



# Real-time High-Res Vision with X1

Real-time High-Res Vision requires high bandwidth support for complex deep learning models operating with small batch sizes in real time. InferX X1 is designed to solve these problems.

## InferX™ X1 Edge Inference Accelerator

The InferX™ X1 Edge Inference Accelerator is optimized for the processing of real-time high-res vision workloads at the edge.

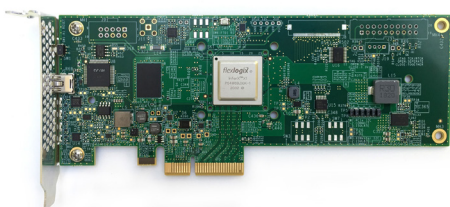
Focused on solving challenging machine vision problems in industrial, smart city, manufacturing, transportation, healthcare, agriculture and other markets. The X1 is designed to run complex AI models with high throughput and accuracy and without the complexity of GPU programming.

High accuracy AI vision models are characterized by deep networks with over a hundred layers, dozens of parallel channels in each layer and multiple operator types. The X1 is ideally suited for processing these workloads in a low power envelope. Also, model accuracy targets may require the use of mixed precisions, including INT8, INT16 and BF16, important capabilities of the X1.

The X1 dynamic tensor processor array offers ASIC speed and efficiency while providing model flexibility, through the use of reconfigurable datapath & control logic technology. The allows the X1 to quickly adopt and deploy new edge inference model technologies via field updates, thus future-proofing designs.

### Benefits

- Supports complex object detection models including YOLOv3, YOLOv4, YOLOv5
- Sub 10W Power Dissipation
- Small form-factor fits in compact space and low power envelopes
- Comparable performance to high power GPU solutions on a range of important object detection models
- Real time image processing
- Dynamic architecture future-proofs designs
- Best Inference/\$\$ and Inference/W
- INT8, INT16, BFloat16 support—can mix between layers
- Trained ONNX models are easily ported to X1 for execution



X1P1



X1M



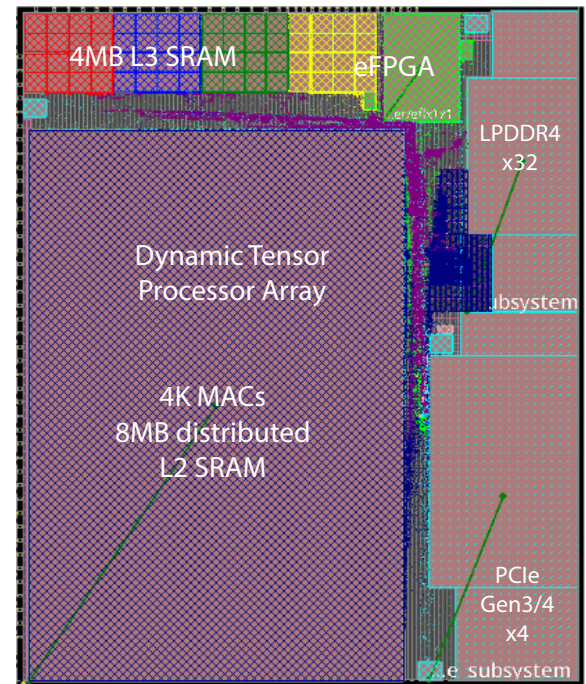
X1

## X1 Architecture


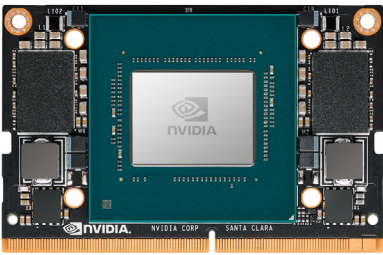
The InferX™ X1 contains a dynamic tensor processor array with 4K MAC units and 12 Megabytes of on-chip SRAM. The X1 also includes connectivity to external LPDDR4 DRAM for model weight, configuration and internal activation storage and Gen3/4 PCI Express for connectivity to a host processor.

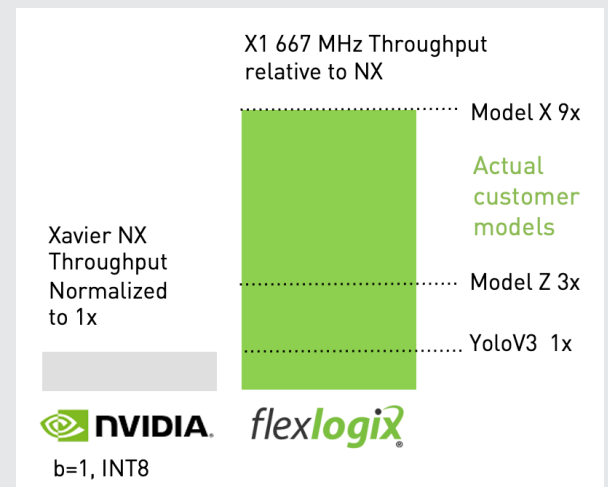
The X1 Edge Inference accelerator approach supports a choice of host architecture (x86, Arm), operating system and system features including easy integration of various sensor input types (cameras, IR, Ultrasonic, RF, etc.) and communication standards (Ethernet, USB, Wi-Fi, etc.).

The InferX X1 accelerator is designed to support both existing and future AI/ML models through the dynamic tensor processor array. The inherent reconfigurability of this architecture provides future-proofing for customers whose edge inferencing workloads are continuing to evolve and improve.



## Performance: InferX Beats Xavier NX at Lower Power, Size

 21 mm	7-13.5W	54mm <sup>2</sup>
 70 mm	15W	350mm <sup>2</sup>



To learn more please contact us: [info@flex-logix.com](mailto:info@flex-logix.com) or visit [www.flex-logix.com](http://www.flex-logix.com)

Copyright © 2022 Flex Logix Technologies, Inc. InferX, Flex Logix, are Trademarks of Flex Logix.  
All other names mentioned herein are trademarks or registered trademarks of their respective owner.