Ten Lessons From Three Generations Shaped Google's TPUv4i

•••

Patrick Legendre & Joergen Fagervik

Agenda

- Evolution of Google TPU designs
- 10 Lessons from 3 Generations TPU designs
- Results
- Conclusion

From TPUv1 to TPUv4i

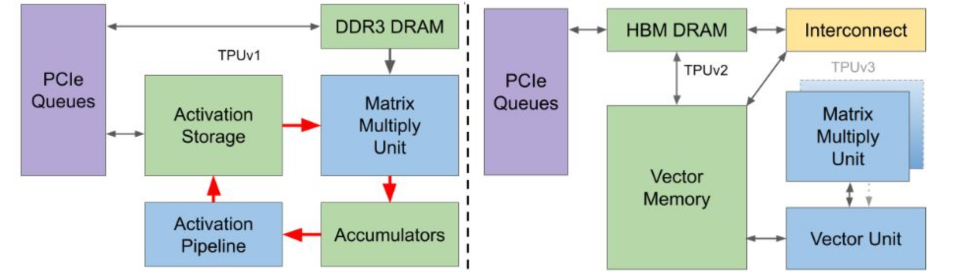
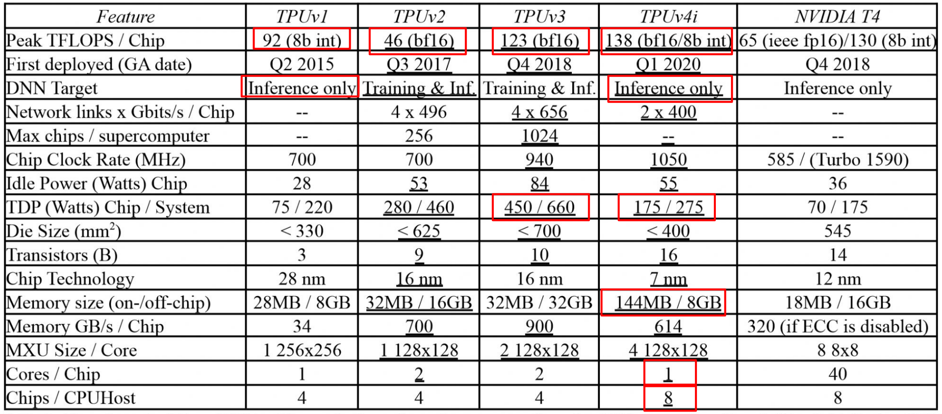


Figure 1. TPUv1 block diagram (left) vs TPUv2/v3.



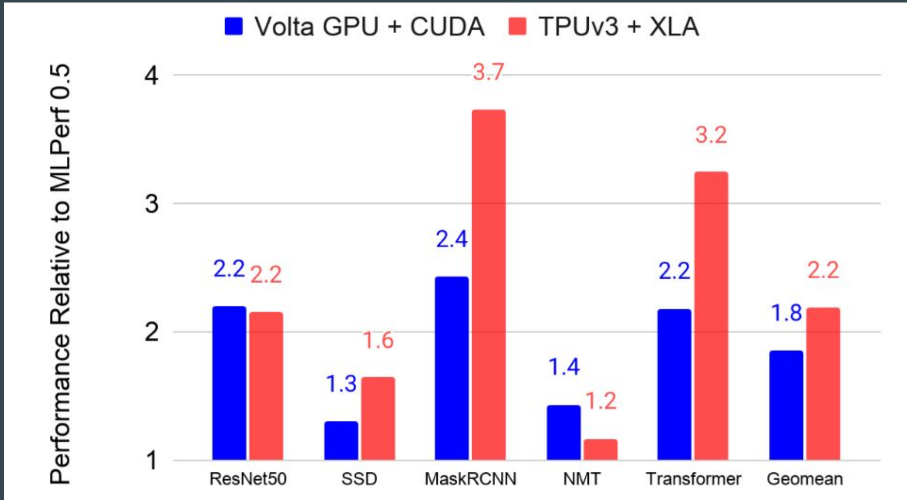
Feature	*TPUv1*	*TPUv2*	*TPUv3*	*TPUv4i*	*NVIDIA T4*
Peak TFLOPS / Chip	92 (8b int)	46 (bf16)	123 (bf16)	138 (bf16/8b int)	65 (ieee fp16)/130 (8b int)
First deployed (GA date)	Q2 2015	Q3 2017	Q4 2018	Q1 2020	Q4 2018
DNN Target	Inference only	Training & Inf.	Training & Inf.	Inference only	Inference only
Network links x Gbits/s / Chip	--	4 x 496	4 x 656	2 x 400	--
Max chips / supercomputer	--	256	1024	--	--
Chip Clock Rate (MHz)	700	700	940	1050	585 / (Turbo 1590)
Idle Power (Watts) Chip	28	53	84	55	36
TDP (Watts) Chip / System	75 / 220	280 / 460	450 / 660	175 / 275	70 / 175
Die Size (mm^2)	< 330	< 625	< 700	< 400	545
Transistors (B)	3	9	10	16	14
Chip Technology	28 nm	16 nm	16 nm	7 nm	12 nm
Memory size (on-/off-chip)	28MB / 8GB	32MB / 16GB	32MB / 32GB	144MB / 8GB	18MB / 16GB
Memory GB/s / Chip	34	700	900	614	320 (if ECC is disabled)
MXU Size / Core	1 256x256	1 128x128	2 128x128	4 128x128	8 8x8
Cores / Chip	1	2	2	1	40
Chips / CPUHost	4	4	4	8	8

Table 1. Key characteristics of DSAs. The underlines show changes over the prior TPU generation, from left to right. System TDP includes power for the DSA memory system plus its share of the server host power, e.g., add host TDP/8 for 8 DSAs per host.

10 lessons

Compiler Compatibility Trumps Binary Compatibility

- Maintain backwards compatibility
- Achieve better instruction level parallelism
- Share source code vs sharing binary files



Target Total Cost over Initial Cost

- TCO vs CapEx

$$TCO = CapEX + n * Opex$$

CapEx (CapitalExpense) = price of an item

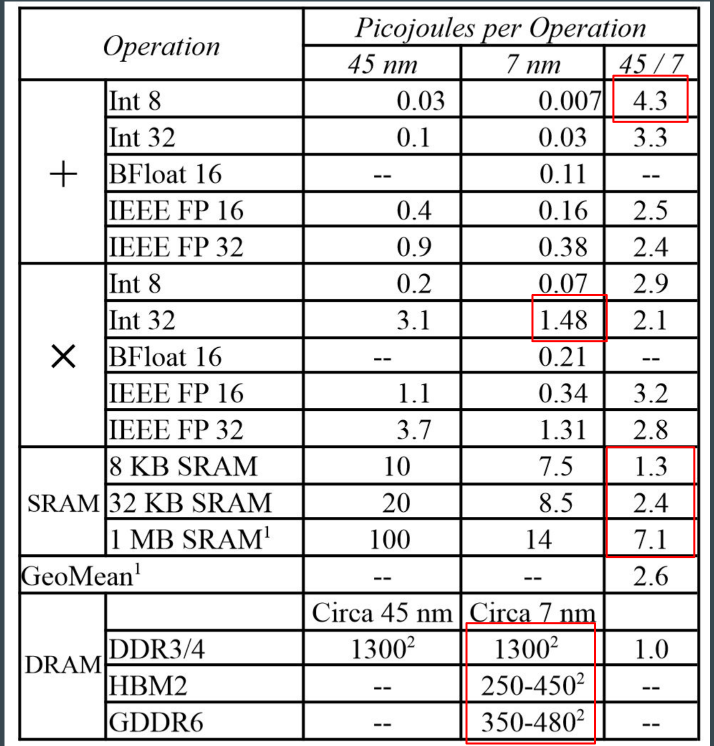
Opex (OperationExpense) = cost of operation (electricity, power provisioning)

TCO = Total cost of Ownership

n = number of years in operation

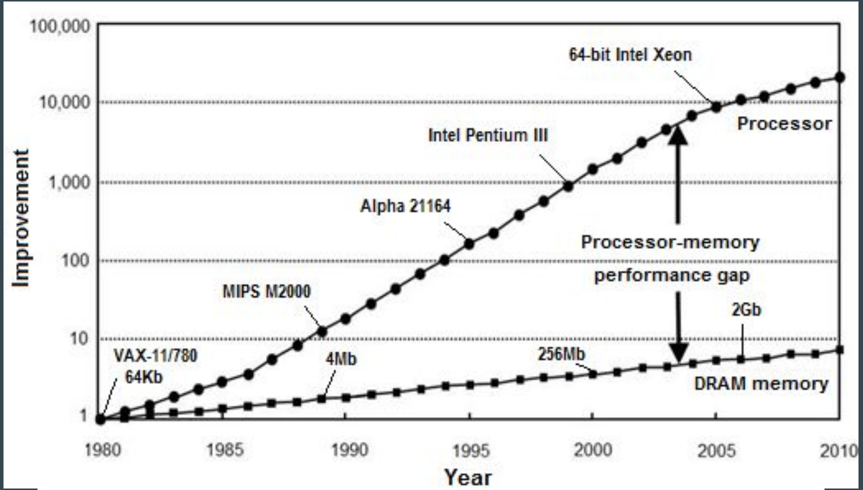
- perf/TCO over raw performance

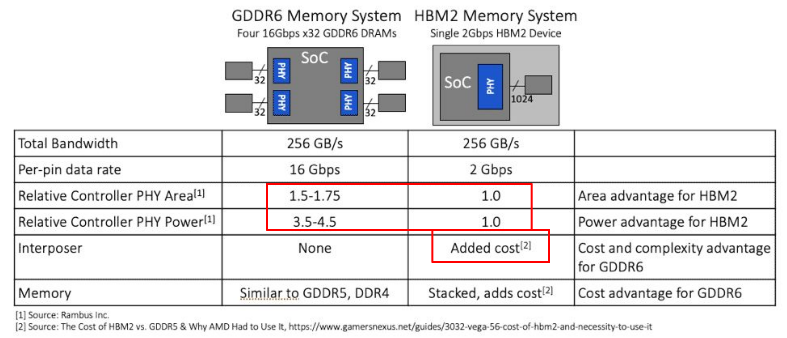
Semi-Conductor Technology Advances Unequally



Operation		*Picojoules per Operation*		
		45 nm	*7 nm*	*45 / 7*
+	Int 8	0.03	0.007	4.3
	Int 32	0.1	0.03	3.3
	BFloat 16	--	0.11	--
	IEEE FP 16	0.4	0.16	2.5
	IEEE FP 32	0.9	0.38	2.4
×	Int 8	0.2	0.07	2.9
	Int 32	3.1	1.48	2.1
	BFloat 16	--	0.21	--
	IEEE FP 16	1.1	0.34	3.2
	IEEE FP 32	3.7	1.31	2.8
SRAM	8 KB SRAM	10	7.5	1.3
	32 KB SRAM	20	8.5	2.4
	1 MB SRAM[1]	100	14	7.1
GeoMean[1]		--	--	2.6
DRAM		Circa 45 nm	Circa 7 nm	
	DDR3/4	1300[2]	1300[2]	1.0
	HBM2	--	250-450[2]	--
	GDDR6	--	350-480[2]	--

Table 2. Energy per Operation: 45 nm [16] vs 7 nm. Memory is pJ per 64-bit access.



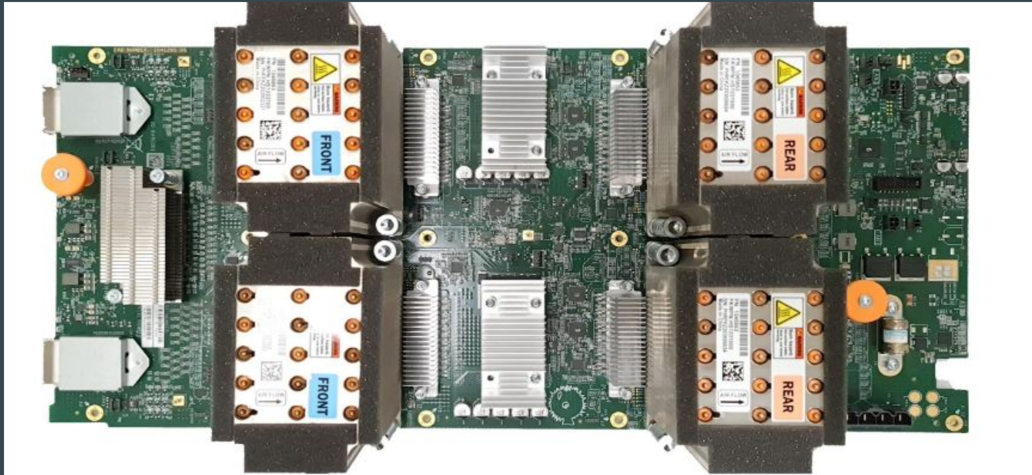


Total Bandwidth	256 GB/s	256 GB/s	
Per-pin data rate	16 Gbps	2 Gbps	
Relative Controller PHY Area[1]	1.5-1.75	1.0	Area advantage for HBM2
Relative Controller PHY Power[1]	3.5-4.5	1.0	Power advantage for HBM2
Interposer	None	Added cost[2]	Cost and complexity advantage for GDDR6
Memory	Similar to GDDR5, DDR4	Stacked, adds cost[2]	Cost advantage for GDDR6

[1] Source: Rambus Inc.
[2] Source: The Cost of HBM2 vs. GDDR5 & Why AMD Had to Use It, https://www.gamersnexus.net/guides/3032-vega-56-cost-of-hbm2-and-necessity-to-use-it

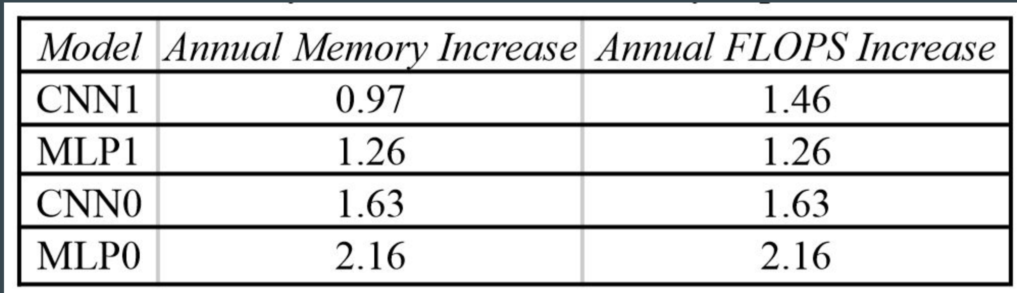
Support Multi-Tenancy

- Support future DNN requirements
- Up to 8 TPUs on 1 host CPU
- Sharing can reduce costs and latency
- Support for multiple batch sizes
- Support fast switching times between models



Deep Neural Networks grow 1.5x Annually

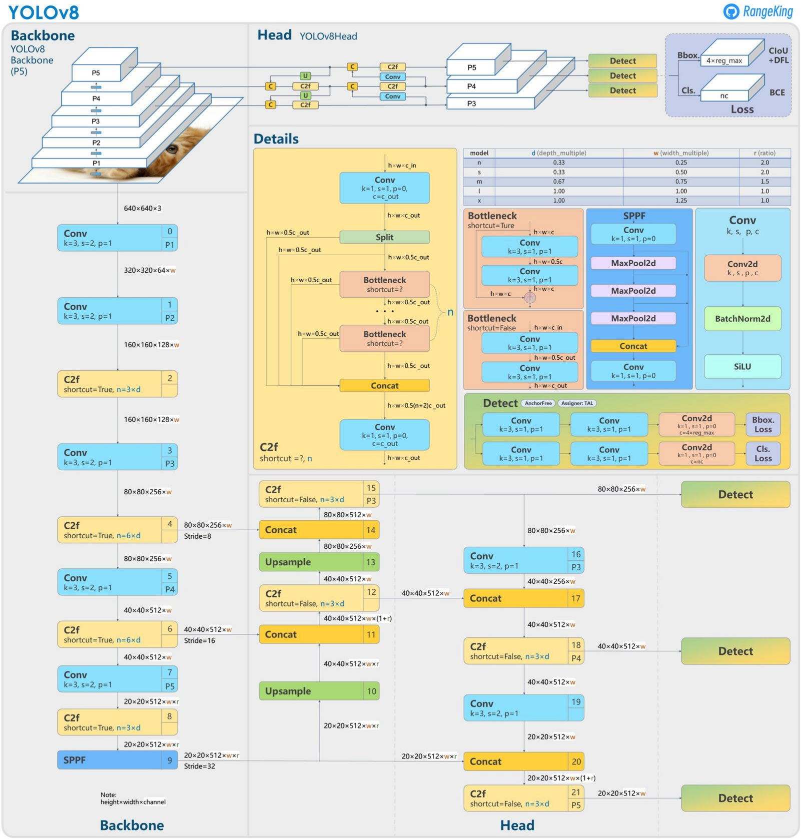
- Production DNNs needs grow as fast as Moore’s law



Model	*Annual Memory Increase*	*Annual FLOPS Increase*
CNN1	0.97	1.46
MLP1	1.26	1.26
CNN0	1.63	1.63
MLP0	2.16	2.16

DNN advances evolve Workloads

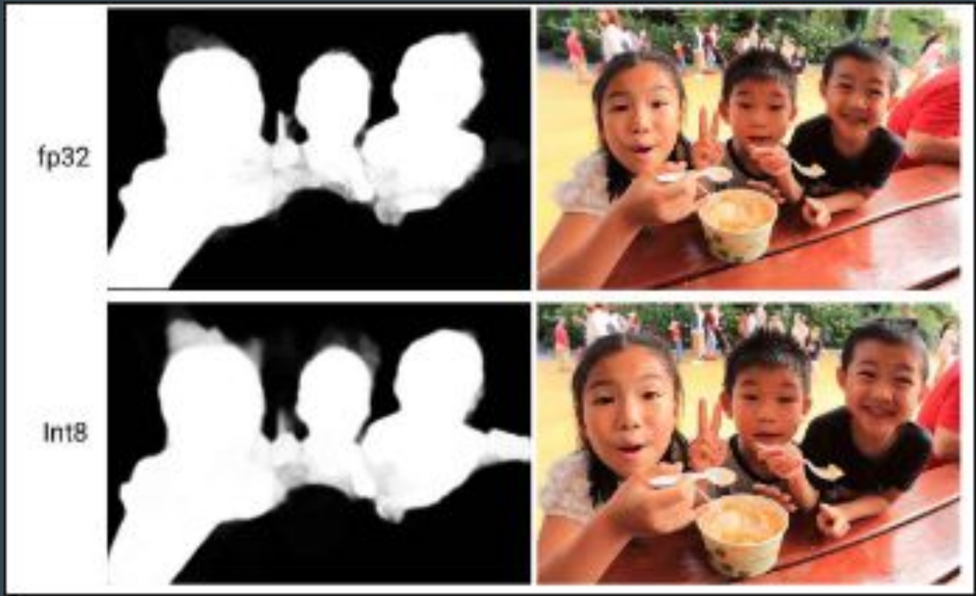
- Ambitious
- Waiting for software and hardware to catch up
- Fast paced area
- Larger and more complex networks



model	d (depth_multiple)	w (width_multiple)	r (ratio)
n	0.33	0.25	2.0
s	0.33	0.50	2.0
m	0.67	0.75	1.5
l	1.00	1.00	1.0
x	1.00	1.25	1.0

Inference requires Floating Point

- Quantization
- Precision
- The 1% drop

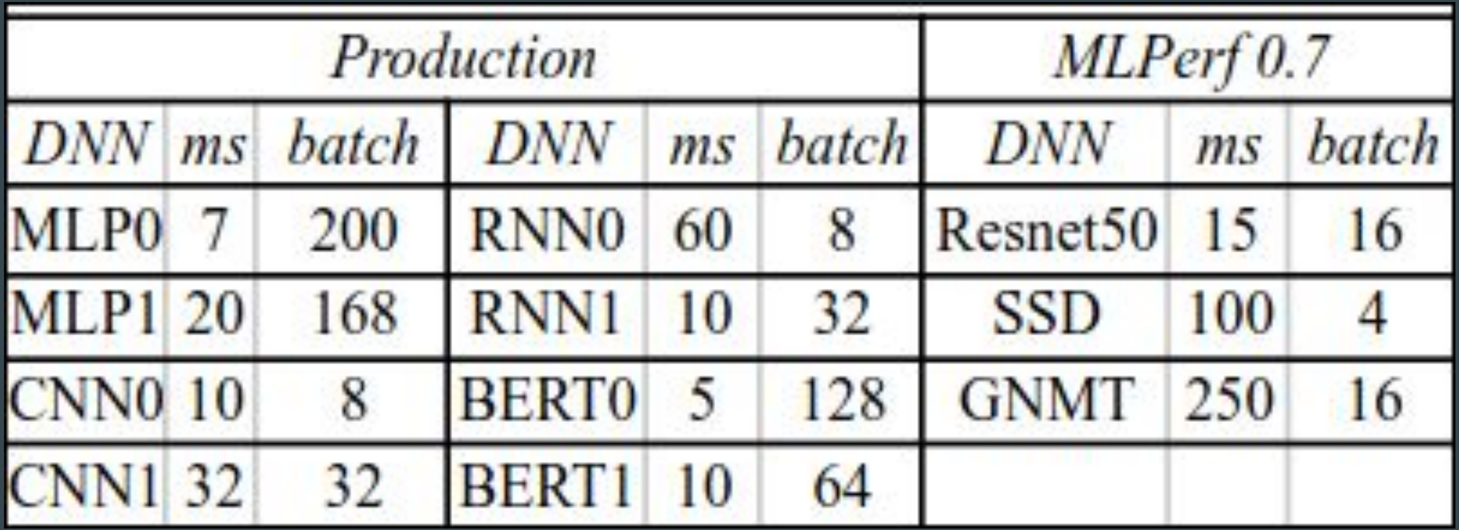


Inference DSAs need Air Cooling

- TPUv3 used liquid cooling
- Data centres
- Storage

Applications limit latency

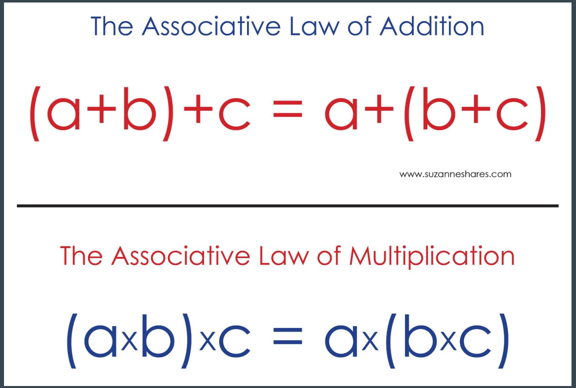
- Latency limit
- Not batch size



Production						*MLPerf 0.7*		
DNN	*ms*	*batch*	*DNN*	*ms*	*batch*	*DNN*	*ms*	*batch*
MLP0	7	200	RNN0	60	8	Resnet50	15	16
MLP1	20	168	RNN1	10	32	SSD	100	4
CNN0	10	8	BERT0	5	128	GNMT	250	16
CNN1	32	32	BERT1	10	64			

Support Backwards compatibility

- Time-to-market constraints
- Same results no matter versions
- The horrors of IEEE754 and associativity
- BFloat16 vs Float32



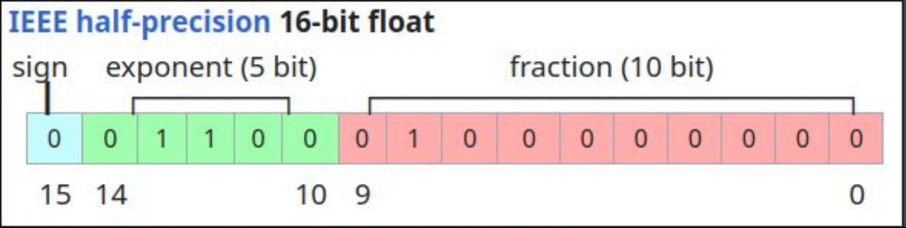
The Associative Law of Addition

$$(a+b)+c = a+(b+c)$$

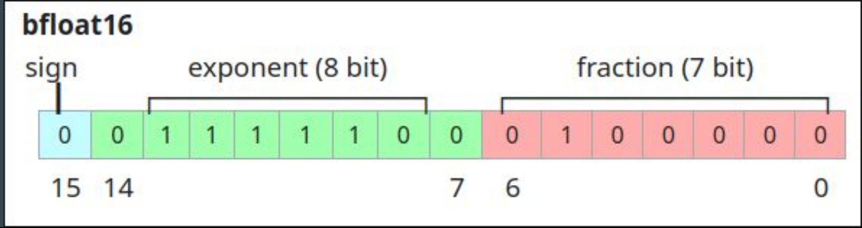
www.suzanneshares.com

The Associative Law of Multiplication

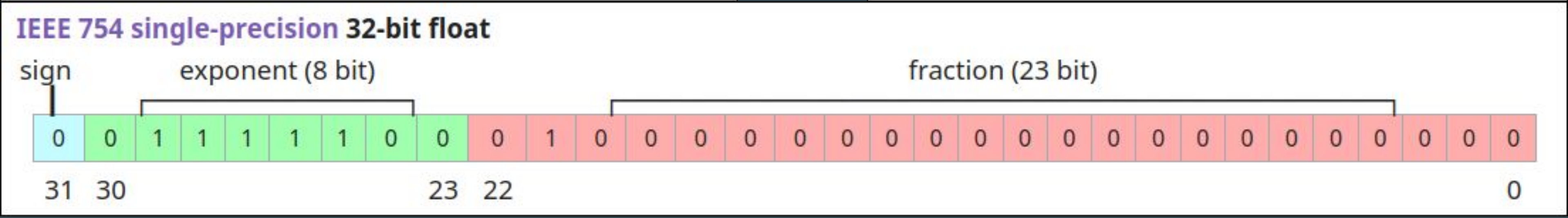
$$(a \times b) \times c = a \times (b \times c)$$



IEEE half-precision 16-bit float

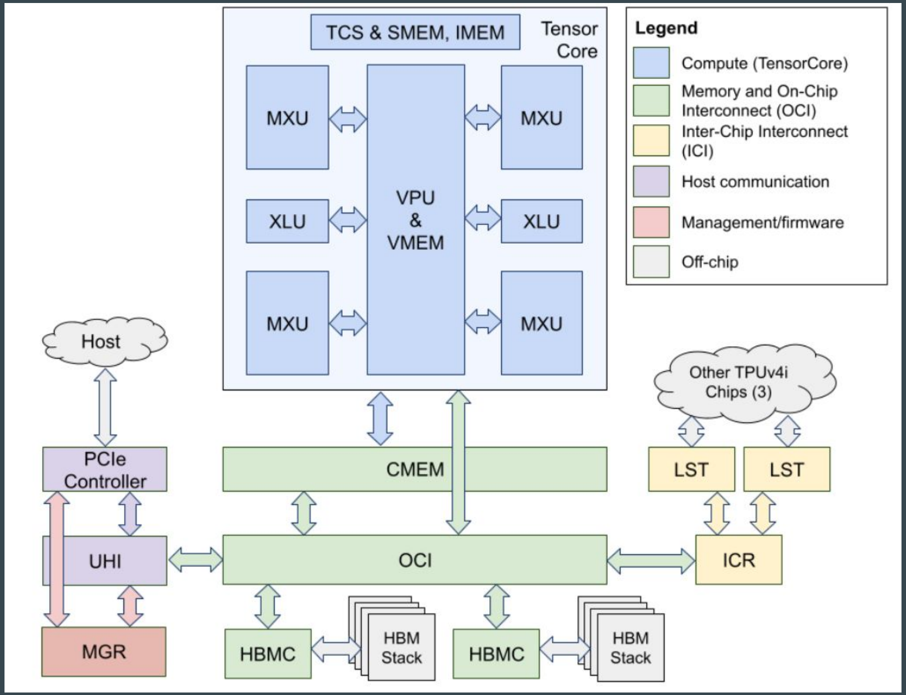


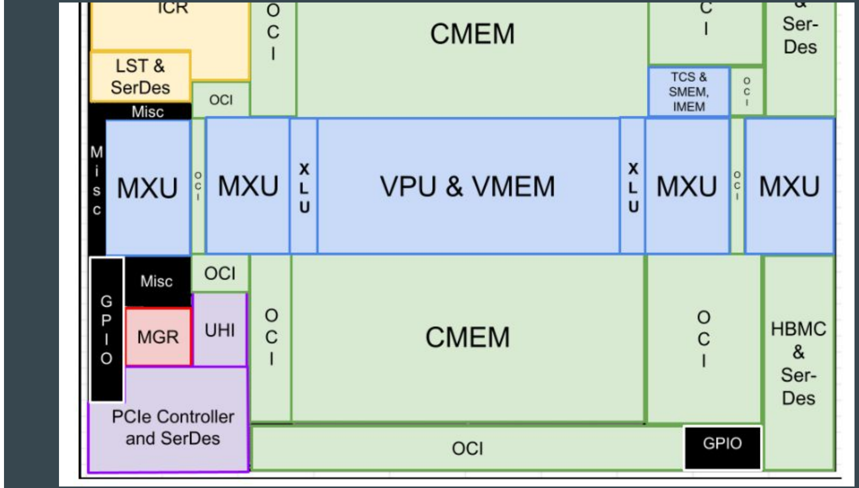
bfloat16



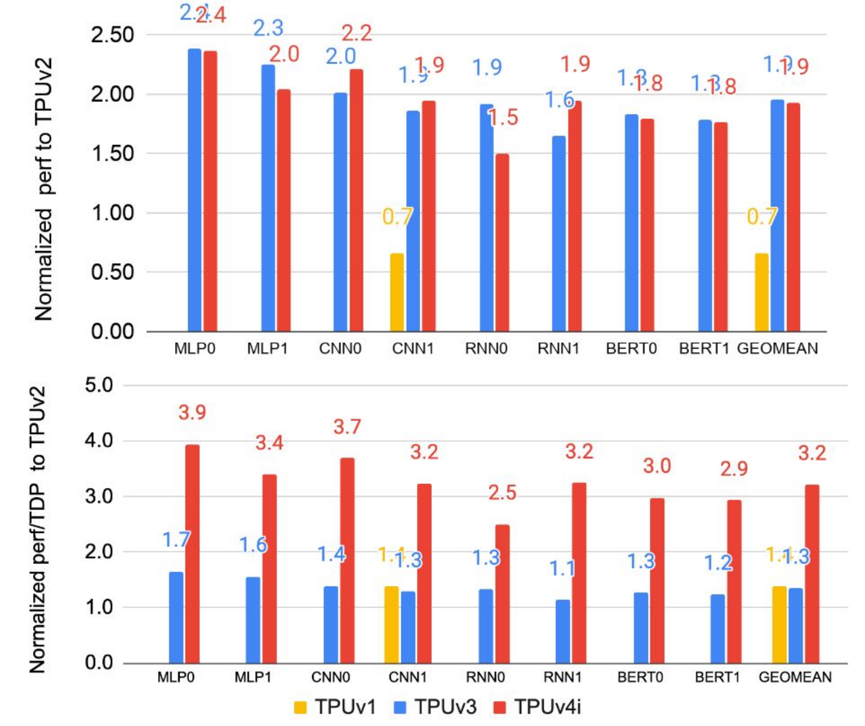
IEEE 754 single-precision 32-bit float

Google TPUv4i

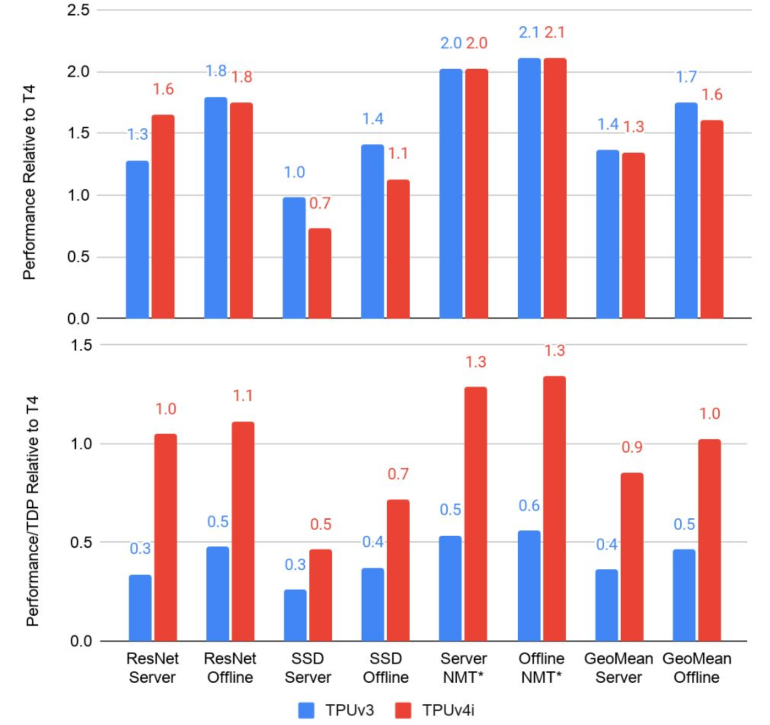




Results



Normalized perf to TPUv2
2.50
2.00
1.50
1.00
0.50
0.00
2.4
2.4
2.3
2.0
2.0
2.2
0.7
1.9
1.9
1.9
1.5
1.6
1.9
1.8
1.8
1.8
1.8
0.7
1.9
1.9
MLP0
MLP1
CNN0
CNN1
RNN0
RNN1
BERT0
BERT1
GEOMEAN
Normalized perf/TDP to TPUv2
5.0
4.0
3.0
2.0
1.0
0.0
3.9
1.7
3.4
1.6
3.7
1.4
3.2
1.4
1.3
2.5
1.3
3.2
1.1
3.0
1.3
2.9
1.2
3.2
1.4
1.3
MLP0
MLP1
CNN0
CNN1
RNN0
RNN1
BERT0
BERT1
GEOMEAN
TPUv1
TPUv3
TPUv4i



Performance Relative to T4
2.5
2.0
1.5
1.0
0.5
0.0
1.3
1.6
1.8
1.8
1.0
0.7
1.4
1.1
2.0
2.0
2.1
2.1
1.4
1.3
1.7
1.6
Performance/TDP Relative to T4
1.5
1.0
0.5
0.0
0.3
1.0
0.5
1.1
0.3
0.5
0.4
0.7
0.5
1.3
0.6
1.3
0.4
0.9
0.5
1.0
ResNet Server
ResNet Offline
SSD Server
SSD Offline
Server NMT*
Offline NMT*
GeoMean Server
GeoMean Offline
TPUv3
TPUv4i

Conclusion

Questions ?

References

- A Survey of Different Approaches for Overcoming the Processor - Memory Bottleneck - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Yearly-improvement-of-processor-and-DRAM-memory-speeds-for-a-time-period-of-three_fig2_340621273 [accessed 20 Oct, 2023]
- Memory Systems for AI, Steven Woo, Available from: https://www.rambus.com/blogs/memory-systems-for-ai-part-6/ [accessed 20 Oct, 2023]
- BFloat16. Available from: https://en.wikipedia.org/wiki/Bfloat16_floating-point_format [accessed 23 Oct, 2023[