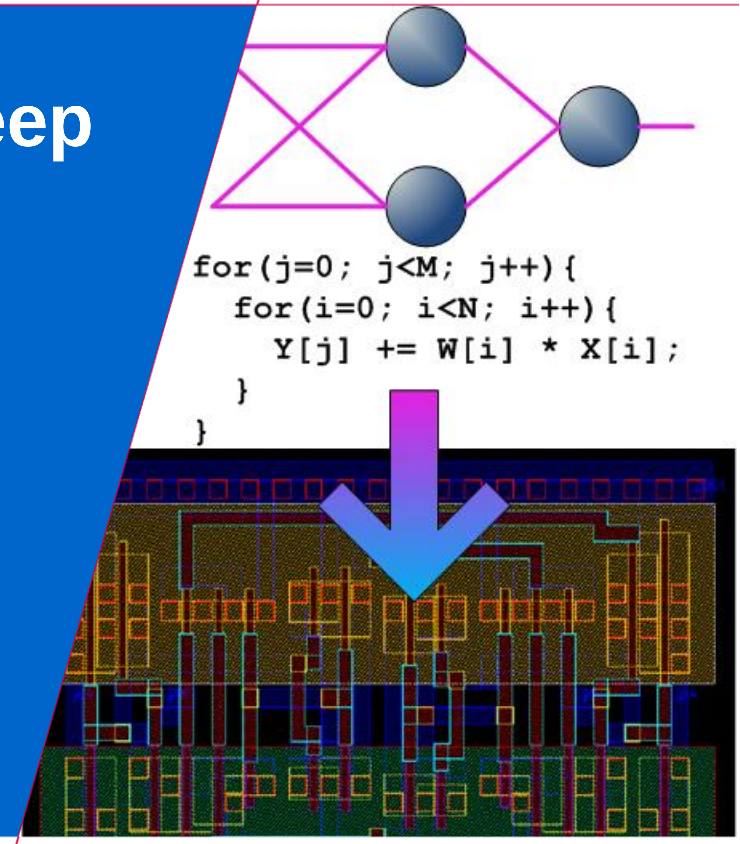
Electronic Systems

Improving the Efficiency of Deep Learning

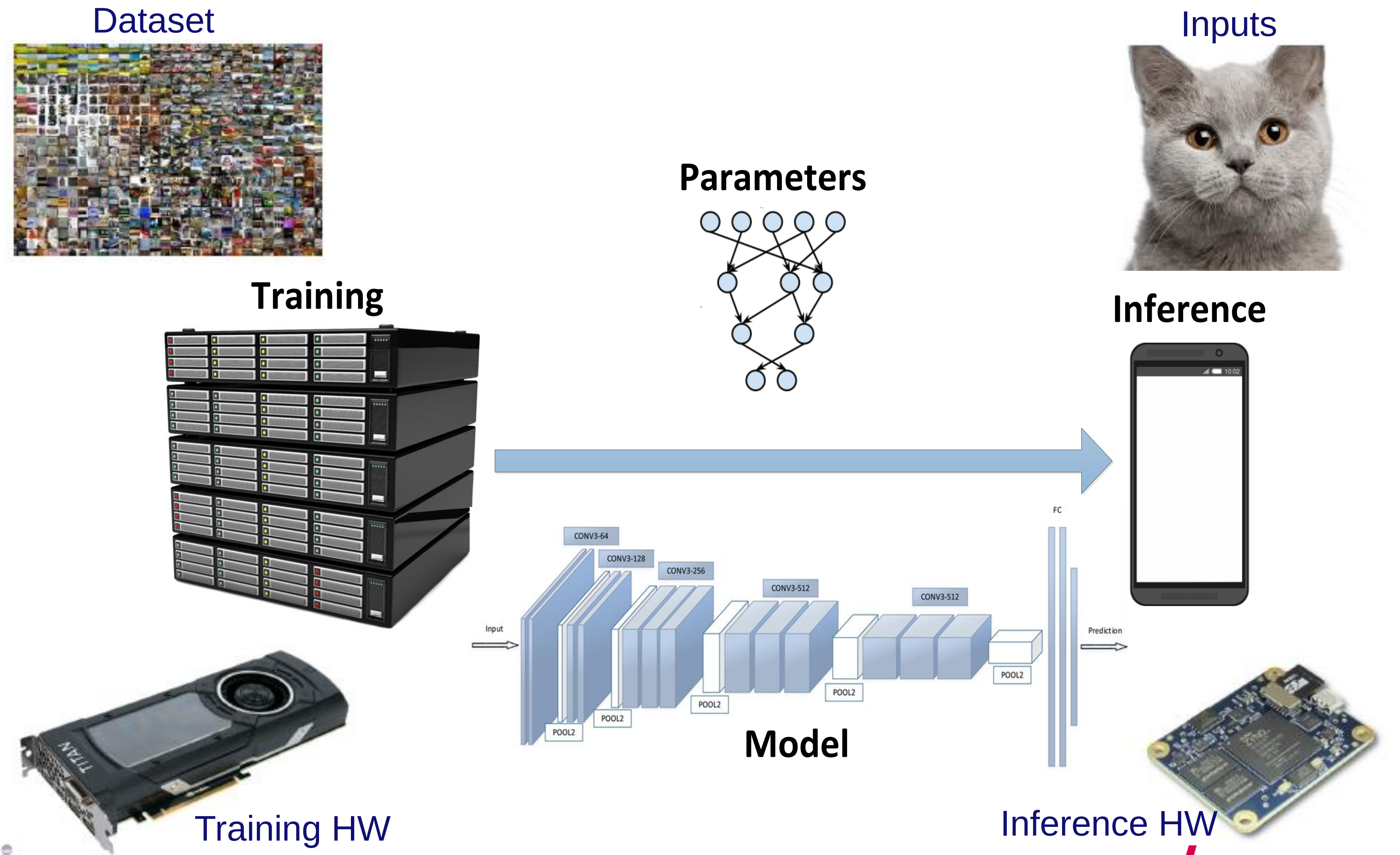
Accelerating Deep Learning Applications



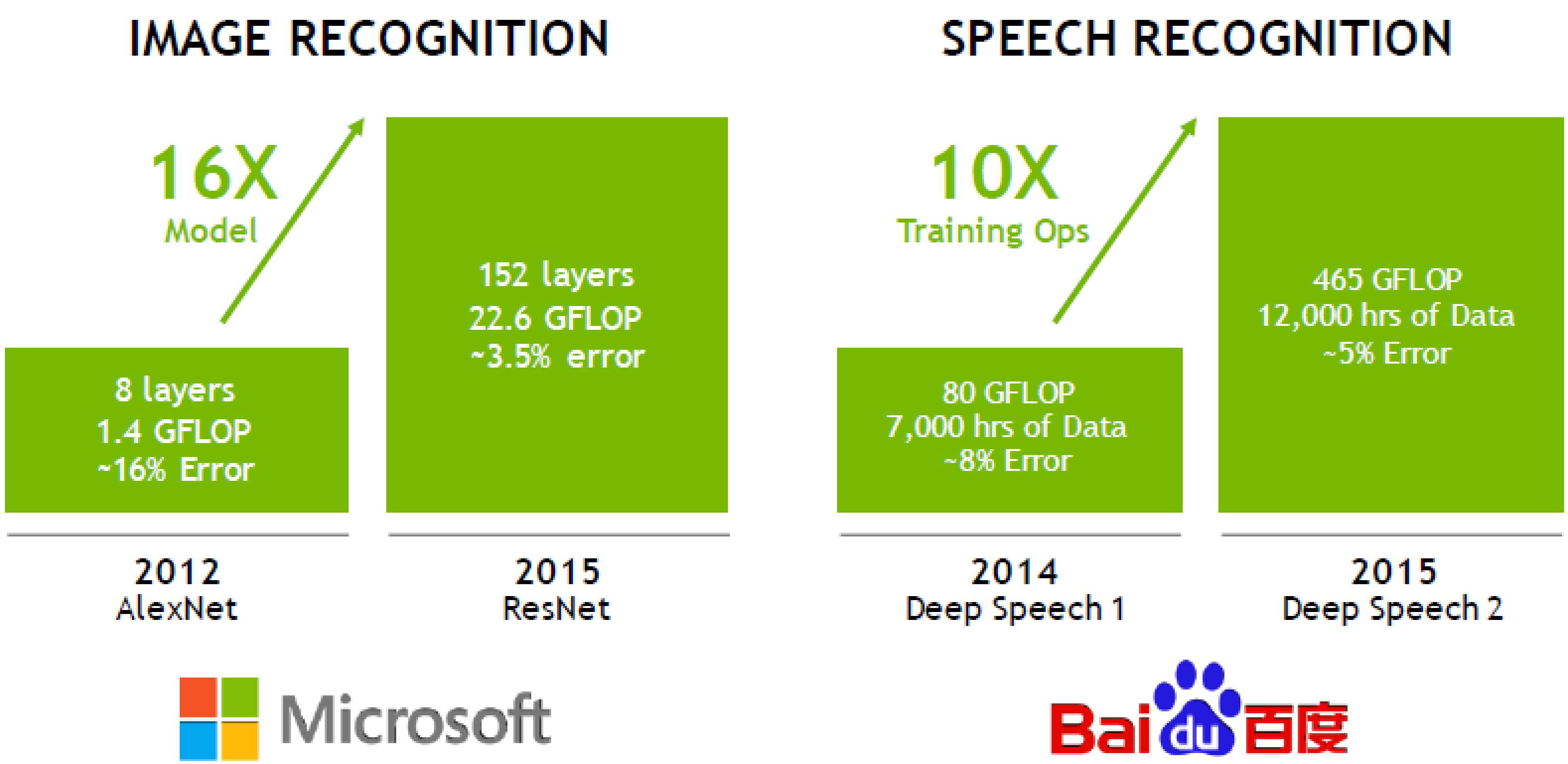
By: Maurice Peemen

Date: 31-5-2017

The deep learning setup



Models are getting larger



Dally, NIPS'2016 workshop on Efficient Methods for Deep Neural Networks

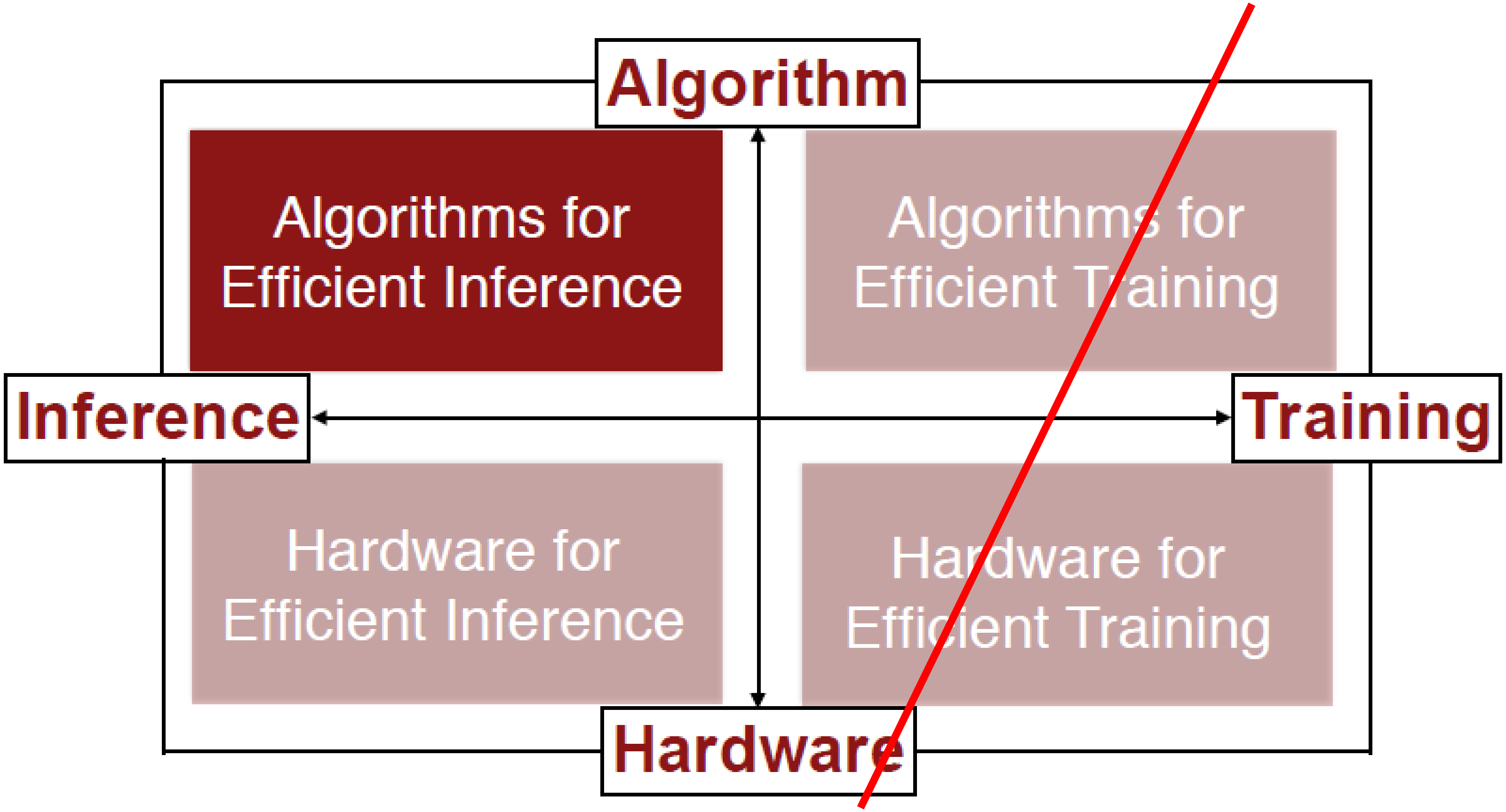
The Efficiency Problem of Deep Learning

- **Computation Intensive**
- **Memory Intensive**
- **Difficult to Deploy**

- **AlphaGo: 1920 CPUs and 280 GPUs**
  $3000 electric bill per game

TU/e Technische Universiteit Eindhoven University of Technology

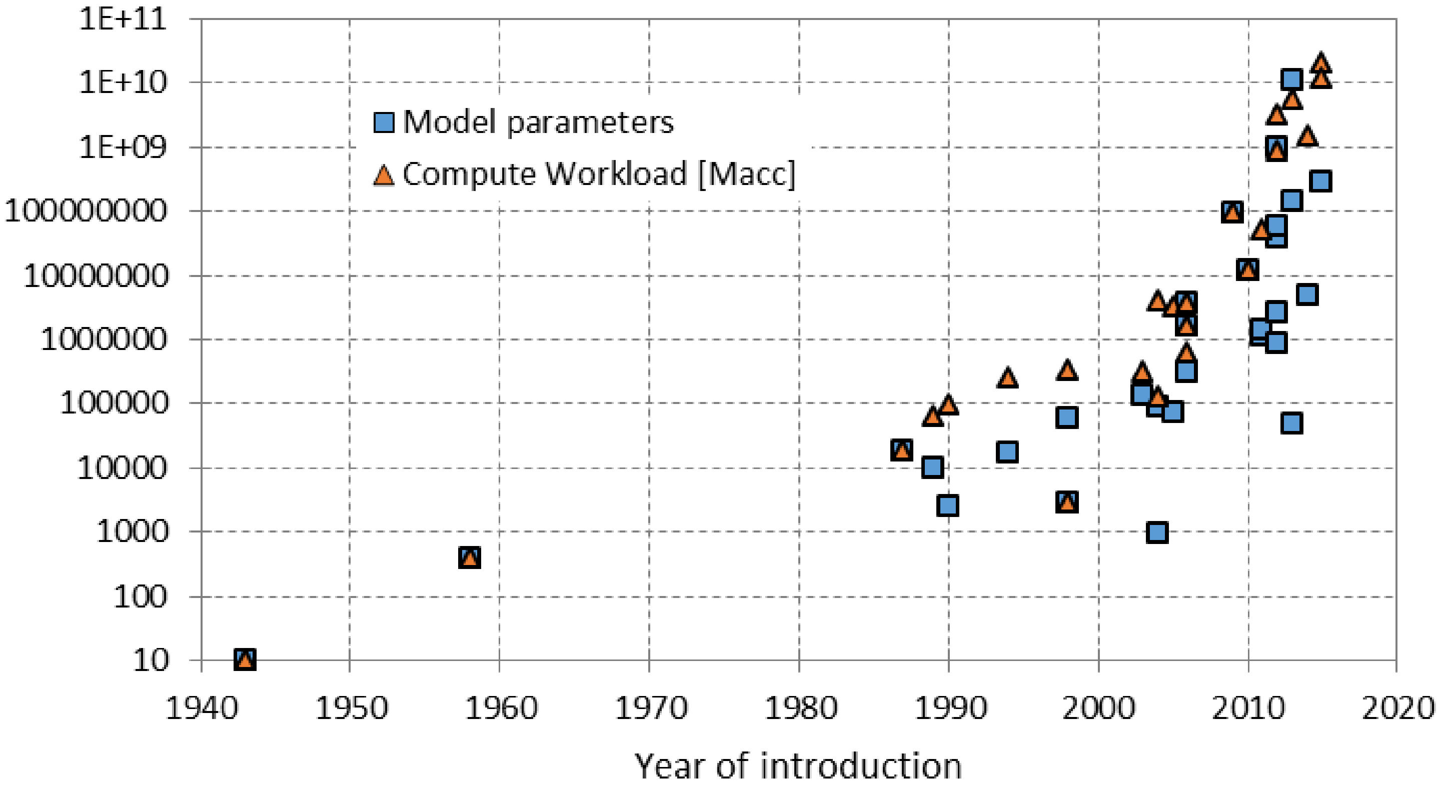
Agenda



Technische Universiteit Eindhoven University of Technology

The Problem of Large DNN Models

- **App developers suffer from the model size**



Technische Universiteit **Eindhoven** University of Technology

Large DNN on Mobile

- **Large models => more references => more energy**

Operation	Energy [pJ]	Relative Cost
32 bit int ADD	0.1	1
32 bit float ADD	0.9	9
32 bit Register File	1	10
32 bit int MULT	3.1	31
32 bit float MULT	3.7	37
32 bit SRAM Cache	5	50
32 bit DRAM Memory	**640**	**6400**

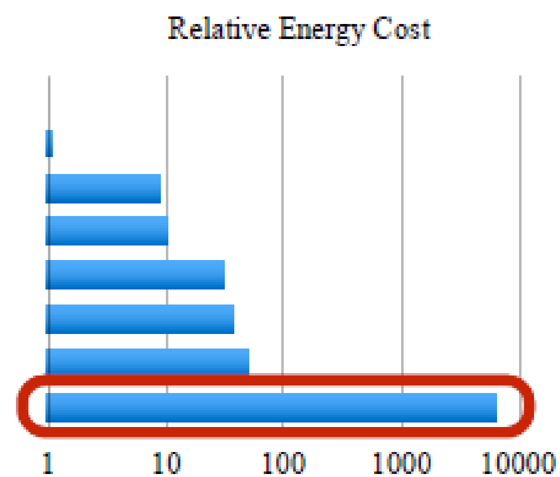


Figure 1: Energy table for 45nm CMOS process. Memory access is 2 orders of magnitude more energy expensive than arithmetic operations.

Advanced Algorithms for Efficient Inference

- **Pruning**
- **Weight sharing**
- **Quantization**
- **Huffman Coding**

- **Best paper ICLR 2016**

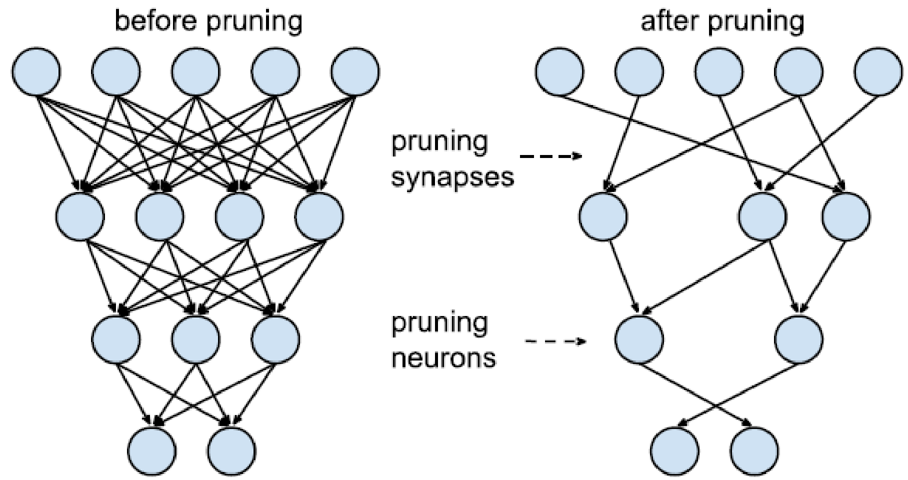
DEEP COMPRESSION: COMPRESSING DEEP NEURAL NETWORKS WITH PRUNING, TRAINED QUANTIZATION AND HUFFMAN CODING

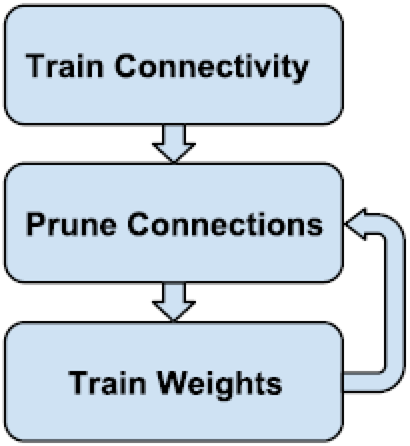
Song Han
Stanford University, Stanford, CA 94305, USA
songhan@stanford.edu

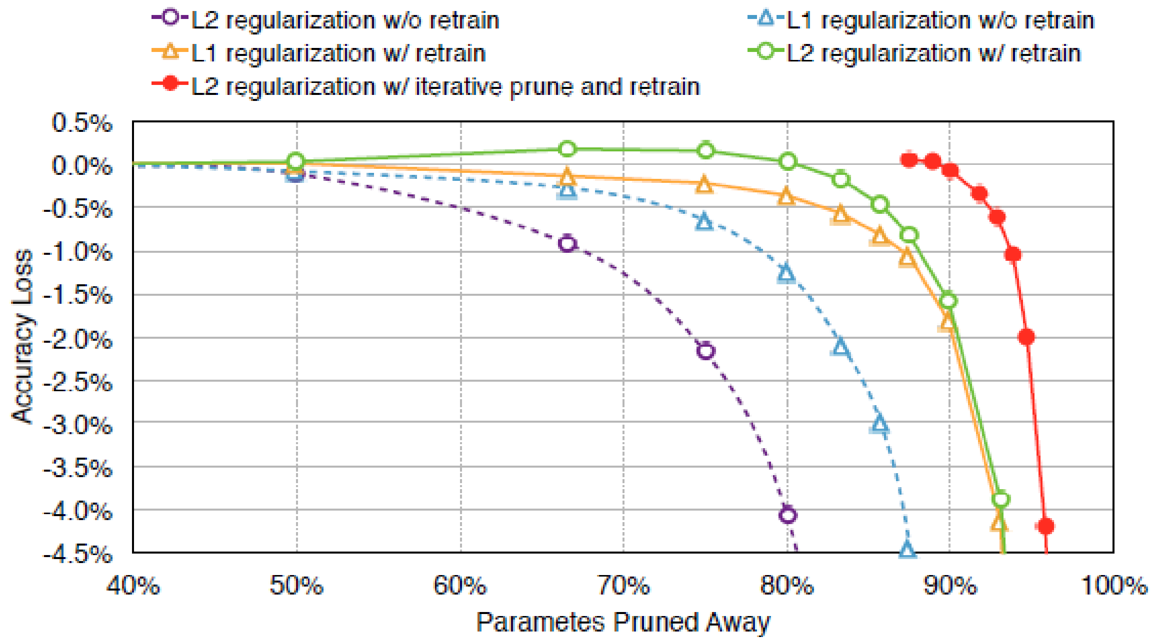
Technische Universiteit **Eindhoven** University of Technology

Pruning Networks

- **Not all parameters are important**
- **Remove some**
- **Retrain to reduce errors**



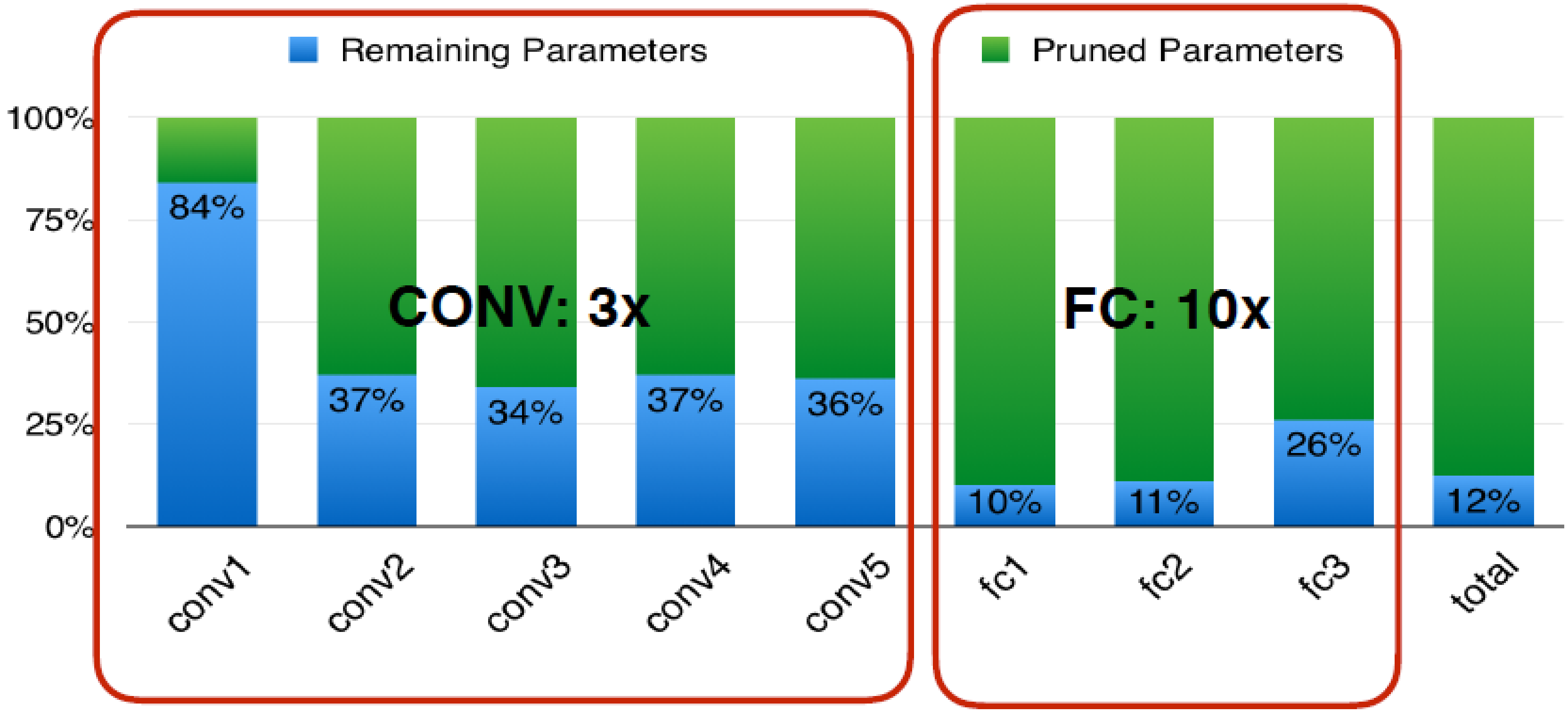




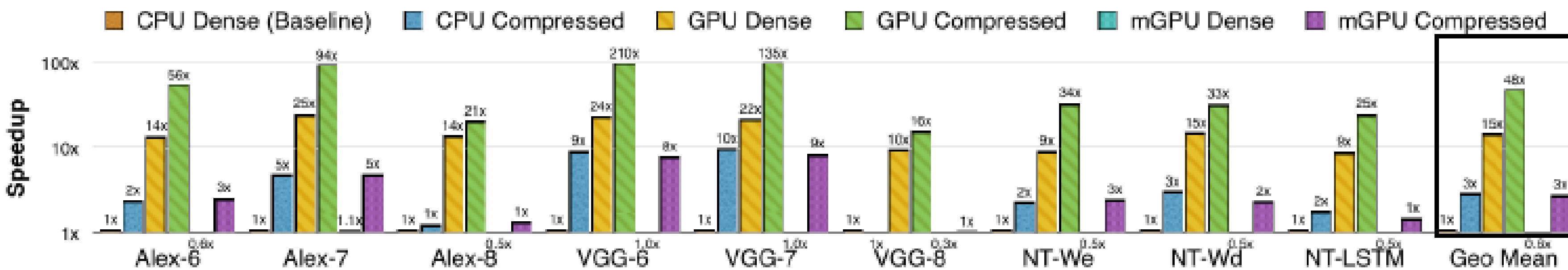
Technische Universiteit Eindhoven University of Technology

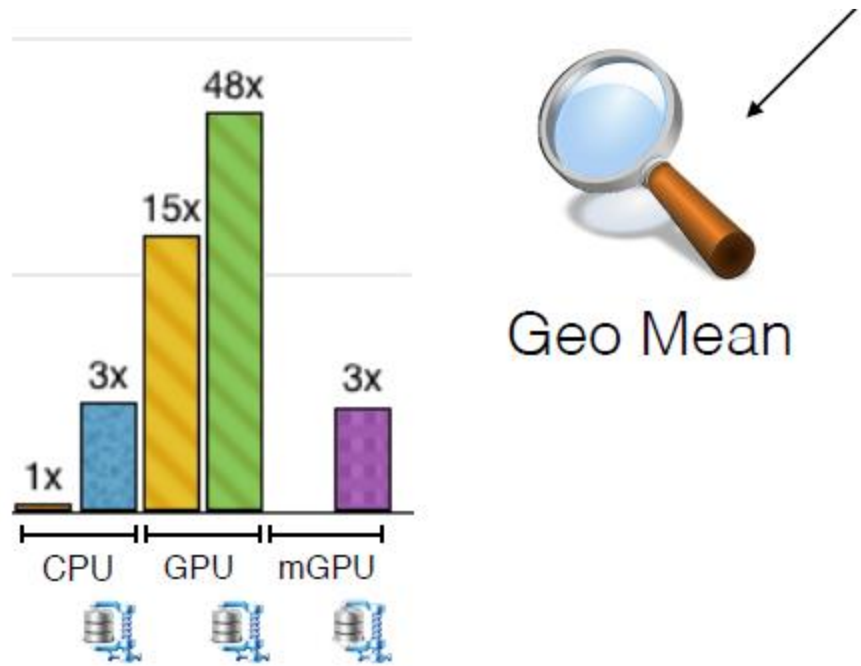
AlexNet Pruned

- **Similar performance with less parameters**



Speedup for Pruned FC layer

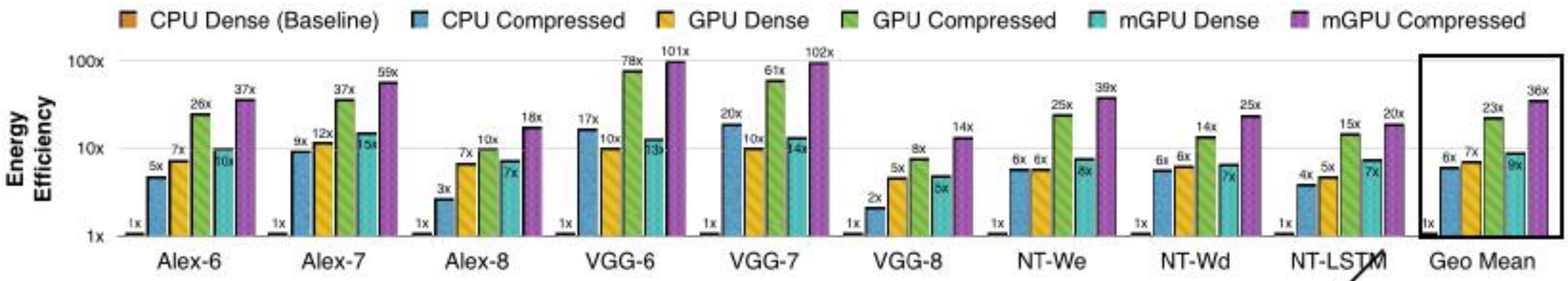




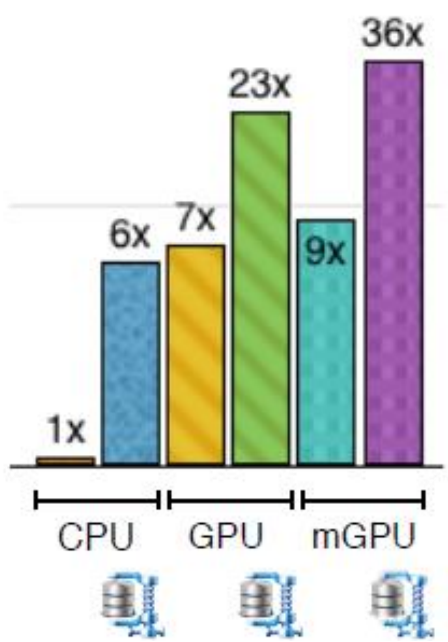
Baseline:

- Intel Core i7 5930K: MKL CBLAS GEMV, MKL SPBLAS CSRMV
- NVIDIA GeForce GTX Titan X: cuBLAS GEMV, cuSPARSE CSRMV
- NVIDIA Tegra K1: cuBLAS GEMV, cuSPARSE CSRMV

Energy Efficiency for Pruned FC Layer



Geo Mean

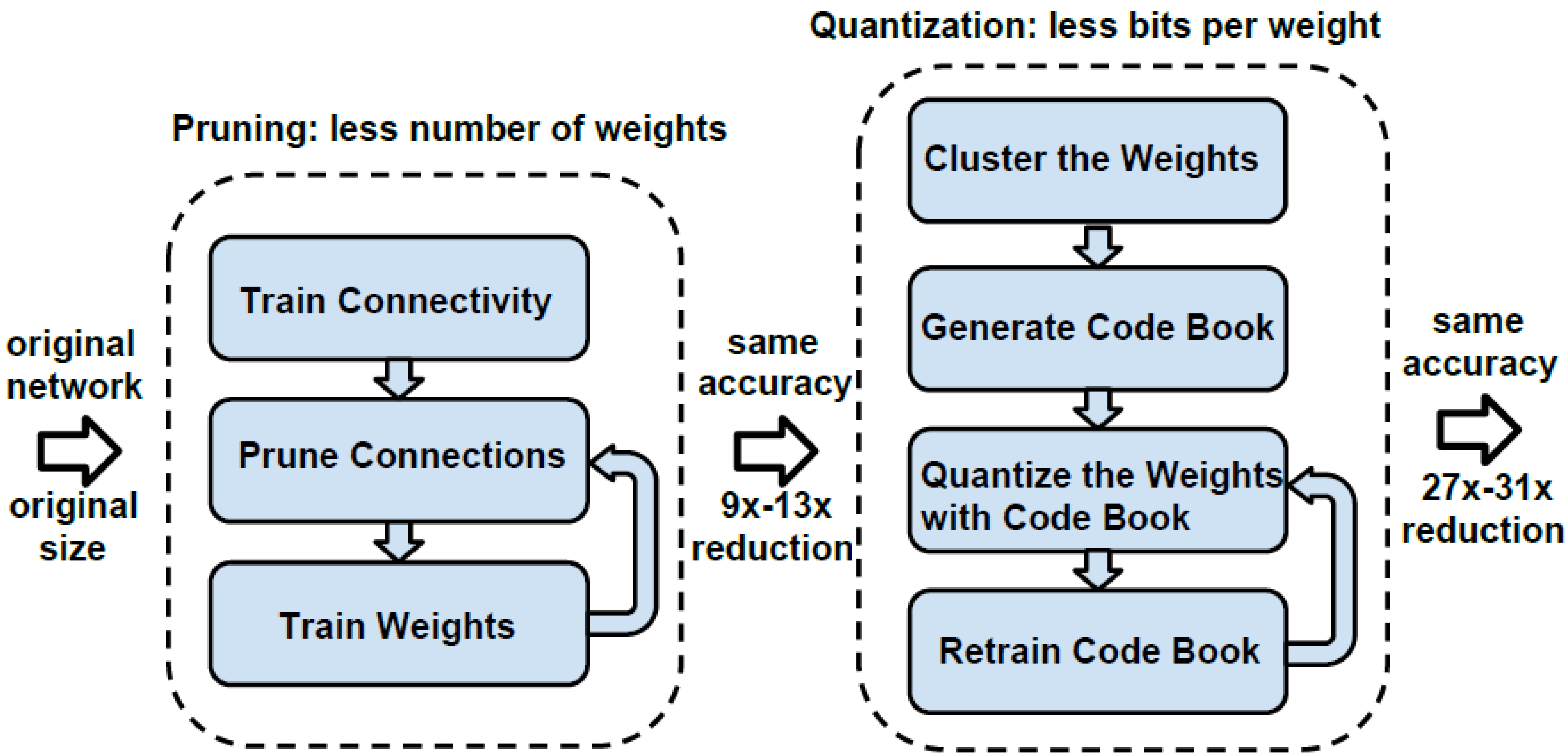


Baseline:

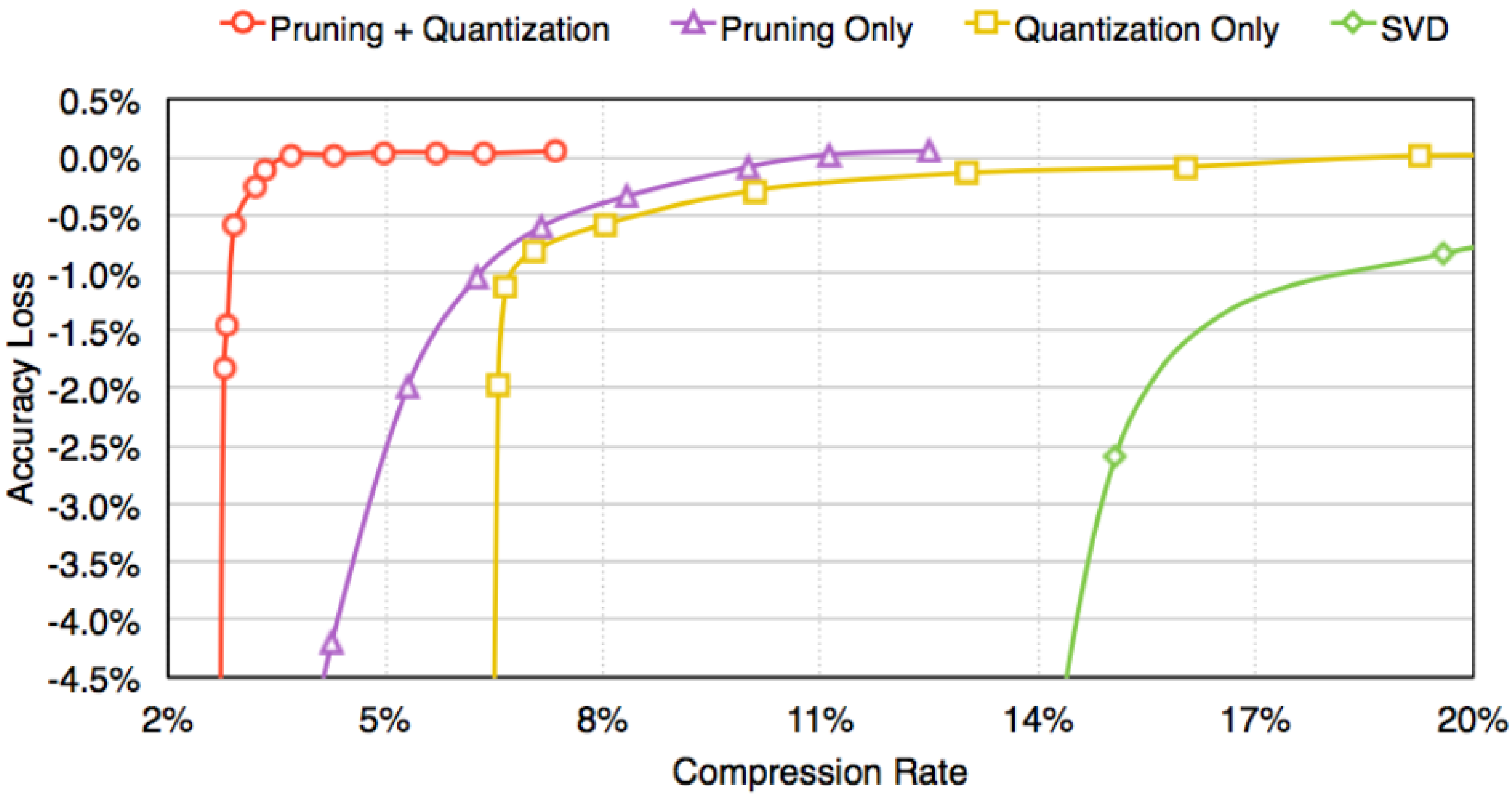
- Intel Core i7 5930K: MKL CBLAS GEMV, MKL SPBLAS CSRMV
- NVIDIA GeForce GTX Titan X: cuBLAS GEMV, cuSPARSE CSRMV
- NVIDIA Tegra K1: cuBLAS GEMV, cuSPARSE CSRMV

Deep Compression Approach

- **Pruning helps the reduction**
- **Advanced quantization reduces even more**



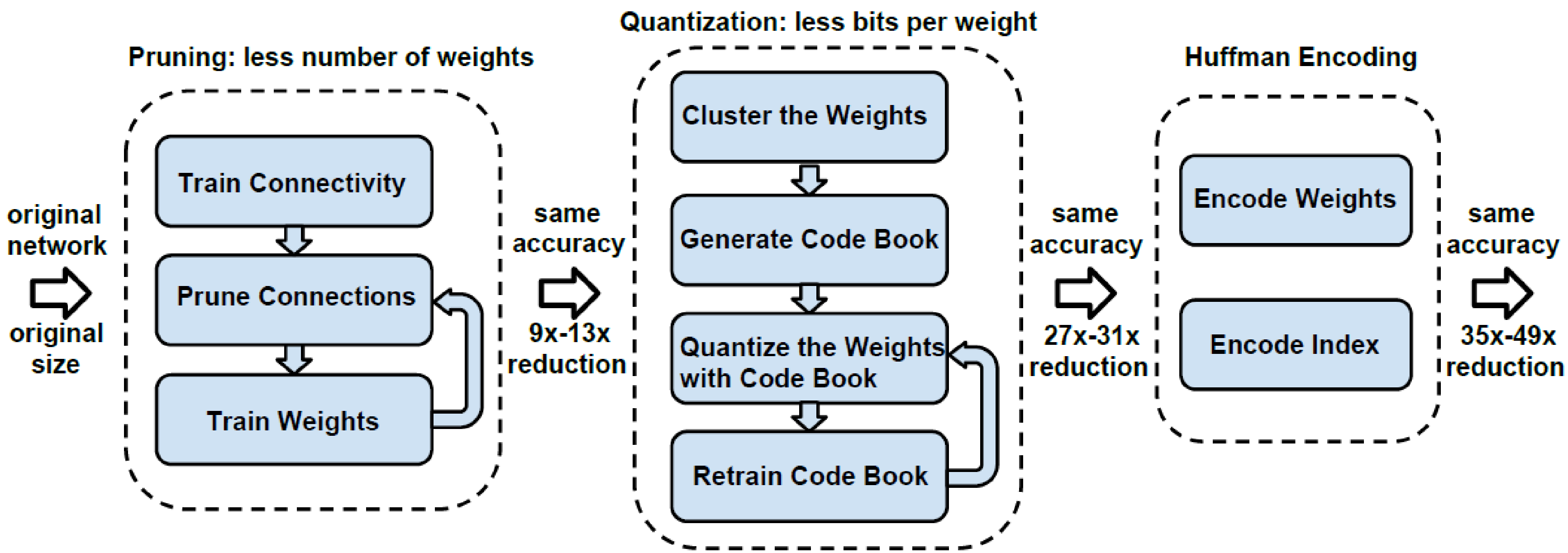
Evaluation of Deep Compression



Technische Universiteit Eindhoven University of Technology

Compression to the extreme

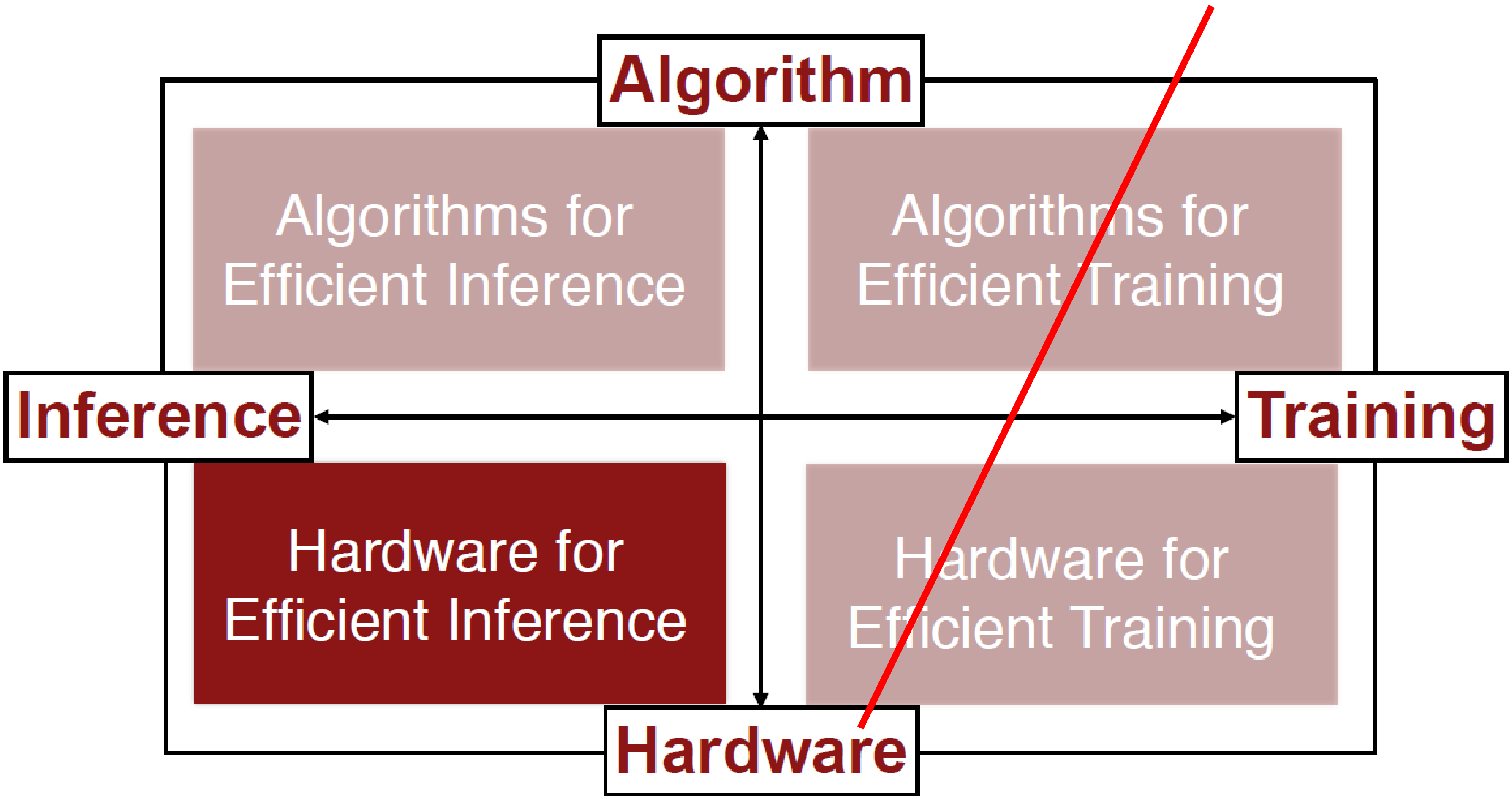
- **Add Huffman Encoding**



Deep Compression Results

Network	Original Size	Compressed Size	Compression Ratio	Original Accuracy	Compressed Accuracy
LeNet-300	1070KB	→ 27KB	**40x**	98.36%	→ 98.42%
LeNet-5	1720KB	→ 44KB	**39x**	99.20%	→ 99.26%
AlexNet	240MB	→ 6.9MB	**35x**	80.27%	→ 80.30%
VGGNet	550MB	→ 11.3MB	**49x**	88.68%	→ 89.09%
GoogleNet	28MB	→ 2.8MB	**10x**	88.90%	→ 88.92%
SqueezeNet	4.8MB	→ 0.47MB	**10x**	80.32%	→ 80.35%

Agenda


Algorithm
Algorithms for Efficient Inference
Algorithms for Efficient Training
Inference
Training
Hardware for Efficient Inference
Hardware for Efficient Training
Hardware

Diannao (Electric Brain)

- **Improved CNN computation efficiency by dedicated functional units + buffers optimized for CNN work**
- **Multiplier + adder tree + shifter + non-linear lookup**
- **Weights in off-chip DRAM**
- **452 GOP/s, 3.02 mm^2, 485 mW**

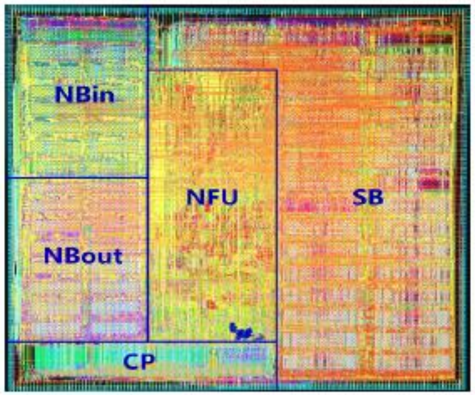


Figure 15. *Layout (65nm).*

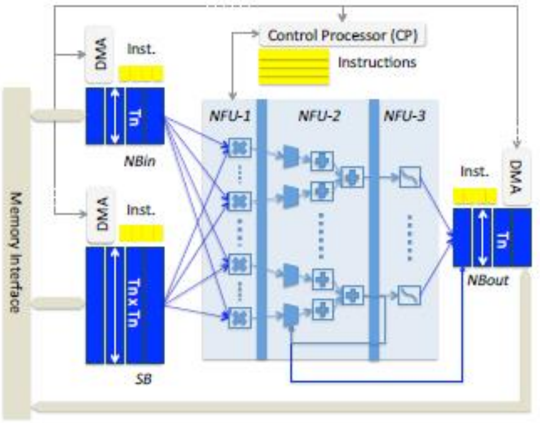


Figure 11. *Accelerator.*

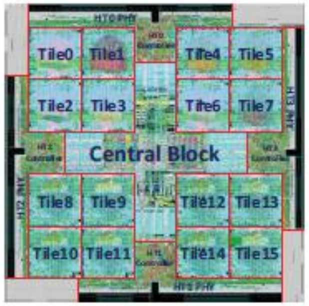
Component or Block	Area in μm^2	(%)	Power in mW	(%)	Critical path in ns
ACCELERATOR	3,023,077		485		1.02
Combinational	608,842	(20.14%)	89	(18.41%)	
Memory	1,158,000	(38.31%)	177	(36.59%)	
Registers	375,882	(12.43%)	86	(17.84%)	
Clock network	68,721	(2.27%)	132	(27.16%)	
Filler cell	811,632	(26.85%)			
SB	1,153,814	(38.17%)	105	(22.65%)	
NBin	427,992	(14.16%)	91	(19.76%)	
NBout	433,906	(14.35%)	92	(19.97%)	
NFU	846,563	(28.00%)	132	(27.22%)	
CP	141,809	(5.69%)	31	(6.39%)	
AXIMUX	9,767	(0.32%)	8	(2.65%)	
Other	9,226	(0.31%)	26	(5.36%)	

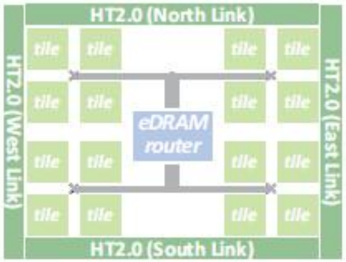
Table 6. *Characteristics of accelerator and breakdown by component type (first 5 lines), and functional block (last 7 lines).*

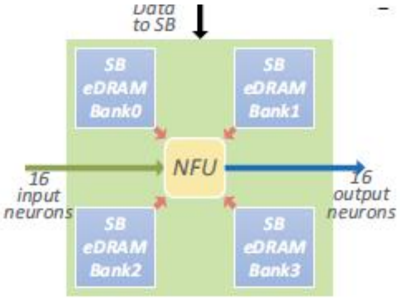
TU/e Technische Universiteit Eindhoven University of Technology

Diannao later variants

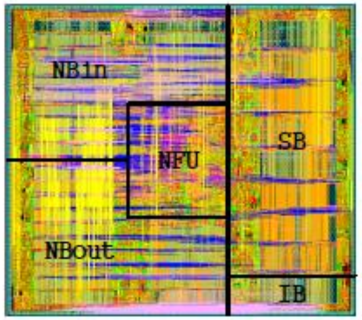
- DaDianNao (Bigger Computer)
  - Multi-processor and EDRAM to fit large models
  - $68mm^2$
  - 16 Watt
  - 12 M parameters

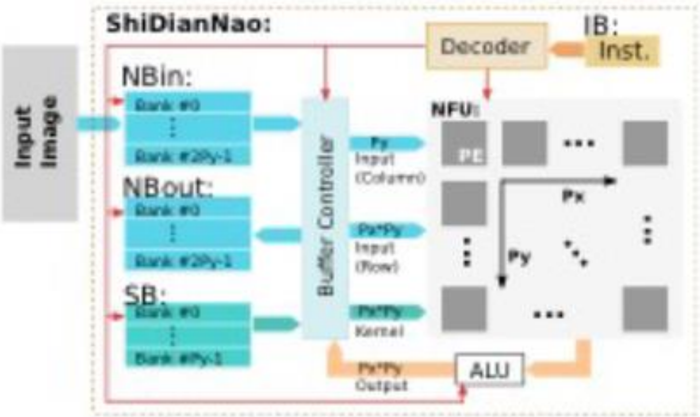






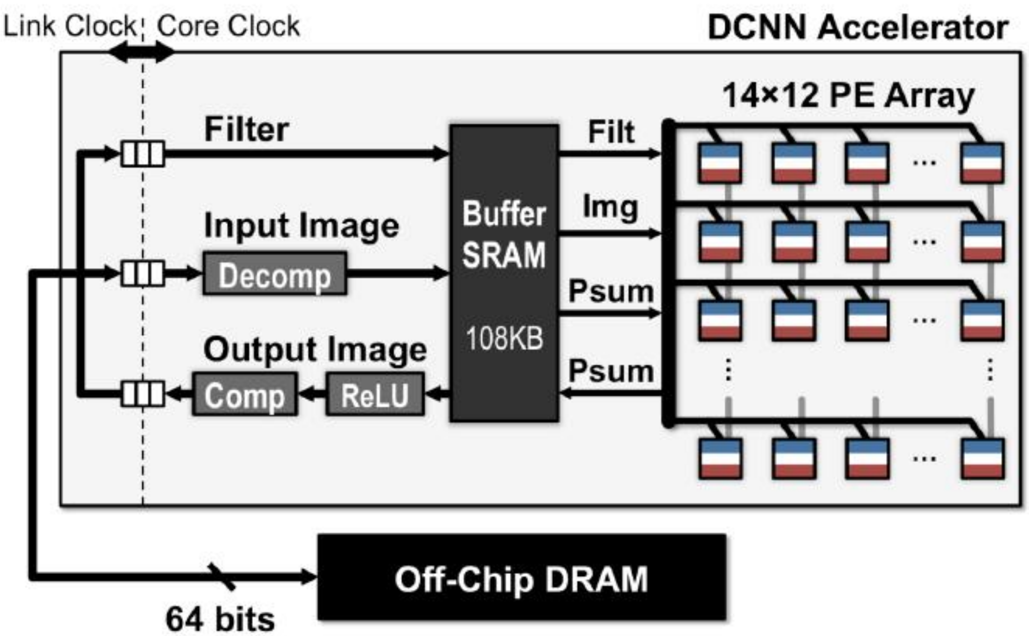
- ShiDiannao (Vision Computer)
  - 2D PE array
  - $4.86\ mm^2$
  - 320 mWatt
  - 64K parameters



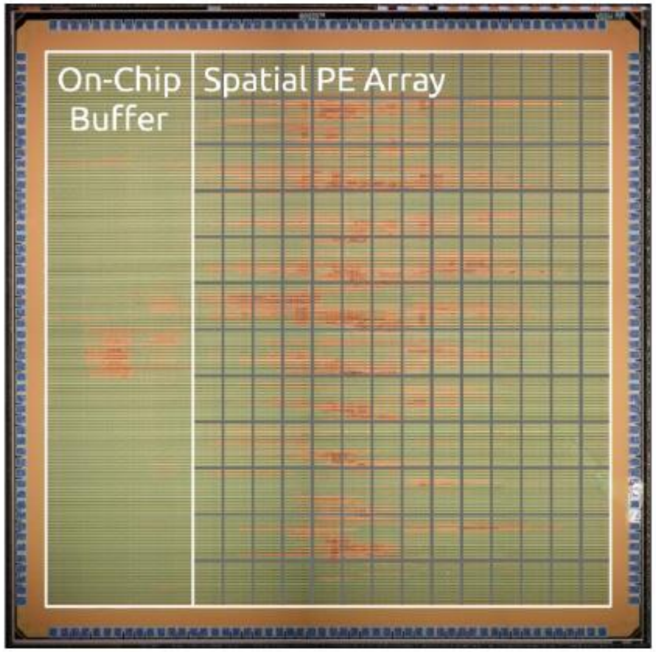


Technische Universiteit Eindhoven University of Technology

Eyeriss: Reduce Access by Row-Stationary Dataflow

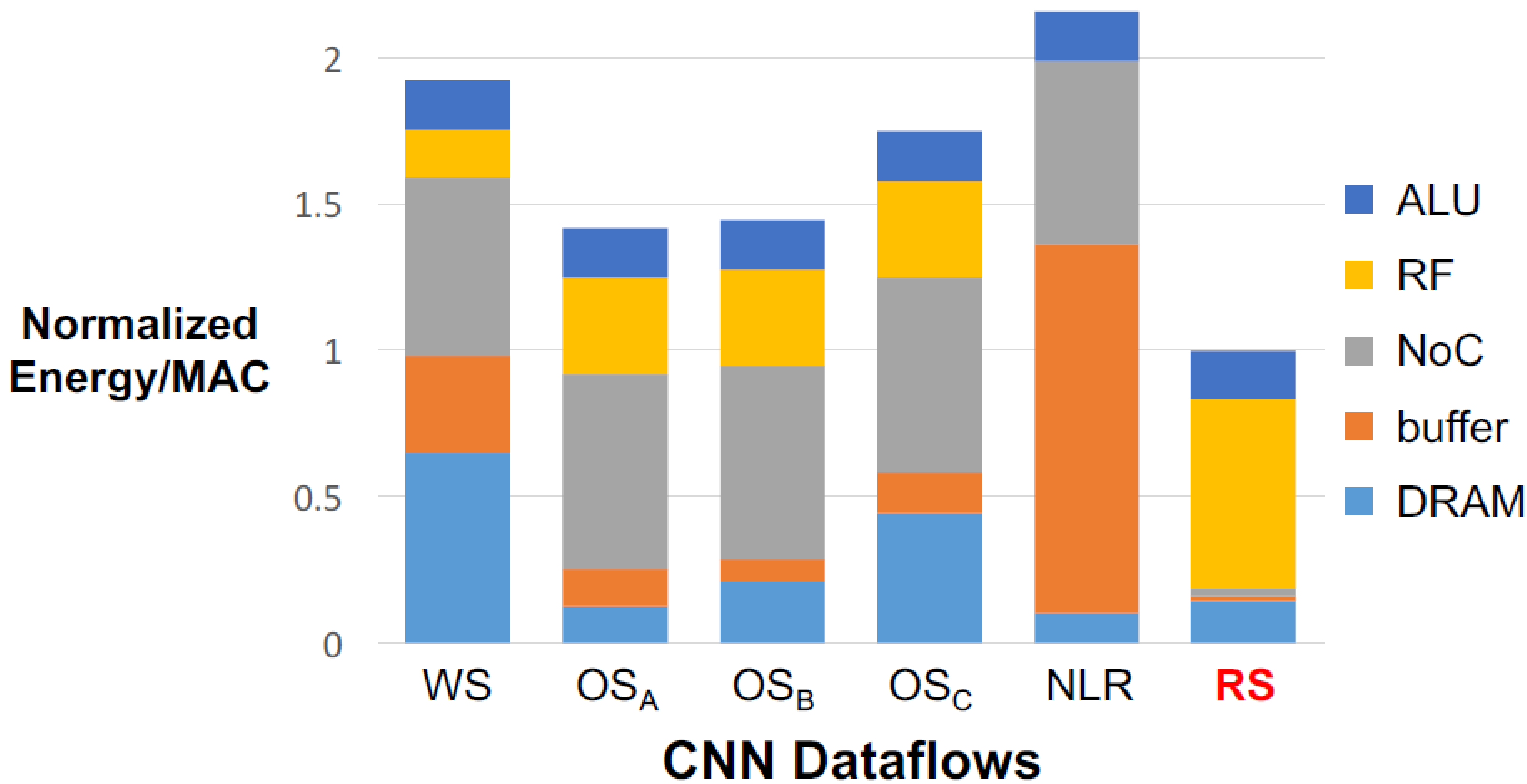


Eyeriss Architecture



Die Photo

Eyeriss: Reduce Memory Access by Row-Stationary Dataflow

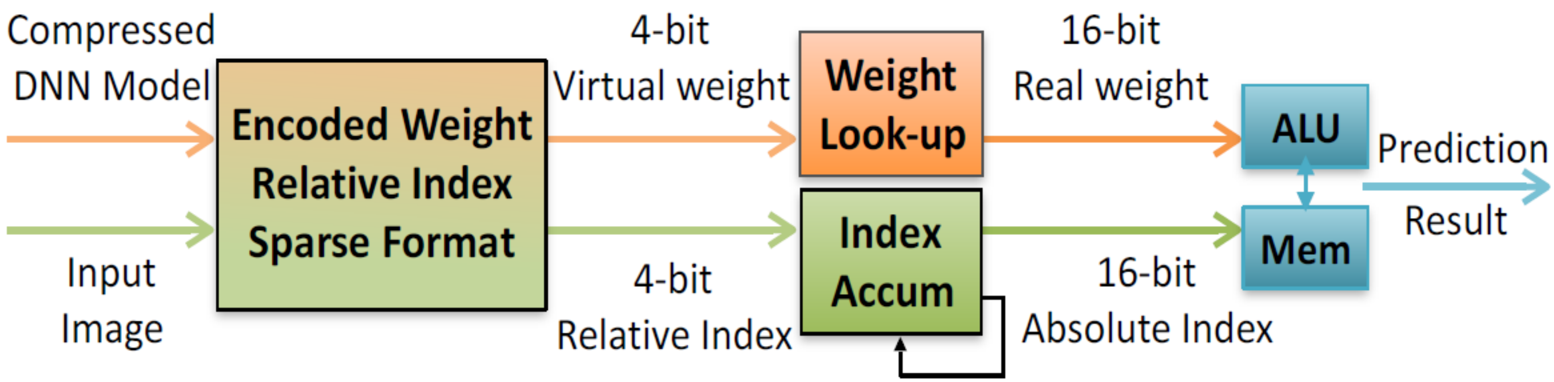


RS uses **1.4× – 2.5× lower** energy than other dataflows

Technische Universiteit **Eindhoven** University of Technology

EIE: Efficient Inference Engine

Weight decode



Address Accumulate

EIE: Efficient Inference Engine on Compressed Deep Neural Network

Song Han* Xingyu Liu* Huizi Mao* Jing Pu* Ardavan Pedram*
Mark A. Horowitz* William J. Dally*†
*Stanford University, †NVIDIA
{songhan,xyl,huizi,jingpu,perdavan,horowitz,dally}@stanford.edu

EIE PE architecture

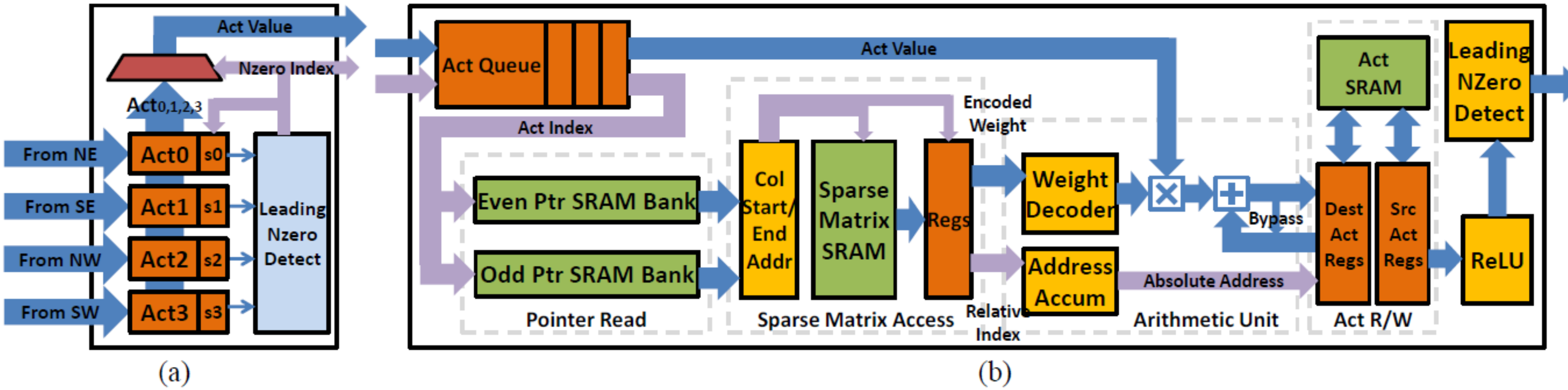
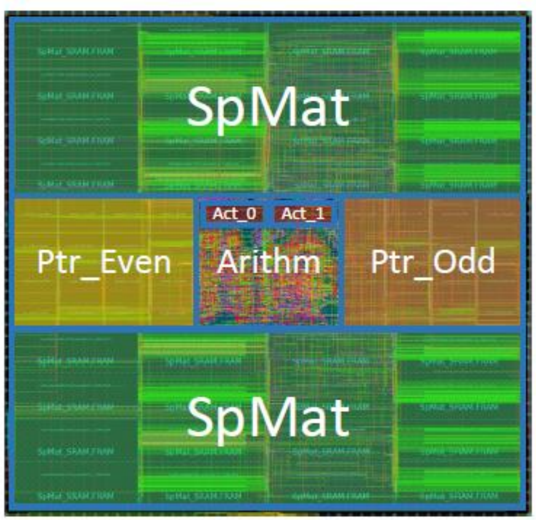


Figure 4. (a) The architecture of Leading Non-zero Detection Node. (b) The architecture of Processing Element.

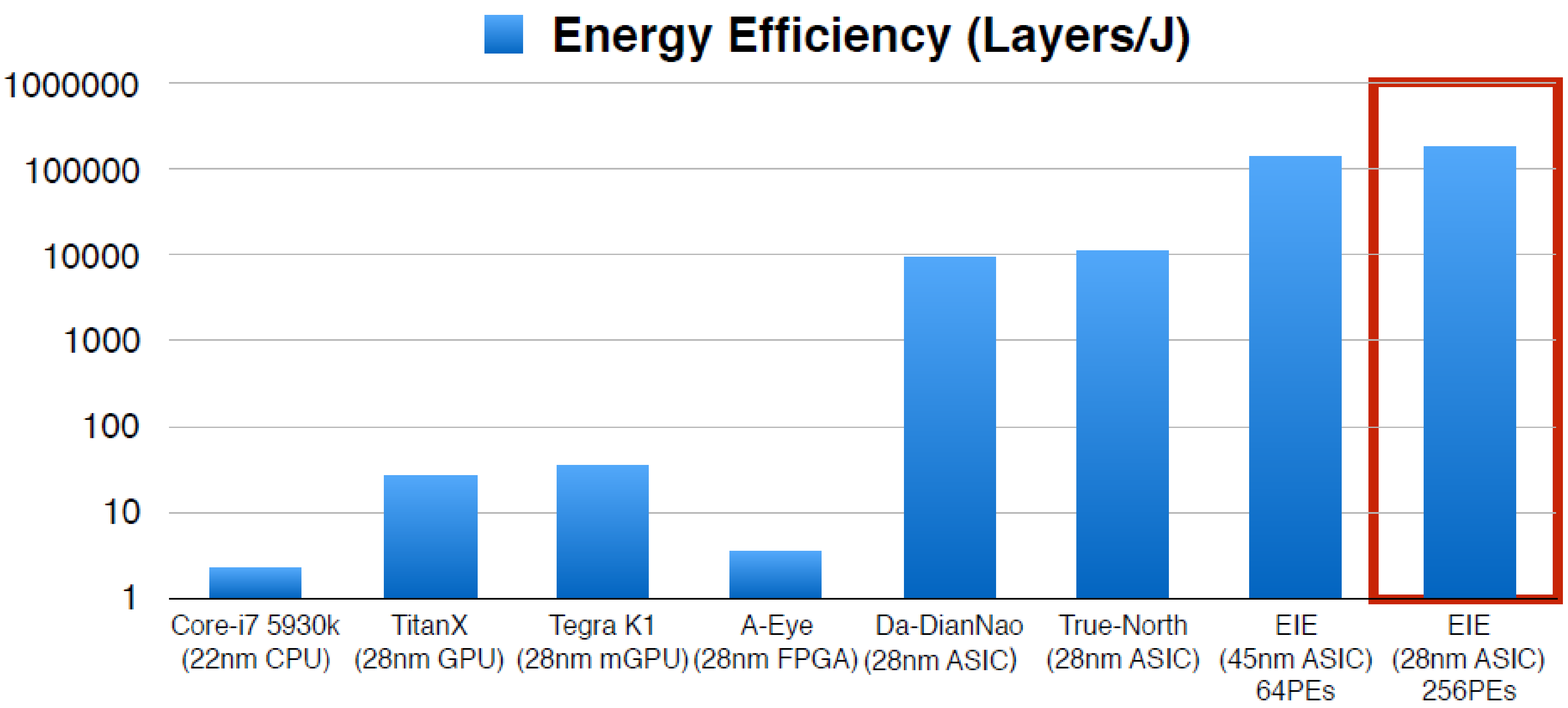


THE IMPLEMENTATION RESULTS OF ONE PE IN EIE AND THE BREAKDOWN BY COMPONENT TYPE (LINE 3-7), BY MODULE (LINE 8-13). THE CRITICAL PATH OF EIE IS 1.15 NS

	Power (mW)	(%)	Area (μm^2)	(%)
Total	9.157		638,024	
memory	5.416	(59.15%)	594,786	(93.22%)
clock network	1.874	(20.46%)	866	(0.14%)
register	1.026	(11.20%)	9,465	(1.48%)
combinational	0.841	(9.18%)	8,946	(1.40%)
filler cell			23,961	(3.76%)
Act_queue	0.112	(1.23%)	758	(0.12%)
PtrRead	1.807	(19.73%)	121,849	(19.10%)
SpmatRead	4.955	(54.11%)	469,412	(73.57%)
ArithmUnit	1.162	(12.68%)	3,110	(0.49%)
ActRW	1.122	(12.25%)	18,934	(2.97%)
filler cell			23,961	(3.76%)

Energy Efficiency Evaluation

- **>10x improvement over Da-DianNao by compression**



Future Intelligence on Mobile

Phones

Drones

Robots

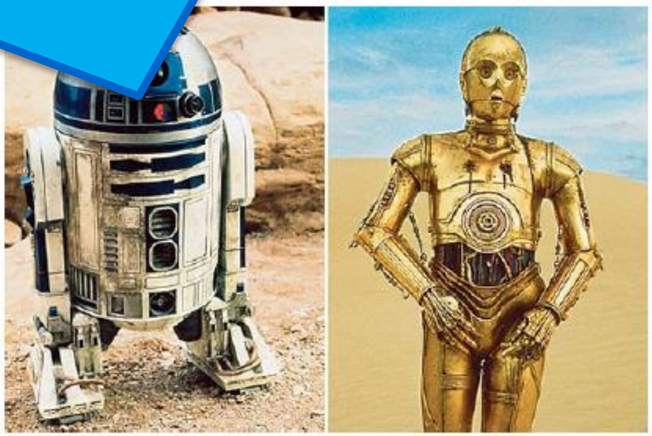
Glasses

Self Driving Cars

Limited Resource
Battery Constrained
Cooling Constrained

Outlook: the Path for Computation

Thank you for your attention



PC | **Mobile-First** | **AI-First**

Computation → Mobile Computation → Brain-Inspired Intelligent Computation