Blaize Graph Streaming Processor

The Revolutionary Graph-native Architecture

The Blaize Graph Streaming Processor™ (GSP) represents a significant inflection point for the industry. Architected for graph-native AI application development, it represents a fundamental way to change how we compute the intense workloads of the future.

Neural Networks Change Computing

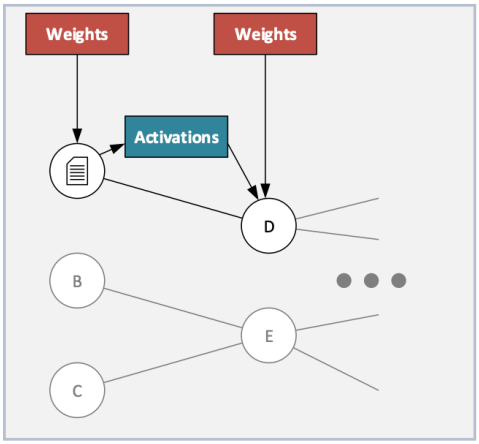
Artificial Intelligence (AI) has taken the world by storm, in many cases surpassing human-level accuracy in areas like image recognition and speech recognition. The convergence of big data, fast hardware processing and advances in AI algorithms have accelerated the adoption of AI in almost every industry. One of the core advancements driving this surge in AI is the emergence and evolution of neural networks, software models that are intended to mimic (in a simple fashion) how neurons in the brain behave.

Neural networks are extremely powerful, but getting these neural networks deployed into real-world applications has many challenges:

- Neural networks need to run efficiently close to where real-world data is collected - environments where computing, memory and energy resources are constrained.
- Neural network architectures, models and methods evolve very rapidly, and developers need to be able to update their models easily and quickly.
- Neural networks are usually part of an overall AI application, so developers need to integrate non-neural network functions along with neural network functions in order to deploy their applications in the real-world.

There are different neural network architectures to accommodate the varied and wide range of AI tasks, but every neural network has one common feature: they are all structured as graphs.

The high-level frameworks used for creating and training models are already graph-oriented, but the underlying hardware used for inference have been stuck in the past. Much of the development effort in implementing a NN model involves adapting the graph to a non-graph architecture. This means both longer development cycles and less efficient processing in the finished product.

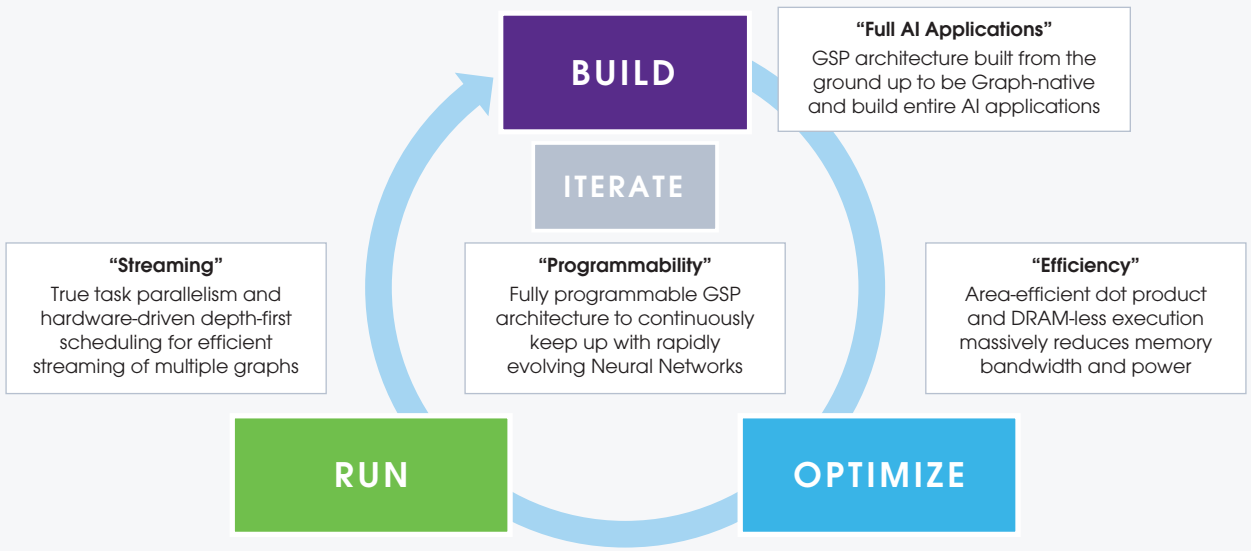


Neural networks consists of multiple layers, each of which has multiple nodes. Processing starts at one end of the network, proceeds through all of the layers, and terminates at the other end of the network. Within each node resides a small computer program, sometimes referred to as a kernel. The kernel performs the functions of that node, manipulating of weights and inputs. The outputs of one layer, also known as activations, become the inputs of the next layer. Each layer is designed to extract some feature, for example lines or shapes from images, and as the layers progress the neural network graph gets built to progressively extract higher and higher level features to ultimately identify something in the input, like objects in an image.

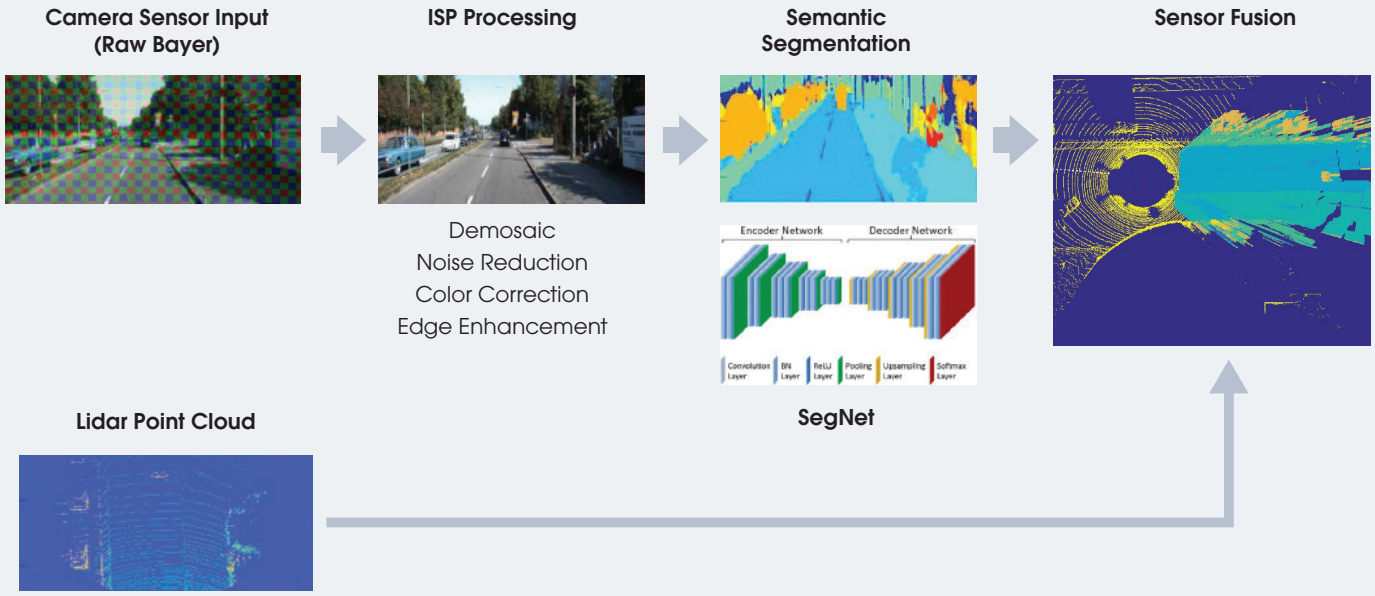
Graph Streaming Processor: Efficiency & Programmability

The Graph Streaming Processor architecture, or GSP, is the first true Graph-native architecture built to address the challenges in efficiently processing neural networks and building complete AI applications. With a fully programmable graph streaming architecture, GSP chips process graphs more efficiently than CPU/GPU architectures. As a result, the GSP architecture enables developers to build entire AI applications, optimize these for edge deployment constraints, run these efficiently in a complete streaming fashion, and continuously iterate to keep up with rapid evolutions in neural networks.

Graph Streaming Processor Architecture Enables Graph-Native AI Application Development



Complete Automotive Sensor Fusion Application



This application consists of non-neural network Image Signal Processing (ISP) functions, neural network processing for semantic segmentation and functions for sensor fusion, which can all be represented as graphs, integrated together to build the full application and run entirely on the GSP architecture efficiently.

A New Way to Compute Neural Nets

The GSP architecture opens up new possibilities for more efficient, lower-power neural-net processing. Drastically reduced data movement both lowers power and improves performance. Task-level parallelism allows multiple nodes from multiple layers to be processed concurrently. Developers can work at the graph level, making development more intuitive and helping to speed AI solutions to market more quickly with the help of a broader range of engineers. And, because it is fully software programmable, changes to trained models can be easily incorporated during development and in the field.

Corporate Headquarters
4370 Town Center Blvd
Suite 240
El Dorado Hills, CA 95762 USA
T: +1 916 347-0050
info@blaize.com

India Office
Block 2, 4th Floor,
DLF Cybercity, APHB Colony,
Gachibowli
Hyderabad-500 032

UK Office
Suite 1, Concept House
Home Park Road
Kings Langley, Herts
WD4 8UD UK

© 2019 Blaize, Inc. All Rights Reserved.

blaize™