Bandwidth-Efficient Deep Learning —— Challenges and Trade-offs

Song Han

Assistant Professor, MIT EECS Department*

77 Massachusetts Avenue, Room 38-344
Cambridge, MA, 02139
songhan@mit.edu

HAN LAB

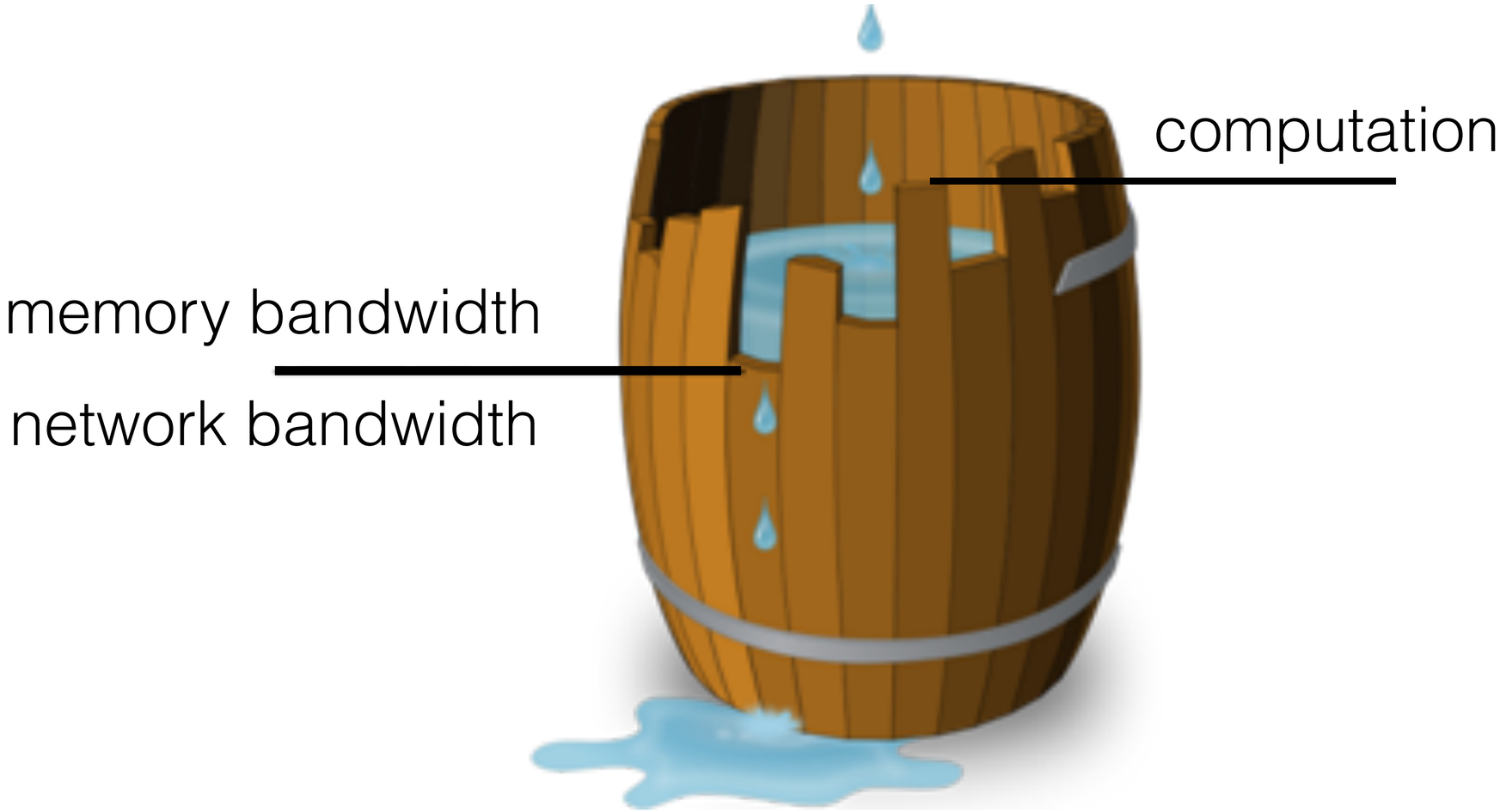
*starting from July 1, 2018

MIT

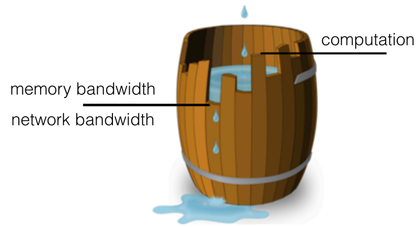
Deep Learning Hardware: Challenges and Trade-offs

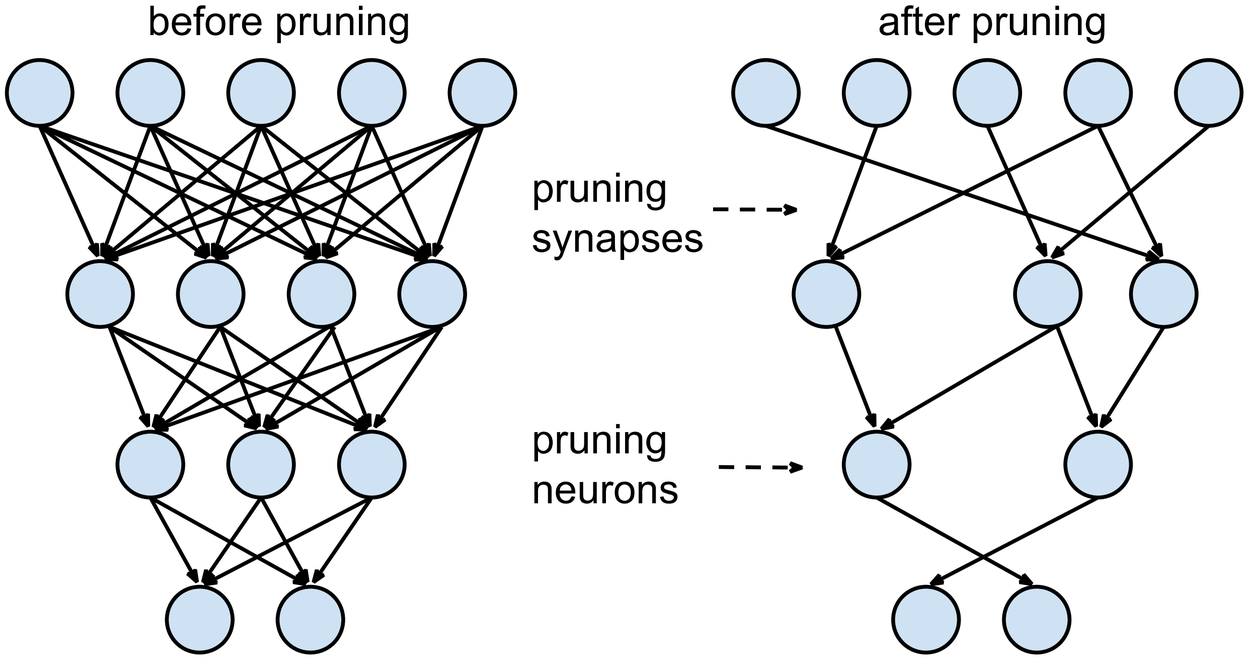
- Training vs. Inference
- Latency vs. Throughput
- Edge vs. Cloud
- Accuracy vs. Model Size
- Flexibility/Programmability vs. Efficiency
- Computation bounded vs. Memory bounded
- Time to market vs. Efficiency
- Mature ecosystem vs. Emerging technology

Data Movement is the Bottleneck



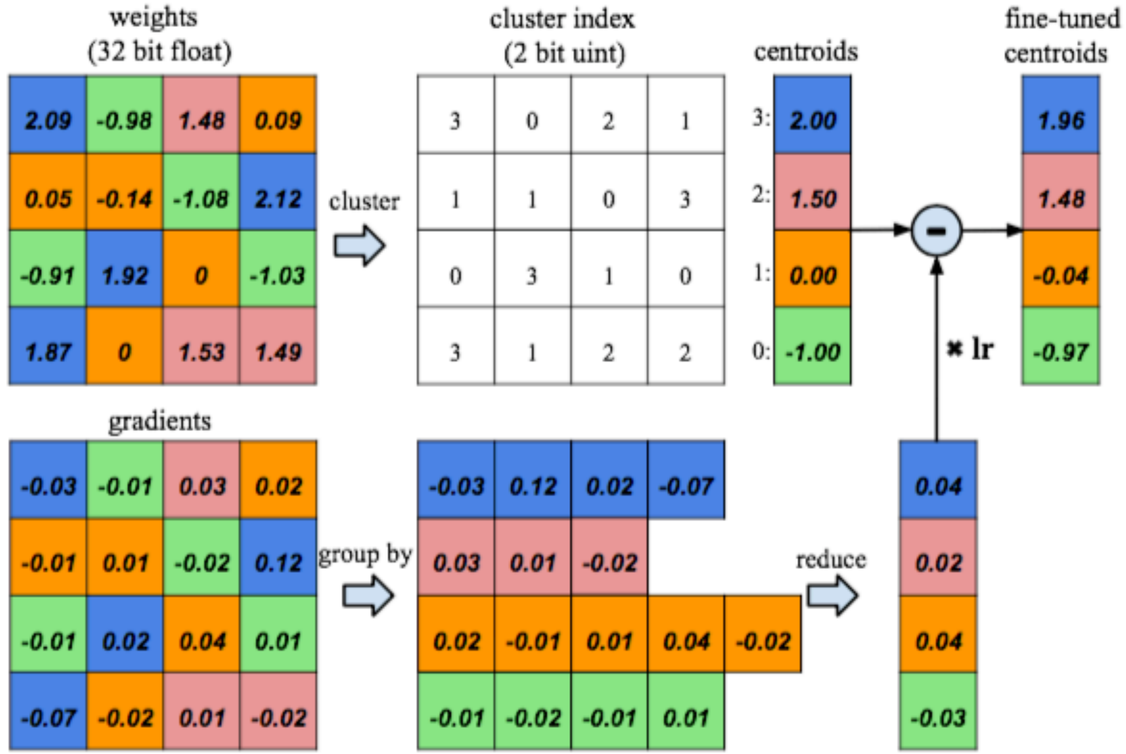
Save Memory Bandwidth: Deep Compression





Pruning

Han et al [NIPS'15]

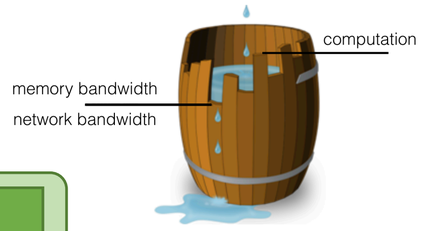


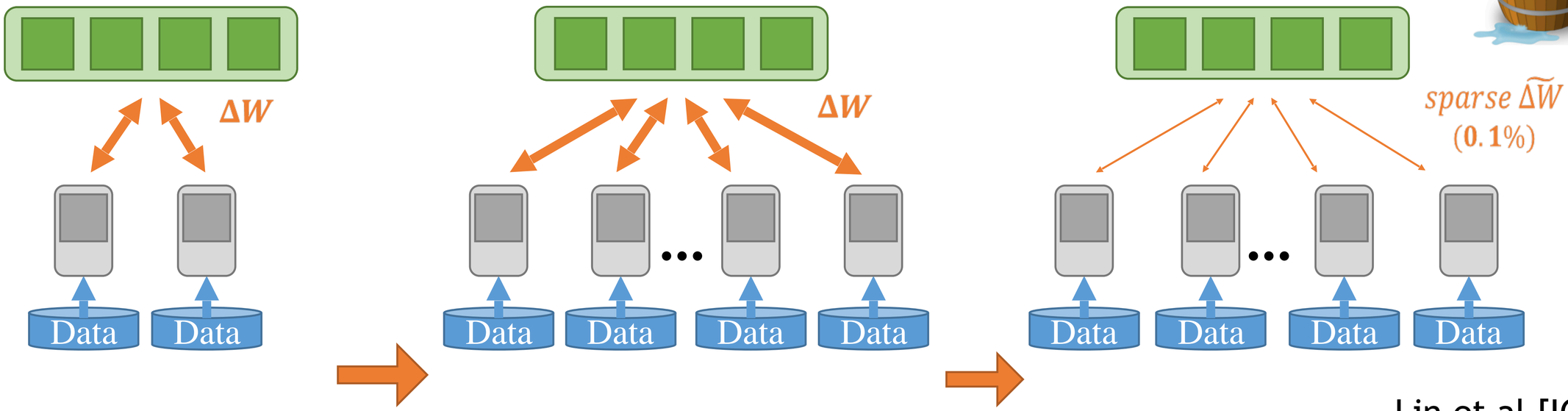
Trained Quantization

Han et al [ICLR'16]

Save Network Bandwidth: Deep Gradient Compression

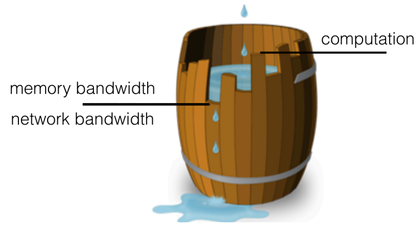
saving the communication bandwidth of distributed training by 500x

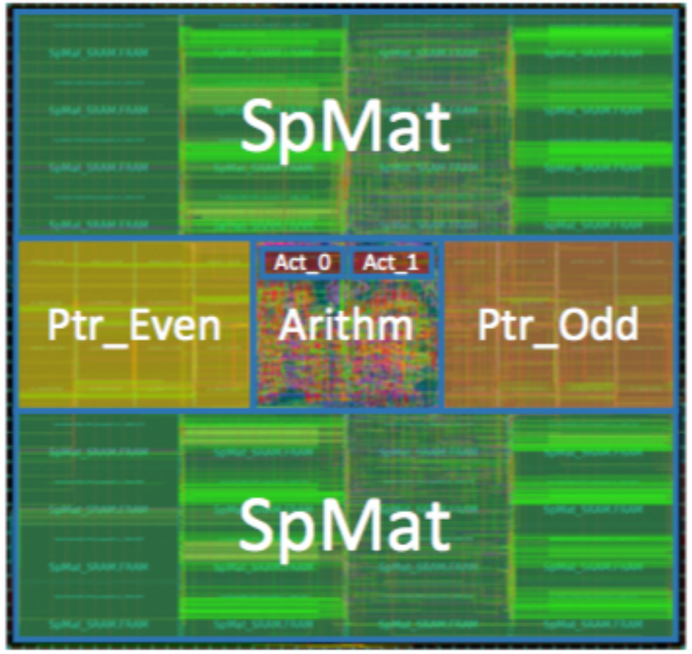




Lin et al [ICLR'18]

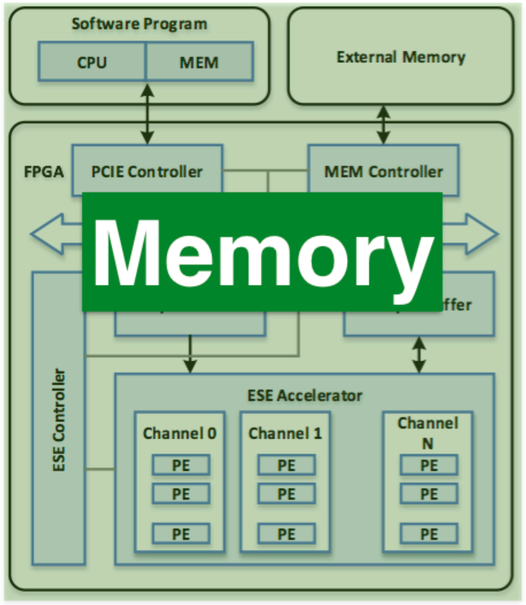
Efficient Inference Engine on Compressed DNN

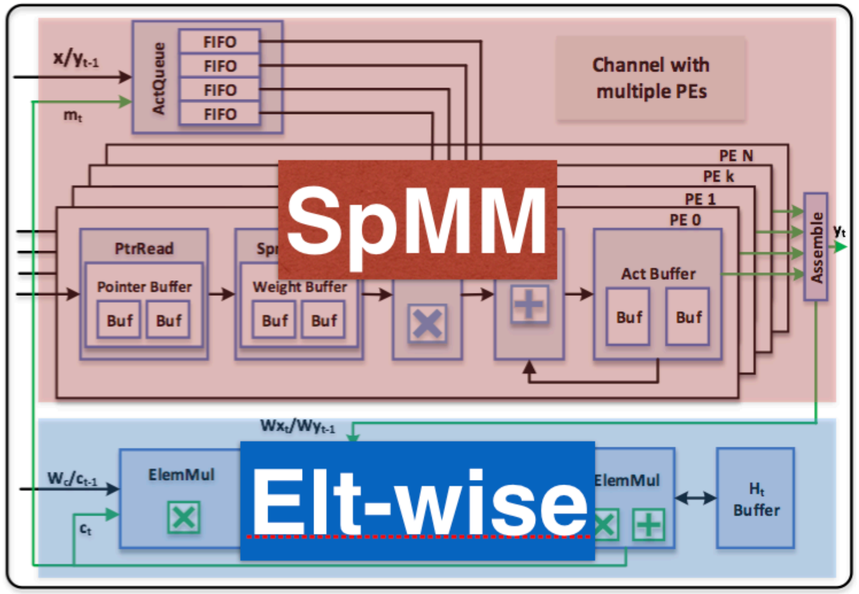




EIE Accelerator

Han et al [ISCA'16]





ESE Accelerator

Han et al [FPGA'17]

Commercialized by DeePhi at AWS

MIT Intelligent Hardware Lab

I’m founding the MIT Intelligent Hardware Lab starting in July 2018.
My research areas include:

- **H**: **H**igh performance, **H**igh energy efficiency **H**ardware
- **A**: **A**rchitectures and **A**ccelerators for **A**rtificial Intelligence
- **N**: **N**ovel algorithms for **N**eural **N**etworks and Deep Learning

Looking forward to industry/academic collaborations and talented PhD students!