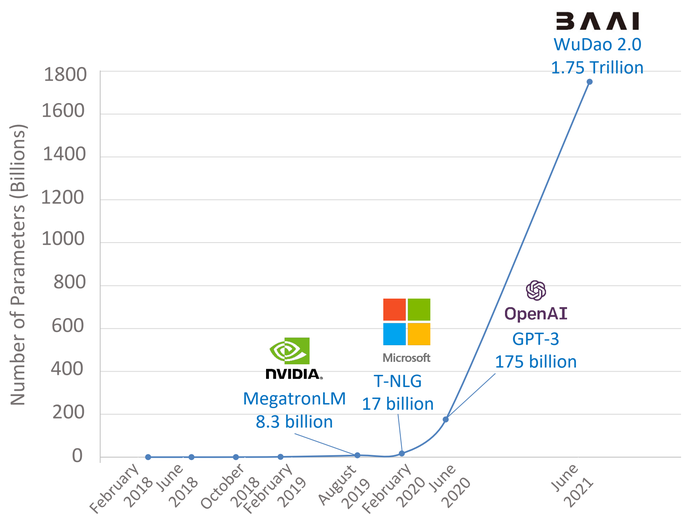
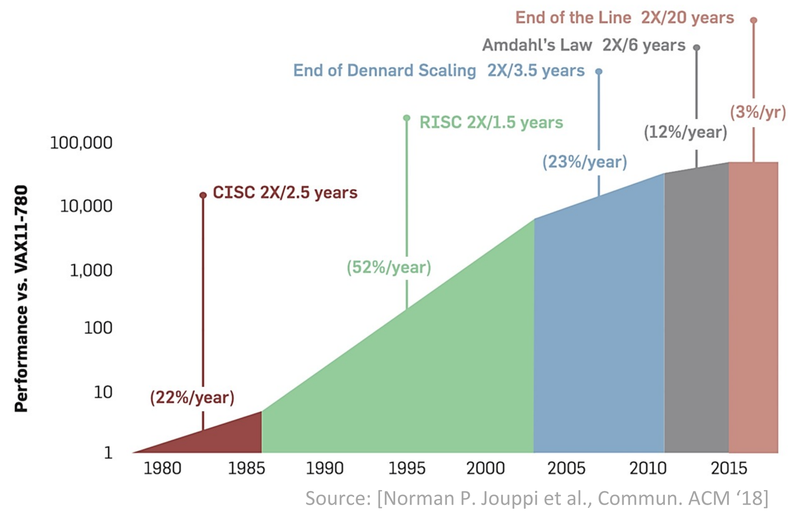
Accuracy Booster

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

Model sizes keep increasing!



Moore's Law is dying



Source: [Norman P. Jouppi et al., Commun. ACM '18]

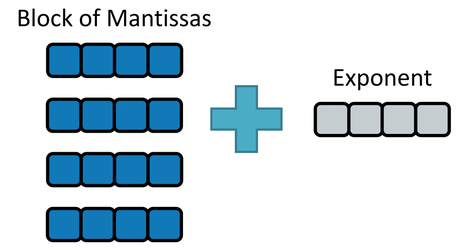
How to make DNN training denser?

Training traditionally uses power-inefficient floating-point arithmetic for accuracy.
HBFP brings fixed-point efficiency to training.

Goal: Training DNNs using 4-bit fixed-point arithmetic with FP32 accuracy

Hybrid Block Floating Point

High accuracy of floating point and the superior hardware density of fixed point



$$a \cdot b = \sum_{i=1}^{N} \left((m_i^a \times 2^{e_a}) \times (m_i^b \times 2^{e_b}) \right)$$
$$= 2^{e_a + e_b} \times \boxed{(m^a \cdot m^b)}$$

Fixed-point dot product

Prior work only studied HBFP's design space for power-of-two mantissa bitwidths (e.g., 2, 4, 8-bits)

The parameter space of HBFP is yet to be explored!

HBFP Parameter Space

- Optimizing block size
  → Study the tensor distribution similarities
- Minimizing mantissa bits
  → Analyze loss landscapes

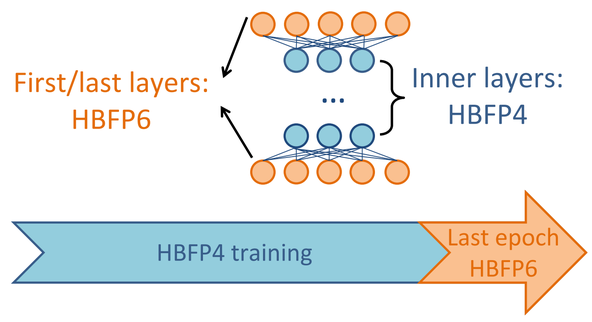
Parameters impact both model accuracy and accelerator density

Mixed-Mantissa HBFP: Accuracy Booster

- A key advantage of BFP is fixed exponent
- We can keep the exponent and vary the mantissa bits

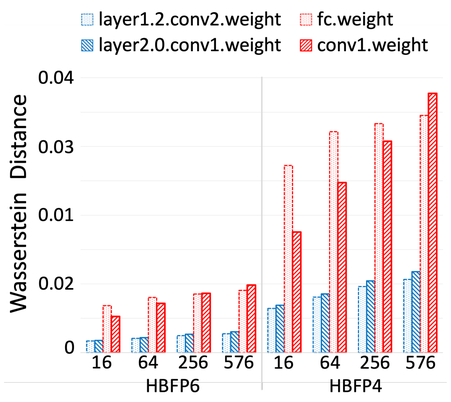
- Epochs have varying precision requirements
- First/last layers of DNN models have a large impact on accuracy

- Accuracy Booster: HBFP6 only in the last epoch and first/last layers, HBFP4 for the rest



Hardware benefits of HBFP4 while maintaining FP32 accuracies

Tensor Distributions: Wasserstein Distance



- Tensor distributions are much more distorted for HBFP4 compared to HBFP6

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

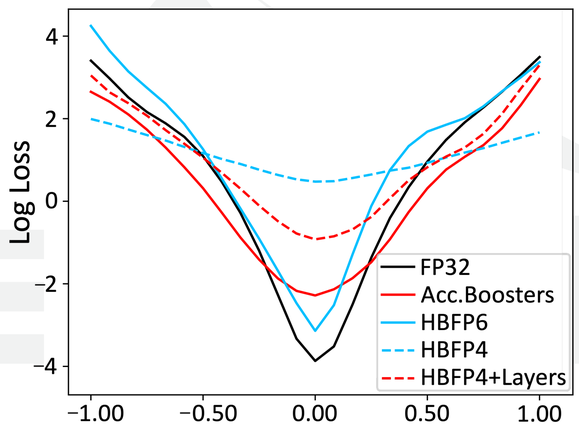
- Wasserstein distances of first/last layers are higher than the other layers

Wasserstein Distance is a viable metric to measure similarity btw. tensors

Analyzing the Loss Landscapes

- Plot the landscape around the current position of the minimizer

- Dimensionality reduction by random directions with filter normalization



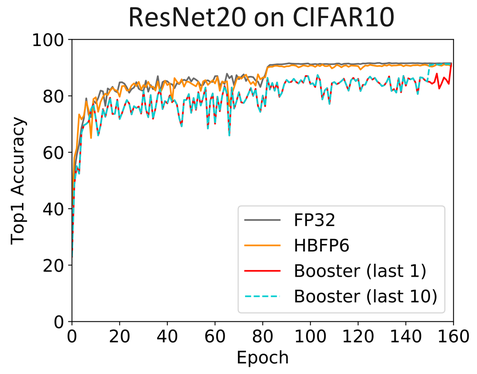
Loss landscapes provides information for the interplay between generalization & optimization!

Accuracy Booster: Experimental Results

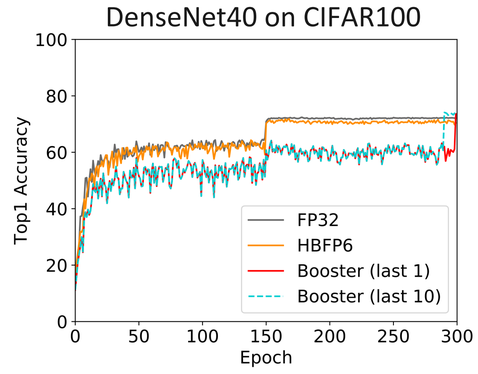
Use HBFP4 for >99% of the operations;
use HBFP6 only for the first/last layers and last epoch.

FP32 accuracy with up to
23 times more area efficient hardware.

ResNet20 on CIFAR10



DenseNet40 on CIFAR100



Transformer-Base trained on IWSLT'14 De→En

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

FP32 level accuracy with 23x higher density

ML&O

PARSA PARALLEL SYSTEMS ARCHITECTURE LAB