SOLUTION BRIEF

Myrtle.ai MAU Accelerator

XILINX

MAU™ Accelerator – Low Latency Inference Accelerator for Data Centers

INTRODUCTION

Highest throughput for latency-constrained inference workloads:

- Deterministic low tail latency
- Improved latency-bounded throughput
- Reduced infrastructure costs
- Enables use of higher quality models under a given latency bound
- Reduced energy consumption

PRODUCT OVERVIEW

Low latency inference acceleration for real-time, memory-bounded workloads including:

- Speech transcription
- Natural language processing
- Speech synthesis
- Time series analysis
- Payment & trading fraud detection
- Recommendation systems

The Challenges

Increased use of AI creates a variety of challenges for companies. Those seeking to achieve high-accuracy inference in real-time applications face significant challenges due to the latency constraints of such applications. Typical approaches to addressing high tail latency involve employing more compute resources. Consequently, companies using cloud services face large increases in costs for more compute capacity. Those with their own data centers need to install more servers, which leads to rising CapEx costs and pressure on physical space.

- Increasing energy demands not only lead to higher OpEx costs but also an increased carbon footprint and potentially issues with power constraints.
- As the complexity of DNNs increases faster than the performance of traditional solutions, these issues become magnified.

XILINX

Compelling Advantages

- Speech Transcription example:
  - 165x higher performance than a CPU-only solution
  - 2.1x higher performance per watt than a GPU solution
  - 29x lower latency than a GPU solution
- Natural Language Processing example:
  - 2.2x lower cost than a CPU-only solution
  - 7.7x smaller carbon footprint than a CPU-only solution
- Speech Synthesis example:
  - 148x higher latency bounded throughput than a GPU solution, at half the latency
  - More advanced model deployment with same throughput

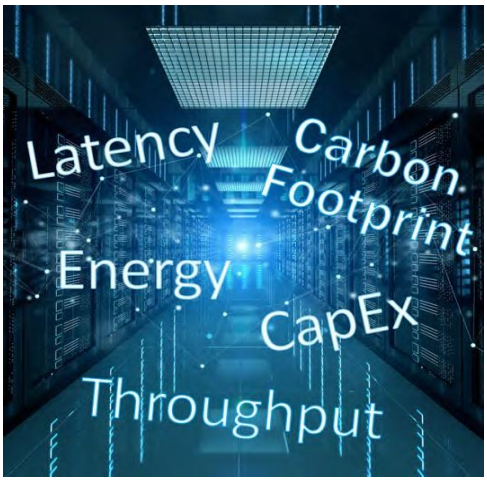
Adaptable. Intelligent.

SOLUTION OVERVIEW

The MAU Accelerator is a low latency inference accelerator for data center machine learning workloads. It achieves both deterministic low tail latency and high throughput, without trading off one against the other. This enables higher quality models to be deployed, providing better services and customer experiences, while significant savings can be made in infrastructure costs and energy consumption.

The MAU Accelerator runs on a server enhanced by a data center accelerator card, the Alveo U250 from Xilinx. These accelerator cards are available today, both in the cloud and for on-premise data centers, facilitating rapid implementation at scale. Neural network models created using popular ONNX supported frameworks such as TensorFlow, PyTorch or MXNet can easily be deployed on the MAU Accelerator, which is ONNX Runtime supported.

Many real-time applications with high throughput, low latency workloads can benefit from the MAU Accelerator today. As development of evermore powerful DNNs continues at pace, offering higher quality results but demanding more compute resources, the benefits of the MAU Accelerator become even more compelling. The reprogrammable nature of the MAU Accelerator means that installed solutions can be upgraded in the field as required in the future.



TensorFlow

mxnet PyTorch

Open-source machine learning frameworks are supported

TAKE THE NEXT STEP

For more information on the MAU Accelerator, please visit myrtle.ai/MAU.
To evaluate what the MAU Accelerator can do for your business, contact Myrtle.ai at hello@myrtle.ai

XILINX