Myrtle.AI launches new domain-specific, sparsity-exploiting inference accelerator at Xilinx Developer Forum

Cambridge, UK, November 12th 2019 – Myrtle, a recognized leader in optimizing DNN inference on FPGAs, today launched a new domain-specific RNN inference accelerator on the Alveo U250 Datacenter Acceleration Card. The MAU Accelerator™ is fully able to exploit unstructured sparsity to deliver performance, power and latency advantages over alternative solutions and achieve cost and energy savings for its customers.

Targeted at developers with latency-bounded throughput challenges, the MAU Accelerator achieves the same throughput as high-end GPUs while consuming a fraction of the energy and with a latency two orders of magnitude lower than a GPU or CPU. Myrtle is engaging today with customers in DNN inference applications such as Natural Language Processing, Query Intent and Recommendation Engines where very high throughput within tight latency bounds is critical to success.

At XDF Europe Myrtle will be demonstrating the MAU Accelerator optimized for the Alveo U250, running an MLPerf.org benchmark. The metrics achieved for this speech transcription workload include an unrivalled throughput of 54 TOPS at 62 watts, with a latency of only 1.09ms.

“FPGAs are addressing the needs of a wide range of AI inference applications especially where there are tight latency bounds”, said Peter Baldwin, CEO, Myrtle. “Traditional AI platforms are burdened with high latency and batch size, which limit the number of workloads which can be processed in a given time window and cause scheduling issues for multiple asynchronous workloads with differing Service Level Agreements.”

Myrtle’s expertise in hardware-software codesign and the quantization, sparsity and compression of machine learning models has been recognized by the MLPerf consortium. Myrtle co-chairs the speech working group, owns the speech transcription workload benchmark and has open-sourced its code to help the industry benchmark new edge and data center hardware more consistently.

More details about how to achieve a step change improvement in throughput for latency-bound applications can be found on www.myrtle.ai or contact Myrtle today on Xilinx_eval@myrtle.ai.

About Myrtle

Myrtle optimizes inference workloads such as speech transcription and recommendation engines for neural networks deployed on Xilinx® FPGAs in data centers and in embedded applications. This enables businesses to reduce costs and offer enhanced, scalable services to their customers. Myrtle is proud to be an Accelerator Partner in the Xilinx Partner Program.
For more information, please visit www.myrtle.ai and follow us on twitter.

Contact:
speech@myrtle.ai