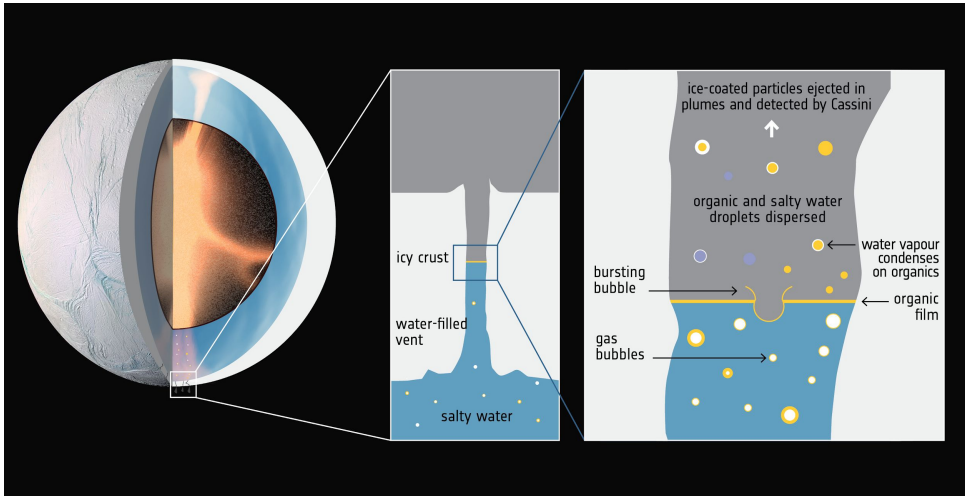
Jet Propulsion Laboratory
California Institute of Technology

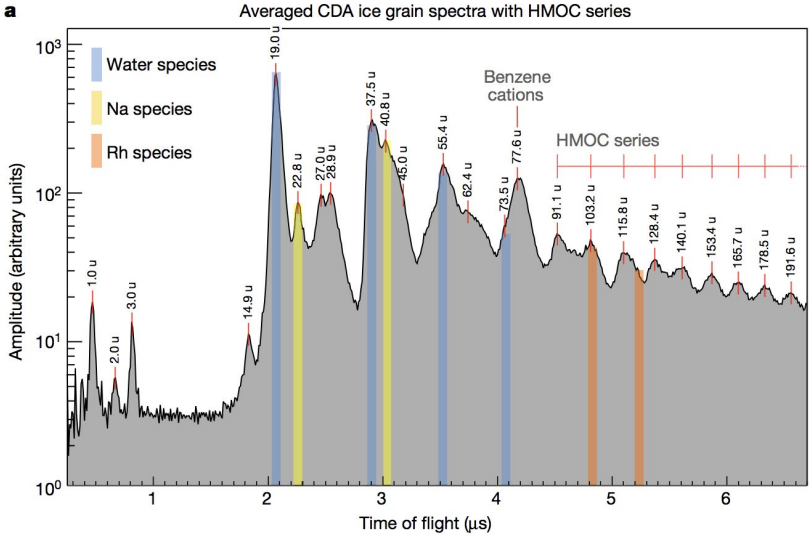
Accelerating autonomy in space using memory-centric architectures for deep neural networks

Kyongsik Yun
NASA Jet Propulsion Laboratory
California Institute of Technology
kyongsik.yun@jpl.nasa.gov
Space Computing Conference
July 31, 2019

© 2019 California Institute of Technology. Government sponsorship acknowledged.
All opinions expressed here are my own, and do not represent the opinion or policy of NASA, JPL, or Caltech.

Enceladus is the likest place to find alien life in the solar system: Astronomers found complex organic molecules, the building blocks of life, on the small icy moon.



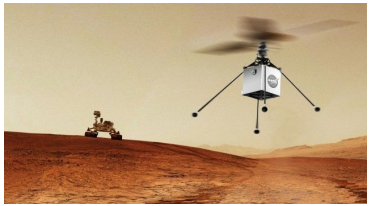


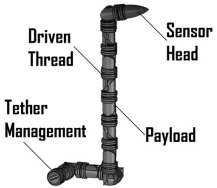
Life Investigation For Enceladus (LIFE)

Postberg et al., Nature (2018)
* High Mass Organic Cations

Jet Propulsion Laboratory
California Institute of Technology

Autonomy challenges - reactive, adaptive resilient autonomy is required for future mission



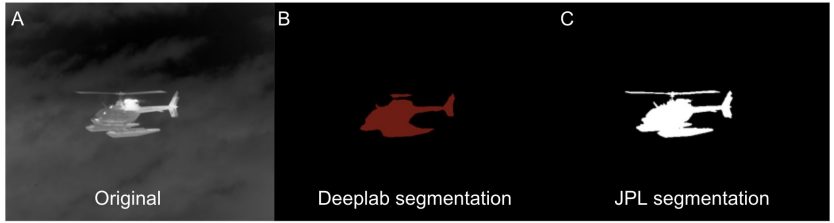


EELS: Exobiology Extant Life Surveyor Accessing Under Ice Habitable Oceans

M2020	Enceladus
2.98m/h (15km/1.25Mars years)	188m/h (3km/16 earth hours)
Deliberative Plan & schedule	Reactive Select activities in response to events/unknown situations
Robust Preprogrammed behaviors	Adaptive Change behaviors based on new situations
Protective Stop activities, go into safe mode, and wait for the ground	Resilient Keep going with remaining capabilities

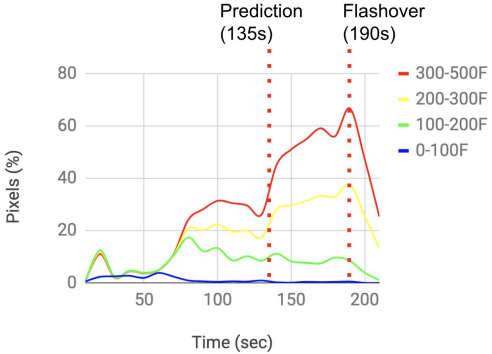
Jet Propulsion Laboratory
California Institute of Technology

Real-time high-precision object recognition and segmentation, and transformation for situational awareness using deep learning

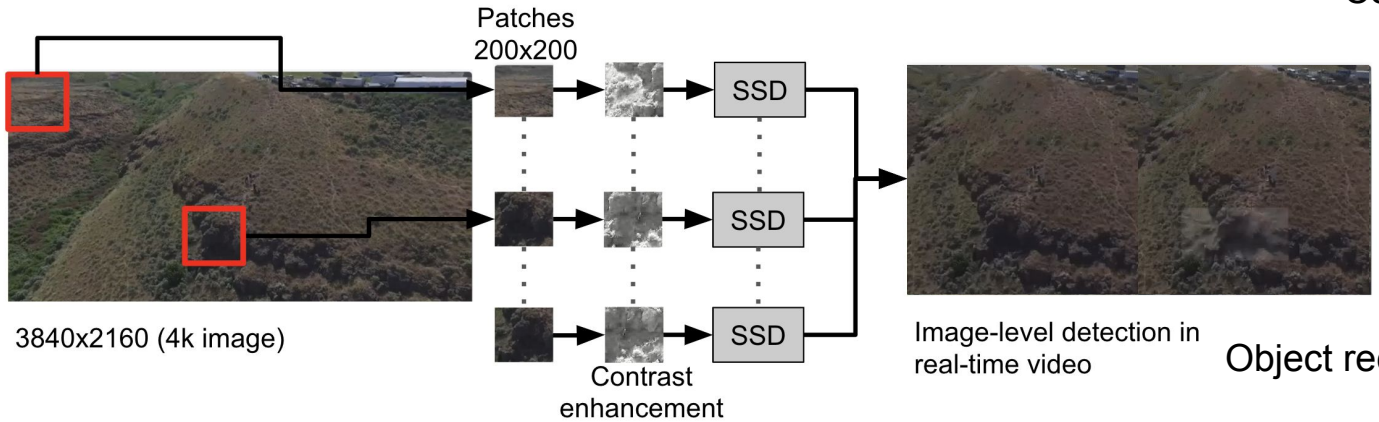


High-precision segmentation, Lu et al. 2018

Original video Enhanced video



Color-visible to IR transformation, Yun et al. 2018



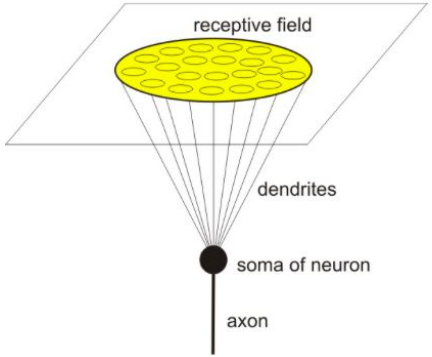
Object recognition for search and rescue mission, Yun et al. 2019

Convolutional neural networks (CNNs) as building blocks of deep learning are inspired by the human vision

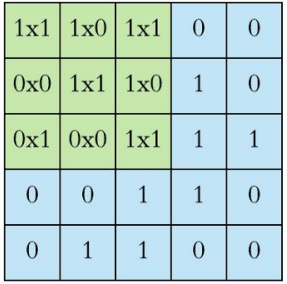
- CNNs are feed forward computational graphs constructed up to 1000s of layers
- Each layer consists of neurons n_i which are interconnected with synapses, associated with weights w_{ij}
- Each neuron computes

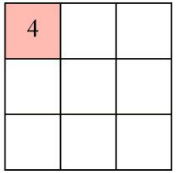
  Linear transform (dot-product of receptive field)

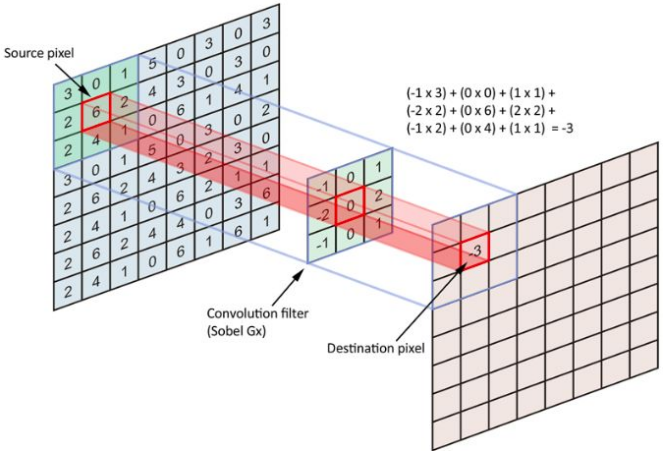
  Followed by a non-linear “activation” function



Convolution in the human brain



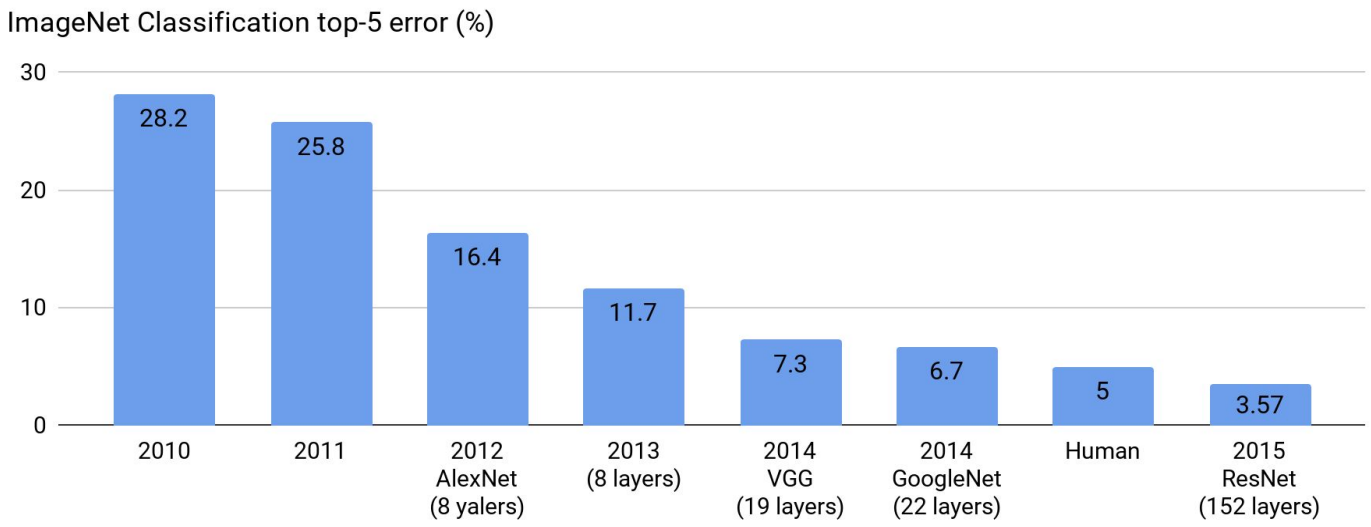




Convolution in the artificial neural networks

Why CNNs are better than other neural nets?

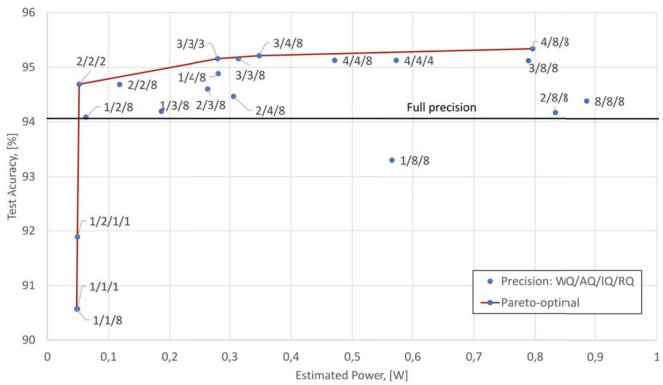
- Convolution takes advantage of local spatial coherence of data
- Pooling layer downscales the data, making CNNs more efficient
- Compared to expert system, CNNs require little or no domain expertise



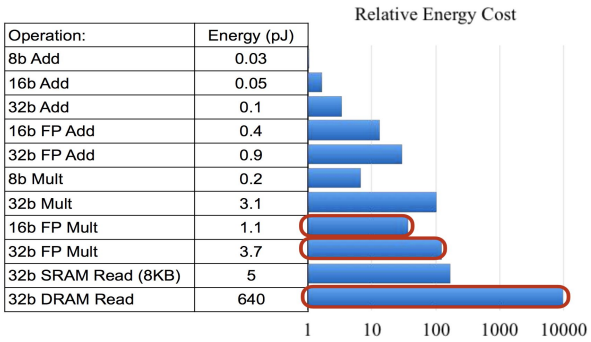
How can CNNs be better? Algorithmic optimization

- Reducing precision (most promising)
  - Reducing precision shrinks arithmetic cost
- Linear reduction in memory footprint (pruning)
  - Reduces weight fetching memory bandwidth
  - NN model may even stay on-chip

Precision	Modelsize (MB) ResNet50
1bit	3.2
8bit	25.5
32bit	102.5

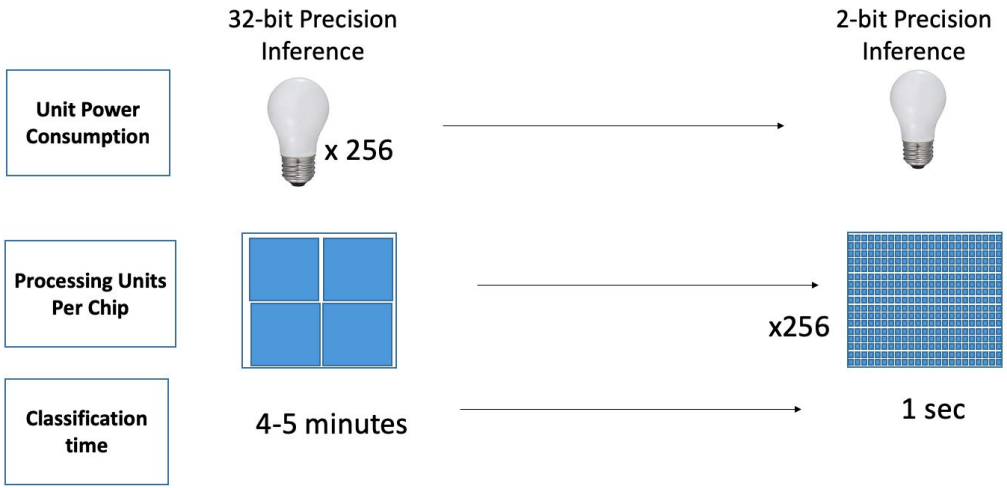


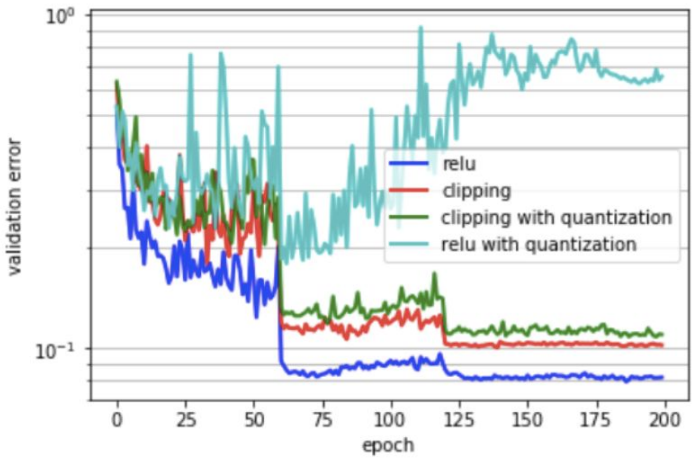
WQ: weight quantization, AQ: activation quantization
IQ: input quantization, RQ: recurrent activation quantization



Rybalkin, Pappalardo 2018
Horowitz ISSCC 2014

The benefits of moving from 32-bit to 2-bit precision for deep learning inference can be exploited for power consumption or performance



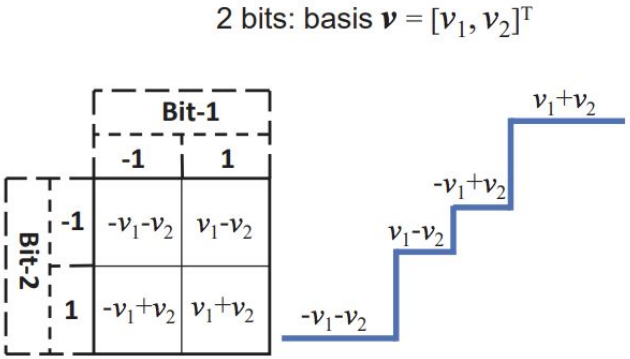


Power consumption and processing speed improve with quantization

Accuracy is compromised with quantization

IBM, NeurIPS 2019

Reduced precision neural nets are improving rapidly



Learnable quantizer on the 2-bit case

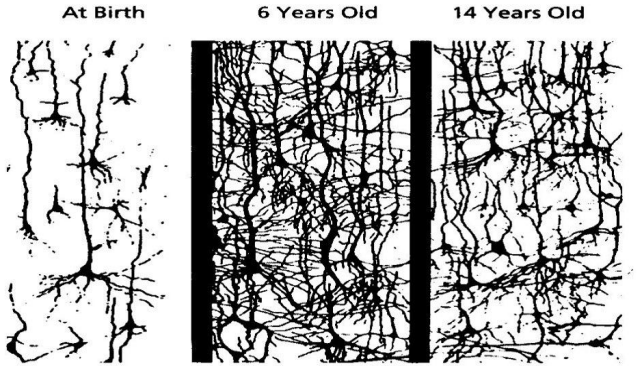
Methods	Bit-width (W/A)	Accuracy(%) Top-1	Top-5
ResNet-50			
FP* [15]	32/32	76.0	93.0
FP-ours	32/32	76.4	93.2
Ours	2/32	**75.1**	**92.3**
Ours	4/32	**76.4**	**93.1**
HWGQ [3]	1/2	64.6	85.9
ABC-Net [33]	5/5	70.1	89.7
Ours	1/2	**68.7**	**88.4**
Ours	2/2	**71.5**	**90.3**
Ours	3/3	**74.2**	**91.6**
Ours	4/4	**75.1**	**92.4**

Bit-width (W/A)	Training time
32/32	1.0×
2/32	1.4×
3/32	1.7×
1/2	2.1×
2/2	2.3×
3/3	3.7×

Learned Quantization (LQ)-Net, Zhang et al. 2018

How can CNNs be better? Pruning

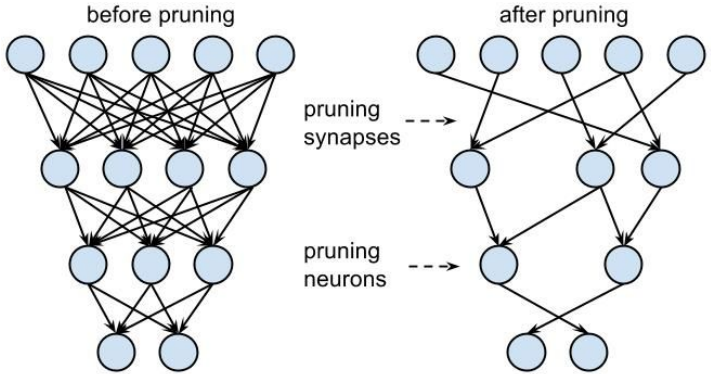
Pruning in human brain



Synaptic density (Chugani).
Children's brains have twice as many synapses as adults' brains

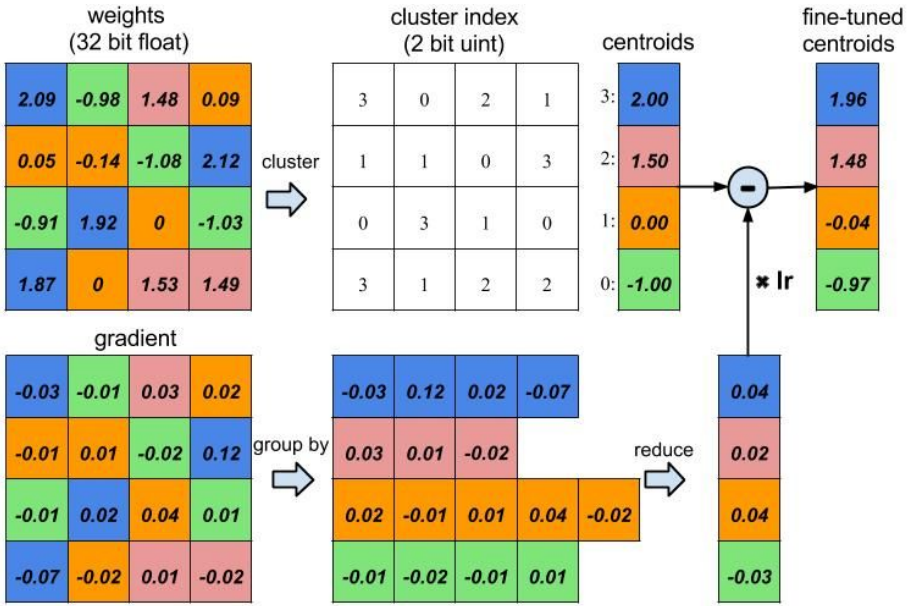
Pruning in artificial neural network with L1 regularization

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2 + \boxed{\lambda \sum_{i=1}^{k} |w_i|}$$



Han 2016

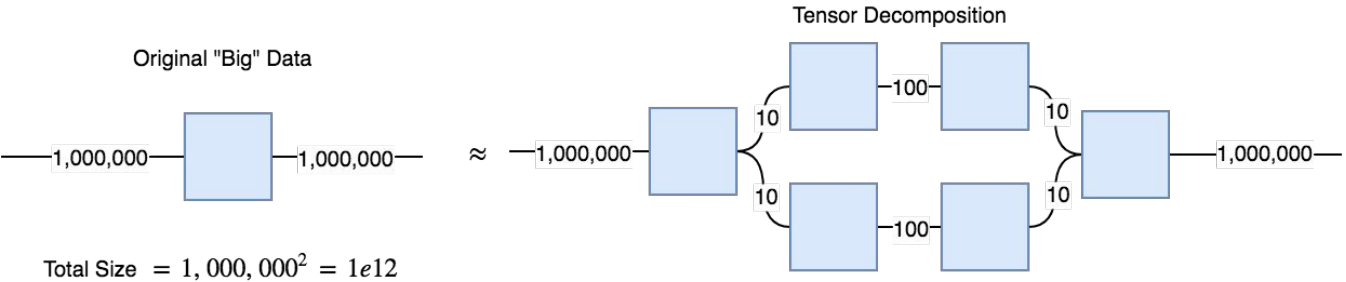
Quantization and pruning without accuracy loss

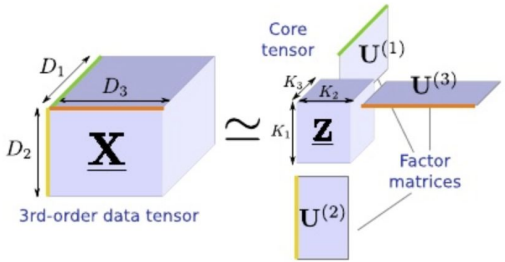


Network	Original Size	Compressed Size	Compression Ratio	Original Accuracy	Compressed Accuracy
Lenet-300-100	1070KB	27KB	**40x**	98.36%	98.42%
Lenet-5	1720KB	44Kb	**39x**	99.20%	99.26%
AlexNet	240MB	6.9MB	**35x**	80.27%	80.30%
VGGNet	550MB	11.3MB	**49x**	88.68%	89.09%
GoogleNet	28MB	2.8MB	**10x**	88.90%	88.92%
SqueezeNet	4.8MB	0.47MB	**10x**	80.32%	80.35%

Han et al. 2016 (Xilinx)

Tensor decomposition for data compression in deep learning



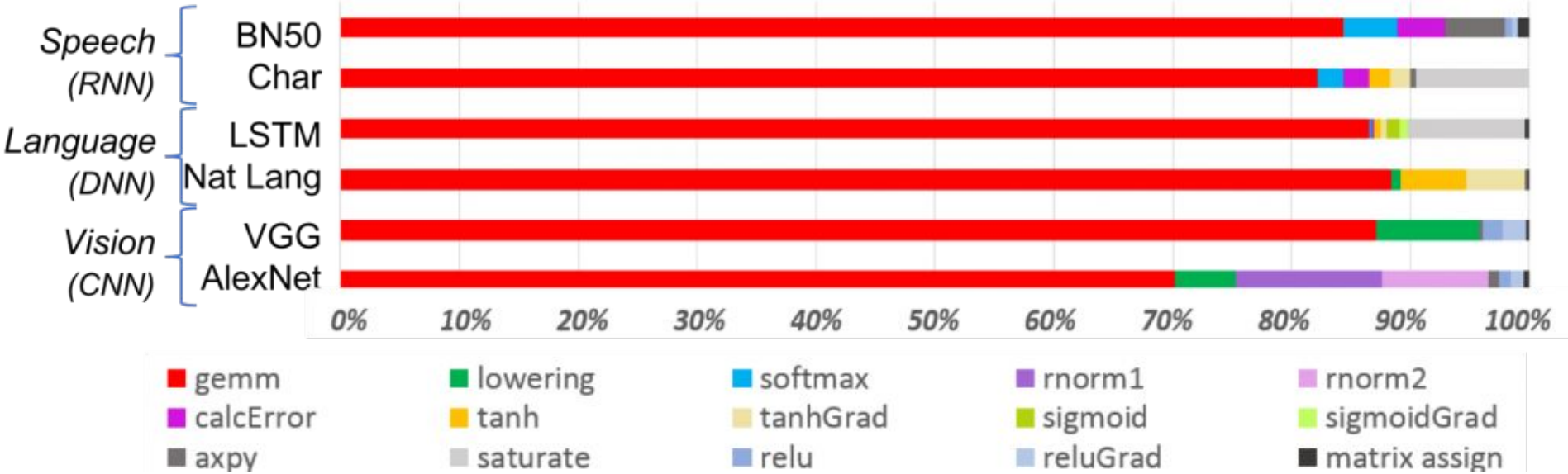


$$x_{ijk} = \sum_{q=1}^{K_1} \sum_{r=1}^{K_2} \sum_{s=1}^{K_3} u_{iq}^{(1)} u_{jr}^{(2)} u_{ks}^{(3)} z_{qrs} + \varepsilon_{ijk}$$

Model	Top-5	Weights	FLOPs	S6		Titan X
AlexNet	80.03	61M	725M	117ms	245mJ	0.54ms
*AlexNet**	78.33	11M	272M	43ms	72mJ	0.30ms
(imp.)	(-1.70)	(×5.46)	(×2.67)	(×2.72)	(×3.41)	(×1.81)
VGG-S	84.60	103M	2640M	357ms	825mJ	1.86ms
*VGG-S**	84.05	14M	549M	97ms	193mJ	0.92ms
(imp.)	(-0.55)	(×7.40)	(×4.80)	(×3.68)	(×4.26)	(×2.01)
GoogLeNet	88.90	6.9M	1566M	273ms	473mJ	1.83ms
*GoogLeNet**	88.66	4.7M	760M	192ms	296mJ	1.48ms
(imp.)	(-0.24)	(×1.28)	(×2.06)	(×1.42)	(×1.60)	(×1.23)
VGG-16	89.90	138M	15484M	1926ms	4757mJ	10.67ms
*VGG-16**	89.40	127M	3139M	576ms	1346mJ	4.58ms
(imp.)	(-0.50)	(×1.09)	(×4.93)	(×3.34)	(×3.53)	(×2.33)

Kim et al. 2016 (Samsung)

Deep learning algorithms are comprised of a spectrum of operations. Although matrix multiplication is dominant, optimizing performance efficiency while maintaining accuracy requires the core architecture to efficiently support all of the auxiliary functions. (***algorithmic improvement direction!***)



Fleischer et al. 2018

Energy per operation, in PJ, for various operations in 45nm
Memory access is >500X arithmetic energy

Operation	16bit (integer)		64 bit (DP-FP)	
	E/op PJ	vs. Add	E/op PJ	vs. Add
ADD	0.18	1.0X	5	1.0X
Multiply	0.62	3.4X	20	4.0X
16-word register file	0.12	0.7X	0.34	0.07X
64-word register file	0.23	1.3X	0.42	0.08X
4K-word SRAM	8	44X	26	5.2X
32K-word SRAM	11	61X	47	9.4X
DRAM	640	3556X	2560	512X

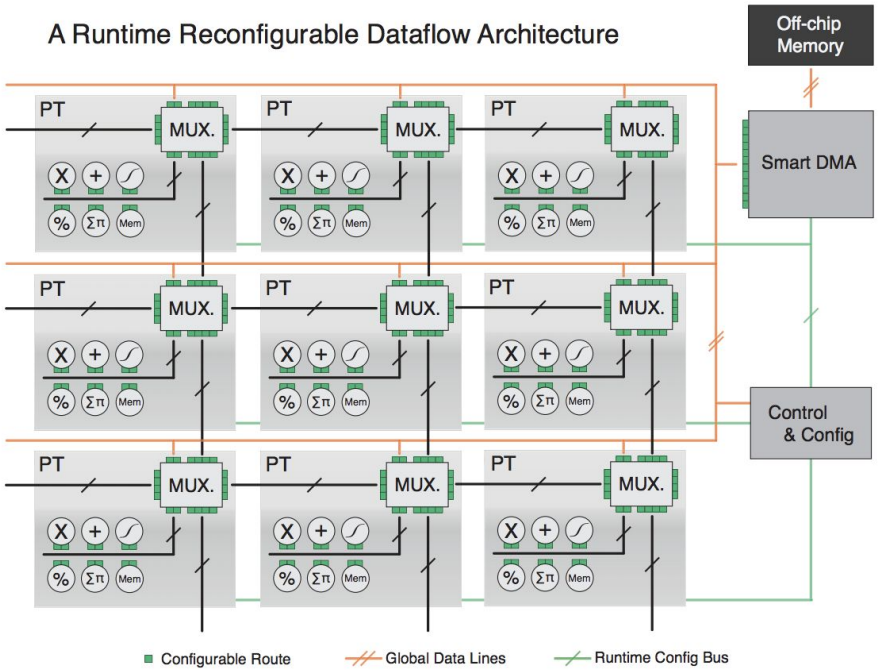
Pedram et al. 2017

Performance comparisons of VGG-CNNs on FPGAs
Hardware improvement for deep learning from CPU to GPU to FPGA

VGG								
Model	Platform	Specification			Precision	GOP/s	GOP/j	Power
		Types	Frequnce	Mermory				
VGG-19 [16]	CPU	Xeon E5-2650v2	2.6GHz	-	float32	119	0.63	95W
VGG-19 [16]	GPU	GTX TITAN X	1002MHz	12G GDDR5	float32	1704	6.82	250W
VGG-16 [23]	FPGA	Stratix-V GSD8	120MHz	32G DDR3	fixed8	117.8	6.17	19.1W
VGG-16 [35]	FPGA	Stratix-V GSD8	200MHz	on-chip	fixed16	821	-	-
VGG-16 [34]	FPGA	Arria 10 SX660	120MHz	- DDR4	8-bit	53	13.9	3.3W
VGG-16 [25]	FPGA	Arria 10 GX 1150	150MHz	8G DDR3L	fixed8/16	645.25	-	-
VGG-16 [38]	FPGA	Arria 10 GX 1150	240MHz	- DDR3	fixed8/16	968.03	-	-
VGG-16 [38]	FPGA	Stratix 10 GX 2800	300MHz	- DDR3	fixed8/16	1604.57	-	-
VGG [31]	FPGA	Arria 10 GX 1150	370MHz	1G DDR4	float-	866	20.75	19.1W
VGG [31]	FPGA	Arria 10 GX 1150	385MHz	1G DDR4	fixed16	1790	47.78	-
VGG-S [36]	FPGA	XCKU115	125MHz	off-chip	fixed32	394.7	14.6	27W
VGG-D [33]	FPGA	Virtex 7 VX690T	200MHz	off-chip	fixed8	1467.6	-	-
VGG-A [33]					fixed8	1500	-	
VGG-16 [37]	15xFPGAs	XC7VX690T	-	off-chip	fixed 16	1197*	37.88	-
VGG-19 [37]					fixed16	1220*	38.13	
VGG-19 [16]	FP-DNN	Stratix-V GSMD5	150MHz	4G DDR3	float32	81	3.24	25W
					float16	364.36	14.57	

Wang et al. 2018

Convolutional neural networks on the edge:
Goal is to minimize memory access (***hardware improvement direction!***)



Xilinx Virtex 6 FPGA platform

Multiplication

Addition

Linear mapper

∑π: parallel convolver with 100 MAC units

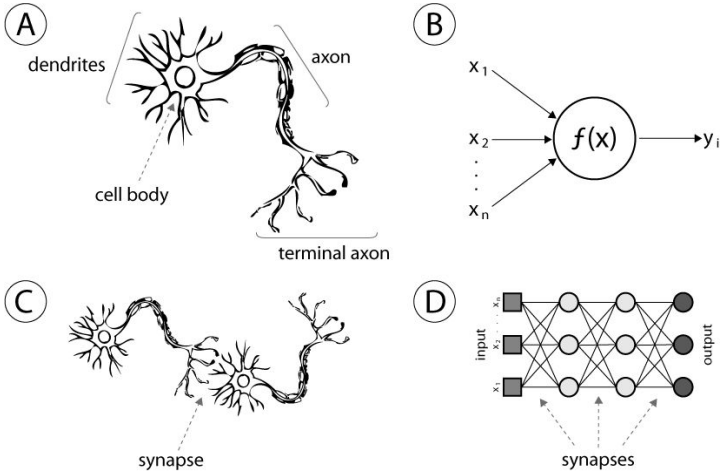
	Intel i7 4cores	NVIDIA GT335m	neuFlow Virtex 6	NVIDIA GTX480	neuFlow IBM 45nm
Peak GOP/sec	40	182	160	1350	1280
Actual GOP/sec	12	54	147	294	1164
FPS	14	67	182	374	1480
Power (W)	50	30	10	250	5
GOP/s/W	0.24	1.8	14.7	1.176	232.8
Precision	Float32	Float32	Fixed16	Float32	Fixed16

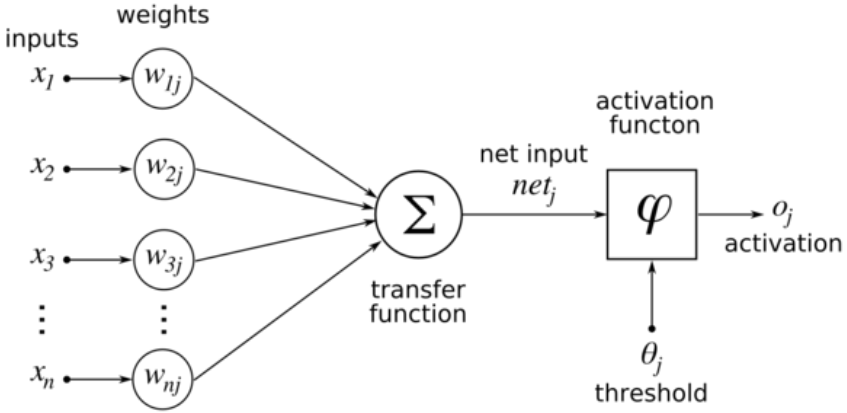
NeuFlow [Farabet et al. 2011]

Motivation of in-memory computing: Neural information processing

"The software written for the brain" executes on billion BPUs (brain processing units=neurons), each connected to several thousand other BPUs

There is no flow control in that the information flows from a BPU to another as needed. No circuitry is wasted in fetching, caching, decoding, and executing code





Near memory computing by incorporating the memory and logic chips in an IC to solve the bandwidth bottleneck

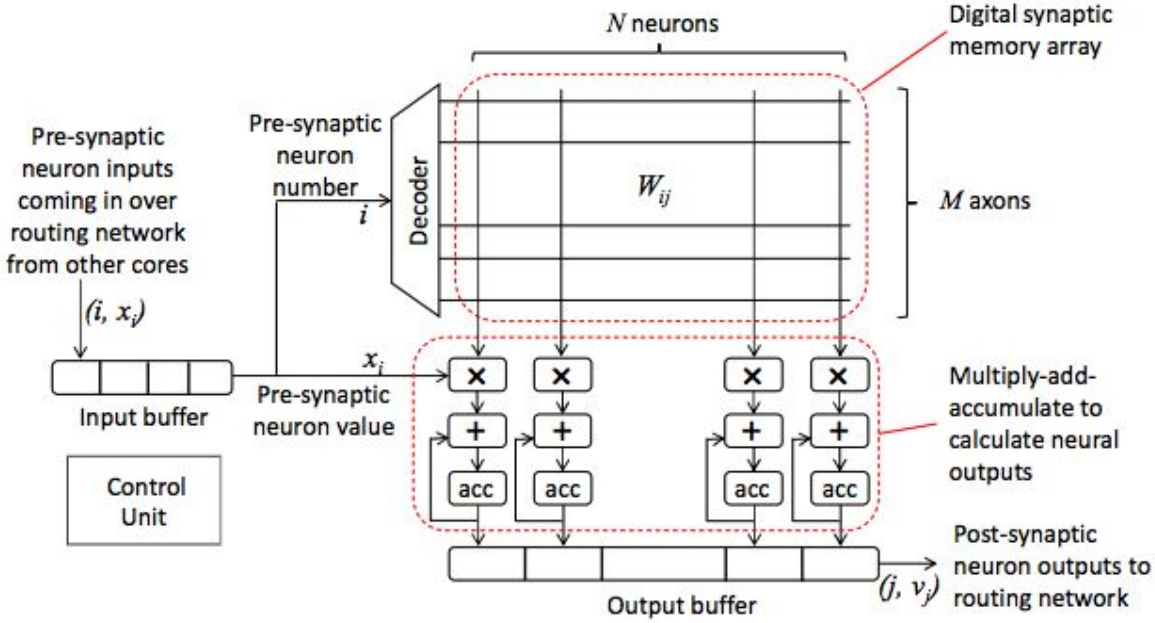


Fig. 1. Proposed neural core architecture.

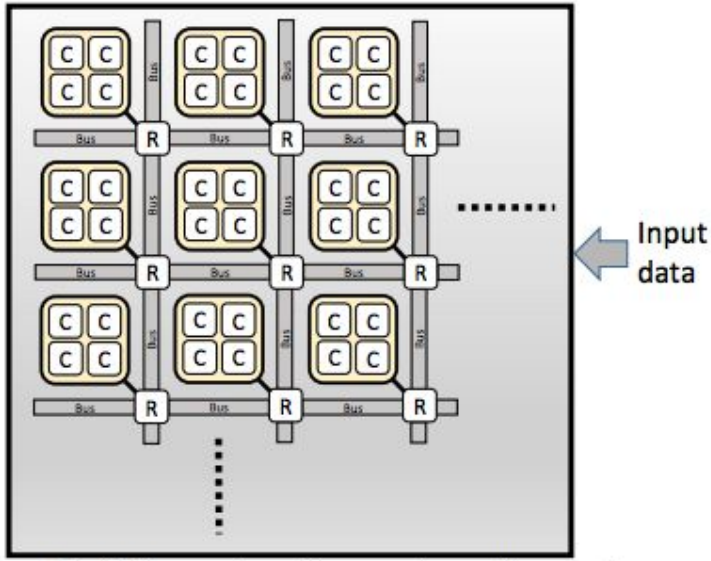
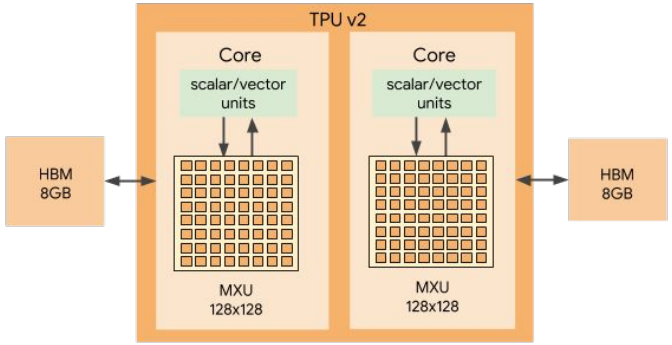
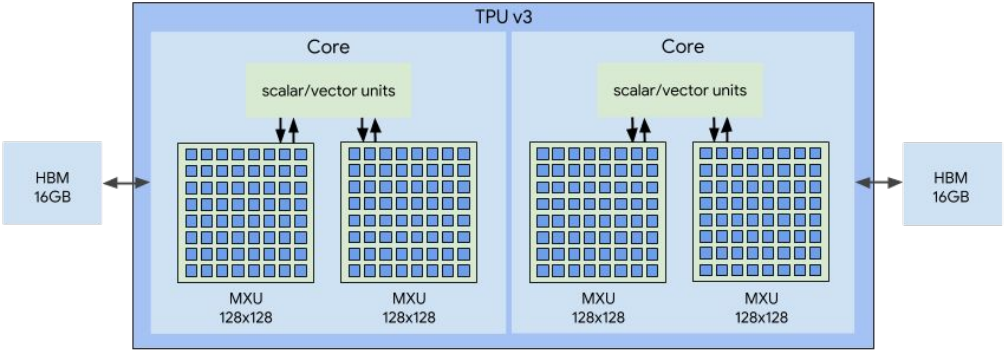


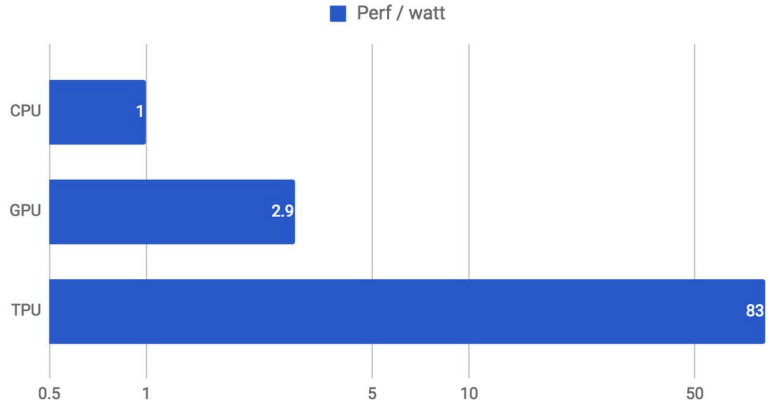
Fig. 2. Proposed multicore system with several neural cores (C) sharing a routing resource (R).

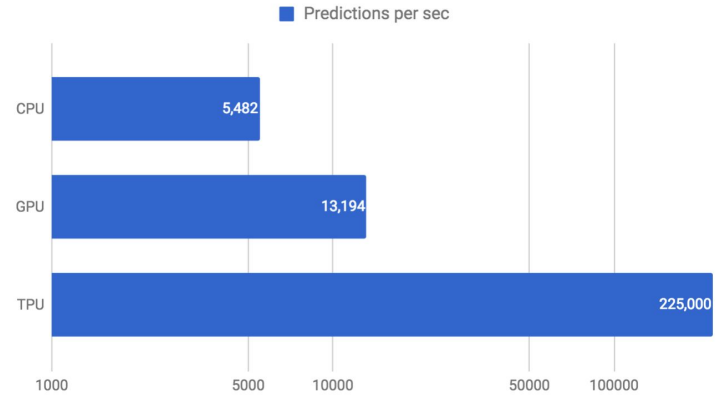
Taha et al. 2013

Performance improvement using near memory computing - TPU



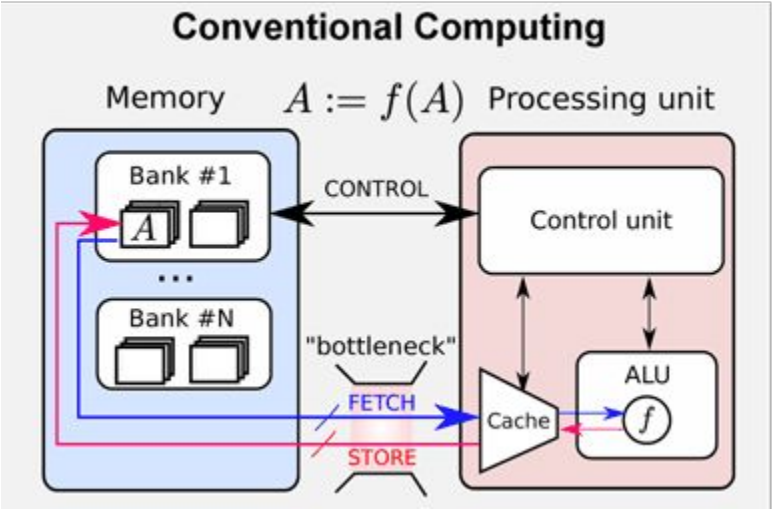


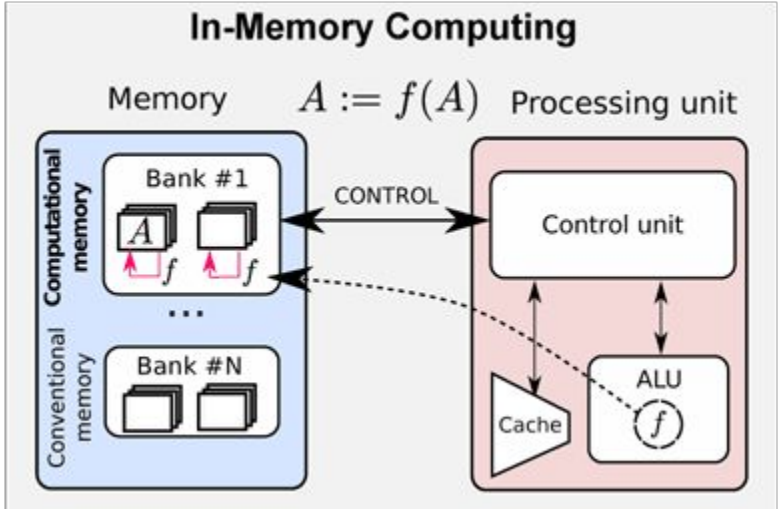




https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

Computational memory (in-memory computing):
Computation is performed in place by exploiting the physical attributes of memory devices organised as a “computational memory” unit.

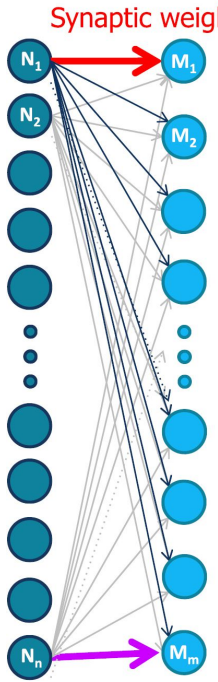


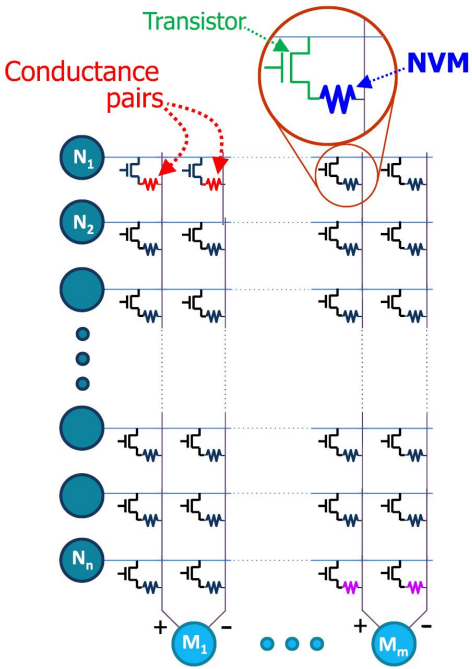


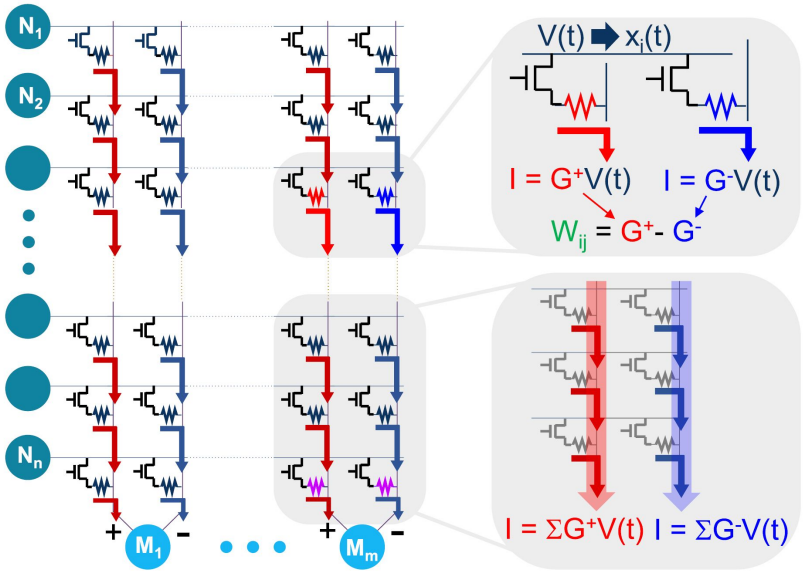
A. Sebastian et al.: “Temporal correlation detection using computational phase-change memory”, Nature Communications 8, 1115, 2017.

In memory computing for deep learning

Store the weights locally and use memory to do vector-matrix multiplication



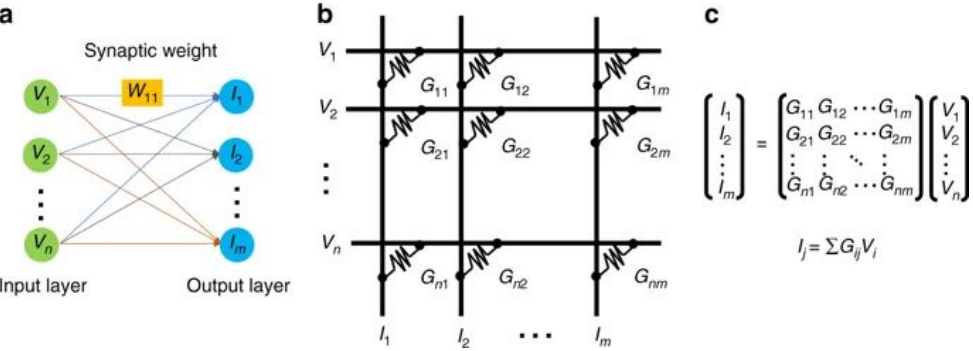


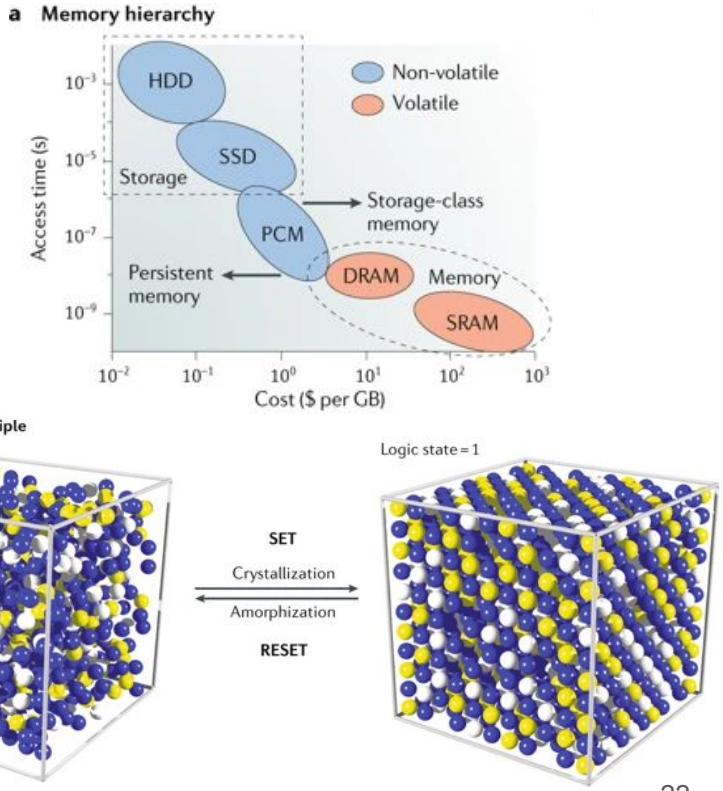


Instant/parallel vector matrix multiplications

Tsai et al. 2018 (IBM)

Phase-change memory for neuro-inspired computing





Neuromorphic computing system based on NVM. a Schematic illustration of one-layer neural network with synaptic weights (*W*) connecting an input layer to an output layer. b A synaptic weight is represented by a conductance value of an NVM element at each cross-point in a crossbar array structure. c Vector-matrix multiplication is performed by sensing the current (*I*) for each column, which is the product of the synaptic weight (*G*) and the input signal (*V*)

Gong et al. Nature Communications 2018

Zhang et al. Nature Reviews Materials 2019

IBM's mixed-precision approach

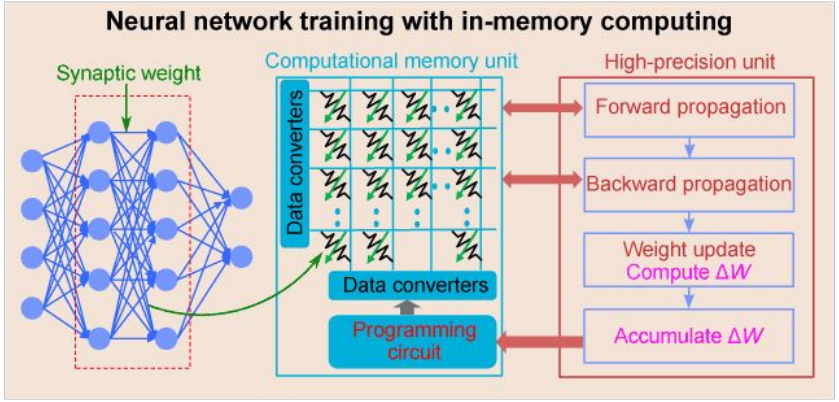
encode the synaptic weights as the conductance values of phase-change memory devices organised in a computational memory unit

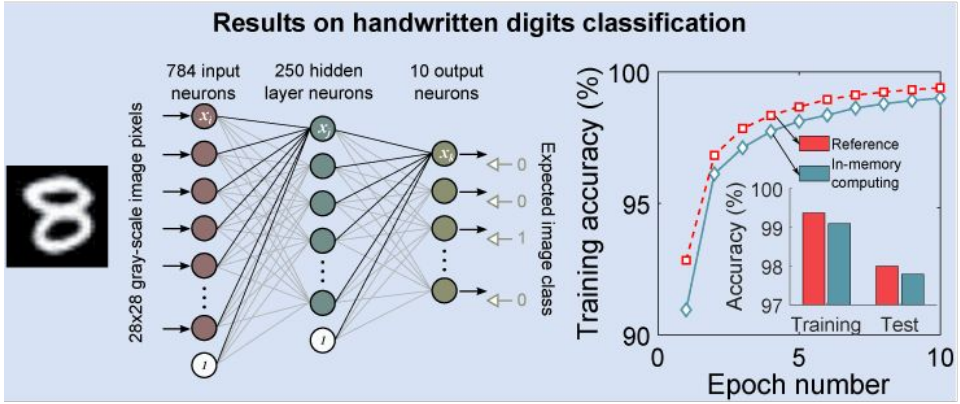
the weight changes are accumulated in high precision

This mixed-precision approach enables training the network to reach high classification accuracy, while performing the bulk of the computation as in-memory computing.

The trained weights will be retained in the computational memory (non-volatile) without any power

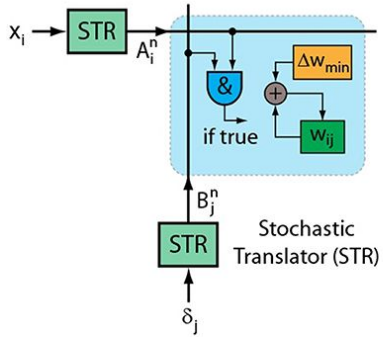
A chip trained in this way can be used for inference tasks within sensor devices at a fraction (<1%) of the power

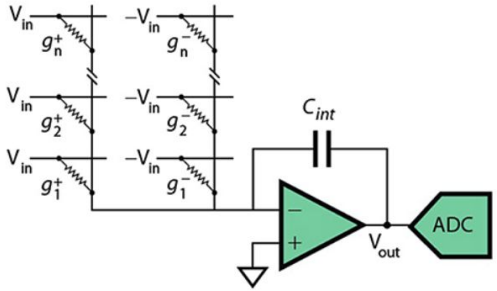


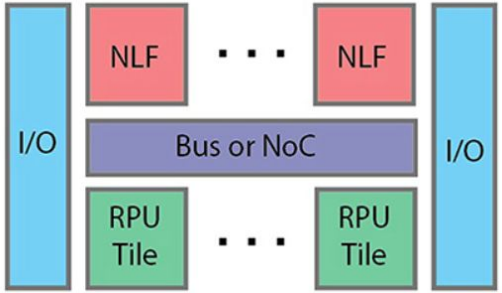


Resistance Processing Unit (RPU) for next generation in-memory computing

Each small RPU mimics a synapse in the brain







System	Throughput (TeraOps/s)	Power (W)	Power efficiency (GigaOps/s/W)	Network size (number of weights)	Acceleration vs. CPU
CPU Power8 12 Cores	0.676	250	2.7	–	1
GPU NVidia Tesla K40	4.3	242	17.8	–	6.4
Design 1	5000	250	20,100	200 M	7400
Design 2	21,000	250	83,800	840 M	31,000
Design 3	420	22	19,000	1680 M	620

Gokmen, Vlasov 2016

CPU+GPU+NPU hybrid system for mobile drives deep learning on the edge

	System-on-chip (SoC)	A11 Bionic	A12 Bionic	Kirin 970	Kirin980
Design	Supplier	Apple		Hisillicon	
	Released date	9.12.2018		8.31.2018	
	64 Bit	Yes			
	manufacturing process	10 nm TSMC	7nm TSMC	10nm TSMC	7nm TSMC
	Transistors	4.3 billion	6.9 billion	5.5 billion	6.9 billion
CPU	CPU Cores	2+4	2+4	4+4	2+2+4
	Performance CPU	Monsoon	New CPU× 2 + 15% performance	Cortex-A73 × 2	Cortex-A76 (2.6GHz) × 2 + Cortex-A76 (1.92GHz) × 2
	Efficiency CPU	Mistral × 4	New CPU× 4 + 50% efficiency	Cortex-A53 × 4	Cortex-A55 × 4
	Max Clock (GHz)	2.4	N/A	2.4	2.6
GPU	GPU	Internally-designed GPU	Internally-designed GPU	Mali-G72 MP12	Mali-G76
	GPU Cores	3	4	12	10
AI Accelerator	AI Processor	2-core Neural Engine	8-core Neural Engine	NPU	Dual NPU
	Performance	600 billion operations per second	5 trillion operations per second	2.22 trillion operations per second	5 trillion operations per second
Memory	Ram Interface	LPDDR4X	LPDDR4X	LPDDR4x	LPDDR4X
	Ram Frequency	N/A	N/A	1833	2133
	Max Bandwidth	N/A	N/A	29.9	34.1

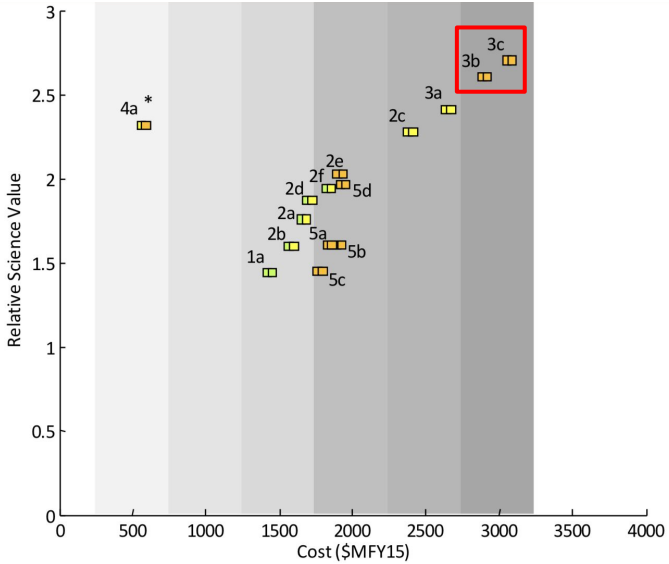
https://syncedreview.com/2018/09/13/ai-chip-duel-apple-a12-bionic-vs-huawei-kirin-980/

Performance comparisons among State-of-the-Art Compute-in-Memory Architectures

	State-of-the-Art Compute-in-Memory Architecture			GPU	CPU
	NVIDIA NVDLA	Mythic IPU	Gyrfalcon Lightspeeur 2801	NVIDIA Pascal Titan-X	6-core Intel i7
Speed (TOPS=trillion operations/second)	2	9.6	5.6	44	0.15
ResNet-50 neural network (224x224X3) Speed (frames/second)	269	900	142	180	2.8
Power (W)	0.291	2	0.3	250	95
Efficiency (TOPS/W)	5.4	4.8	9.6	0.18	0.001

Autonomy in space performance requirement

- Mars helicopter 50GOPS, 1TOPS/W AI
- Next generation autonomous rover requires >3,000GOPS (63X of Mars2020)
- In-memory computing will reduce risk of the Enceladus lander mission



- 1a Enc. Flyby
- 2a Simple Enc. Orbiter
- 2b Simple Enc. Orbiter Short Ops
- 2c Enhanced Enc. Orbiter
- 2d Simple Enc. Orbiter Freeflying Mag
- 2e Simple Enc. Orbiter Small Landers
- 2f Simple Enc. Orbiter Impactor
- 3a High-Perf Enc. Orbiter
- 3b High-Perf Enc. Orbiter Small Landers
- 3c High-Perf Enc. Orbiter Smart Lander
- 4a Titan-Enceladus Connection
- 5a Enc. Sample Return
- 5b Enc. Sample Return Solar/SEP
- 5c Enc. Sample Return 4 km/s
- 5d Enc. Sample Return Extra Payload

Risk Legend

Mission Risk (Left-hand color of data symbols): reduction in science return

Implementation Risk (Right-hand color of symbols): consumption of project margins (cost, schedule, perf.)

- Mission failure/Overrun
- Significant
- Moderate
- Small to moderate
- Minimal

NASA Enceladus Mission Concept Study 2010

Challenges of in-memory architectures for deep learning

- Hardware demand
  - Speed and power in parallel weights update
  - Large memory required
- Persistency
  - Unstable analog memory states (low precision)
  - Write endurance (flash memory limitation)
- Integration
  - Need ADC/DAC to connect with other functions (activation, pooling)
  - Programming analog memory states

Summary

- Deep learning is the core of next generation autonomy in space computing
- Currently GPU is the best tool for deep learning
- Specialized ASICs (TPU) are currently being implemented, performing 10X than GPU
- In next 5 years, in-memory computing (IMC) will be the most efficient tool for deep learning, achieving 100X performance than GPU
- IMC-based deep learning will be crucial for accelerating autonomy in space