Panel – The Impact of AI Workloads on Datacenter Compute and Memory

Marc Tremblay
Microsoft

Distringuished Engineer, Azure H/W

Responsible for silicon / systems roadmap for AI

Full stack expert from application requirement to silicon

Previously CTO of Microelectronics @ Sun

Sumit Gupta
IBM

Vice President
AI, ML and HPC business

Responsible for strategy and HW/SW products for Watson ML accelerator and Spectrum compute

Previously GM AI & GPU data center business, Nvidia

Cliff Young
Google

Software Engineer
Google Brain team

HW/SW Codesign and Machine Learning

One of the TPU designers

Previously at D.E. Shaw and Bell Labs

Maxim Naumov
Facebook

Research Scientist
Facebook AI Research

Deep Learning, Parallel Algorithms and Systems

Co-developed many of Nvidia GPU-accelerated libraries

Previously at Nvidia Research and Intel

Greg Diamos
Baidu

AI Research Lead
Silicon Valley AI Labs

Co-developed Deep Speech and Deep Voice

Contributor to Volta SIMT scheduler, Compiler and Microarchitecture

Previously at Nvidia

Rob Ober
Nvidia

Chief Platform Architect
Datacenter products

GPU deployments for AI and DL @ Hyperscalers

CPU architecture, storage systems, SSD, N/W, wireless and pwr mgmt.

Previously at Sandisk, LSI, Apple etc.

Anand Iyer
Samsung

Dir, Planning/Technology, Semiconductor products

Data center and AI HBM/Accelerators/CPU

CPU, Network infra, near memory compute

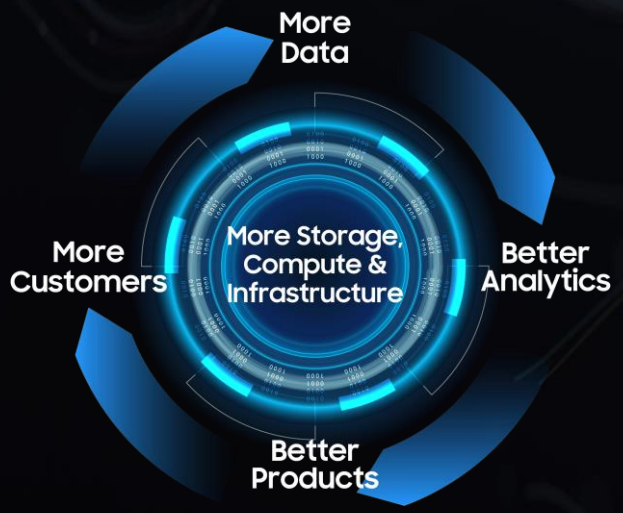
Previously at Broadcom, Cavium and Digital

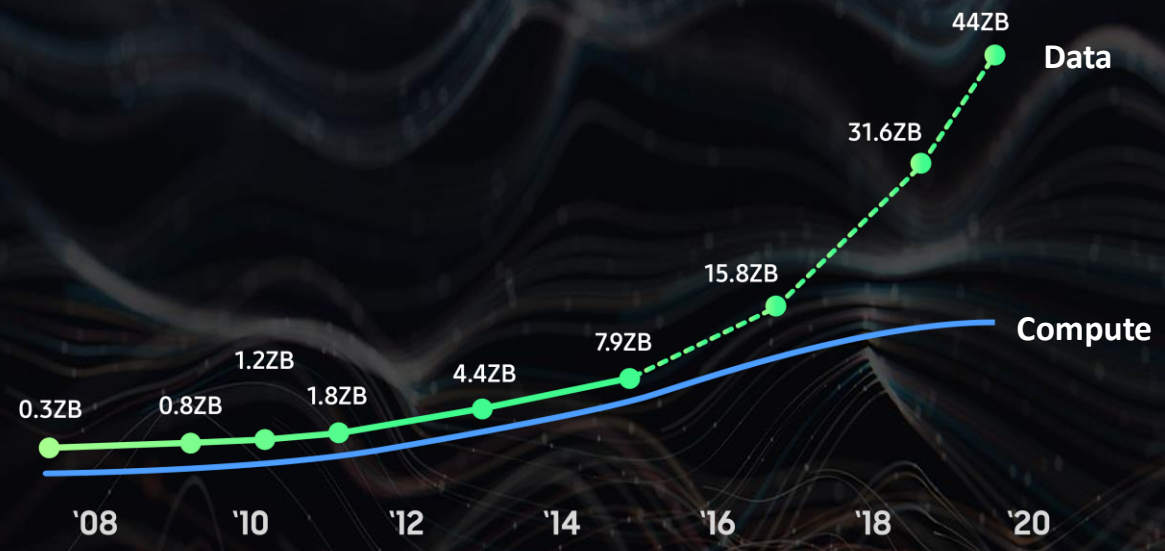
Samsung @ The Heart of Your Data

Unparalleled Product Breadth & Technology Leadership

DRAM | GDDR (SAMSUNG GDDR6) | HBM (SAMSUNG HBM) | NAND (SAMSUNG 3D V-NAND) | Foundry

Virtuous cycle of big data vs compute growth chasm





Data Curve from IDC/EMC Digital Universe reports 2008-2017, Compute curve HPE analysis

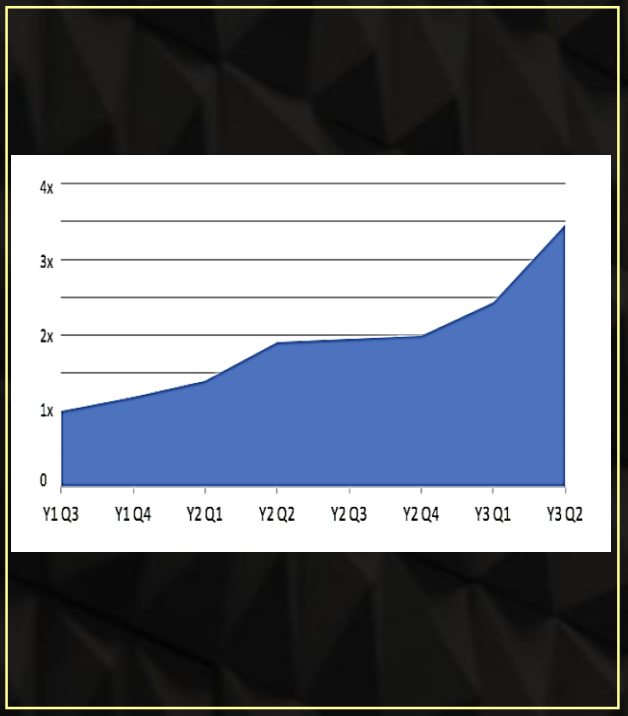
350M photos
4PB / day

2 Trillion searches / year

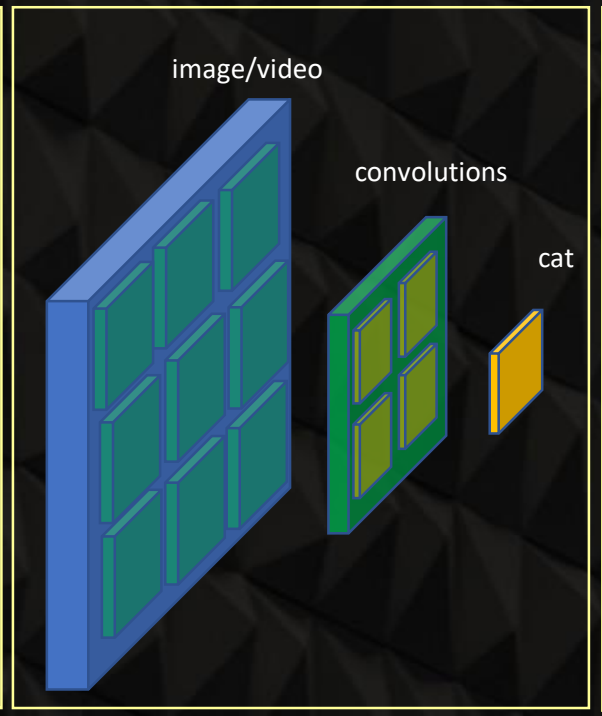
Flu Outbreak: Google knows before CDC

Data is being created faster than our ability to make sense of it.

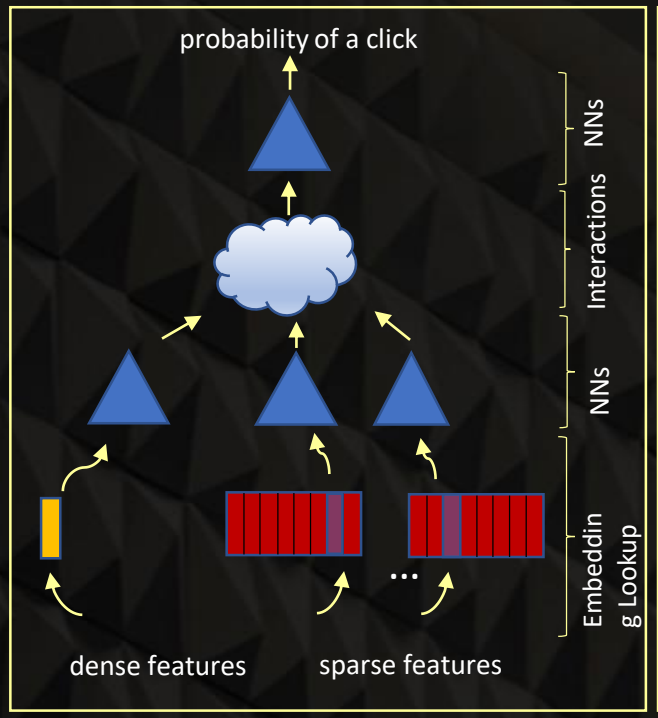
ML/DL Systems – Market Demand and Key Application Domains



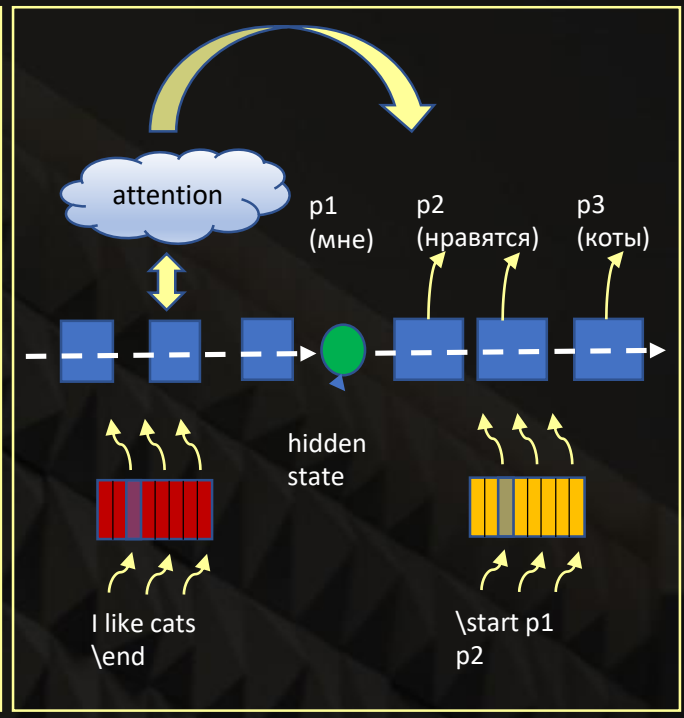
DL inference in data centers [1]



Vision



Recommenders



NMT

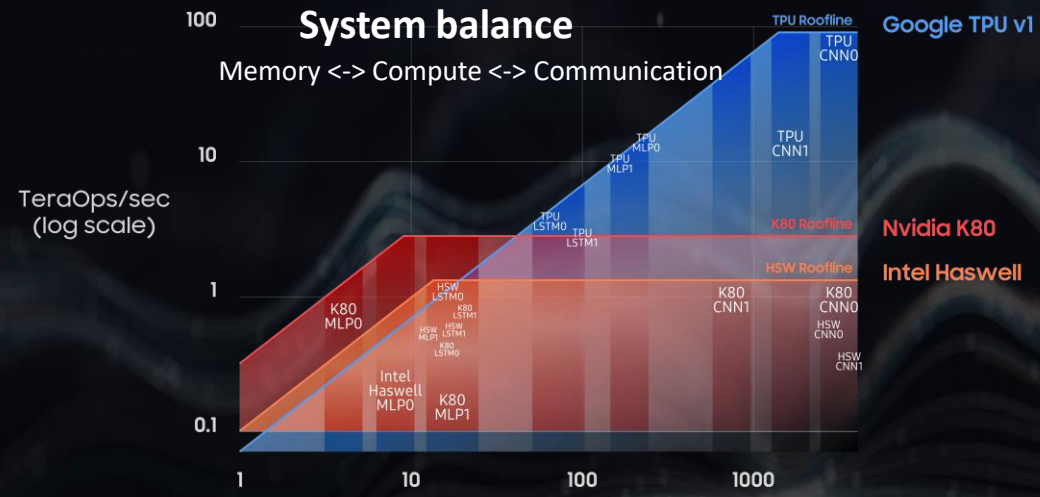
[1] “Deep Learning Inference in Data Centers: Characterization, Performance Optimizations and Hardware Implications”, ArXiv, 2018
https://code.fb.com/ai-research/scaling-neural-machine-translation-to-bigger-data-sets-with-faster-training-and-inference
https://code.fb.com/ml-applications/expanding-automatic-machine-translation-to-more-languages

Resource Requirements

Category	Model Types	Model Size (W)	Max. Activations	Op. Intensity (W)	Op. Intensity (X and W)
RecSys	FCs	1-10M	> 10K	20-200	20-200
	Embeddings	>10 Billion	> 10K	1-2	1-2
CV	ResNeXt101-32x4-48	43-829M	2-29M	Avg. 380/Min. 100	Avg. 188/Min. 28
	Faster-RCNN-ShuffleNet	6M	13M	Avg.3.5K/Min. 2.5K	Avg. 145/Min. 4
	ResNeXt3D-101	21M	58M	Avg. 22K/Min. 2K	Avg. 172/ Min. 6
NLP	seq2seq	100M-1B	>100K	2-20	2-20

AI Application Performance

For the foreseeable future, off-chip memory bandwidth will often be the constraining resource in system performance.



N. Jouppi, et.al., "In-Datacenter Performance Analysis of a Tensor Processing Unit"

Research + Development

- Time to train and accuracy
- Multiple runs for exploration, sometimes overnight

Production

- Optimal work/$
- Optimal work/watt
- Time to train

Memory Access for:

- Network / program config and control flow
- Training data mini-batch compute flow

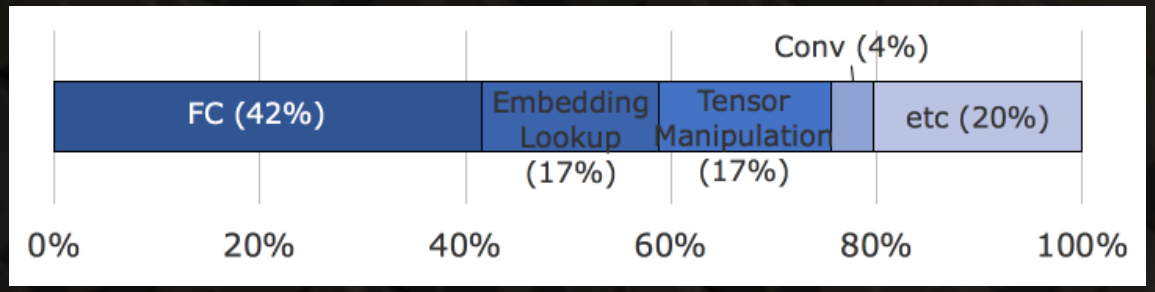
Compute consumes:

- Mini-batch data

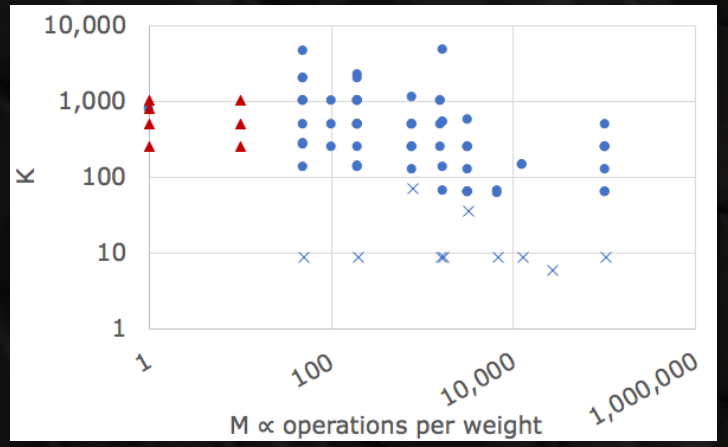
Communication for:

- All reduce
- Embedding table insertion

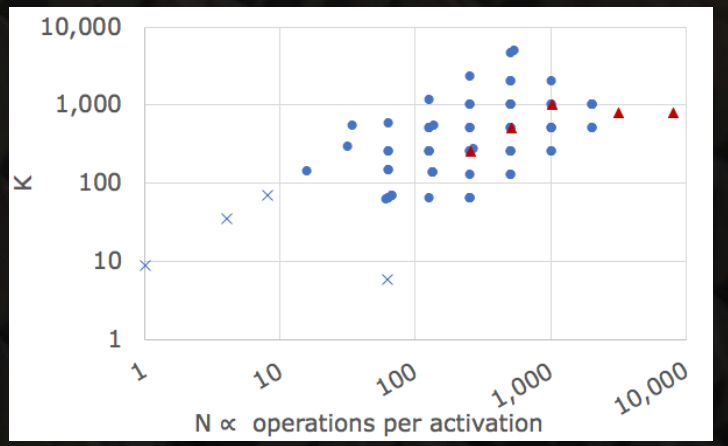
Hardware Trends



Time spent in caffe2 operators in data centers [1]



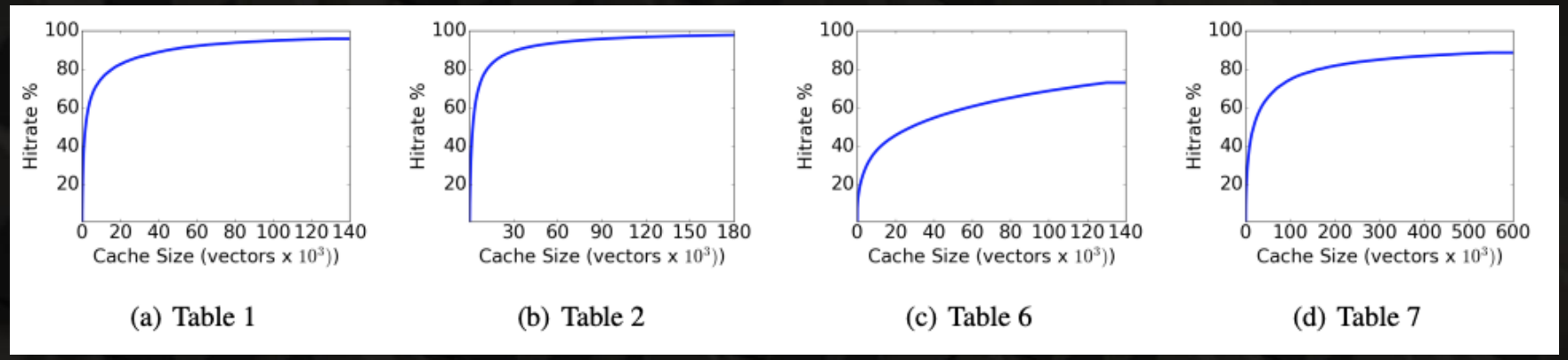
- High memory bandwidth and capacity for embeddings
- Support for powerful matrix and vector engines
- Large on-chip memory for inference with small batches
- Support for half-precision floating-point computation

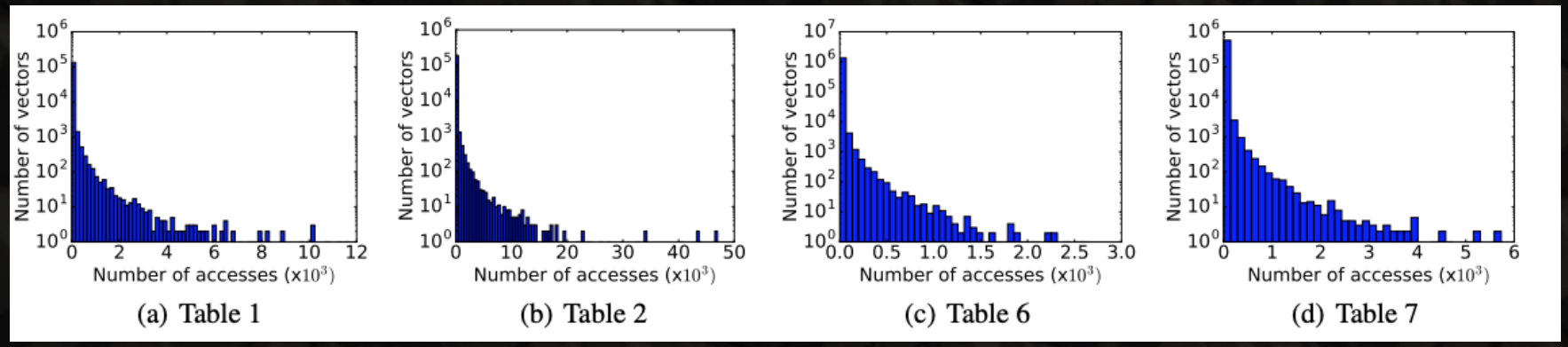


Common activation and weight matrix shapes ($\mathbf{X}_{MxK}\mathbf{W}^{T}_{KxN}$) [1]

[1] "Deep Learning Inference in Data Centers: Characterization, Performance Optimizations and Hardware Implications", ArXiv, 2018

Sample Workload Characterization





Embedding table hit rates and access histograms [2]

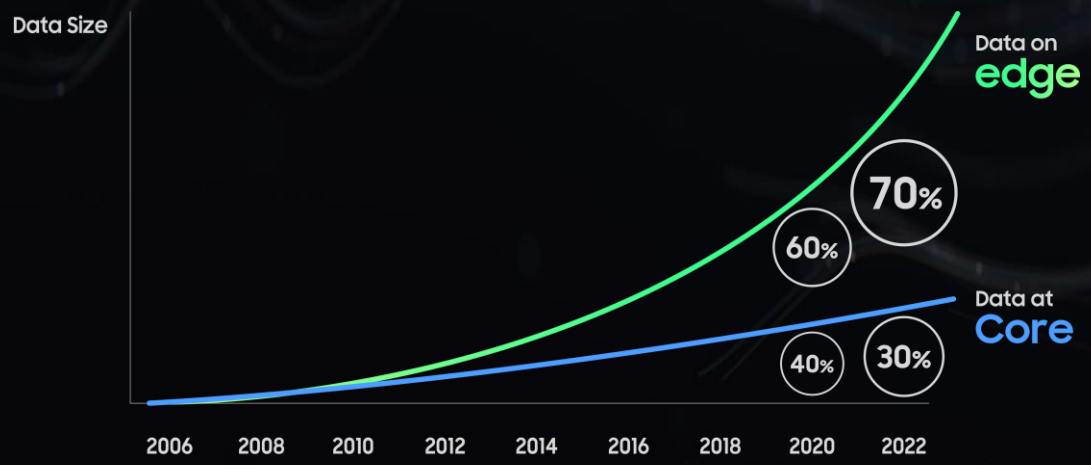
[2] “Bandana: Using Non-Volatile Memory for Storing Deep Learning Models”, SysML, 2019

MLPerf

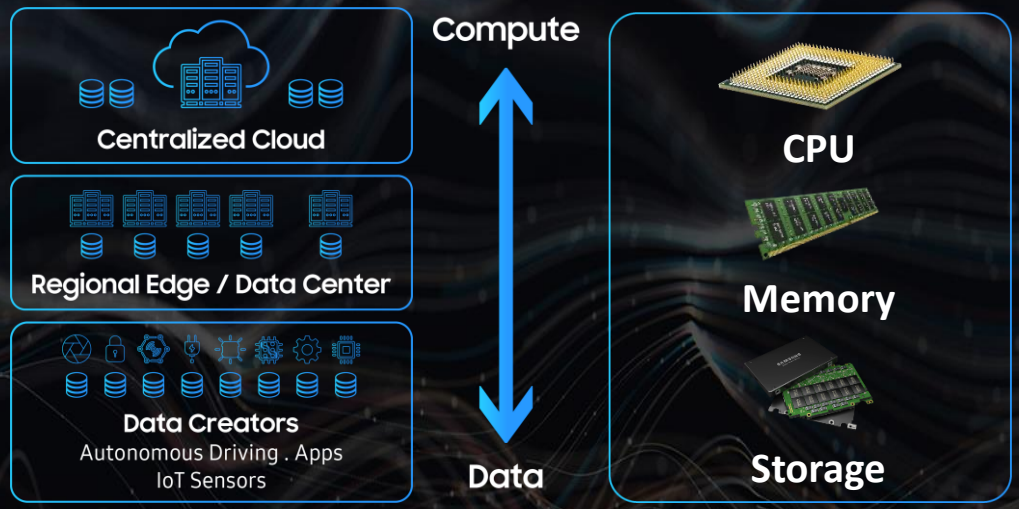
Closed Division Speedups

#	Submitter	Hardware	Chip count and type		Software	Benchmark results (speedup relative to reference implementation)							Cloud Scale	Power (unofficial, submitter-provided)
						Image classification	Object detection, light-weight	Object detection, heavy-wt.	Translation, recurrent	Translation, non-recur.	Recommendation	Reinforcement Learning		
						ImageNet	COCO	COCO	WMT E-G	WMT E-G	MovieLens-20M	Pro games		
						ResNet-50 v1.5	SSD w/ ResNet-34	Mask-R-CNN	NMT	Transformer	NCF	Mini Go		
Available in cloud														
1	**Reference**	Pascal P100	1	a	Unoptimized reference	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	n/a
2	**Google**	TPUv2.8	4	a	TF 1.12	29.3	8.5		28.1				2.6	n/a
3		TPUv2.512 + TPUv2.8	260	a	TF 1.12	781.5							171.6	n/a
4		TPUv3.8	4	a	TF 1.12	48.2	11.1		43.1				4.2	n/a
5		8x Volta V100	8	a	TF 1.12, cuDNN 7.4	64.1							11.4	n/a
Available on premi														
6	**Intel**	1x 2S SKX8180	2	c	Intel Caffe 1.1.2a	0.85							n/a	none
7		8x 2S SKX8180	16	c	Intel Caffe 1.1.2a	6.7							n/a	none
8		4x 4S SKX8180	16	c	Intel Caffe 1.1.2a	6.6							n/a	none
9		1x 2S SKX8180	2	c	BigDL 0.7.0						1.6		n/a	none
10		1x 2S SKX8180	2	c	TensorFlow 1.10.1							6.3	n/a	none
11		1x 4S SKX8180	4	c	TensorFlow 1.10.1							9.9	n/a	none
12	**NVIDIA**	DGX-1	8	a	ngc18.11_MXNet, cuDNN 7.4	65.6							n/a	none
13		DGX-1	8	a	ngc18.11_pyTorch, cuDNN 7.4		30.8	15.5	62.0	57.2	93.4		n/a	none
14		8x DGX-1	64	a	ngc18.11_pyTorch, cuDNN 7.4		127.3	61.7					n/a	none
15		24x DGX-1	192	a	ngc18.11_pyTorch, cuDNN 7.4					301.6			n/a	none
16		32x DGX-1	256	a	ngc18.11_pyTorch, cuDNN 7.4				405.2				n/a	none
17		80x DGX-1	640	a	ngc18.11_MXNet, cuDNN 7.4	1,424.4							n/a	none
18		DGX-2	16	a	ngc18.11_MXNet, cuDNN 7.4	119.5							n/a	none
19		DGX-2	16	a	ngc18.11_pyTorch, cuDNN 7.4		52.1	28.4	108.0	88.2	116.8		n/a	none
20		DGX-2h	16	a	ngc18.11_MXNet, cuDNN 7.4	126.2							n/a	none
21		DGX-2h	16	a	ngc18.11_pyTorch, cuDNN 7.4		58.7	30.0	115.8	97.4	116.8		n/a	none
22		4x DGX-2h	64	a	ngc18.11_pyTorch, cuDNN 7.4			69.3					n/a	none
23		8x DGX-2h w/ 8 V100s	64	a	ngc18.11_pyTorch, cuDNN 7.4		147.8						n/a	none
24		16x DGX-2h	256	a	ngc18.11_pyTorch, cuDNN 7.4				420.2				n/a	none
25		32x DGX-2h	512	a	ngc18.11_MXNet, cuDNN 7.4	1,193.4							n/a	none
Research														
26	**Google**	TPUv3.32 + TPUv2.8	20	a	TF 1.12	147.4	46.5		117.0				n/a	none
27		TPUv3.512 + TPUv2.8	260	a	TF 1.12	1,243.8							n/a	none
28	**Intel**	1x 2S SKX8180	2	c	custom TensorFlow 1.10.1							6.9	n/a	none
29		1x 4S SKX8180	4	c	custom TensorFlow 1.10.1							12.9	n/a	none

The Move to The Edge



By 2022, 7 out of every 10 bytes of data created will never see a data center.



Considerations

- Compute closer to Data
- Smarter Data Movement
- Faster Time to Insight

Let Data Speak for Itself!

Panel – The Impact of AI Workloads on Datacenter Compute and Memory

Marc Tremblay
Microsoft

Distringuished Engineer, Azure H/W

Responsible for silicon / systems roadmap for AI

Full stack expert from application requirement to silicon

Previously CTO of Microelectronics @ Sun

Sumit Gupta
IBM

Vice President
AI, ML and HPC business

Responsible for strategy and HW/SW products for Watson ML accelerator and Spectrum compute

Previously GM AI & GPU data center business, Nvidia

Cliff Young
Google

Software Engineer
Google Brain team

HW/SW Codesign and Machine Learning

One of the TPU designers

Previously at D.E. Shaw and Bell Labs

Maxim Naumov
Facebook

Research Scientist
Facebook AI Research

Deep Learning, Parallel Algorithms and Systems

Co-developed many of Nvidia GPU-accelerated libraries

Previously at Nvidia Research and Intel

Greg Diamos
Baidu

AI Research Lead
Baidu SVAIL

Co-developed Deep Speech and Deep Voice

Contributor to Volta SIMT scheduler, Compiler and Microarchitecture

Previously at Nvidia

Rob Ober
Nvidia

Chief Platform Architect
Datacenter products

GPU deployments for AI and DL @ Hyperscalers

CPU architecture, storage systems, SSD, N/W, wireless and pwr mgmt.

Previously at Sandisk, LSI, Apple etc.

Anand Iyer
Samsung

Dir, Planning/Technology, Semiconductor products

Data center and AI HBM/Accelerators/CPU

CPU, Network infra, near memory compute

Previously at Broadcom, Cavium and Digital

Samsung @ The Heart of Your Data

Unparalleled Product Breadth & Technology Leadership

DRAM | GDDR (SAMSUNG GDDR6) | HBM (SAMSUNG HBM) | NAND (SAMSUNG 3D V-NAND) | Foundry

Cloud AI

6X faster training time

8X training cost effectiveness

HBM

Edge computing AI

2X faster data access

2X hot data feeding

GDDR6

On-device AI

1.5X bandwidth

privacy & fast response

LPDDR5

Samsung @ The Heart of Your Data

Visit us at Booth #726