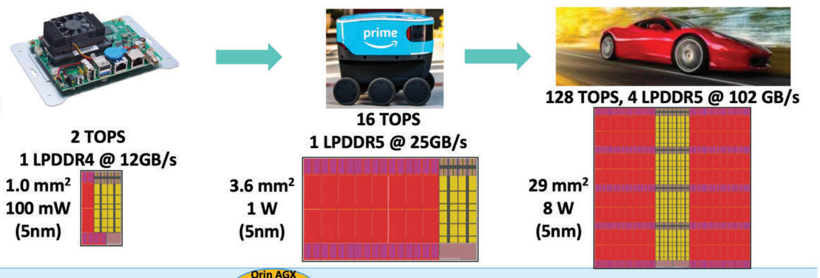
InferX AI Solution

World Class AI in your SoC

InferX IP is linearly scalable (computed based on 5nm, 1 GHz)

Over **4x TOPS efficiency** (IPS/TOPS) compared to Orin AGX


128 TOPS, 4 LPDDR5 @ 102 GB/s

2 TOPS
1 LPDDR4 @ 12GB/s
1.0 mm^2
100 mW
(5nm)

16 TOPS
1 LPDDR5 @ 25GB/s
3.6 mm^2
1 W
(5nm)

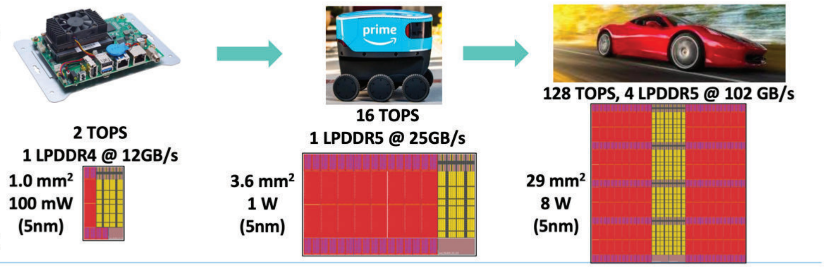
29 mm^2
8 W
(5nm)

YOLOv5s (640×640)	35 IPS	Orin AGX 135 TOPs 125 IPS	234 IPS		1250 IPS
YOLOv5l6 (1280×1280)	1.6 IPS		14 IPS	Orin AGX 135 TOPs 31 IPS	115 IPS
ResNet50 (512x512)	16 IPS		144 IPS		1060 IPS
ResNet50 (1024x1024)	4 IPS		34 IPS		230 IPS
DETR 2020 (Transformer) (512x512)	17 IPS		211 IPS		603 IPS
DETR 2020 (Transformer) (1024×1024)	2.5 IPS		33 IPS		178 IPS

InferX IP uses bandwidth efficiently in a shared-memory system

InferX does **not** require dedicated DRAM: friendly to shared-memory system

- 4K transfers to/from DRAM
- Performance not very sensitive to latency
- Use a fraction of available BW
- Over **4x DDR efficiency** (IPS/GB) compared to Orin AGX


128 TOPS, 4 LPDDR5 @ 102 GB/s

2 TOPS
1 LPDDR4 @ 12GB/s
1.0 mm^2
100 mW
(5nm)

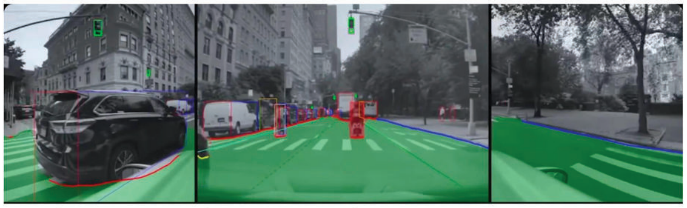
16 TOPS
1 LPDDR5 @ 25GB/s
3.6 mm^2
1 W
(5nm)

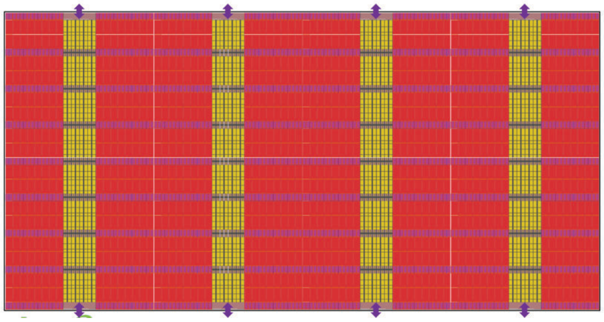
29 mm^2
8 W
(5nm)

	DRAM BW & Capacity		DRAM BW & Capacity		DRAM BW & Capacity
YOLOv5s (640×640)	3 GB/s, 256MB (35 IPS)	Orin AGX 204.8 GB/s 125 IPS	14 GB/s, 256MB (234 IPS)		65 GB/s, 256MB (1250 IPS)
YOLOv5l6 (1280×1280)	2 GB/s, 512MB (1.6 IPS)		15 GB/s, 512MB (14 IPS)	Orin AGX 204.8 GB/s 31 IPS	60 GB/s, 512MB (115 IPS)
ResNet50 (512x512)	2 GB/s, 128MB (16 IPS)		18 GB/s, 128MB (144 IPS)		52 GB/s, 256MB (1060 IPS)
ResNet50 (1024x1024)	2 GB/s, 256MB (4 IPS)		16 GB/s, 256MB (34 IPS)		67 GB/s, 256MB (230 IPS)
DETR 2020 (Transformer) (512x512)	2 GB/s, 128MB (17 IPS)		14 GB/s, 128MB (211 IPS)		85 GB/s, 256MB (603 IPS)
DETR 2020 (Transformer) (1024×1024)	2 GB/s, 128MB (2.5 IPS)		6 GB/s, 128MB (33 IPS)		28 GB/s, 256MB (178 IPS)

InferX IP scales to over 1000 TOPS for next-generation ultra-HD AI models

1024 TOPS
HBM @ 820 GB/s
240 mm^2
64 W
(5nm)





YOLOv5l6 – HD (1280×1280)	576 IPS
YOLOv5l6 – Ultra HD (5120×2560)	120 IPS
ResNet50 – HD (1024x1024)	1950 IPS
ResNet50 – Ultra HD (4096x2048)	232 IPS
DETR 2020 (Transformer) – HD (1024x1024)	1950 IPS
DETR 2020 (Transformer) – Ultra HD (2048x2048)	120 IPS

Other Nodes

Ask about AI benchmarks on other nodes such as N3 and 18A.

InferX #1 AI PPA

TOPS is a measure of PEAK AI throughput (2 times the number of Multiplier-Accumulator operations per second).

There are Dense TOPS and Sparse TOPS - Sparsity sacrifices accuracy. InferX TOPS are Dense TOPS.

What matters in your SoC is getting the inferences/second your application needs for the NN model and image size you want at the smallest silicon area and power.

As the data to the left shows, InferX outperforms Orin AGX using far fewer TOPS. InferX is more efficient. InferX is 4 to 10 times more inferences/second than Orin AGX for the same number of TOPS.

Adapt to New Models in Field

When you are designing your AI SoC you may be focused on beating Nvidia, but your IP options are DSP-derived VLIW architectures.

Unlike these other architectures, InferX is very programmable so it is possible to upgrade post-silicon, in the field to run any new operator and model that is created during the operating life of your chip.

InferX incorporates eFPGA as well as Tensor Processors. The eFPGA can be programmed to run anything.