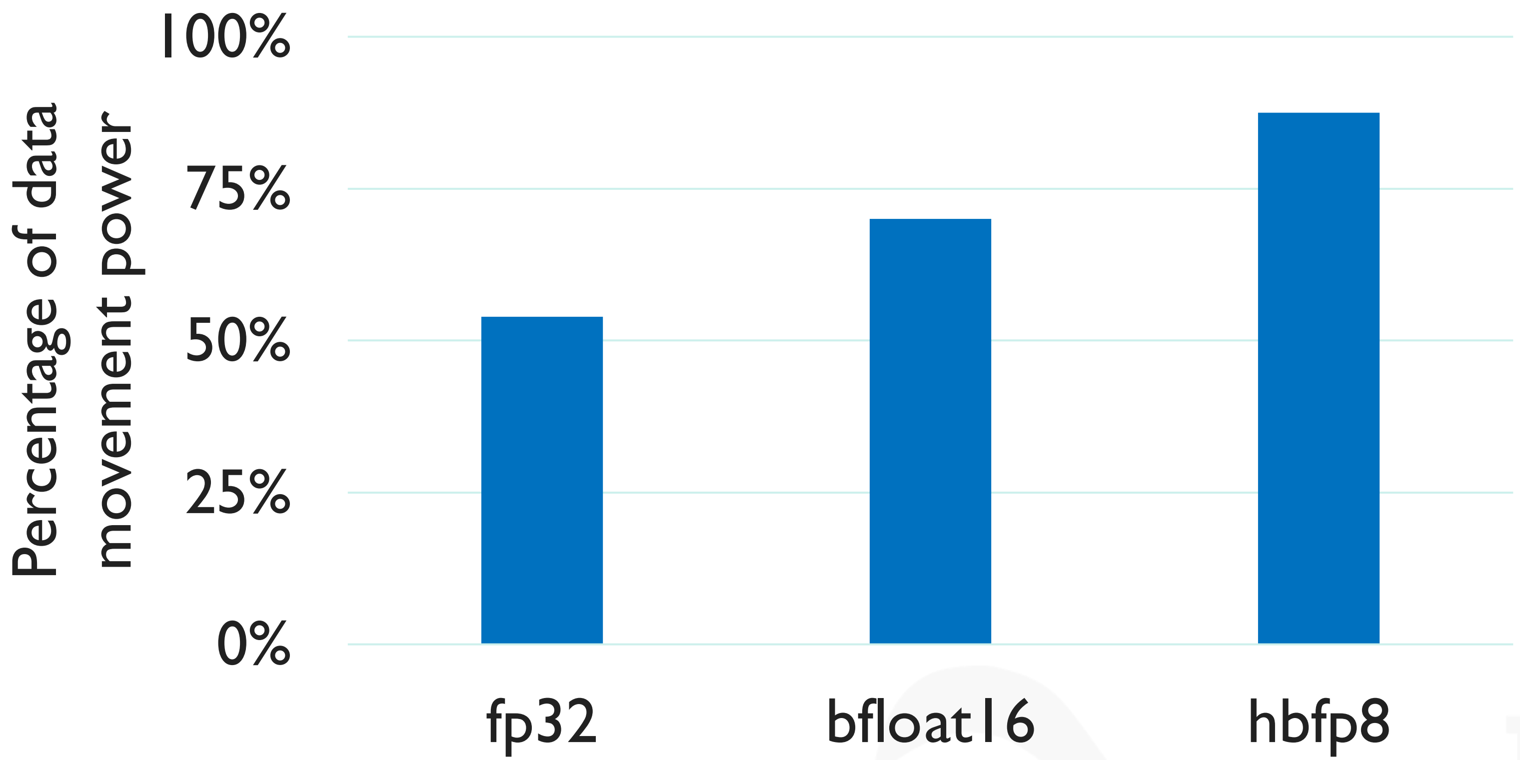
hbfp6 w/ smaller blocks over 1.6x better than hbfp8!

Outline

- Overview of EcoCloud
- DNN Accelerators
  - Inference/Training Divergence
  - Encoding
  - Inference Accelerator
- Summary

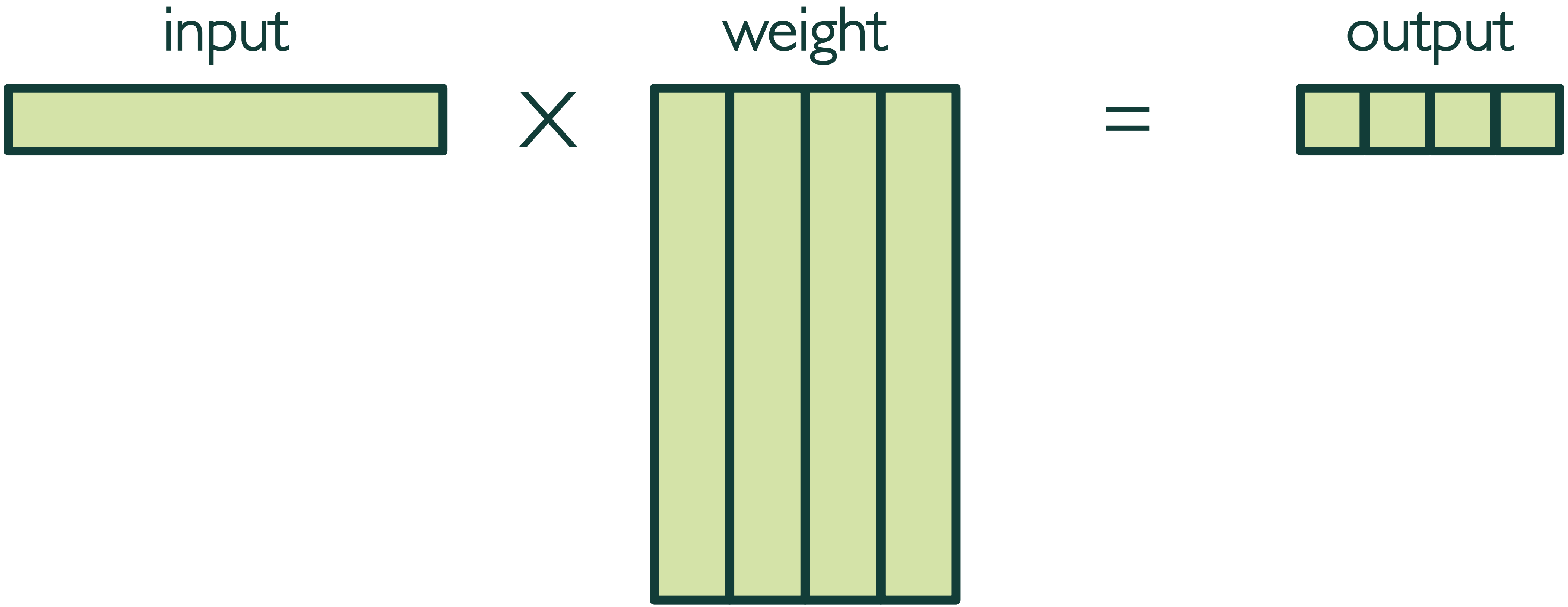
Fixed point is Bound by Movement

- Lower precision → more power for data movement



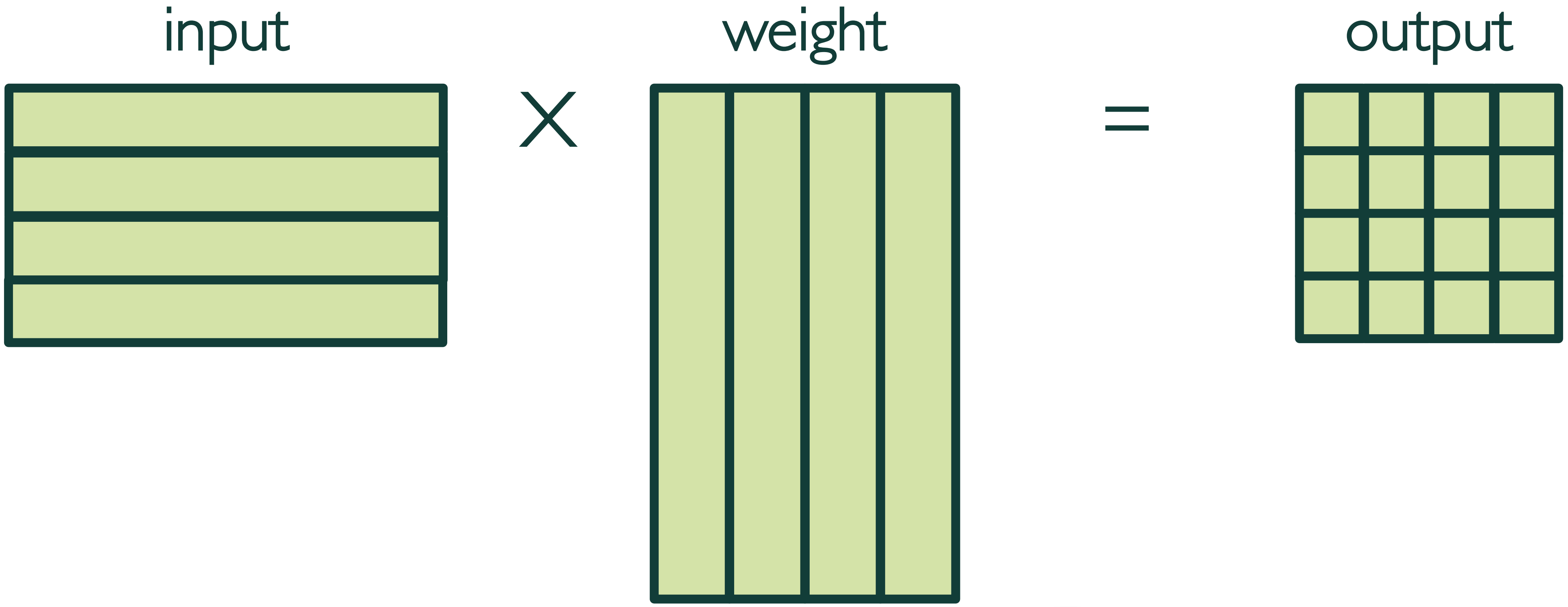
Fixed-point DNN accelerators must exploit reuse for efficiency

Exploiting Reuse in DNNs



Vector-matrix multiplication has input reuse but no weight reuse

Exploiting Reuse in DNNs



Matrix(-matrix) multiplication has both weight and input reuse

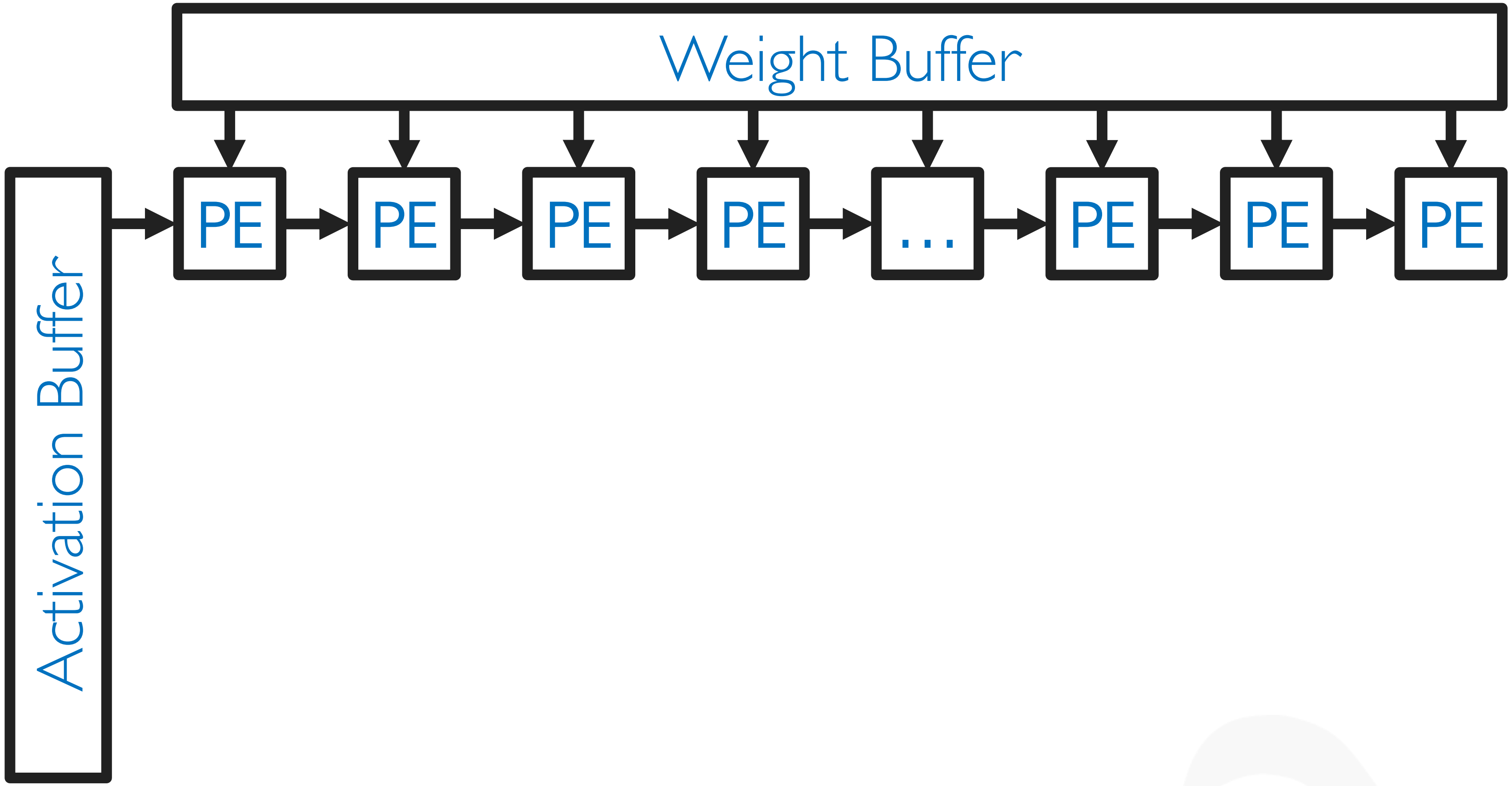
Reuse vs. Latency Tradeoff w/ Batching

- Many models (MLPs, LSTMs) are vector-matrix multiplication based
- Batching converts vector-matrix to matrix multip.
  - Execute inputs in groups, reuse weights across inputs
  - Recovers power lost in data movement

- But, forces requests to wait for batch formation
  - Online inference services have tight latency requirements
  - From tens of microseconds to few milliseconds

Exploit slack in latency & power to maximize accelerator throughput!

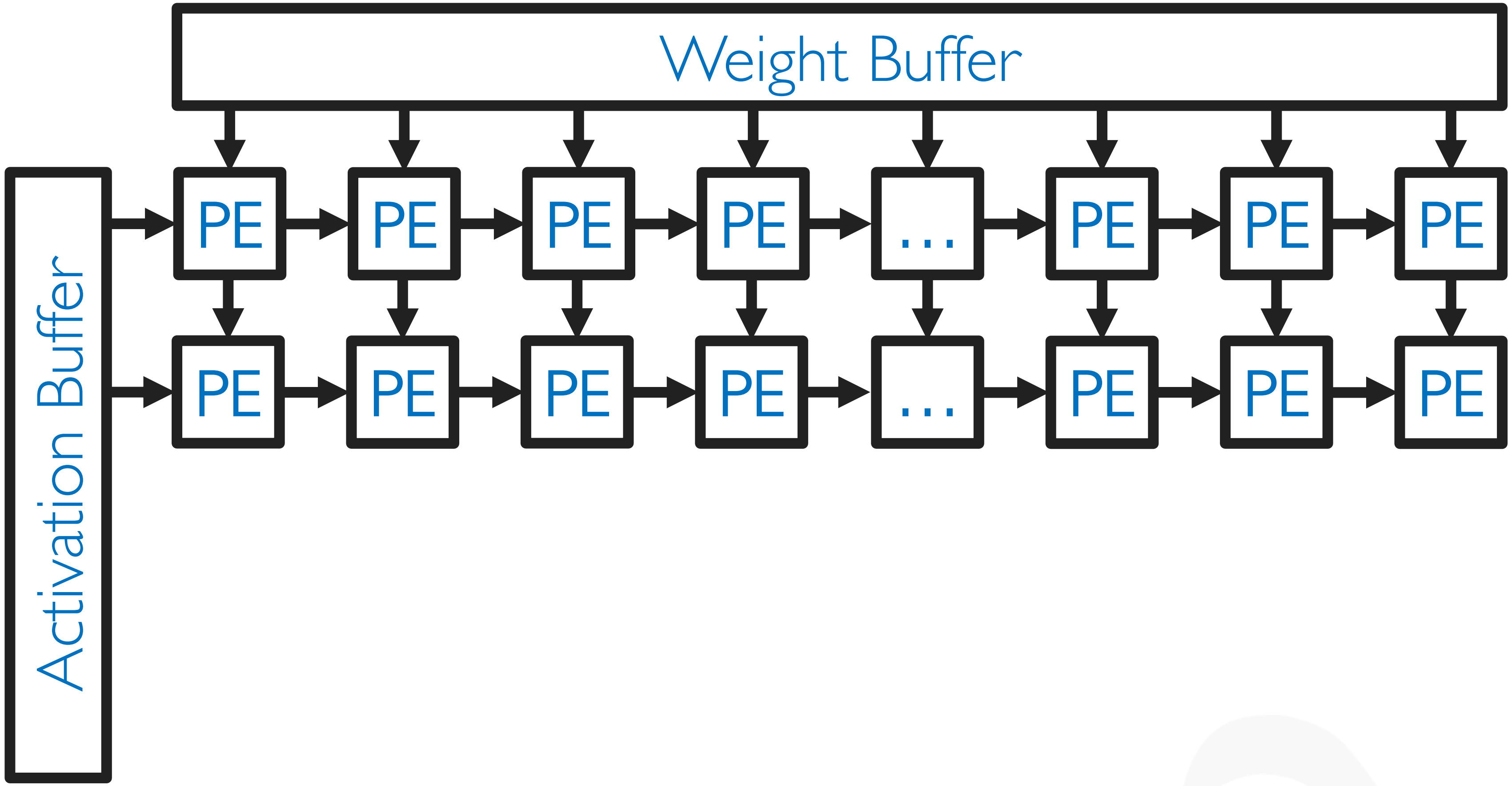
Single-Row Array: Latency vs. Throughput vs. Power



Weight Power	10n
Act. Power	10
PE Power	n
Total Power	11n
Throughput	n
Latency	1

Assume SRAM access consume 10x more power than a MAC

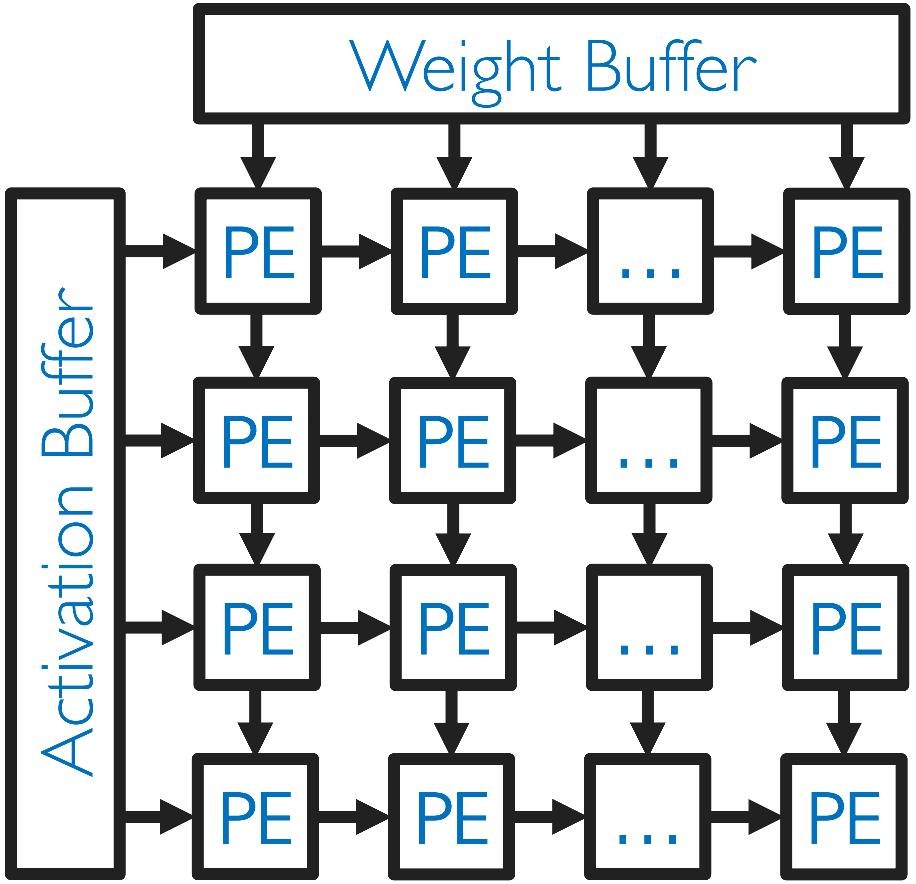
2x Rows → 2x Throughput, ~1x power



Weight Power		10n
Act. Power	2x	20
PE Power	2x	2n
Total Power	~1x	12n
Throughput	2x	2n
Latency		1

Moderate batching increases throughput with little effect on latency

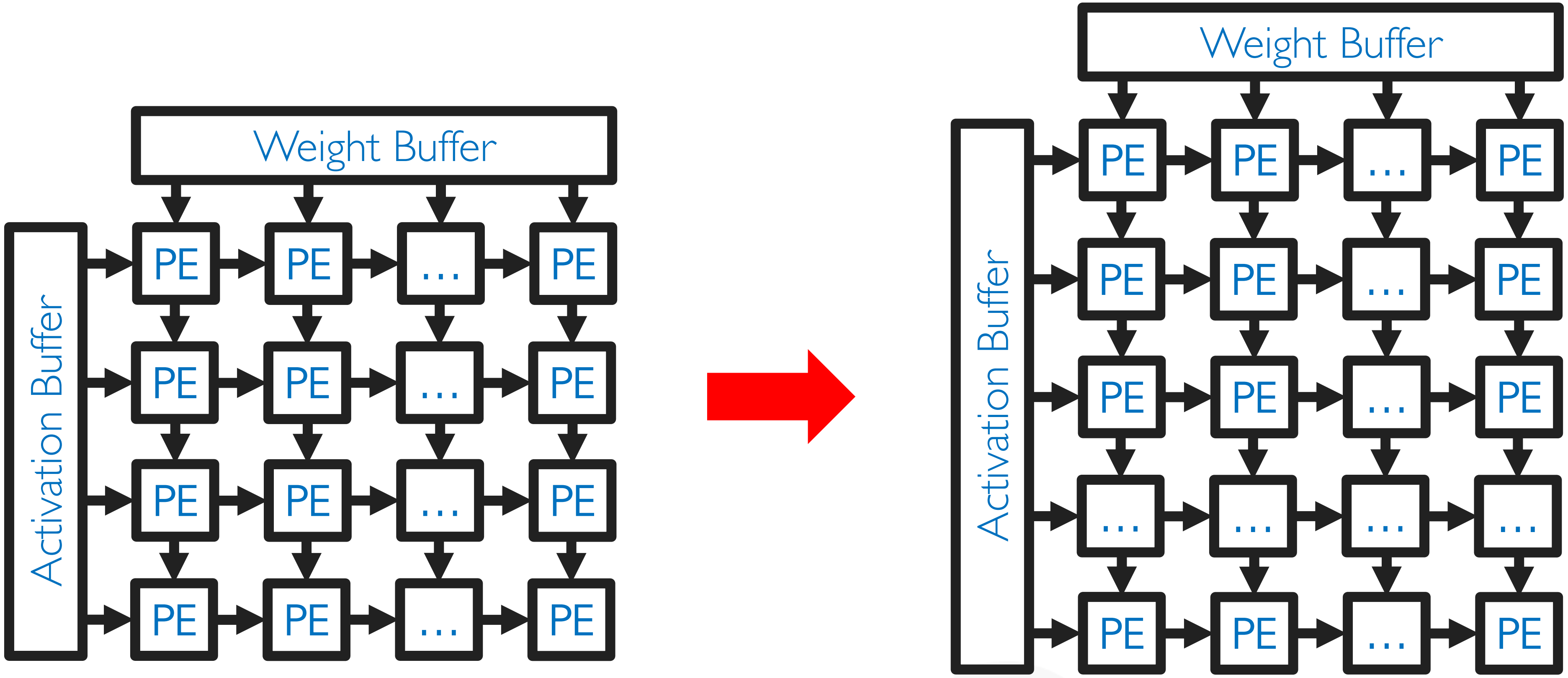
0.5x Columns → 2x Latency, ~0.5x power



Weight Power	0.5x	5n
Act. Power	2x	40
PE Power	2x	2n
Total Power	~0.5x	7n
Throughput	2x	2n
Latency	2x	2

Small factor in μs increase in latency, iso-throughput halves power!

Give the Power Back to PEs



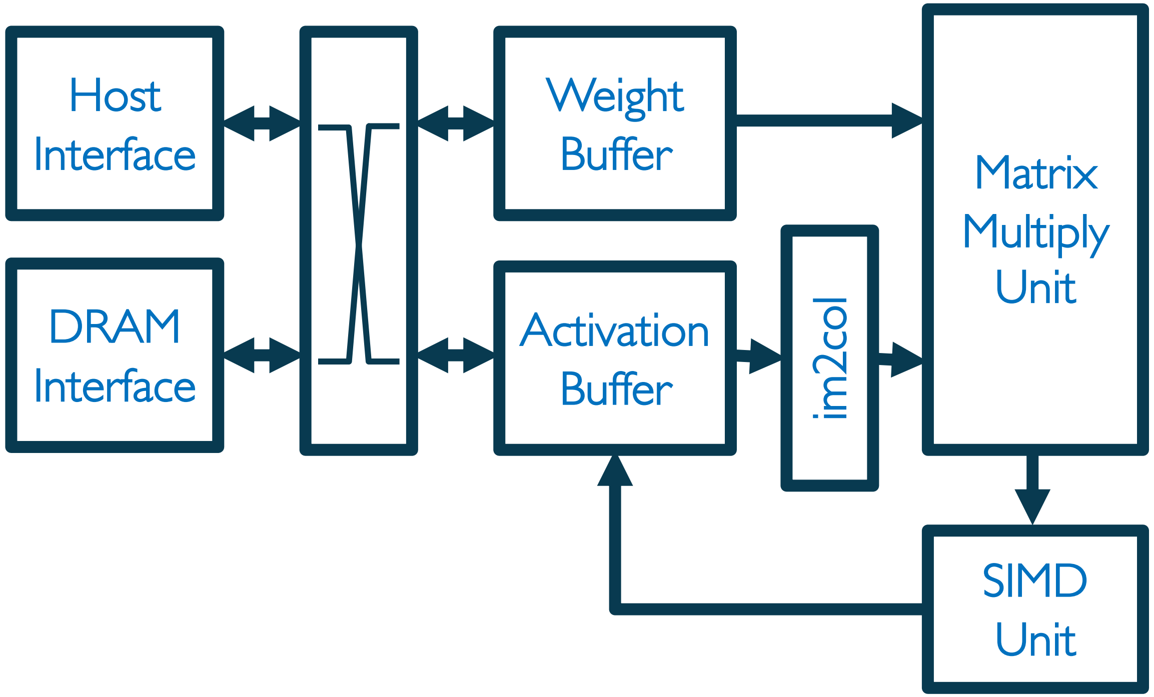
Power = 7n
Throughput = 2x
Latency = 2

Power = 11n
Throughput ~3x
Latency = 2

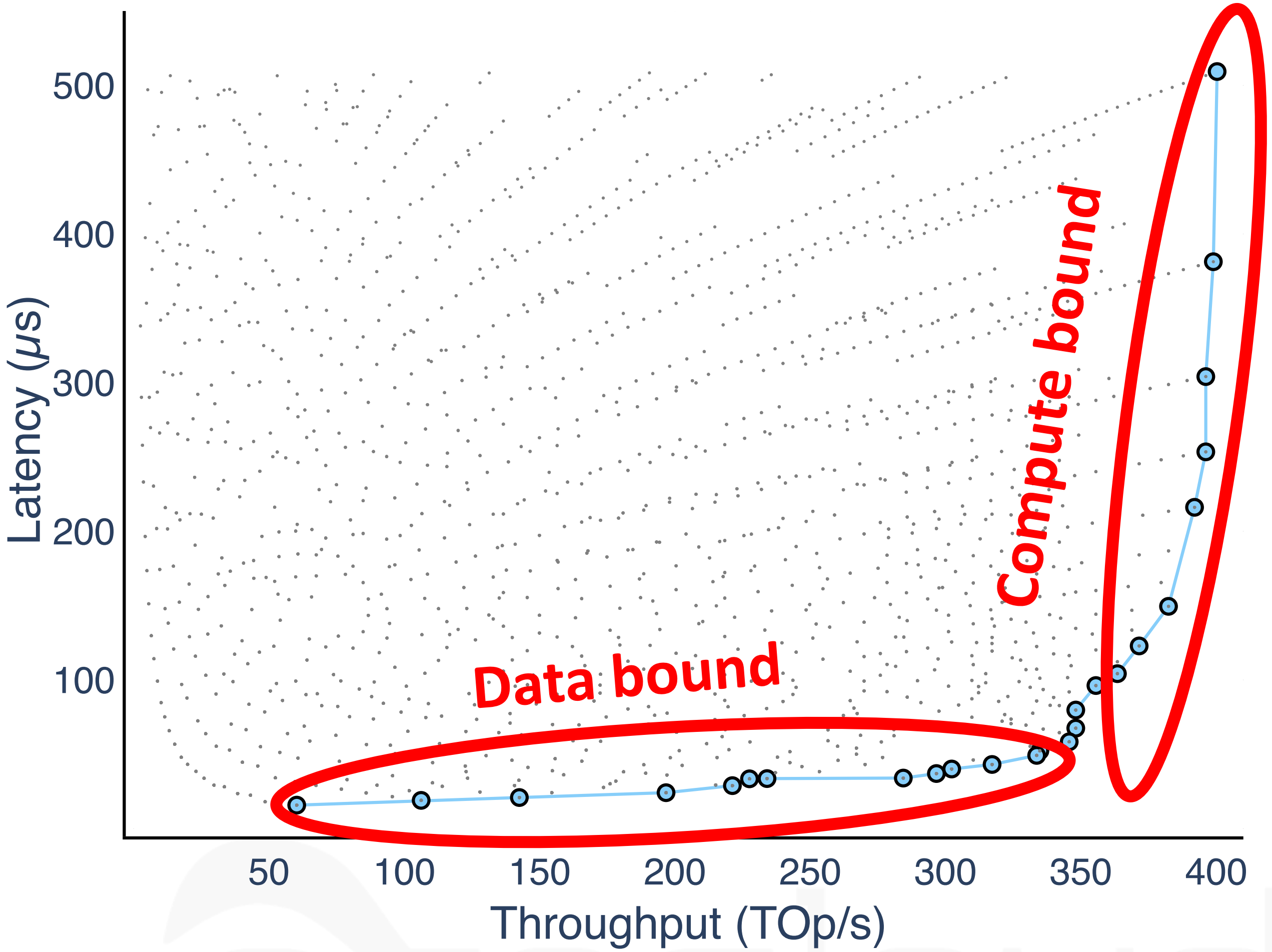
Equinox Evaluation Methodology

- Analytical model
  - 1D array of systolic arrays
  - 75MB of SRAM
  - Vary array height and operating frequency
- Calculate optimal ALU array dimensions for max area & power
  - 300 m^2 @ 75 Watts
- Calculate inference throughput and latency for each design
  - 15-step, 2048-wide LSTM DNN as a reference workload
- Modeled both hbfp8 (HBFP with 8-bit mantissas) and bfloat16 arrays

HBFP8 Design Space [MICRO'21]

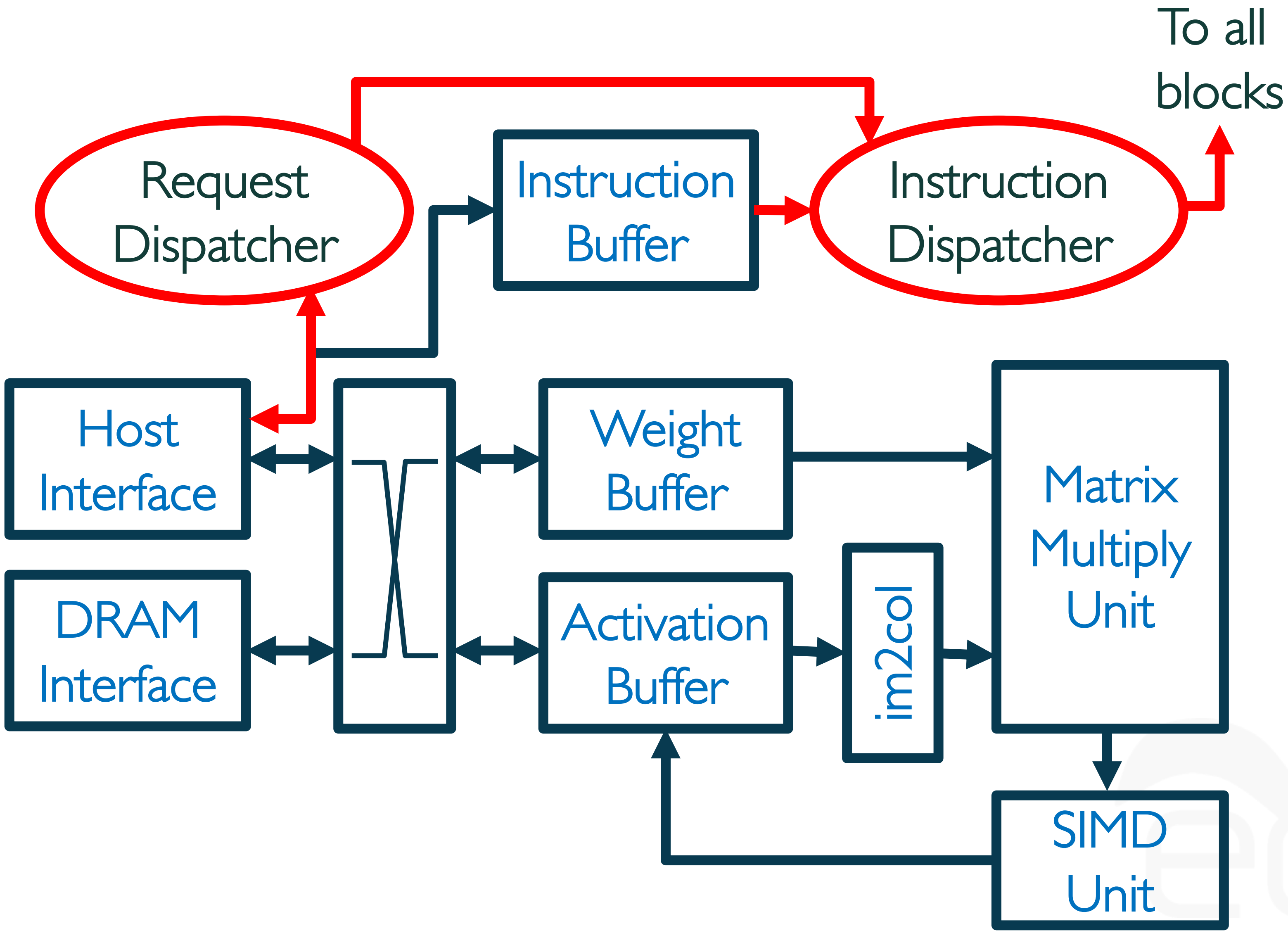


Canonical Systolic Array Based Inference Accelerator



Pareto Optimal Frontier

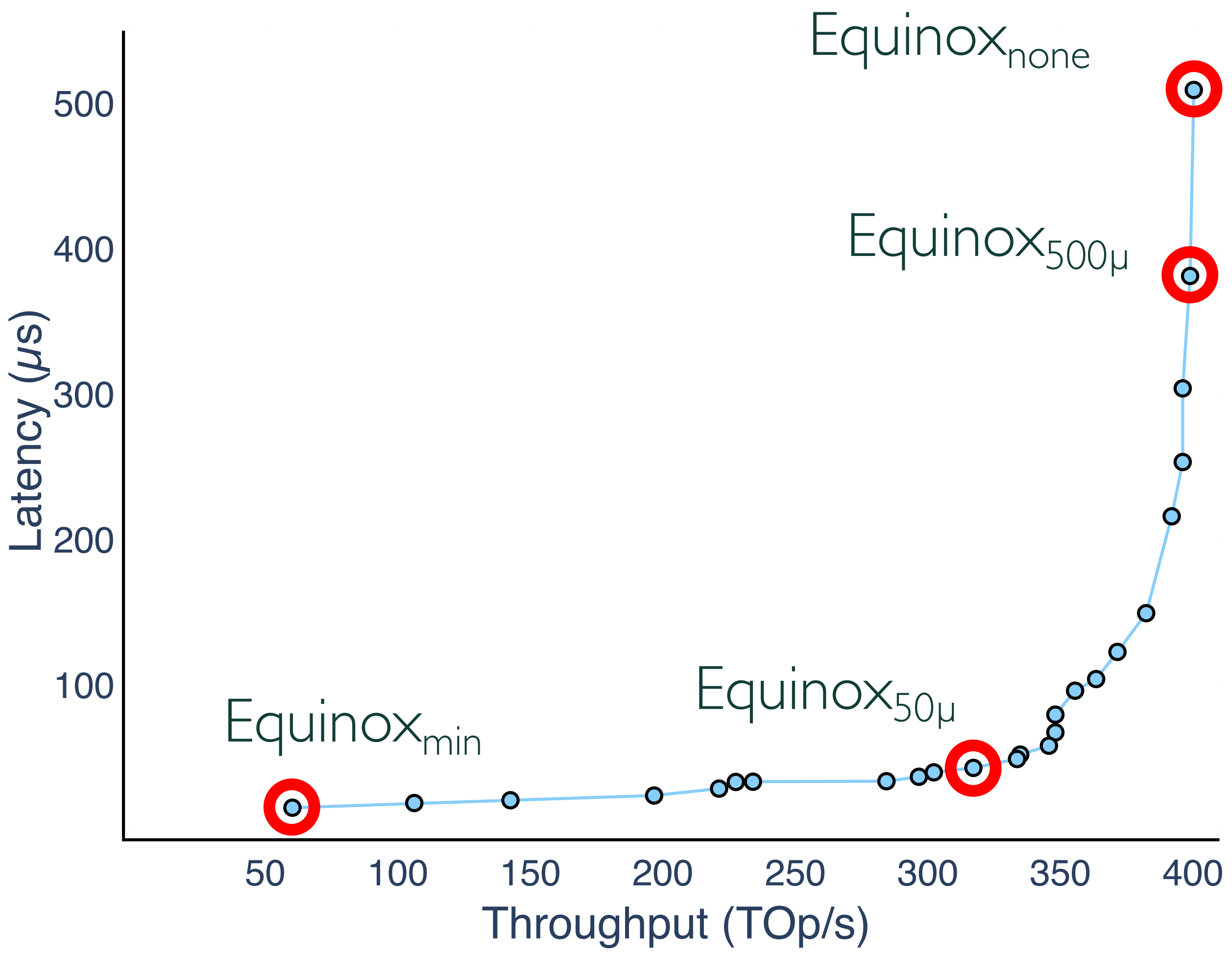
Equinox: Piggybacking Training on Inference



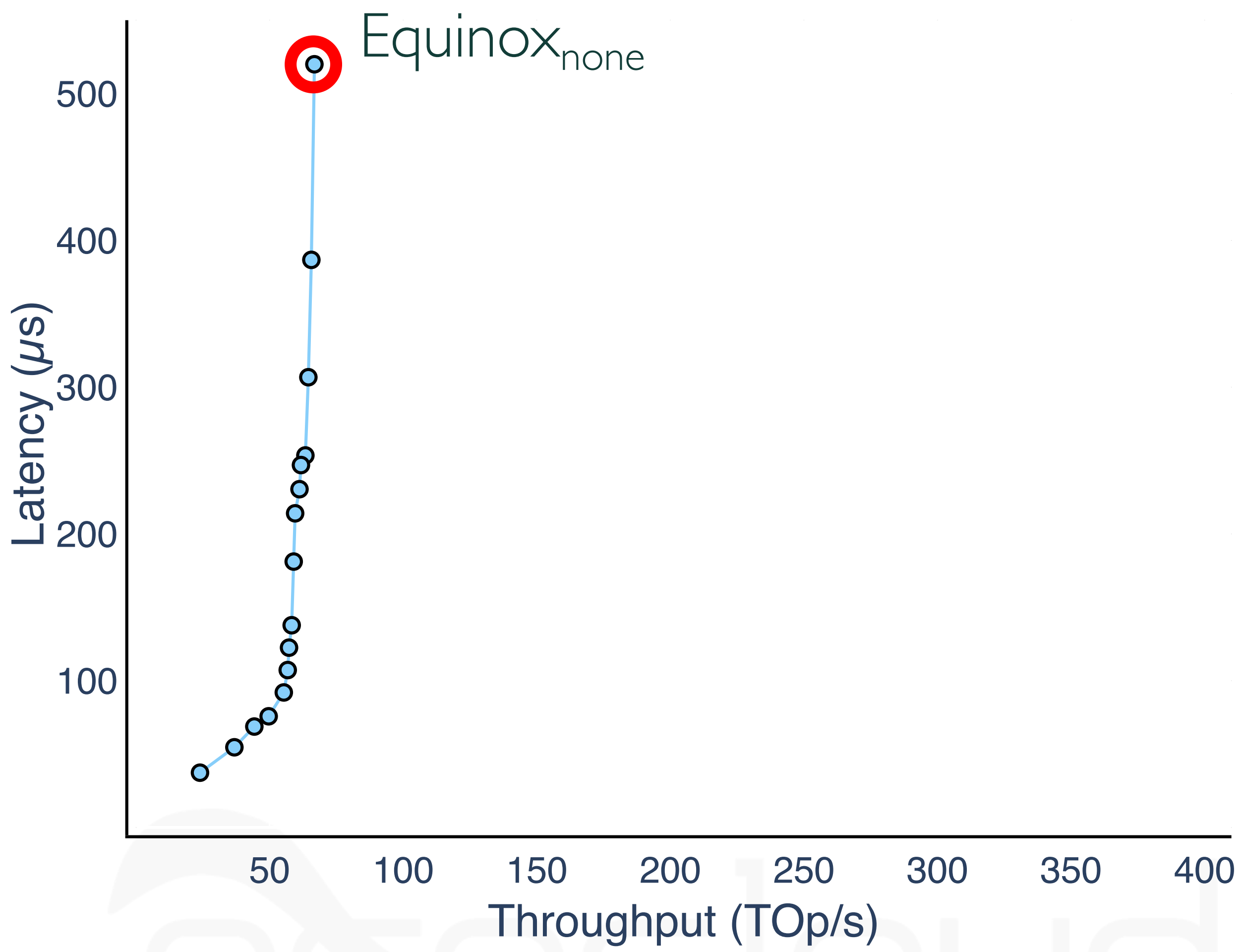
Prioritized inference/training request dispatch

- Runs requests in units of batches
- Runs requests to completion
- Guarantees tail latency on inference requests
- Dispatch similar to Microsoft Brainwave

Equinox with Various Latency Constraints



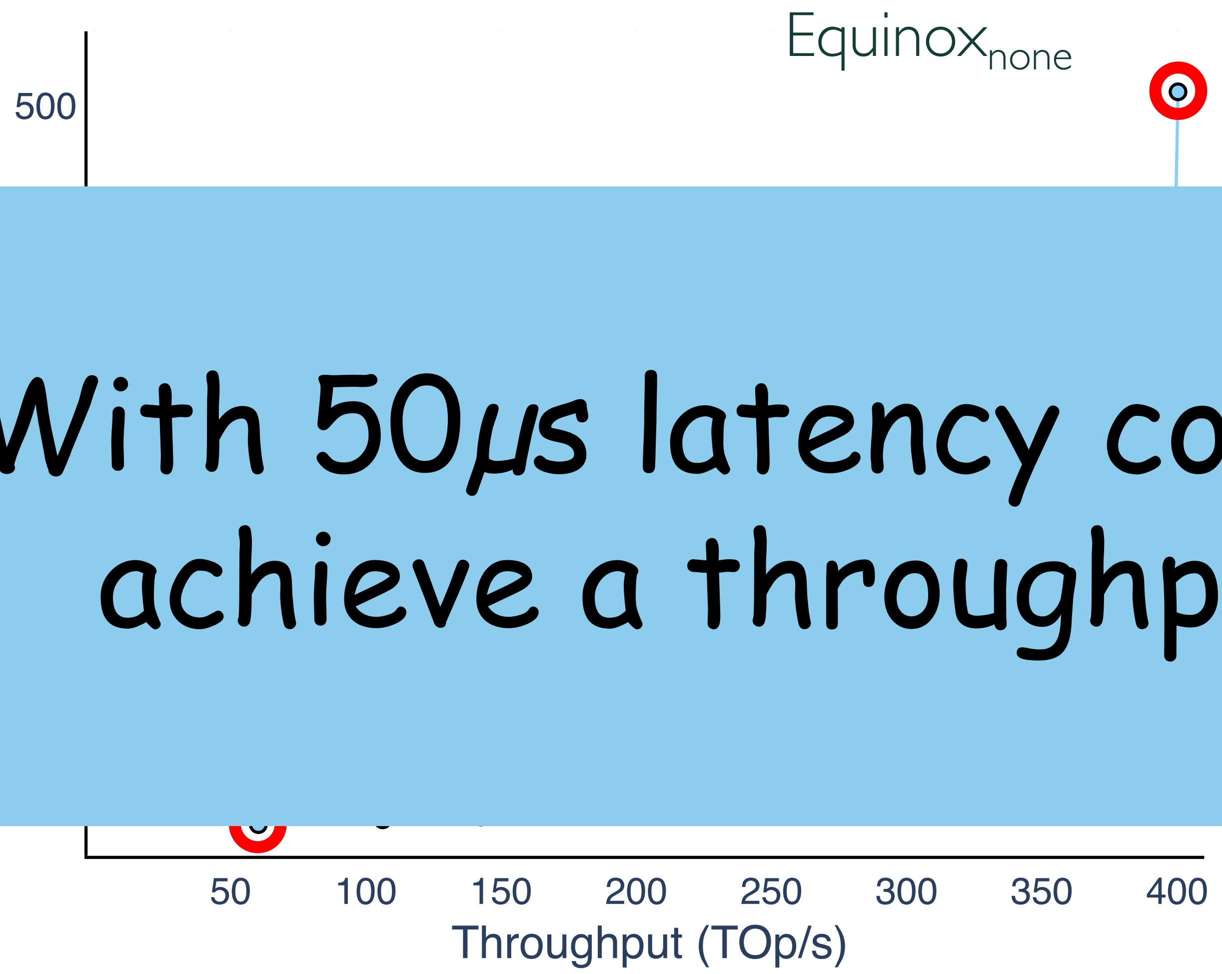
hbfp8



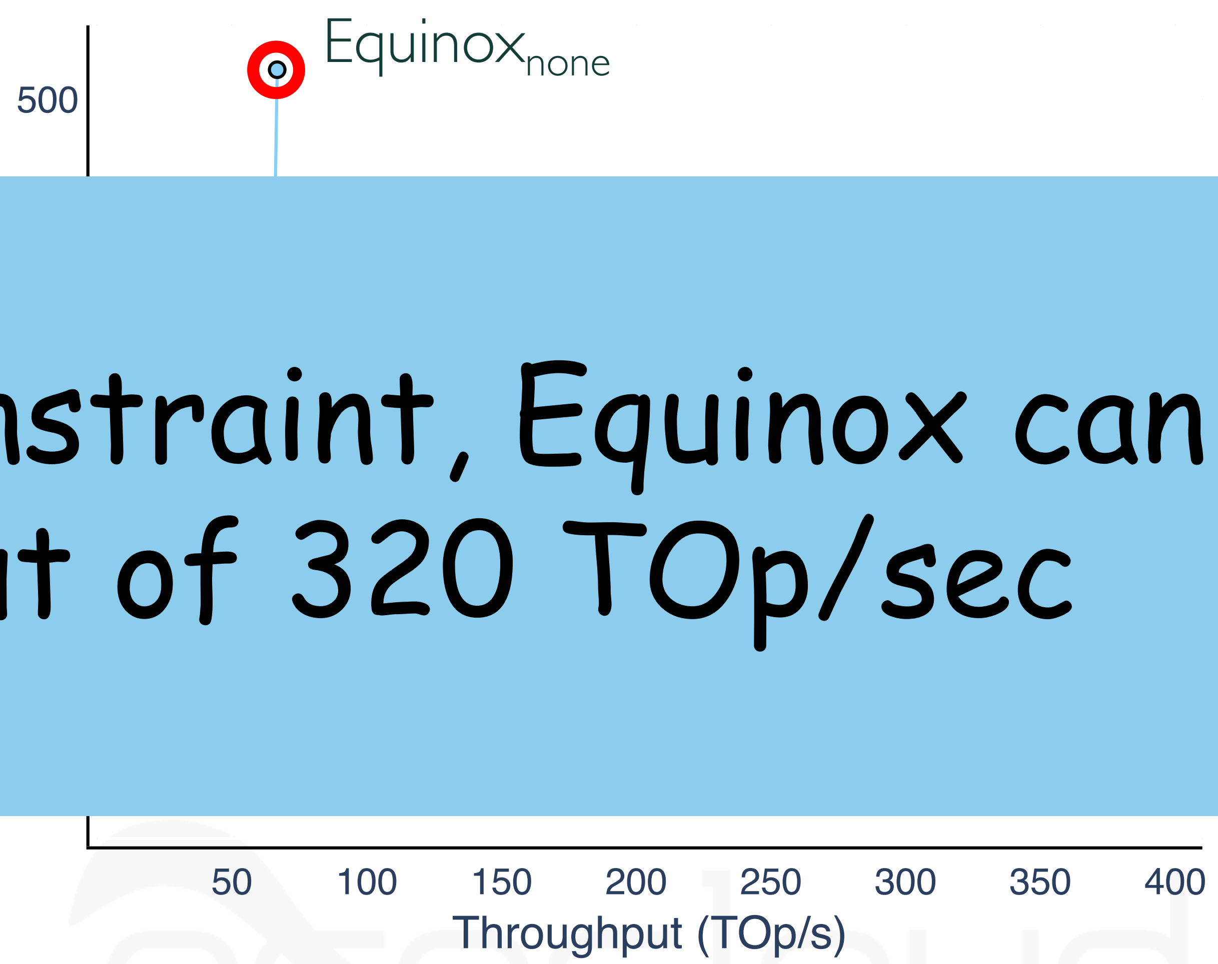
bfloat16

Equinox with Various Latency Constraints

With 50μs latency constraint, Equinox can achieve a throughput of 320 TOp/sec



hbfp8



bfloat16

Summary

There is a divergence in AI infrastructure

Can we bridge the gap?

- hbfp enables training with fixed point
- Batching presents a trade-off between latency and power, throughput in custom inference accelerators
- Can prioritize request dispatch to honor latency constraints while training (for free) on an inference accelerator
- For results on piggybacked training see the MICRO'21 paper

See

- parsa.epfl.ch/coltrain
- github.com/parsa-epfl/HBFPEmulator

Thank You!

For more information please visit us at
ecocloud.ch