Accuracy Boosters

Epoch-Driven Mixed-Mantissa Block Floating Point for DNN Training

Simla Burcu Harma, Ayan Chakraborty,
Babak Falsafi, Martin Jaggi, Yunho Oh

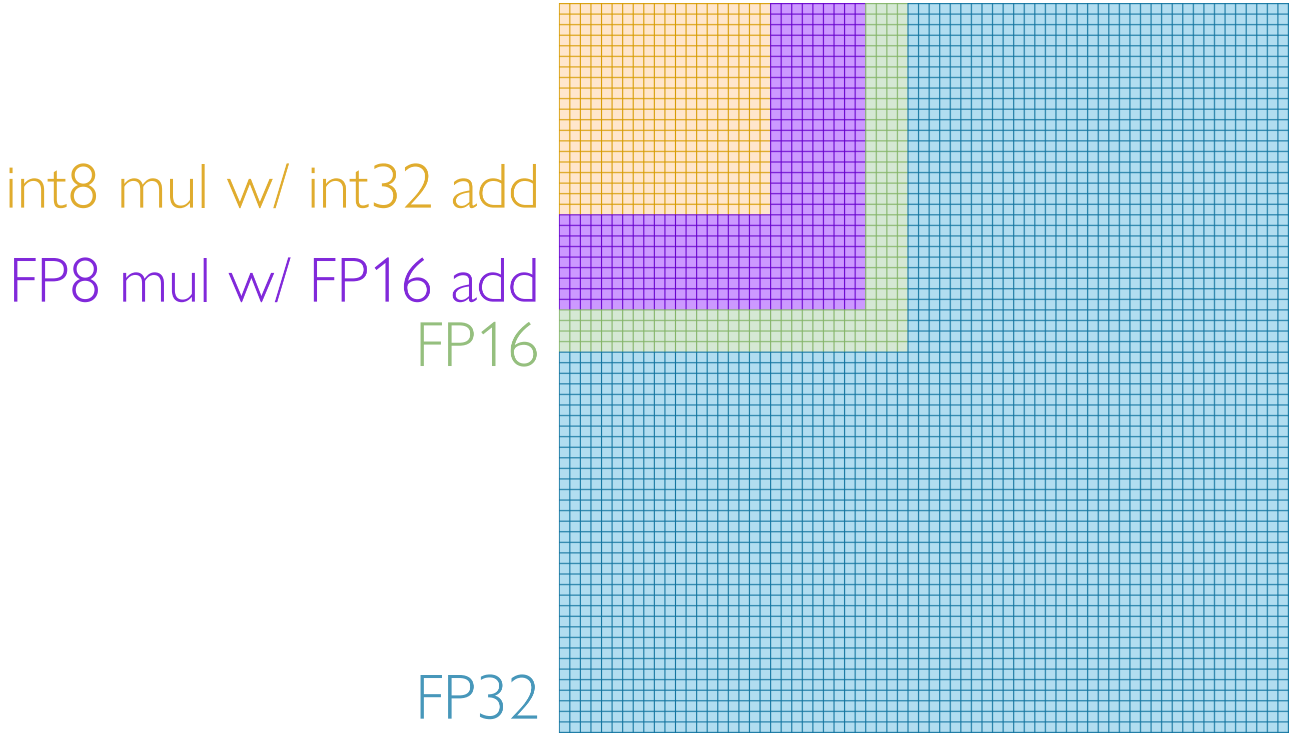
PARALLEL SYSTEMS ARCHITECTURE LAB

parsa.epfl.ch/coltrain

EPFL

Dense and Accurate DNN Training

PARALLEL SYSTEMS ARCHITECTURE LAB



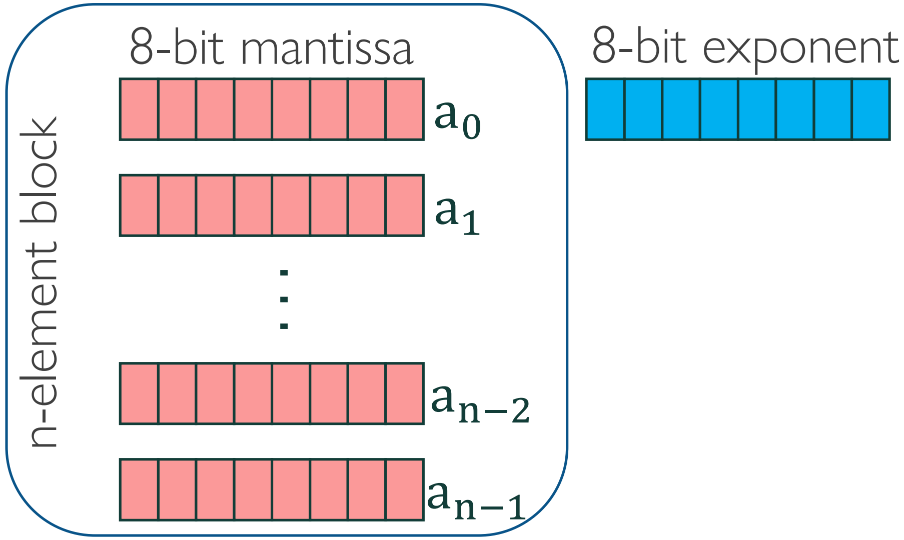
Relative logic area of multiply-and-accumulate (MAC) using different datatypes on the same silicon [Fox et al., ICLR'21]

Goal: Training DNNs using fixed-point arithmetic with FP32 accuracy

A Narrow Bitwidth Format: HBFP

- High accuracy of floating point
- The superior hardware density of fixed point

- Block Floating Point (BFP) for dot products (> 90% ops)
- Floating Point for activations and other arithmetic

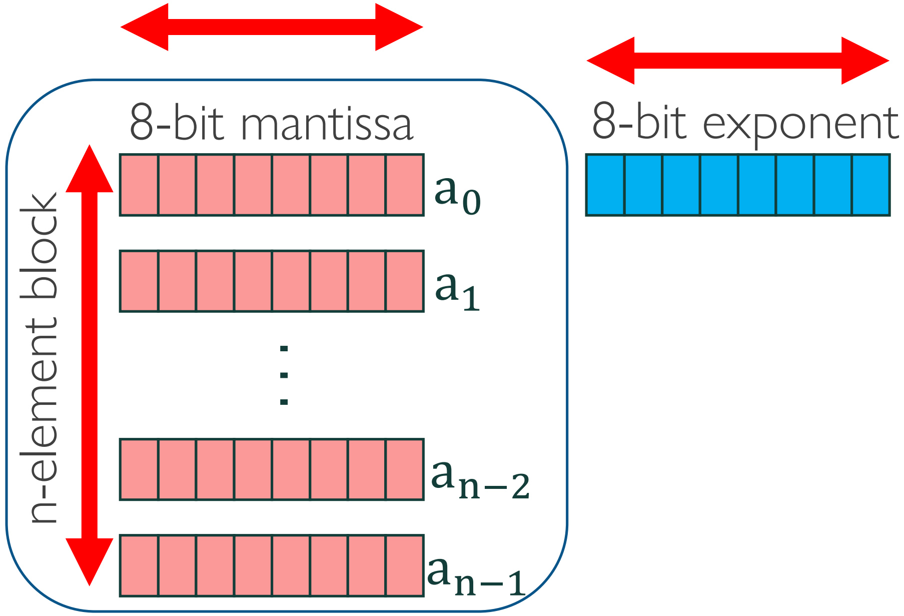


BFP representation with an exponent per tensor

Drumond et al., Training DNNs with Hybrid Block Floating Point, NeurIPS'18

A Narrow Bitwidth Format: HBFP

- Explore the HBFP parameter space
  - Maximizing block size
  - Minimizing mantissa bits
  - ⇒ Study the tensor distribution similarities
  - ⇒ Analyze the loss landscapes



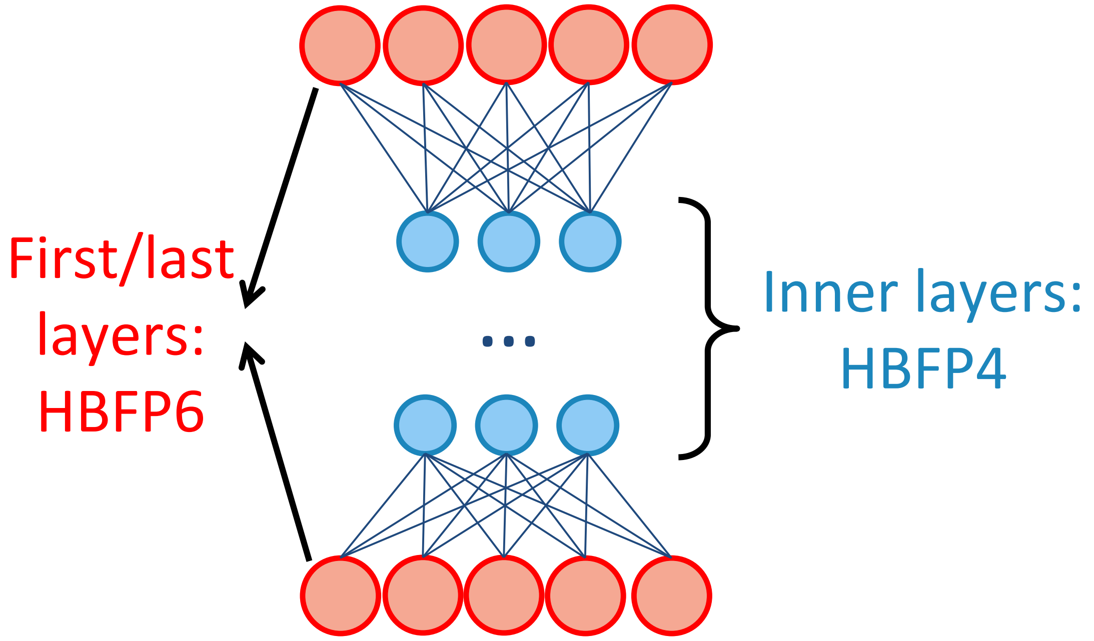
BFP representation with an exponent per tensor

The parameter space of HBFP is yet to be explored!

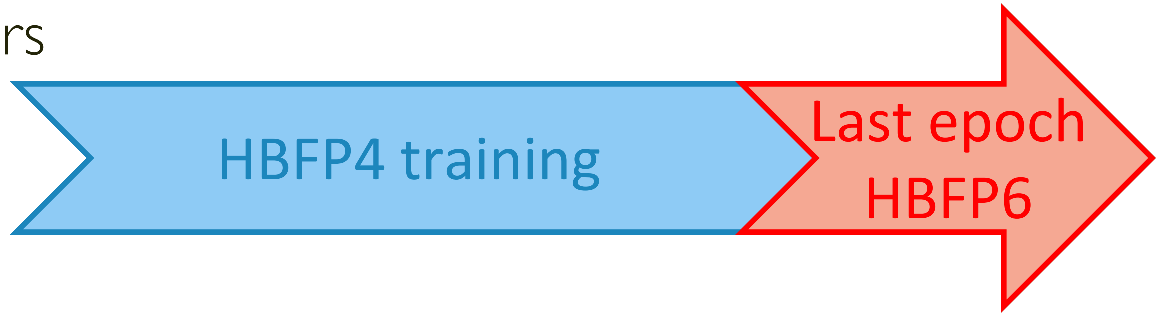
Contributions

PARALLEL SYSTEMS ARCHITECTURE LAB

- Explore the HBFP parameter space
  - Maximizing block size
  - Minimizing mantissa bits
  - ⇒ Study the tensor distribution similarities
  - ⇒ Analyze the loss landscapes

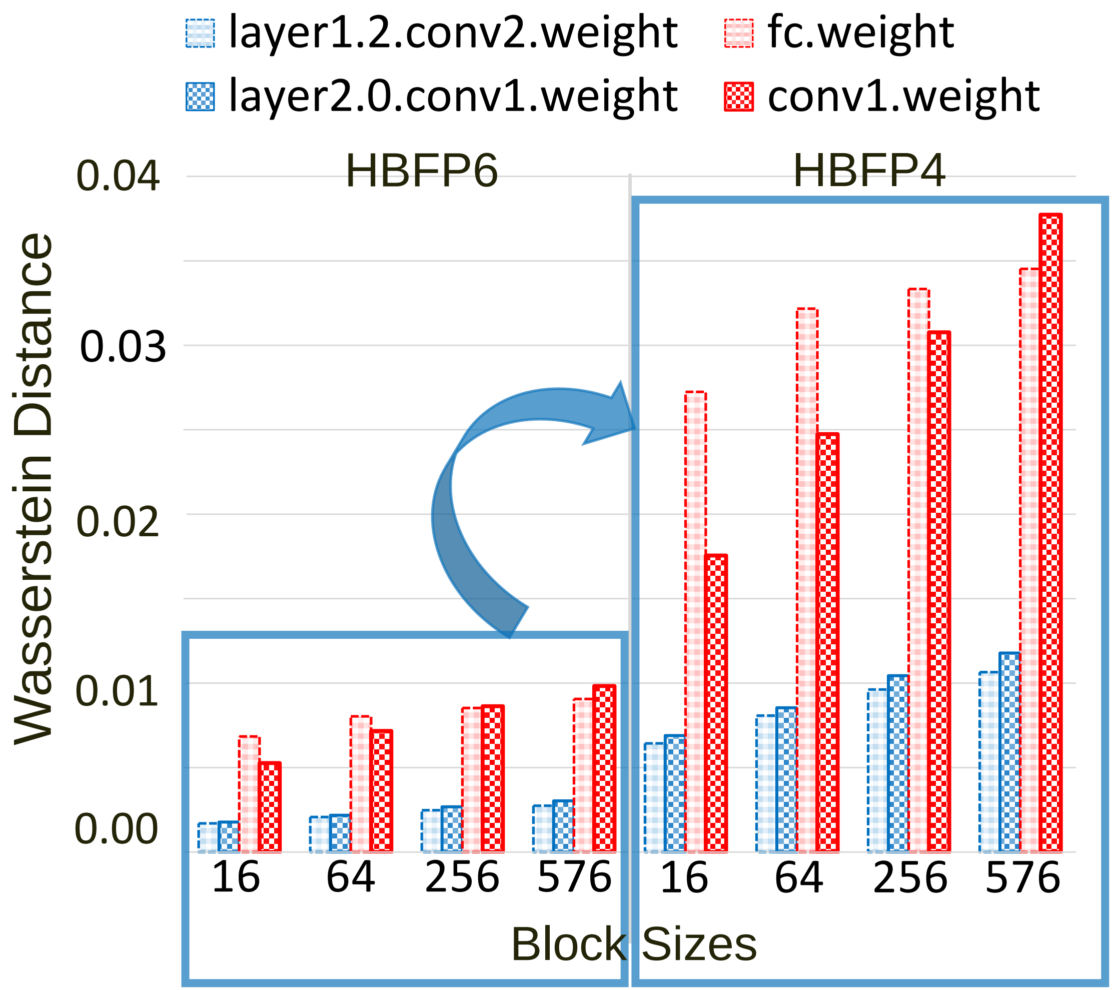


- Accuracy Boosters
  - HBFP6 only in the last epoch and first/last layers
  - HBFP4 for the rest (99.7% of ops)



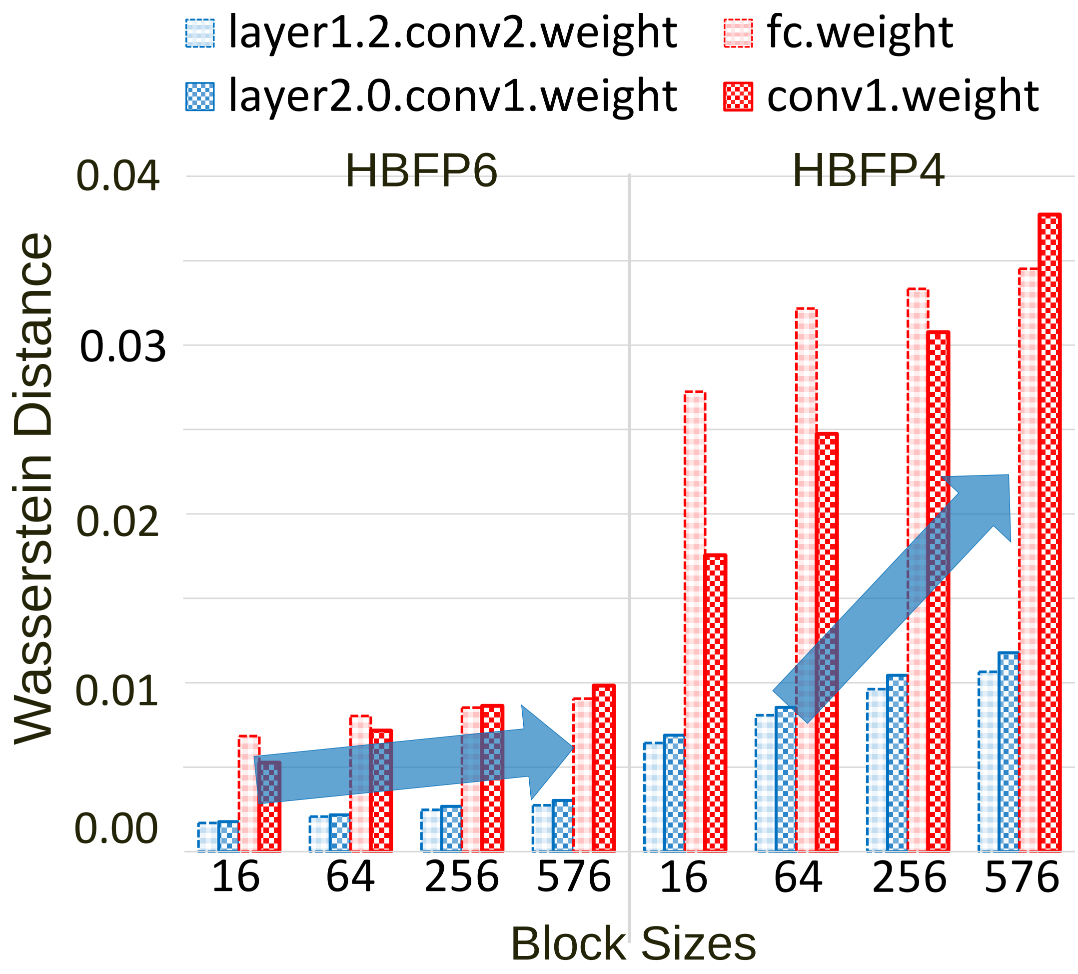
We can get the HW benefits of HBFP4 while maintaining FP32 accuracies

Tensor Distributions: HBFP4 vs. HBFP6



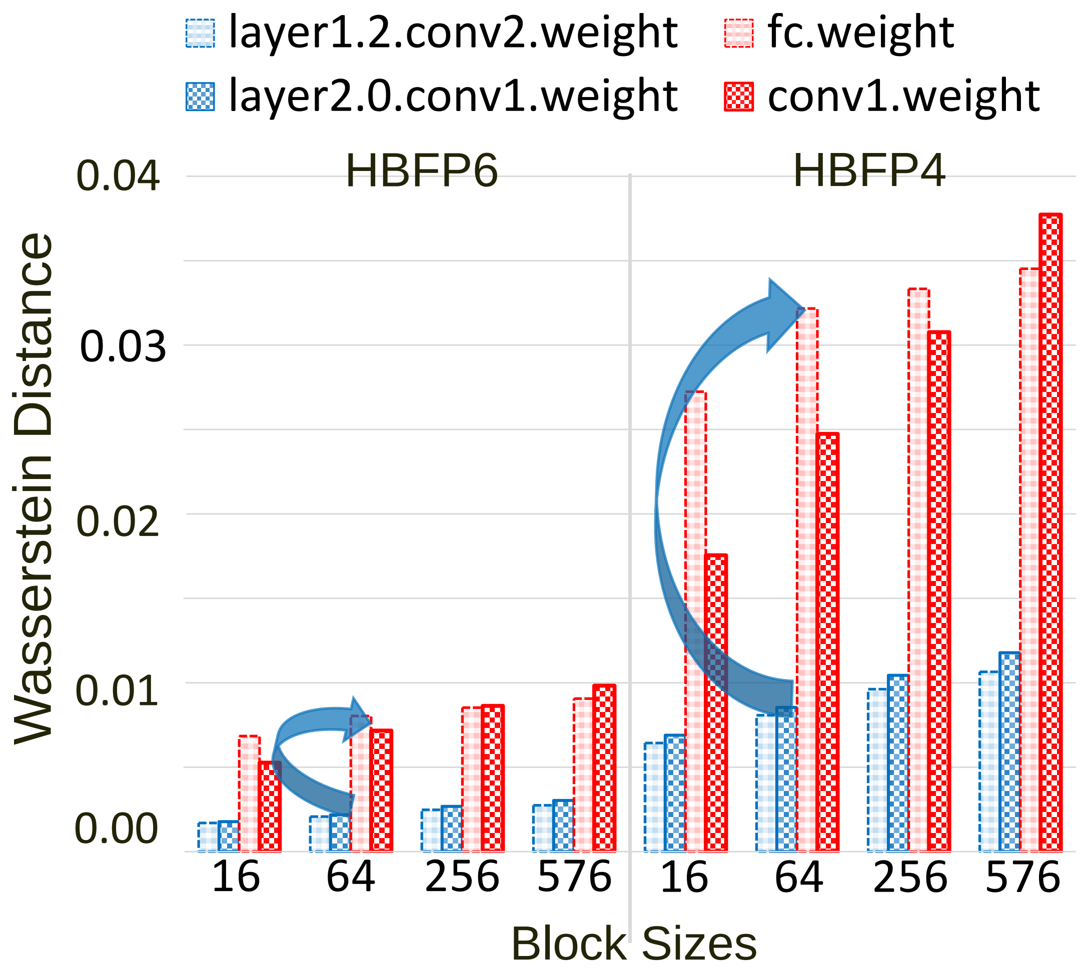
Tensor distributions are much more distorted for HBFP4 compared to HBFP6

Tensor Distributions: Block Sizes



HBFP6 is not sensitive to the block size, while HBFP4 is sensitive

Tensor Distributions: First/Last Layers

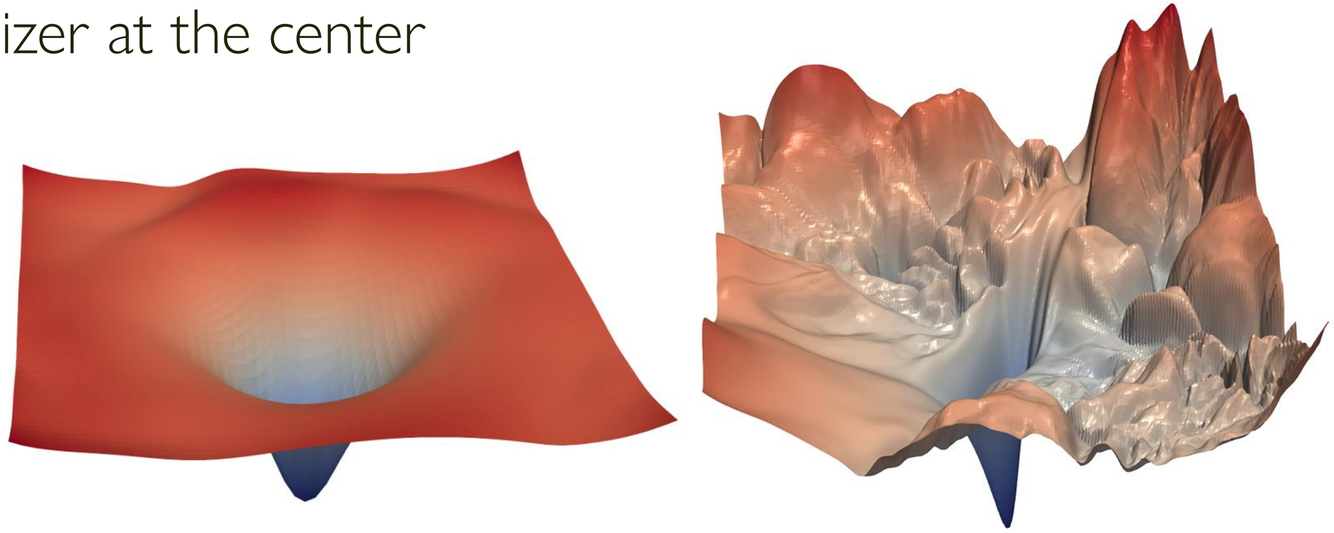


Wasserstein distance of first/last layers is higher than the other layers

Analyzing the Loss Landscapes

- Plot the landscape around the current position of the minimizer

- Dimensionality reduction
  - Pick two random directions and form a plane
  - Add a third dimension → will be the loss value calculated at each point within that plane
  - Position the current state of the minimizer at the center



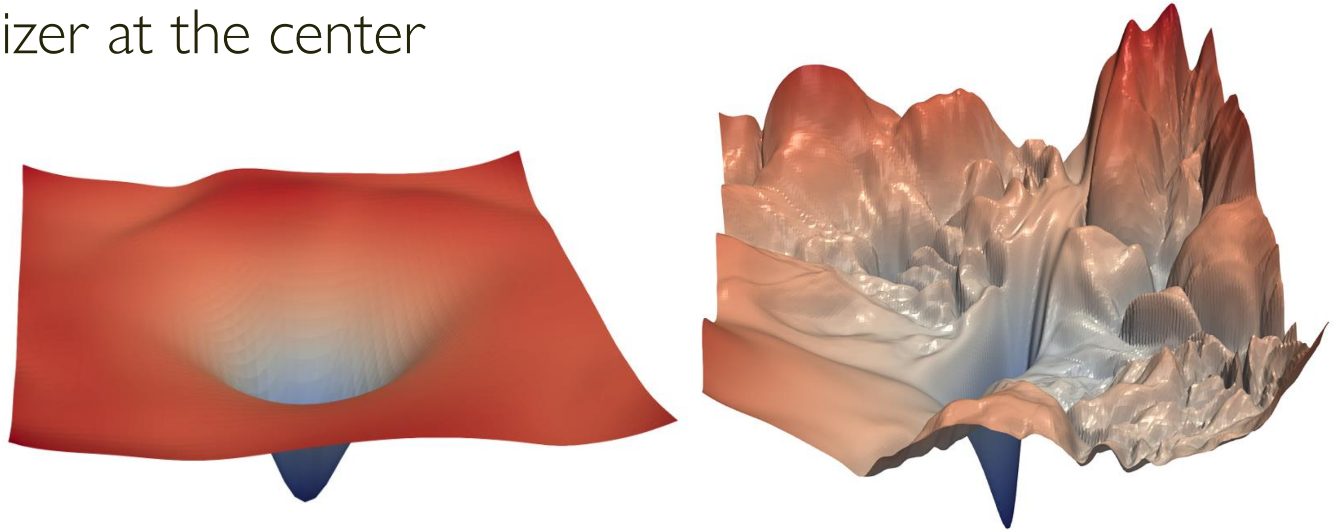
Li et al., Visualizing the Loss Landscape of Neural Nets, NeurIPS'18

Analyzing the Loss Landscapes

- Plot the landscape around the current position of the minimizer

- Dimensionality reduction
  - Pick two random directions and form a plane
  - Add a third dimension → will be the loss value calculated at each point within that plane
  - Position the current state of the minimizer at the center

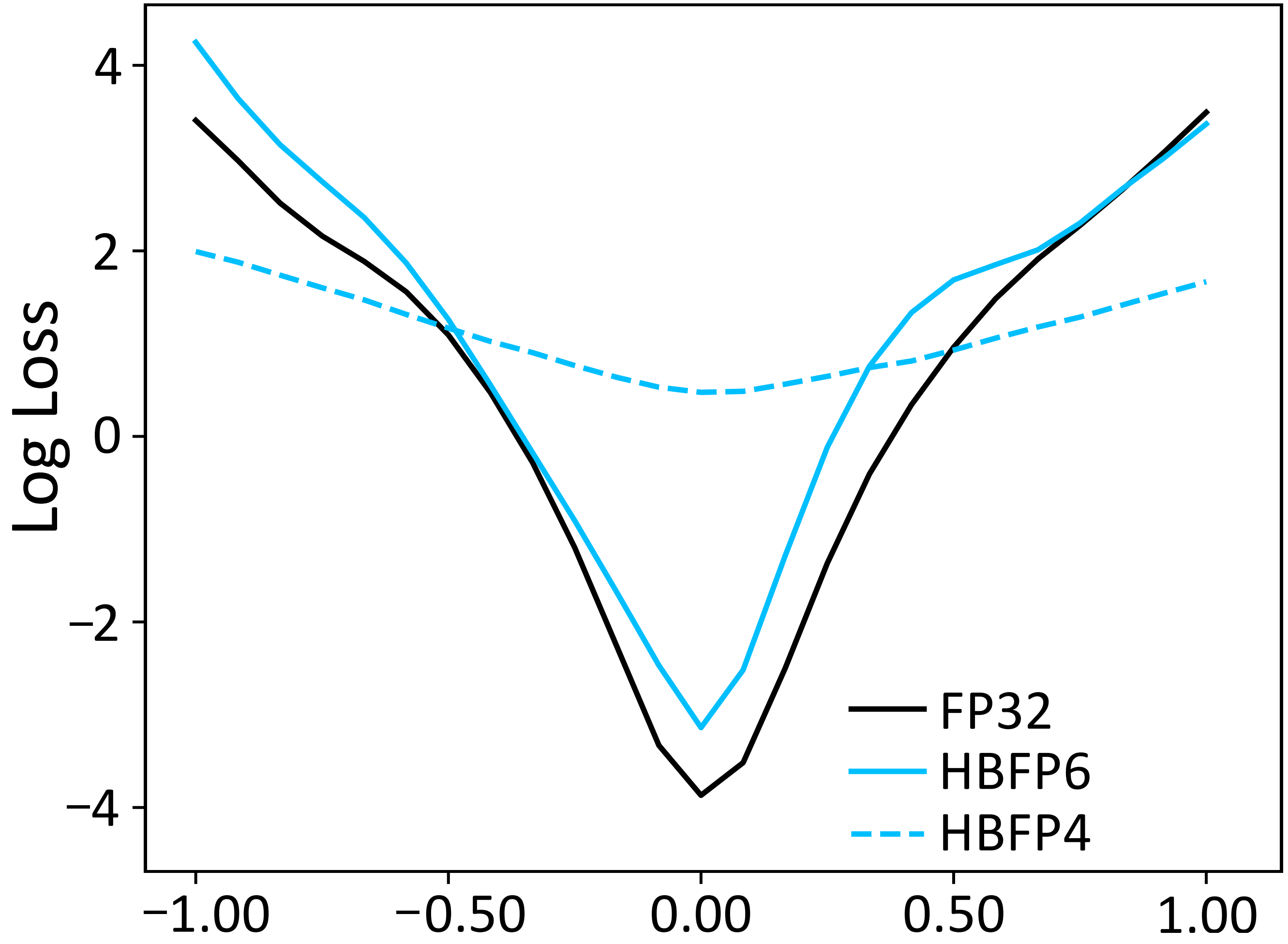
- Loss value → Optimization
- Flatness → Generalization



Loss landscapes provides information for the interplay between generalization & optimization

Li et al., Visualizing the Loss Landscape of Neural Nets, NeurIPS'18

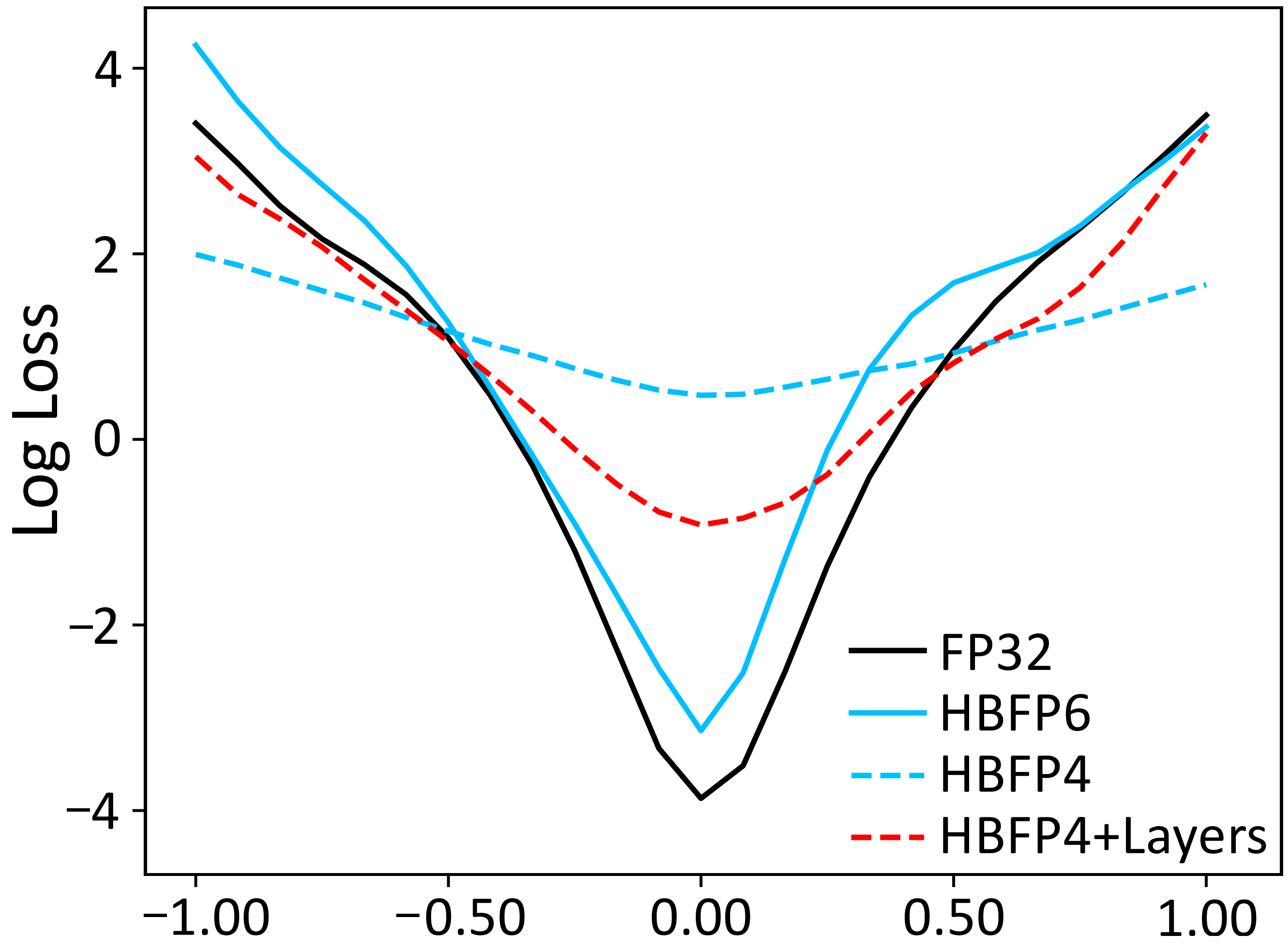
Loss Landscapes: FP32 vs Standalone HBFP



HBFP4 fails to converge to good minimum in contrast to HBFP6

Loss Landscapes: HBFP4+Layers

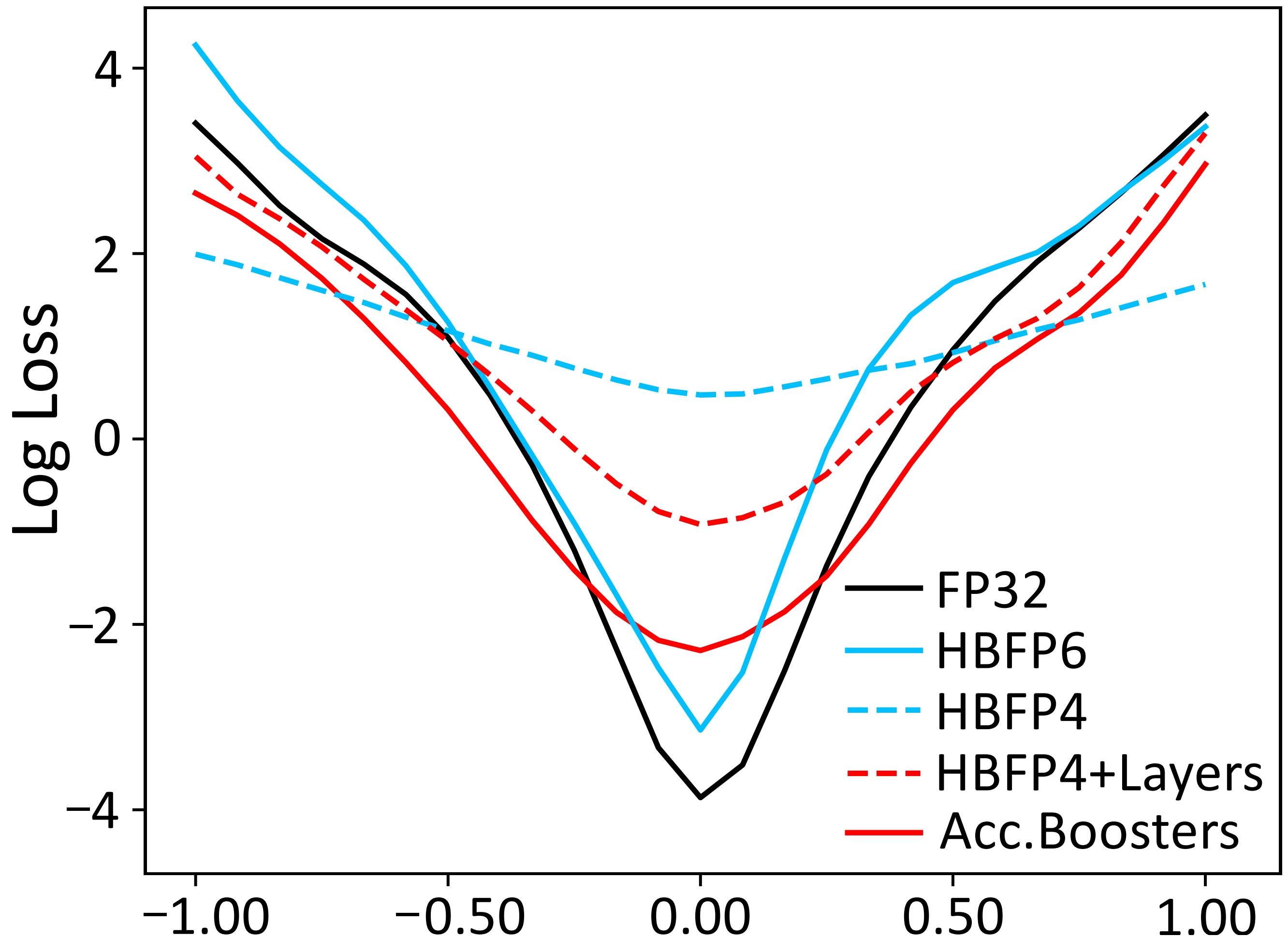
PARALLEL SYSTEMS ARCHITECTURE LAB



Increase in accuracy but there is still imbalance btw. optimization and generalization

Loss Landscapes: Accuracy Boosters

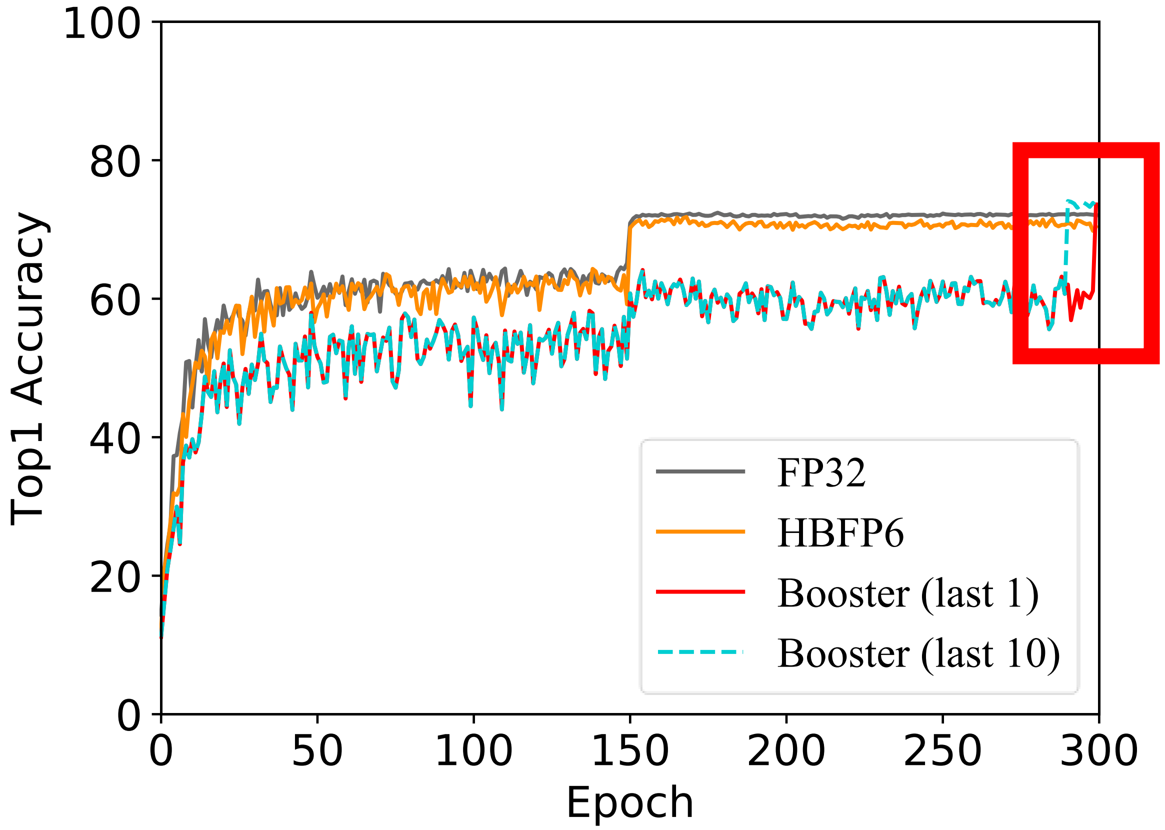
PARALLEL SYSTEMS ARCHITECTURE LAB



Accuracy Boosters is the recipe to reach the sweet spot between generalization & optimization

Accuracy Boosters: Model Accuracies

DenseNet40 on CIFAR100

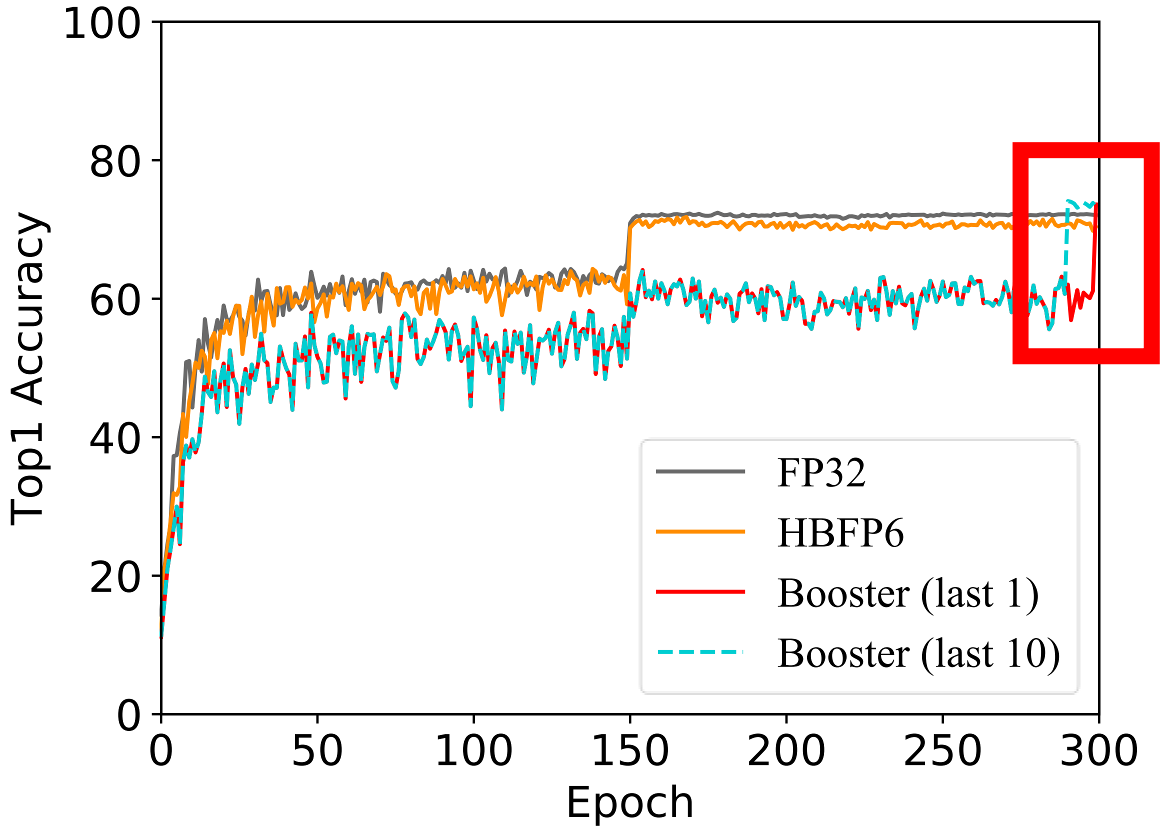


FP32 level accuracy while using HBFP4 for majority of operations with 21.3x higher density

Accuracy Boosters: Model Accuracies

PARALLEL SYSTEMS ARCHITECTURE LAB

DenseNet40 on CIFAR100



Transformer-Base trained on IWSLT’14 De⟶En

Configuration	BLEU Score
FP32	34.77
HBFP6	34.47
HBFP4	32.64
Booster	**36.08**

FP32 level accuracy while using HBFP4 for majority of operations with 21.3x higher density

Summary

PARALLEL SYSTEMS ARCHITECTURE LAB

- HBFP has a rich parameter space → Opportunities to increase arithmetic density

- Explore HBFP parameters
  - Block size ⇒ Tensor distribution similarities
  - Mantissa bitwidth ⇒ Loss landscapes

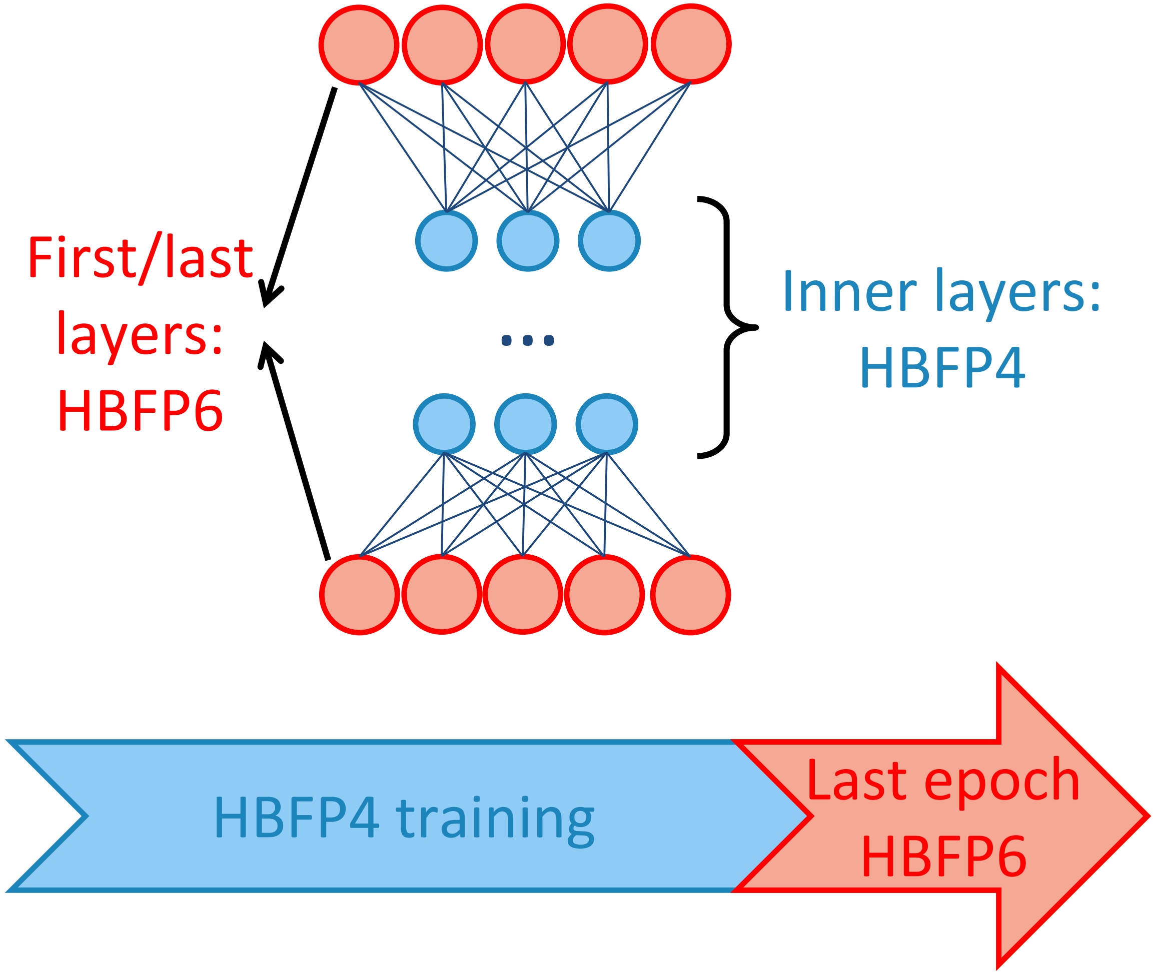
- Accuracy Boosters: Mixed-mantissa BFP across layers and epochs

- Accuracy Boosters employs HBFP4 for the 99.7% of total operations
  - FP32-level accuracies
  - Up to 21.3× higher arithmetic density over FP32

Thank You!

For more information please visit us at parsa.epfl.ch

or contact me via simla.harma@epfl.ch



Wasserstein Distance

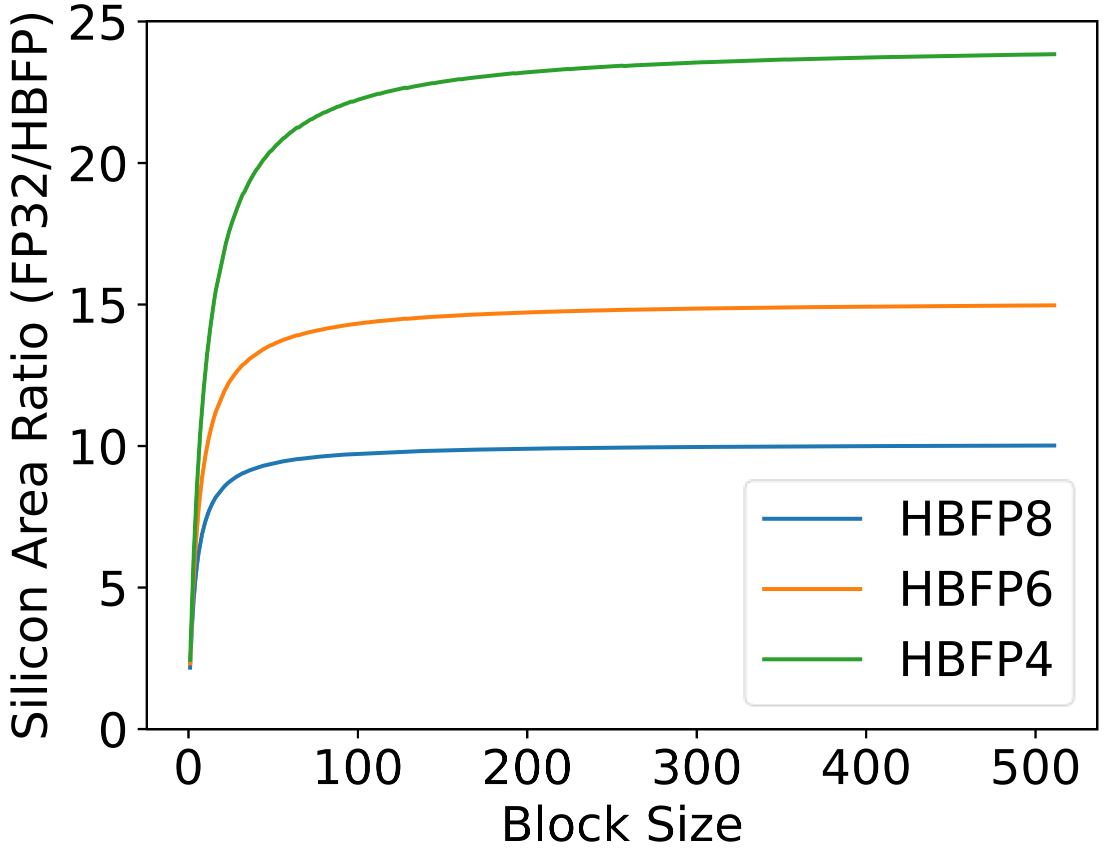
$$W(P,Q) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}[||x-y||],$$

where $\Pi(P,Q)$ is the set of all joint distributions $\gamma(x,y)$ whose marginal distributions are equal to P and Q

- $\gamma(x,y)$ can be interpreted as the amount of mass that must be transported from x to y to transform P to Q

HBFP Parameter Space: Why Minimize?

PARALLEL SYSTEMS ARCHITECTURE LAB



Considerable power and area savings!

Hardware Support

- HBFP4 hardware can support HBFP6 operations in 4 steps

- Lower HBFP6 operations into HBFP4:

$$A \times B = (2^4 \cdot A_{\mathrm{HI}} + A_{\mathrm{LO}}) \times (2^4 \cdot B_{\mathrm{HI}} + B_{\mathrm{LO}})$$

$$A \times B = 2^8 \cdot A_{\mathrm{HI}} \times B_{\mathrm{HI}} + 2^4 \cdot (A_{\mathrm{HI}} \times B_{\mathrm{LO}} + A_{\mathrm{LO}} \times B_{\mathrm{HI}}) + A_{\mathrm{LO}} \times B_{\mathrm{LO}}$$

- Support 2^4 and 2^8 by modifying the BFloat16 accumulators
  - Offset the exponent by 4 or 8
  - With little hardware can achieve lower HBFP6 throughput