Accelerate low-latency edge AI inference with higher system-level efficiency and lower power vs. CPU, GPU and FPGA solutions

The Blaize® Xplorer® X1600E accelerator is an enterprise grade designed for AI inference at the edge via easy plug-in EDSFF interface, enabling servers and custom products to easily integrate AI inference. The X1600E accelerator is based on the Blaize Graph Streaming Processor (GSP®) architecture that enables new levels of processing power with energy efficiency ideal for AI inferencing workloads at the edge. The EDSFF X1600E accelerator easily integrates into a 1U rack system for large AI inference deployment.

With low power, low latency, and more efficient use of memory, the X1600E can be used for computer vision applications and new AI inferencing solutions across a range of edge smart vision use cases, like autonomous optical inspection, traffic and parking management and more.

### Programmability to Build Complete AI Apps, Keep Pace with Rapid Evolution of AI Models

The X1600E, is a software defined AI inference accelerator, making it easy to update and maintain after deployment. The X1600E GSP architecture is designed to run efficiently in a streaming fashion, and it is fully programmable via the Blaize® Picasso® SDK and AI Studio. The hardware and software are purpose-built to enable developers to build entire edge AI inference applications, optimized for deployment and consistent updates by end users.
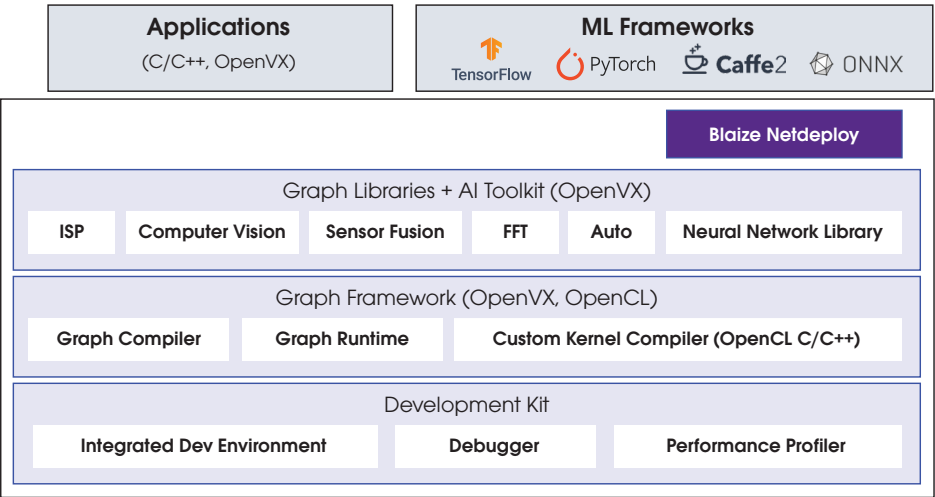
Edge and Enterprise Servers & Applications

- Smart Parking & Traffic Management
- Smart Retail
- Warehouse and Factory Safety
- Autonomous Optical Inspection
- Network Video Recorders
- Security Systems

Features

- 1 Blaize 1600 SoC with 16 GSP cores, providing 16 TOPs
- Soft ISP available to run on Blaize 1600 SoC
- 4 GB LPDDR4
- PCIe Gen 3.0, 4 lanes
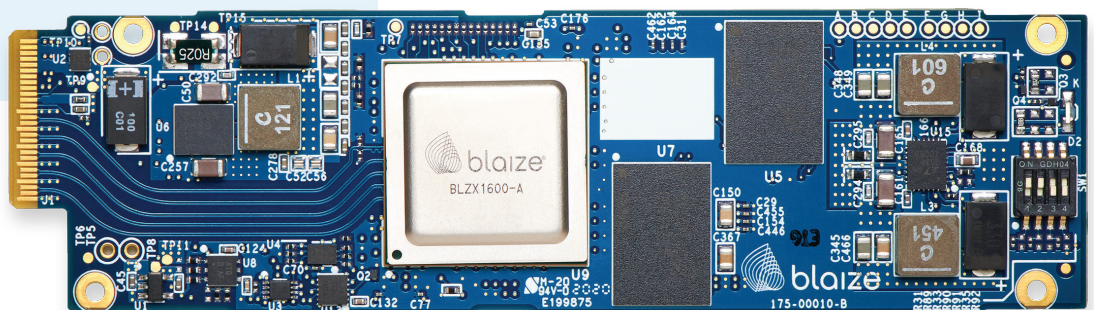
Blaize® Picasso® Software Development Kit Accelerates AI Development Cycle

- Extensive support for Machine Learning frameworks and modern neural networks
- NetDeploy toolkit to automatically quantize, compress and optimize neural networks for resource-constrained environments
- Extensive libraries to build end-to-end AI applications, integrate soft ISP, Tracker, Sensor Fusion and FFT
- Standard Languages: OpenVX extended, OpenCL C/C++ kernels to build custom layers

Applications (C/C++, OpenVX)	ML Frameworks: TensorFlow, PyTorch, Caffe2, ONNX

Blaize Netdeploy

Graph Libraries + AI Toolkit (OpenVX)

ISP	Computer Vision	Sensor Fusion	FFT	Auto	Neural Network Library

Graph Framework (OpenVX, OpenCL)

Graph Compiler	Graph Runtime	Custom Kernel Compiler (OpenCL C/C++)

Development Kit

Integrated Dev Environment	Debugger	Performance Profiler

Specifications

Features	Description	X1600E
GSP	1 Blaize GSP 1600 SoCs, Data formats INT8, INT16, BF16, FP16, FP32, FP64	16 TOPs
ISP	Programmable GSP accelerated Soft ISP	●
Memory	Blaize 1600 SoC memory	4GB LPDDR4
Power	Typical / Max	7W / 15W
Storage	Quad SPI NOR Flash	1MB
Communications Interface	PCIe Gen 3.0, 4 lanes	1U short form factor Asymmetric
Temperature Range	Ambient temperature	10°C to +35°C
Thermal	Thermal solution provided	Passive
Compliance	RoHS, WEE, CE, FCC	●

blaize®

4659 Golden Foothill Parkway, Suite 206, El Dorado Hills, CA 95762 USA
www.blaize.com | info@blaize.com