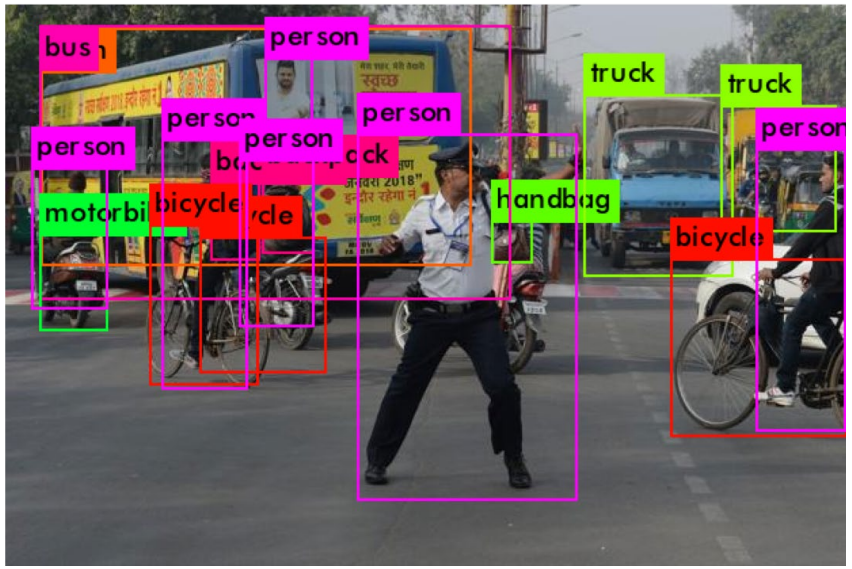
Accelerator Evaluation on Real Edge-Inference Applications

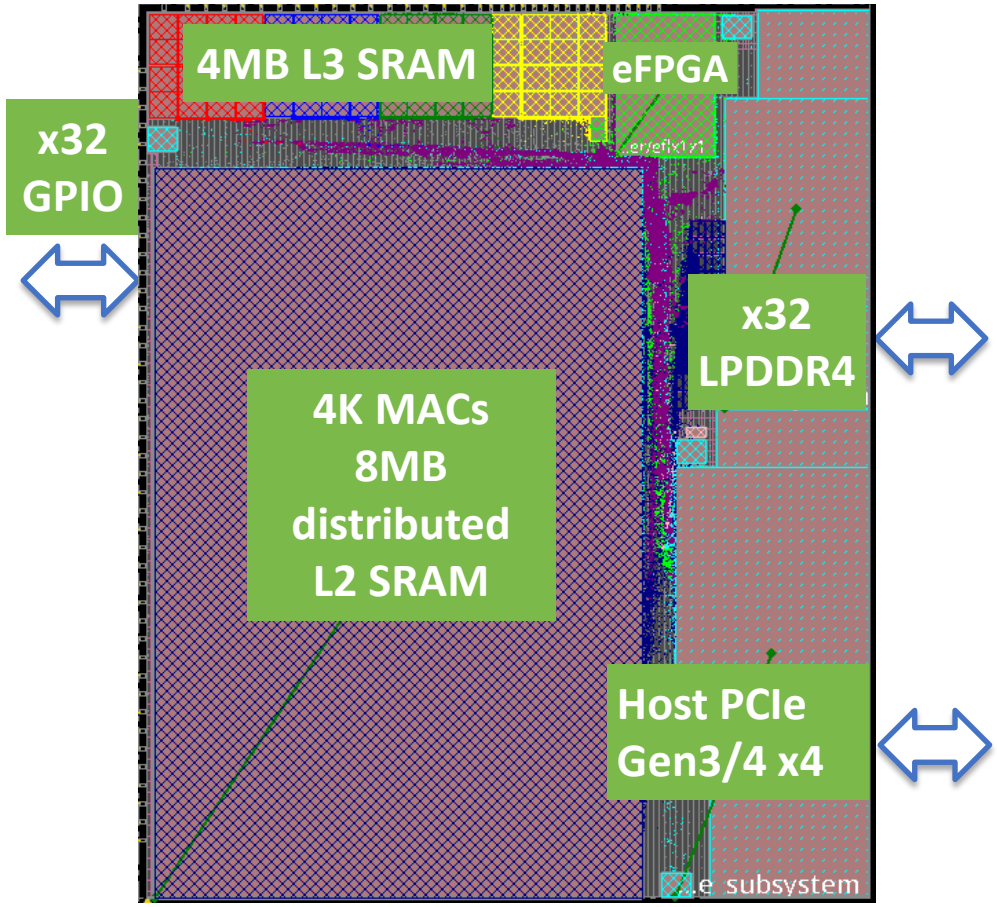
Vinay Mehta, Inference Technical Marketing Manager
Flex Logix Technologies, Inc.

vmehta@flex-logix.com

Linley Spring Processor Conference
April 6-9, 2020, Santa Clara, CA



InferX™ X1



- *54mm² TSMC 16FFC*
- *933MHz Operation*

- 4K MACs @ INT8
  - 2K MACs @ BF16
  - Winograd acceleration for INT8
- 8MB L2 SRAM + 4MB L3 SRAM
- x32 LPDDR4 (14.9GB/s peak BW)

- 13.5 W (max)

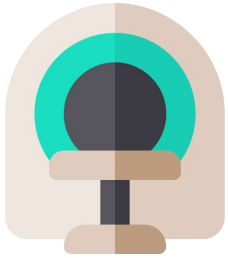
- Partners: TSMC, GUC, Synopsys, Arteris, Analog Bits, Cadence, Mentor

- Available as Chip & PCIe card Q3

Segmenting Edge Customers by CNN Complexity

CCTV with shoplifting alerts

Attention monitoring for ADAS



Medical segmentation and classification

Quality assurance and inspections

Image denoising

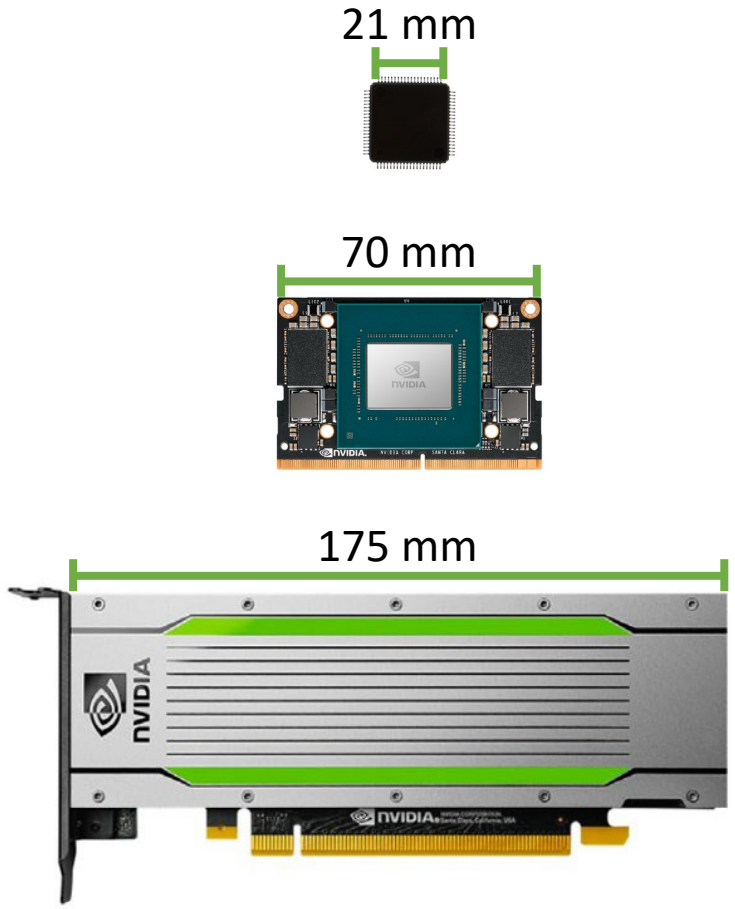
Perception — **Learned DSP**

All share requirements for real-time (**streaming: batch=1, low latency**) with large input size

Evaluating Like a Customer

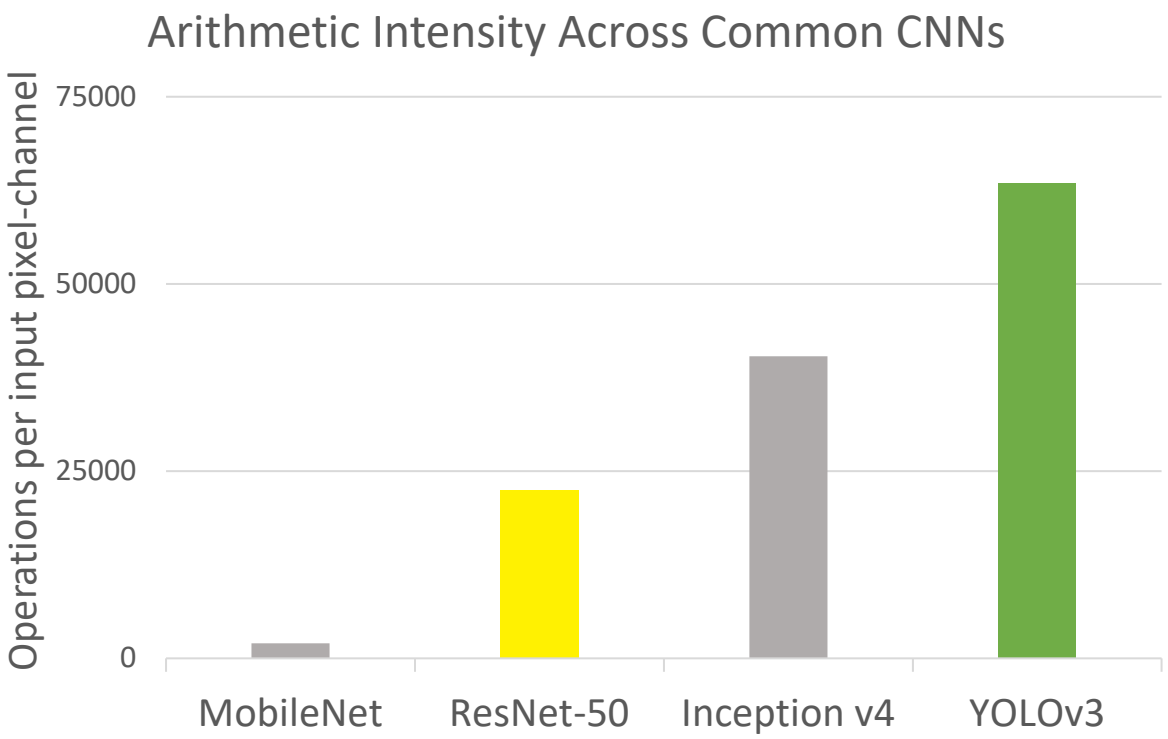
- Choose a representative workload
- Be clear with metrics (eg, latency and throughput)
- Tie performance back to the design requirements

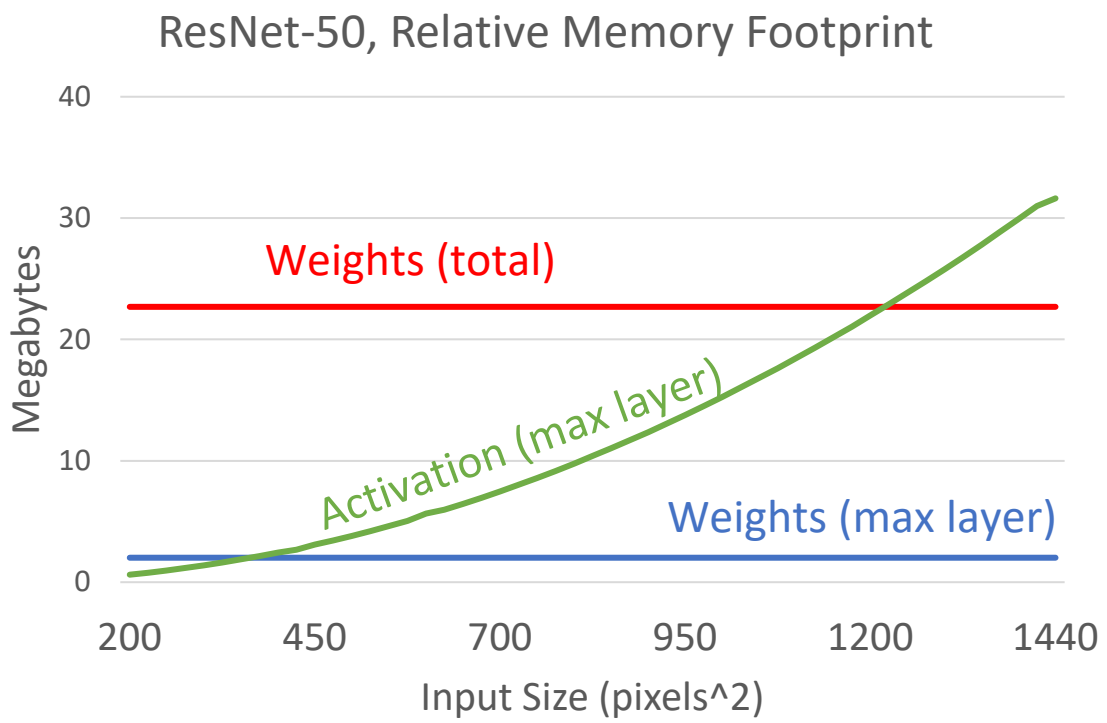
Customers' Evaluation Involves More Than Performance



	TDP	Die Size
InferX X1	7-13.5 W	54 mm^2
Nvidia Xavier NX	15 W	350 mm^2
Nvidia Tesla T4	75 W	545 mm^2

Right Benchmark: Characterizing Models (Actual Model Still Best!)



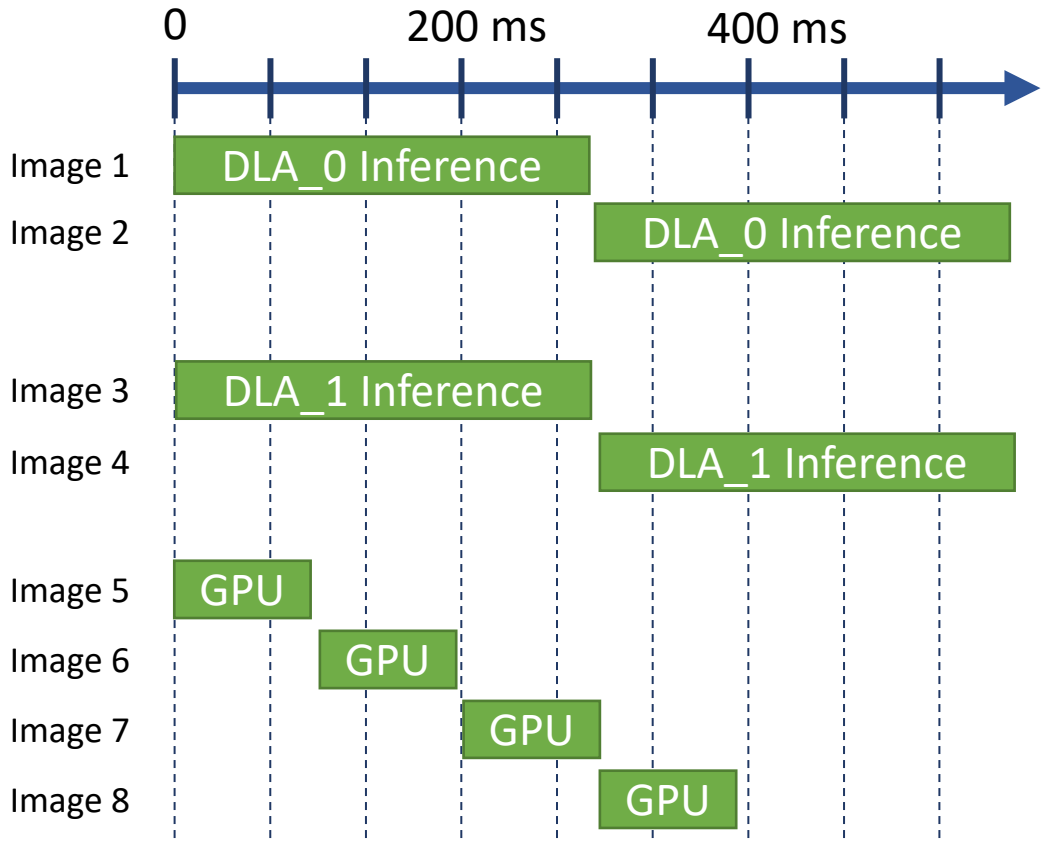


Reporting Benchmarks: Single Stream vs Pooling

YOLOv3-1440 INT8, b=1 on Nvidia Jetson NX

	Latency (ms)	FPS
DLA_0	290	3.4
DLA_1	290	3.4
GPU	95	10.5
		+
		"17.3 FPS"

Assumes pool of data



Reporting Benchmarks: Input Data is a Stream

YOLOv3-1440 INT8, b=1 on Nvidia Jetson NX

	Latency (ms)	FPS
DLA_0	290	3.4
DLA_1	290	3.4
GPU	95	10.5
		+
		"17.3 FPS"

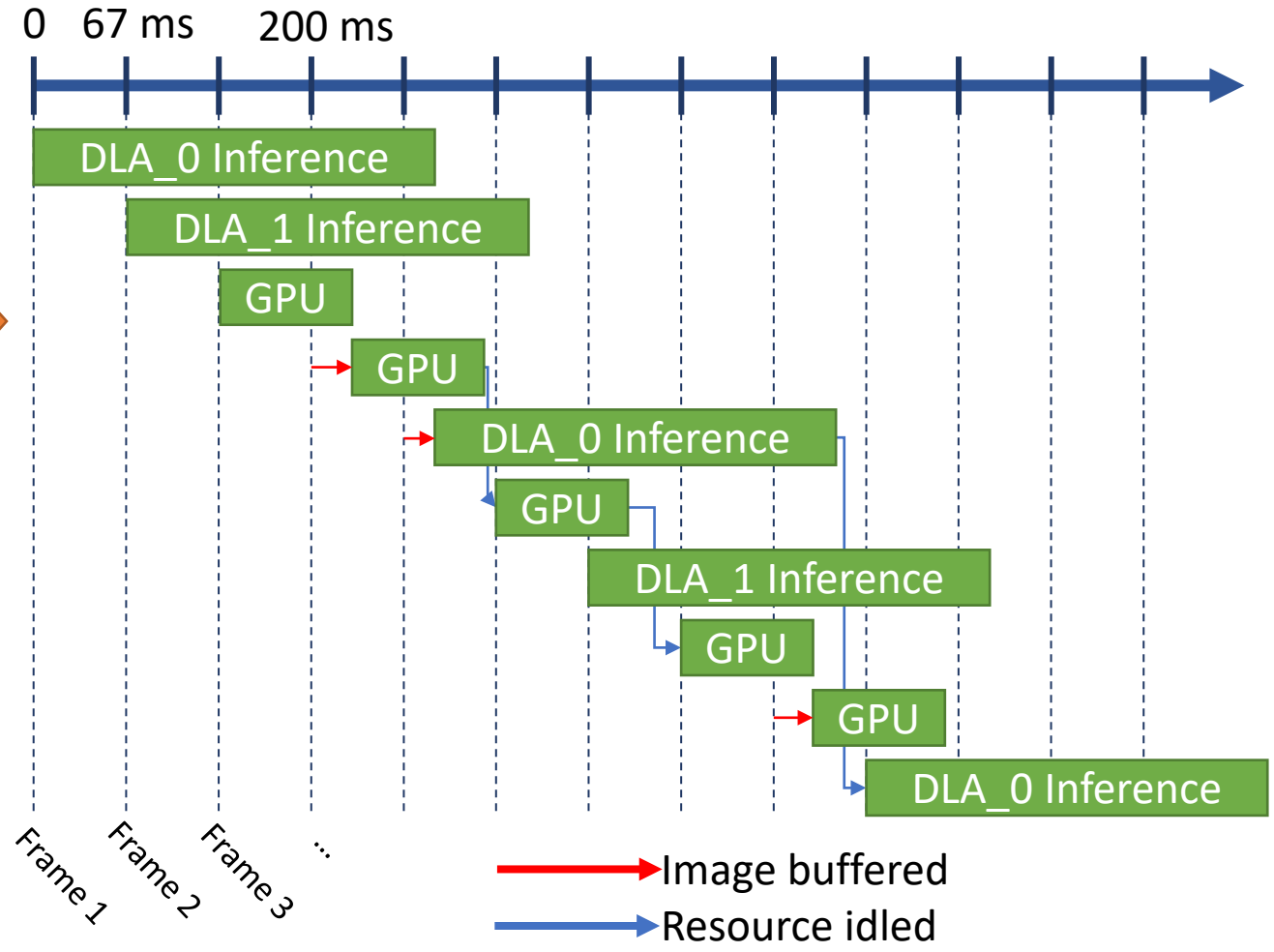
Reporting Benchmarks: Input Data is a Stream

YOLOv3-1440 INT8, b=1 on Nvidia Jetson NX

	Latency (ms)	FPS
DLA_0	290	3.4
DLA_1	290	3.4
GPU	95	10.5
	+	
	"17.3 FPS"	

Can I run inf. on a 15 FPS sensor?

(no, not as you expect)



Throughput Does Not Correspond to Effective Latency

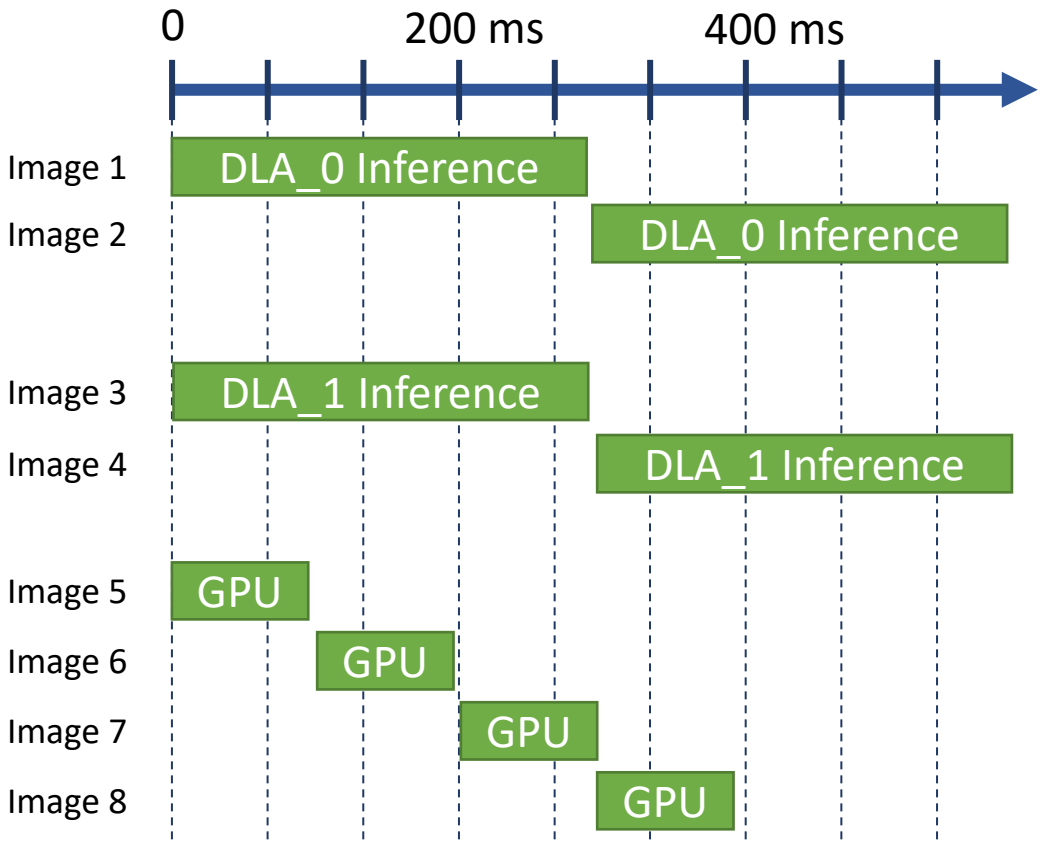
YOLOv3-1440 INT8, b=1 on Nvidia Jetson NX

	Latency (ms)	FPS
DLA_0	290	3.4
DLA_1	290	3.4
GPU	95	10.5

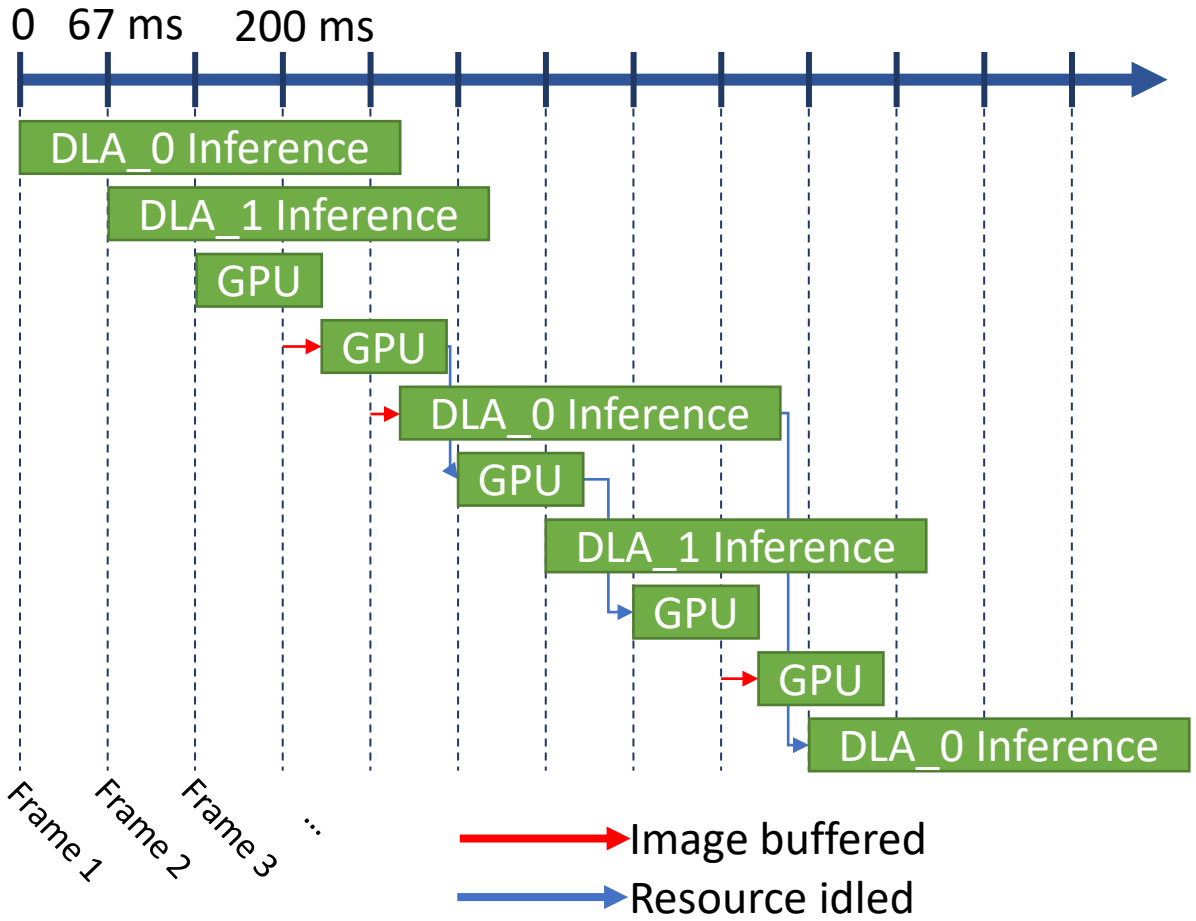
- Cannot use all available resources on same inference
- Difficult to schedule processing engines
- Accessible performance demonstrated by latency

Datacenter vs Edge Benchmarks: Summary

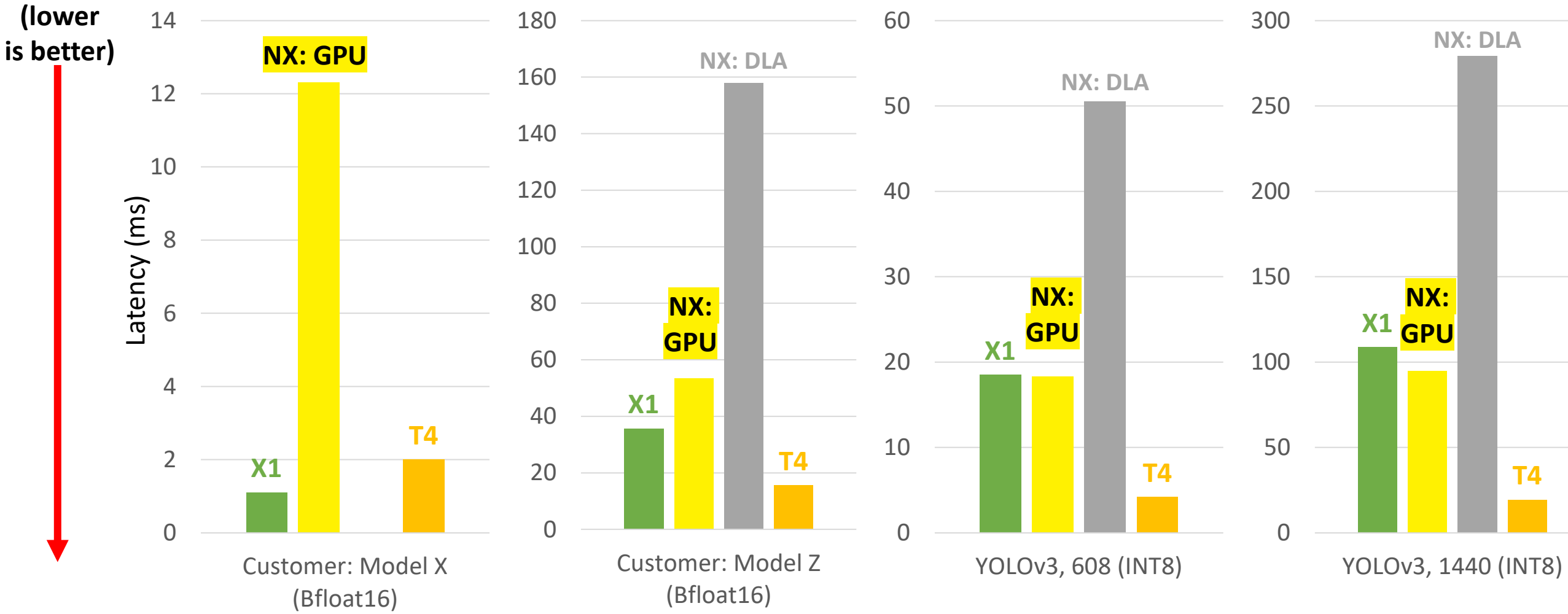
Datacenter Inference: Out-of-Order, No Fixed FPS



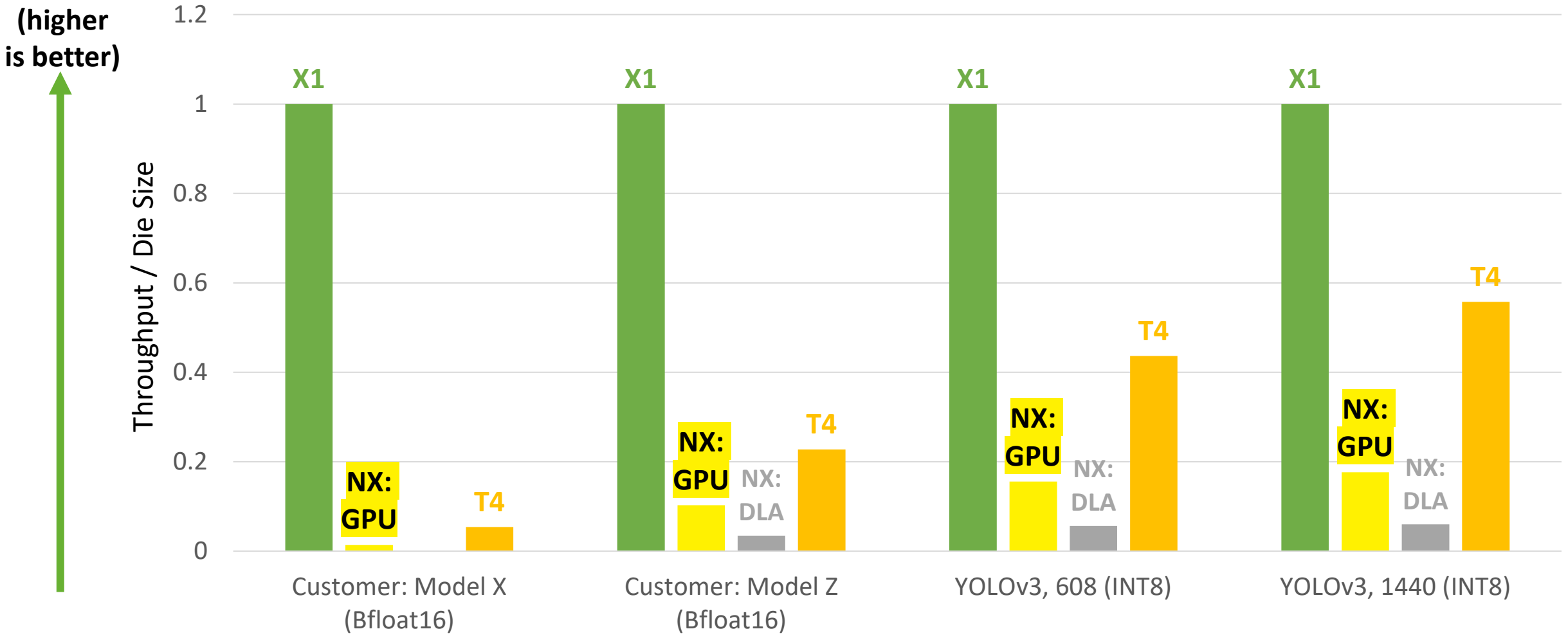
Edge Data: Single Stream, 15 FPS



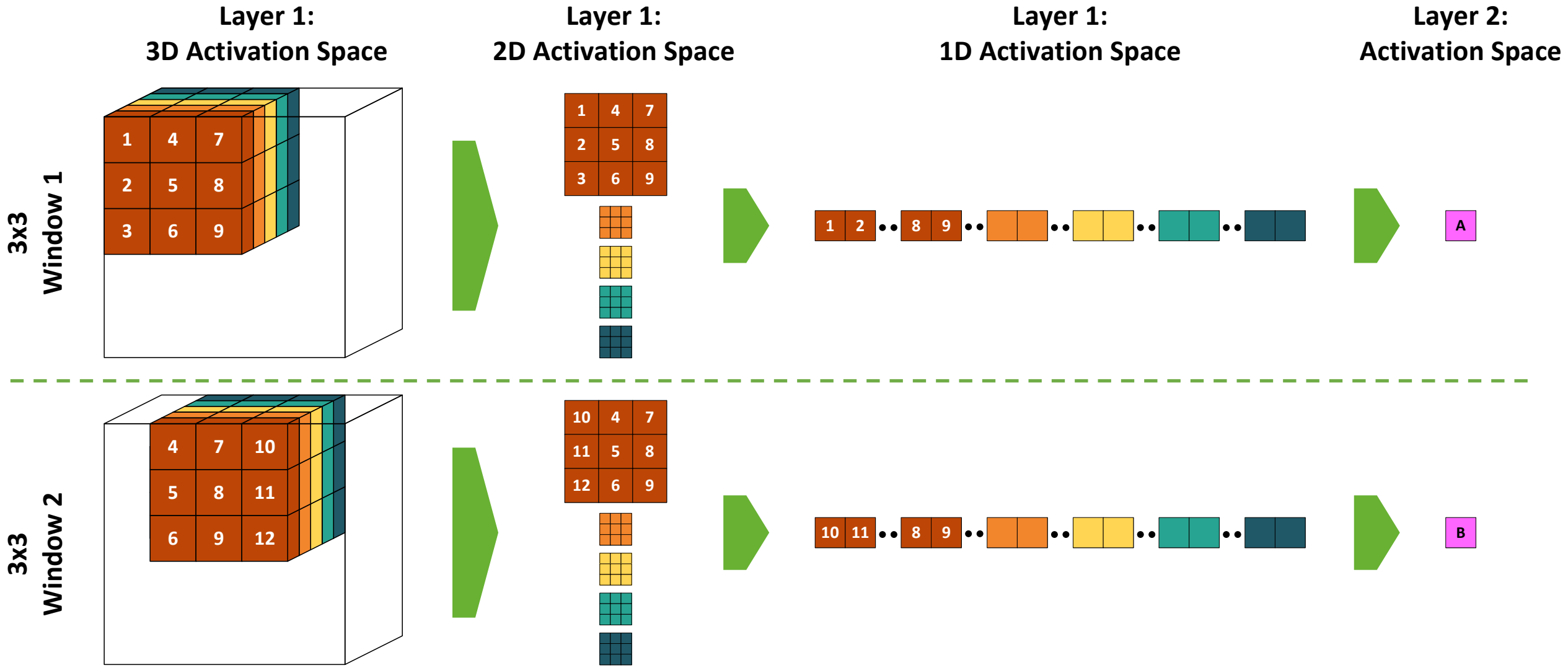
Real World Benchmark: Latency (If Power and Cost Didn't Matter…)



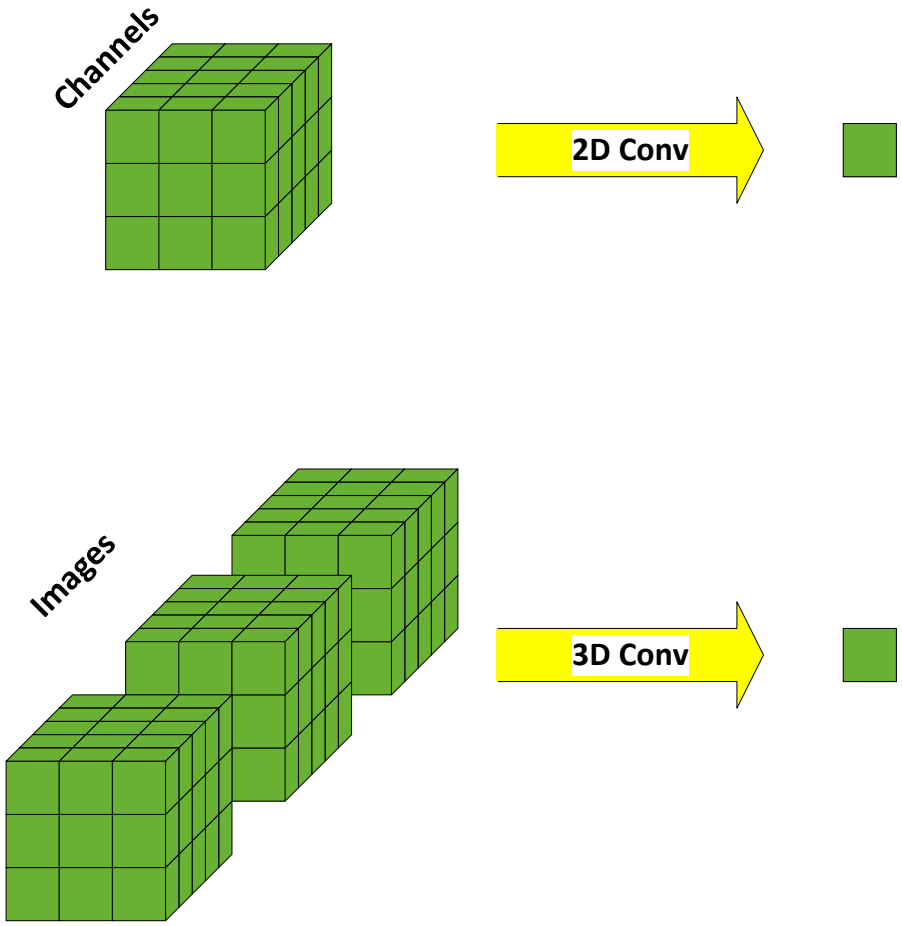
InferX X1 Has Superior Performance for the Price

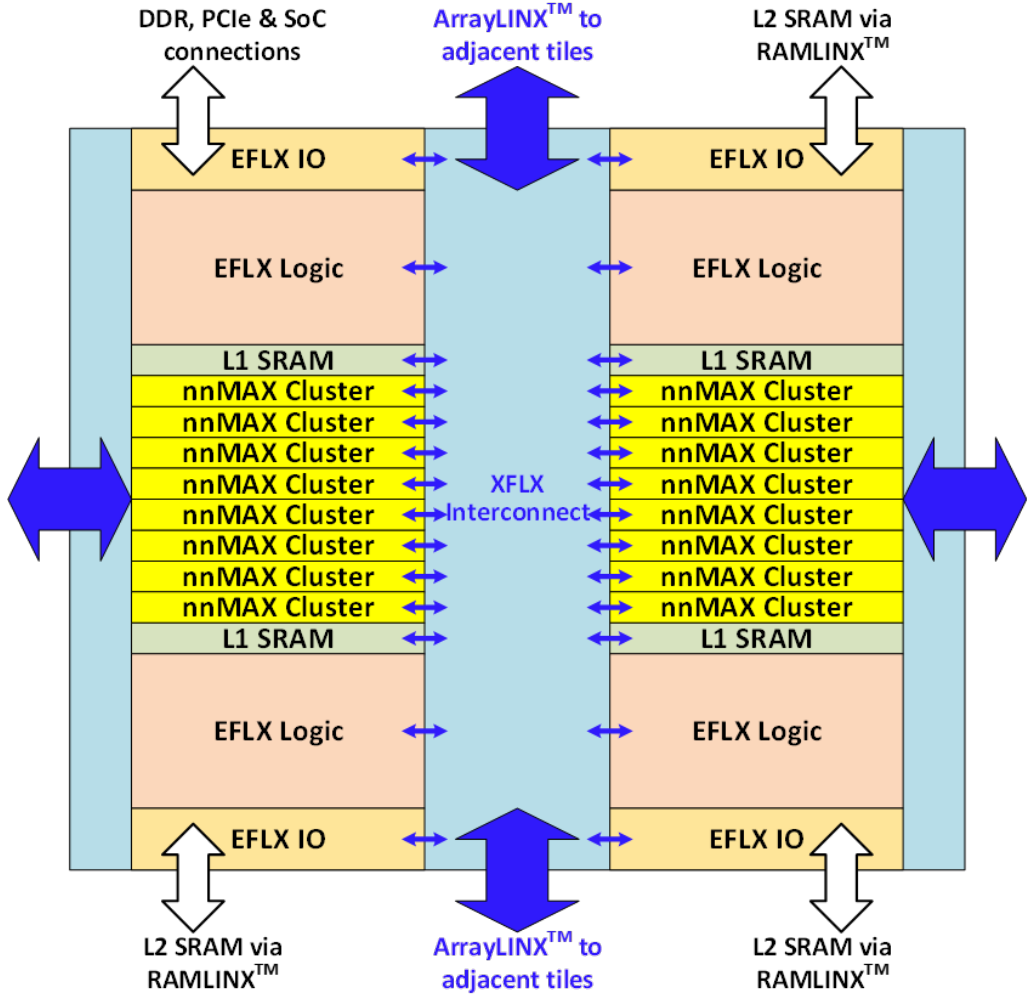


Key to X1 Efficiency is in Data Packing

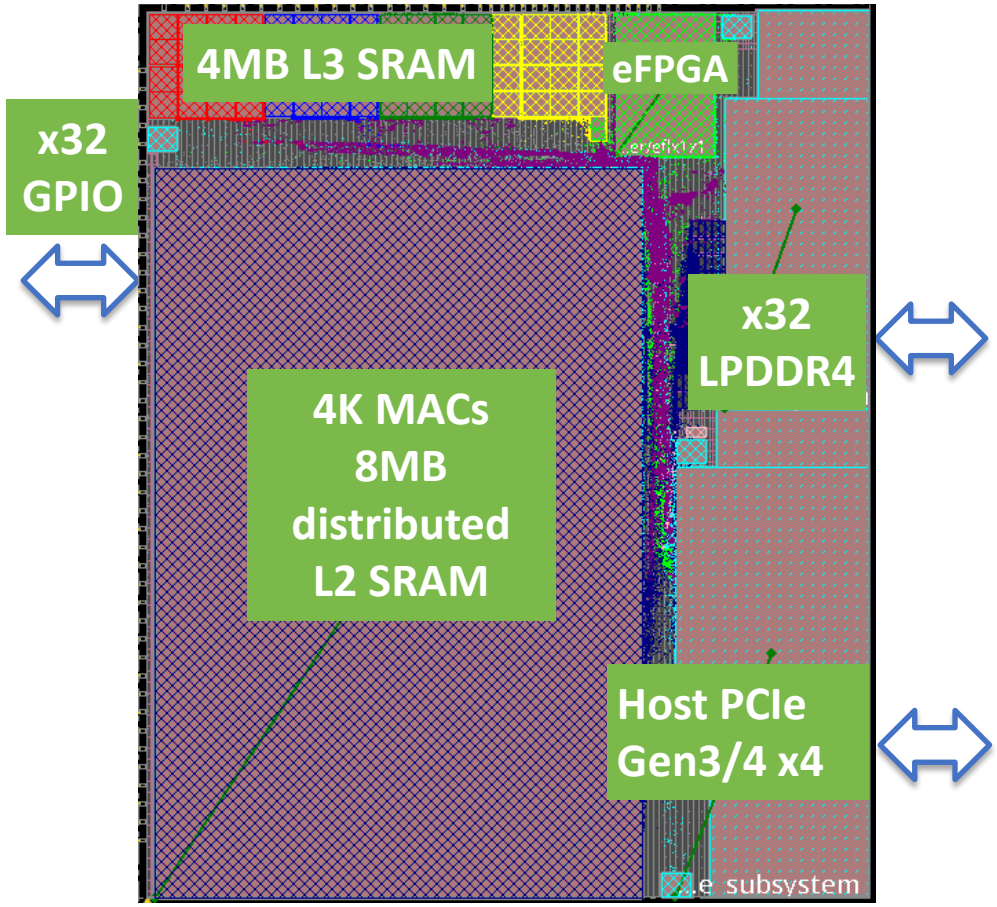


X1 Flexibility Achieves Efficiency Where Others Can't, Eg. 3D Convolutions





InferX™ X1



- *54mm² TSMC 16FFC*
- *933MHz Operation*

- 4K MACs @ INT8
  - 2K MACs @ BF16
  - Winograd acceleration for INT8
- 8MB L2 SRAM + 4MB L3 SRAM
- x32 LPDDR4 (14.9GB/s peak BW)

- 13.5 W (max)

- Partners: TSMC, GUC, Synopsys, Arteris, Analog Bits, Cadence, Mentor

- Available as Chip & PCIe card Q3

Appendix

Benchmark Data

		Latency (ms)					Throughput (inf/s)					Throughput/Die Size		
		NX-DLA (1)	**NX-DLA (2)**	**NX-GPU**	**T4**	**X1***	**NX-DLA (1)**	**NX-DLA (2)**	**NX-GPU**	**T4**	**X1***	**NX-DLA (1)**	**NX-GPU**	**T4**
Customer Model Z	BF/FP16	157.8	163.1	53.5	15.54	35.7	6.3	12.3	18.7	64.4	28.0	0.0349	0.1030	0.2276
Customer Model X	BF/FP16	[redacted]		12.3	2.01	1.1	[redacted]		81.3	497.5	909.1	[redacted]	0.0138	0.0542
YOLOv3-608	INT8	50.5	53.7	18.3	4.2	18.5	19.8	37.2	54.6	238.1	54.1	0.0565	0.1560	0.4364
YOLOv3-1440	INT8	279.2	289.8	94.9	19.3	108.6	3.6	6.9	10.5	51.8	9.2	0.0600	0.1766	0.5575

*performance estimate