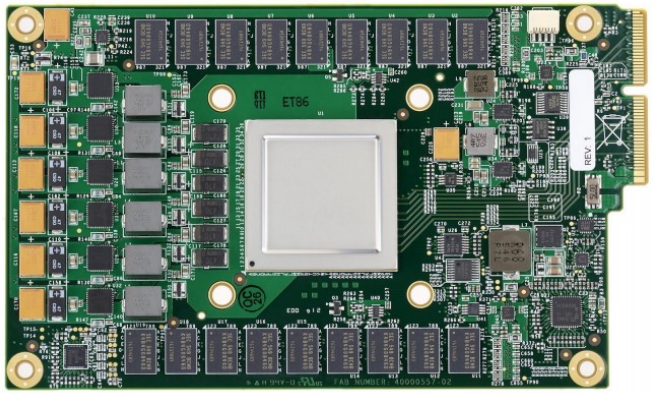
Application-Specific Hardware

…in the real world

Google

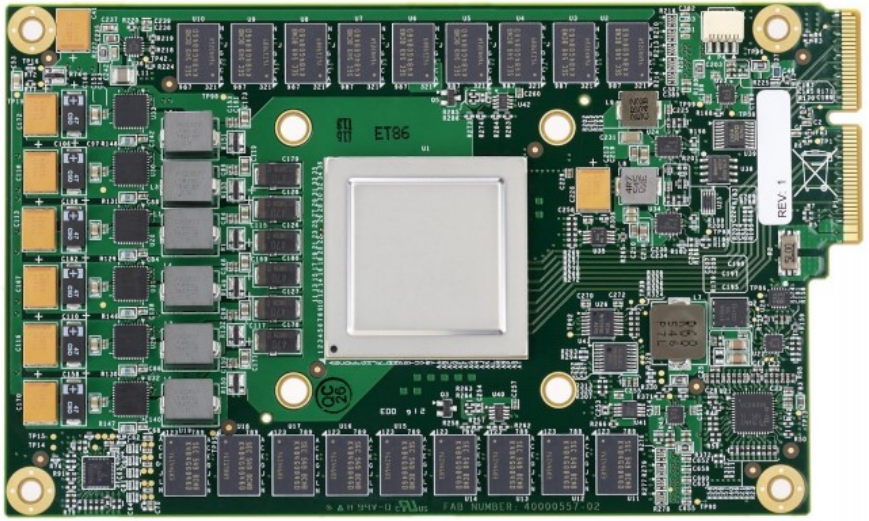


Microsoft

http://warfarehistorynetwork.com/wp-content/uploads/Military-Weapons-the-Catapult.jpg

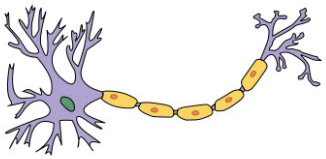
Google - Tensor Processing Unit

- Why?
  - 2006:
    - First considered datacenter ASIC/FPGA/GPU, decided excess capacity would suffice
  - 2013 projection:
    - Search by voice for 3min/day using DNNs → **double** datacenter computation needs
- Goals:
  - 10x better cost-performance vs GPUs
  - Deployment ASAP

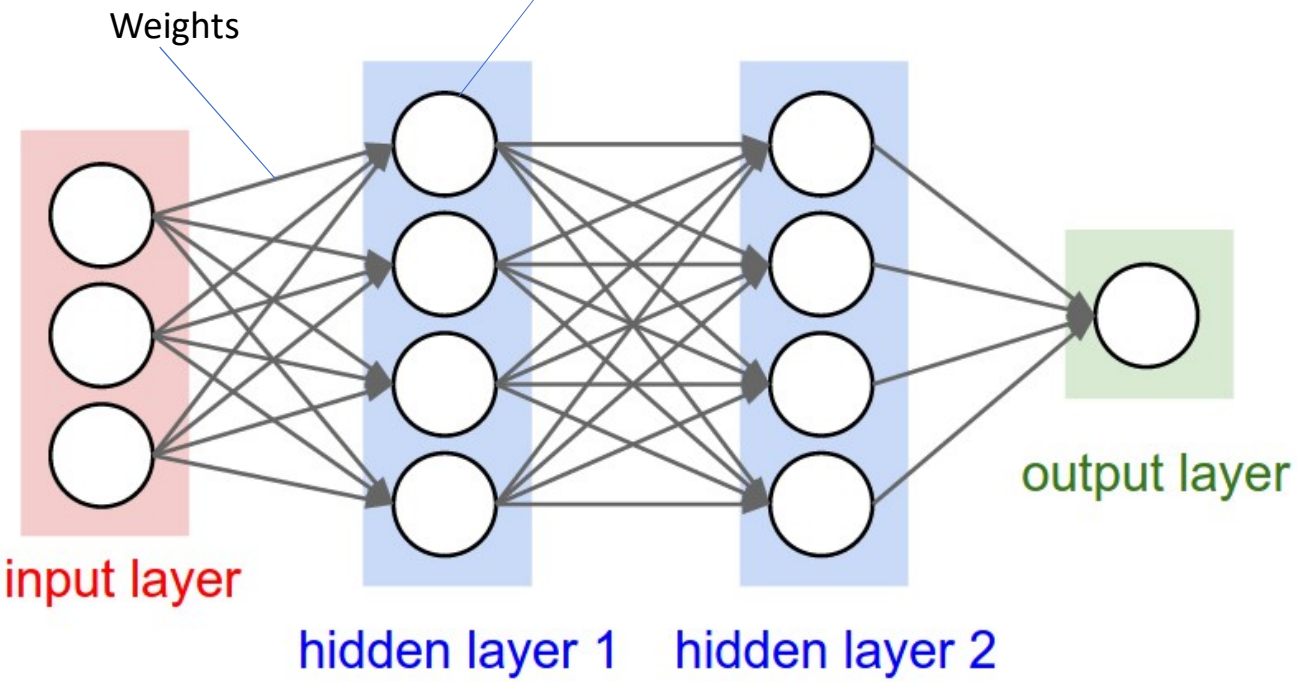


All images in this section are from TPU paper at ISCA 2017: https://www.cs.virginia.edu/~smk9u/CS6354F17/TPU.pdf

Neural nets



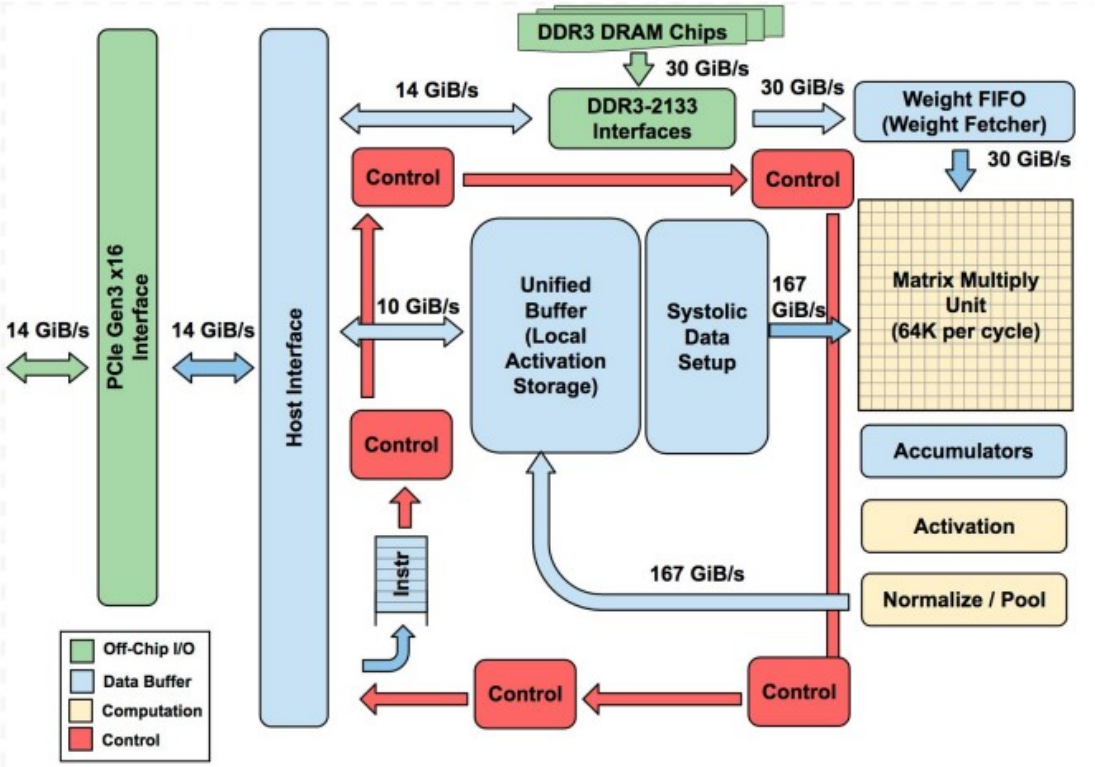
https://upload.wikimedia.org/wikipedia/commons/b/be/Derived_Neuron_schema_with_no_labels.svg



http://cs231n.github.io/neural-networks-1/

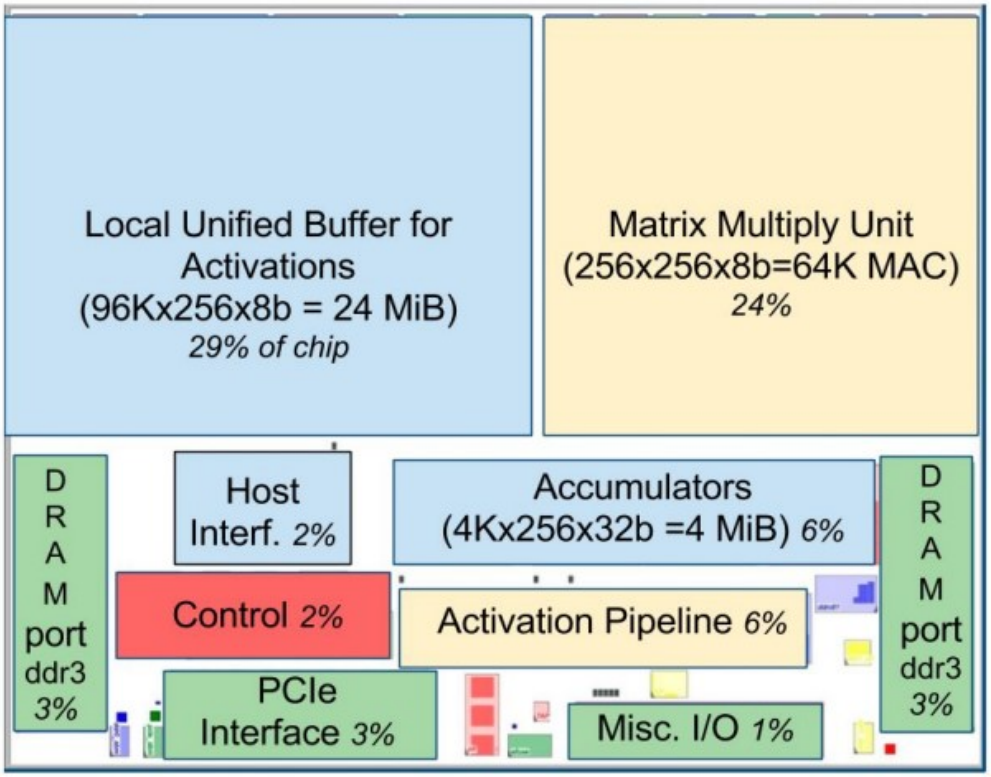
TPU architecture

- PCIe coprocessor
- No internal instruction fetch
  - CISC-like instructions from host:
    - Read_Host_Memory
    - Read_Weights
    - MatrixMultiply/Convolve
    - Activate
    - Write_Host_Memory
- Off-chip DDR3 weight memory

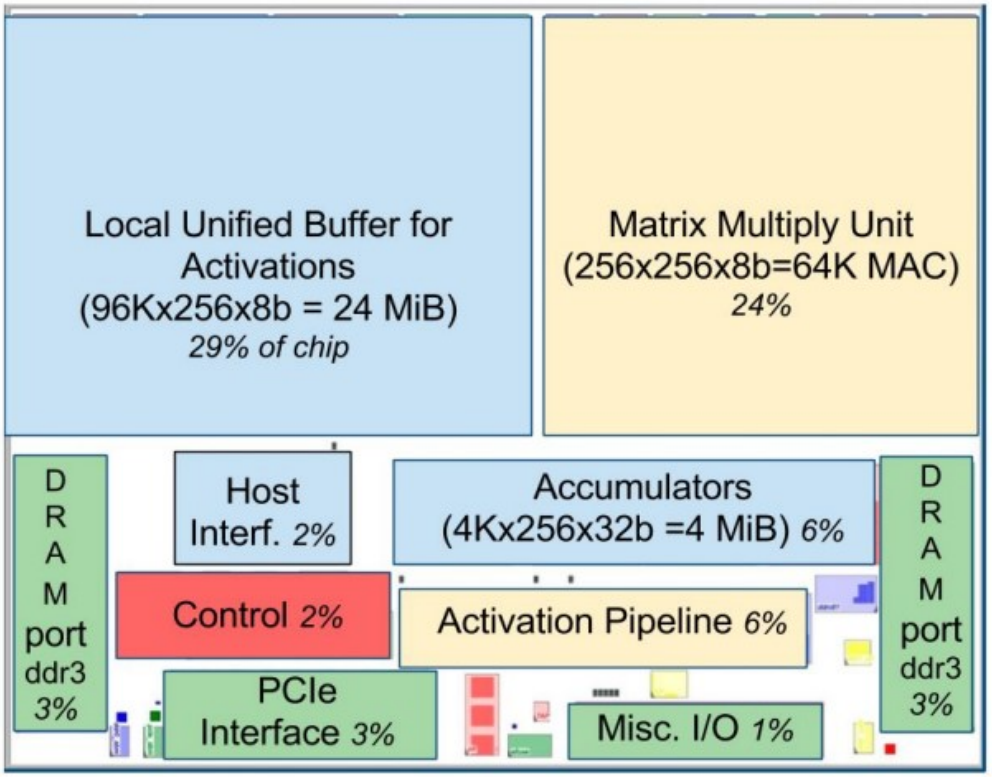


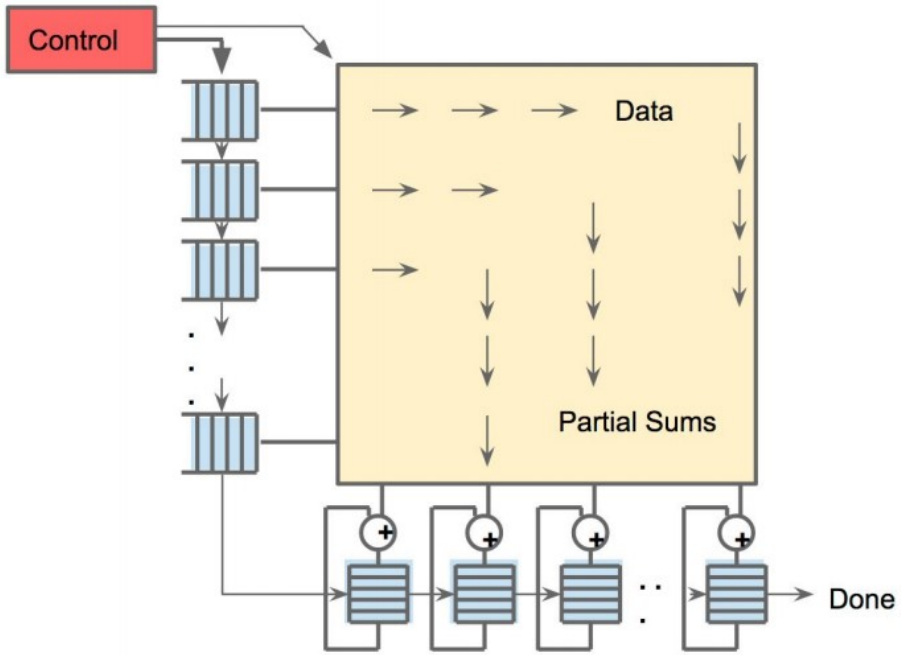
TPU architecture

- MACs for core computation
- 24MB Unified Buffer
  - Store intermediate results
  - Sized to match pitch of matmult unit, simplify compilation w/ specific apps
- Tiny control logic



TPU architecture – systolic structure



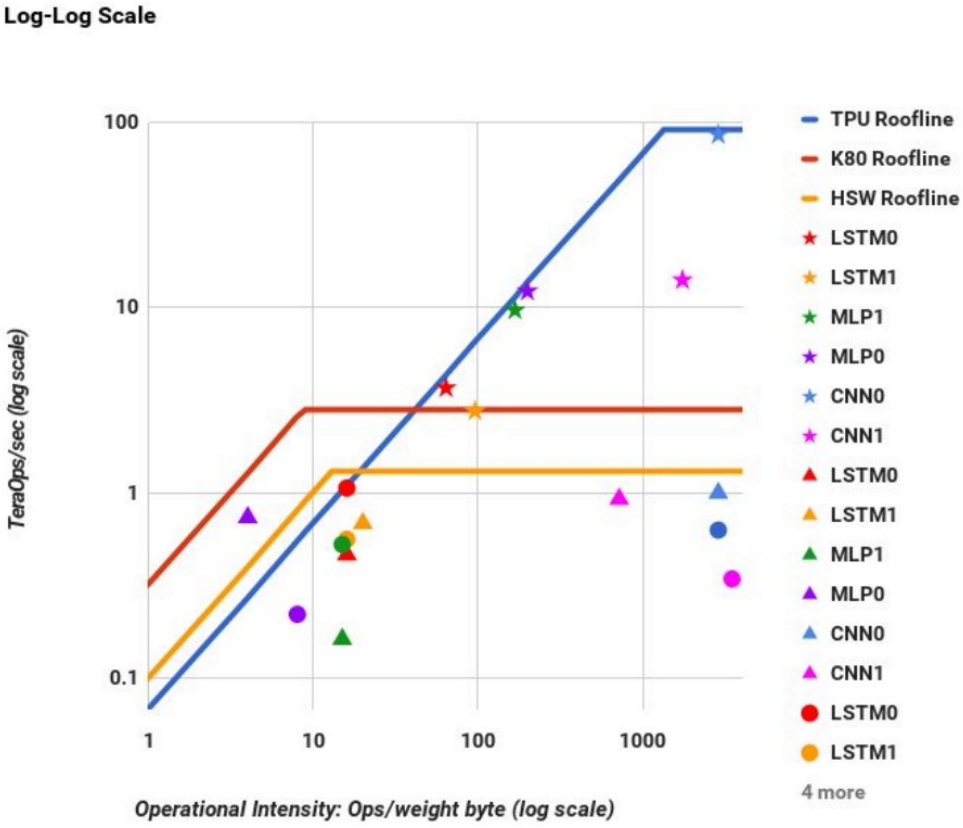


System configurations

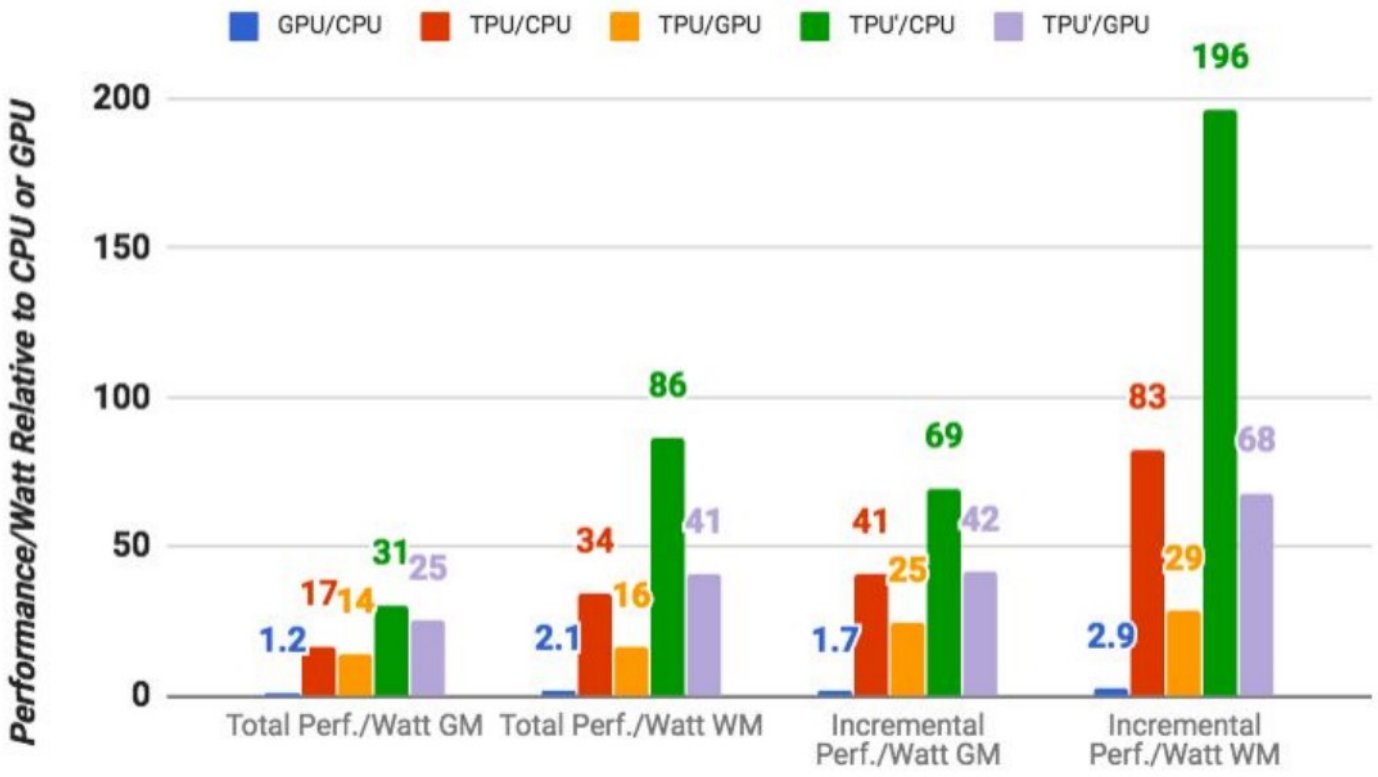
Model	Die										Benchmarked Servers				
	mm^2	nm	MHz	TDP	Measured		TOPS/s		GB/s	On-Chip Memory	Dies	DRAM Size	TDP	Measured	
					Idle	Busy	8b	FP						Idle	Busy
Haswell E5-2699 v3	662	22	2300	145W	41W	145W	2.6	1.3	51	51 MiB	2	256 GiB	504W	159W	455W
NVIDIA K80 (2 dies/card)	561	28	560	150W	25W	98W	--	2.8	160	8 MiB	8	256 GiB (host) + 12 GiB x 8	1838W	357W	991W
TPU	<331*	28	700	75W	28W	40W	92	--	34	28 MiB	4	256 GiB (host) + 8 GiB x 4	861W	290W	384W

Performance

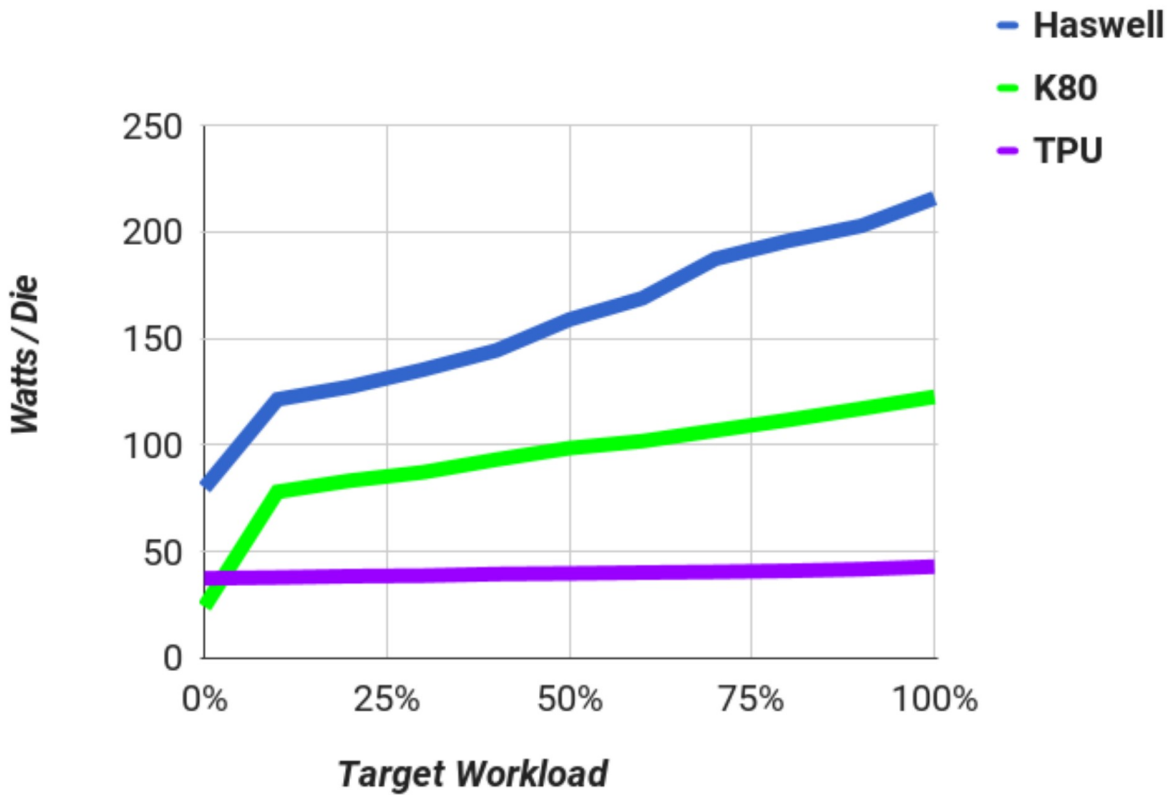
- “Roofline curves” – computation vs memory-intensity
  - “Ridge point” at intensity where app becomes compute-bound
  - Before ridge = memory-bound
  - After ridge = compute-bound
- Below curve = response time-constrained



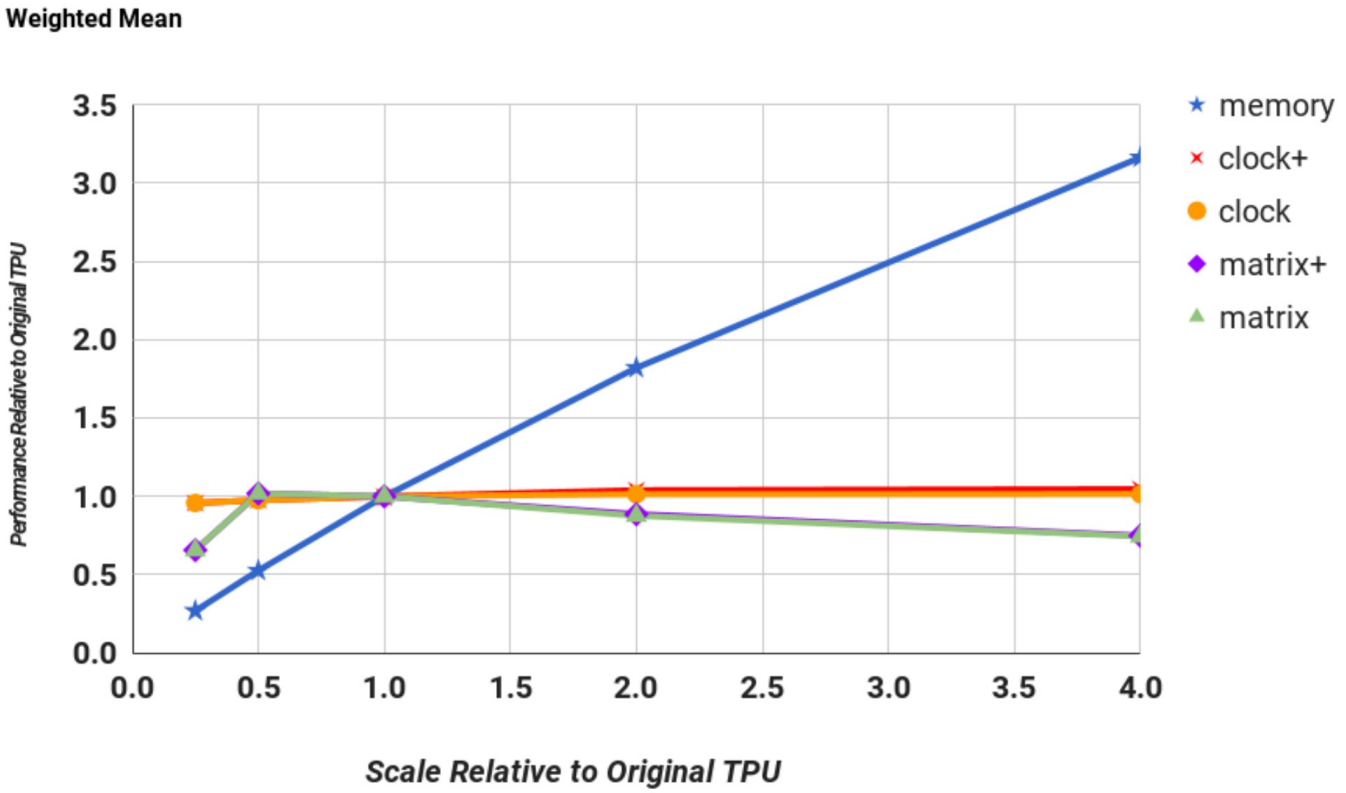
Performance – energy efficiency



Performance – energy proportionality

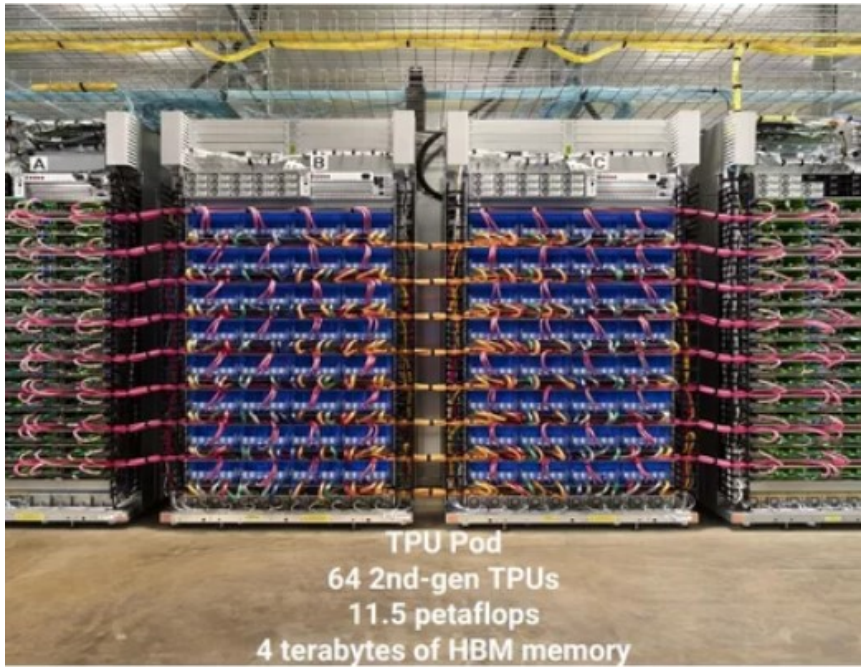


Design space exploration



TPU v2

- At HotChips 2017:
  - 2x 128x128x32b “mixed multiply units” (MXUs)
  - 64GB HBM
  - 64x TPU modules per “pod” → 4TB HBM
  - Some available in TensorFlow cloud svc



http://www.tomshardware.com/news/tpu-v2-google-machine-learning,35370.html

Microsoft - Catapult

Google v. Microsoft

- Why Google ASIC? Why Microsoft FPGA?
- Flexibility? Programmability?
- Cost and usefulness over time?

Takeaways

- Industry and academia have very different constraints

Takeaways

- Industry and academia have very different constraints

"Your **system-level** ideas needs to **deliver 2x or more**, or someone else should fund it"

"Get me 10X in 15 months"

Google

Takeaways

- Industry and academia have very different constraints
- Different goals may require fundamentally different tech

"Your **system-level** ideas needs to **deliver 2x or more**, or someone else should fund it"

"Get me 10X in 15 months"

Google

Takeaways

- Industry and academia have very different constraints
- Different goals may require fundamentally different tech
- Time and money dominate
  - (In academia, too!)

"Your **system-level** ideas needs to **deliver 2x or more**, or someone else should fund it"

"Get me 10x in 15 months"

Google