EDGECORTIX®

Dynamic Neural Accelerator®

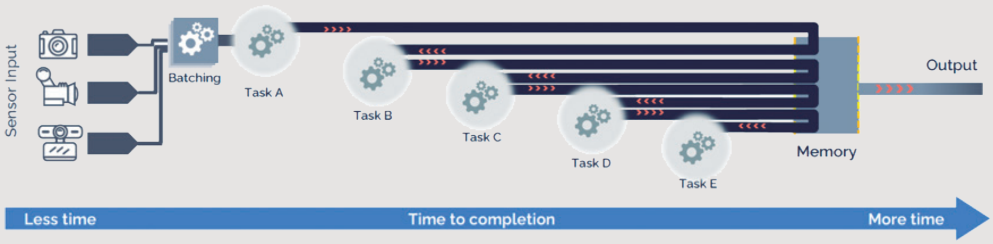
Run-time Reconfigurable Neural Network IP

EdgeCortix Dynamic Neural Accelerator II (DNA-II) is a highly-efficient and powerful neural network IP core that can be paired with any host processor. Achieving exceptional parallelism and efficiency through run-time reconfigurable interconnects between compute elements, DNA-II has support for both convolutional and transformer networks, and is ideal for a wide variety of edge AI applications. DNA-II provides scalable performance starting with 1K MACs supporting a wide range of target applications and SoC implementations.

DNA-II Efficiency

Tera Operations Per Second (TOPS) ratings are typically quoted using optimal conditions, with compute units fully parallelized. When AI models are mapped to other vendors hardware, the parallelism drops - and achievable TOPS falls to a fraction of their claimed peak capability. EdgeCortix reconfigures data paths between DNA engines to achieve better parallelism and reduce on-chip memory bandwidth, using a patented runtime reconfigurable datapath architecture.

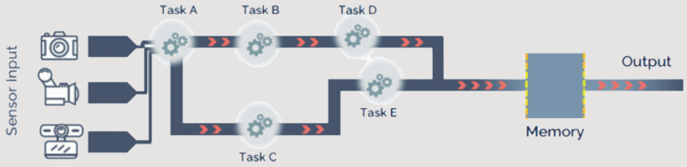
Typical AI Inference Flow in Industry IP Cores



IP Core Inefficiencies

- Slower processing due to batching
- Higher power consumption due to higher re-use of resources
- Low compute utilization resulting in lower efficiency

DNA with Runtime Reconfigurable Datapath

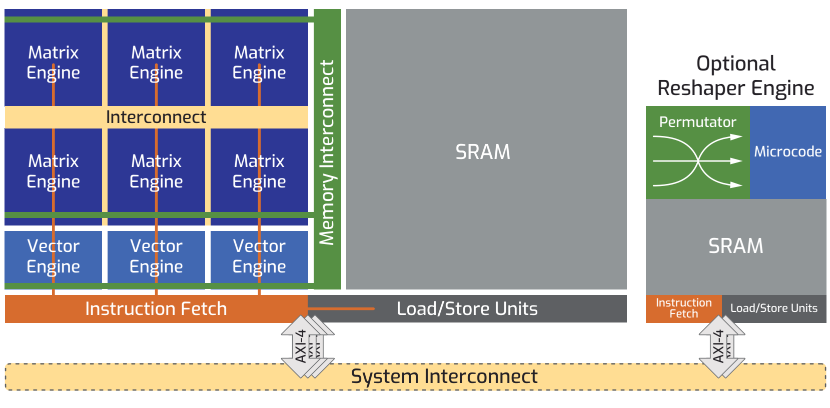


DNA Datapath Advantages

- Much higher utilization and efficiency
- Significantly faster processing due to task and model parallelism
- Very low power consumption for edge AI use cases

The MERA Compiler and Software Framework optimizes computation order and resource allocation in scheduling tasks for neural networks, even with several models running simultaneously in the same DNA engine, resulting in better performance, greater energy-efficiency and lower latency.

DNA-II IP Subsystem



Key Benefits of DNA IP with MERA

Customizable IP Implementation: Achieve desired performance (TOPS), size, and power for target implementation and process technology

Optimized for Generative AI: Supports popular Generative AI models including LLMs and LVMs

Efficient AI Compute: Achieves very high AI compute utilization, resulting in exceptional energy efficiency

Real-Time Data Streaming: Optimized for low-latency operations with batch=1

Arbitrary Activation Function Support: Hardware accelerated approximation provides enhanced adaptability

Advanced Precision: Software-enabled mixed-precision provides near FP32 accuracy

Efficient Memory Utilization: Reduces memory footprint and optimizes DRAM bandwidth, using efficient memory compression

Power Management: Advanced power management enables ultra-high efficiency modes. DNA Cores can be optimized for power/performance trade-offs

Flexible Compute Blocks: Compute block uses optimized systolic arrays from 32x32 to 64x64 elements. Separate vector unit powers neural network activation and scaling

Optional Reshaper IP: Integrated tensor reshaper engine minimizes host CPU load

DNA IP Provides Industry Leading Performance and Power

The table below shows examples of IP capabilities based in the 12nm TSMC FinFET process.

IP can be customized to meet specific user requirements.

Example DNA IP Implementations

Configuration	1	2	3	4	5
MACs	1K	2K	4K	8K	12K
Matrix Engines	1	2	4	2	3
Vector Engines	0.5	1	2	1	1
Channels	32	32	32	64	64
Memory (MB)	2	2.5	4	5	6
TOPS (@ 800MHz*)	1.6	3.3	6.6	13.1	19.7

*DNA-II supports clock frequencies from 100MHz

Deliverables

- DNA IP Source Code with customizable multi-core options
- MERA Compiler and Software Framework
- Large set of predefined models
- Test Bench example
- Complete documentation
- Technical support
- Optional Reshaper Engine IP

Developing with the EdgeCortix DNA IP and MERA

MERA is the compiler and software framework enabling deep neural network graph compilation and AI inference with the DNA IP. With built-in support for the open-source Apache TVM compiler framework, it provides the tools, APIs, code-generator and runtime needed to deploy a pre-trained deep neural network bitstream to the DNA IP. MERA supports a model development workflow using tools including PyTorch, TensorFlow, TensorFlow Lite, and ONNX.

Learn more about DNA AI IP Cores

Explore the MERA Compiler Framework