SOLUTION BRIEF

Data Center
Myrtle Sparse RNN Acceleration Solution

Accelerating Speech Workloads for the Data Center

Myrtle MAU accelerator cores, implemented in Intel® FPGAs, optimize data center infrastructure, reducing costs and increasing performance

Intel® FPGA Programmable Acceleration Card D5005

Executive summary

Speech applications—for call center automation, translation services, and more—are improving services and reducing operating costs for numerous businesses across multiple industries but are placing growing demands on data center resources. Speech to text and other recurrent neural network (RNN) workloads require a tremendous amount of computational power and memory, leading to data center bottlenecks, escalating energy demands, and rising costs.

Accelerators based on GPUs or FPGAs are being installed in the cloud and in on-premise data centers to process these workloads more efficiently. While GPUs provided an early solution for graphics acceleration, newer FPGA-based accelerators such as the Intel® FPGA Programmable Acceleration Card (Intel® FPGA PAC) D5005 can deliver superior results for speech applications, supporting 165x the number of simultaneous voice channels of software-only implementation.[1]

Myrtle, a leading expert in machine learning acceleration, has developed an RNN solution for the Intel® FPGA PAC D5005 that achieves a compute performance comparable to that of an NVIDIA Tesla V100* GPU but with a 2x reduction in power consumption. The solution also reduces the latency, or response delay—which is critical for real-time voice applications such as interactive customer-facing voice services—by 29x.[1]

For businesses that depend on, or plan to implement, speech synthesis, speech transcription, machine translation, or other RNN workloads, the Myrtle solution offers a smart way to reduce the total cost of ownership of data center infrastructure while creating capacity to handle higher peaks in demand, facilitate higher levels of automation, or add new revenue-generating services quickly and cost-effectively.

While this solution brief focuses on speech, other RNN workloads associated with genomics and financial applications also benefit from the Myrtle acceleration solution.

myrtle.ai

Business challenge: Rising data center demands

Demand is rising among both consumers and businesses for advanced voice-related services such as speech recognition, natural language processing, and speech to text. These RNN workloads—which comprise 29 percent of all data center deep learning inference workloads[2]—place exceptional demands on data centers.

In 2013, Google projected that it would need to double its number of data centers to meet the computational requirements of people using just three minutes of voice search per day.[2] In its data centers, Facebook researchers say that a "significant fraction of the future demand is expected to come from workloads corresponding to DL [deep learning] inference."[3]

At the same time as demand is growing for deep learning inference models, the models are becoming more sophisticated and demanding, leading to higher compute and memory requirements. RNN models can be scaled to contain millions or even billions of parameters, resulting in rapidly escalating computational and energy costs. Innovation is also driving a continual influx of new, even larger RNN models, such that hardware optimized for today's models can quickly become irrelevant and inefficient.

In response to these issues, businesses can scale out their data centers, adding more infrastructure and computing hardware, but doing so is extremely expensive and requires potentially untenable increases in electricity consumption. Another approach is for businesses to build their own application-specific integrated circuits (ASICs), but this option can take more than 18 months and cost millions of dollars, and it cannot evolve. By the time the product is deployed, the rapid evolution of machine learning applications threatens to make the system architecture and unmodifiable circuitry suboptimal and potentially obsolete.

These and related challenges are why businesses are looking for technology solutions for RNN and other machine learning workloads that are reconfigurable, deliver better performance per watt, and are low latency and low cost.

Solution: Highly optimized accelerators for Intel FPGAs

Taking advantage of its years of deep learning experience, Myrtle has developed its proprietary accelerator core, MAU, that is optimized to process some of today's most demanding RNN workloads and scale as needed to meet emerging demands.

Each MAU accelerator core is optimized for high-performance computation on unstructured sparse matrices. Cores contain features to support RNN computation, bidirectional LSTM pointwise operations, and nonlinearities that are required by machine learning speech algorithms. Several cores can be combined to support wide matrix multiplication, allowing flexible configuration of the accelerator grid to target different neural network descriptions.

After analyzing a given RNN model and workload such as speech to text, Myrtle configures a network of its MAU accelerator cores to take full advantage of the features and capacity of the Intel FPGA PAC D5005 board.

The Intel FPGA PAC D5005 running in an Intel® Xeon® processor-based server creates a heterogeneous computing platform with different compute engines (CPU and FPGA) that allow the workload to be partitioned and optimized, running the heavy computational aspects of the RNN on the Intel FPGA PAC D5005 while allowing the CPU to focus on those aspects of the workload to which it is best suited.

Tests show that the combined Myrtle and Intel FPGA PAC D5005 solution can support more than four thousand real-time voice channels on a single chip, comparable to the performance of an NVIDIA Tesla V100 GPU. Compared with an Intel Xeon processor-only implementation[4] supporting 28 real-time voice channels, RNN processing on the Intel FPGA PAC D5005 runs at 165x the number of real-time channels, offering a step-change in processing capacity for the same server socket infrastructure.

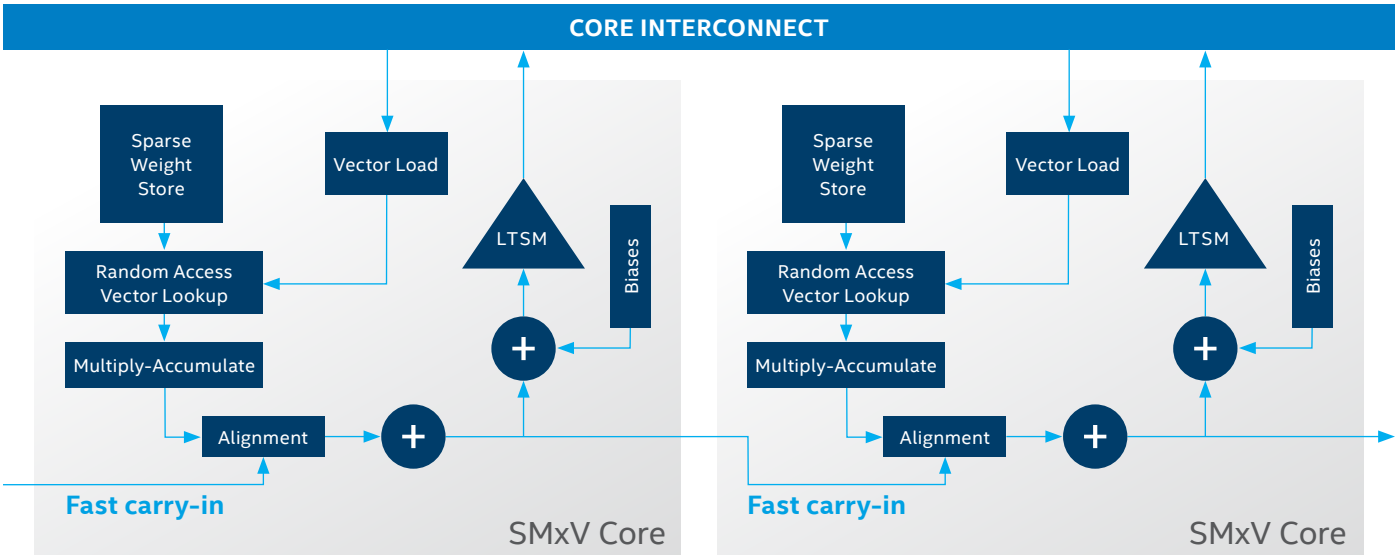

Figure 1. Myrtle's MAU accelerator cores are optimized for and deployed on Intel® FPGA PAC D5005s.

Unlike fixed devices, Intel FPGA PAC D5005 boards are dynamically reconfigurable and can be quickly updated to run the latest RNN models and numerics. In addition, unlike ASICs, FPGAs can be reconfigured to perform other tasks during periods of low load or updated with new algorithms to adapt to the workload's changing needs.

The Intel FPGA PAC D5005 is supported by the Intel® Acceleration Stack for Intel® Xeon® CPUs with FPGAs, which provides a common developer interface and includes drivers, application programming interfaces, and an FPGA Interface Manager. The Intel® Acceleration Stack works with industry-leading operating systems and virtualization and orchestration software, providing a common interface for software developers to get fast time to revenue, simplified management, and access to a growing ecosystem of acceleration workloads.

Myrtle has expertise in compressing RNN models without loss of accuracy and scaling to meet different performance and cost requirements, making it possible to target not only data center applications but also real-time applications at the edge, giving businesses opportunities to grow as voice and speech services proliferate on mobile devices, in vehicles, and in other new applications.

Solution components

- Myrtle MAU accelerator cores, high-performance sparse linear algebra accelerators with features for RNN processing
- Intel® FPGA PAC D5005, a high-performance PCI Express* (PCIe*)-based FPGA acceleration card for data centers
- Intel® Acceleration Stack for Intel® Xeon® CPUs with FPGAs, which provides a common interface, drivers, APIs, and an FPGA Interface Manager to save developers time
- Intel® Xeon® processor-based servers

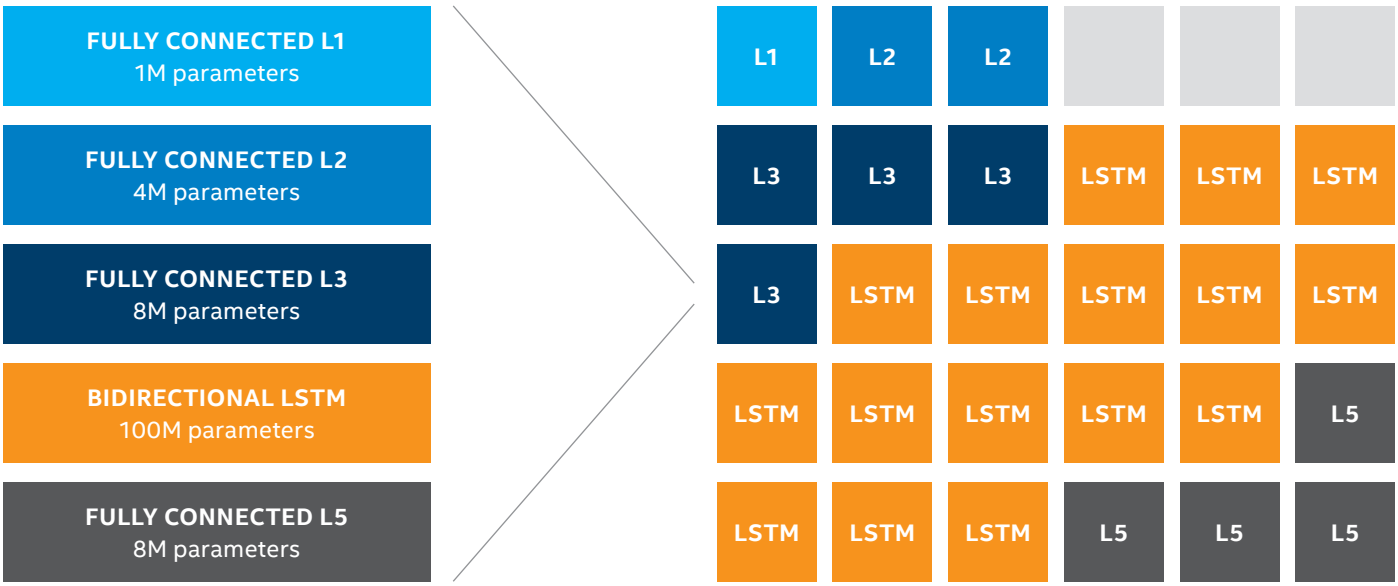

Figure 2. Myrtle's scalable MAU accelerator cores integrated on Intel® FPGA PAC D5005 boards for a DeepSpeech* speech-to-text workload.

Use case: FPGA-based acceleration for speech to text

Myrtle partnered with the Intel Network and Custom Logic Group (NCLG) to develop RNN-optimized speech recognition algorithms on high-performance Intel FPGA PAC D5005 boards. The solution was tested on DeepSpeech*, an open source speech-to-text model from Baidu that is typical of RNNs deployed in production systems. The solution was compared with an Intel Xeon processor-only solution and a standard GPU processing model based on an NVIDIA Tesla V100 GPU.

For the DeepSpeech solution, Myrtle harnessed two optimization techniques—sparsity and quantization—to optimize throughput performance, reduce latency, and increase energy efficiency. Myrtle's in-house machine learning team trained features into the DeepSpeech model, allowing efficient computation on an FPGA with negligible loss in model accuracy.

Sparsity

Sparsity decreases the total effective number of parameters by explicitly setting some values to zero, a technique that cannot be efficiently exploited by GPUs or CPUs. With suitable hardware architectures created on the FPGA fabric, the use of sparsity significantly improves the effective arithmetic intensity of a system and reduces memory requirements because the parameters do not need to be stored. For an FPGA platform, that means weights can be stored entirely in on-chip RAM, enabling extremely high-bandwidth access to model parameters and a highly power-efficient solution.

The test showed that Myrtle can tune MAU accelerator cores to very high levels of sparsity with negligible loss of accuracy—reaping the performance rewards of sparsity without reducing the user experience.

Quantization

Quantization transforms floating-point numbers into narrow integers, reducing the number of bits used during inference. In the test, Myrtle showed that it could quantize to eight-bit integers, which reduces the size of the RNN model by a factor of four, resulting in an immediate 4x improvement in arithmetic intensity and a 4x decrease in the data bandwidth requirement. Again, this technique was used with negligible loss of accuracy.

By treating machine learning workloads as an algorithm-accelerator codesign challenge, Myrtle demonstrated a substantial increase in system performance, allowing significant algorithm offload from the Intel Xeon processor-based server to the Intel FPGA PAC D5005 board.

	MYRTLE MAU ACCELERATOR	GPU	MYRTLE MAU ACCELERATOR ADVANTAGE
Platform	Intel® FPGA PAC D5005[a]	NVIDIA Tesla V100*[b]	
Frequency (MHz)	250	1530	
Sparsity (%)	96	0	
Quantization	8-bit integer	16-bit fp	
Batch size	1	256	
Effective thoughput (TOPS)	54	53.37[c]	Comparable
Power (W)	102	216	2.1x lower power
Performance per watt (effective GOPS/W)	529	247	2.1x greater GOPS/W
Latency per 1s input audio (ms)	0.343	126	365x lower latency

Figure 3. Testing shows a solution that combines an Intel® Xeon® processor with a Myrtle MAU accelerator running on an Intel® FPGA PAC D5005 implements equivalent throughput compared to the NVIDIA Tesla V100, but at half the power. Latency is reduced by 365x.

When compared to a GPU acceleration model, the FPGA achieves performance and latency simultaneously, with better performance per watt. For GPU acceleration, performance must be traded against latency requirements, reducing achievable performance for real-time applications. See Figure 4 for a comparison of the CPU and GPU when the system is optimized for latency.

	MYRTLE MAU ACCELERATOR	GPU	MYRTLE MAU ACCELERATOR ADVANTAGE
Platform	Intel® FPGA PAC D5005[a]	NVIDIA Tesla V100*[b]	
Frequency (MHz)	250	1530	
Sparsity (%)	96	0	
Quantization	8-bit integer	16-bit fp	
Batch size	1	1	
Effective thoughput (TOPS)	54	1.12	48x throughput advantage
Power (W)	102	191.8	1.8x lower power
Performance per watt (effective GOPS/W)	529	5.84	90x higher GOPS/W
Latency per 1s input audio (ms)	0.343	10.1	29x lower latency

Figure 4. Testing shows that even when the GPU system is optimized for latency, the MAU running on the Intel® FPGA PAC D5005 performs with a latency 29x faster and a throughput 48x greater than the NVIDIA Tesla V100.

a. Intel® FPGA PAC D5005 card measurements taken in conjunction with Intel® Xeon® i7-7700K processor at 4.20 GHz, RAM 4 x 16 GB at 2800 MHz, 1 TB M.2 PCIe* SSD, PRIME Z270-P* motherboard, 650 W PSU, Ubuntu*.

b. NVIDIA Tesla V100* measurements taken using NVIDIA Tesla V100 instance on Google Cloud in conjunction with Intel® Xeon® CPU at 2.30 GHz, CPU(s): 12, 4 x 16 GB RAM.

c. Peak throughput of 53.37 TOPS measured over shorter input duration of 200 ms. When measuring latency over a 1s input period, peak throughput drops to 23 TOPS.

Solution value: High performance and low latency without sacrificing accuracy

Tests show that implementing Myrtle MAU accelerator cores on Intel FPGA PAC D5005 boards results in several benefits that are critical for businesses that provide voice and speech services:

- **Efficient performance:** The solution provided comparable throughput and 2x greater performance per watt than the NVIDIA Tesla V100 GPU, leading to lower power consumption and operational cost.[1]
- **Low latency:** The solution delivered a 29x improvement in latency compared to the NVIDIA Tesla V100 GPU.[1] For low-latency processing, the effective throughput of the GPU fell dramatically while the MAU accelerator provided high performance and low latency simultaneously. This enables multichannel interactive speech services to be achieved without noticeable lag to the end user, and within a cost envelope that makes these services viable in large-scale deployments.
- **Accuracy:** The extremely high performance levels were attained with less than 0.23 percent loss of accuracy. This enables a step change in platform performance without compromising the end user experience.
- **Expandability:** The solution can support RNN processing at 165x that of software-only implementations.[1] This vastly reduces the operational cost per voice channel, which can translate to orders of magnitude greater profitability. Moreover, a significant reduction in server infrastructure requirements enables on-premise deployment of multichannel voice solutions that were previously impractical due to physical space and infrastructure restrictions.
- **Flexibility:** Because FPGAs are dynamically reconfigurable, the solution can be continually updated and optimized to meet future demands. This prolongs the useful life of deployed hardware and is particularly important in the rapidly evolving field of machine learning.

Figure 5. Reducing server infrastructure requirements substantially reduces TCO, and enables on-premise solutions that previously were not feasible.

Adding Intel FPGA PAC D5005 boards to existing Intel Xeon processor-based server deployments can increase the processing capability of the server by 10x to 80x overall, freeing up the Intel Xeon processors to be utilized for other functionality.[5] Relative to a GPU-based implementation, the Intel FPGA PAC D5005 implementation can help reduce OpEx by 27 percent.[6] In applications where latency is important, the Intel FPGA PAC D5005 offers 48x throughput, adding significant CapEx savings as many multiples of GPUs and servers would be needed for equivalent performance.

Conclusion

To meet stringent latency, power, and cost requirements, speech to text and other RNN workloads require tightly coupled hardware and software solutions. Compared with traditional CPU- and GPU-based solutions, Myrtle's MAU accelerator cores running on Intel FPGA PAC D5005 boards enable businesses to run speech applications on fewer servers, reducing infrastructure and operating costs while meeting more stringent constraints on data center power and floor area. Myrtle's scalable solution then frees up server capacity so businesses can handle higher peaks in demand, facilitate higher levels of automation, or add new revenue-generating services quickly and cost-effectively.

With Myrtle and Intel, businesses can reduce costs and increase the quality and range of speech services they offer their customers while retaining the flexibility to reap the benefits of the latest advances in machine learning for speech applications as they emerge.

About Myrtle

Myrtle is a technology company based in Cambridge, UK, that has developed powerful new software to accelerate deep learning inference on FPGAs. Myrtle is a leading expert in the creation of optimized implementations for speech applications in data centers. The company's codebase and models for speech inference are being used to benchmark new edge and data center hardware for the MLPerf consortium, an industry-led machine learning benchmarking effort.

Learn more

Myrtle's MAU accelerator is available for deployment on Intel PACs. Contact Myrtle at stratix_eval@myrtle.ai to explore the solution and evaluate the benefits for your business.

Learn more about Intel FPGA solutions and about Intel Programmable Acceleration Cards and the Acceleration Stack.

Performance results referenced in this document are based on running DeepSpeech* software on Stratix® 10 FPGA in 2Q19. These tests were run by Myrtle.

1. M. Ashby, C. Baaij, P. Baldwin, M. Bastiaan, O. Bunting, A. Cairncross, et al., "Exploiting unstructured sparsity on next-generation datacenter hardware." https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/wp/myrtlle-unstructured-sparsity-wp.pdf
2. N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit," in *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2017, pp. 1–12.
3. J. Park, M. Naumov, P. Basu, S. Deng, A. Kalaiah, D. Khudia, J. Law, P. Malani, A. Malevich, S. Nadathur et al., "Deep Learning Inference in Facebook Data Centers: Characterization, Performance Optimizations and Hardware Implications," *arXiv preprint arXiv: 1811.09886*, 2018.
4. DeepSpeech workload run on dual-socket Intel® Xeon® 6140M Gold running at 2.3 GHz processes 28 real-time voice channels with an effective 0.327 TOPS. Intel® FPGA PAC D5005 (at 54.0 TOPS) shows 165x higher throughput.
5. For DeepSpeech 1, Myrtle accelerated 98.7 percent of the neural network on the Intel® FPGA PAC D5005. The remaining 1.2 percent running on the CPU represents an 83x computational offload of the CPU. This excludes audio preprocessing and inference decoding stages of a full speech-to-text implementation.
6. Based on a 10x reduction in server infrastructure with cost estimate for server plus Intel® FPGA PAC D5005 at 1.5x server-only costing; capital expenditure is 1.5/10 of original cost. Based on a comparable number of servers with either an NVIDIA Tesla V100 or Intel® FPGA PAC D5005, assuming operating expenditure is proportional to power and the power of server-only system is 200W. Server + GPU = 416W, server + Intel® FPGA PAC D5005 = 302 W.

1219/YLR/CMD/PDF