Scaling of Machine Learning

Scaled ML

March 24, 2018

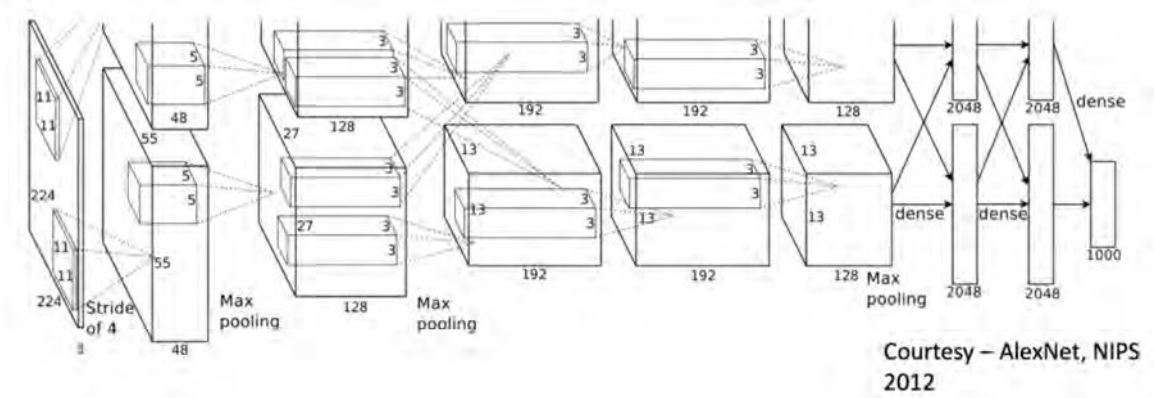
Bill Dally

Chief Scientist and SVP of Research, NVIDIA Corporation

Professor (Research), Stanford University

The AI Revolution has been Enabled by Hardware

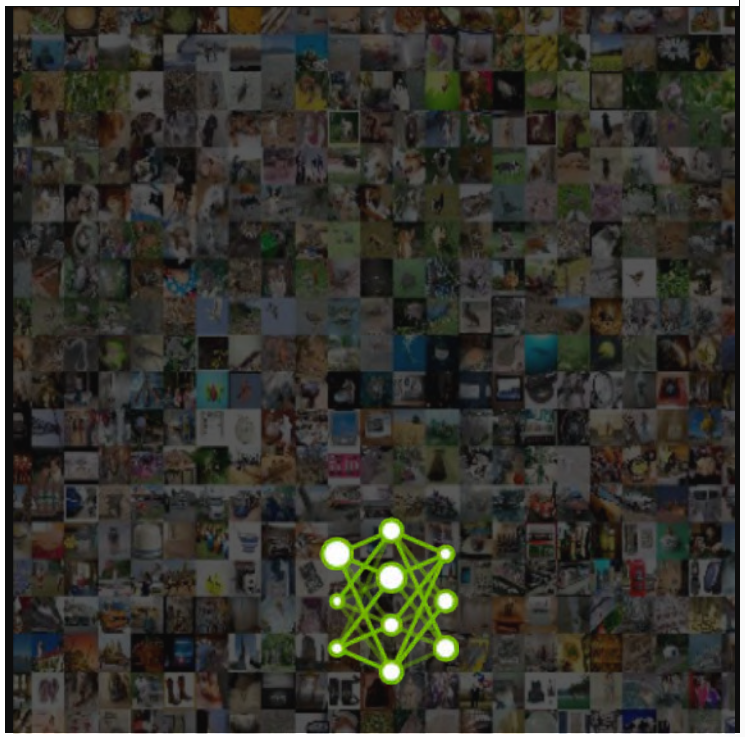
Hardware and Data enable DNNs



Courtesy – AlexNet, NIPS 2012

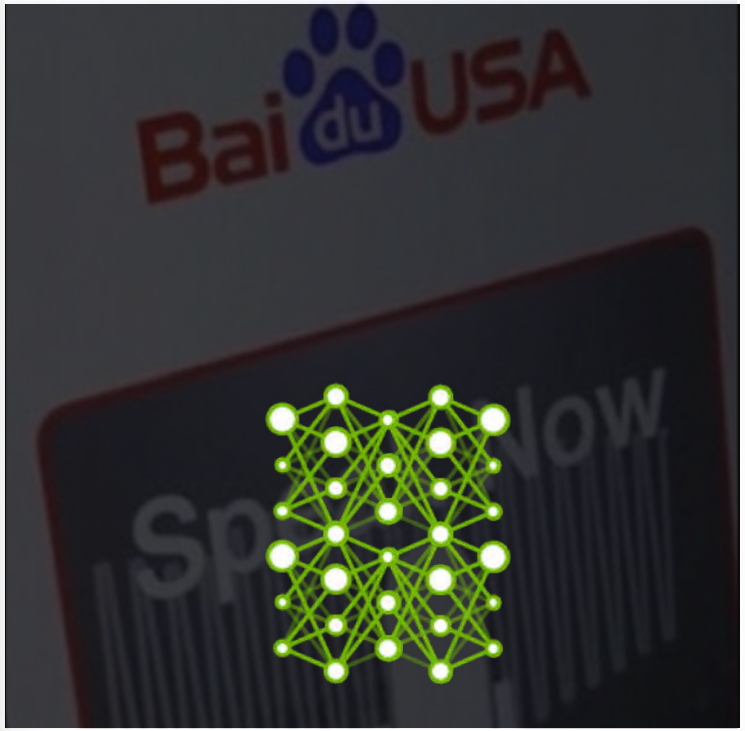
Evolution of DL is Gated by Hardware

7 ExaFLOPS
60 Million Parameters



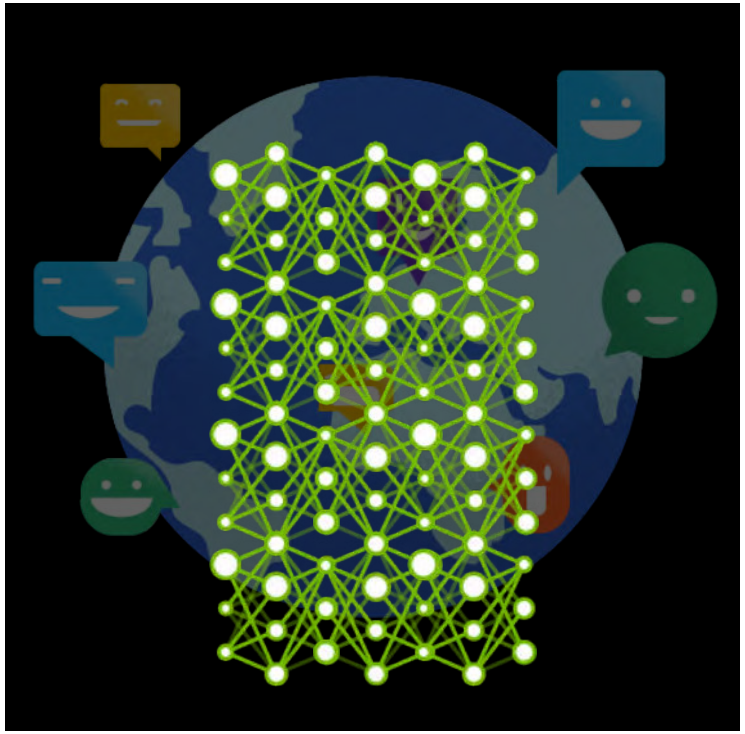
2015 – Microsoft ResNet
Superhuman Image Recognition

20 ExaFLOPS
300 Million Parameters



2016 – Baidu Deep Speech 2
Superhuman Voice Recognition

100 ExaFLOPS
8.7 Billion Parameters



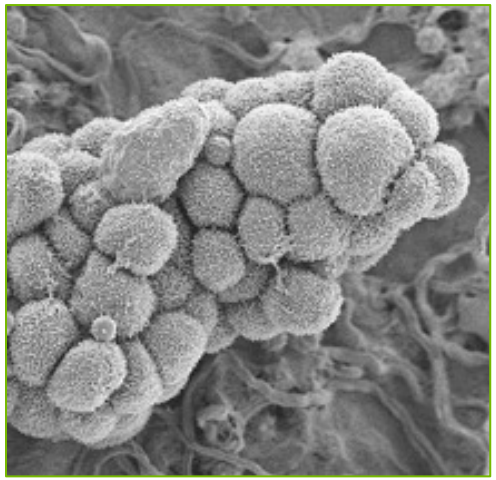
2017 – Google Neural Machine Translation
Near Human Language Translation

Many Dimensions of Scaling

Scaling Applications

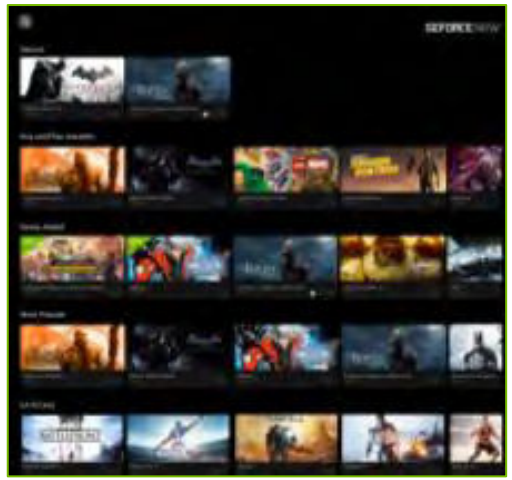
INTERNET & CLOUD

Image Classification
Speech Recognition
Language Translation
Language Processing
Sentiment Analysis
Recommendation



MEDICINE & BIOLOGY

Cancer Cell Detection
Diabetic Grading
Drug Discovery



MEDIA & ENTERTAINMENT

Video Captioning
Video Search
Real Time Translation

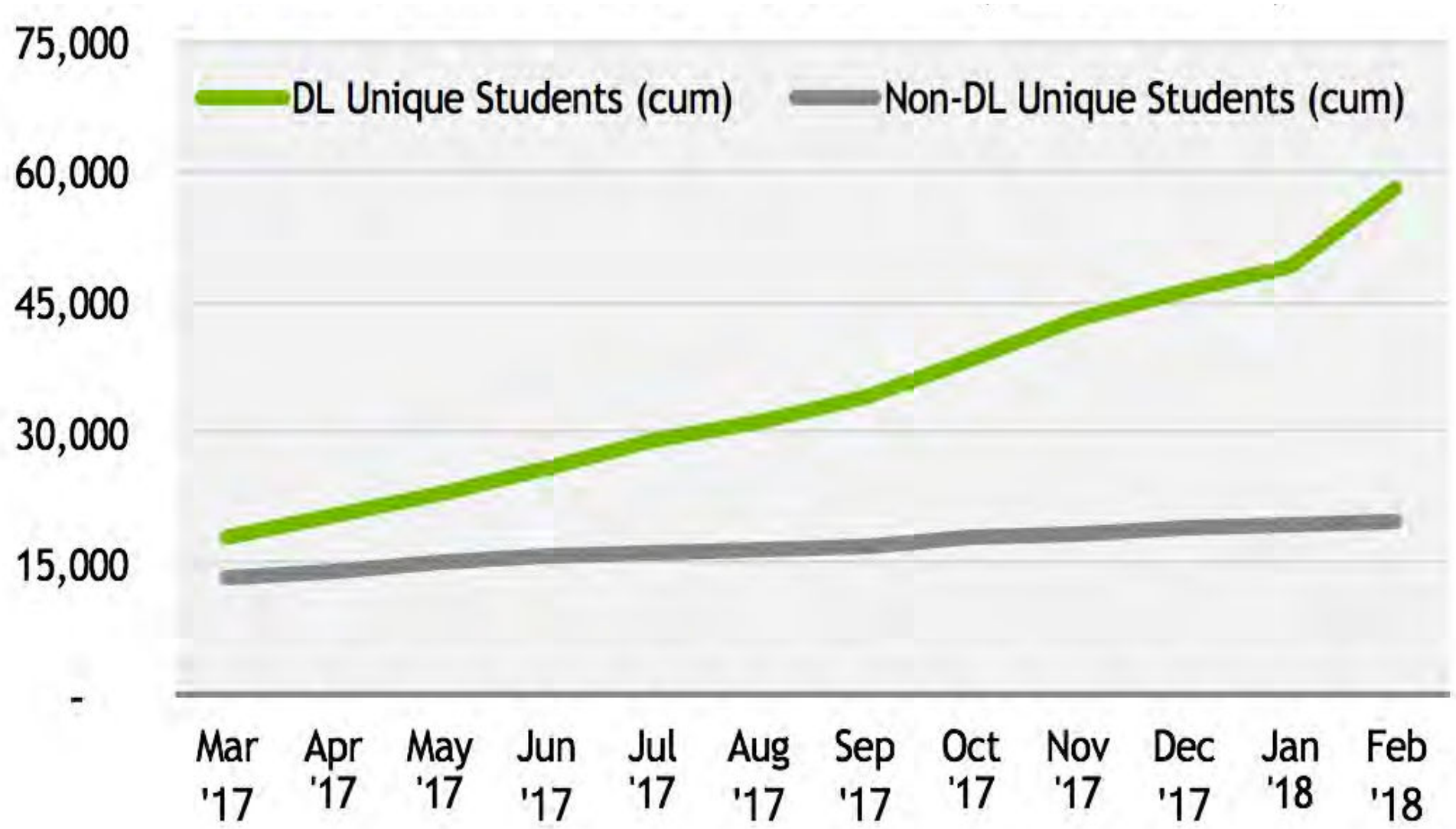
SECURITY & DEFENSE

Face Detection
Video Surveillance
Satellite Imagery

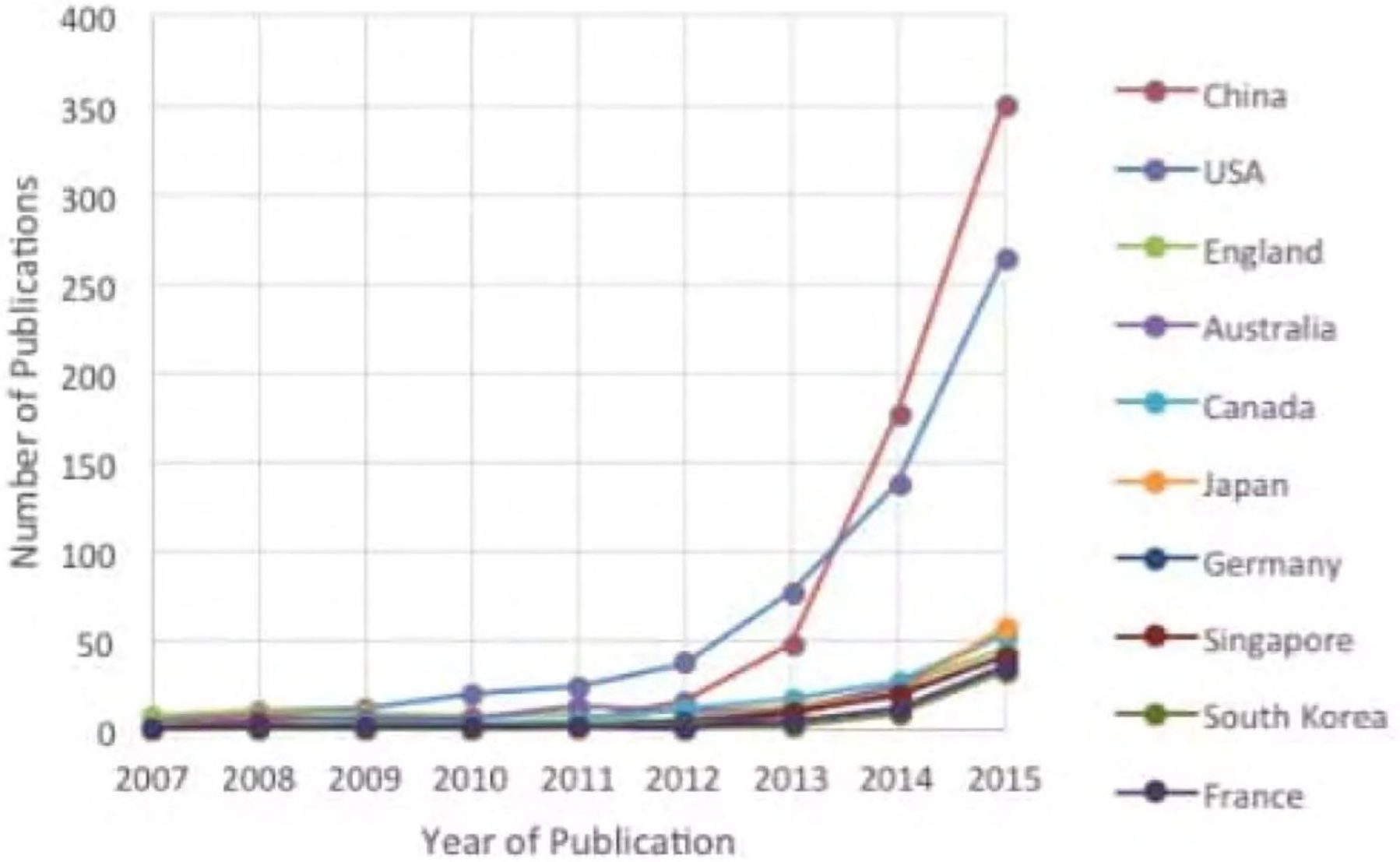
AUTONOMOUS MACHINES

Pedestrian Detection
Lane Tracking
Recognize Traffic Sign

Scaling DL Practitioners



Number of Deep Learning Papers



Washington Post 10/13/2016

Scaling of Models and Data

Important Property of Neural Networks

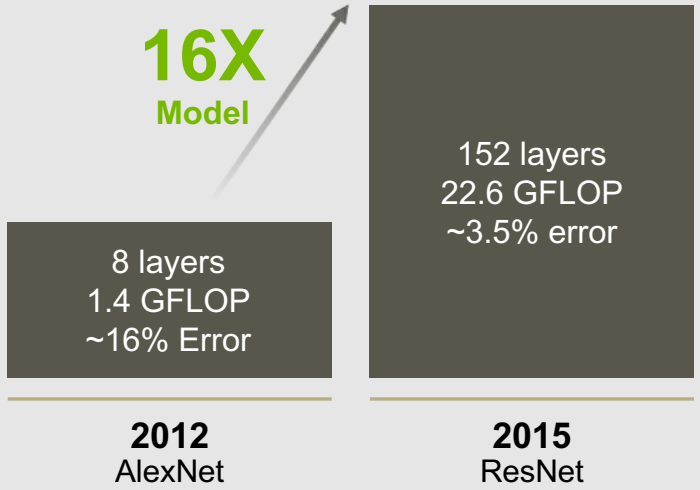
Results get better with

more data +
bigger models +
more computation

(Better algorithms, new insights and improved techniques always help, too!)

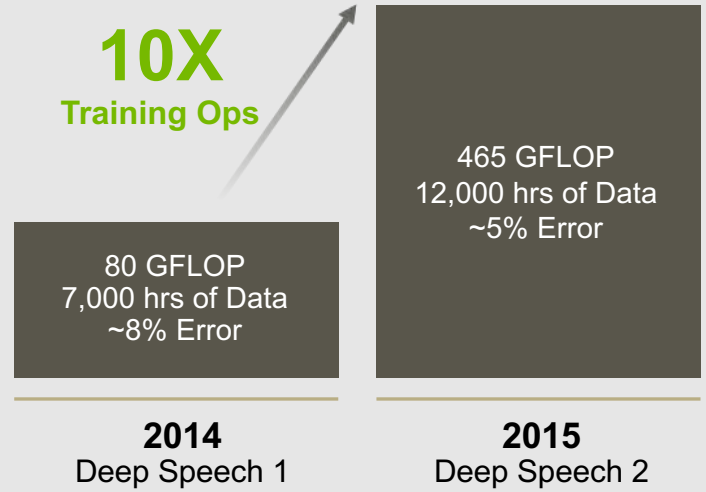
Google

IMAGE RECOGNITION



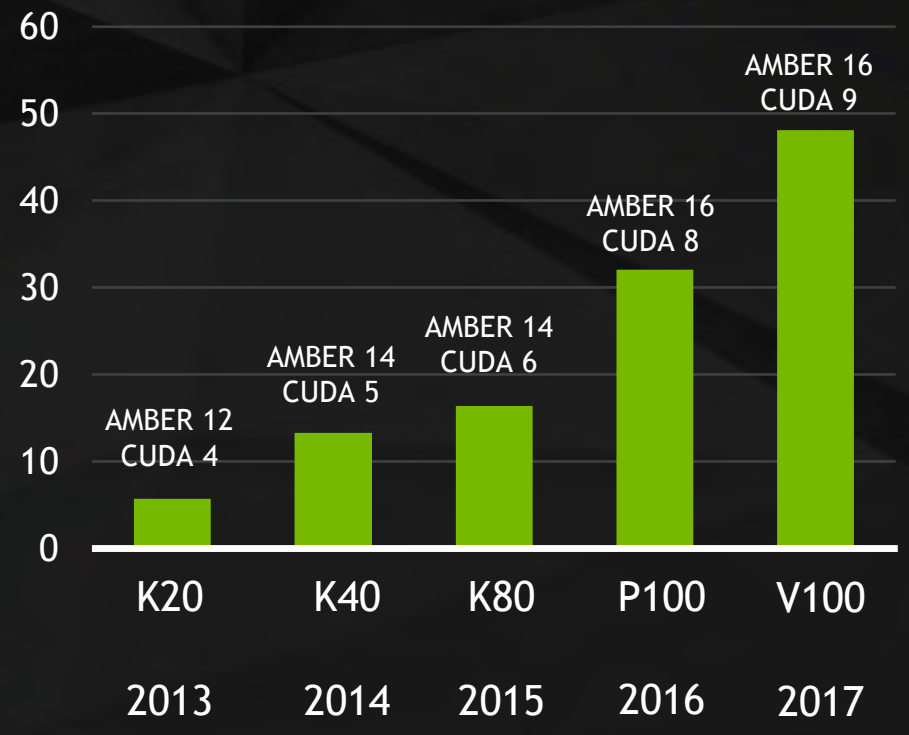
Microsoft

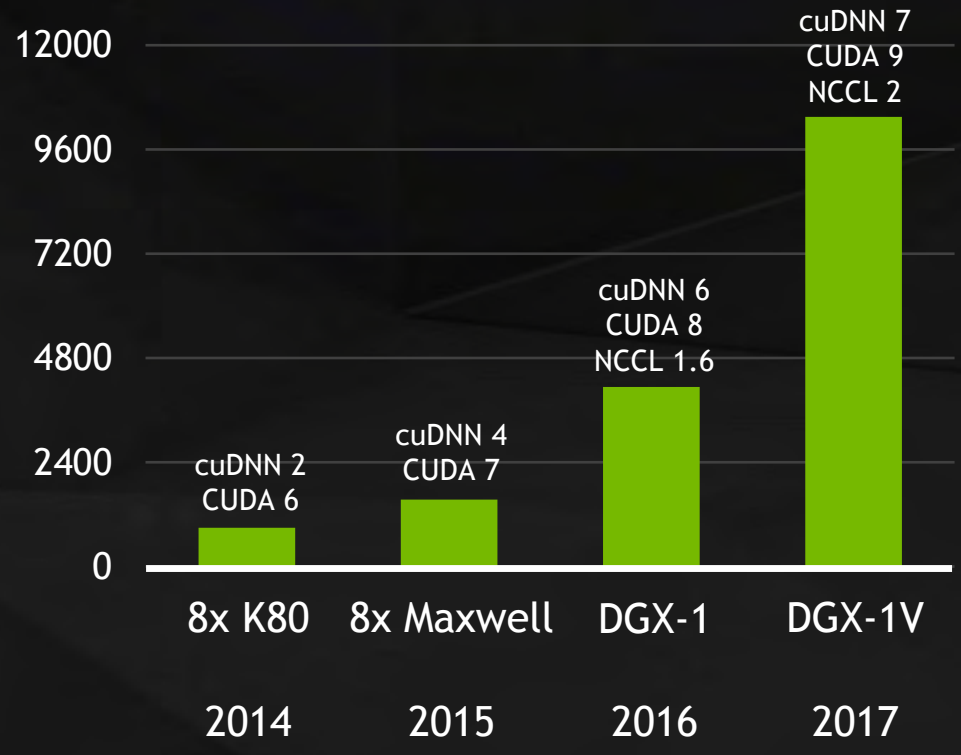
SPEECH RECOGNITION



Baidu 百度

Scaling of GPU Performance





V100, DGX-1V Performance measured on Pre-production Hardware

Scaling the Number of GPUs

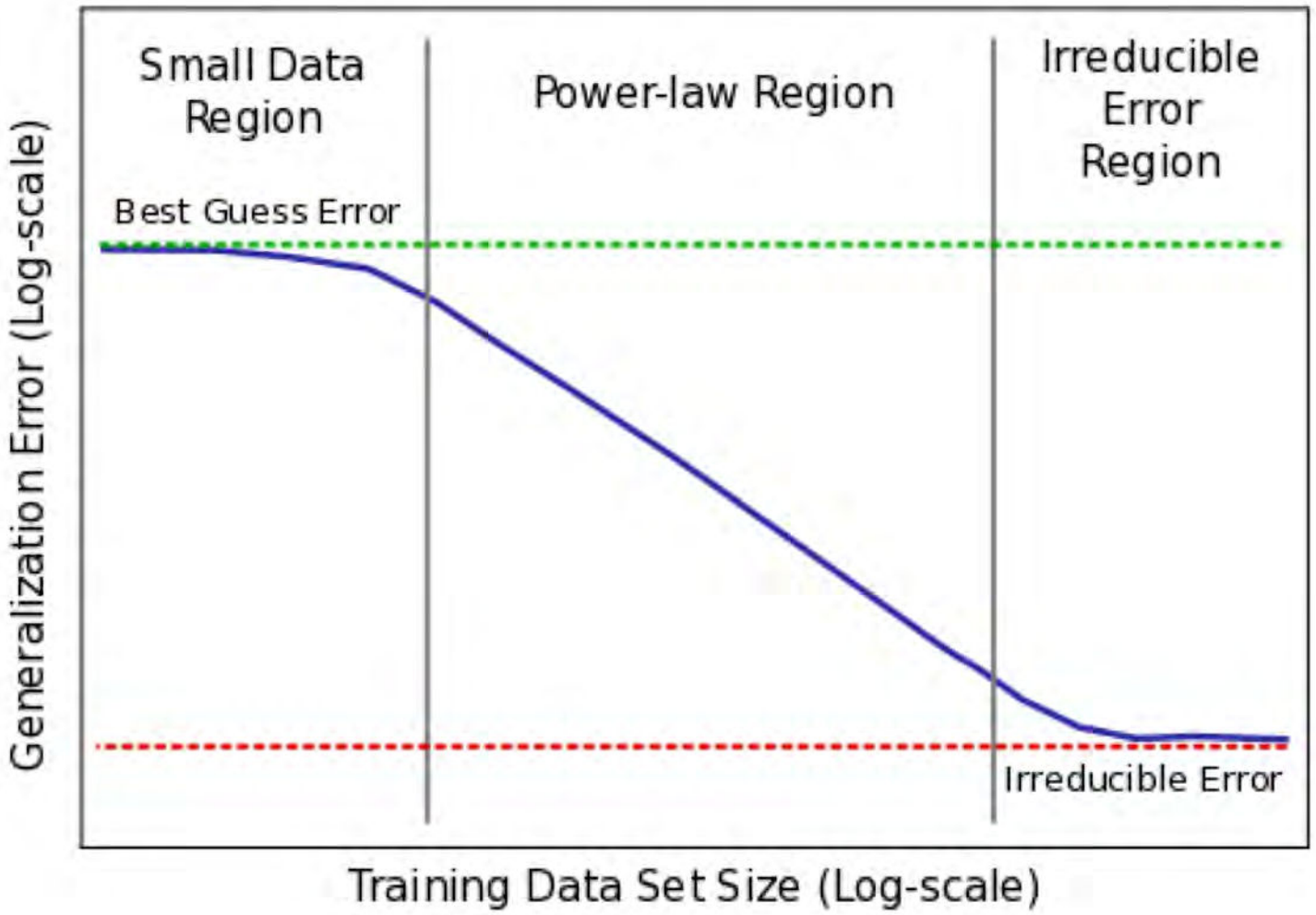
Table 1: 90-epoch training time and single-crop validation accuracy of ResNet-50 for ImageNet reported by different teams.

Team	Hardware	Software	Minibatch size	Time	Accuracy
He *et al.* [5]	Tesla P100 × 8	Caffe	256	29 hr	75.3 %
Goyal *et al.* [4]	Tesla P100 × 256	Caffe2	8,192	1 hr	76.3 %
Codreanu *et al.* [3]	KNL 7250 × 720	Intel Caffe	11,520	62 min	75.0 %
You *et al.* [10]	Xeon 8160 × 1600	Intel Caffe	16,000	31 min	75.3 %
This work	Tesla P100 × 1024	Chainer	32,768	15 min	74.9 %

Akiba et al. “Extremely Large Minibatch SGD: Training ResNet-50 on ImageNet in 15 Minutes,” arXiv.

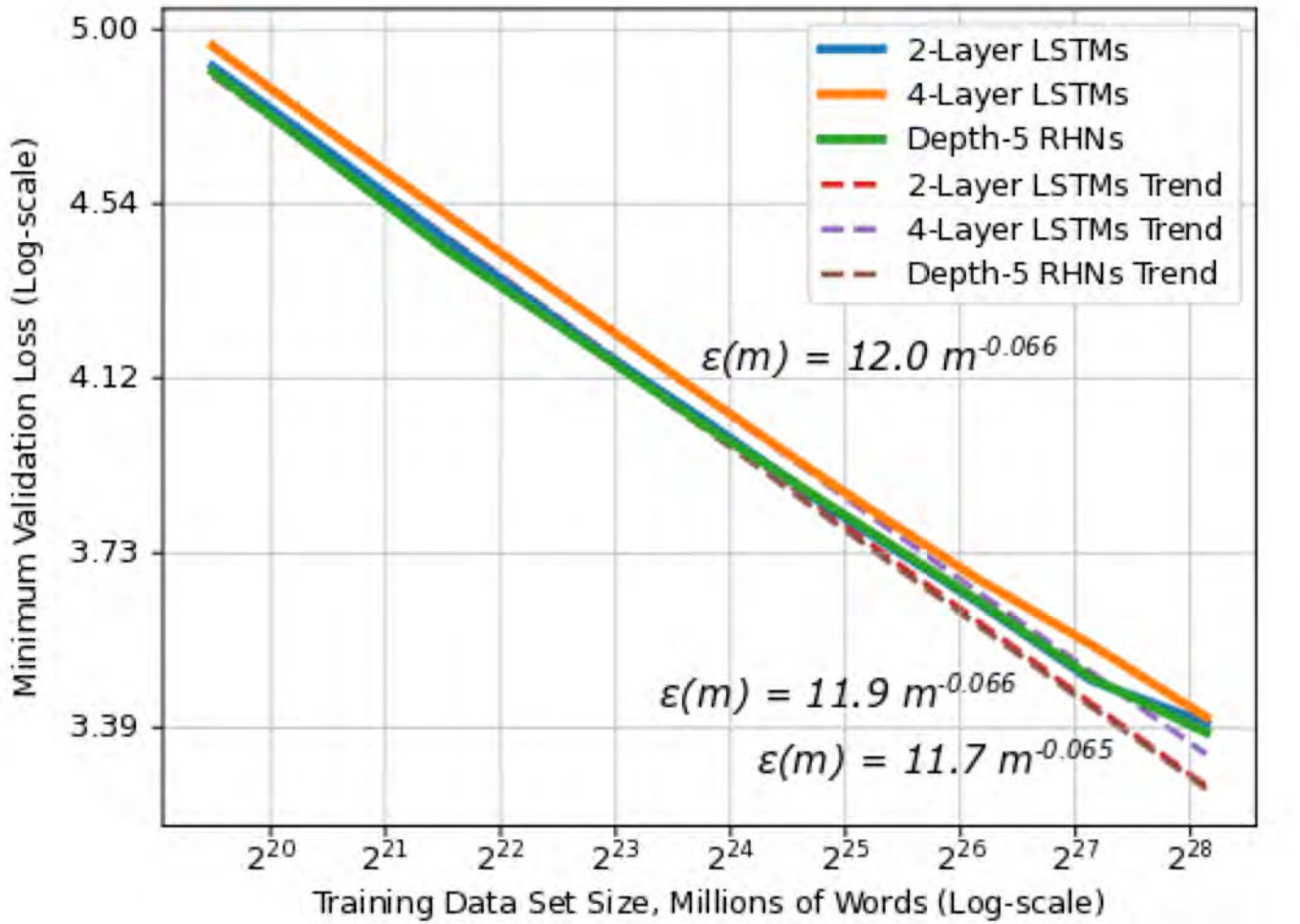
Power-Law Scaling of Deep Learning

Power-Law Scaling of DNNs



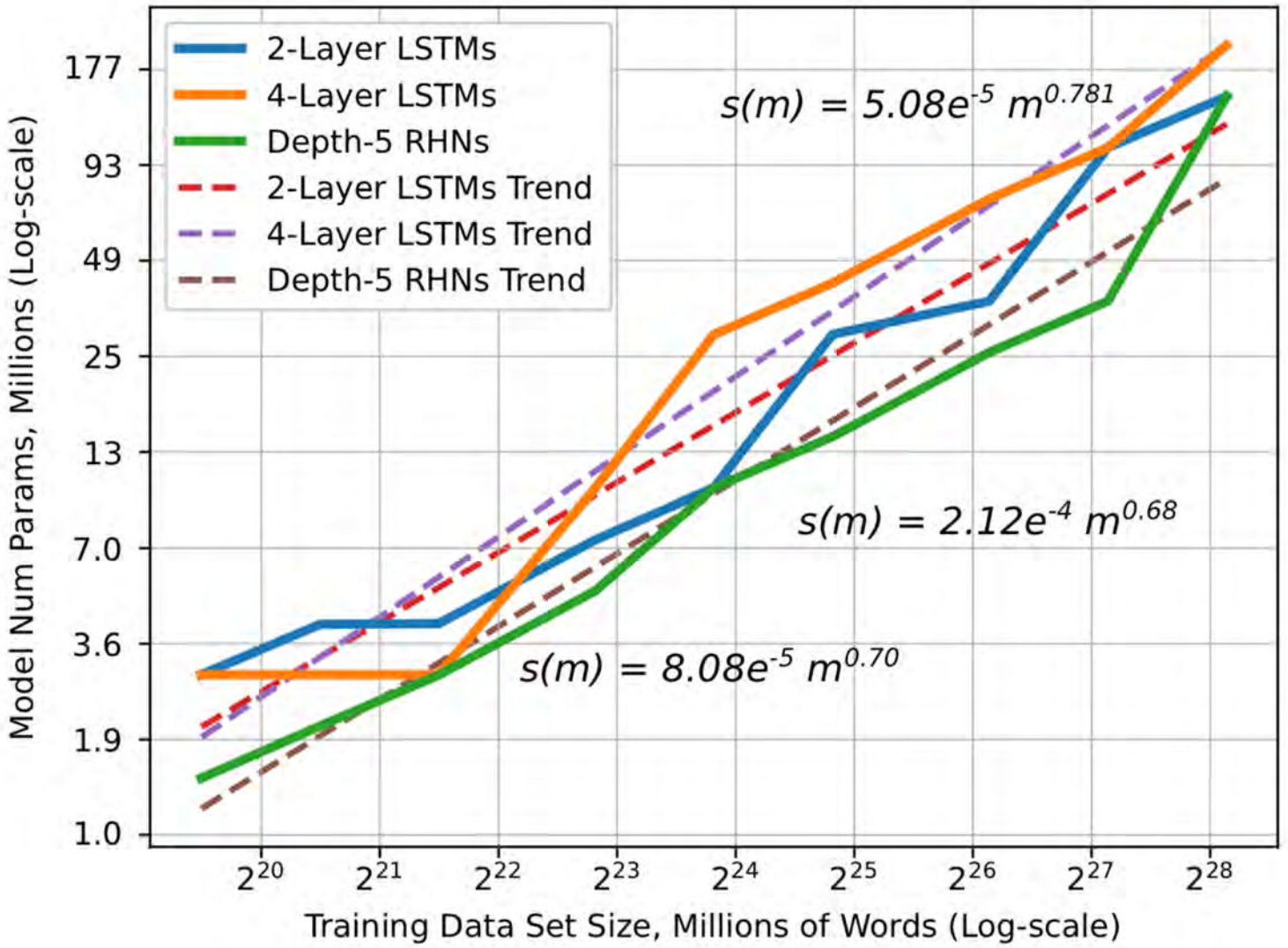
Hestness et al., “Deep Learning Scaling is Predictable, Empirically,” arXiv:1712.00409

More Data, More Accuracy



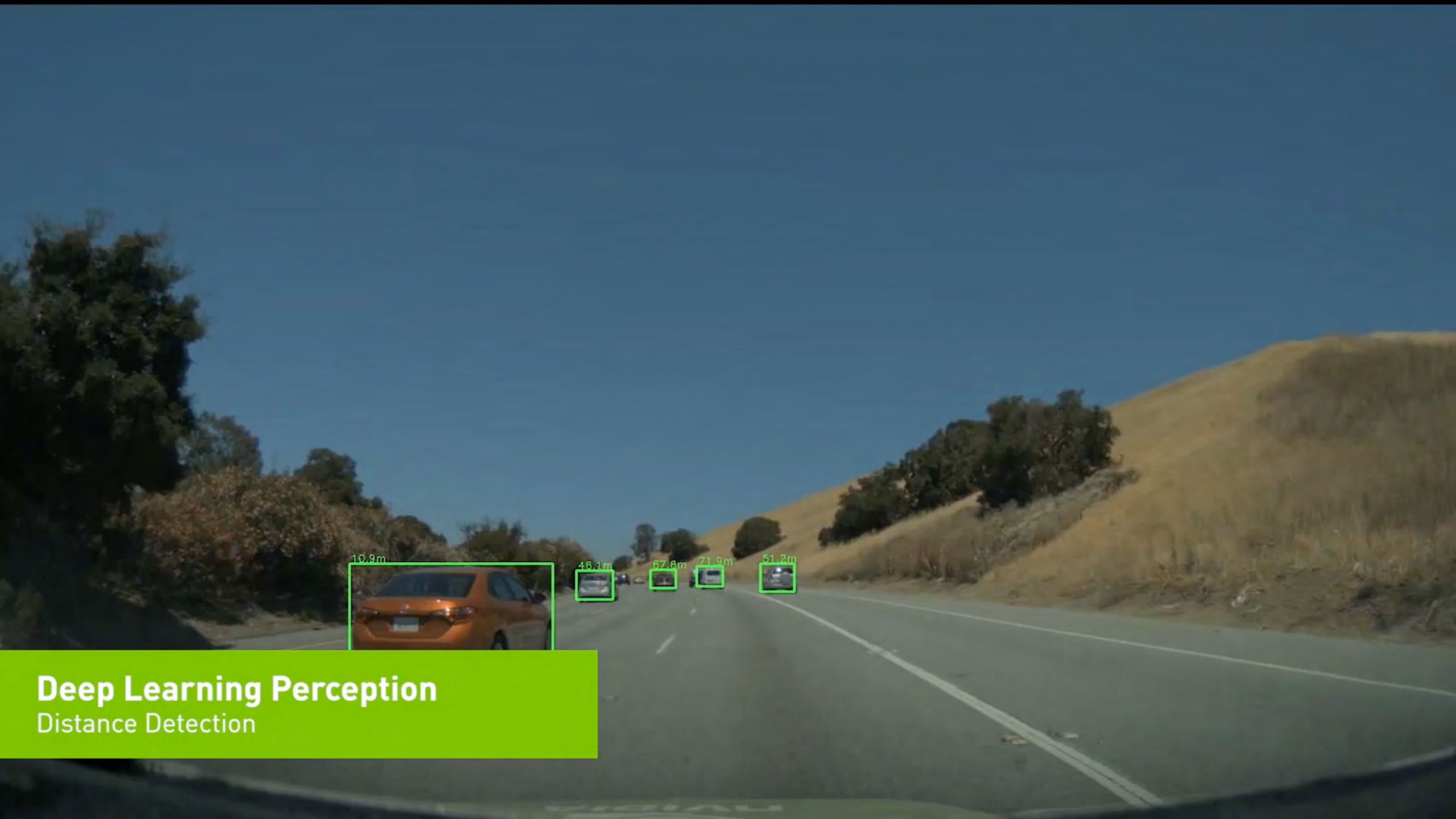
Hestness et al., “Deep Learning Scaling is Predictable, Empirically,” arXiv:1712.00409

More Data, Larger Model



Hestness et al., “Deep Learning Scaling is Predictable, Empirically,” arXiv:1712.00409

Inference Performance Required Today



Deep Learning Perception
Distance Detection

ResNet-50 requires 7.72 Billion operations to process one 225x225 image

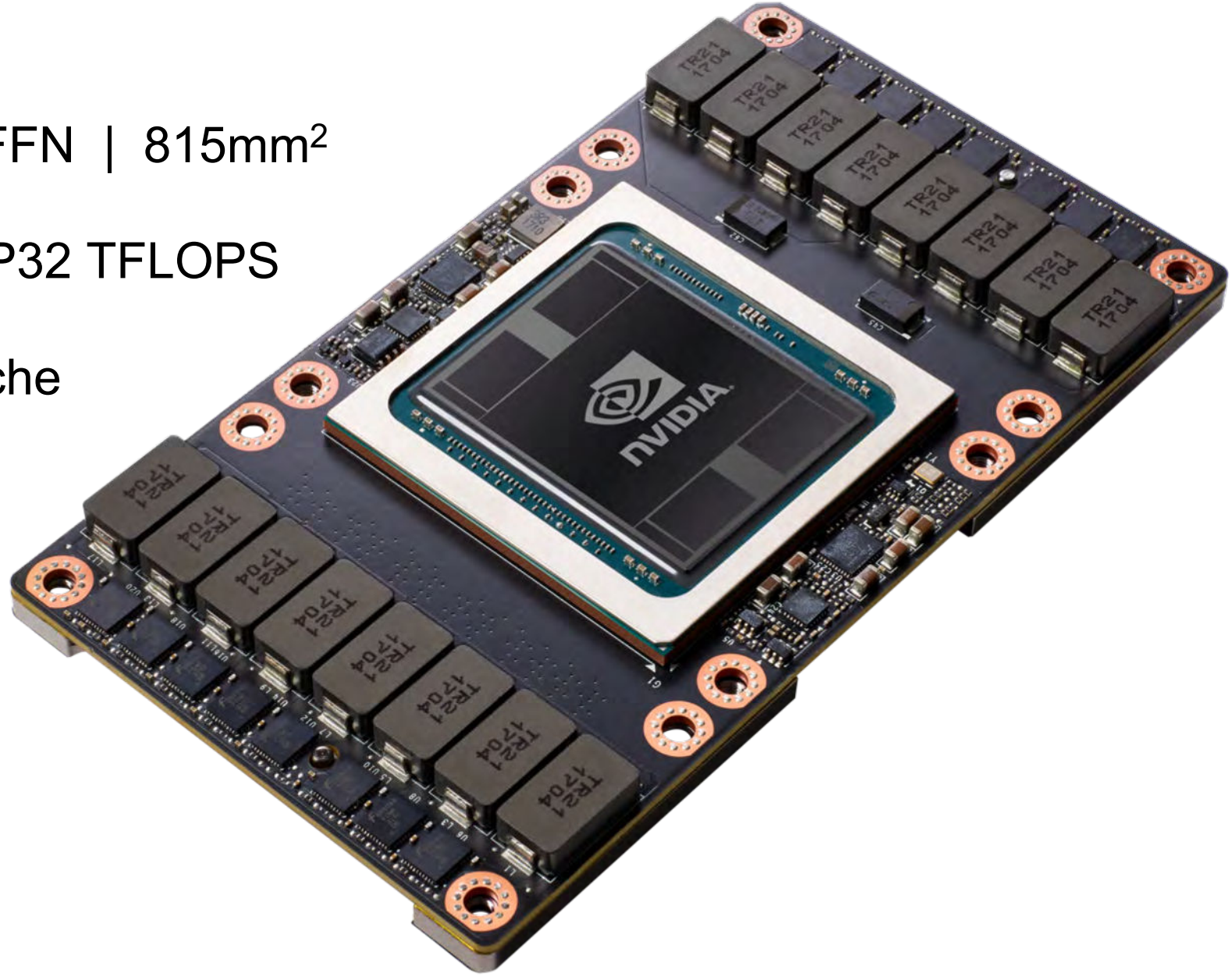
230Gops for 30fps
9.4Tops for HD

12 cameras, 3 nets each -> 338Tops

GPUs

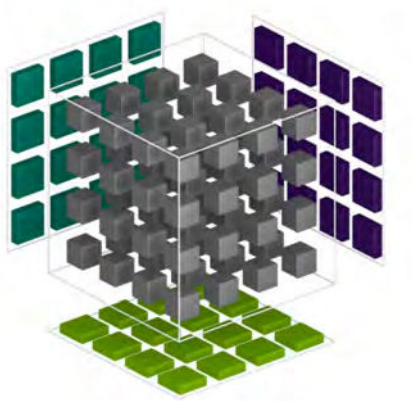
Volta V100

21B xtors | TSMC 12nm FFN | 815mm^2
5,120 CUDA cores
7.5 FP64 TFLOPS | 15 FP32 TFLOPS
120 Tensor TFLOPS
20MB SM RF | 16MB Cache
16GB HBM2 @ 900 GB/s
300 GB/s NVLink



Tensor Core

Mixed Precision Matrix Math
4x4 matrices



$$D = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

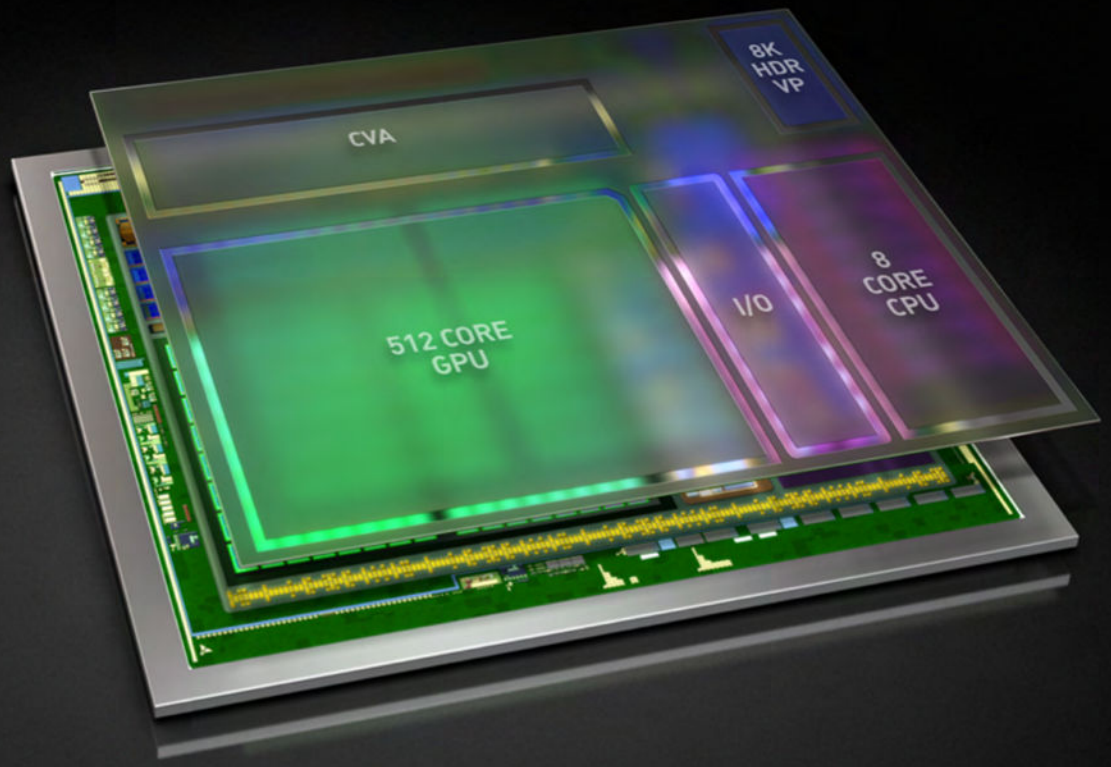
FP16 FP16 FP16 or FP32

$$D = AB + C$$

Specialized Instructions Amortize Overhead

Operation	Energy**	Overhead*
HFMA	1.5pJ	2000%
HDP4A	6.0pJ	500%
HMMA	110pJ	27%

*Overhead is instruction fetch, decode, and operand fetch – 30pJ
**Energy numbers from 45nm process



XAVIER

AI SUPERCOMPUTER SOC

7 Billion Transistors 16nm FF

8 Core Custom ARM64 CPU

512 Core Volta GPU

Deep Learning Accelerator

Dual 8K HDR Video Processors

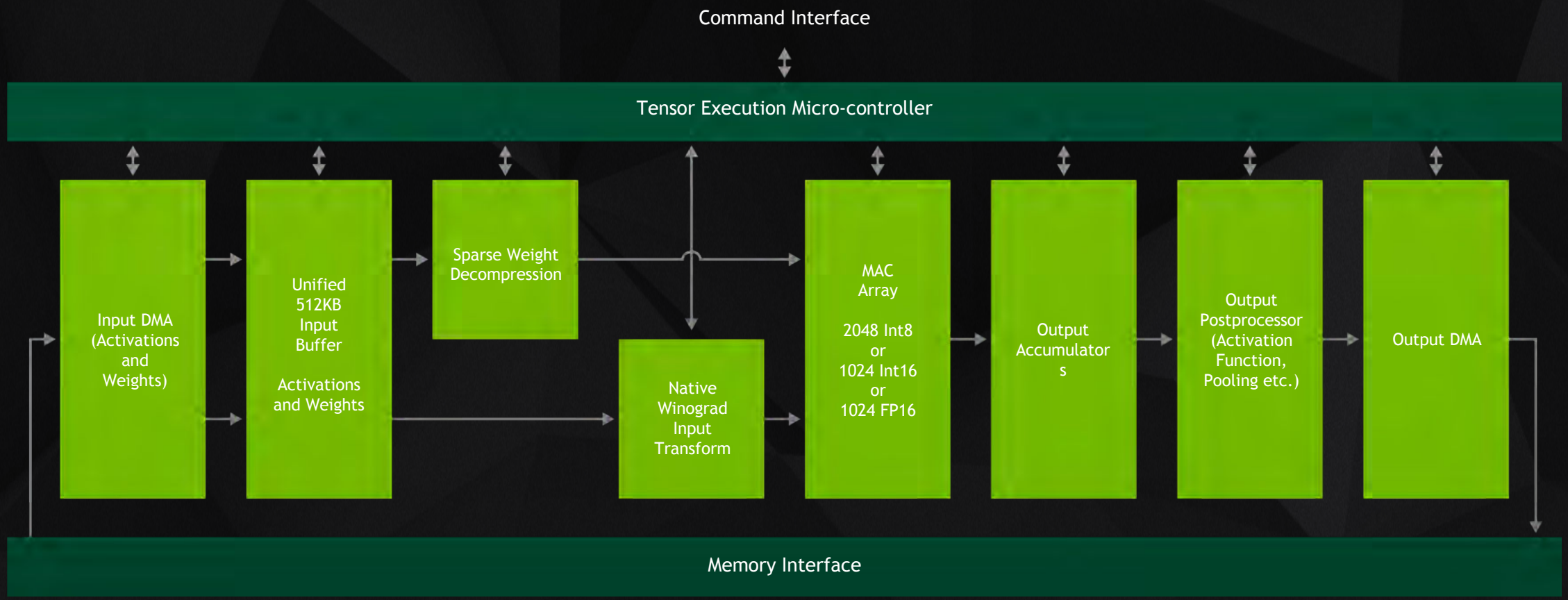
Designed for ASIL C Functional Safety

30 TOPS DL

160 SPECINT

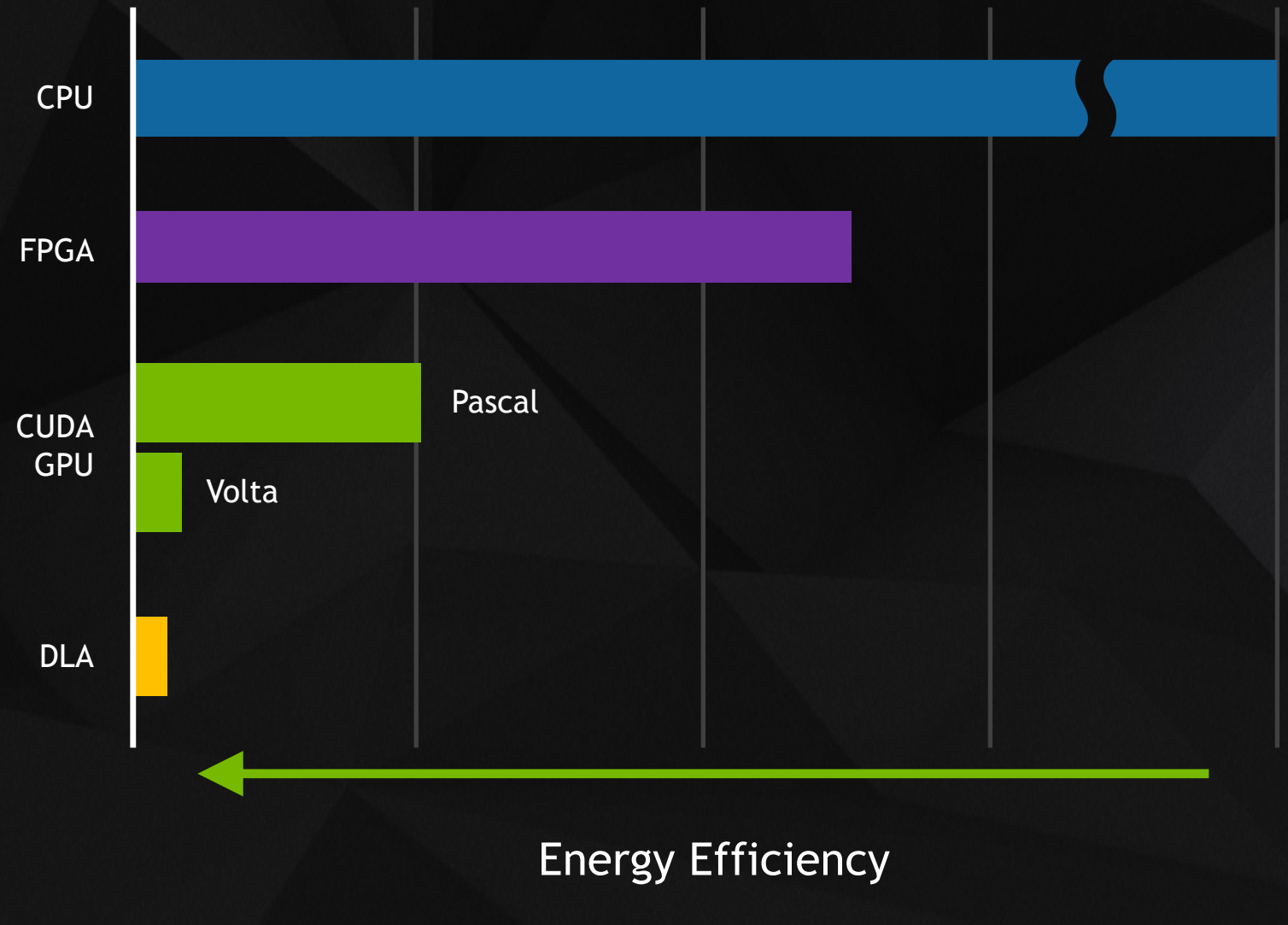
30W

NVIDIA DLA



Open-sourced at nvdla.org

Comparison of Energy Efficiency



How Do We Continue to Scale Deep Learning Performance, Data Size, Model Size Now that Moore's Law is Over

Train with More GPUs

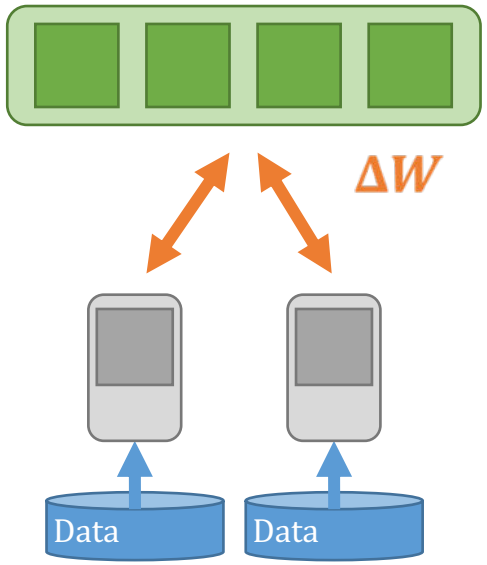
SCALING THE NUMBER OF GPUS

Table 1: 90-epoch training time and single-crop validation accuracy of ResNet-50 for ImageNet reported by different teams.

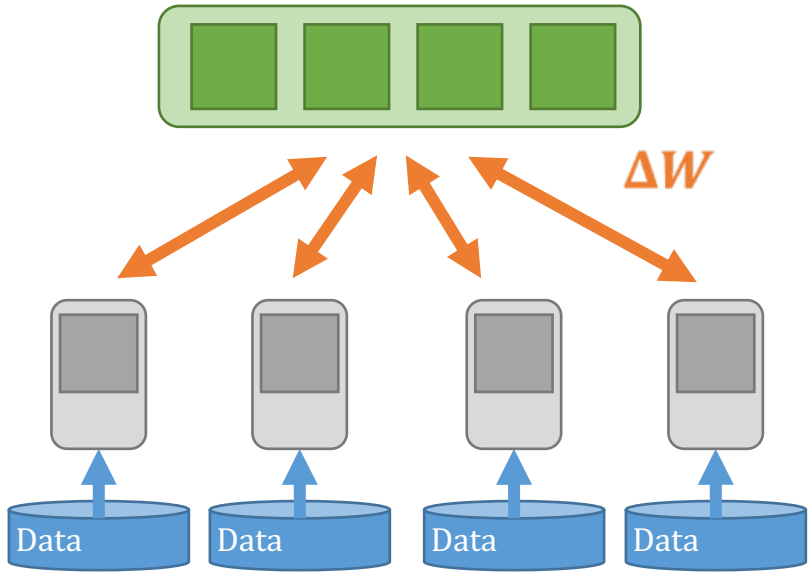
Team	Hardware	Software	Minibatch size	Time	Accuracy
He *et al.* [5]	Tesla P100 × 8	Caffe	256	29 hr	75.3 %
Goyal *et al.* [4]	Tesla P100 × 256	Caffe2	8,192	1 hr	76.3 %
Codreanu *et al.* [3]	KNL 7250 × 720	Intel Caffe	11,520	62 min	75.0 %
You *et al.* [10]	Xeon 8160 × 1600	Intel Caffe	16,000	31 min	75.3 %
This work	Tesla P100 × 1024	Chainer	32,768	15 min	74.9 %

Akiba et al. “Extremely Large Minibatch SGD: Training ResNet-50 on ImageNet in 15 Minutes,” arXiv.

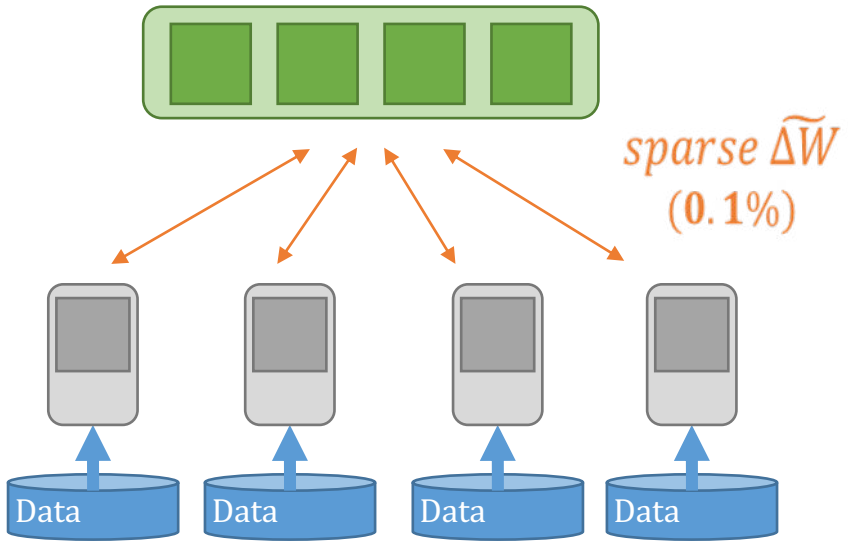
Distributed training



Increased Parallelism

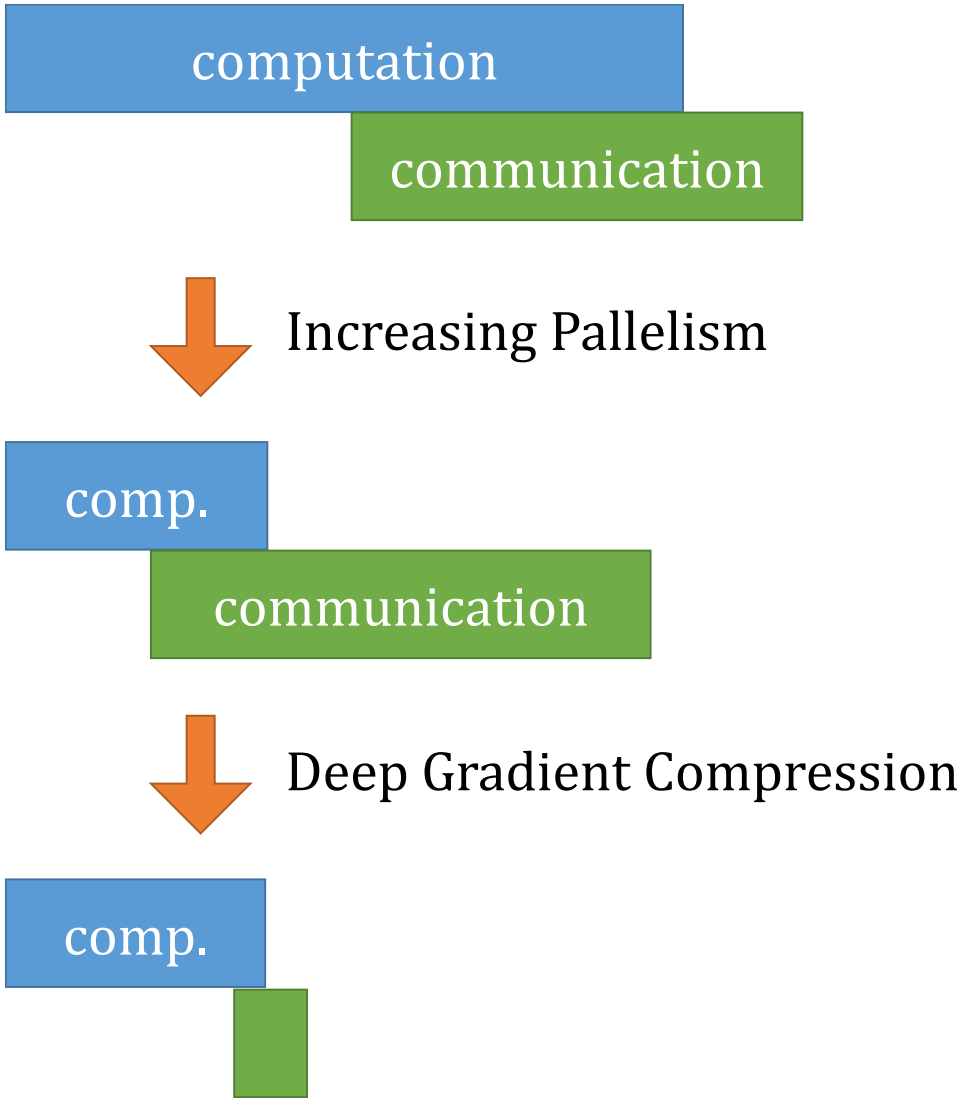


Deep Gradient Compression



Lin, Han, Mao, Wang, Dally, “Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training”, ICLR 18

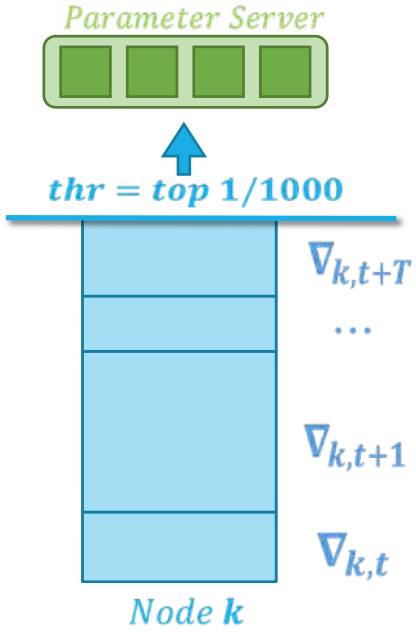
Deep Gradient Compression



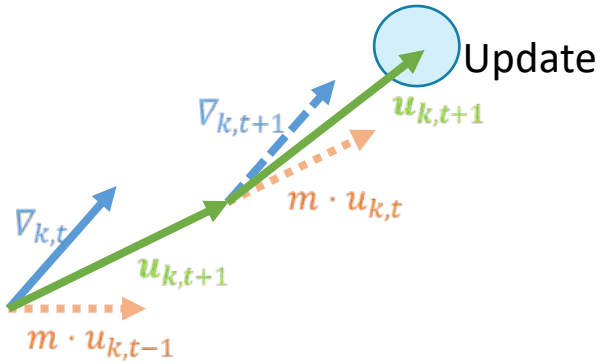
Lin, Han, Mao, Wang, Dally, “Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training”, ICLR 18

Deep Gradient Compression

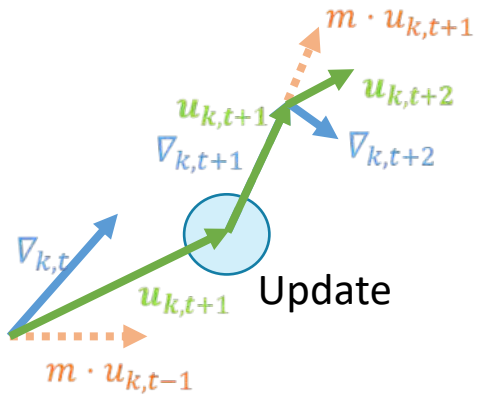
Local Gradient Accumulation



Momentum Correction



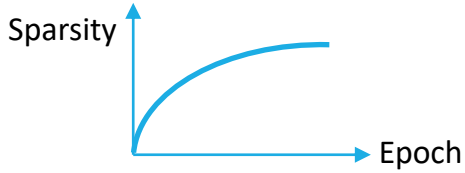
Momentum Factor Masking



Local Gradient Clipping

- $thr_{|G_{k,t}|_2} = N^{-1/2} \cdot thr_{|G_t|_2}$

Warm-up Training



Lin, Han, Mao, Wang, Dally, “Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training”, ICLR 18

200-600x Reduction in Communication
No Loss of Accuracy

Task		Baseline	Ours
ResNet-50 On ImageNet	Top-1 Accuracy	75.96%	**76.15% (+0.19%)**
	Top-5 Accuracy	92.91%	**92.97% (+0.06%)**
	Compression Ratio	1 ×	**277 ×**
5-Layer GRU On LibriSpeech Corpus	Word Error Rate (WER) On test-clean	9.45%	**9.06% (-0.39%)**
	Word Error Rate (WER) On test-other	27.07%	**27.04% (-0.03%)**
	Compression Ratio	1 ×	**608 ×**
2-Layer LSTM Language Model On Penn Treebank	Perplexity	72.30	**72.24 (-0.06)**
	Compression Ratio	1 ×	**462 ×**

Lin, Han, Mao, Wang, Dally, “Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training”, ICLR 18

How Do We Continue to Scale Deep Learning Performance, Data Size, Model Size Now that Moore's Law is Over

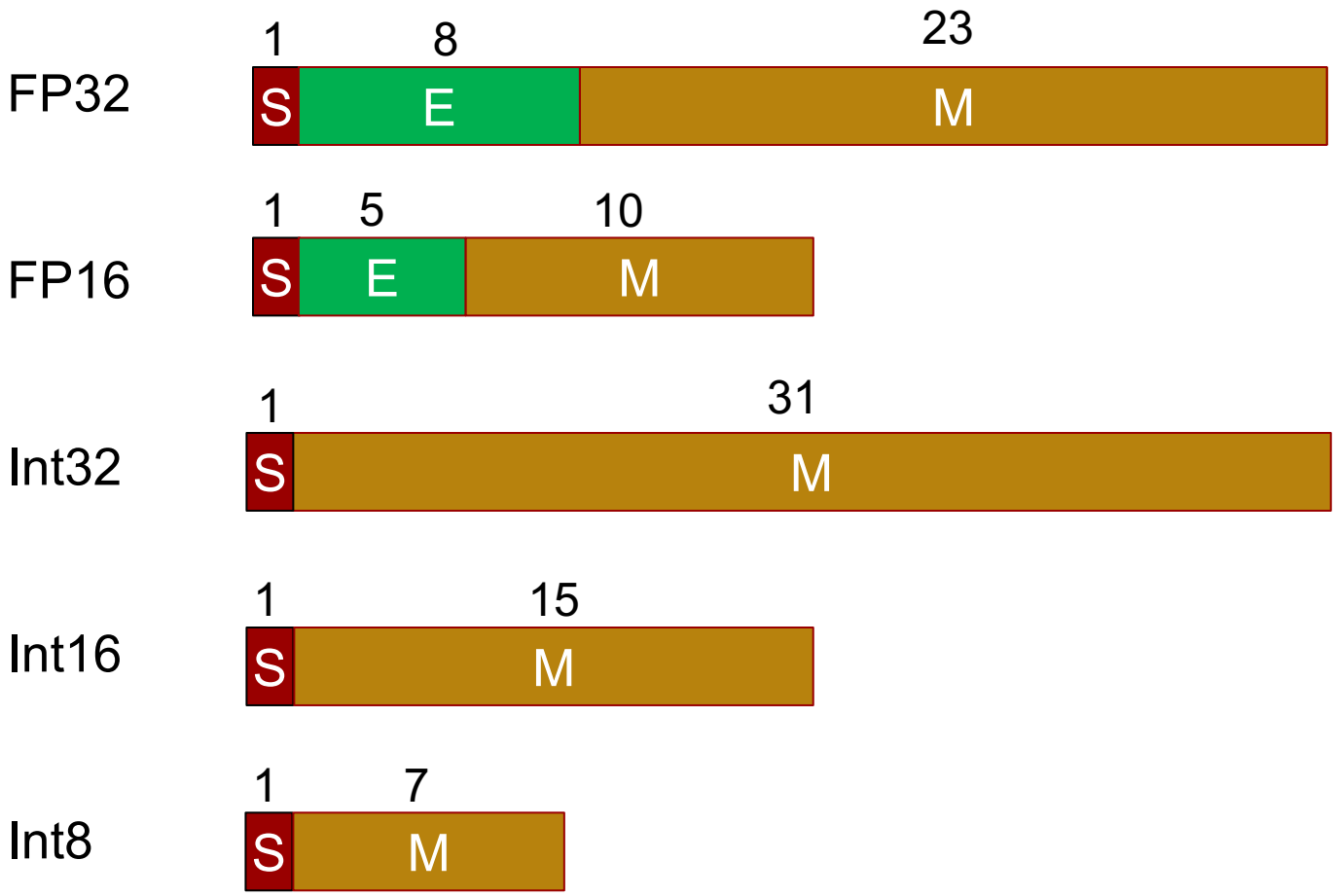
Small is Efficient

Fewer Bits per Weight

Fewer Weights

Reduced Precision

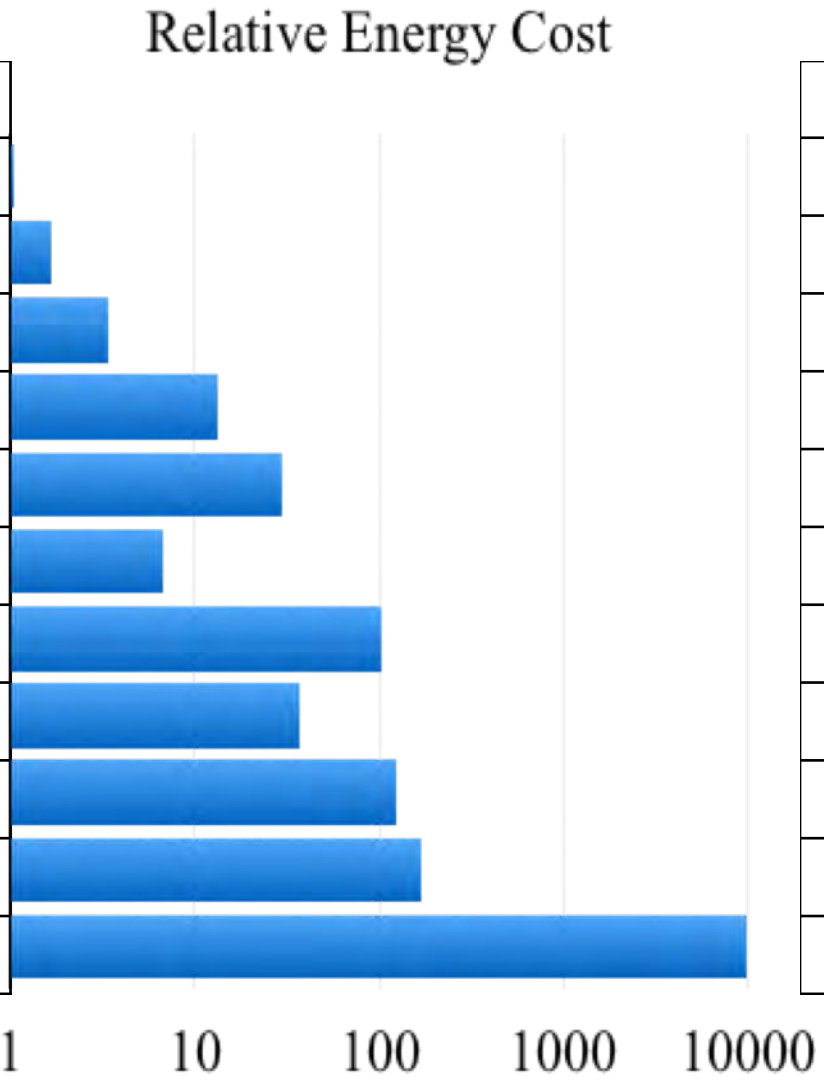
Number Representation

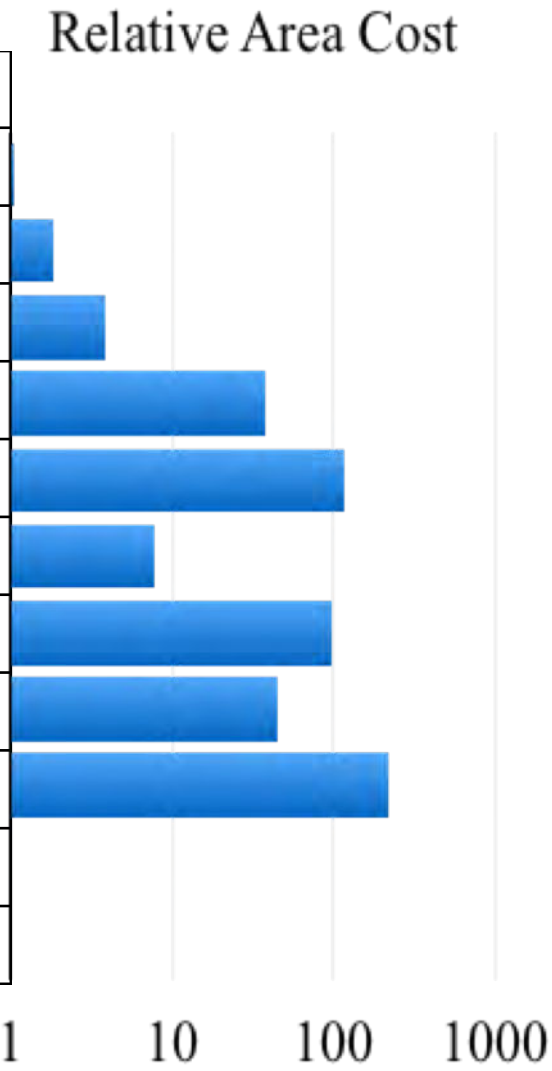


	Range	Accuracy
FP32	10^{-38} - 10^{38}	.000006%
FP16	$6x10^{-5}$ - $6x10^{4}$	.05%
Int32	0 – $2x10^{9}$	½
Int16	0 – $6x10^{4}$	½
Int8	0 – 127	½

Cost of Operations

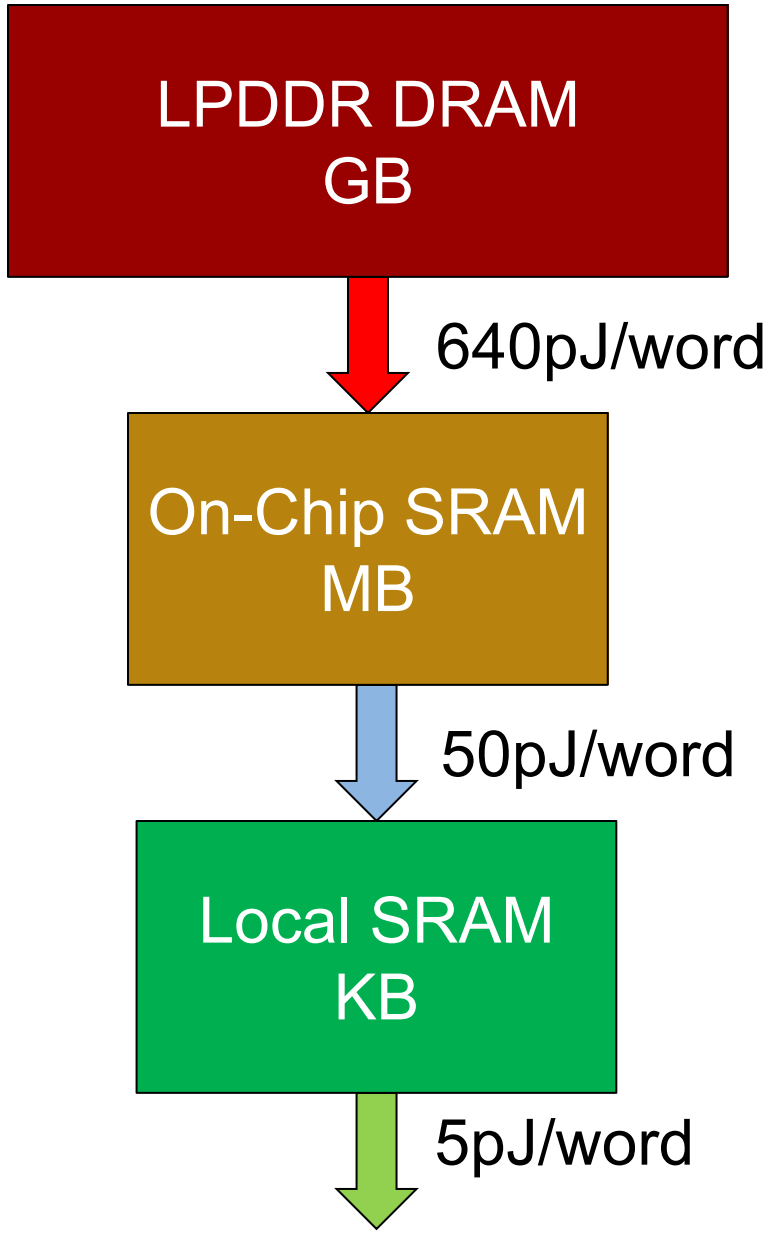
Operation:	Energy (pJ)	Area (μm²)
8b Add	0.03	36
16b Add	0.05	67
32b Add	0.1	137
16b FP Add	0.4	1360
32b FP Add	0.9	4184
8b Mult	0.2	282
32b Mult	3.1	3495
16b FP Mult	1.1	1640
32b FP Mult	3.7	7700
32b SRAM Read (8KB)	5	N/A
32b DRAM Read	640	N/A



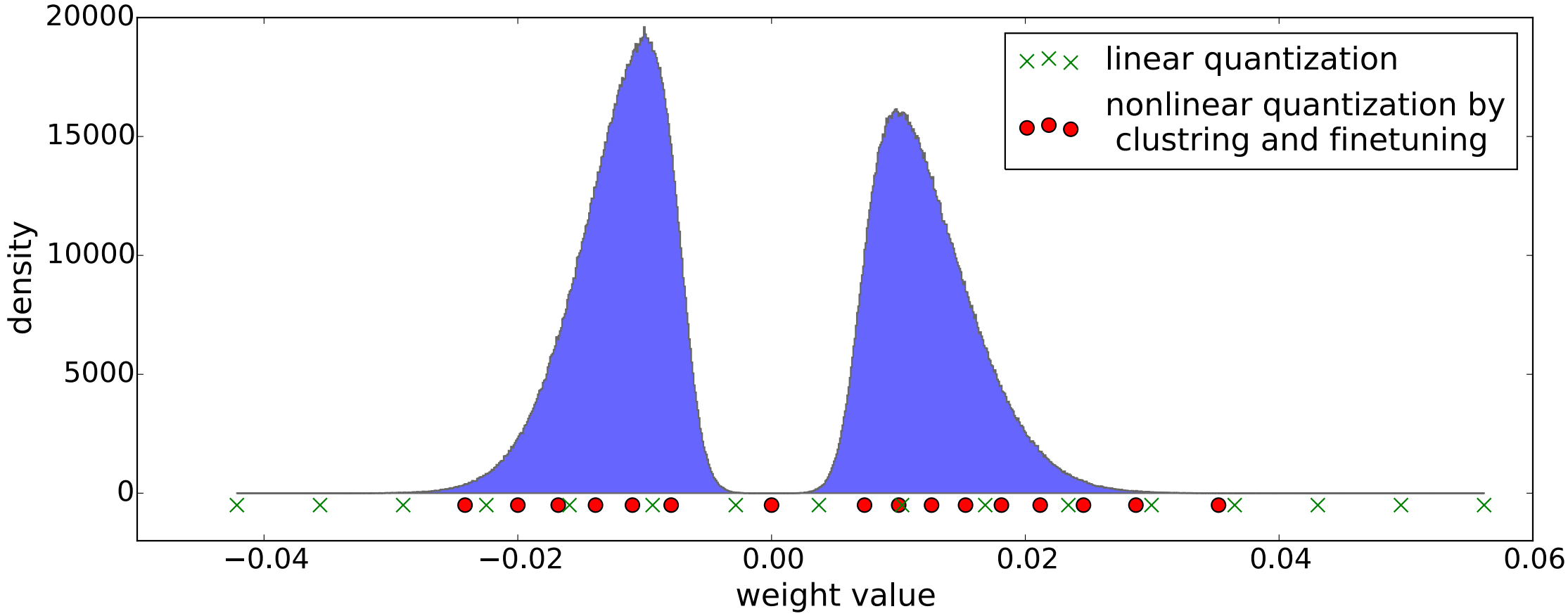


Energy numbers are from Mark Horowitz “Computing’s Energy Problem (and what we can do about it)”, ISSCC 2014
Area numbers are from synthesized result using Design Compiler under TSMC 45nm tech node. FP units used DesignWare Library.

The Importance of Staying Local

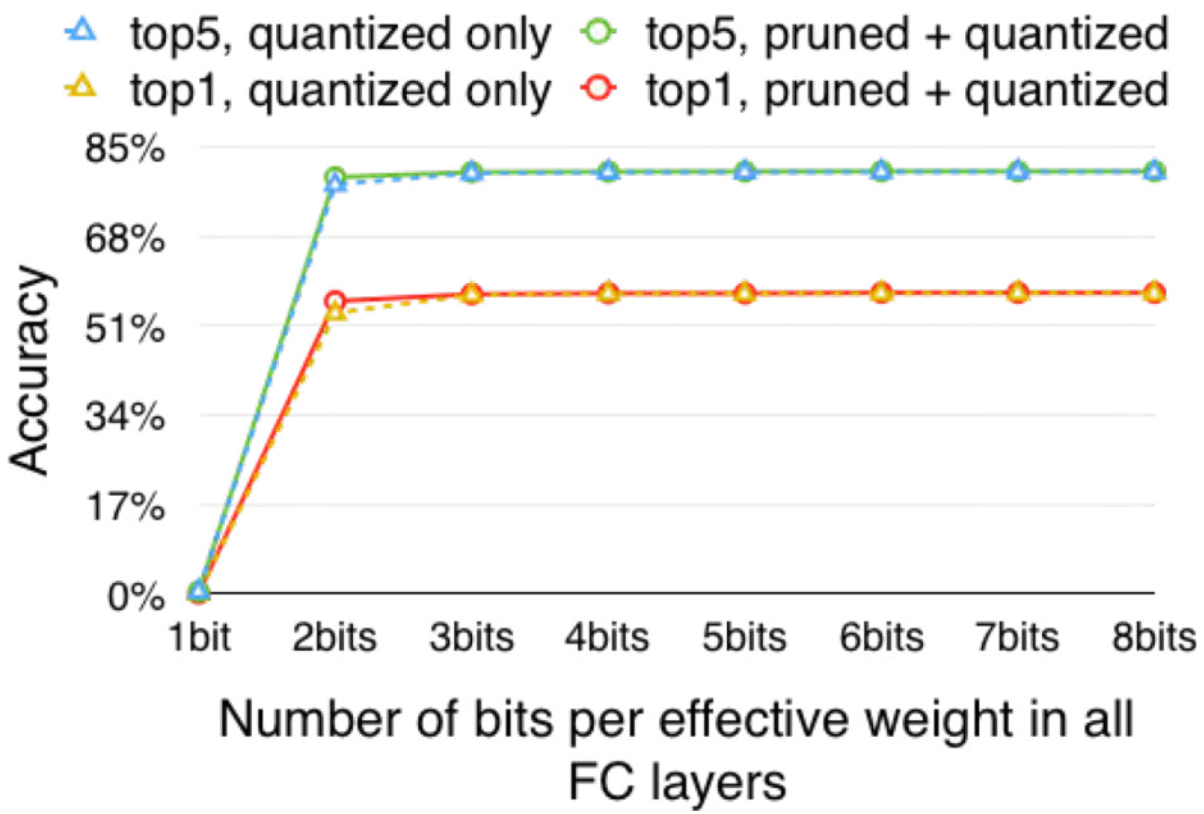


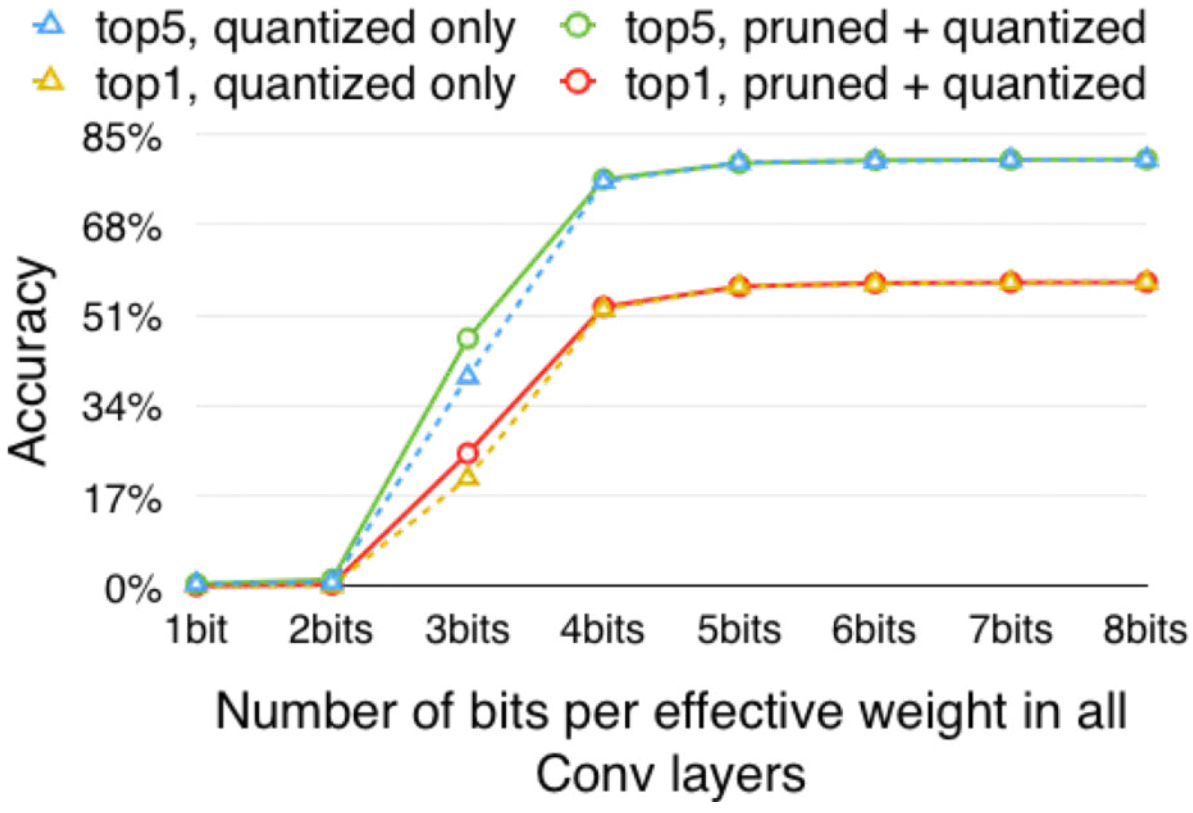
Trained Quantization



Han et al. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding, arXiv 2015

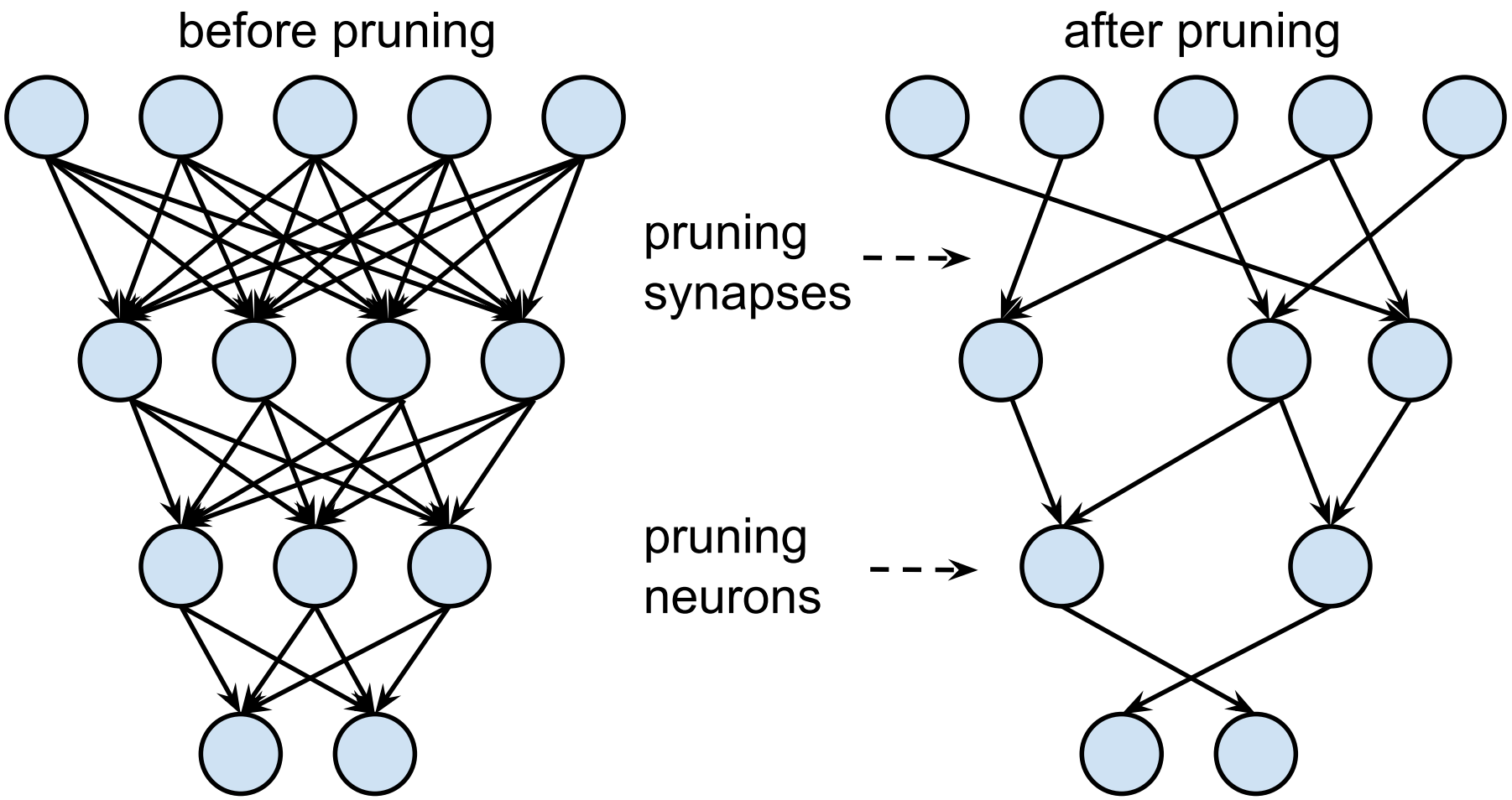
Bits per Weight





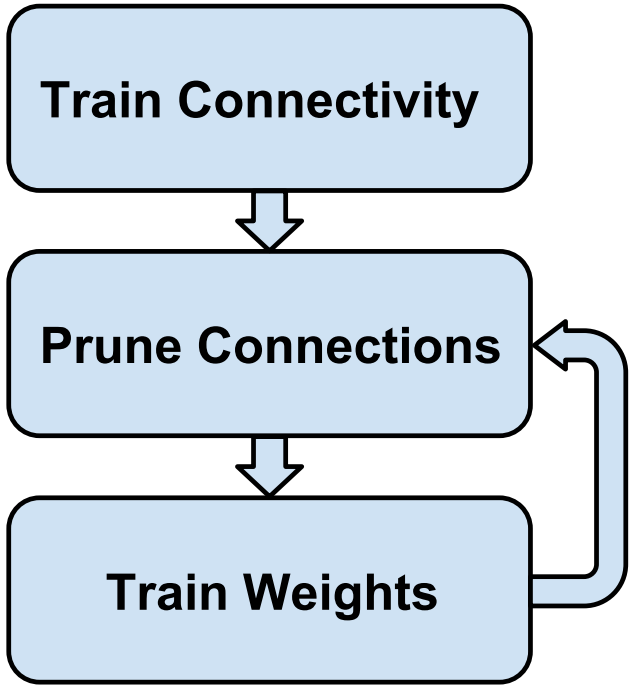
Pruning

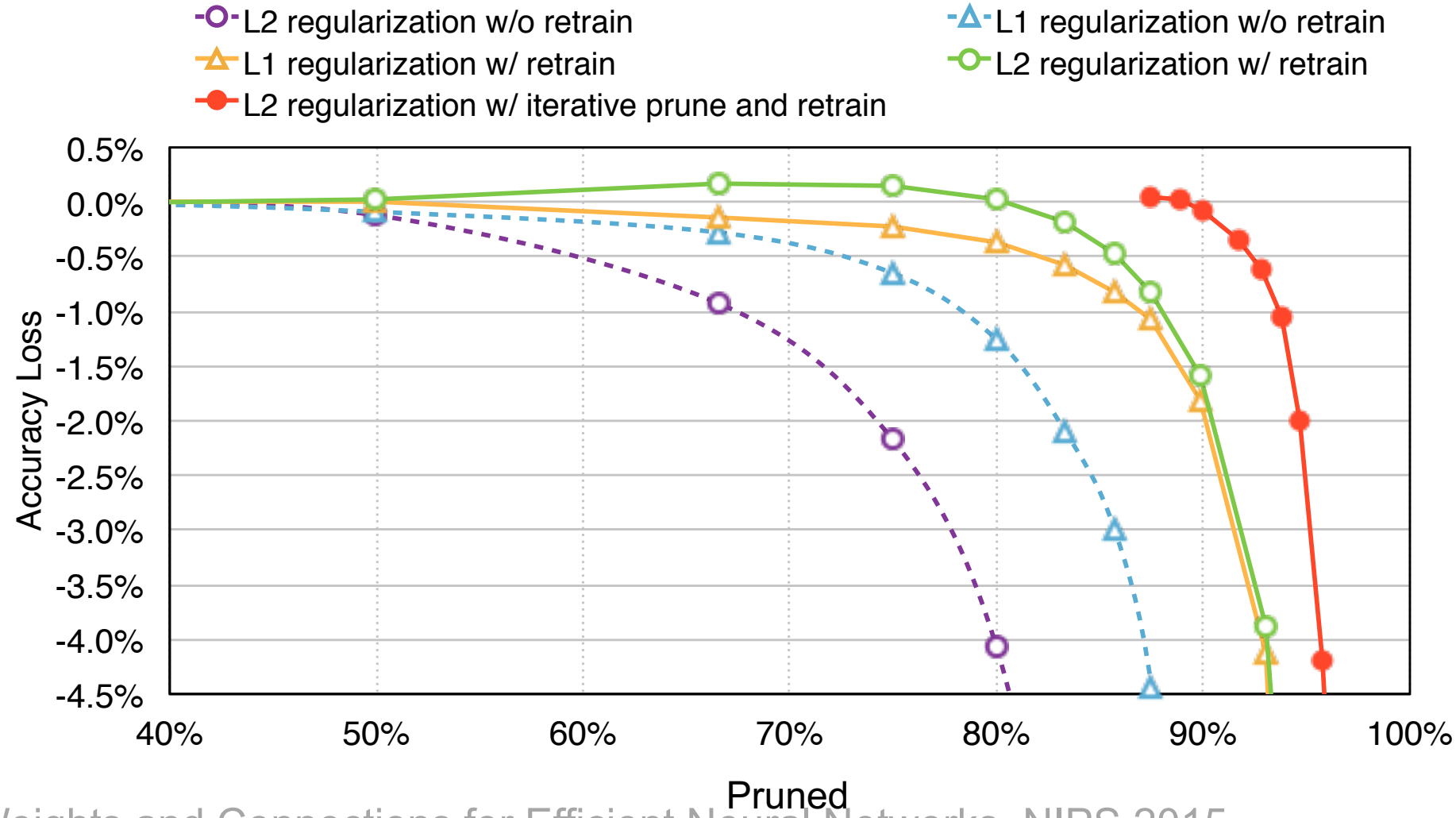
Pruning



Han et al. Learning both Weights and Connections for Efficient Neural Networks, NIPS 2015

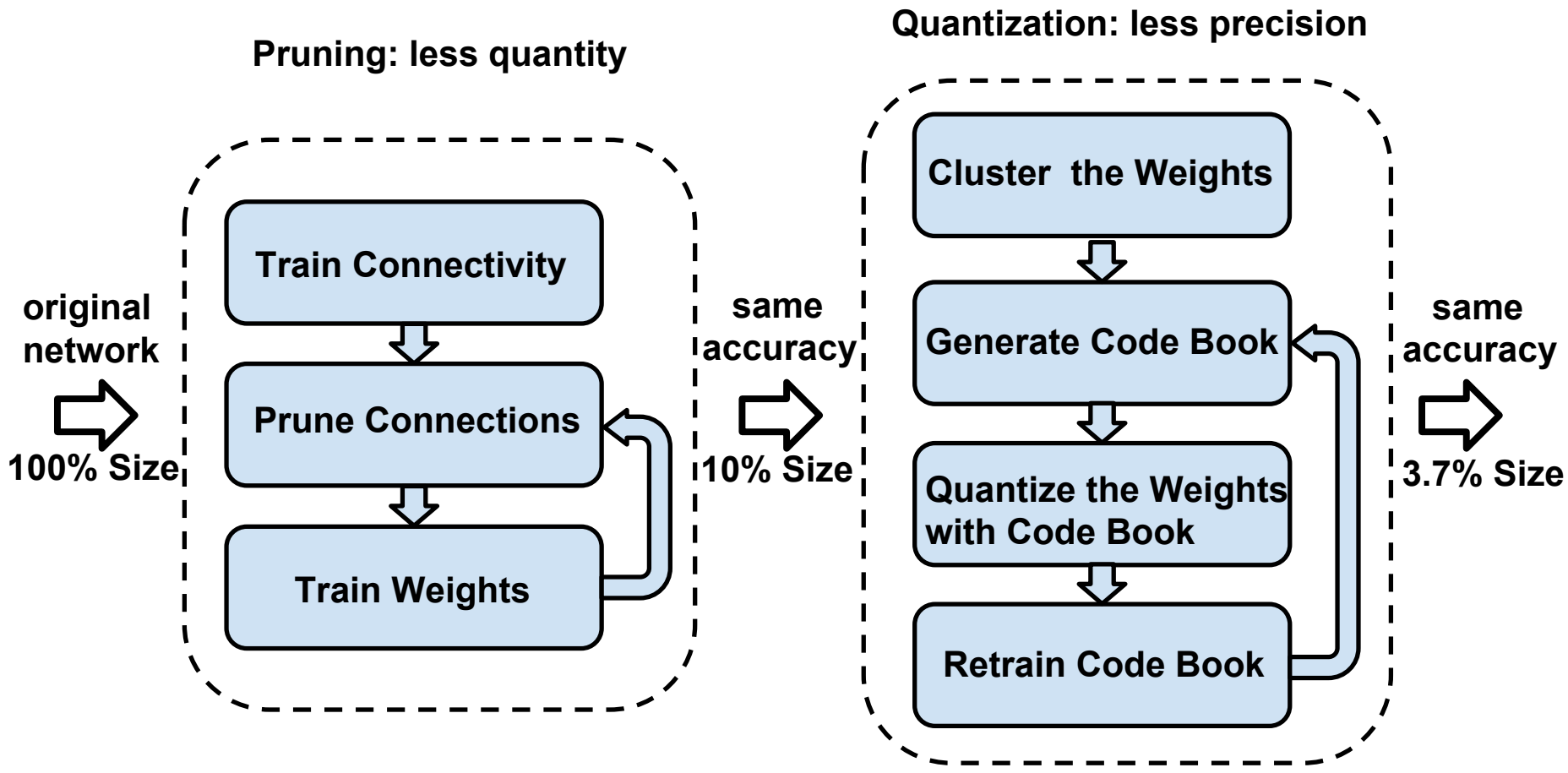
Retrain to Recover Accuracy





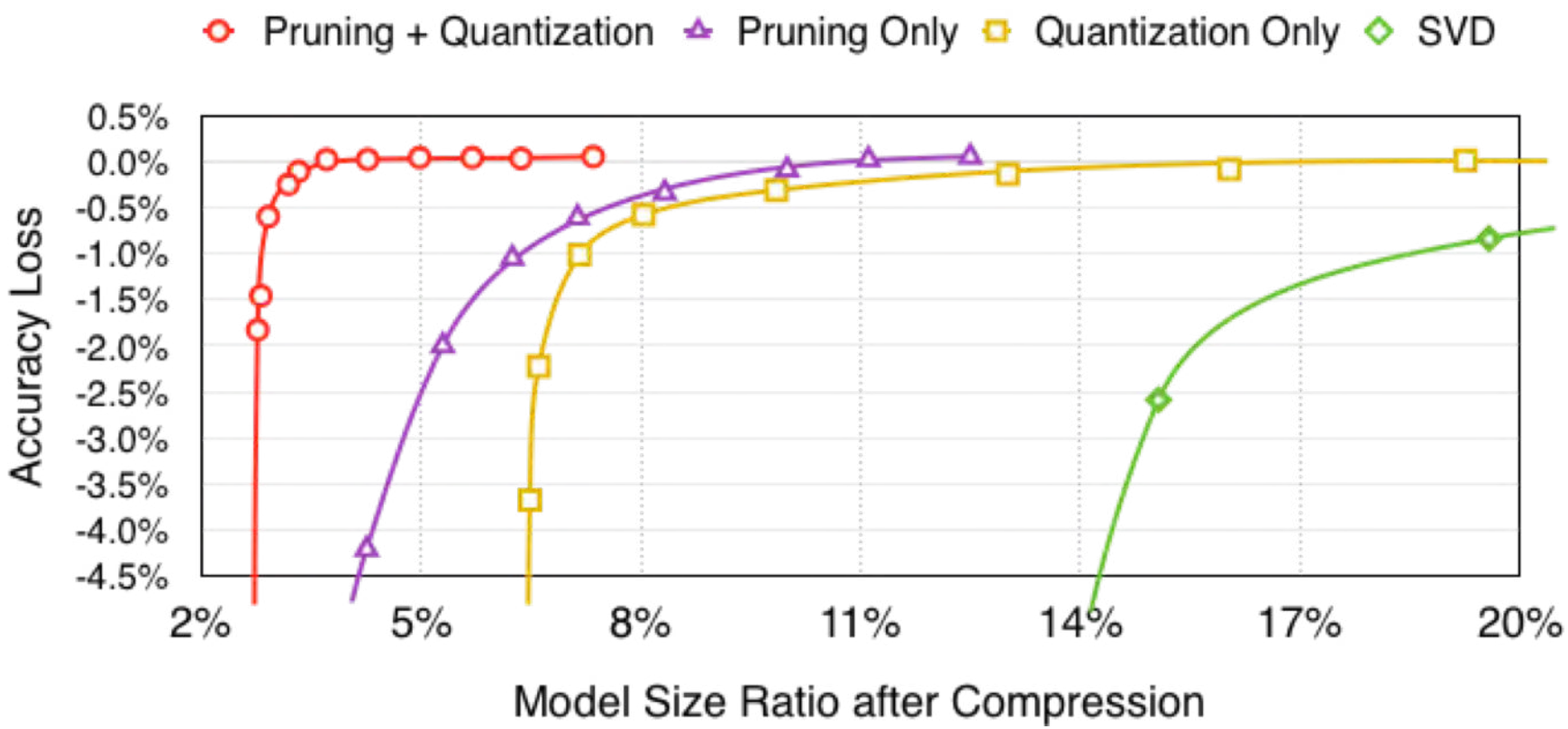
Han et al. Learning both Weights and Connections for Efficient Neural Networks, NIPS 2015

Deep Compression
Combining Pruning with Trained Quantization



Han et al. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding, arXiv 2015, ICLR 2016 – Best Paper

Pruning + Trained Quantization



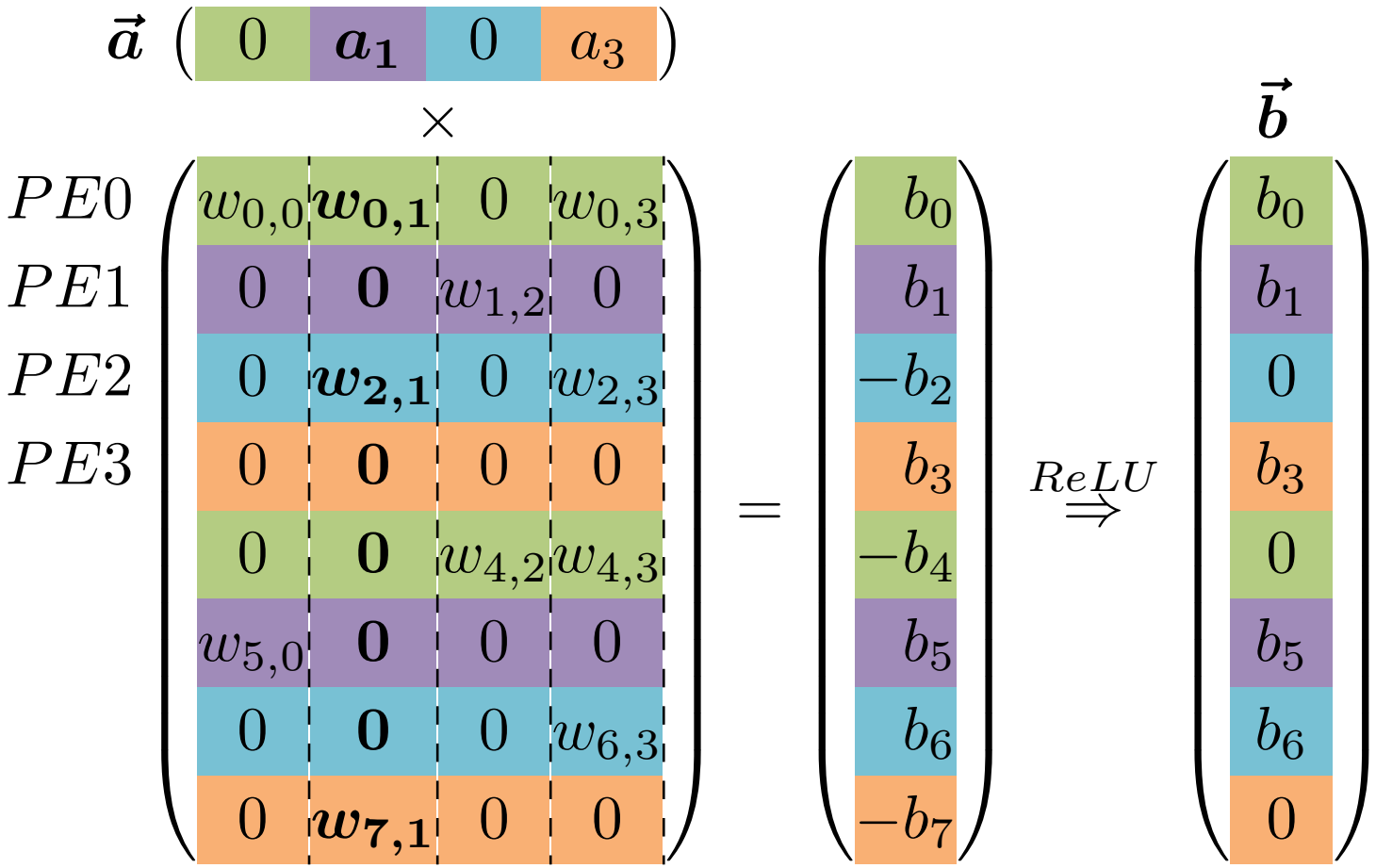
30x – 50x Compression Means

- Complex DNNs can be put in mobile applications (<100MB total)
  - 1GB network (250M weights) becomes 20-30MB
- Memory bandwidth reduced by 30-50x
  - Particuarly for FC layers in real-time applications with no reuse
- Memory working set fits in on-chip SRAM
  - 5pJ/word access vs 640pJ/word

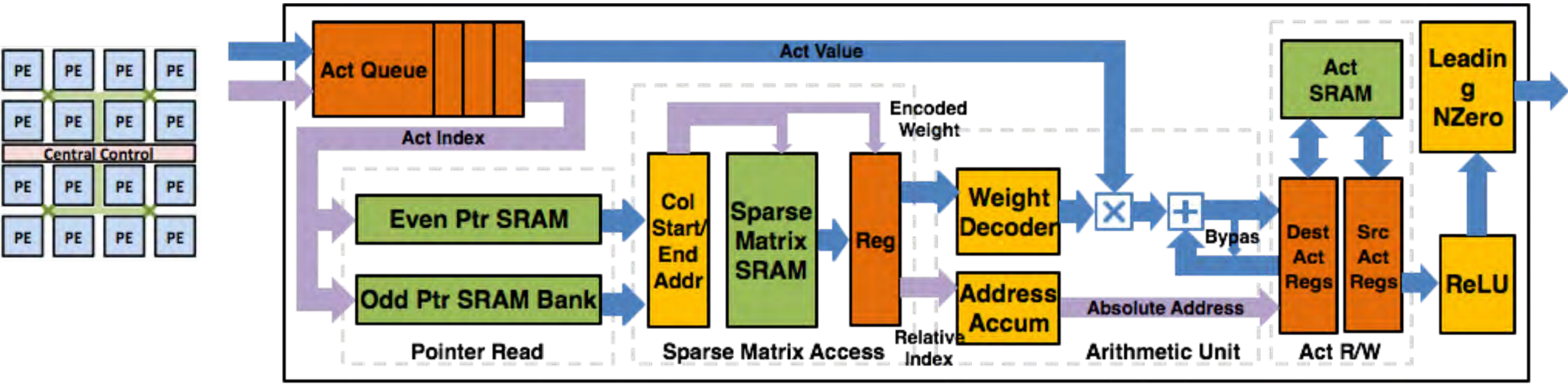
Efficient Inference Engine

Supporting Sparse Weights and Codebooks

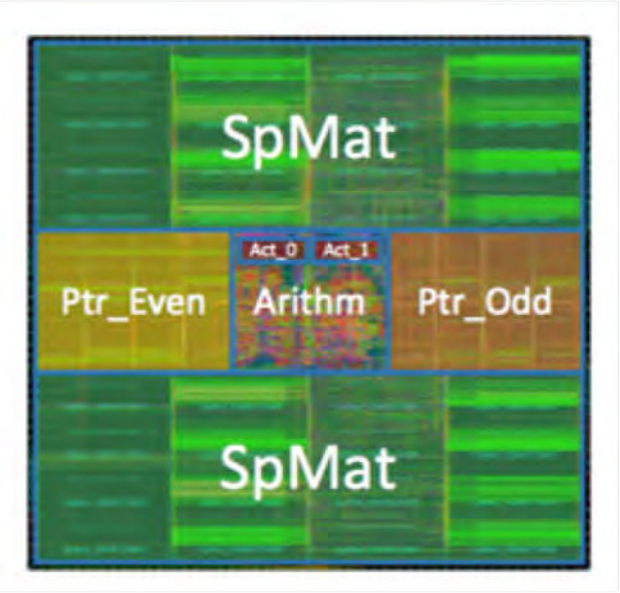
Sparse Matrix Representation



EIE Architecture



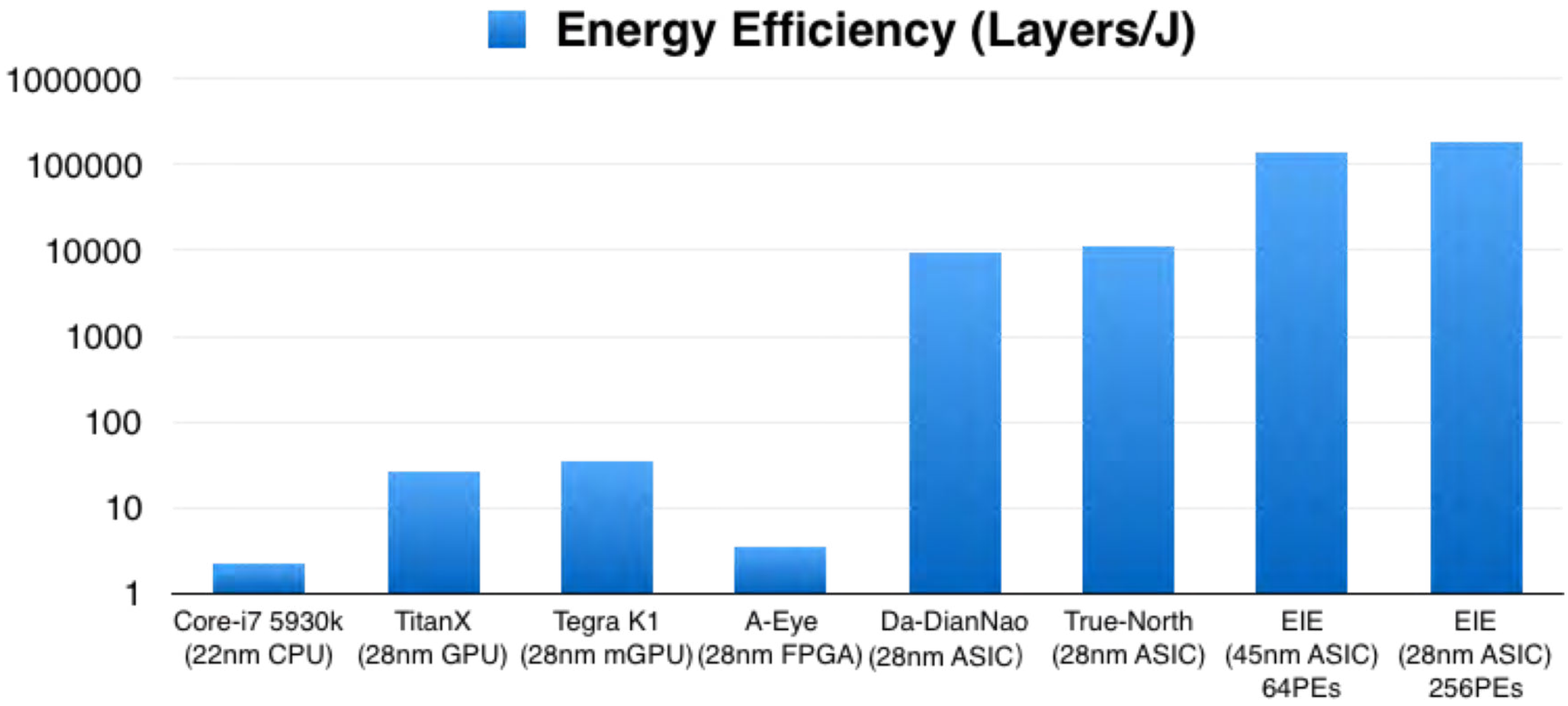
Implementation



Technology	45 nm
# PEs	64
on-chip SRAM	8 MB
Max Model Size	84 Million
Static Sparsity	10x
Dynamic Sparsity	3x
Quantization	4-bit
ALU Width	16-bit
Area	40.8 mm^2
MxV Throughput	81,967 layers/s
Power	586 mW

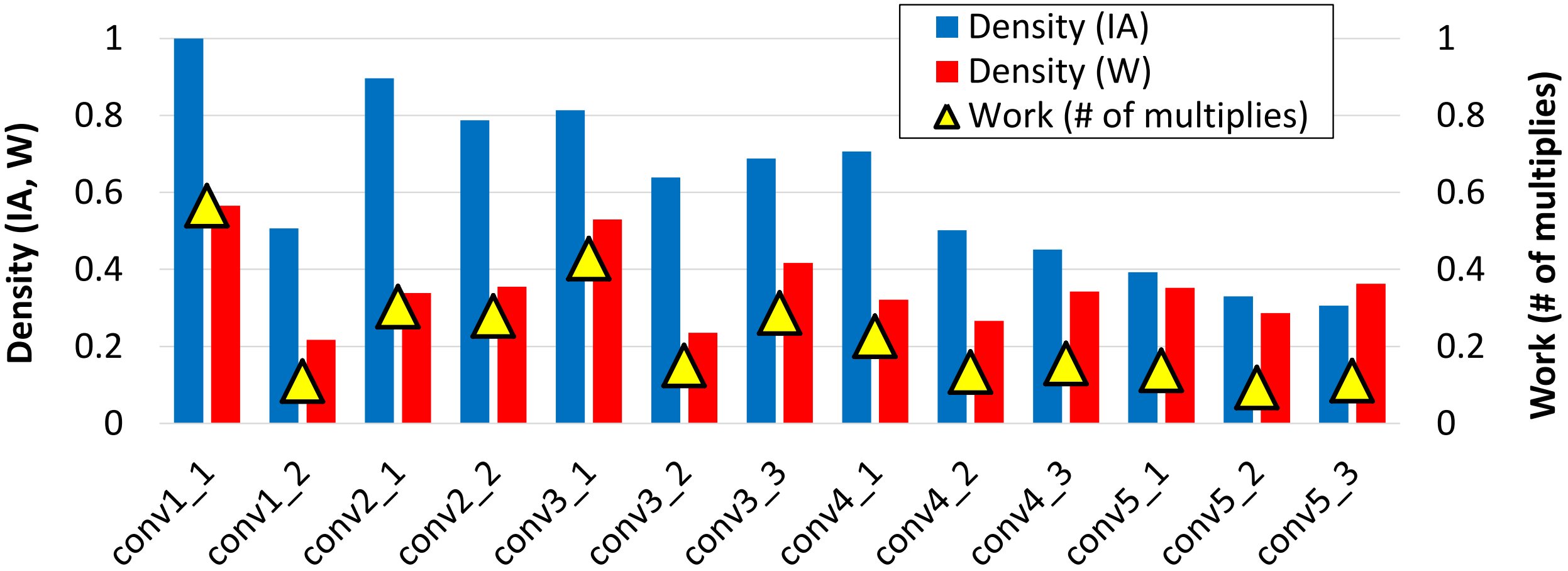
1. Post layout result
2. Throughput on AlexNet FC-7

Comparison: Energy Efficiency



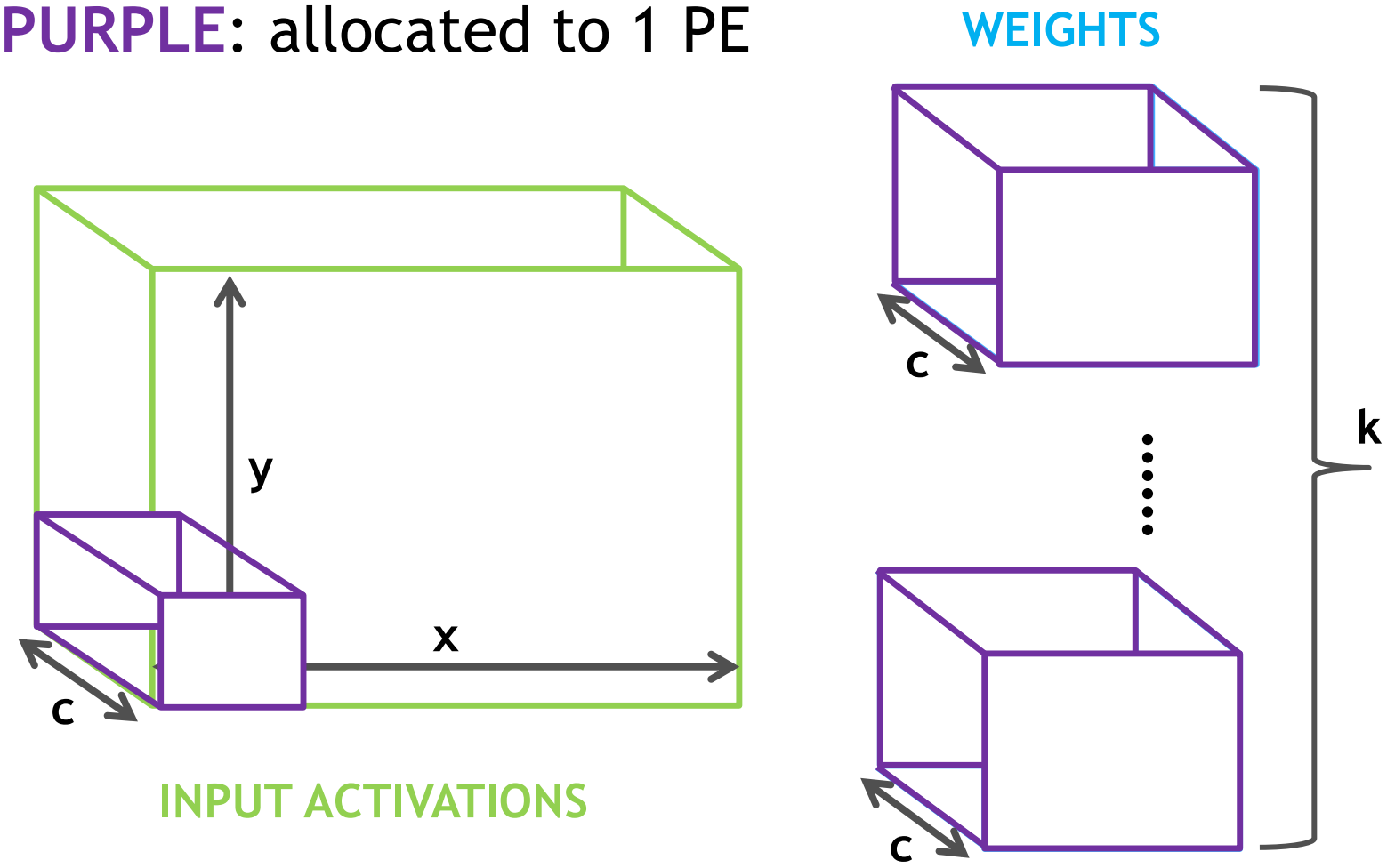
Sparse Convolutional Accelerator

Density of VGGNet



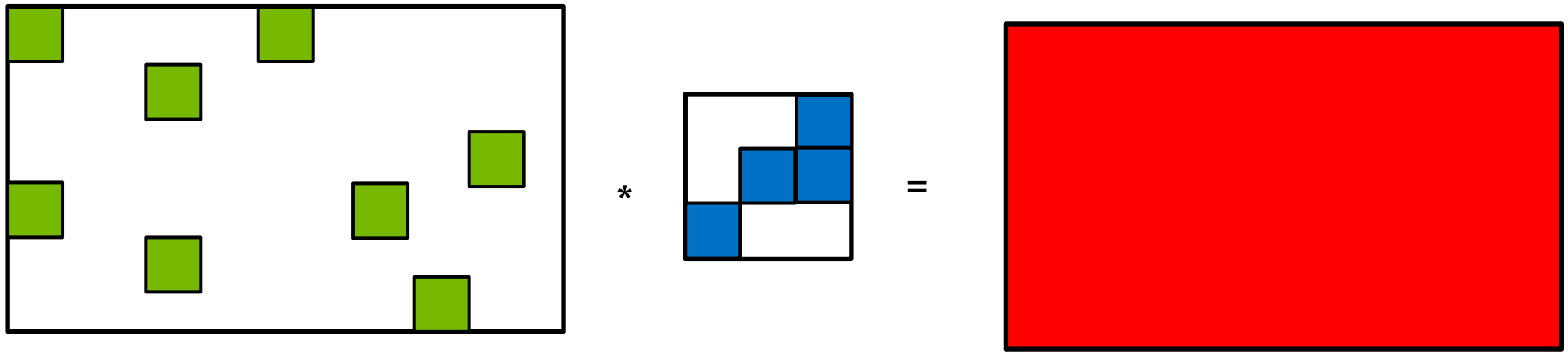
Blocking CNN Inference

PURPLE: allocated to 1 PE



Sparse Convolution

- Only compute where both operands are nonzero
- 10-30x Reduction in work



Sparse Convolution Engine

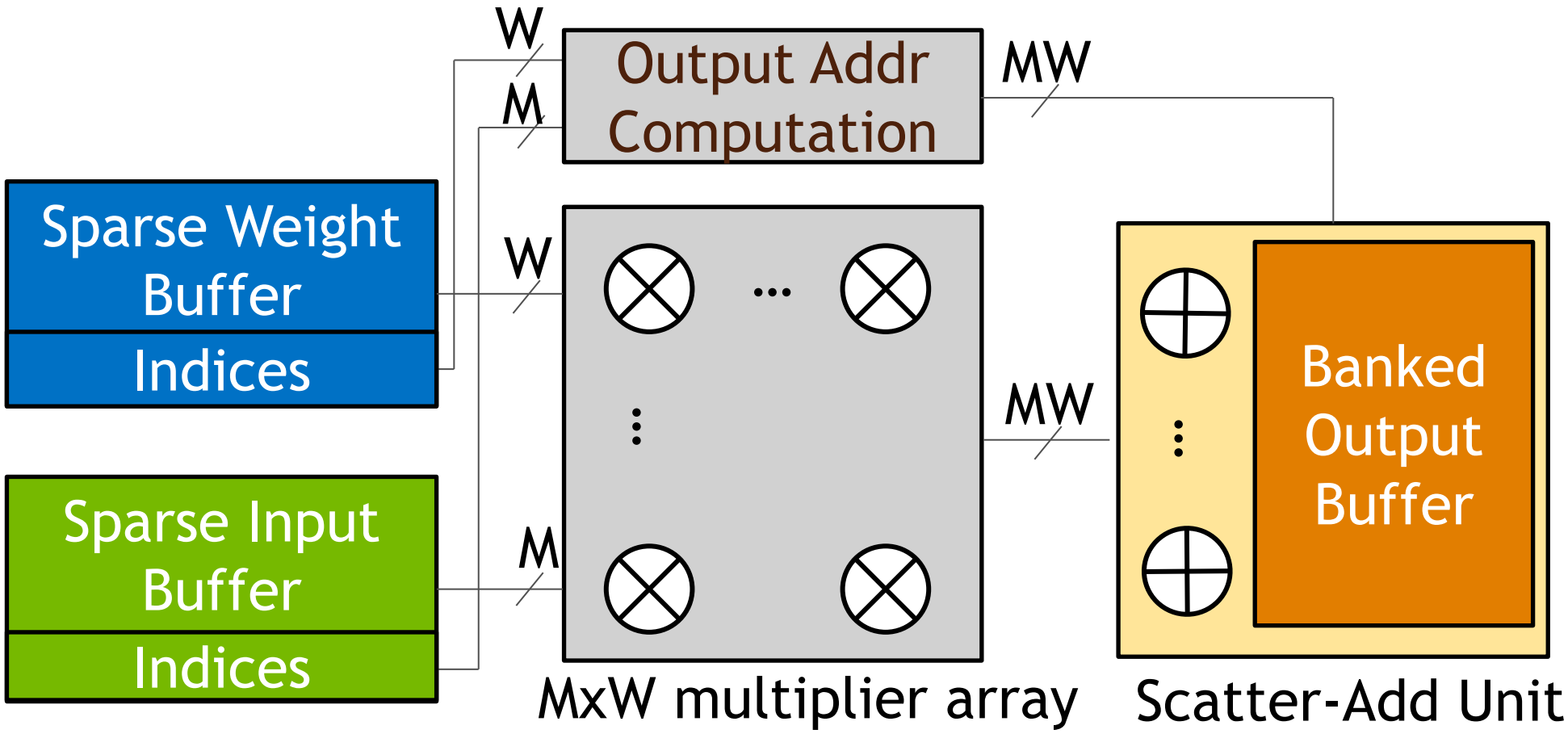
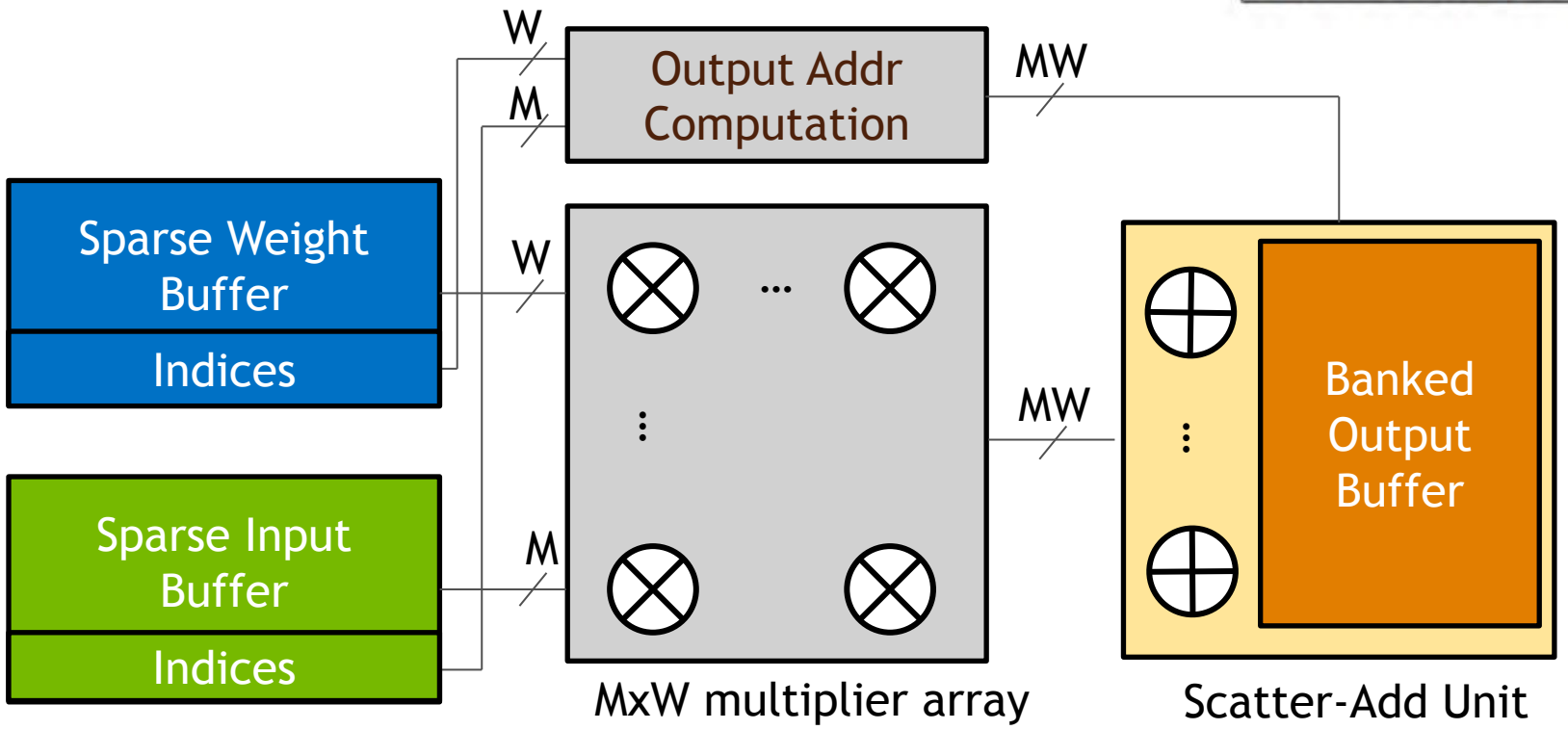
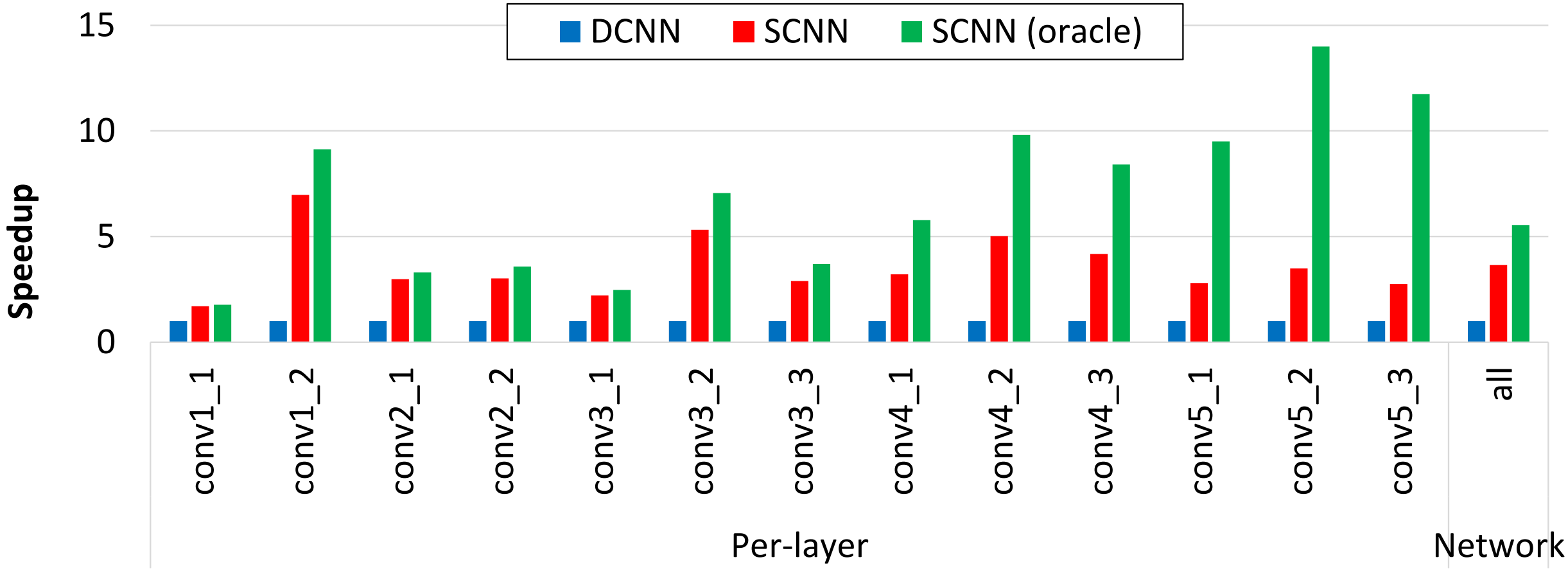


Table 3: SCNN PE area breakdown.

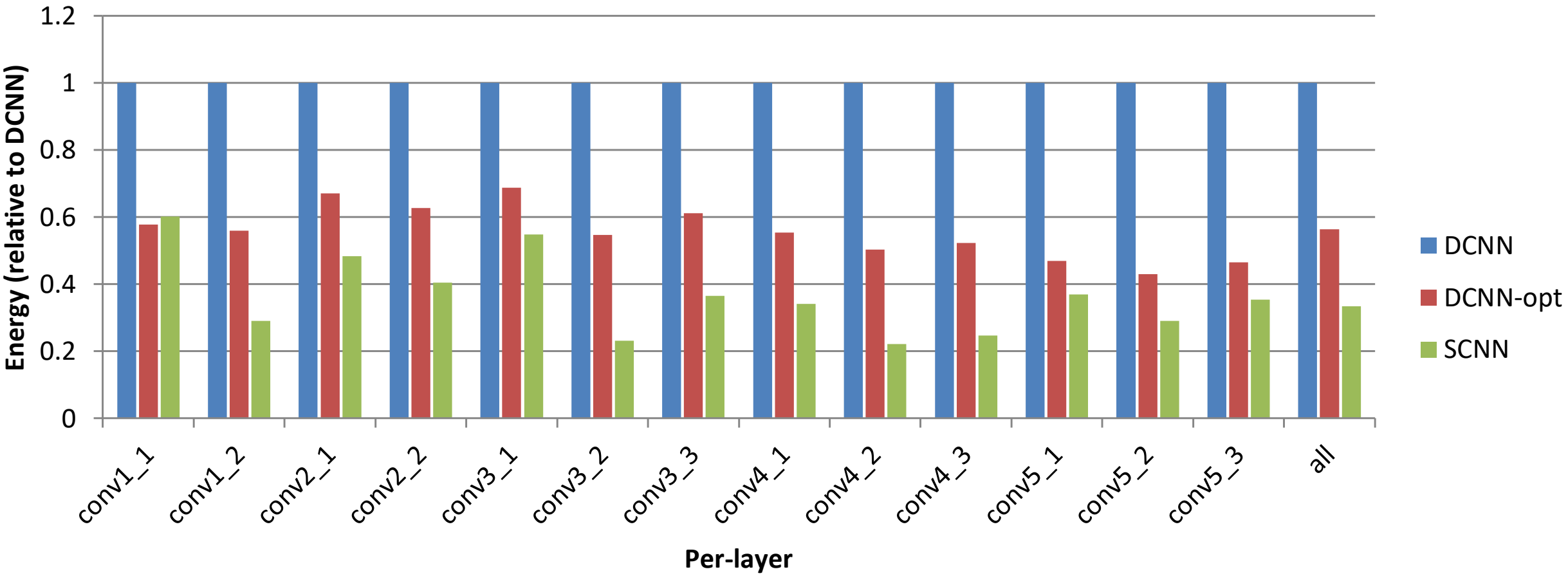
PE Component	Size	Area (mm^2)
IARAM + OARAM	20 KB	0.026
Weight FIFO	0.5 KB	0.004
Multiplier array	16 ALUs	0.005
Scatter network	16×32 crossbar	0.023
Accumulator buffers	6 KB	0.038
Other	—	0.019
Total	—	0.115
Accelerator total	64 PEs	7.4



Speedup for VGGNet



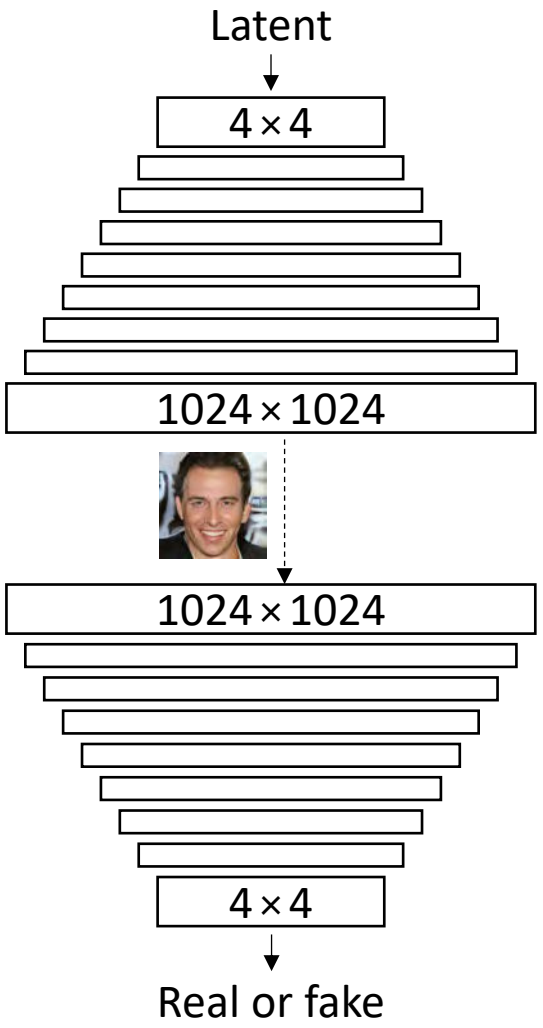
Energy Efficiency



Scaling Applications

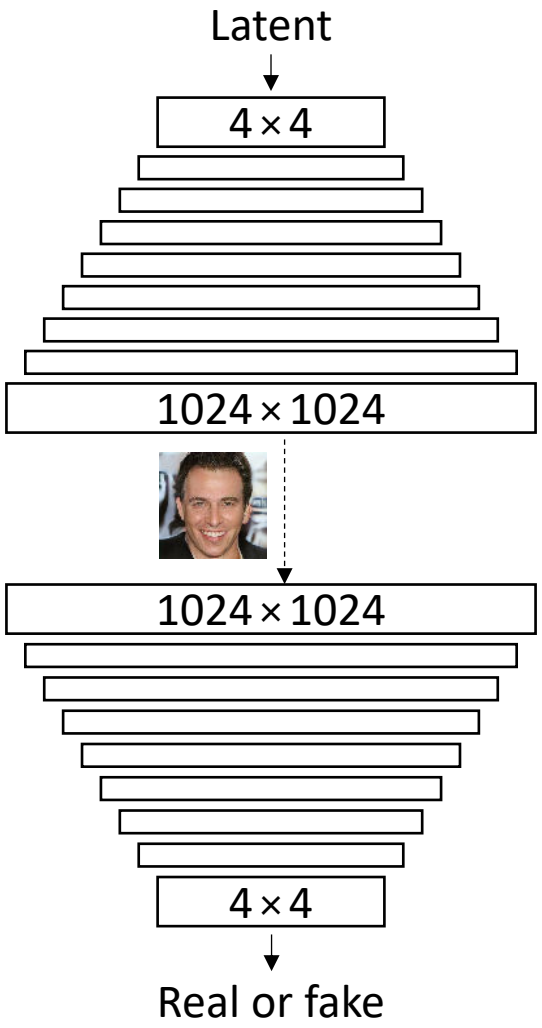
Progressive GAN

Traditional approach

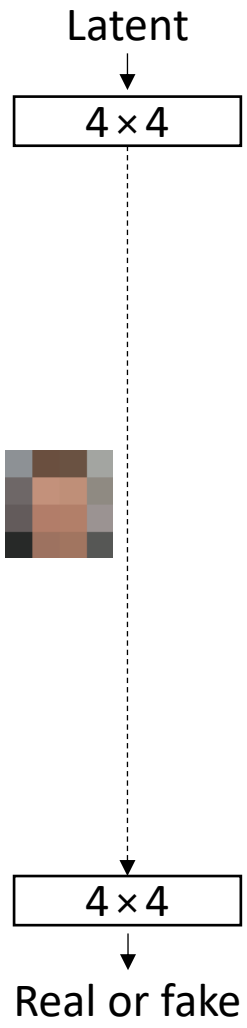


- High resolution ⇒ deep networks
- All layers are initialized to random weights
- Neither of the two networks has any idea what it's supposed to do!

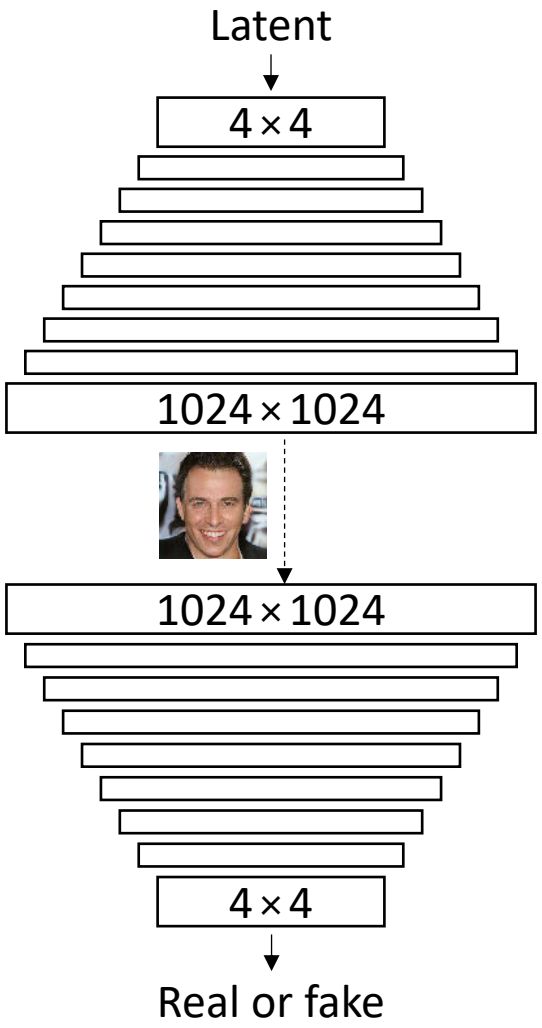
Traditional approach



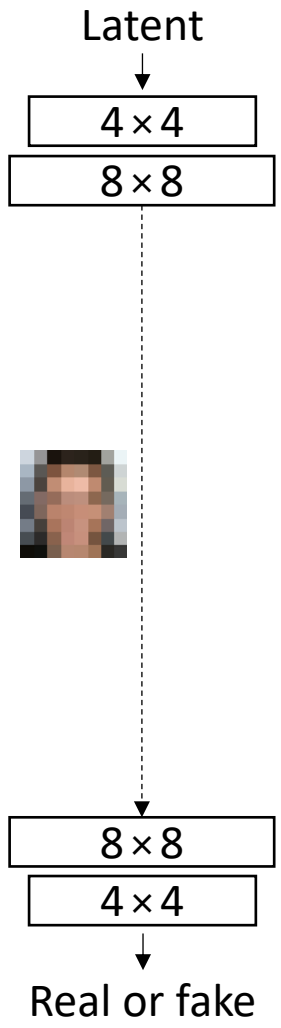
Progressive growing



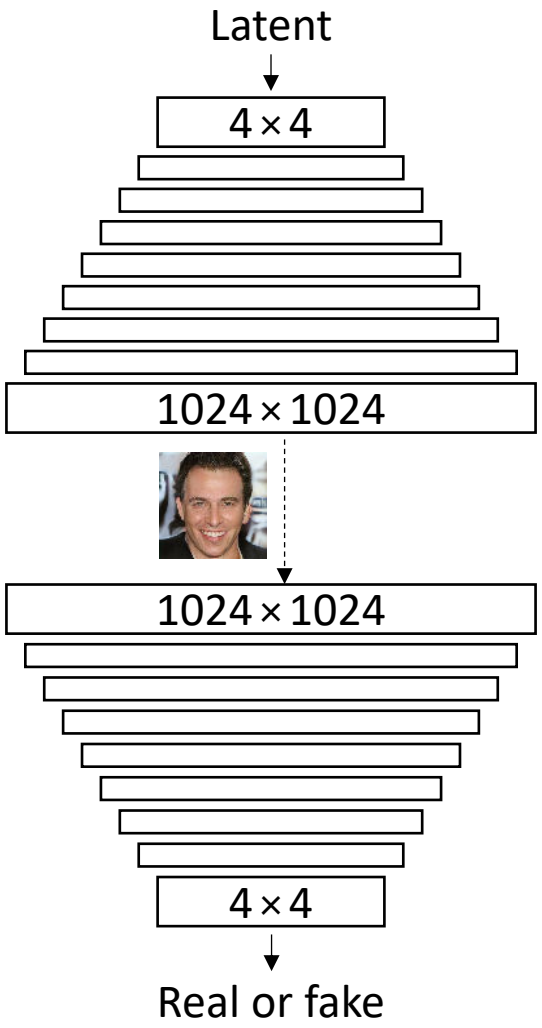
Traditional approach



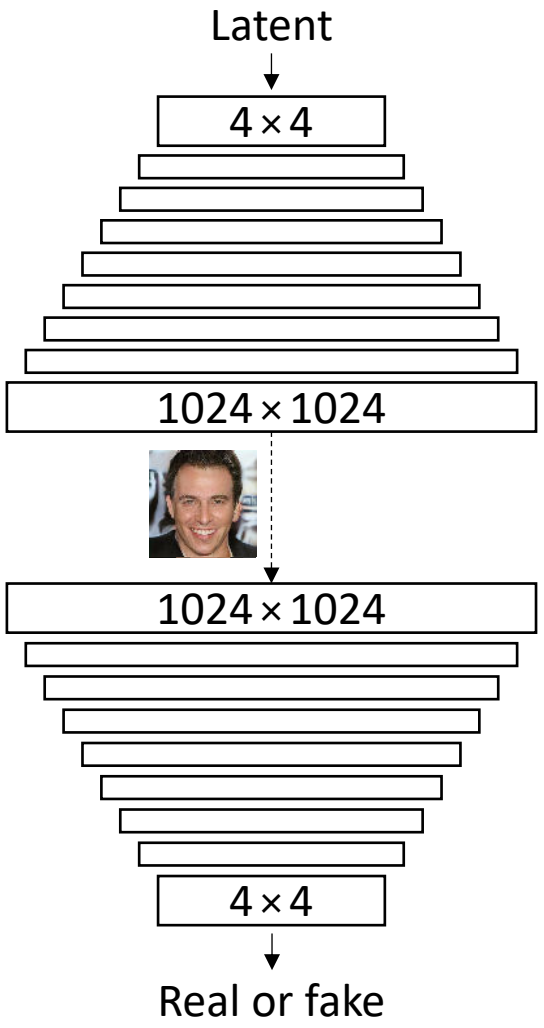
Progressive growing



Traditional approach



Progressive growing

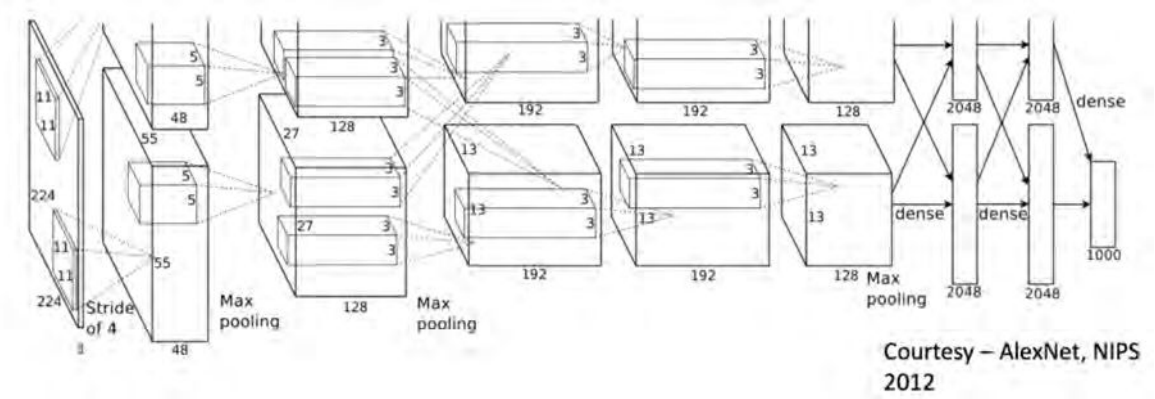


Conclusion

Summary

- **Deep Learning** is fueling a revolution in AI - **Transportation, Health Care, Education, and Graphics**
- **Hardware** has **Enabled** this revolution – and is **Limiting Scaling**
- **Power-Law Scaling** - Accuracy improves with more data, larger models needed
- **Demanding Today** – 9Tops for HD video – 100s for AV
- **GPUs**
  - GV100 120TF HP, 400 GF/W, 900GB/s
  - Xavier 20DL TOPS, 2 TOPS/W
- **Continue Scaling** after Moore's Law
- **Sparsity** – only 30% of activations are non-zero, can prune to 30% weight density
- **Trained Quantization** – only need 4-6 bits to represent weights
- **Accelerators**
  - **EIE** – exploits pruning and TQ for FC layers – **24,000x efficiency gain**
  - **SCNN** – exploits sparsity in Conv layers
- **Better AI** through **Faster Hardware**

Thank You



Courtesy – AlexNet, NIPS 2012