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

The Blaize® Xplorer® X600M accelerator is a commercial grade device designed for AI inference at the edge via easy plug-in M.2 interface, enabling industrial PC and custom products to easily integrate AI inference. The X600M 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.

With low power, low latency, and more efficient use of memory, the X600M 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 X600M, is a software defined AI inference accelerator, making it easy to update and maintain after deployment. The X600M GSP architecture is designed to run efficiently in a streaming fashion, and it is fully programmable via the Blaize® Picasso® SDK and Blaize 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.

Blaize® Xplorer® X600M M.2 Small Form Factor Accelerator

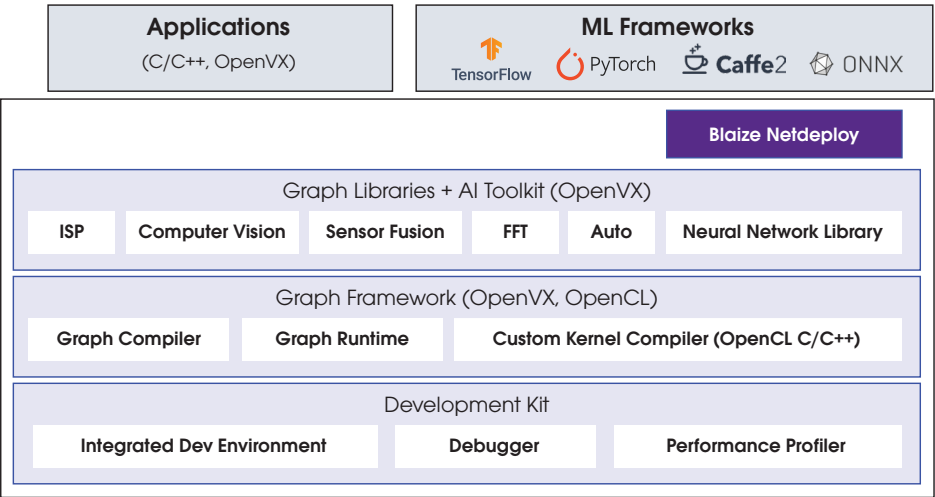### Edge and Enterprise Servers & Applications

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

Features

- 1 Blaize 1600 SoC with 16 GSP cores, providing 6 TOPs
- Low power – 3W average, 8W peak
- Commercial grade
- Soft ISP available to run on Blaize 1600 SoC
- 2 GB LPDDR4
- PCIe Gen 3.0, 4 lanes
- PCI-SIG M.2
- H.265/H.264 Video Encode / Decode up to 4K@30FPS

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	X600M
GSP	1 Industrial Blaize GSP 1600 SoCs, Data formats INT8, INT16, BF16, FP16, FP32, FP64	6 TOPs
ISP	Programmable GSP accelerated Soft ISP	●
Memory	Blaize 1600 SoC memory	2GB LPDDR4
Power	Typical / Max	3W / 8W
Storage	Quad SPI NOR Flash	1MB
Communications Interface	PCIe	Gen 3.0, 4 lanes
Temperature Range	Ambient temperature	0°C to +70°C
Thermal	Thermal solution provided	Passive
Compliance	RoHS, WEE, CE, FCC	●

blaize®