BLAIZE: AI FOR THE EDGE

INTRODUCTION

While NVIDIA dominates the market for AI-specific silicon accelerating the training of neural networks, many AI startups are developing silicon to accelerate inference processing, both for data center and edge applications. CPUs have typically been the choice for inference processing, but this is changing rapidly as the size of neural networks grows exponentially and applications are emerging that require multiple neural networks to solve complex problems. This far surpasses a CPU's processing power. One of the critical challenges for inference processors is the selection of the right balance of performance, cost and power consumption for a specific set of applications – one size will not fit all. Into this mix jumps California-based startup Blaize with the announcement of its first generation of production-ready platforms that, it contends, provide that balance for targeted edge applications.

Blaize has been at this for five-plus years under the leadership of co-founder and CEO Dinakar Munagala, developing chips and software alongside strategic investors, including Denso, Daimler and Samsung. These and other early customers have helped Blaize understand and meet the AI inference hardware and software needs for low-power, high-performance solutions. This paper will explore Blaize products, strategy and early customer engagements, as well as the challenges the company may face.

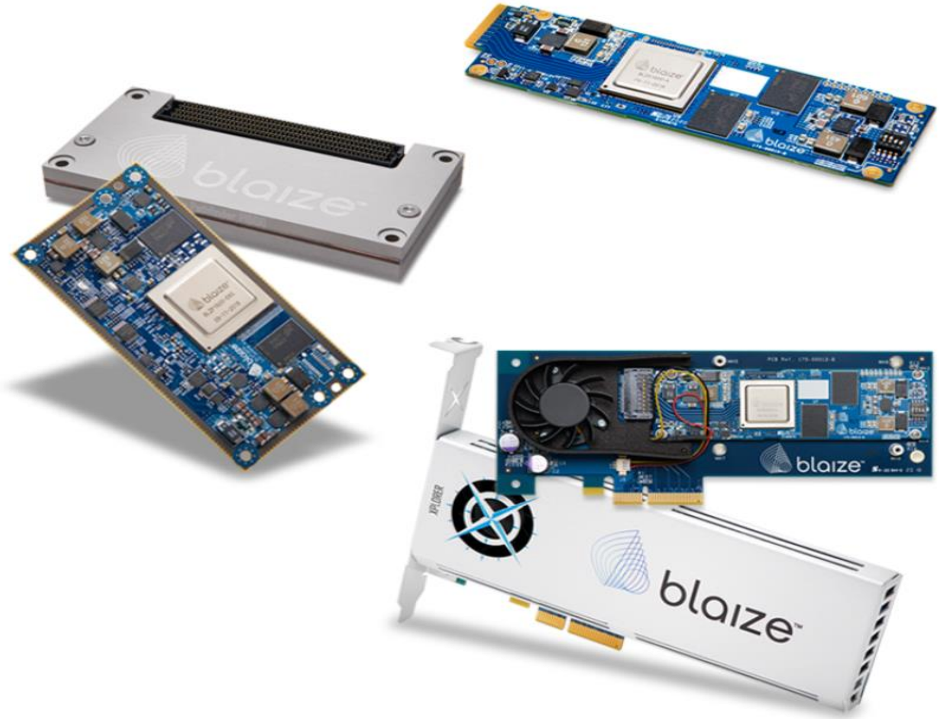
BLAIZE TARGET MARKETS AND PRODUCTS

We anticipate that the market for inference processing will expand rapidly, with revenues likely exceeding that of AI training by 2025. Some applications will require high-performance, multi-function platforms. In contrast, some will demand minimal power and cost for mass edge deployment, while others will fall somewhere in the middle of that spectrum.

Blaize is targeting the middle ground that demands significant performance at modest power and general programmability to lower costs. The company states that its first chip can deliver 60x higher system efficiency at 10x lower latency, consuming 50x less memory bandwidth and delivering 16 trillion operations per second (8-bit integer TOPS) at 7 watts, compared to GPUs typically considered for the same class of tasks. The foundational technology is a graph streaming processor (GSP) that offers full programmability, not just multiply-accumulate cores needed for matrix operations. Initial target markets include image-centric applications in industrial/manufacturing, smart city, sensor fusion, retail monitoring/analysis and last-mile delivery segments.

If 16 TOPS seems to be somewhat modest, consider that higher-performing inference chips such as the Tenstorrent Grayskull or NVIDIA Orin (expected in late 2021 and 2022, respectively) consume far more power and are targeting applications demanding far more performance. In an AI processor for an autonomous vehicle, for example, 45 watts, or perhaps even 800 watts for a robotaxi, may be feasible. However, in an application that requires an AI processor attached to a lamp-post video camera or factory floor robotic monitors, 45 watts could be a showstopper. Blaize believes it has landed on a balance of performance, power and cost for the specific use cases it is targeting.

Figure 1: Blaize's Packaging Options for its GSP Chip for Standalone and Host-connected Applications



Source: Blaize

Blaize embraces a customer-centric go-to-market approach, co-developing applications with select customers and investors and refining its hardware and software designs accordingly. To accelerate time-to-adoption, Blaize is offering its technology as a family of ready-to-deploy cards and workstations instead of merchant silicon. There are three cards in final testing, which should be generally available later this year.

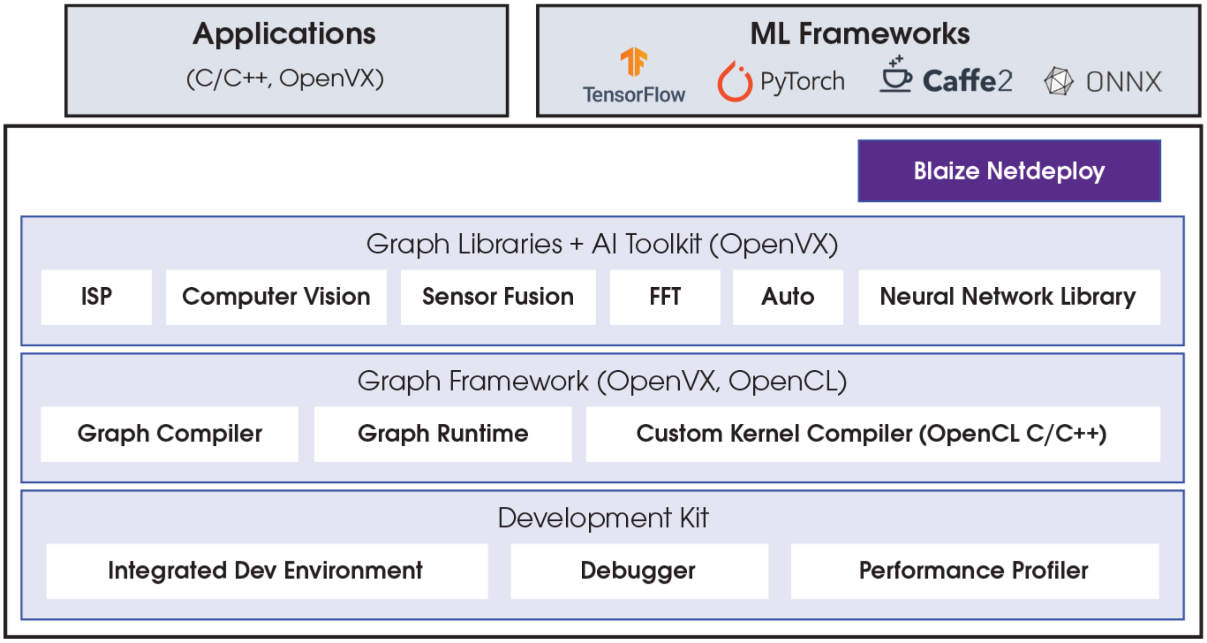
THE PATHFINDER P1600 EMBEDDED SYSTEM ON MODULE (SOM)

The Pathfinder P1600embedded system-on-module, priced at $399 in volume, is a complete credit card-size system that includes an Arm CPU, Blaize GSP and I/O. The ARM CPU programmability enables functions such as communication, planning and action initiation based on intelligent analysis of sensor input(s).

THE XPLORER X1600 ACCELERATORS

Blaize is offering two form factors for host-attached accelerators: the X1600E and X1600P. For low-power embedded markets, the Blaize 16 TOPS X1600E's EDSFF card, priced at $299, plugs into a PC or server appliance for environments such as convenience stores or factory floors. The X1600P is a half-height/half-length PCIe card that supports applications in cloud edge servers with up to 4 GSPs at 64 TOPS. The GSP chip consumes 7 watts, while the TDP for the 1-GSP card runs about 25 watts.

FIGURE 2: BLAIZE PICASSO SOFTWARE STACK



Source: Blaize

The Blaize Software Suite: Picasso and AI Studio

Software is critical for any AI endeavor. For a startup, the software stack is essential to enable customers to execute current models and develop new applications. The Picasso Software Development Kit (SDK), announced last fall, includes support for open AI frameworks such as PyTorch and TensorFlow, as well as the compilers, libraries and runtime routines needed to generate and execute code efficiently on the GSP cores. Notably, Blaize software innovations include the AI Studio and Netdeploy.

Netdeploy automates optimization and compression of the graph that represents the neural network model, reducing work from weeks to hours according to the company. Optimization includes handling model sparsity as well as quantization, selecting the minimum precision format(s) needed for the different model layers. Pruning the network compresses the model, requiring less computation. To further accelerate application development, Picasso also includes a library of pre-trained and optimized neural networks for many frequently used models.

The new AI Studio provides a visual model development toolset, allowing domain experts, who are not typically coders, to create new models. To enable “code-free” development, AI Studio poses a series of multiple-choice questions to the subject matter expert, building a deep neural network model which is then trained and optimized for execution (inference) on the Blaize GSP platform.

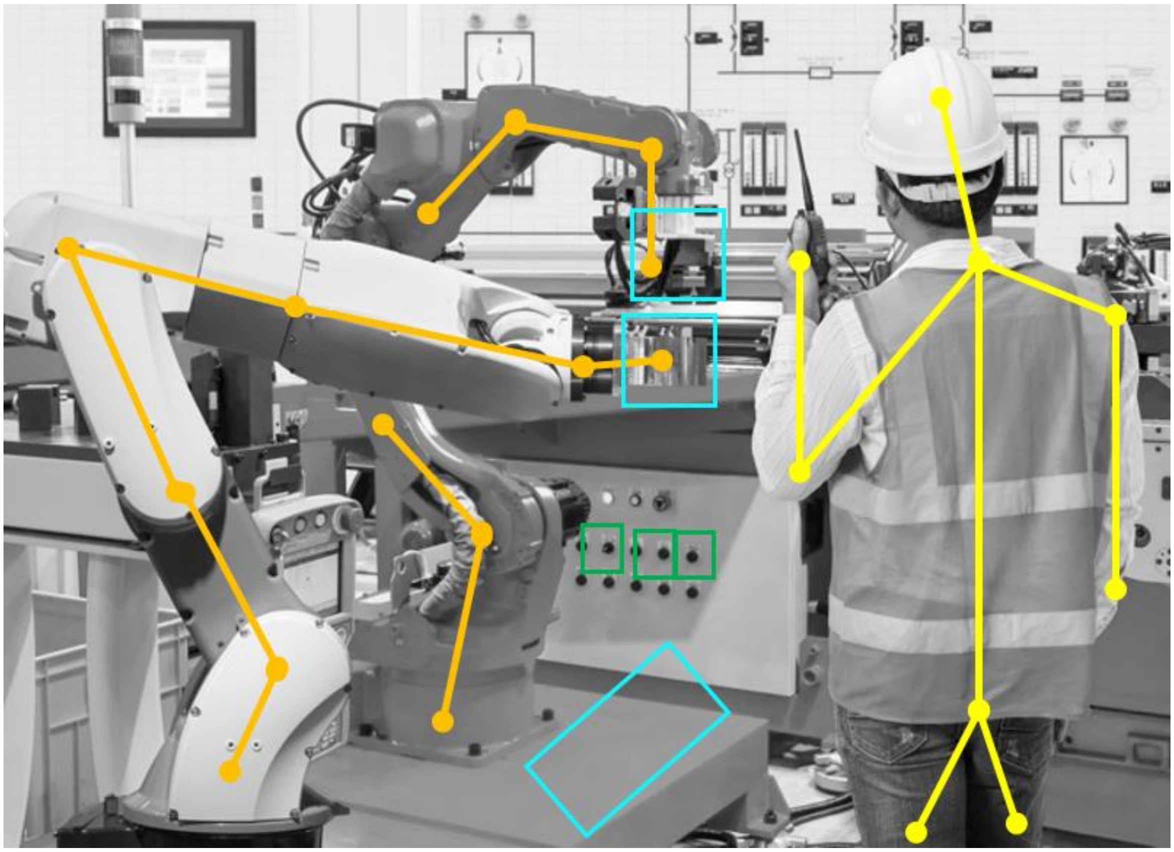
Early Blaize Customer Projects

Blaize stands out in the crowded AI silicon startup field because it has been nurturing client opportunities since it began to raise venture capital in 2017. We believe this effort could result in better alignment of the company’s offerings and the needs of its target markets as well as accelerate its revenue ramp. Based on the five use cases Blaize has shared thus far, image processing is central to its initial market thrust, especially in high-frame-rate and multi-stream applications. The company has shown demonstrations that tell the stories quite well, but here is a synopsis of a few of the workloads now in pilot testing. Blaize management expects that some of these could be in production by the end of 2020.

Industrial Monitoring

On the factory floor, the Blaize GSP is running five independent neural networks at 50FPS with less than 100ms aggregate latency to monitor human and robot pose position, measure production volume and monitor product quality. Notice the green bounding boxes in Figure 3, where the system can follow switch settings of the robot to verify proper operation.

FIGURE 3: INDUSTRIAL MONITORING

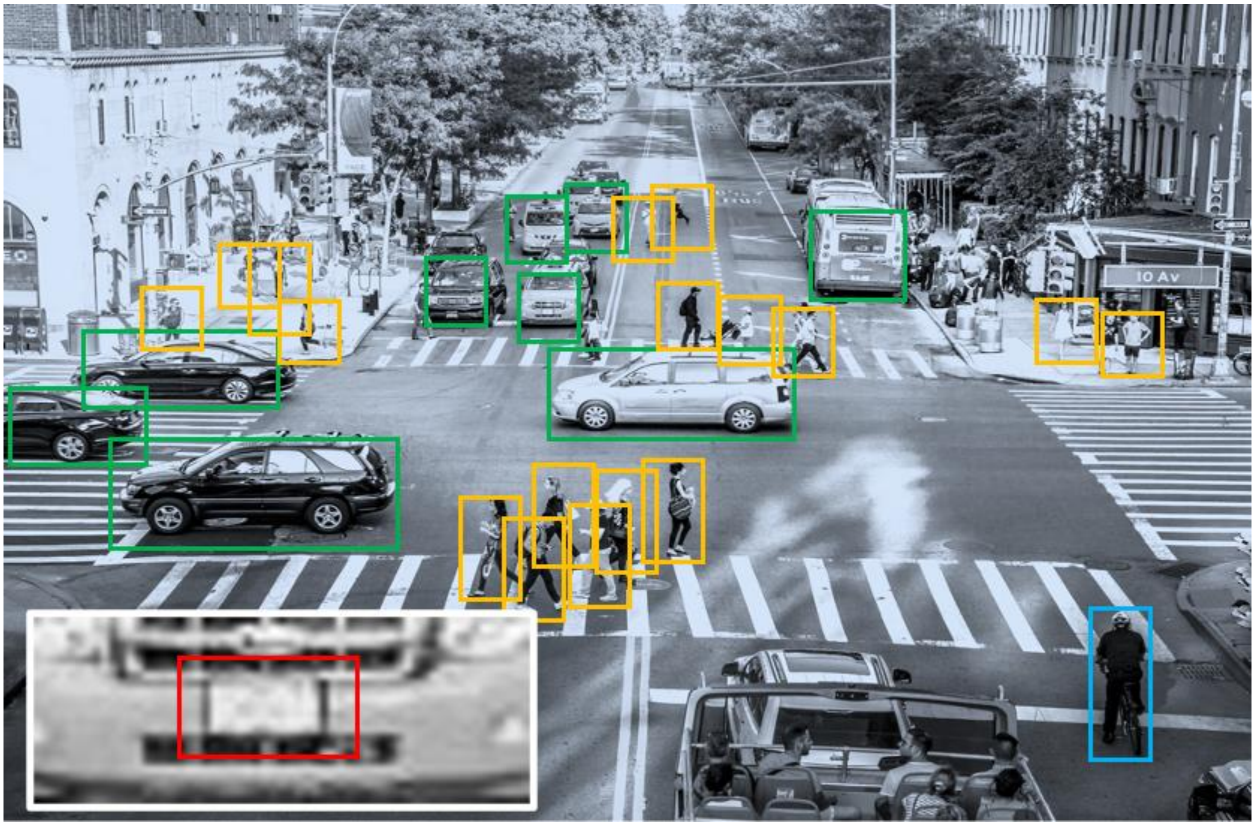


Source: Blaize

SMART CITY

In this pilot project, the Blaize GSP is running three independent neural networks at 50FPS, monitoring human detection, pose, position, auto-detection and traffic intersection safety and security. The system not only detects problems but can take action to reduce risk to life and property. For example, the system could identify a person who falls or walks into the path of traffic. The intelligent system could determine whether events are noteworthy and take appropriate action, such as turning all traffic lights to red or calling emergency responders for aid. And of course, the detection of a license plate belonging to a stolen vehicle could initiate a message to authorities. Safely employing such automated power would require adequate security to prevent malicious hacking.

FIGURE 4: SMART CITY APPLICATIONS FOR TRAFFIC FLOW AND PUBLIC SAFETY



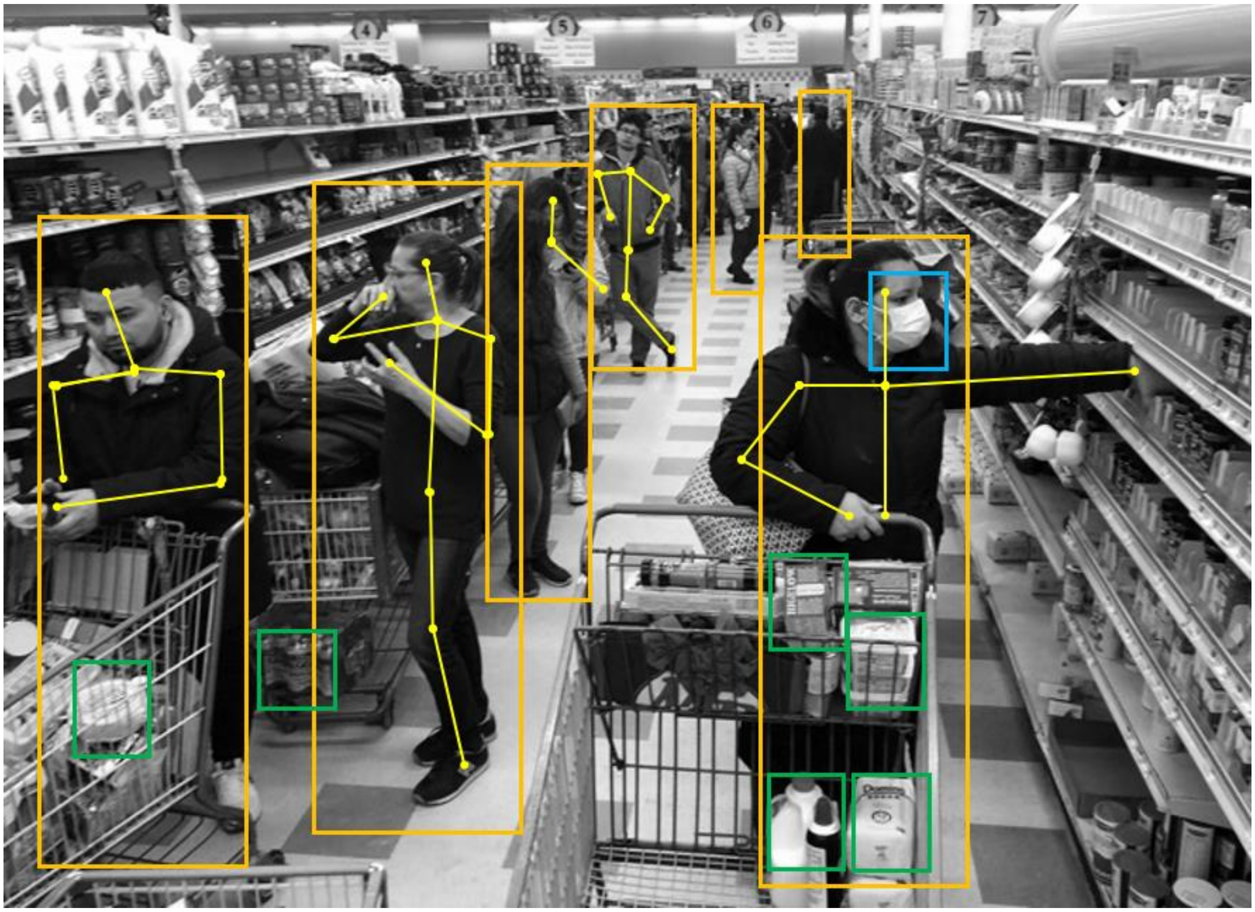
Source: Blaize

RETAIL

In retail applications, analysis of shoppers' activities can reduce theft, optimize stocking management and product placement and improve flow and safety. In today's reality of the coronavirus pandemic, retail establishments also need to ensure compliance with regulations regarding face coverings. In a current customer retail pilot, Blaize GSPs are processing four independent neural networks at 50FPS to improve operations, lower costs, increase safety and improve revenue.

Blaize envisions servers with multiple 4-GSP PCIe cards processing an array of streams from various sensors deployed across a facility, neighborhood or city. Since a single GSP can handle up to five streams of high-resolution, high-frame-rate input, this approach could help reduce the costs to stores, cities or factories seeking to improve operations.

FIGURE 5: RETAIL APPLICATIONS



Blaize sees significant opportunity for AI in retail environments, and the GSP can handle multiple DNN models simultaneously to improve security, safety and operations.

Source: Blaize

CONCLUSIONS AND RECOMMENDATIONS

Acceleration of AI inference at the edge is in its infancy. Blaize believes co-development engagements with clients has helped them create a “no-compromise” solution, delivering high-performance AI processing at reasonable power and costs, all underpinned by a rich AI software stack. We are aware of other startups that may soon claim better performance – although consuming more energy – targeting high-end edge applications such as natural language processing and conversational interaction. Similarly, other processors may consume far less power, even under a watt. Still, these are unlikely to match Blaize’s performance. And such low-end, fixed-function chips do not provide general programmability, which allows a more flexible design to accelerate the entire application and lower costs.

The outcomes of the Blaize pilots will provide useful milestones by which one can measure the company's success. These pilots will be far more instructive than typical benchmarking and performance claims as they will demonstrate whether Blaize can solve the real-world problems its clients are trying to address. The success of the pilots will depend not just on the company's silicon, but on the capabilities and ease of use of the Picasso and AI Studio stack that Blaize has created. Pilots are demanding and always seem to take longer than anyone hopes or expects, but these early projects will form the litmus tests that Blaize must pass.

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTOR

Karl Freund, Senior Analyst at Moor Insights & Strategy

PUBLISHER

Patrick Moorhead, Founder, President, & Principal Analyst at Moor Insights & Strategy

INQUIRIES

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

LICENSING

DISCLOSURES

This paper was commissioned by Blaize, Inc. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.