Coverage Initiation: Blaize samples its GSP graph processor, eyes low-cost AI inference at the edge

Analysts - John Abbott

Publication date: Thursday, October 15 2020

Introduction

AI chip startup Blaize, which came out of stealth at the end of 2019, has begun sampling the first embedded and accelerator platforms built around its Graph Streaming Processor (GSP). It has also launched its first software development tools, including the Picasso software development kit and the code-free AI Studio. Blaize is specifically targeting artificial intelligence at the edge with its dataflow processor. It aims to provide the performance, low latency and energy efficiency that alternative platforms, such as CPUs, GPUs and FPGAs, may struggle to match due to their original design points. To support AI, edge applications need to offer full programmability and enough compute power to handle the workloads, but at low cost and in a form factor that's easy to productize.

The 451 Take

There are so many AI chip startups that it has become hard to differentiate one from the other. However, Blaize has a solid story, especially in the techniques it uses to minimize unnecessary data movement without having to include massive amounts of expensive on-chip memory. AI inference is a larger opportunity than the more established area of AI training, and the market is currently wide open, with few incumbents. Blaize has a solid software offering as well that utilizes standards where it can, but it does need to develop a healthy partner ecosystem.

Context

Blaize, originally known as ThinCI, was founded in 2010, with its headquarters in El Dorado Hills, California and a large development team in Hyderabad, India. It also picked up some silicon expertise in the UK by hiring teams from MIPS Technologies and Sega. The founding team (CEO Dinakar Munagala, CTO Satyaki Koneru and chief scientist Ke Yin) all previously worked at Intel, as did chief

software architect Val Cook. Funding has reached $87m, including a series C round of $65m in September 2018, the last publicly disclosed. Strategic investors include Daimler and automotive component makers Denso and Magma, as well as the Toyota-backed Mirai Creation Fund. Venture funds include GGV Capital, Samsung Catalyst Fund, SGInnovate, Temasek and Wavemaker Partners. The company began early access customer engagements in 2018 in the automotive, smart vision and enterprise computing segments.

Products

Products fall into three categories: the GSP chip itself, boards and cards that use it, and software programming tools. Codenamed 'El Cano,' the 16-core Blaize graph streaming processor has been designed from scratch, with claimed peak performance of 16 trillion operations per second (TOPS) of 8-bit integer operations. It's being manufactured in 14nm process by Samsung and typically operates within a 7W power envelope; 7nm is on the roadmap. Blaize calls El Cano a second-generation part because it has had a 28nm test chip available for prototyping since 2017. The graph-oriented design avoids the use of off-chip memory through streaming in order to minimalize data movement that would otherwise limit performance, add to latency and increase power consumption. A key element is the on-chip hardware scheduler, which looks ahead in the graph and schedules computation as soon as enough data is available. This avoids the need to process an entire image or layer at the same time. Multiple neural networks, or multiple nodes from multiple layers, can be run either sequentially or in parallel.

Boards based on the GSP range in price from $300 to $1,000. The Pathfinder P1600 embedded system on module (SOM) is a PCIe plug-in card for edge AI applications such as cameras, machines with sensors and network edge equipment. It's available in commercial or industrial grades. No host processor is required because the stand-alone SOM includes its own dual ARM A53 CPUs to run the operating system and application software – as well as the ARM Crypto Secure Engine, video encoders/decoders, and embedded interfaces to camera, display, Ethernet, PCIe and USB. In contrast, the Xplorer Accelerator Platform X1600E PCIe 3.0 plug-in card works as an accelerator inside a host server or appliance (perhaps an industrial PC or a rack of cards in a 1U server). It comes as a 16-core EDSFF card for board and server mounting, or as a four-GSP (64-core) half-height/half-length accelerator for commercial and enterprise use.

Sophisticated software is key to all of this. There's a code-free visual development tool called AI Studio, but it's the Picasso software development kit that provides a more traditional tool for use with C/C++ and all the standard frameworks. It includes an AI toolkit (for computer vision, deep learning, linear algebra, etc.), pre-built models, and an automated, edge-aware deployment tool called NetDeploy, which converts and optimizes applications into graph-native format. The Picasso Graph Framework (which supports standards such as OpenVX and OpenCL) then compiles the graphs for execution on the GSP.

Strategy

Blaize says it's addressing three gaps in the current story of AI processing: that AI-enabled edge computing is unique enough to require a new, graph-oriented architecture rather than existing architectures; that new levels of efficiency are a requirement that currently can't be met (at least while providing enough compute power); and that widespread deployment of new AI technologies is currently a big barrier to adoption. Although Blaize is a chip company and will sell at the chip level in volume (i.e., to automotive customers), it says that most customers just want to plug something in and go, so cards and boards will be its primary route to market at this stage. The company will work with industrial OEMs on integration projects. Smart retail, smart city, smart manufacturing, industrial automotive, robotics, security, smart vision, edge inference and IoT are on its list of target applications. No partners are currently announced, perhaps because many of them are relatively

small and operating in the fragmented industrial space. Automotive could prove to be the route to volume shipments, and the company's strategic backers could be first in the queue.

Competition

NVIDIA is the inevitable incumbent in the AI processor market, but its mainstream GPUs are less obviously suited to edge deployments, where power efficiency and graph-native operation are potential advantages for Blaize. The Pathfinder module is somewhat comparable to NVIDIA's Xavier module, which claims to offer performance of 32 TOPS in a 10W power envelope. The Blaize GSP is similar in some senses to GraphCore's IPU, though like NVIDIA, GraphCore has focused initially on high-end training, and its use of on-chip memory makes it necessarily more expensive. Wave Computing, which owned MIPS but which has filed Chapter 11 and appears to be inactive, was one of the recent advocates of dataflow processing for AI. Mythic also has a dataflow architecture, but its longer-term vision is analog computing for AI, which may be harder to achieve. Well-funded SambaNova Systems talks about 'reconfigurable dataflow,' but so far hasn't revealed many details. Hailo's focus is on edge inference at low power though a distributed memory pipeline. If automotive turns out to be the sweet spot for Blaize, then it's likely to compete with established players in that market, primarily Intel's Mobileye and Movidius.

SWOT Analysis

Strengths	Weaknesses
Blaize has had 10 years to establish its architecture and has offered a prototype chip since 2017. That's why its software offerings look relatively mature.	Edge is undoubtably a hot market opportunity, but it's also an ill-defined term and a fragmented market, so real money-generating projects aren't as numerous yet as might be imagined.
Opportunities	**Threats**
Rapid time to deployment (plug in and go) could be a significant advantage as the market matures, and the need for efficient AI inference is now growing rapidly.	Blaize needs a few anchor chip-level customers with volume requirements. If it gets caught up in too many semi-custom integration projects with smaller industrial companies, then volume shipments and revenue might be delayed.

Source: 451 Research, LLC