Determine Inference Efficiency using TOPS, SRAM, DRAM & Throughput

July 2019

Terminology

- MAC = multiply and accumulate unit
- TOPS = trillions or tera operations/second
  - TOPS is a measure of peak or maximum achievable performance
  - TOPS = compute frequency * operations/compute unit (1 MAC = 2 operations)
- Throughput is a measure of actual performance
  - For a specified model, batch size, image size
- Throughput/TOPS is a partial measure of the efficiency of a chip
  - TOPS is an indicator of the size of the compute hardware but does not indicate the size of the on-chip or off-chip memory

Inference Throughput and Cost Contributors

- MACs – more MACs increase TOPS and silicon area
- Architecture
  - organization of the MACs affects utilization
  - interconnection of MACs with memory affects utilization
- SRAM – highest bandwidth, greater cost/bit
- DRAM – lowest cost/bit, need bandwidth not bits; PHYs/BGA balls add cost

- GOAL: get the highest throughput, on a representative model, using the least MACs, least SRAM Megabytes & least DRAMs

Inference Chips ranked in descending order of ResNet-50 throughput

	TOPS (INT8)	Number of DRAM	ResNet-50 (batch=1)	ResNet-50 (batch=10++)
Groq	400	?	26,000	
Nvidia Tesla T4	130	8	961	4,365 (batch=128)
Nvidia Xavier AGX	32	8	480	
InferX X1	8.5	1 (8MB SRAM)	363	1,134 (batch=8)
Novumind	15	1	n/a	120 (batch=?)
Horizon Robotics Journey 2.0	1?	?	n/a	94 (batch=?)
Nvidia Jetson Nano	0.5 (FP)	2?	36 (FP16, b=1)	
Google Edge TPU	4	1?	21 (batch=1?)	

- Inference chips listed have published TOPS and ResNet-50 performance for some batch size
- ResNet-50 is a poor benchmark because it uses 224x224 images (megapixel is what people want) but it is the only benchmark given by most inference suppliers
- Unfortunately, almost no one provides information on the size of on-chip SRAM

Throughput/TOPS: Efficiency of MAC Utilization for ResNet-50

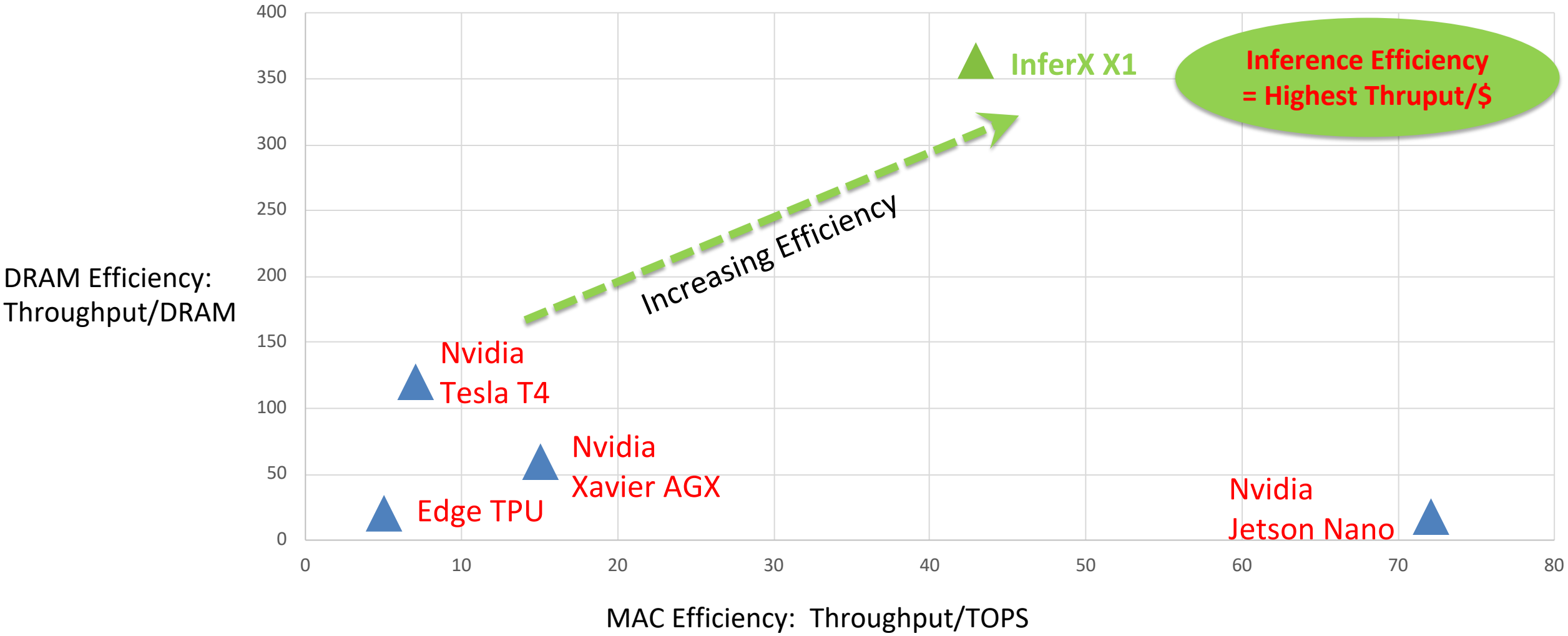
Throughput/TOPS (INT8)	Batch=1	Batch=10++
Nvidia Jetson Nano (FP)	72	
Groq	65	
InferX X1	43	133
Horizon Robotics Journey 2.0		94
Hailo-8	26	Can't do large batches
Nvidia Xavier AGX	15	
Nvidia Tesla T4	7	34
Google Edge TPU	5	
Novumind		8

Some architectures get more than 10x the throughput for similar numbers of TOPS!
Remember, the throughput is a function of (TOPS, SRAM, DRAM, architecture), not just TOPS

Throughput/DRAM: Efficiency of DRAM Bandwidth Utilization for ResNet-50

Throughput/DRAM	Batch=1	Batch=10++
InferX X1	363	1134
Nvidia Tesla T4	120	546
Nvidia Xavier AGX	60	
Novumind		120
Google Edge TPU	21	
Jetson Nano	18	

Throughput/TOPS & Throughput/DRAM for ResNet-50, batch=1



Throughput/TOPS & Throughput/DRAM for ResNet-50, *batch=10++*