Morgan Lewis

SILICON VALLEY FIRST CUP OF COFFEE SEMINAR SERIES

UPCOMING SEMINARS:

Artificial Intelligence (AI) Boot Camp

January 12	Computer-Implemented Inventions in Biotechnology and Healthcare, Patentability from European and US Perspective
January 13	M&A and Investment into AI Companies
January 19	Software As a Medical Device: US FDA Regulatory and Legal Framework
January 20	Patent and Trade Secret Protection for Inventions That Use AI
January 21	AI in Hiring and Recruiting
January 28	AI and Copyright

Morgan Lewis

SILICON VALLEY FIRST CUP OF COFFEE SEMINAR SERIES

UPCOMING SEMINARS:

Artificial Intelligence (AI) Boot Camp

February 2	The Ethics of Artificial Intelligence for the Legal Profession
February 3	AI and Data Privacy
February 4	Patents for MedTech AI: Opportunities and Pitfalls
February 9	IP Landscape of AI Hardware Startups
February 11	AI in Digital Advisory Offerings: Regulatory Considerations
February 16	Bias Issues and AI
February 25	The Risks of Bias and Errors in AI-Enabled Decision-Making

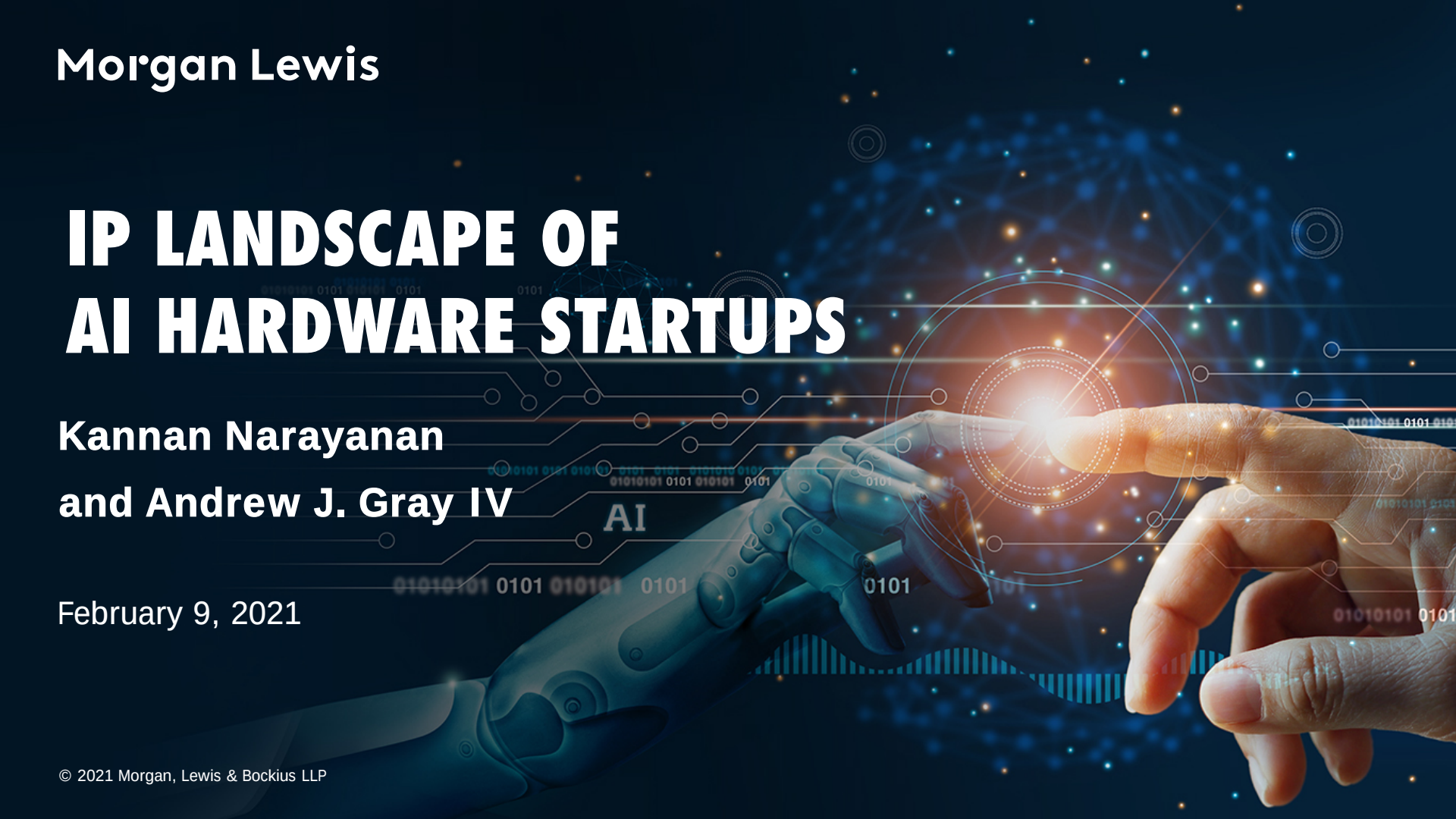
Morgan Lewis

IP LANDSCAPE OF AI HARDWARE STARTUPS

Kannan Narayanan

and Andrew J. Gray IV

February 9, 2021



Presenters

Kannan Narayanan

Andrew J. Gray IV

Morgan Lewis

Short Background

Kannan Narayanan
Silicon Valley

1994 – 1998: Computer Science & Engineering Bachelors (IIT)
1998 – 1999: Computer Science Masters (U Pitt)
2000 – 2017: R&D, Engineering, and Leadership Roles
(Cisco Systems, Intel, AMD, and Hardware Startups)
2014 – 2018: J.D. at Santa Clara University
2018 – 2021: IP Associate, Morgan Lewis

Presentation Overview

Part 1: Background (Technology & IP Overview)

1. Growth in AI Applications & Data
2. Limitations of Conventional Hardware
3. Need for Specialized Hardware
4. AI Hardware Technologies

Part 3: Other Topics

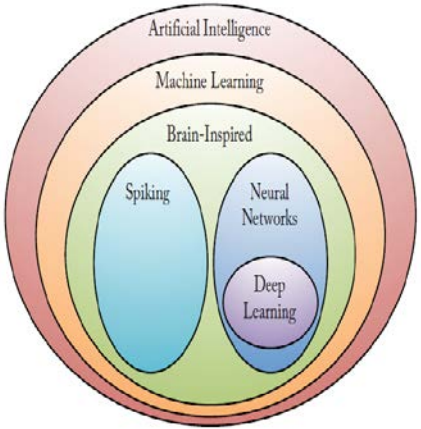
1. Open-Source Software / Hardware
2. Trade Secret Protection
3. Defensive Publications
4. AI Hardware-Related Patent Litigation
5. Conclusion

Part 2: Case Studies (Patenting Strategies)

1. Training (Nvidia, Intel, Cerebras, Graphcore)
2. Cloud Computing (Google)
3. Datacenter (Facebook)
4. FPGA (AMD + Xilinx)
5. Neuromorphic Computing (Brainchip)
6. Optical Computing (Luminous)
7. Fully Homomorphic Encryption (Cornami)
8. Analog Compute-in-Memory (Mythic)
9. Inference (Qualcomm)
10. FP Conversion (Cambricon)

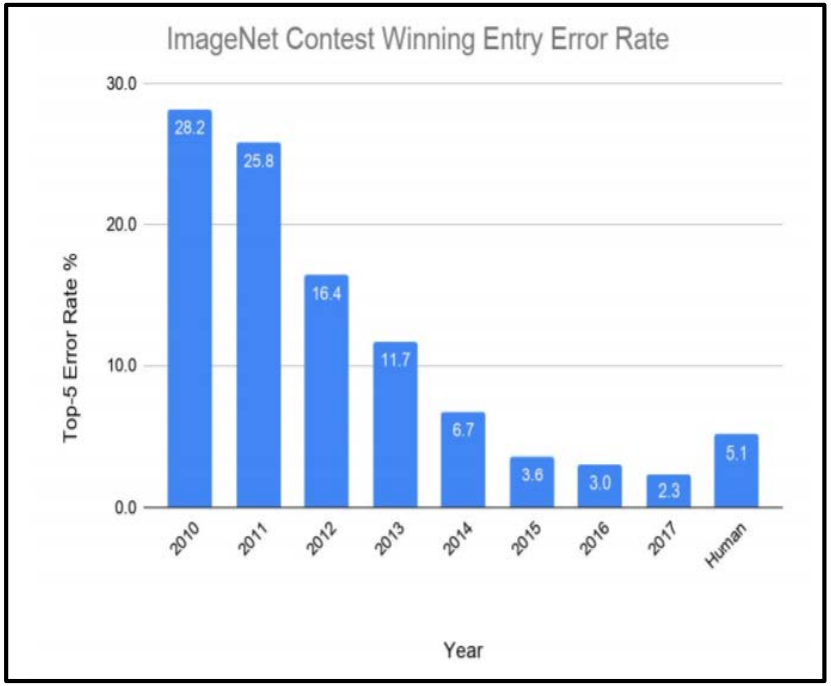
Part 1: Background (Technology & IP Overview)

1. Growth in AI Applications and Data
2. Limitations of Conventional Hardware
3. Need for Specialized Hardware
4. AI Hardware Technologies

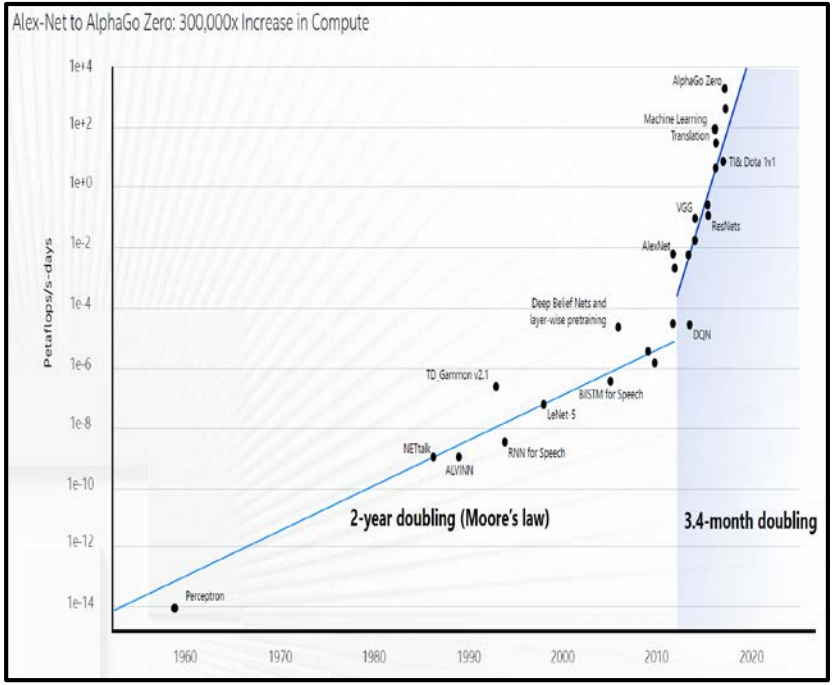


Source: *Efficient Processing of Deep Neural Networks*, Sze et al.

Growth in AI Applications and Data

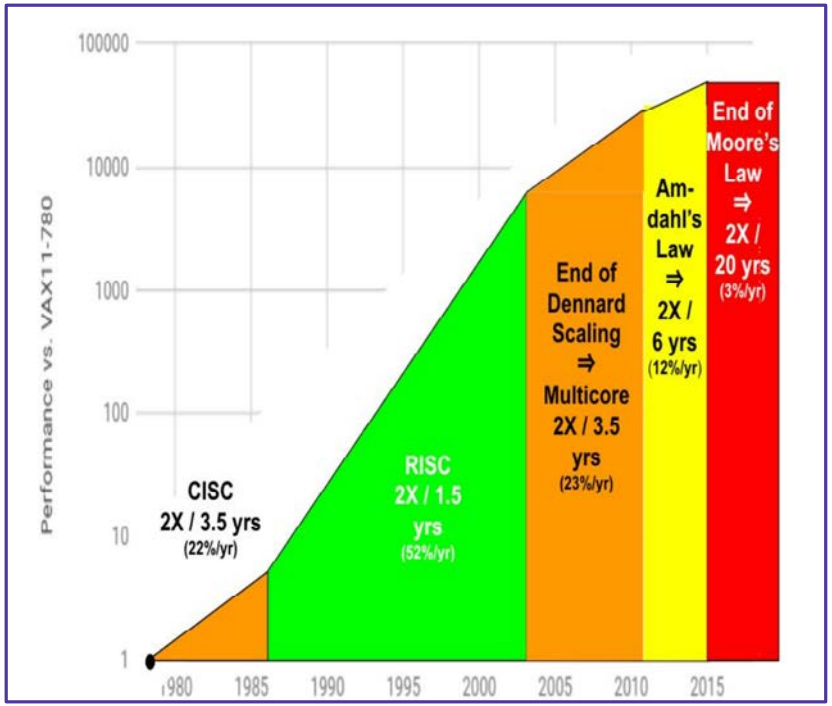


Source: Jeff Dean (Google)
https://arxiv.org/ftp/arxiv/papers/1911/1911.05289.pdf

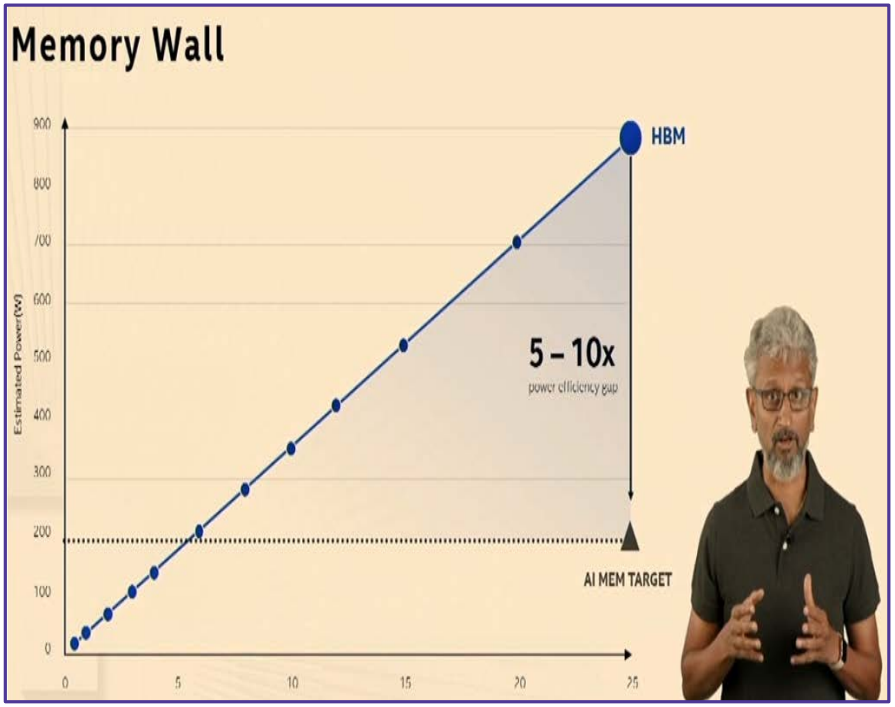


Source: AI and Compute (November 2019)

Limitations of Conventional Hardware



Source: Computer Architecture: A Quantitative Approach, 6/e 2018, John Hennessy and David Patterson,



Source: Raja Koduri (Intel), Hot Chips 2020

Need for Specialized Hardware

Deep learning models have three properties:

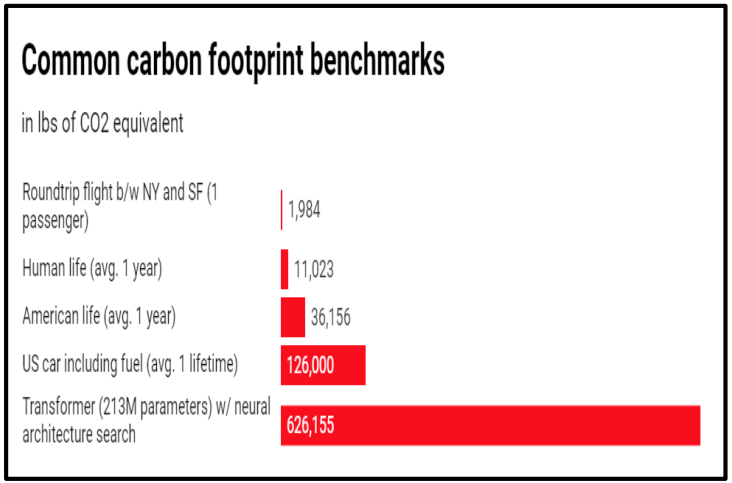
1. Tolerance for reduced-precision computations
2. Computations are mostly compositions of matrix multiplies, vector operations, applications of convolutional kernels, and dense linear algebra operations.
3. Not a significant use of branch predictors, speculative execution, hyper-threaded execution processing cores and deep cache memory hierarchies and TLB

Need for Specialized Hardware (Training vs. Inference)

- **Training and inference have very different compute and responsiveness characteristics**
  - For **training**, larger models on larger data sets require multiple chips
    - Need larger-scale systems, accelerators, and high-performance interconnects.
  - For **inference**, 8-bit integer-only calculations are sufficient for many models.
    - Single-chip inference on low-power devices in areas like speech or vision target low-precision linear algebra computations at high performance/Watt.

Need for Reducing Power Consumption

Embedded devices have limited battery capacity, **data centers** have a power ceiling due to cooling cost.



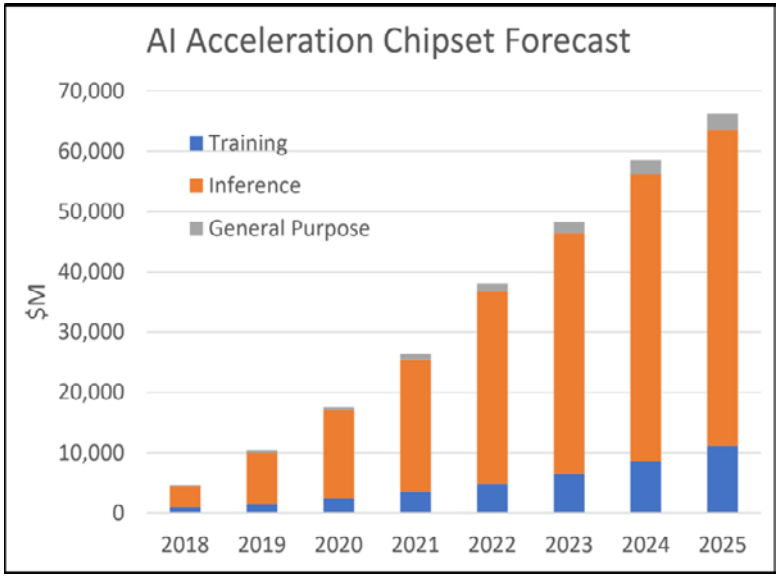
Source: MIT Technology Review

Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Multiply	0.2
32b Multiply	3.1
16b FP Multiply	1.1
32b FP Multiply	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

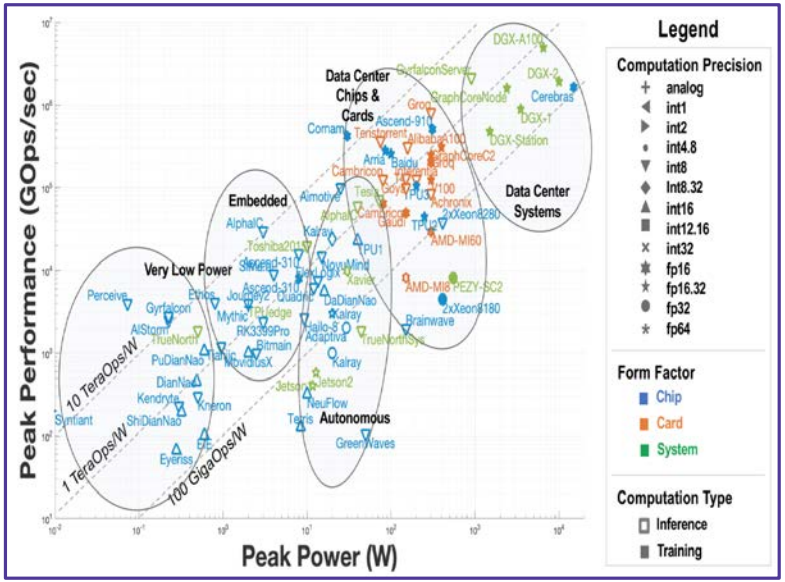
Relative Energy Cost

Source: Horowitz, ISSCC 2014

Major Semiconductor Disruption Is Underway

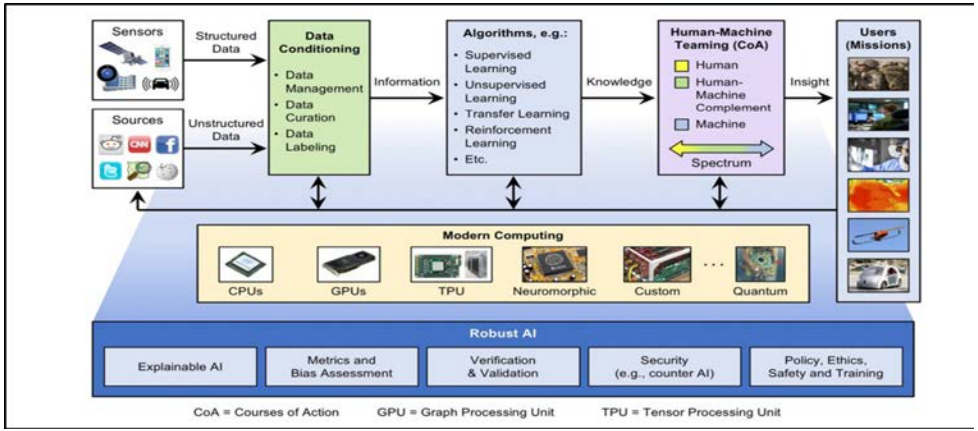


Source: Tractica Deep Learning Chipsets



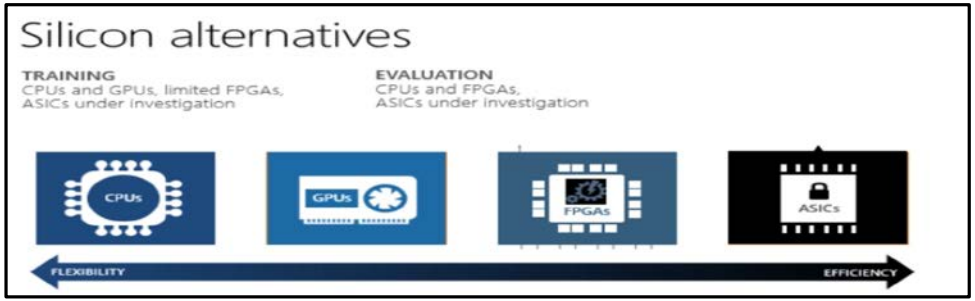
Source: *Survey of Machine Learning Accelerators*, Reuther et al.

AI Architecture

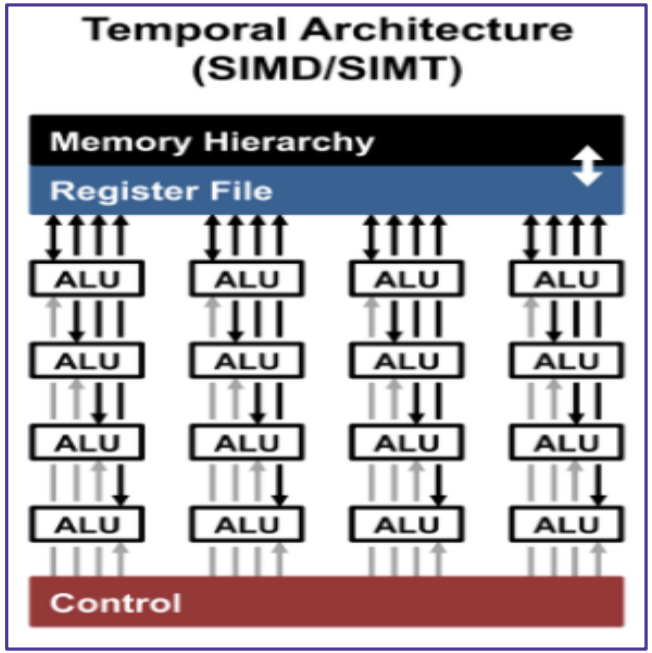


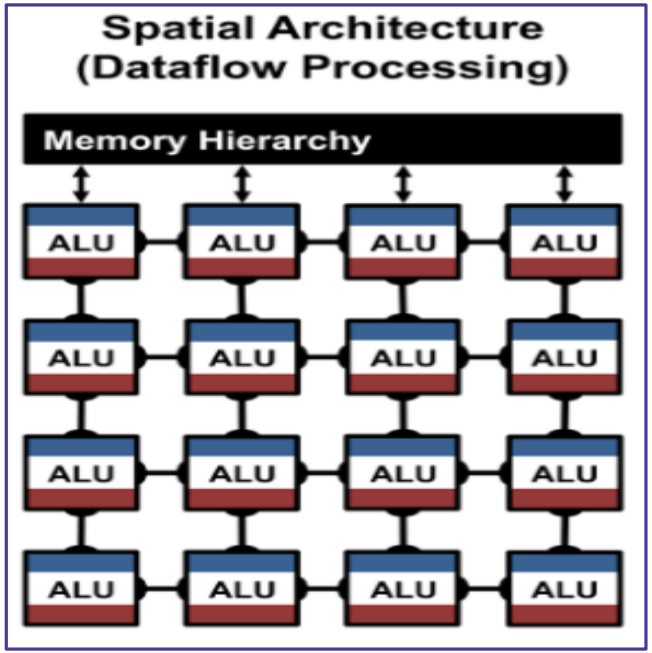
Source: *Survey of Machine Learning Accelerators*, Reuther et al.

Source: Microsoft (venturbeat.com)



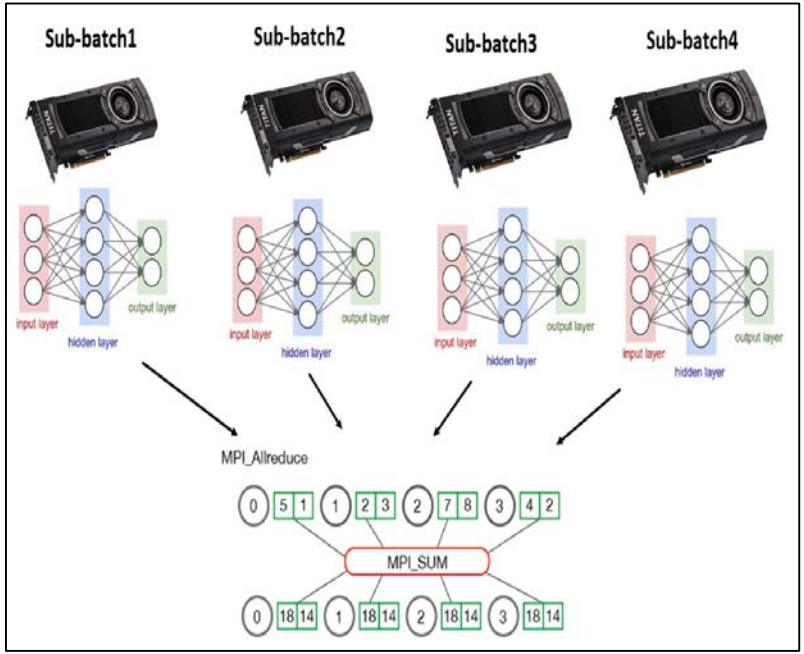
AI Hardware Architecture
Temporal vs Spatial Architecture



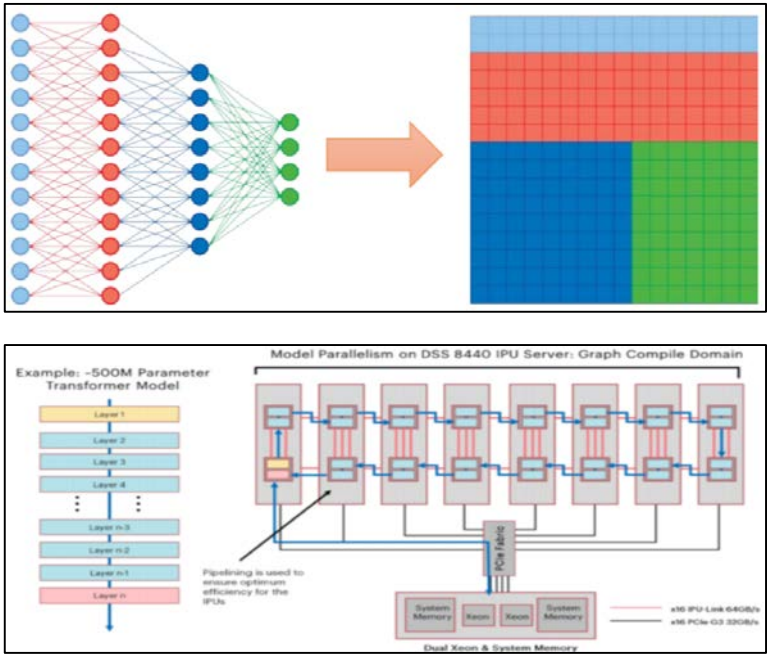


Model vs Data Parallelism

Data Parallelism

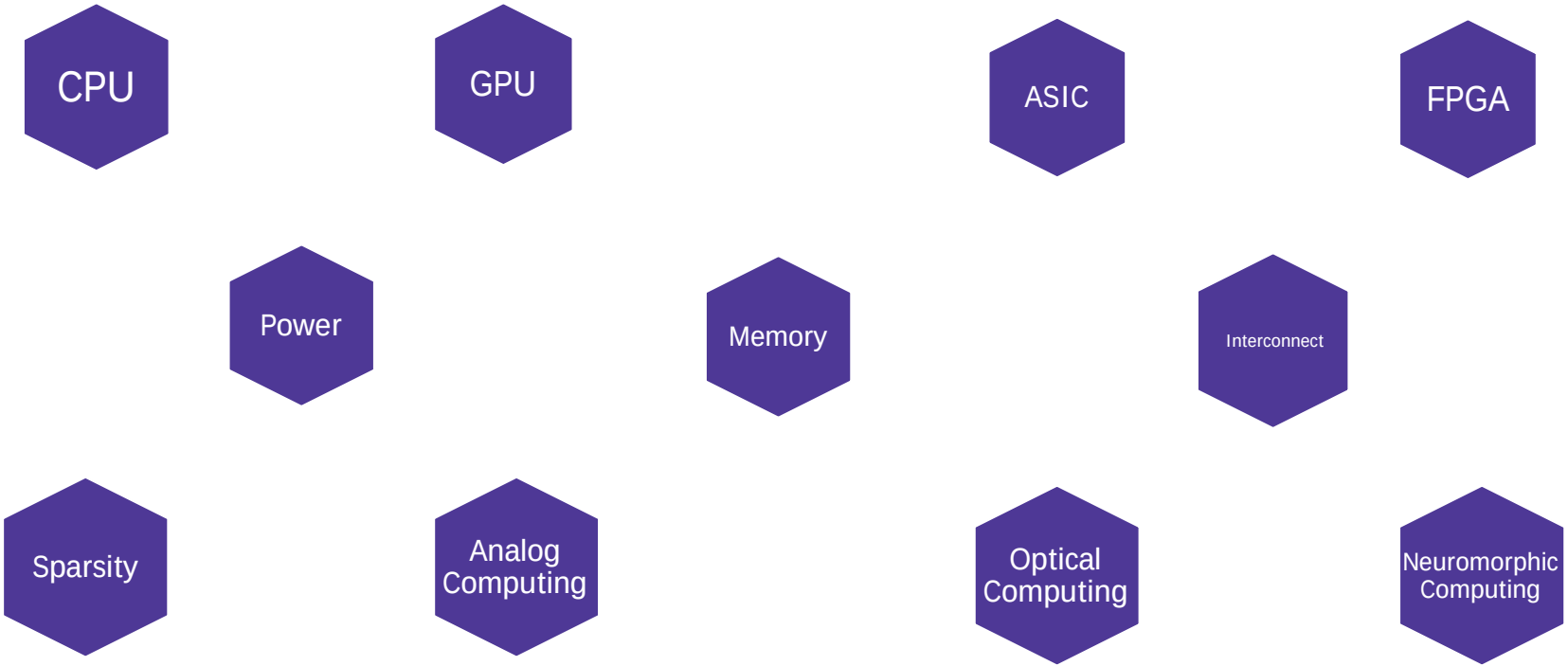


Model Parallelism



Source: *TPU vs GPU vs Cerebras vs Graphcore: A Fair Comparison between ML Hardware*, Mahmoud Khairy

Innovation Landscape



Innovation Landscape

Technology Area	Importance	Example Innovations	Example Players
CPU	Application logic, system performance, and efficiency, majority of AI inferencing on CPU	Faster CPUs, special instructions, data format (e.g., bfloat16)	Intel, ARM, AMD, IBM
GPU	Majority of AI training is on GPU, process technology, software support / ecosystem	Specialized MAC cores	Nvidia, Intel, AMD, Imagination
Memory & Interconnect	Training requires lots of data, and data movement	Near memory, Resistive RAM (ReRAM), NOR Flash technology, RDMA over Converged Ethernet (RoCE)	Nvidia, Intel, Cerebras, GraphCore

Innovation Landscape

Technology Area	Importance	Example Innovations	Example Players
Power Optimizations	Lower energy costs, improve reliability in data centers, inferencing and edge computing need low power	Embedded FPGA for edge inferencing, custom NN, dataflow, NOR Flash for in-memory, exploit sparsity, low bit quantization, reduce memory bottlenecks	Flex Logix, SiMa.ai, Horizon Robotics, BrainChip, Hailo, LeapMind, Movidius, Lattice Semiconductor, Gyrfalcon Technology
Analog computation	Low-power	Flash components for in-memory computation, resistive materials, “charge-domain” technology	Mythic, LightOn, AIStorm, Analog Inference
Optical computing	Speed of light data transfer, power savings	Use light beams, which interact linearly through the superposition effect, for computing linear algebra	IBM, LightOn, Fathom Computing, Lightelligence, Lightmatter, Luminous

Innovation Landscape

Technology Area	Importance	Example Innovations	Example Players
Neuromorphic computing	Lower energy costs, process visual and auditory stimuli as efficiently as brain	Spiking or Pulsed Neural Networks emulate brain functions	IBM TrueNorth, Intel Loihi SynSense, General Vision, BrainChip
Sparsity	Network weights ~ 0 and contribute negligibly to the overall results. Sparse matrix techniques ignore the computation associated with zero and near-zero values	Pruning, model sparsity, sparse evolutionary training	GrAI Matter Labs, Tenstorrent

Evaluation Metrics for ML Accelerators

- Accuracy
- Throughput (high volume data, real-time)
- Latency (for interactive applications)
- Energy and Power (embedded devices, data center cooling costs)
- Hardware Cost
- Flexibility (range of models and tasks)
- Scalability

Source: *Efficient Processing of Deep Neural Networks*, Sze et al.

Example Metrics for ML Accelerators (Training)

	#	Metric	Google TPU v3 (2 Shelves)	DGX-2H V100	DGX A100	Cerebras CS-1	GraphCore DELL IPU
Raw Metrics	1	Description	16x TPU chips (4 boards)	16x V100 chips	8x A100 chips	1x WSE chip	16x IPU chips (8 cards)
	2	Technology node	>12nm (16nm est.)	TSMC 12nm	TSMC 7nm	TSMC 16nm	TSMC 16nm
	3	Total Compute Transistor Count (B)	176 (est.)	336	432	1200	376
	4	Server Size (rack unit factor)	6U (est.)	10U	6U	15U	4U
	5	Total Compute memory (GB)	512	512	320	18	4.8
	6	Theoretical PFLOPS (16 bit)	1.9	2	2.5	2.5	2
	7	Scale-up Interconnect	Unkown fabric	NVSwitch (2.4 TB/sec)	NVSwitch (4.8 TB/sec)	on-chip 2D mesh (100 Pb/sec)	Bi-directional ring (300 GB/sec/chip)
	8	Scale-out Interconnect	Unkown fabric	8x Infiniband 100 Gb/sec	8x Infiniband 200 Gb/sec	12x Ethernet 100 Gb/sec	4x Ethernet 100 Gb/sec
	9	Built-in CPU & Storage	Yes (2 CPUs + unkown mem + no storage)	Yes (2 CPUs + 1.5TB mem + 30 TB SSD)	Yes (2 CPUs + 1 TB mem + 15 TB SSD)	No	Yes (2 CPUs + 0.7 TB mem + 16 TB SSD)
	10	Max compute TDP (watts)	7,200	7,200	3,200	20,000	2,400
	11	Max system TDP (watts)	9,250	12,000	6,500	Unkown	4,800
	12	Cloud Price (USD/chip/hour)	2	2.48	3.1	N/A	0.8
	13	On-premise purchase (USD)	N/A	399K	199K	2M (est.)	105K
Efficiency Metrics	14	Performance of MLPerf-Resent training throughput (images/sec)	29,163 (29 mins)	24,164 (35 mins)	21,143 (40 mins)	Unkown	Unkown
	15	Model/Memory size (GB)	16	32-512	40-320	18	4.8
	16	Transistor Density (BTran/U)	29.3	33.6	72	80	94
	17	Theoretical Performance/Area (Theoretical PFLOPs/U)	0.26	0.2	0.42	0.16	0.5
	18	Achievable Performance/Area (Resnet train throughput/U)	4,986	2,416	3,320	Unkown	Unkown
	19	Power Density (watts/U)	1,541	1,200	1,083	1,333	1,200
	20	Theoretical Performance/Watts (Theoretical TFLOPs/watts)	0.26	0.27	0.78	0.125	0.83
	21	Achievable Performance/Watts (Resnet train throughput/watts)	4.15	3.35	6.2	Unkown	Unkown
	22	Theoretical Performance/Cloud Price (Theoretical PFLOPs/USD)	0.95	0.8	0.8	N/A	2.5
	23	Achievable Performance/Cloud Price (Resnet train throughput/USD)	911	608	852	Unkown	Unkown

Source: *TPU vs GPU vs Cerebras vs Graphcore: A Fair Comparison between ML Hardware*, Mahmoud Khairy

Benchmarks for ML Accelerators (MLPerf)

Results from other rounds:

- MLPerf Inference v0.5 results
- MLPerf Inference v0.6 skipped to align with training
- MLPerf Inference v0.7 results

- (●) Datacenter, Closed Division
- ( ) Datacenter, Open Division
- ( ) Edge, Closed Division
- ( ) Edge, Open Division
- ( ) Mobile Phones, Closed Division
- ( ) Mobile Phones, Open Division
- ( ) Mobile Notebooks, Closed Division
- ( ) Mobile Notebooks, Open Division

Closed Division Times

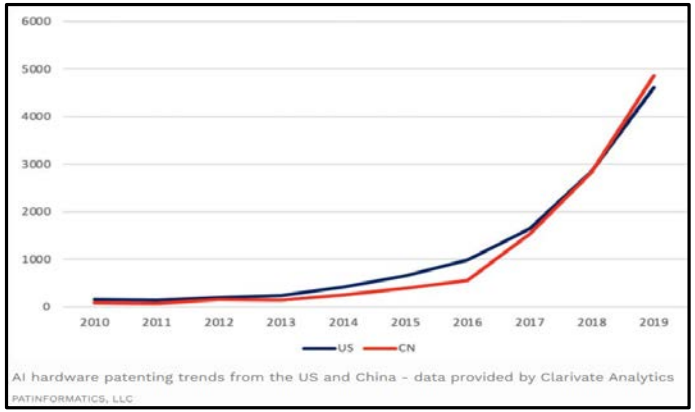
#	Submitter	System	Processor	#	Accelerator	#	Software	Benchmark results (minutes)							
								Image classifi-cation	Object detection, light-weight	Object detection, heavy-wt.	Translation, recurrent	Translation, non-recur.	Recom-mendation	Reinforce-ment Learning	
								ImageNet	COCO	COCO	WMT E-G	WMT E-G	MovieLens-20M	Go	
								ResNet-50 v1.5	SSD w/ ResNet-34	Mask-R-CNN	NMT	Transformer	NCF	Mini Go	
Available in cloud															
0.6-1	Google	TPUv3.32			TPUv3	16	TensorFlow, TPU 1.14.1.dev	42.19	12.61	107.03	12.25	10.20	[1]		
0.6-2	Google	TPUv3.128			TPUv3	64	TensorFlow, TPU 1.14.1.dev	11.22	3.89	57.46	4.62	3.85	[1]		
0.6-3	Google	TPUv3.256			TPUv3	128	TensorFlow, TPU 1.14.1.dev	6.86	2.76	35.60	3.53	2.81	[1]		
0.6-4	Google	TPUv3.512			TPUv3	256	TensorFlow, TPU 1.14.1.dev	3.85	1.79		2.51	1.58	[1]		
0.6-5	Google	TPUv3.1024			TPUv3	512	TensorFlow, TPU 1.14.1.dev	2.27	1.34		2.11	1.05	[1]		
0.6-6	Google	TPUv3.2048			TPUv3	1024	TensorFlow, TPU 1.14.1.dev	1.28	1.21			0.85	[1]		
Available on-premise															
0.6-7	Intel	32x 2S CLX 8260L	CLX 8260L	64			TensorFlow						[1]	14.43	
0.6-8	NVIDIA	DGX-1			Tesla V100	8	MXNet, NGC19.05	115.22					[1]		
0.6-9	NVIDIA	DGX-1			Tesla V100	8	PyTorch, NGC19.05		22.36	207.48	20.55	20.34	[1]		

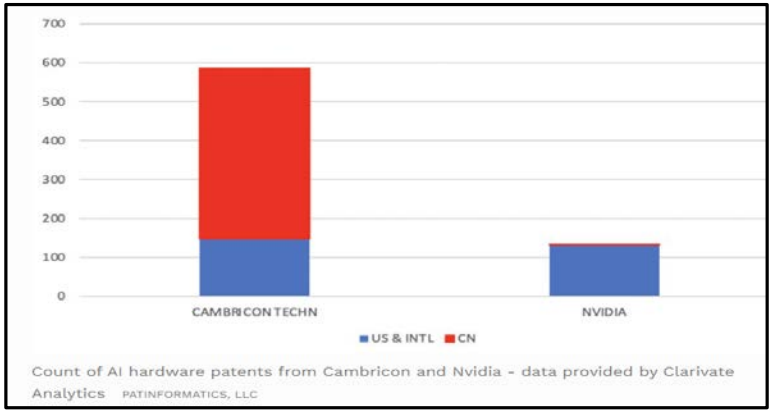
Results

#	Accelerator	#	Software	Task	Image classification		Object detection (large)		Medical imaging		Speech-to-text		Natural Language Processing				Recommendation			
				Data	ImageNet		COCO		BraTS 2019		LibriSpeech		SQuAD v1.1				1TB Click Logs			
				Model	ResNet		SSD-Large		3D-UNet		RNN-T		BERT				DLRM			
				Accuracy (%FP32 ref)		99.00		99.00	99.00	99.90		99.00		99.00		99.90		99.00		99.90
				Scenario	Server	Offline	Server	Offline	Offline	Offline	Server	Offline	Server	Offline	Server	Offline	Server	Offline	Server	Offline
				Units	queries/s	samples/s	queries/s	samples/s	samples/s	samples/s	queries/s	samples/s	queries/s	samples/s	queries/s	samples/s	queries/s	samples/s	queries/s	samples/s
2	NVIDIA T4	4	TensorRT 7.2, CUDA 11.0 Update 1		20,002	22,973	450	518	27	27	3,397	5,489	1,099	1,621	549	669				
2	NVIDIA T4	6	TensorRT 7.2, CUDA 11.0 Update 1		34,603	37,268			44	44							189,019	206,499	189,019	206,499
2	NVIDIA A100-PCIe	2	TensorRT 7.2, CUDA 11.0, cuDNN 8		52,425	62,885	1,544	1,671	74	74	14,600	16,982	5,100	5,610	2,428	2,791	385,085	454,409	385,085	454,409
2	NVIDIA GRID T4-16Q	4	TensorRT 7.2, CUDA 11.0 Update 1		20,401	22,726	425	536												
2	NVIDIA Quadro RTX 6000	4	TensorRT 7.2, CUDA 11.0, cuDNN 8		52,527	59,130	1,379	1,449	44	44	11,102	15,052	4,236	4,849	2,278	2,511	279,364	322,834	279,364	322,834
2	NVIDIA Quadro RTX 6000	10	TensorRT 7.2, CUDA 11.0 Update 1		123,200	150,388	3,497	3,609	163	163	29,608	38,454	10,653	12,029	5,604	6,259				
2	NVIDIA Quadro RTX 8000	3	TensorRT 7.2, CUDA 11.0, cuDNN 8		41,489	44,750	1,044	1,066	54	54	8,996	11,860	2,807	3,629	1,649	1,879	233,103	270,418	233,103	270,418
2	NVIDIA Quadro RTX 8000	8	TensorRT 7.2, CUDA 11.0 Update 1		99,004	119,210	2,598	2,881	142	142	24,628	30,963	8,568	9,567	4,496	4,964				
2	NVIDIA Quadro RTX 8000	10	TensorRT 7.2, CUDA 11.0, cuDNN 8		134,316	149,124	3,521	3,589	163	163	29,528	38,333	10,804	11,944	5,604	6,217	777,956	929,411	777,956	929,411
1	NVIDIA T4	4	TensorRT 7.2.0.14, CUDA 11.0.207		21,506	23,290	450	535	28	28	4,096	5,712	1,249	1,708	629	715	126,514	126,287	126,514	126,287
2	NVIDIA T4	4	TensorRT 7.2.0.14, CUDA 11.0.207		21,805	23,844	470	546	29	29	4,196	5,875	1,349	1,739	679	743	126,015	131,571	126,015	131,571

Source: *mlperf.org*

Global Patent Landscape





Source: Forbes.com

Cambricon, once Huawei's core AI chip supplier, eyes $400M IPO

Source: TechCrunch (June 24, 2020)

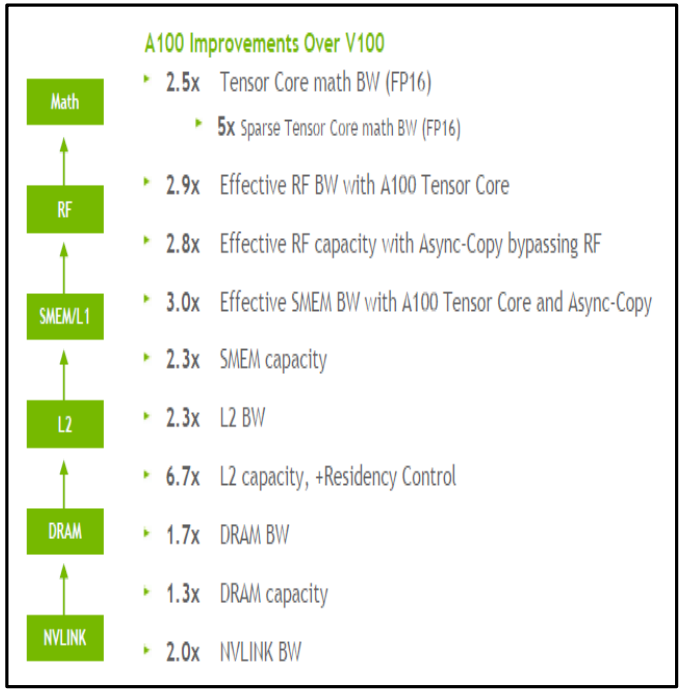
Source: enterpriseai.news

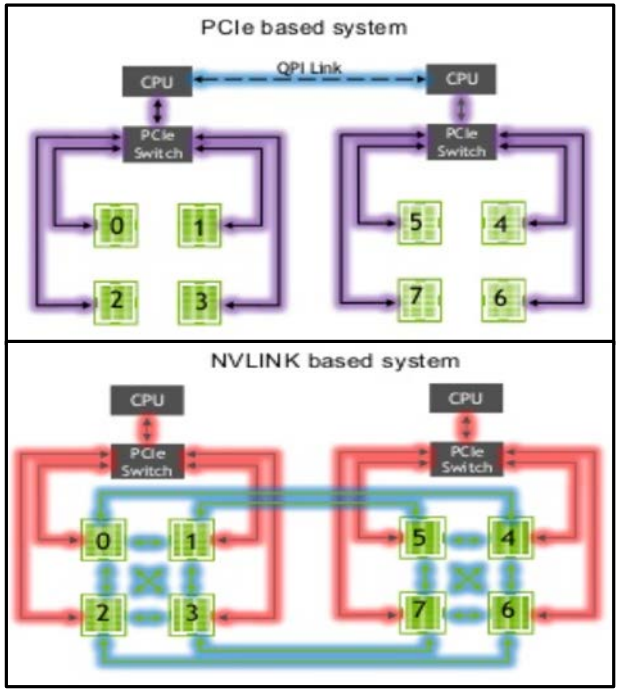
Part 2: Case Studies

Agenda (claiming strategies, continuation strategies, legal hurdles, overcoming the legal hurdles through claim amendments)

- Training (Nvidia, Intel, Cerebras, Graphcore)
- Cloud Computing (Google)
- Datacenter (Facebook)
- FPGA (AMD + Xilinx)
- Neuromorphic Computing (Brainchip)
- Optical Computing (Luminous)
- Fully Homomorphic Encryption (Cornami)
- Analog Compute-in-Memory (Mythic)
- Inference (Qualcomm)
- FP Conversion (Cambricon)

Case Study: Nvidia A100 (Training)





Source: NVIDIA

Case Study: Nvidia Mellanox (Traning)

MELLANOX TECHNOLOGIES

ALLOWANCE RATE

91.1%

TOTAL NUMBER OF APPLICATIONS

501
TOTAL

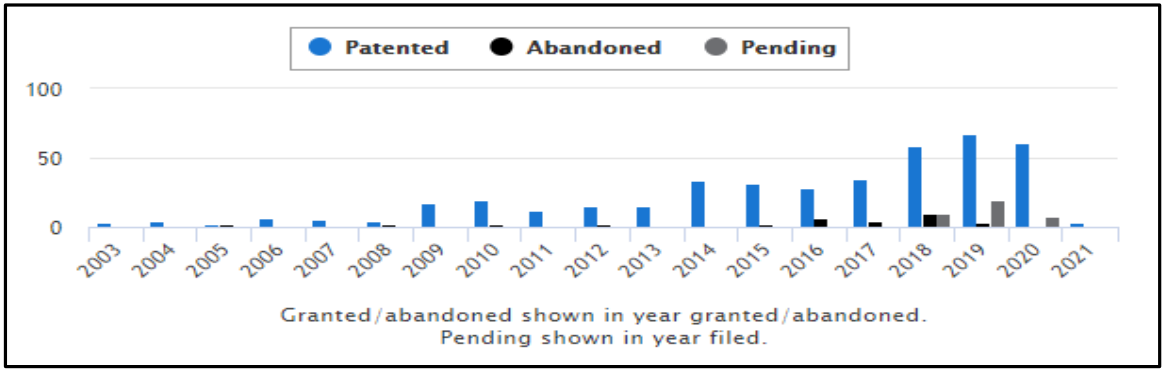
421 PATENTED APPLICATIONS
41 ABANDONED APPLICATIONS
39 PENDING APPLICATIONS

Source: PatentAdvisor

Israel News | Tech News

Worth $7 Billion? This Israeli Chip Maker Is Now at the Heart of Nvidia's New Strategy

Mellanox's new processor looks to be the central component in Nvidia's move into the data market – and one of the company's most important products



Source: PatentAdvisor

Case Study: Intel Nervana vs Intel Habana (Training)

- **Nervana NNP-T training chip**

  ~ high-end GPUs, supports HBM2, 16 GB (adds cost and manufacturing complexity),

  Model parallelism through a fast, low-latency fabric on the die, **not based on industry-standard Ethernet**

  NNP-T chip was being developed on TSMC, while NNP-I chip was being manufactured on Intel's 10-nm facility.

- **Habana**

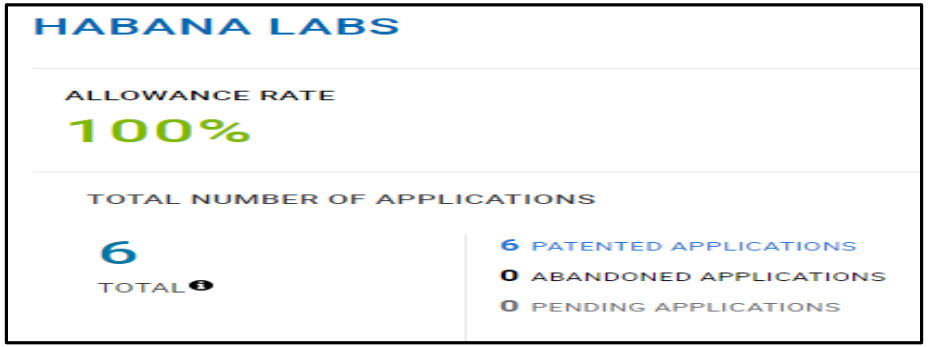
  Fabric supports RDMA over Converged Ethernet (RoCE)

  Converged architecture for inference and training

  Nervana software stack already supports model parallelism.

Case Study: Intel Habana (Training)

- Intel acquired Habana Labs for $2 Billion in December 2019.
- At the time of acquisition, Habana Labs had 3 issued patents and 3 pending applications.



Source: PatentAdvisor

Application	Filing Date	Status	Issue Date	Number of Office Actions between Filing Date and Patent Issuance	
15/700,213	2017-09-11	Patented Case	2020-12-01	1 office actions (Examiner average is 1.4)	Hiding latency of multiplier-accumulator using partial results
16/150,299	2018-10-03	Patented Case	2020-12-01	2 office actions (Examiner average is 1.6)	Processor Suspension Buffer and Instruction Queue
16/136,294	2018-09-20	Patented Case	2020-07-14	1 office actions (Examiner average is 1.4)	Hardware accelerator for outer-product matrix multiplication
15/883,119	2018-01-30	Patented Case	2019-11-26	1 office actions (Examiner average is 1.8)	LARGE-SCALE COMPUTATIONS USING AN ADAPTIVE NUMERICAL FORMAT
16/024,862	2018-07-01	Patented Case	2019-11-26	1 office actions (Examiner average is 1.7)	DATA COMPRESSION SCHEME UTILIZING A REPETITIVE VALUE WITHIN THE DATA STREAM
15/700,207	2017-09-11	Patented Case	2019-11-26	2 office actions (Examiner average is 1.8)	MATRIX MULTIPLICATION ENGINE

Intel Habana (Training) – Example Claming Strategy

- App. No. 15/700,213 (Patent No. 10,853,448, issued 12/01/2020)

  Title: Hiding latency of multiplier-accumulator using partial results

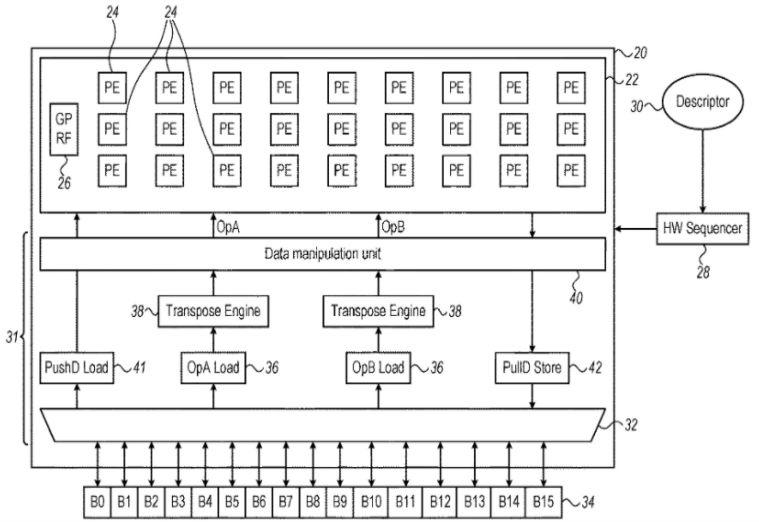
  Filed 09/11/2017; 1 Office Action (Art Unit 2182)

1. Computational apparatus, comprising:

a memory, which is configured to contain multiple matrices of input data values;

an array of processing elements, each configured to perform multiplications of respective first and second input operands and to accumulate products of the multiplication to generate respective output values; and

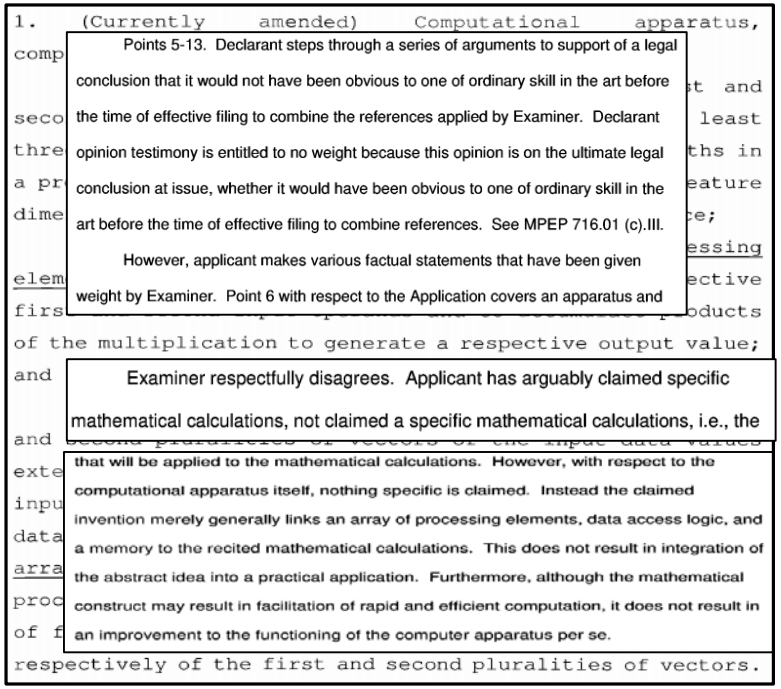
data access logic, which is configured to select from the memory a plurality of mutually-disjoint first matrices and a second matrix, and to distribute to the processing elements the input data values in a sequence that is interleaved among the first matrices, along with corresponding input data values from the second matrix, so as to cause the processing elements to compute, in the interleaved sequence, respective ~~convolutions~~ matrix multiplications of each of the first matrices with the second matrix, wherein the processing elements have a latency of a given number of clock cycles for each multiply-accumulate operation, and wherein the data access logic. is configured to select and distribute the input data values in the interleaved sequence from a number of the first matrices that is equal to the given number.



Intel Habana – Overcoming Section 101 Rejection

- App. No. 15/700,207 (Patent No. 10,489,479, issued 11/26/2019)

  Title: Matrix Multiplication Engine; Filed 09/11/2017; 2 Office Actions (Art Unit 2182)



1. (Currently amended) Computational apparatus, comp … st and … seco … least … thre … ths in … a pr … eature … dime … ce; … essing … elem … ective … firs … oducts … of the multiplication to generate a respective output value; and … and … exte … inpu … data … arra … proc … of f … respectively of the first and second pluralities of vectors.

Points 5-13. Declarant steps through a series of arguments to support of a legal conclusion that it would not have been obvious to one of ordinary skill in the art before the time of effective filing to combine the references applied by Examiner. Declarant opinion testimony is entitled to no weight because this opinion is on the ultimate legal conclusion at issue, whether it would have been obvious to one of ordinary skill in the art before the time of effective filing to combine references. See MPEP 716.01 (c).III.

However, applicant makes various factual statements that have been given weight by Examiner. Point 6 with respect to the Application covers an apparatus and

Examiner respectfully disagrees. Applicant has arguably claimed specific mathematical calculations, not claimed a specific mathematical calculations, i.e., the

that will be applied to the mathematical calculations. However, with respect to the computational apparatus itself, nothing specific is claimed. Instead the claimed invention merely generally links an array of processing elements, data access logic, and a memory to the recited mathematical calculations. This does not result in integration of the abstract idea into a practical application. Furthermore, although the mathematical construct may result in facilitation of rapid and efficient computation, it does not result in an improvement to the functioning of the computer apparatus per se.

1. (Currently amended) Computational apparatus, comprising:

a memory, which is configured to contain first and second input matrices of input data values, having at least three dimensions including respective heights and widths in a predefined sampling space and a common depth in a feature dimension, orthogonal to the predefined sampling space;

an array of processing elements arranged in hardware logic as a grid of array rows and array columns, each processing element configured to perform a multiplication of respective first and second input operands and to accumulate products of the multiplication to generate a respective output value; and

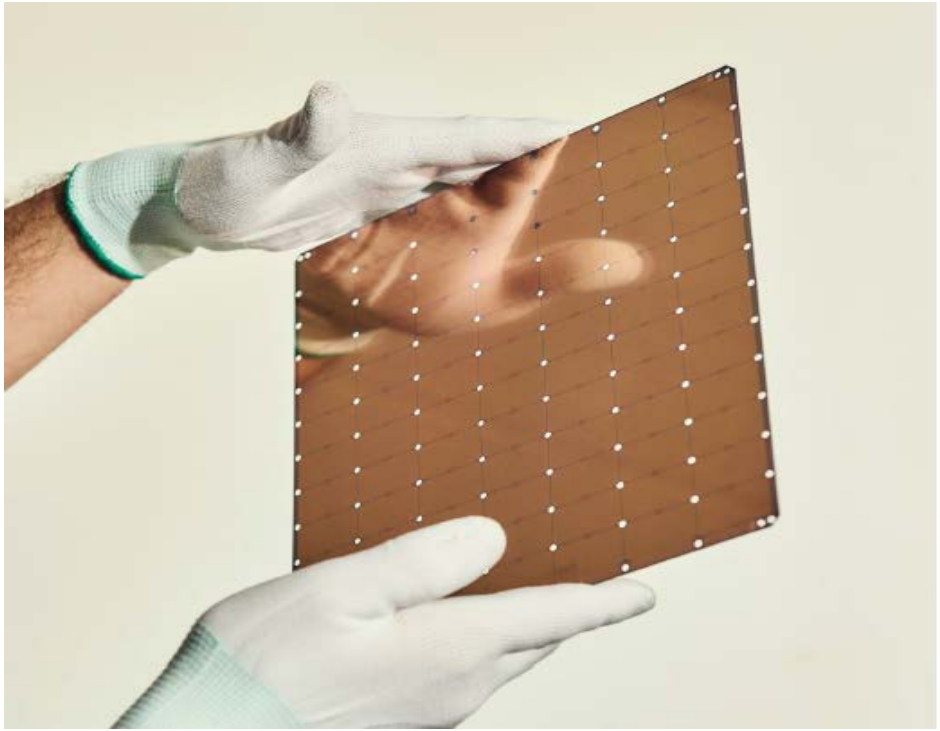
data access logic, which comprises:

at least one load unit, which is configured to extract from the memory first and second pluralities of vectors of the input data values extending in the feature dimension from the first and second input matrices, respectively; [[,]] and

a data manipulation unit, which is configured to distribute the input data values ~~from the extracted~~ in the first plurality of the vectors in sequence to at least a first array row in the array of the processing elements, and to distribute the input data values in the second plurality of the vectors in sequence to at least a first array column in the array of the processing elements,

so as to cause the processing elements in the array to compute a convolution of first and second two-dimensional (2D) matrices composed respectively of the first and second pluralities of vectors.

Case Study: Cerebras (Training)



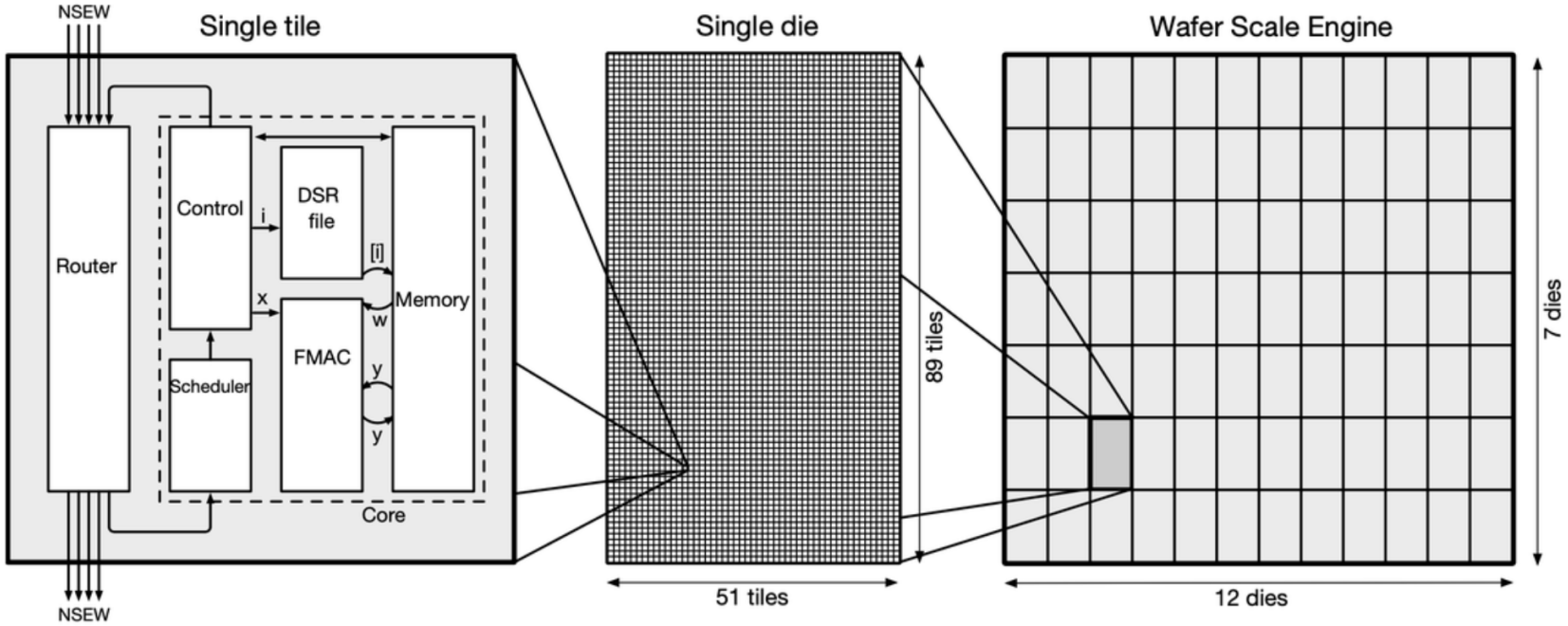
Cerebras Wafer Scale Engine (WSE)

The Most Powerful Processor for AI

400,000 AI-optimized cores
46,225 mm² silicon
1.2 trillion transistors
18 Gigabytes of On-chip Memory
9 PByte/s memory bandwidth
100 Pbit/s fabric bandwidth
TSMC 16nm process

Source: Cerebras (HotChips 2020)

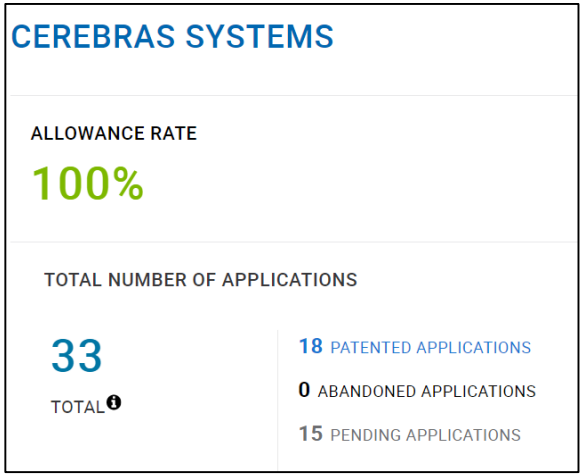
Case Study: Cerebras (Training)

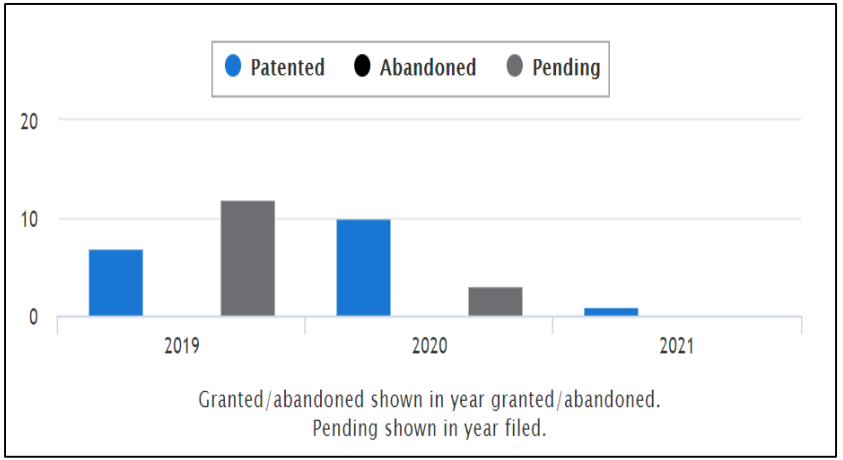


Source: Cerebras Systems (SC 2020)

Case Study: Cerebras (Training)

Source: PatentAdvisor





Source: Techcrunch.com

Problems solved:

- communicating across the scribe lines between chip
- handling yield
- thermal expansion,
- packaging and
- power/cooling

Cerebras (Training) – Continuation Stratgy

- App. No. 16/019,882 (Patent No. 10,366,967, issued 7/30/2019)

  Title: Apparatus and Method for Multi-Die Interconnection;

  Filed 06/27/2018; 2 Office Actions (Art Unit 2124)

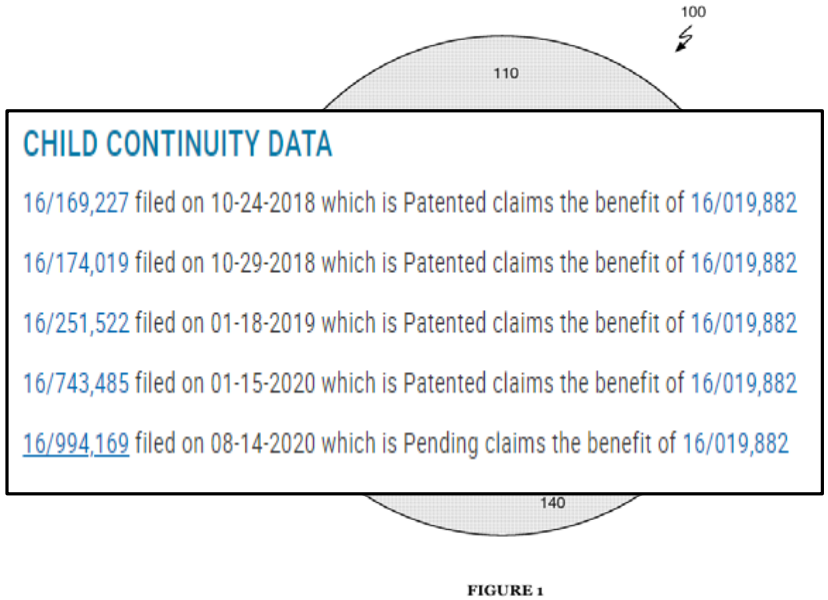


FIGURE 1

1. (CURRENTLY AMENDED) ~~A semiconductor wafer~~ An integrated circuit comprising:

a ~~body, the body of the~~ semiconductor ~~wafer comprising a single, continuous~~ substrate ~~formed with semiconductor material~~;

a plurality of ~~distinct~~ die lithographically formed within the ~~body of the~~ semiconductor substrate ~~wafer~~, ~~wherein each of~~ the plurality of ~~distinct~~ die ~~is integrally formed with the semiconductor material of the body of the semiconductor wafer and is not diced from the body~~ comprising a first die defining a plurality of die edges; and

a circuit layer formed at the plurality of die, the circuit layer comprising a plurality of inter-die ~~circuit~~ connections that communicatively connect die of the plurality of die, the plurality of inter-die connections comprising a first set of inter-die connections associated with the first die ~~each:~~

~~(i) extend across the semiconductor wafer and (ii) connect distinct pairs of die of the plurality of distinct die formed within the body of the semiconductor wafer;~~

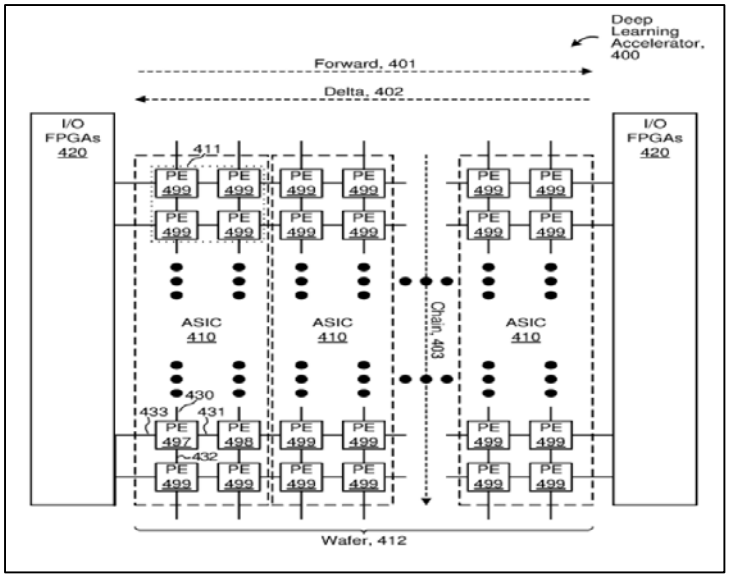
wherein the first set of inter-die connections comprises, for each die edge of the first die, a respective inter-die connection that crosses the die edge.

Cerebras (Training) – Software Claim Strategy

- App. No. 16/463,091 (Patent No. 10,614,357, issued 7/4/2020)

  Title: Dataflow Triggered Tasks for Accelerated Deep Learning;

  Filed 5/22/2019; 1 Office Action (Art Unit 2123)



3. (currently amended) A method comprising:

sending a fabric packet by a sending processing element to a fabric, the fabric packet comprising a virtual channel specifier and a fabric packet payload;

routing the fabric packet via the fabric from the sending processing element to a receiving processing element via zero or more routing processing elements, the routing in accordance with the virtual channel specifier;

in the receiving processing element, receiving the fabric packet from the fabric, reading one or more instructions from a memory of the receiving processing element at an address based at least in part on the virtual channel specifier, and using at least a portion of the fabric packet payload as an input operand to execute at least one of the one or more instructions;

wherein the virtual channel specifier is one of a plurality of virtual channel specifiers, each of the plurality of virtual channel specifiers is associated with a respective set of one or more sets of fabric packets, and the receiving comprises associating the fabric packet with the respective set associated with the virtual channel specifier; and

~~The method of claim 2,~~ wherein a block/unblock state is maintained for each of the virtual channel specifiers, and the block/unblock state of a particular one of the virtual channel specifiers is set to a block state in response to a block instruction specifying the particular one of the virtual channel specifiers and the block/unblock state of the particular one of the virtual channel specifiers is set to an unblock state in response to an unblock instruction specifying the particular one of the virtual channel specifiers.

Case Study: Graphcore (Training)

GRAPHCORE

ALLOWANCE RATE

100%

TOTAL NUMBER OF APPLICATIONS

76

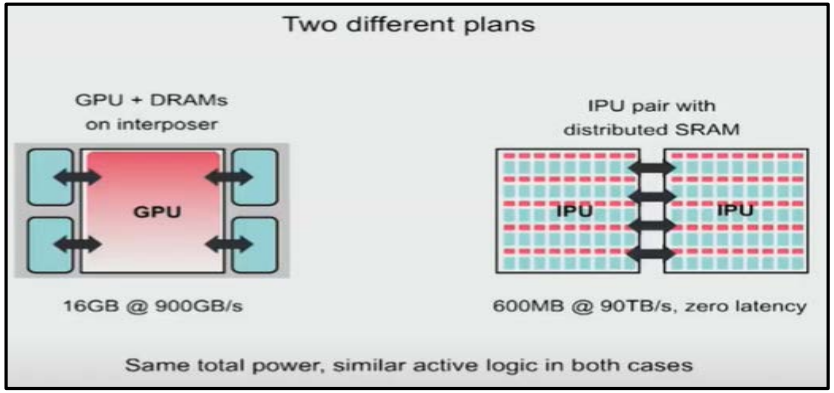
TOTAL

17 PATENTED APPLICATIONS

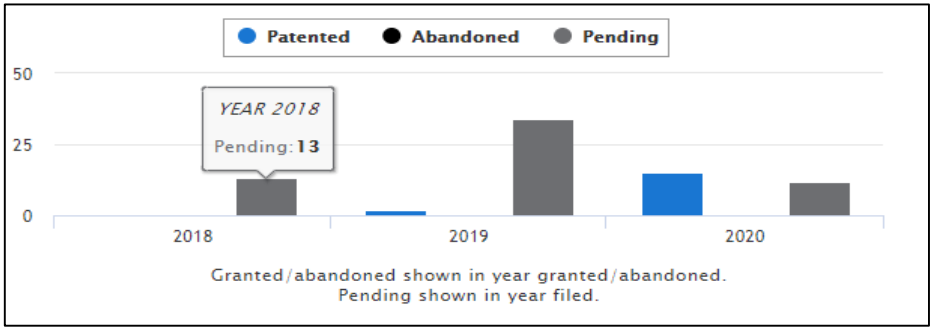
0 ABANDONED APPLICATIONS

59 PENDING APPLICATIONS

Source: PatentAdvisor



Source: GraphCore (Stanford 2018)



Morgan Lewis

Graphcore (Training) – Software Claim Strategy

- App. No. 15/886,053 (Patent No. 10,802,536, issued 10/13/2020)

  Title: Compiler Method; Filed 2/1/2018; 1 Office Action (Art Unit 2186)

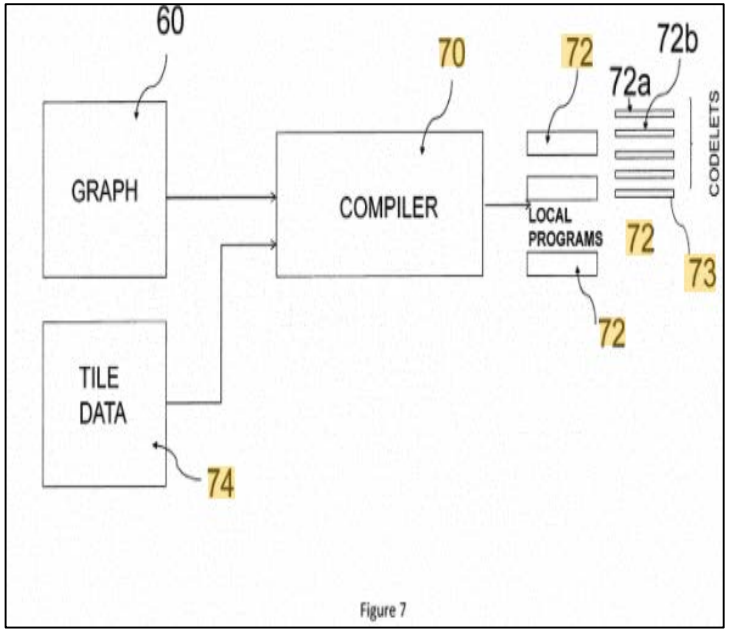


Figure 7

1. (Original) A computer implemented method of generating multiple programs to deliver a computerised function, each program to be executed in a processing unit of a computer comprising a plurality of processing units each having instruction storage for holding a local program, an execution unit for executing the local program and data storage for holding data, a switching fabric connected to an output interface of each processing unit and connectable to an input interface of each processing unit by switching circuitry controllable by each processing unit, and a synchronisation module operable to generate a synchronisation signal, the method comprising:

generating a local program for each processing unit comprising a sequence of executable instructions;

determining for each processing unit a relative time of execution of instructions of each local program whereby a local program allocated to one processing unit is scheduled to execute with a predetermined delay relative to a synchronisation signal a send instruction to transmit at least one data packet at a predetermined transmit time, relative to the synchronisation signal, destined for a recipient processing unit but having no destination identifier, and a local program allocated to the recipient processing unit is scheduled to execute at a predetermined switch time a switch control instruction to control the switching circuitry to connect its processing unit wire to the switching fabric to receive the data packet at a receive time.

Google (Cloud Computing) – Example Circuit Claim

- App. No. 15/389,202 (Patent No. 9,710,748, issued 7/18/2017)

  Title: Neural Network Processor; Filed 12/22/2016; 1 Office Action (Art Unit 2124)

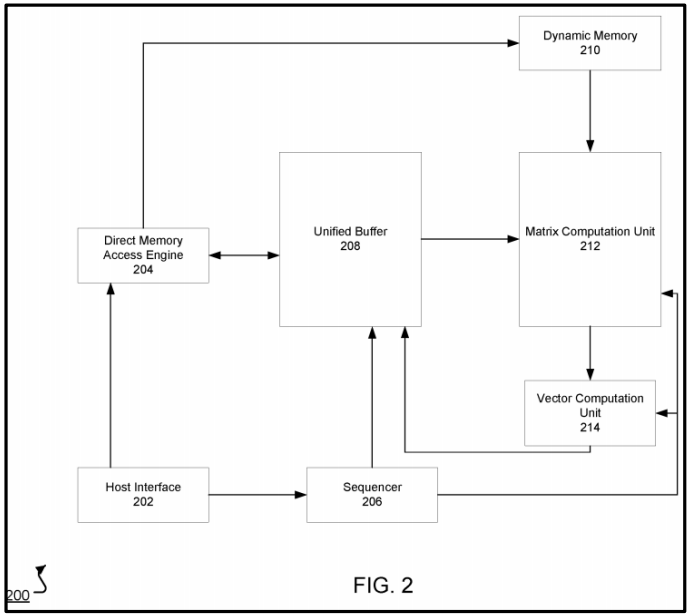


FIG. 2

1. (Currently Amended) A circuit for performing neural network computations for a neural network comprising a plurality of neural network layers, the circuit comprising:

a matrix computation unit configured to, for each of the plurality of neural network layers:

receive a plurality of weight inputs and a plurality of activation inputs for the neural network layer, and

generate a plurality of accumulated values based on the plurality of weight inputs and the plurality of activation inputs,

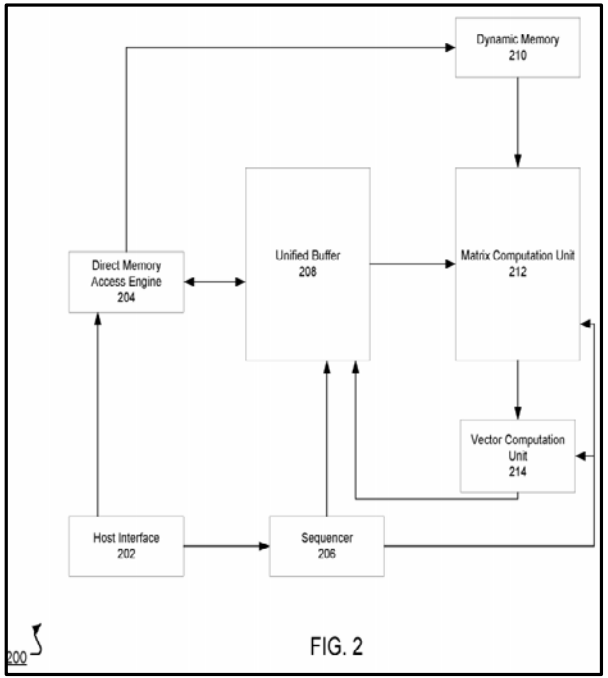
wherein the matrix computation unit is configured as a two dimensional systolic array comprising a plurality of cells, wherein the plurality of weight inputs is shifted through a first plurality of cells along a first dimension of the systolic array, and wherein the plurality of activation inputs is shifted through a second plurality of cells along a second dimension of the systolic array; and

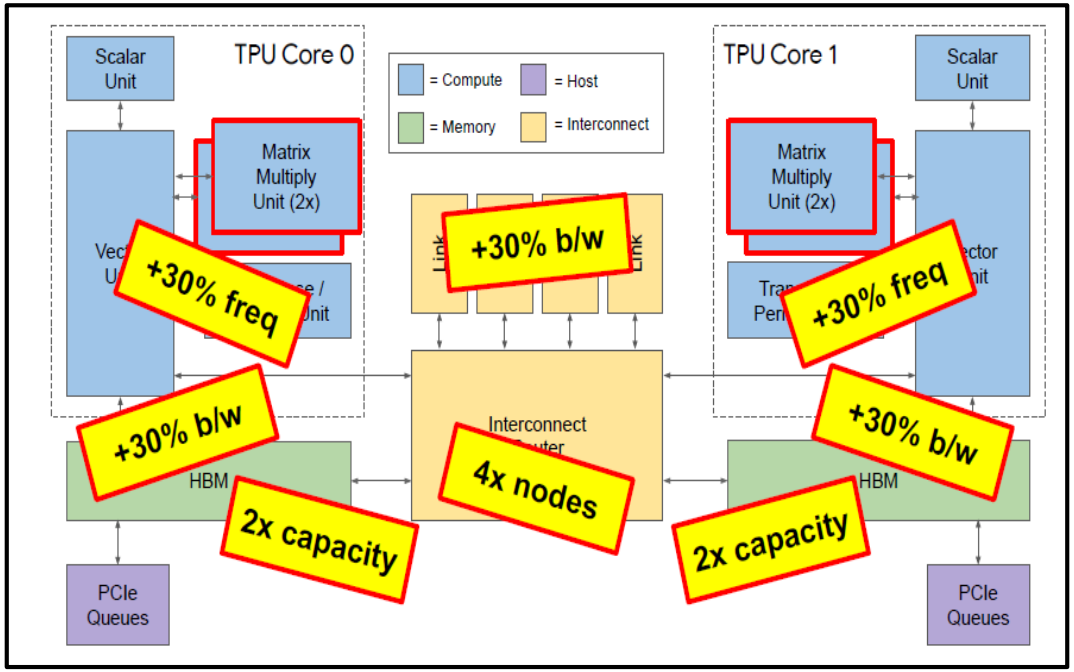
a vector computation unit communicatively coupled to the matrix computation unit and configured to, for each of the plurality of neural network layers:

apply an activation function to each of the plurality of accumulated ~~value~~ values for the neural network layer generated by the matrix computation unit to generate a plurality of activated values for the neural network layer.

Case Study: Google (Cloud Computing)

Google Tensor Processing Unit TPUv3





Source: Google (HotChips 2020)

Case Study: Facebook (Datacenter)

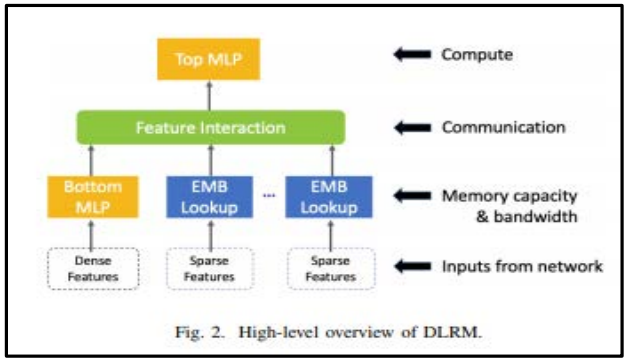


Fig. 2. High-level overview of DLRM.

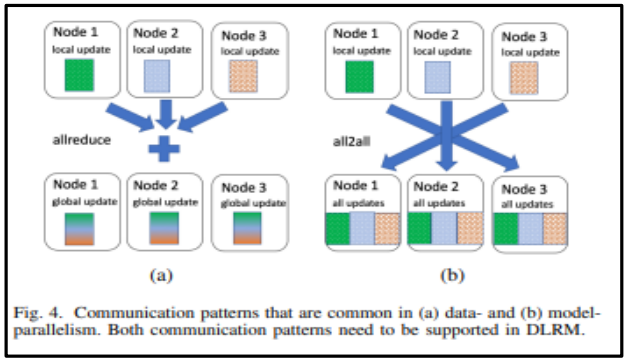


Fig. 4. Communication patterns that are common in (a) data- and (b) model-parallelism. Both communication patterns need to be supported in DLRM.

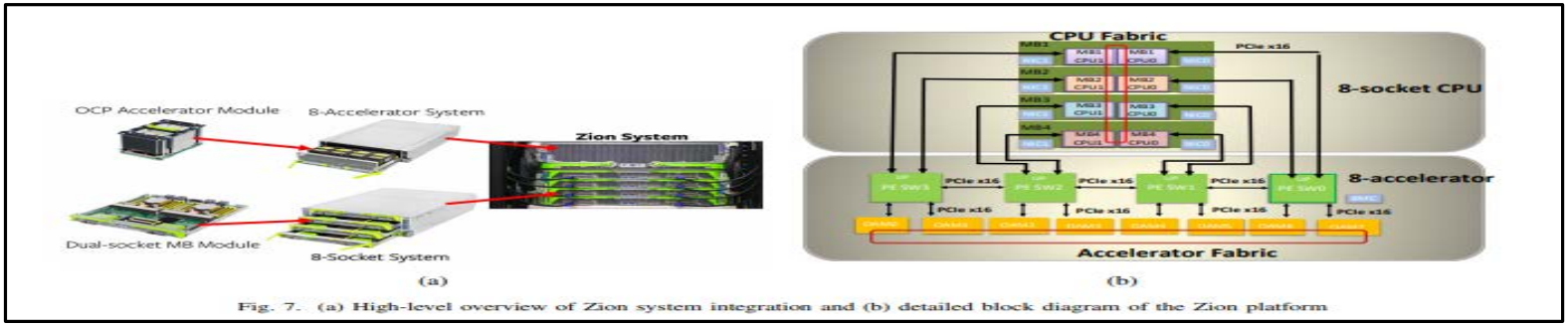


Fig. 7. (a) High-level overview of Zion system integration and (b) detailed block diagram of the Zion platform

Source: Deep Learning Training in Facebook Data Centers: Design of Scale-up and Scale-out Systems

Case Study: Facebook (Datacenter)

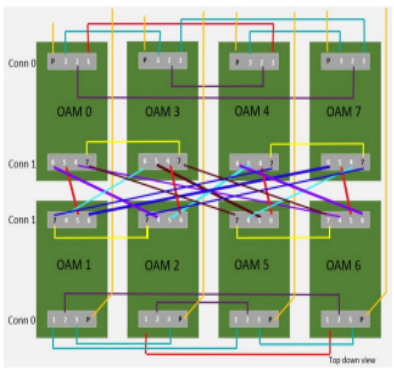


Fig. 8. Accelerator fabric interconnecting layout in Zion.

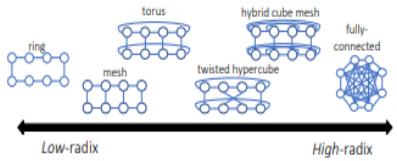


Fig. 9. Different topology design space for an 8-node system.

TABLE II
HIGH-LEVEL COMPARISON OF INTERCONNECTION NETWORKS.

	High Performance Interconnect	Accelerator Fabric
Topology	low-diameter, high (bisection) bandwidth	high (node) bandwidth
Routing	adaptive routing	deterministic routing
Flow Control	cut-through	store & forward
Fabric Design	router centric	node centric

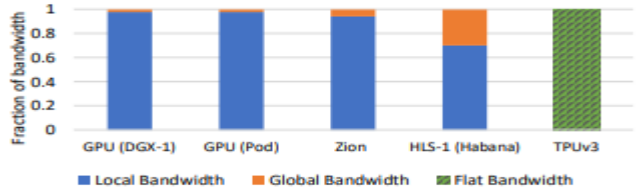


Fig. 12. Comparison of local vs global bandwidth for different scale-out systems. The TPU system does not differentiate between local and global bandwidth since it is flat topology.

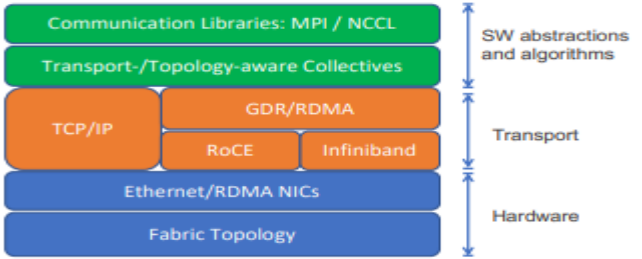


Fig. 13. A view of interconnect software and hardware stack

Source: Deep Learning Training in Facebook Data Centers: Design of Scale-up and Scale-out Systems

AMD + Xilinx (FPGA) – Own Prior Art?

- App. No. 16/451,804 (pending)

  Title: Method and apparatus for efficient programmable instructions in computer systems;

  Filed: 06/25/2019 (1 Office Action)

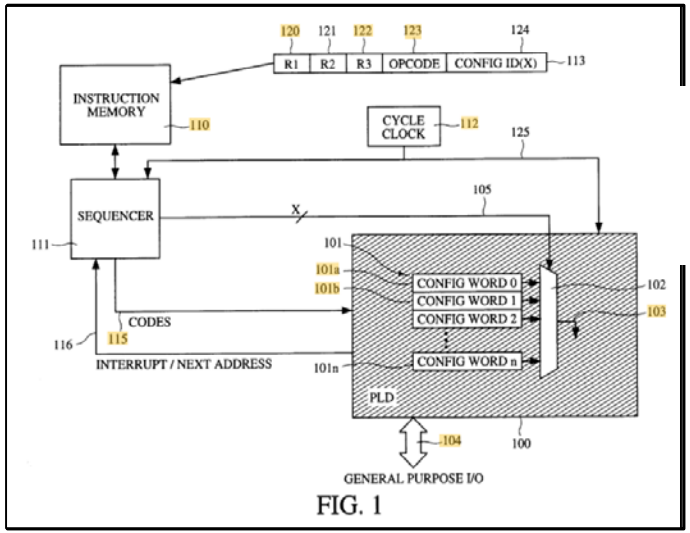


FIG. 1

==

??

In a common implementation, a FPGA on a peripheral component interconnect express (PCIe) slot is coupled to a processor via the PCIe bus. Sending computation to the FPGA over the PCIe bus is a long-latency event, with routine accesses to the device going through a device driver and potentially taking hundreds of cycles. When the FPGA is finished with the computation, the FPGA typically generates an interrupt, and handling the interrupt can cause additional delay. Accordingly, more efficient ways of performing computations using programmable logic devices are

1. A processor comprising:

a first programmable execution unit;

a dispatch unit; and

a memory;

wherein the processor is configured to:

load a first program of an application into the memory;

detect a bitfile portion of the first program;

program the first programmable execution unit with the bitfile portion of the first program;

program the dispatch unit to map a first set of specialized instructions to the first programmable execution unit; and

during execution of the first program, dispatch any specialized instruction of the first set to the first programmable execution unit for execution.

Case Study: Brainchip (Neuromorphic Computing)

Source: PatentAdvisor

BRAINCHIP

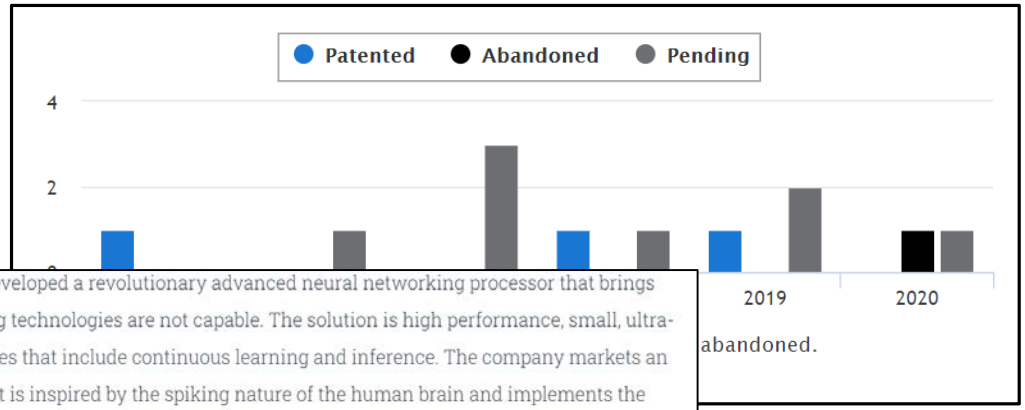
ALLOWANCE RATE

75%

TOTAL NUMBER OF APPLICATIONS

12

TOTAL



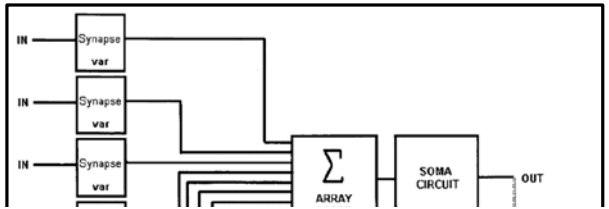
BrainChip is a global technology company that has developed a revolutionary advanced neural networking processor that brings artificial intelligence to the edge in a way that existing technologies are not capable. The solution is high performance, small, ultra-low power and enables a wide array of edge capabilities that include continuous learning and inference. The company markets an innovative event-based neural network processor that is inspired by the spiking nature of the human brain and implements the network processor in an industry standard digital process. By mimicking brain processing BrainChip has pioneered an event domain neural network processor, called Akida™, which is both scalable and flexible to address the requirements in edge devices. At the edge, sensor inputs are analyzed at the point of acquisition rather than transmission to the cloud or a datacenter. The Akida neural processor is designed to provide a complete ultra-low power Edge AI network processor for vision, audio and smart transducer applications. The reduction in system latency provides faster response and a more power efficient system that can help reduce the large carbon footprint of datacenters.

Source: Brainchip

Brainchip (Neuromorphic Computing) – Non-Patent Literature Prior Art

- App. No. 12/243,697 (Patent No. 8,250,011, issued 8/21/2012)

  Title: AUTONOMOUS LEARNING DYNAMIC ARTIFICIAL NEURAL COMPUTING DEVICE AND BRAIN INSPIRED SYSTEM; Filed 2/1/2018; 1 Office Action (Art Unit 2122)



IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 17, NO. 1, JANUARY 2006 211

A VLSI Array of Low-Power Spiking Neurons and Bistable Synapses With Spike-Timing Dependent Plasticity

Giacomo Indiveri, *Member, IEEE*, Elisabetta Chicca, and Rodney Douglas

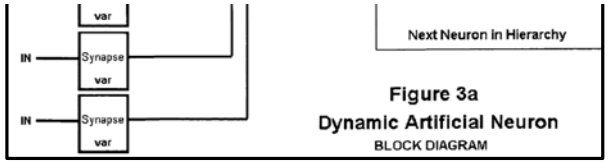


Figure 3a
Dynamic Artificial Neuron
BLOCK DIAGRAM

1. (currently amended) An information processing system intended for use in artificial ~~intelligence, consisting of~~ intelligence and having a plurality of digital artificial neuron circuits connected in an array, the system comprising

a ~~first~~ plurality of digital dynamic synapse circuits, wherein each digital dynamic synapse circuit contains a binary register that stores a value representing neurotransmitter type and level, wherein the digital dynamic synapse circuits comprise ~~comprising~~ a means of learning and responding to input signals signals, either by producing [[a]] or compounding ~~strength value~~ the value, thereby simulating behavior of a biological synapse; and ~~a biological Post Synaptic Potential,~~

a temporal integrator circuit that integrates and combines each individually simulated synapse neurotransmitter type and value ~~Post Synaptic Potential values~~ over time, wherein time is dependent on the neurotransmitter type stored in each digital dynamic synapse circuit. ~~and thus constitutes an artificial membrane potential value,~~

~~a second plurality of dynamic soma circuits each capable of producing one or more pulses when the integrated membrane potential value has reached or exceeded a stored variable threshold value,~~

Case Study: Luminous Computing (Optical Computing)

LUMINOUS COMPUTING

ALLOWANCE RATE

100%

TOTAL NUMBER OF APPLICATIONS

5

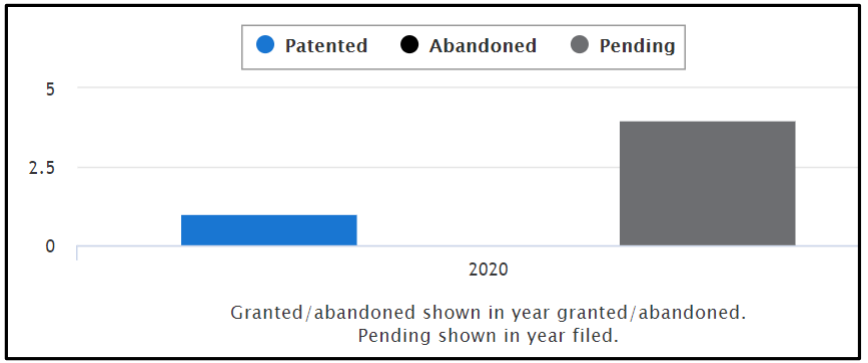
TOTAL

1 PATENTED APPLICATIONS

0 ABANDONED APPLICATIONS

4 PENDING APPLICATIONS

Source: PatentAdvisor



Computing / Microchips

Bill Gates just backed a chip startup that uses light to turbocharge AI

Luminous Computing has developed an optical microchip that runs AI models much faster than other semiconductors while using less power.

by **Martin Giles** June 13, 2019

Optical solution

Luminous sees light as the answer. It uses lasers to beam light through tiny structures on its chip, known as waveguides. By using different colors of light to move multiple pieces of data through waveguides at the same time, it can outstrip the data-carrying capabilities of conventional electrical chips.

The ability to transport very large amounts of information swiftly means optical processors are ideally suited to handling the vast number of computations that drive AI models. They also require far less power than electrical ones.

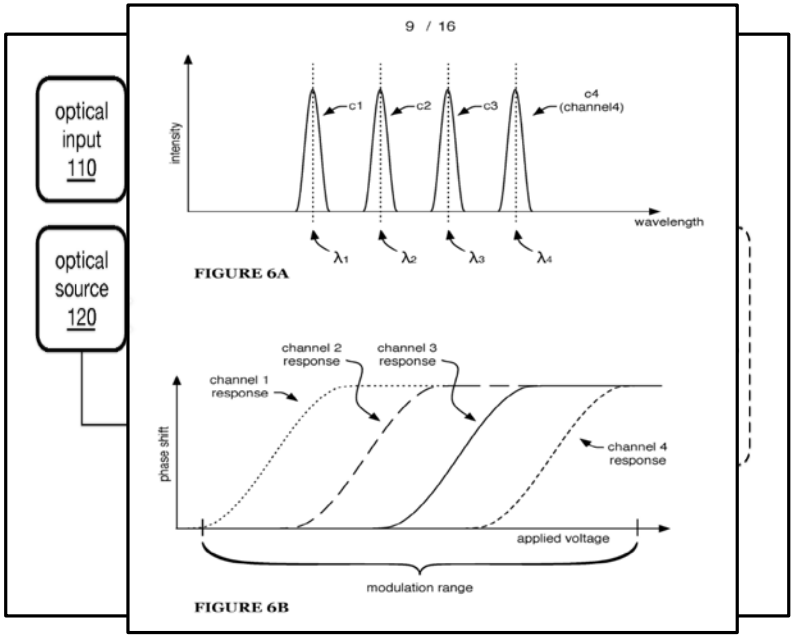
Source: MIT Technology Review (June 2019)

Luminous Computing (Optical Computing) – Method Claim for Analog-to-Digital Conversion

- App. No. 16/826,008 (Patent No. 10,837,827, issued 11/17/2020)

  Title: System and Method for Photonic Analaog-to-Digital Conversion;

  Filed 03/20/2020; No Office Action (Art Unit 2845)

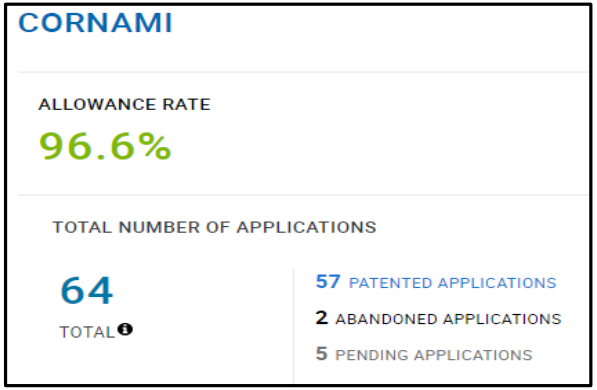


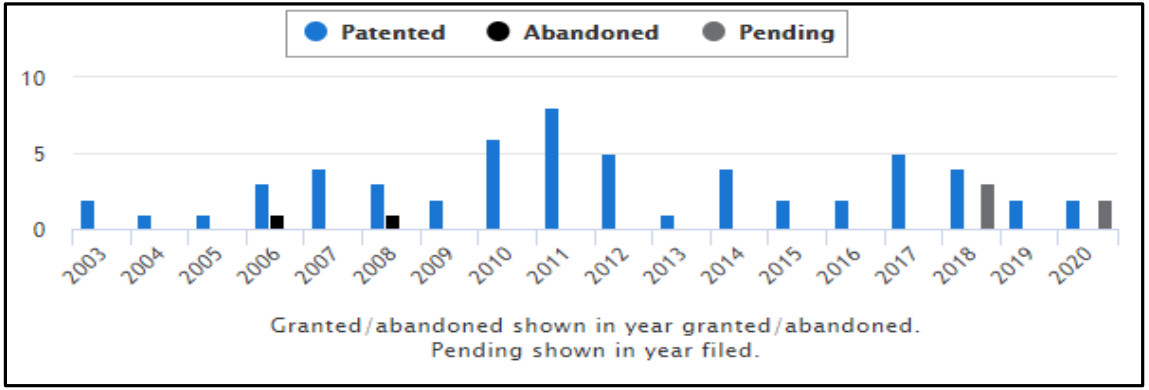
1. A method for analog-to-digital conversion, the method comprising:
   - receiving an analog input signal and a first optical carrier;
   - based on the analog input signal, modulating a phase of the first optical carrier to generate a phase-modulated optical signal;
   - at a photonic circuit, receiving the phase-modulated optical signal and a second optical carrier;
   - at the photonic circuit, generating a spatially-separated plurality of optical outputs based on the phase-modulated optical signal, comprising, at a coupler of the photonic circuit, interfering the phase-modulated optical signal with the second optical carrier;
   - at a detector bank comprising a plurality of detectors, receiving the spatially-separated plurality of optical outputs, wherein each optical output of the spatially-separated plurality of optical outputs is received by a different detector of the plurality of detectors; and
   - at the detector bank, generating a set of binary outputs, comprising, for each optical output of the spatially-separated plurality of optical outputs: generating, based on the optical output, a respective binary output of the set;

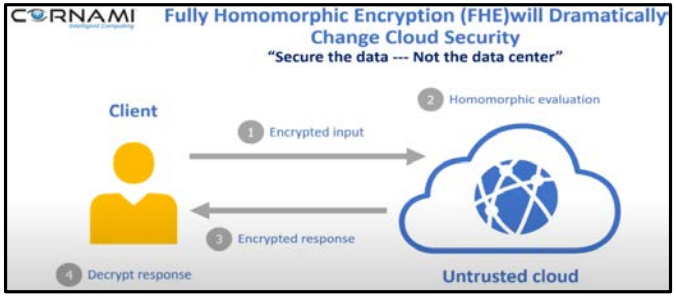
wherein the set of binary outputs is indicative of a value associated with the analog input signal.

Case Study: Cornami (Fully Homomorphic Encryption)

Source: PatentAdvisor







Source: Cornami

Homomorphic Encryption: The 'Golden Age' of Cryptography

The ability to perform complex calculations on encrypted data promises a new level of privacy and data security for companies in the public and private sectors. So when can they get started?

Source: Darkreading.com

Cornami (Fully Homomorphic Encryption) – How to Draft Claims to Avoid 3600 Art Unit ?

- App. No. 16/743,257 (Pending)

  Title: Method and Apparatus for Configuring a Reduced Instruction Set Computer Processor Architecture to Execute a Fully Homomorphic Encryption Algorithm

  Filed 01/15/2020 (Art Unit **3600!**)

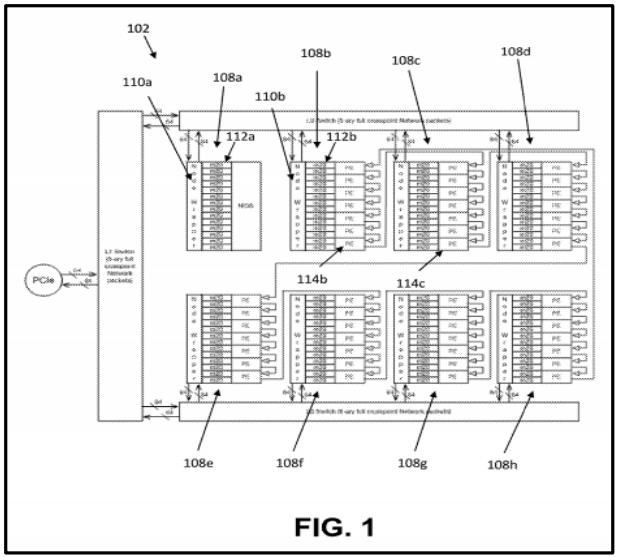


FIG. 1

1. A method for configuring a reduced instruction set computer processor architecture to process a Discrete Fourier Transform (DFT) of a finite-length sequence N, wherein the computer processor architecture includes a plurality of primary processing cores defined by RISC processors, each primary processing core comprising a main memory, at least one cache memory, and a plurality of arithmetic logic units, each primary core having an associated node wrapper, the node wrapper including access memory associated with each arithmetic logic unit, a load/unload matrix associated with each arithmetic logic unit, the method comprising:

(a) applying a Decimation-in-Frequency algorithm to the DFT to decompose the DFT of a finite-length sequence N into two derived DFTs each of a length N/2;

(b) constructing a logic element equivalent of each stage of the derived DFTs in which inputs and outputs are composed of real and imaginary components;

(c) repeating (a) and (b) for each stage of the DFT except for the endpoint stages of the DFT;

(d) for each endpoint stage of the DFT constructing a logic element equivalent of the corresponding stage of the derived DFTs in which inputs and outputs are composed of only real components;

(e) configuring at least one primary core of the computer processor architecture to implement the logic element equivalents of each stage of the DFTs in a manner which operates in a streaming mode wherein data streams out of corresponding arithmetic logic units into the main memory and other ones of the plurality arithmetic logic units; and

(f) configuring the computer processor architecture to couple the output of each stage on the DFT to the input of a subsequent stage.

Case Study: Mythic (In-Memory Computation)

Source: PatentAdvisor

MYTHIC

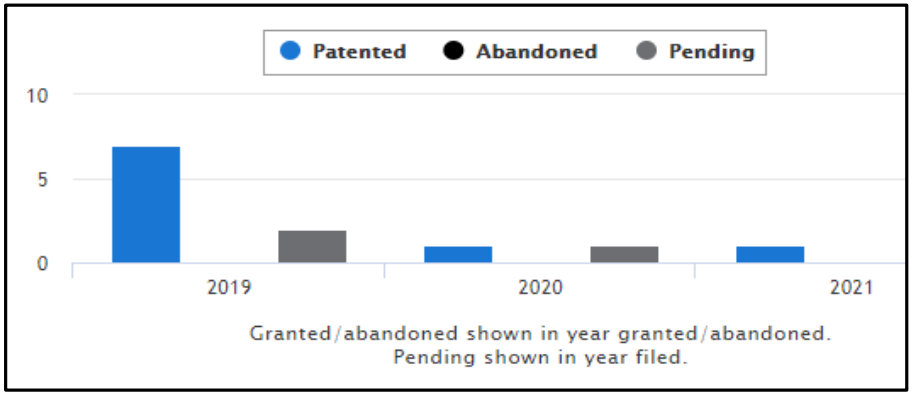
ALLOWANCE RATE

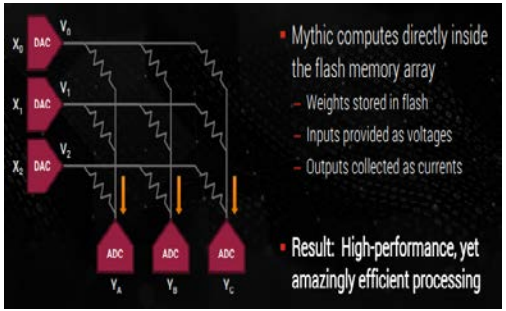
100%

TOTAL NUMBER OF APPLICATIONS

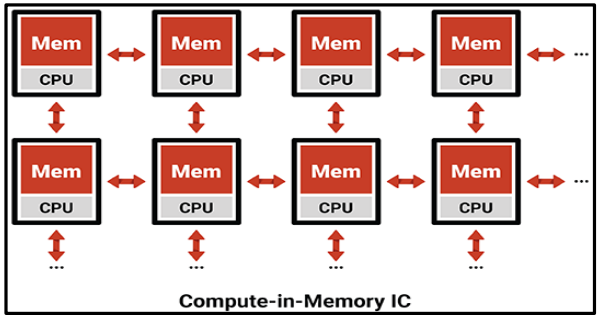
12
TOTAL

9 PATENTED APPLICATIONS
0 ABANDONED APPLICATIONS
3 PENDING APPLICATIONS





Mythic products are based on a unique tile-based AI compute architecture that features three fundamental hardware technologies – Compute-in-Memory, Dataflow Architecture, and Analog Computing. For AI developers, the Mythic SDK streamlines the preparation of trained neural networks for edge and low-latency datacenter deployments, and also performs automatic optimization and compilation of dataflow graphs for our unique architecture.



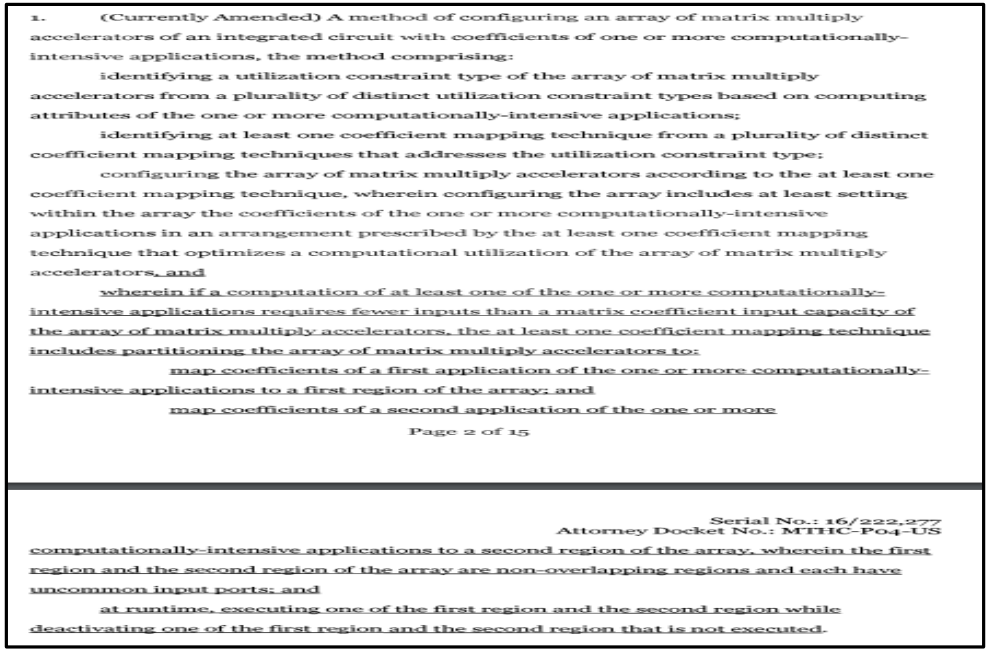
Source: Mythic-ai.com

Mythic (Analog Compute-in-Memory) – Method Claim to Map Compute Close to Data

- App. No. 16/222,277 (Patent No. 10,409,889, issued 09/10/2019)

  Title: SYSTEMS AND METHODS FOR MAPPING MATRIX CALCULATIONS TO A MATRIX MULTIPLY ACCELERATOR

  Filed 12/17/2018 (Art Unit 2182)



1. (Currently Amended) A method of configuring an array of matrix multiply accelerators of an integrated circuit with coefficients of one or more computationally-intensive applications, the method comprising:

identifying a utilization constraint type of the array of matrix multiply accelerators from a plurality of distinct utilization constraint types based on computing attributes of the one or more computationally-intensive applications;

identifying at least one coefficient mapping technique from a plurality of distinct coefficient mapping techniques that addresses the utilization constraint type;

configuring the array of matrix multiply accelerators according to the at least one coefficient mapping technique, wherein configuring the array includes at least setting within the array the coefficients of the one or more computationally-intensive applications in an arrangement prescribed by the at least one coefficient mapping technique that optimizes a computational utilization of the array of matrix multiply accelerators, and

wherein if a computation of at least one of the one or more computationally-intensive applications requires fewer inputs than a matrix coefficient input capacity of the array of matrix multiply accelerators, the at least one coefficient mapping technique includes partitioning the array of matrix multiply accelerators to:

map coefficients of a first application of the one or more computationally-intensive applications to a first region of the array; and

map coefficients of a second application of the one or more

Page 2 of 15

Serial No.: 16/222,277
Attorney Docket No.: MTHC-P04-US

computationally-intensive applications to a second region of the array, wherein the first region and the second region of the array are non-overlapping regions and each have uncommon input ports; and

at runtime, executing one of the first region and the second region while deactivating one of the first region and the second region that is not executed.

Mythic (Analog Compute-in-Memory) – Example Claim for Mixed-Signal Integrated Circuit

- App. No. 16/127,488 (Patent No. 10,409,889, issued 09/10/2019)

  Title: SYSTEMS AND METHODS FOR MIXED-SIGNAL COMPUTING

  Filed 9/11/2018 (Art Unit 2122)

1. A mixed-signal integrated circuit comprising:

a reference signal source that generates a plurality of analog reference signals based on digital input,

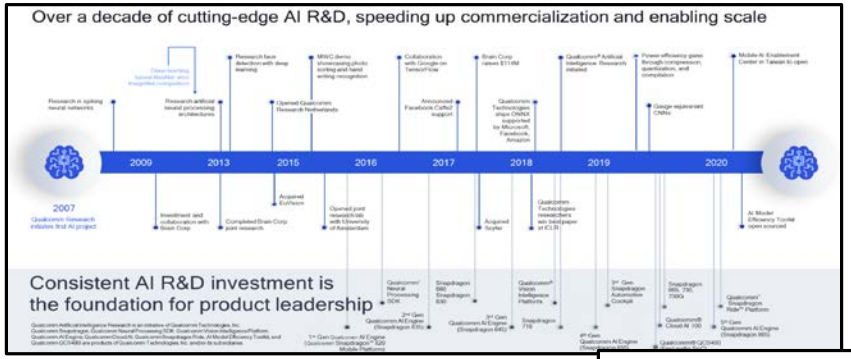
wherein an output terminal of the reference signal source is electrically connected to a shared signal path, and wherein the reference signal source sources the plurality of analog reference signals to the shared signal path;

a plurality of local signal accumulators arranged along the shared signal path and each of the plurality of local signal accumulators having an input terminal electrically connected to the shared signal path, wherein each of the plurality of local signal accumulators:

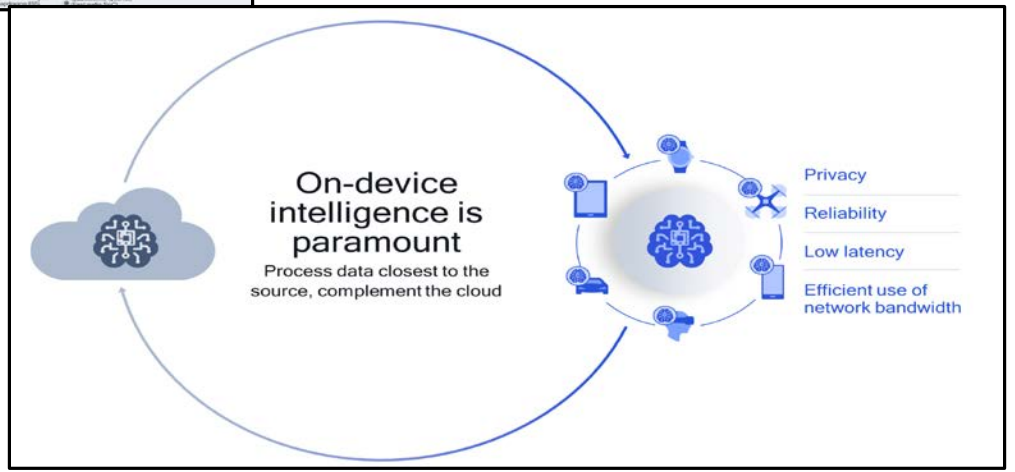
collects, via the shared signal path, the plurality of analog reference signals from the reference signal source; and

stores a sum of the plurality of electrical charges over a predetermined number of clock cycles.

Case Study: Qualcomm (Inference)



Source: Qualcomm



Case Study: Qualcomm (Inference) – A Very Short But Broad Claim

- App. No. 16/556,094 (Notice of Allowance, dated 1/15/2021)

  Title: METHOD, APPARATUS, AND SYSTEM FOR AN ARCHITECTURE FOR MACHINE LEARNING ACCELERATION

  Filed 08/29/2019 (Art Unit 2184)

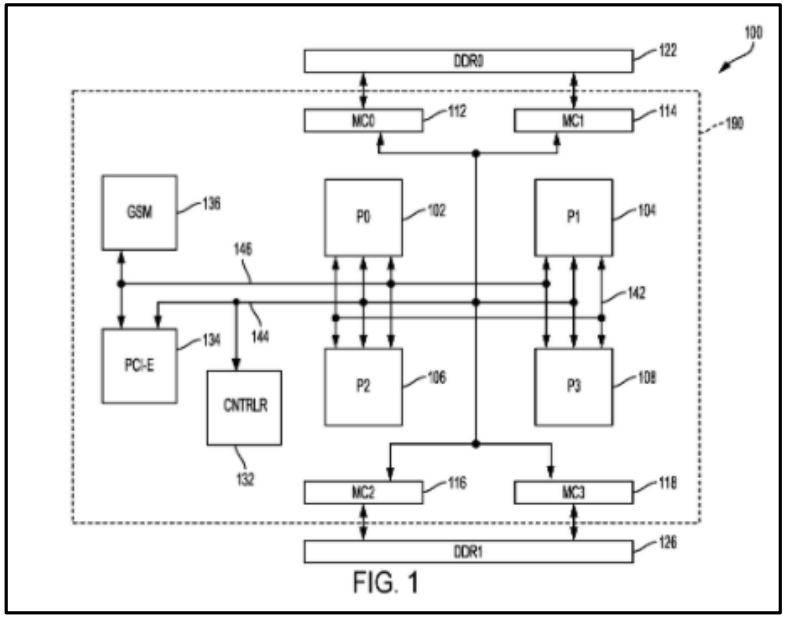


FIG. 1

1. (Currently amended) An inference accelerator comprising:

a memory system;

a plurality of processing elements, each processing element:

having a corresponding tightly coupled memory (TCM);

coupled to the memory system; and

adapted to access the memory system; and

a global synchronization manager (GSM) module coupled to the plurality of processing elements and to the memory system, the GSM adapted to synchronize operations of the plurality of processing elements and memory system using corresponding synchronization modules of each of the plurality of processing elements.

Case Study: Cambricon (FP Conversion)

CAMBRICON TECHNOLOGIES

ALLOWANCE RATE

84.8%

TOTAL NUMBER OF APPLICATIONS

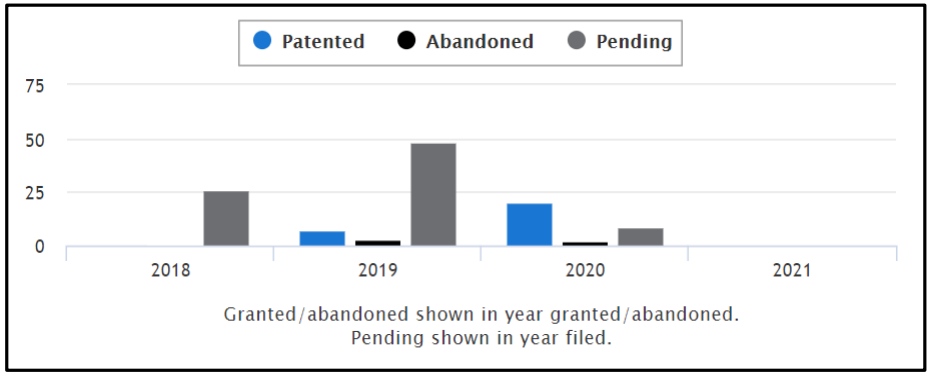
116

TOTAL

28 PATENTED APPLICATIONS

5 ABANDONED APPLICATIONS

83 PENDING APPLICATIONS



Source: PatentAdvisor

Case Study: Cambricon (FP Conversion) – Example of A Successful Math Claim With Circuit

- App. No. 16/508,139 (Patent No. 10,726,336, issued 07/28/2020)

  Title: Apparatus And Method for Compression Coding for Artificial Neural Network

  Filed 7/10/2019 (Art Unit 2123)

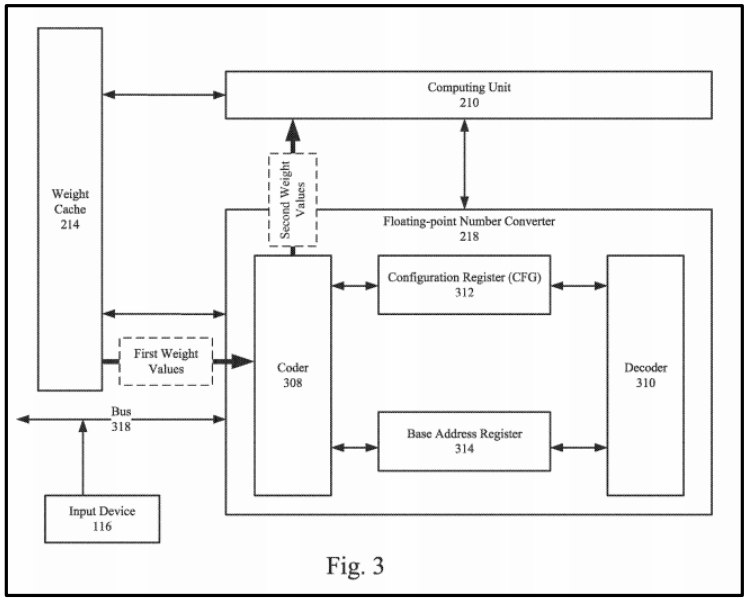


Fig. 3

1. (Currently Amended) A neural network processor, comprising:

a floating-point number converter configured to:

receive one or more first weight values of a first bit length and first input neuron data, and

convert the one or more first weight values to one or more second weight values of a second bit length,

wherein the second bit length is less than the first bit length,

wherein each of the first weight values includes a first sign bit, a first exponent field, and a first mantissa field, and

wherein each of the second weight values includes a second sign bit, a second exponent field, and a second mantissa field, wherein a bit length of the second mantissa field is less than a bit length of the first mantissa field; and

a computing unit configured to:

receive the first input neuron data, and

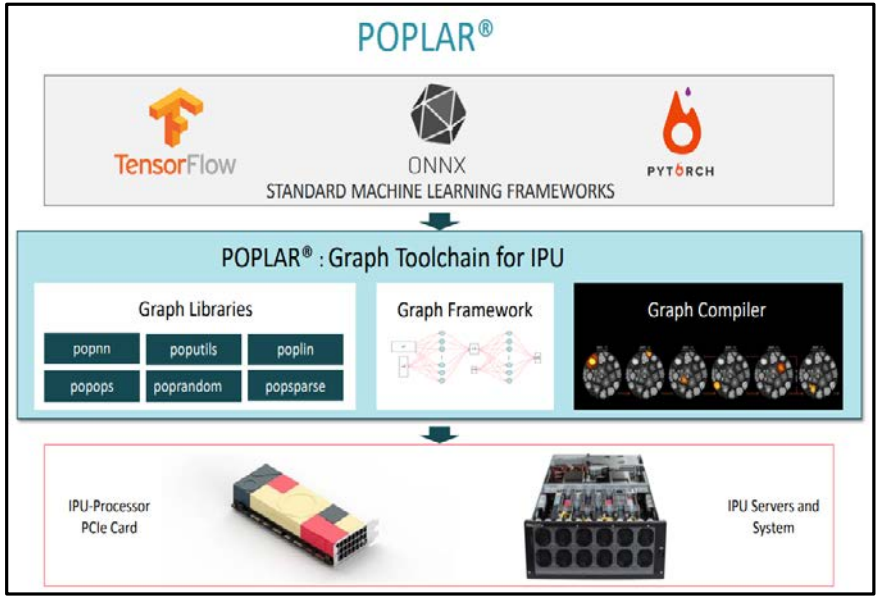
calculate first output neuron data based on the first input neuron and the second weight values.

Part 3: Other Topics

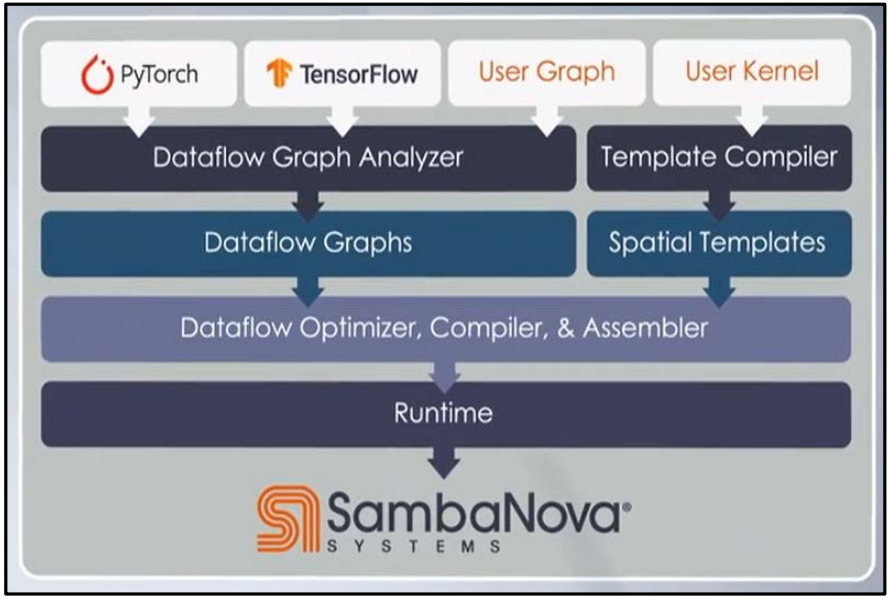
Agenda

- IP Legal Strategy Has to Include Open-Source Software / Hardware
- Trade Secret Protection – Legal Considerations
- Defensive Publications Can Complement Patents
- AI Hardware-Related Patent Litigation (Patent Infringement and All-Elements Rule)
- AI Hardware Startups-Related Lawsuits & Proceedings
- Conclusion

IP Legal Strategy Has to Include Open Source

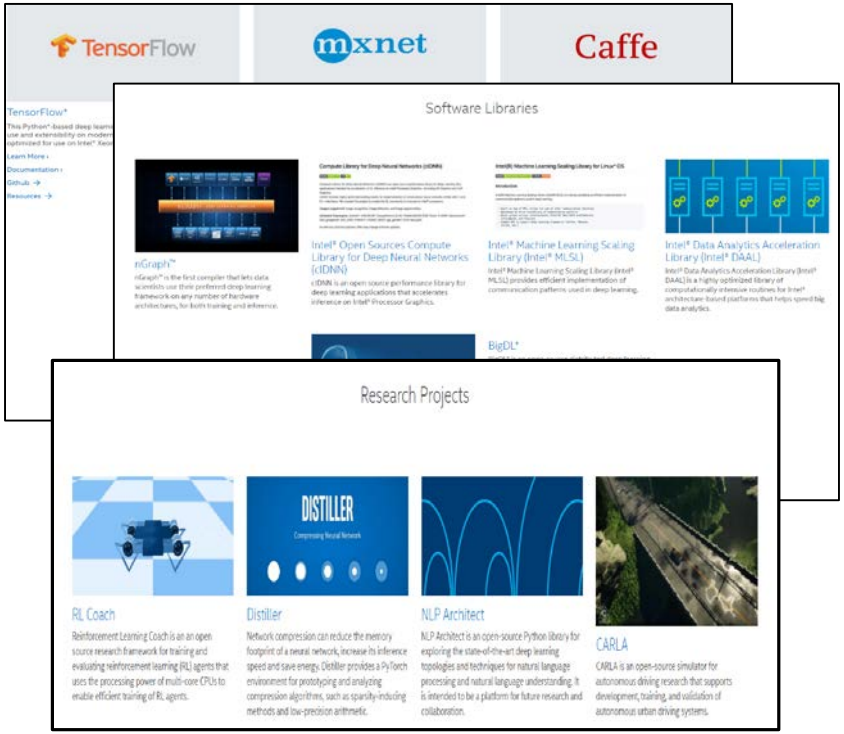


Source: GraphCore (Stanford 2018)



Source: SambaNova Systems

IP Legal Strategy Has to Include Open Source



Source: Intel

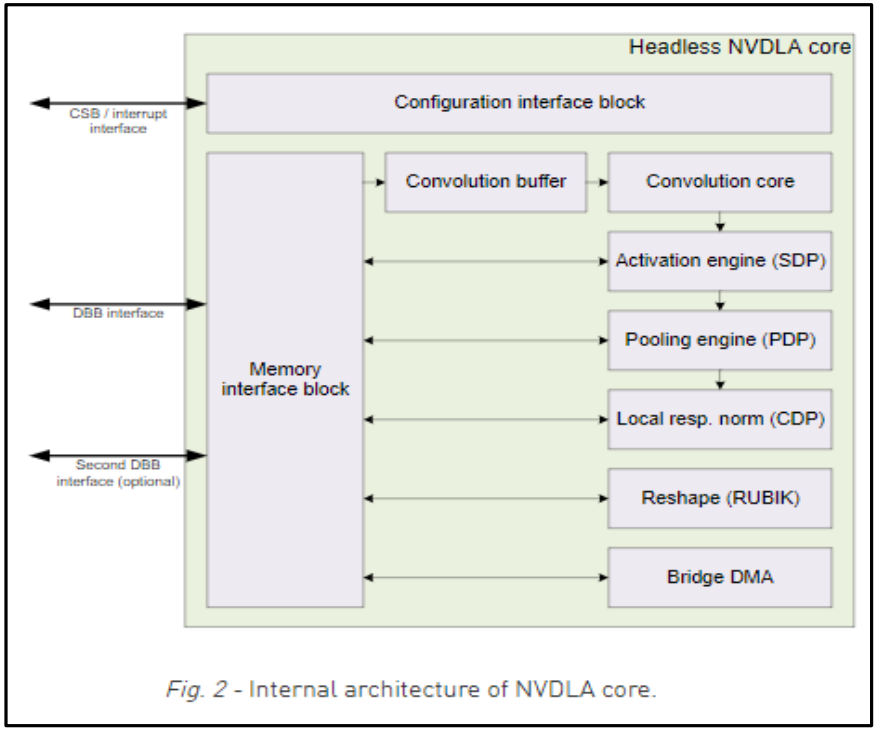
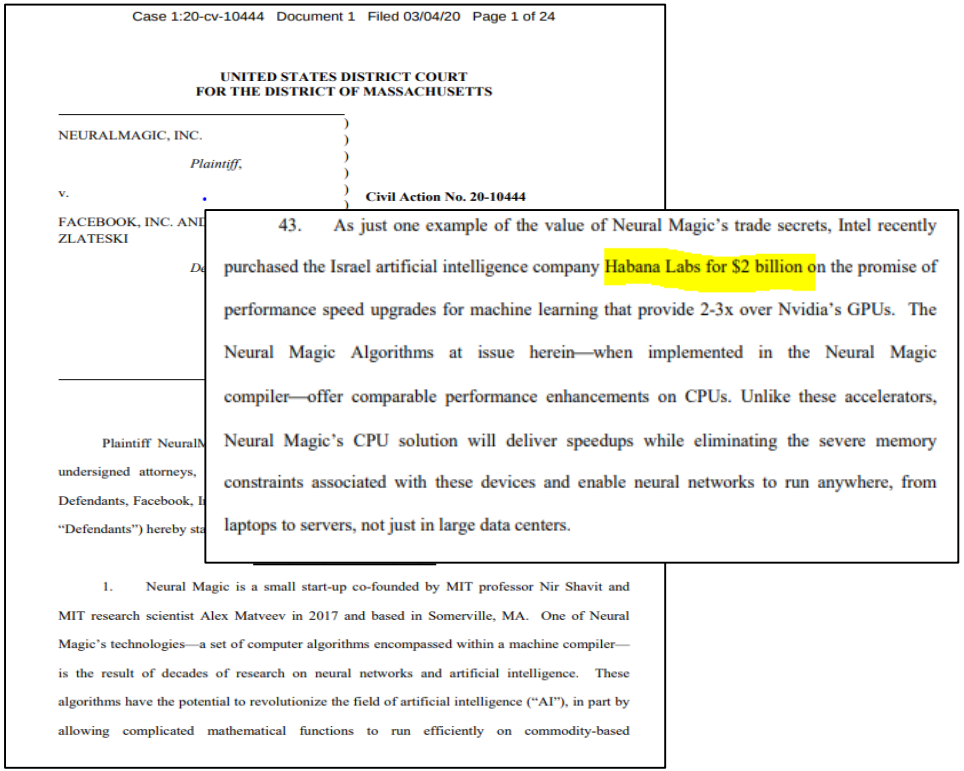


Fig. 2 - Internal architecture of NVDLA core.

Source: Nvidia

Trade Secret Protection – Legal Considerations



Case 1:20-cv-10444 Document 1 Filed 03/04/20 Page 1 of 24

UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF MASSACHUSETTS

NEURALMAGIC, INC.

Plaintiff,

v.

Civil Action No. 20-10444

FACEBOOK, INC. ANI
ZLATESKI

De

Plaintiff NeuralM
undersigned attorneys,
Defendants, Facebook, I
"Defendants") hereby sta

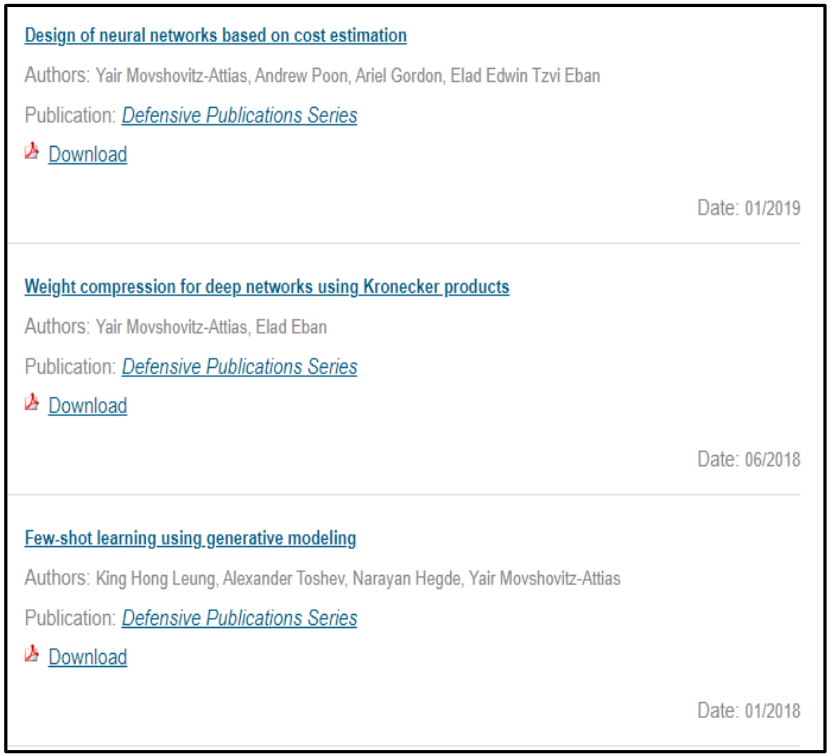
1. Neural Magic is a small start-up co-founded by MIT professor Nir Shavit and MIT research scientist Alex Matveev in 2017 and based in Somerville, MA. One of Neural Magic's technologies—a set of computer algorithms encompassed within a machine compiler—is the result of decades of research on neural networks and artificial intelligence. These algorithms have the potential to revolutionize the field of artificial intelligence ("AI"), in part by allowing complicated mathematical functions to run efficiently on commodity-based

43. As just one example of the value of Neural Magic's trade secrets, Intel recently purchased the Israel artificial intelligence company Habana Labs for $2 billion on the promise of performance speed upgrades for machine learning that provide 2-3x over Nvidia's GPUs. The Neural Magic Algorithms at issue herein—when implemented in the Neural Magic compiler—offer comparable performance enhancements on CPUs. Unlike these accelerators, Neural Magic's CPU solution will deliver speedups while eliminating the severe memory constraints associated with these devices and enable neural networks to run anywhere, from laptops to servers, not just in large data centers.

- Are reasonable steps taken to prevent disclosure?
- Does invention have potential/actual economic value
- Is invention generally known, or easily reverse engineered?
- Is invention "readily ascertainable"?
- How fast is technology changing?

Defensive Publications Can Complement Patents

- A publication of a disclosure that provides defensive benefits, such as the creation of prior art against others as of the publication date.
- Takes many forms
  (informal / self-published / formal)



Design of neural networks based on cost estimation

Authors: Yair Movshovitz-Attias, Andrew Poon, Ariel Gordon, Elad Edwin Tzvi Eban

Publication: *Defensive Publications Series*

Download

Date: 01/2019

Weight compression for deep networks using Kronecker products

Authors: Yair Movshovitz-Attias, Elad Eban

Publication: *Defensive Publications Series*

Download

Date: 06/2018

Few-shot learning using generative modeling

Authors: King Hong Leung, Alexander Toshev, Narayan Hegde, Yair Movshovitz-Attias

Publication: *Defensive Publications Series*

Download

Date: 01/2018

Source: tdcommons.org

AI Hardware-Related Patent Litigation – Legal Hurdles (Example case: ACS v. NVIDIA)

IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF DELAWARE

ADVANCED CLUSTER SYSTEMS, INC., Plaintiff, v. NVIDIA CORPORATION, Defendant.	Civil Action No. **DEMAND FOR JURY TRIAL**

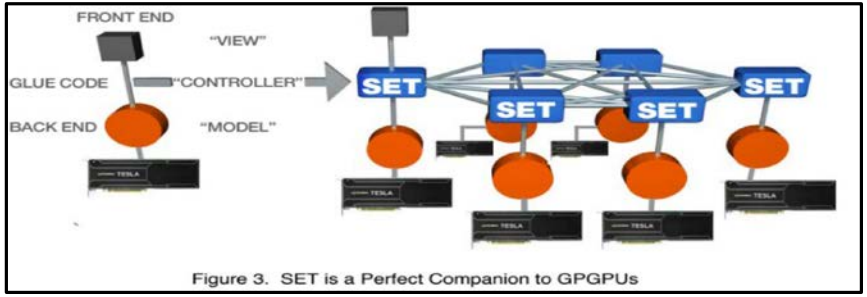


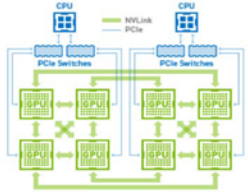
Figure 3. SET is a Perfect Companion to GPGPUs

Patents-in-Suit

- United States Patent No. 8,082,289 titled "Cluster Computing Support for Application Programs"
- United States Patent No. 8,140,612 titled "Cluster Computing Support for Application Programs"
- United States Patent No. 8,676,877 titled "Cluster Computing Using Special Purpose Microprocessors"

38. ACS is informed and believes, and thereon alleges that NVLink is an interconnect architecture that facilitates data and control transmission between multiple GPUs and Central Processing Units ("CPUs"), combined together in a hybrid cube mesh as shown below.

NVIDIA® NVLink™ Hybrid Cube Mesh



In general terms, the NVLink architecture implements the same GPU-to-GPU communication architecture claimed in the Patents-in-Suit. The details are set forth in the exemplary claim charts provided at Exhibits E-H hereto.

Example Claim Chart from Lawsuit (Patent Infringement – All Elements Rule)

Patent-in-Suit: Patent No. 8,082,289 titled "Cluster Computing Support for Application Programs" (Exhibit E, claim chart)

Claim 29 Language	**Alleged NVIDIA's Infringing Products** **(e.g.,** DGX Station, DGX-1, DGX-2, HGX-1, HGX-2, Tesla V100, and Tesla P100)
A method of evaluating a command on a computer cluster comprising:	Each Accused Server Product practices a method of evaluating a command on a computer cluster as follows.
communicating a command from at least one of a user interface or a script to one or more cluster node modules within the computer cluster;	For example, each Accused Server Product includes a CPU or host executing a program (user interface or script) that communicates commands to a cluster of GPU Accelerators, each of which include a cluster node module.
for each of the one or more cluster node modules, communicating a message based on the command to a respective kernel module associated with the cluster node module;	For example, each GPU Accelerator cluster node module communicates a message based on the command received from the CPU or host to a CUDA kernel associated with the cluster node module.
for each of the one or more cluster node modules, receiving a result from the respective kernel Module	For example, each GPU Accelerator cluster node module receives results from the CUDA kernel associated with the cluster node module.
for at least one of the one or more cluster node modules, responding to messages from other cluster node modules.	For example, at least one GPU Accelerator cluster node module responds to messages from other GPU Accelerator cluster node module using the NVLink peer-to-peer communication function.

AI Hardware Startups-Related Lawsuits & Proceedings

Hailo Technologies
- 4:17-cv-05097-PJH Hailo Technologies LLC v. NO NDA, INC.
  - Patent infringement (dismissed voluntarily)
- 2:19-cv-00751-TSZ Hailo Technologies LLC v. Anker Innovations Limited
  - Patent infringement (dismissed voluntarily)
- 2:19-cv-00958-RAJ-BAT Hailo Technologies LLC v. Moovn Technologies LLC
  - Patent infringement (dismissed voluntarily)
- *** many

Recently, Hailo Technologies LLC – after three continuous attempts at suing the manufacturer of Anker Roav SmartCharge Car Kit – sued Best Buy, Target, and Walmart for infringement through the sale of Anker's devices. While this clearly looks like a move made after extreme frustration, things might not go well for these online retailers.

Source: https://www.greyb.com/ecommerce-infringement-through-sales/

Gyrfalcon Technology
- 3:18-cv-06361-JD Synopsys, Inc. v. Gyrfalcon Technology Inc.
  - copyright infringement (dismissed - settled)
  - Allegation: GTI enabled access to Synopsys' EDA software by entering a key code/password provided by Synopsys, but that it did not have authorization from Synopsys.

INSPUR ELECTRONIC INFORMATION INDUSTRY CO. LTD.,
- 2:20-cv-00019-JRG Longhorn HD LLC. v. Inspur Group Co. Ltd.
  - Patent infringement (ongoing)

Wave Computing (Ch. 11)
- Ch. 11 bankruptcy proceedings

HiSilicon Technologies (Huawei)
- Court of Appeals Docket #: 18-1979
  - Huawei Technologies Co., Ltd. v. Samsung Electronics Co., Ltd.
  - Patent infringement
- 3:16-cv-02787-WHO Huawei Technologies, Co, Ltd et al v. Samsung Electronics Co, Ltd.
  - Patent infringement (dismissed-settled)
- 3:17-cv-06451-EMC Cohen v. TSMC North America Corp. et al
  - Patent infringement (dismissed-settled)
- 1:17-cv-00189-RC Cohen v. TSMC North America et al
  - Patent infringement (transfer)
- 4:19-cv-00731-SDJ Vantage Micro LLC v. Huawei Device USA, Inc. et al
  - Patent infringement (dismissed)

Bitmain Technologies
- 1:18-cv-25106-KMW United American Corp. v. Bitmain, Inc. et al
  - antitrust (dismissed lack of jx)
- 2:18-cv-01626-TSZ Bitmain Technologies Ltd. v. Doe
  - 18:1030 Computer Fraud and Abuse Act (dismissed vol.)

General Vision
- 2:13-cv-00915-MCE-CKD Cognimem Technologies, Inc. et al v. Paillet et al
  - Trademark infringement (Lanham Act) - Dismissed - Voluntarily
- 2:13-cv-00915-MCE-CKD Cognimem Technologies, Inc. et al v. Paillet et al
  - Trademark infringement (Lanham Act) - Dismissed - Voluntarily

Conclusion

Startups should ask:

- What’s is your niche?
- What is the SW/HW ecosystem, including open source?
- What is the right IP strategy?
- What is the right patent claim / continuation strategy?

Coronavirus COVID-19 Resources

We have formed a multidisciplinary **Coronavirus/COVID-19 Task Force** to help guide clients through the broad scope of legal issues brought on by this public health challenge.

To help keep you on top of developments as they unfold, we also have launched a resource page on our website at **www.morganlewis.com/topics/coronavirus-covid-19**

If you would like to receive a daily digest of all new updates to the page, please visit the resource page to **subscribe** using the purple "Stay Up to Date" button.

Morgan Lewis

Biography

Kannan Narayanan
Silicon Valley
+1.650.843.7251
kannan.narayanan@morganlewis.com

Drawing on 18 years of R&D experience in the technology industry and a background in computer science and engineering, Kannan Narayanan works with clients to build strong patent portfolios, preparing and prosecuting US and foreign patents, performing patent due diligence, and providing non-infringement and invalidity opinions and freedom to operate in a variety of technology areas, including artificial intelligence (AI), natural language processing, data visualization, computer architecture, robotic process automation, genetic programming, cloud computing, social networking, wireless power transmission, fraud detection, semiconductor device manufacturing, computer networking, additive manufacturing, image processing, medical and healthcare related technologies, and consumer products.

Biography

Andrew J. Gray IV
Silicon Valley
+1.650.843.7575
andrew.gray@morganlewis.com

Serving as the leader of Morgan Lewis's semiconductor practice and as a member of the firm's fintech and technology practices, Andrew J. Gray IV concentrates his practice on intellectual property (IP) litigation and prosecution and on strategic IP counseling. Andrew advises both established companies and startups on Blockchain, cryptocurrency, computer, and Internet law issues, financing and transactional matters that involve technology firms, and the sale and licensing of technology. He represents clients in patent, trademark, copyright, and trade secret cases before state and federal trial and appellate courts throughout the United States, before the US Patent and Trademark Office's Patent Trial and Appeal Board, and before the US International Trade Commission.

Our Global Reach

Africa
Asia Pacific
Europe
Latin America
Middle East
North America

Our Locations

Abu Dhabi
Almaty
Beijing*
Boston
Brussels
Century City
Chicago
Dallas
Dubai
Frankfurt
Hartford
Hong Kong*
Houston
London
Los Angeles
Miami
Moscow
New York
Nur-Sultan
Orange County
Paris
Philadelphia
Pittsburgh
Princeton
San Francisco
Shanghai*
Silicon Valley
Singapore*
Tokyo
Washington, DC
Wilmington

Morgan Lewis

THANK YOU

Morgan, Lewis & Bockius UK LLP is a limited liability partnership registered in England and Wales under number OC378797 and is a law firm authorised and regulated by the Solicitors Regulation Authority. The SRA authorisation number is 615176.

Our Beijing and Shanghai offices operate as representative offices of Morgan, Lewis & Bockius LLP. In Hong Kong, Morgan, Lewis & Bockius is a separate Hong Kong general partnership registered with The Law Society of Hong Kong. Morgan Lewis Stamford LLC is a Singapore law corporation affiliated with Morgan, Lewis & Bockius LLP.