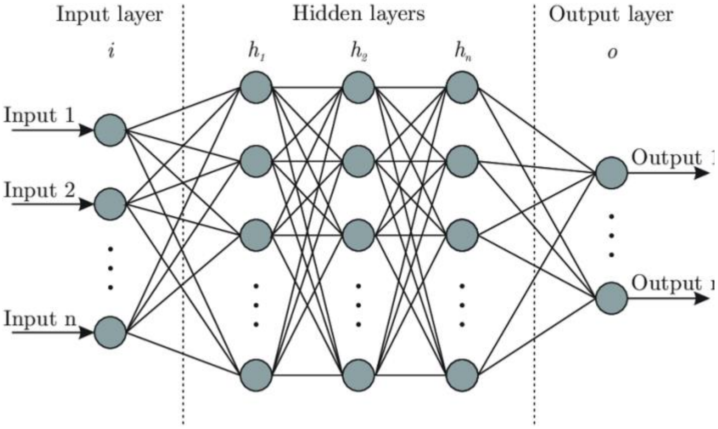
Roviero Graph Processor

Roviero

FULL STACK AI ACCELERATION



CortiOne

World's first native graph processor – smallest and lowest power consuming, digital, fully programmable neural network.

CortiCore AI Processor Family

Why neural network acceleration?

Invention of silicon based programmable processors in early 70s started a revolution that led to complex programs being developed for solving complex problems. But as problems got complex with millions of cases to cover the programs became untenable. Artificial intelligence or neural networks are emerging as the solution to the complex untenable programs. The neural networks train on example data, much like human beings and applies that learned knowledge to provide solutions like object detection, object localization and identification, natural language processing etc. to create autonomous cars and robots etc.

The challenge with the neural networks is the amount of computation required. As shown in Figure 1 the compute required has grown 300,000x in the past 6 years itself and it continues to grow. This compute requires a hardware-based compute acceleration.

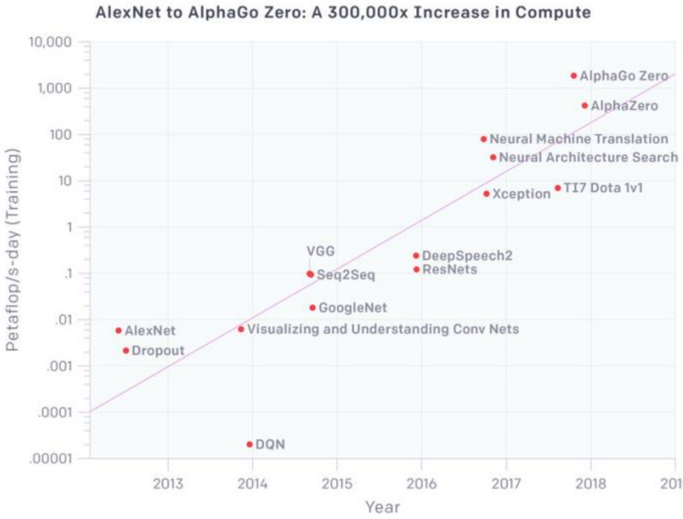


Figure 1.

The energy problem

Large compute in short order of time and processing of large neural net models stored in memory consumes tremendous amount of energy. Following table shows the energy consumption as a factor of compute. Note the throughput in this table is in GOP/s whereas real applications require TOP/s.

Platform	Throughput	Power	Throughput per power	Source
ASIC	74.6 GOP/s	278 mW	0.268 TOP/s/W	Chen et al. (2016)
Xilinx Zynch ZC706	137 GOP/s	9.63 W	0.0142 TOP/s/W	Qiu et al. (2016)
NVIDIA TK1	155 GOP/s	10.2 W	0.0152 TOP/s/W	Chen et al. (2016)
Titan X	3.23 TOP/s	250 W	0.0129 TOP/s/W	Han et al. (2016a)

Table 1: Throughput and power consumption of different accelerator platforms. All works implement an ImageNet network.

**https://nuit-blanche.blogspot.com/2016/05/thesis-ristretto-hardware-oriented.html*

Coupled with the large compute is large models. Large models stored in the memory consume majority of the power.

CortiOne

Various applications have different power and performance requirements. It is critical to be able to provide different acceleration for optimal power consumption. We have devised a Platform approach with one hardware and one software for all applications and performances as shown in Figure 2 below.

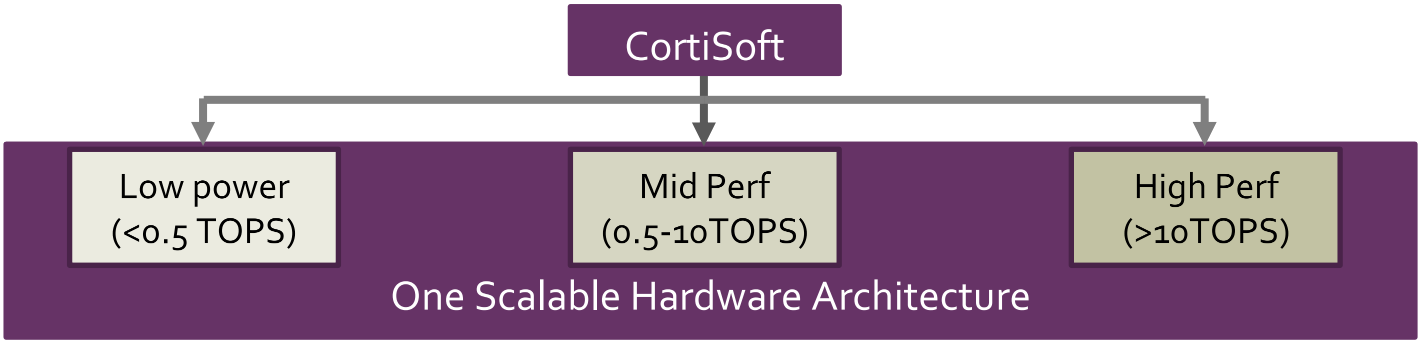


Figure 2.

The challenge with AI acceleration is not limited to the hardware, instead we believe software is the biggest challenge. Being able to run any neural network at high utilization (>80%) and low memory usage is the challenge. Feeding the 3-dimensional neural network data using a traditional instruction set makes the compiler intangible in terms of achieving utilization and there by low power and size. CortiCore architecture provides the solution via its unique instruction set that dramatically reduces the compiler complexity. The approach allows us to create a compiler that achieves >80% utilization with 16X reduced memory (compared to currently available solutions) on all neural networks – demonstrated on our FPGA platforms.

Key features

- Any frameworks, any NN, any backbone
- GRAPH SIMD instruction set - makes compiler possible
- AI Data movement and compute-oriented instructions
- >80% compute utilization
- Highly parallel design - high performance at low frequency of operation
- Unique memory architecture - 16X smaller activation memory
- Memory arch reduces data movement dramatically
- Model and activation memory stays in sleep most of the time
- Data traversal-based activation memory reduction >10X - under compiler control
- Implements sparse NN efficiently, reducing model size and compute requirement by >3x
- Efficiently handle both input-stationery and weight-stationary
- All digital logic - implement in any process node
- Very low host code support to run the AI processing job
- Scalable from 0.1 TOPS to 100 TOPS