**Publication date:**
11 Nov 2020
**Author:**
Aditya Kaul
Director, AI & Intelligent Automation

# On the Radar: Blaize

Brought to you by Informa Tech

# Table of Contents:

# Summary

## Catalyst

The edge is where the action is in AI silicon. The emergence of AI at the edge, whether on-device, on an edge server, or on a specialized edge appliance, has seen both silicon incumbents and a range of startup companies positioning themselves in the market with new products. Blaize (formerly Thinci) is one such AI silicon startup targeting the edge, and has in the last year come out of stealth mode. Blaize provides a blank slate approach to AI chip architecture, and a full-fledged software suite that targets AI edge developers, and has currently got customers trialing its chipset across a range of application markets and form factors that includes modules and PCIe cards.

## Omdia View

Omdia believes that the AI edge silicon market is ripe for heterogeneous solutions and innovative approaches that balance performance and power efficiency. AI edge has multiple sub-segments depending on the power budget. There are 3 large segments which can be addressed under AI Edge: low power (<5W), medium power (5-50W), and high power (50W>). The low power segment is largely smartphone-driven and is off limits for most AI silicon vendors, with most smartphone vendors like Apple, Samsung, and Huawei building their own chipsets. This leaves the medium and high-power markets, which include devices ranging from robots, drones, security cameras, PCs, and laptops, to edge servers and edge appliances. Overall, Omdia forecasts the AI edge chipset market to reach $51.9 billion dollars in revenue by 2025, of which $23.4 billion (45% of the market) is for applications other than mobile phones, where the majority of the third-party chipset opportunity lies.

There are a range of requirements, power budgets, form factors, and use cases that apply across the AI edge, which is ripe for innovative solutions that provide best-in-class products addressing specific needs. Unlike the AI cloud datacenter market where incumbent silicon vendors like NVIDIA and Intel dominate, the AI edge market has lower barriers to entry. While both Intel and NVIDIA have solutions targeting the AI edge, the heterogeneous nature of chipset solutions required along with novel applications for AI edge provides ample opportunity for new startups to innovate and disrupt.

## Why put Blaize on your radar?

Blaize comes with a blank-slate approach to AI edge silicon specifically targeted at processing AI models with an inherent graph-based architecture, such as neural networks, delivering high performance at a low power budget. Rather than using general purpose hardware like CPUs or GPUs which have issues with latency and power, Blaize technology is purpose-built for the AI edge. While Blaize is not unique with a blank-slate approach to AI silicon, as other startups are building purpose-built AI chipsets, Blaize has had more than 5 years working on the hardware technology and now has products shipping alongside a fully-functional software suite, which includes a code-free development platform and an optimization tool that can help accelerate AI edge deployments. Nevertheless, Blaize does have a lot to prove in the market including the robustness and performance of its technology as it puts its product in customers hands and as the applications become production ready.

# Market Context

## Definition of AI Edge

As per Omdia, the “AI edge” consists of devices that provide AI algorithm computing with latency of less than 20ms. This means that the processing could be done on the device (on-device), on-premises (near-edge), or on cloud elements located within the 20ms range (edge cloud). The terminology is somewhat loose; broadly, all the devices fall into these three categories and Omdia has used latency as a guiding factor rather than location. On-device computing is the most prominent category, followed by near-edge or on-premises computing. The edge cloud, which is a mature computing category in areas like telecoms, has yet to see a large-scale deployment of AI-enabled edge clouds and therefore is considered a niche category for AI edge.

## Market Landscape

The vast majority of chipset vendors for AI edge solutions are focused on the on-device and near-edge categories, although on-device is largely dominated by DIY players like Google, Huawei, Samsung, and Apple who develop their own chipsets with AI accelerators embedded within their smartphone processors. There is a wider market of on-device AI opportunities for third-party chipset vendors that cover robots, security cameras, drones, automotive, AR/VR, smart speaker devices, PCs, and laptops. Within the on-device market, the medium power (5W-50W) category is where most of the third-party chipset opportunity is expected to be. Chipset vendors see higher profit margins in the medium power category compared to lower power devices, allowing for AI accelerators to be added onto the BOM (Bill of Materials) cost. Also, one is less likely to see device OEMs in this category build their own chipsets as we see in smartphones.

### Incumbent Players

Established incumbents like NVIDIA, Intel, Qualcomm, and Xilinx (acquired by AMD) continue to build their capabilities in the AI edge chipset market. NVIDIA is the largest player in the AI edge market, and has built a reputation as the ‘go to’ silicon vendor for AI, with their GPU platform being repurposed for running AI and specifically deep learning algorithms. NVIDIA’s Jetson Xavier competes in 5W-50W category, and combined with their popular Jetpack software suite, it continues to drive adoption across smart city, surveillance, industrial, and automotive markets. Intel’s Myriad 2 Vision Processing Unit is its primary offering in AI edge, and has found success in the drone and smart camera markets, while Intel’s Mobileye is being widely used in automotive. Xilinx provides the Zynq SoC platform (with FPGA cores) along with their Vitis AI software suite. Xilinx has focused on the automotive market for AI, but is seeing increasing adoption for edge video analytics within smart cameras and video appliances. Qualcomm’s Snapdragon is the most widely used chipset for mobile, but Qualcomm has also tied together its AI and ML offerings with the Snapdragon Neural Processing Engine SDK. This SDK allows for heterogeneous execution of deep learning workloads across Snapdragon, CPUs, GPUs, or DSPs. Qualcomm has also been expanding its Snapdragon into the robotics space targeted at AI vision applications.

## Startups

There are several AI edge chipset startups that have emerged in the last few years including Gyrfalcon, Mythic, and Syntiant amongst many others. Omdia has estimated that there are more than 50 companies that are competing in the AI edge silicon market, although very few have shipped products to customers. Blaize is part of a burgeoning AI edge startup ecosystem, most of who are creating discrete chipsets or ASICs specifically designed to accelerate AI workloads at the edge.

# Product Overview

## Graph Streaming Processor

The Blaize Graph Streaming Processor (GSP) is a graph-native, fully programmable processor with 16 cores, providing 16 TOPS of AI inference performance (8-bit INT) in a 7W power envelope. The Blaize GSP claims to use 50x less memory bandwidth, 10x lower latency, and 60x better efficiency than competing GPU and CPU solutions. The GSP is built for task-level parallelism, where calculations can be performed dynamically using a hardware scheduler across multiple graph nodes and threads as data becomes available. The GSP architecture also saves on DRAM memory bandwidth, as data flows dynamically through nodes with data processed in parallel rather than sequentially. The lack of sequential processing reduces the need for data to be moved off the chip into an external DRAM bringing power and performance advantages. The task parallelism for the GSP is offered in addition to instruction-level, thread-level, and data-level parallelism, making it uniquely capable of streaming neural network graphs while achieving a balance between high performance and lower power budget, ideal for AI edge applications.

## Modules and Cards

Blaize offers the GSP processor across a range of module, cards, and development kit form factors. These include:

- **Pathfinder P1600 Embedded System on Module** - A credit card-sized module with 1 GSP chip providing 16 TOPS at 7W. The module has dual ARM A53 CPU cores, a video encode/decode codec, and Ethernet, CSI, and USB I/O interfaces for commercial and industrial applications such as in-camera, sensors, or network edge equipment. This is priced at $399 in volume.
- **Xplorer X1600 EDSFF PCIe3 Small Form Factor Accelerator Platform** - A small form factor accelerator with PC/server board mount options, using 1 GSP chip providing 16 TOPs at 7W. 2 GB or 4 GB, dual-channel, 32-bit wide LPDDR4. A total of 18 of these cards can be combined into a 280 TOPS 1U edge server appliance, ideal for retail and factory shopfloors where PC/servers are available. This is priced at $299 in volume.
- **Xplorer X1600P PCIe Accelerator Platform** - A PCIe3, half-height, half-length, accelerator for commercial and enterprise applications, with 1 GSP chip providing 16 TOPs at 7W. Up to 4 GSP chips on one card provide 16-64 TOPs at <30W, or 8 of these cards can be combined into a 500 TOPS edge server appliance, with 8GB, 16GB or 32GB, dual-channel, 32-bit wide LPDDR4.
- **Pathfinder EST-1600 Embedded Kit** – A development kit with the Pathfinder P1600 Embedded System on Module, preloaded with Blaize Picasso Software Development Kit, code samples, and multiple external peripheral options including camera, display, USB, SD card, and Ethernet.

## Software Suite

The hardware is complemented with the Blaize AI Software Suite, arguably one of the most extensive software platforms being offered for AI edge application development. The software suite includes Blaize's Picasso SDK for developers, with libraries to build complete AI edge applications.

Apart from providing a library of pre-trained models, the software suite also includes AI Studio which is a code—free visual development tool meant for domain experts that are not necessarily technically versed in AI/ML coding. Using a survey mechanism, the tool selects and builds a neural network model for the appropriate application at hand, which is then trained and optimized for running on the appropriate hardware.

Netdeploy, a key component of both software offerings, automates the optimization and compression of graphs (i.e. neural networks). This includes optimization of model sparsity, quantization, precision across model layers, and pruning of models to have them run efficiently on edge devices.

# Company Information

## Background

Blaize (formerly Thinci) was founded in 2010 by ex-Intel employees Dinakar Munagala, Satyaki Koneru, Ke Yin, & Val G. Cook. At the end of 2020, Blaize had raised $87 million in funding with investors including Denso, Temasek, Daimler, and Samsung Catalyst amongst others. Blaize is headquartered out of El Dorado Hills, CA with offices spread out across the US, India and the UK.

Blaize was founded in 2010 and work on the GSP began before the AI boom seen post-2017, with the initial impetus being to build hardware that removes constraints of legacy GPU and CPU architectures. As a result, the GSP, like the CPU or GPU, can take on a whole host of applications that go beyond AI and include non-AI workloads such as image or speech signal processing, allowing for the hardware to process a full application end to end, rather than rely on other processors. The general applicability of the GSP across applications including both AI and non-AI tasks gives Blaize a unique position in the market and puts them head to head with Intel and NVIDIA. With their wide applicability across a full application workflow, Blaize is targeting vision applications at the edge with a specific focus on industrial monitoring, smart city, retail, automotive, and mobility.

## Current Position

Blaize came out of stealth mode in Nov 2019 and through the course of 2020 has been providing its GSP chipset to select customers. In Aug 2020, Blaize announced details of the software suite and the various modules that are now available for sampling, along with details of specific customer trials and applications. As of 4Q 2020, the company continues to ship products to customers mainly for trial purposes, with the expectation that it will start to ship these modules in production sometime before the end of 2020.

The customer examples and demos provided have been in the areas of retail security, smart city surveillance, industrial monitoring, and last-mile autonomous vehicle delivery. The focus has been on AI vision use cases that fall under object detection, pose and position detection, face recognition, ID or license plate reading, and anomaly detection. Sensor fusion has been another workload that Blaize has been showcasing across its customer examples, where feeds from camera, Lidar, and  radar, along with non-AI workloads like Image Signal Processing (ISP), can all be run as a singular graph-native application end to end using its Picasso SDK. This is a powerful value proposition for customers that usually need to segment the end to end workflow across CPUs, GPUs, and ASICs. The end to end workflow capability has only been showcased by some incumbents like NVIDIA, and most startups have only targeted the specific AI workloads. If Blaize does prove its capabilities from an end to end perspective, it has the potential to disrupt the AI edge market displacing incumbents like NVIDIA and Intel, as well as the niche AI edge startups that may not be able to compete with a holistic AI edge solution.

For go-to-market, most new players like Blaize are looking at partnering with system integrators and channel partners across the various verticals like retail, automotive, and industrial, rather than going straight to device OEMs (including server and appliance OEMs).  This feels like the sensible strategy, considering that startups don’t have established relationships with OEMs, who are generally locked into specific vendors, and there is a long gestation period for OEMs to bring new chipset solutions into their fold. Blaize hopes that its chipset will integrate into existing boards, adding AI acceleration, and therefore it is targeting SIs rather than OEMs.

## Future Plans

Blaize is looking at expanding its proposition to go beyond supporting vision applications, to also showcase its capabilities in audio and voice/speech recognition applications.

## Key facts

**Table 1: Data sheet: Blaize**

| | | | |
|---|---|---|---|
| **Product name** | Graph Streaming Processor (GSP) | **Product classification** | AI embedded and accelerator chipset |
| **Version number** | NA | **Release date** | Nov 2019 |
| **Industries covered** | Retail, Automotive, Industrial, Smart City | **Geographies covered** | All |
| **Relevant company sizes** | Large, small, and medium | **Purchase options** | Modules, developer kit, cards |
| **URL** | www.blaize.com | **Routes to market** | Direct and channel |
| **Company headquarters** | El Dorado Hills, CA, USA | **Number of employees** | 300+ |

Source: Omdia

# Analyst Comment

The AI edge market is highly competitive, with this trend expected to accelerate in 2021. Incumbents like NVIDIA have already made their intention clear that, while they have a strong foothold in the AI cloud and datacenter market, they are going to continue to push the boundaries for GPUs at the edge. The NVIDIA-ARM acquisition is another data  point that  argues for  a stronger move for NVIDIA towards the edge with ARM's leadership position in embedded and mobile, merging their enormous developer communities under one umbrella that dominates AI compute from cloud through to edge. The AMD-Xilinx acquisition can also be seen as a positive for AMD in this space, with Xilinx's AI edge proposition having been strengthened of late, giving AMD reach into edge markets.

Despite the threat from NVIDIA and AMD and their impending acquisitions, the AI edge market is heterogeneous, and has room for new architectures with varying value propositions around power, accuracy, performance, speed to market, and reconfigurability. The key to winning in this market is providing developers and customers an ability to build applications with the least amount of friction, while providing robust, high-quality solutions. Although Blaize is a startup technically it has been in stealth mode for close to 9 years, which means that there has been a long run-up towards perfecting the technology and the IP that is at the heart of the Blaize GSP. Therefore, Blaize is not your 'run of the mill' AI hardware startup.

Blaize brings several key elements to the table which give it an edge.

- This includes a unique, fully programmable chipset architecture built from the ground up to process graph-oriented neural networks.
- The ability to process end to end workloads, from AI through to non-AI.
- A compelling full suite of software for AI edge applications that doesn't require technical knowledge of AI.
- A diverse range of modules and cards that can be used for plugging into existing equipment.

More than the technology and the performance of its products, Blaize needs to focus on building an ecosystem of developers and SI partners that it can bring into their fold. Being a new player in the space, it has only limited relationships with OEMs or SIs and is building a developer ecosystem from scratch. There are new opportunities emerging in the 'grey area' between the edge cloud and edge appliance, where Internet players like AWS (Wavelength) and Microsoft (Azure Edge Zones) have started to build an ecosystem of partners. Of course, almost every other AI silicon startup is in the partnership building game and trying to compete with incumbent marketing budgets is going to be hard for startups like Blaize. However, with its product portfolio largely nailed down, its priority should be to use future funding rounds to secure its ecosystem, developer, and partner strategy.

Also, we have yet to see customers deploying the Blaize solution beyond proof-of-concepts. There is a lot to prove for Blaize, but the fundamentals look strong, and it's now up to the market to decide whether the Blaize proposition is worth investing over incumbents like NVIDIA or Intel. We are still in the early days of the AI edge market with no clear winner or preferred solution, and there is a general appetite for experimenting with new solutions.

# Appendix

## On the Radar

On the Radar is a series of research notes about vendors bringing innovative ideas, products, or business models to their markets. On the Radar vendors bear watching for their potential impact on markets as their approach, recent developments or strategy could prove disruptive and of interest to tech buyers and users.

## Further reading

Artificial Intelligence Edge Devices - 2020 (May 2020)

Deep Learning Chipsets - 2020 (July 2020)

## Author

Aditya Kaul, Research Director, AI & Intelligent Automation

askananalyst@omdia.com

## Citation policy

Request external citation and usage of Omdia research and data via citations@omdia.com.

## Omdia consulting

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Omdia's consulting team may be able to help you. For more information about Omdia's consulting capabilities, please contact us directly at consulting@omdia.com.

## Copyright notice and disclaimer

The Omdia research, data and information referenced herein (the "Omdia Materials") are the copyrighted property of Informa Tech and its subsidiaries or affiliates (together "Informa Tech") and represent data, research, opinions or viewpoints published by Informa Tech, and are not representations of fact.

The Omdia Materials reflect information and opinions from the original publication date and not from the date of this document. The information and opinions expressed in the Omdia Materials are subject to change without notice and Informa Tech does not have any duty or responsibility to update the Omdia Materials or this publication as a result.

Omdia Materials are delivered on an "as-is" and "as-available" basis. No representation or warranty, express or implied, is made as to the fairness, accuracy, completeness or correctness of the information, opinions and conclusions contained in Omdia Materials.

To the maximum extent permitted by law, Informa Tech and its affiliates, officers, directors, employees and agents, disclaim any liability (including, without limitation, any liability arising from fault or negligence) as to the accuracy or completeness or use of the Omdia Materials. Informa Tech will not, under any circumstance whatsoever, be liable for any trading, investment, commercial or other decisions based on or made in reliance of the Omdia Materials.

## CONTACT US

omdia.com

askananalyst@omdia.com