InferX™ Vision AI IP+SW
Orin AGX or better in your SoC

flexlogix reconfigurable®

Fast vision AI Inference at low cost, low power for your SoC

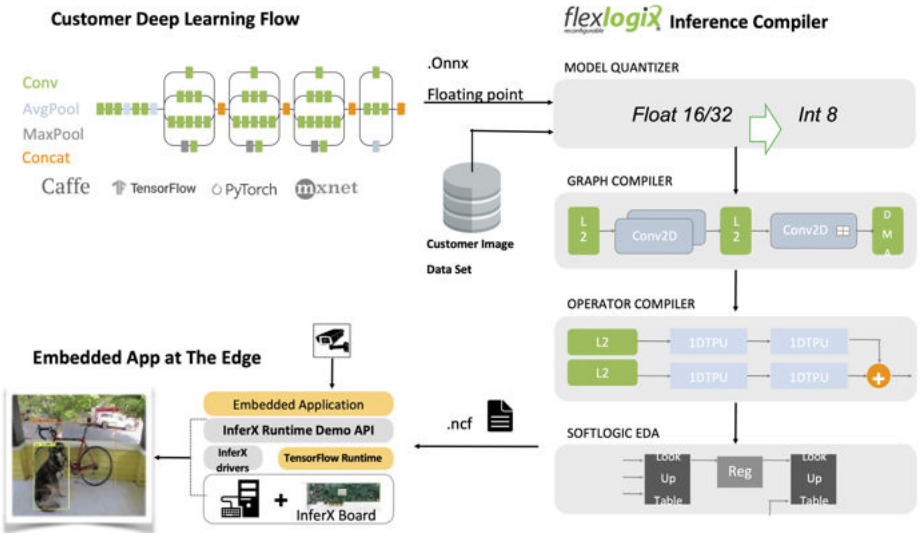
- ✔ 80% hardwired, 100% reconfigurable delivers GPU performance at less $/W/size
- ✔ Only 10's of square millimeters of TSMC 16, 7, 5, 4 and 3nm nodes
- ✔ 10x cheaper, 10x lower power and much smaller than Orin AGX with less DRAM BW
- ✔ Run super-resolution models like Yolov5L6 1280x1280 pixels at 30 frames/second
- ✔ Optimized for batch=1 and very high accuracy.
- ✔ The InferX Compiler converts high level neural network models to InferX code
- ✔ InferX can run multiple models simultaneously
- ✔ eFPGA core means we can always adapt for any new operators & activations
- ✔ InferX runs Transformers efficiently and accurately
- ✔ AXI interface, -40C to +125C design, very high DFT coverage, full test vectors
- ✔ Dozens of chips have worked first time with our IP and Software

InferX Inference Compiler

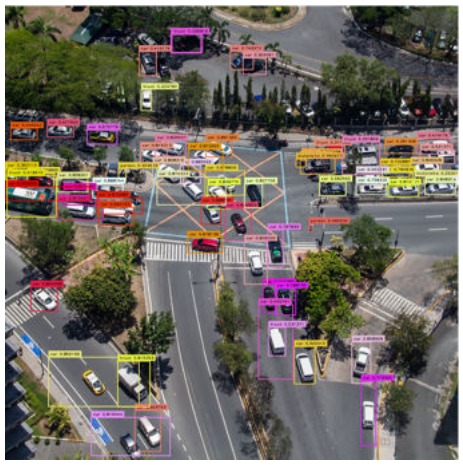
InferX Compiler takes in Neural Network models in high level formats. Scripts convert to INT8 for high performance with accuracy within <1% of the mAP of FP16. Quantization is done automatically. No shortcuts are taken on precision: no pruning or other modifications of your model.

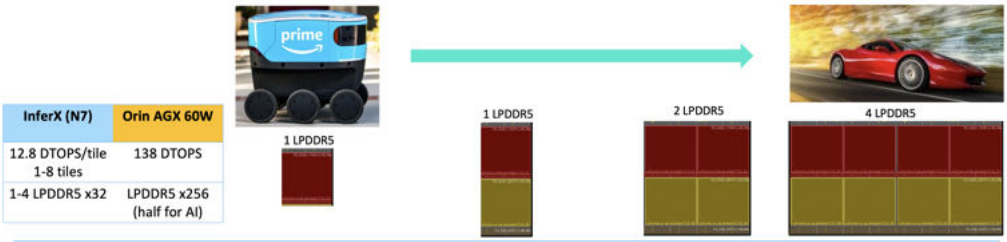
The InferX compiler generates the InferX code for the model which you can load and execute on InferX.

You can compile multiple models and instruct InferX which one to execute when. The InferX Compiler is in Beta and is available for evaluation under Software License.



Run super-resolution models on megapixel images (N7, batch=1)





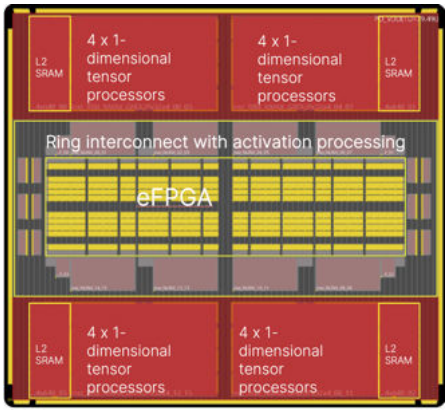
InferX (N7)	Orin AGX 60W
12.8 DTOPS/tile 1-8 tiles	138 DTOPS
1-4 LPDDR5 x32	LPDDR5 x256 (half for AI)

		1 LPDDR5	2 LPDDR5	4 LPDDR5
DETR 2020 (Transformer) (1024×1024)		19 IPS	39 IPS	77 IPS	147 IPS
YOLOv5s (640×640)	Orin AGX 125 IPS	200 IPS	330 IPS	716 IPS	1123 IPS
YOLOv5l6 (1280×1280)		12 IPS	19 IPS	Orin AGX 31 IPS 43 IPS	100 IPS
ResNet50 (1024x1024)		29 IPS	39 IPS	84 IPS	197 IPS

InferX Hardware

Scalable & Reconfigurable to Optimize for your Needs

InferX Tensor Processing Tile



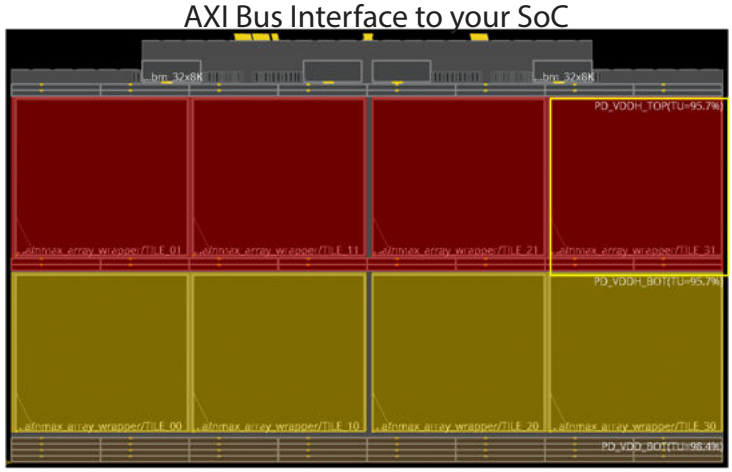
InferX is a tensor processing tile. A single tile has 16 hard-wired 1-dimensional tensor processors (TPU) capable of processing in INT8, INT16 and BF16 modes. Each TPU contains a 2-dimensional matrix for holding AI weights. The TPUs connect via a programmable interconnect so they can be configured differently for each DSP or AI operation to achieve maximum performance. Reconfiguration takes a few microseconds. Each Tile has 16 Dense TOPs (N5): efficient utilization of TOPs and minimum DRAM bandwidth also impact throughput. We are more efficient than Orin AGX: more throughput from fewer Dense TOPs.

InferX uses eFPGA for programmable state machine control of the tensor processors . This allows configuration of the hardware for optimal execution of a given task. And since eFPGA is programmable InferX can easily adapt to new operators as they are developed (common in AI). Flex Logix does the coding of the eFPGA with our SoftLogic team for AI and DSP operators.

InferX, like all Flex Logix IPs, is designed to operate over SS-FF, full voltage range and -40 to +125C Tj. InferX is available on the main TSMC Finfet nodes: 16, 7, 5/4 and 3nm. We have designed IP in 16/7/5.

InferX Arrays

AXI Bus Interface to your SoC



InferX is delivered as an array of tiles from 1x1 to as large as desired. The array has an AXI bus interface to connect to your SoC. An 8 tile array, organized as 2x4, is shown below. In addition to the AXI bus interface, there are interfaces for DFT, configuration and clocking.

InferX performance is linear: An array with N tiles will run about N times faster than a single tile array. InferX is silicon proven in TSMC 16FFC (below).

TSMC IP Alliance Member

Flex Logix® is a TSMC IP Alliance Member based on the work it has done with TSMC over many years to develop IP meeting TSMC9000 compliance for design methodology, validation in silicon & documentation. Flex Logix has implemented IP in TSMC 40, 28, 16, 12 and 7nm; and has started on 5nm. Dozens of customer chips are working in silicon with our IP; dozens more are in design.

About Flex Logix

- Our eFPGA is licensed for 40 chips with >20 working in Silicon; our software has been used for years
- Our CEO has managed up to 500 people and took a startup from 4 people to IPO to $2 Billion Market Cap
- Our Executives have extensive industry experience and industry recognition, including the Outstanding Paper Award at ISSCC
- Our technical team is a combination of silicon engineering, software development, and architecture
- We have >60 issued US patents and patent applications; as well we have patents in Europe and China
- We have raised >$90 Million; our lead investors are Lux Capital, Eclipse Ventures and Mithril Capital.

www.flex-logix.com

03/2023