Reduce Speech Transcription Costs by up to 90% with CAI (WP030)

Achronix
Data Acceleration

May 24, 2022 **White Paper**

Introduction to CAI

What is Conversational AI?

Conversational artificial intelligence (CAI) uses deep learning (DL), a subset of machine learning (ML), to automate speech recognition, natural language processing and text to speech using machines. The CAI pipeline is usually depicted with three key functional blocks:

1. Speech to text (STT), also known as automatic speech recognition (ASR)
2. Natural language processing (NLP)
3. Text to speech (TTS) or speech synthesis

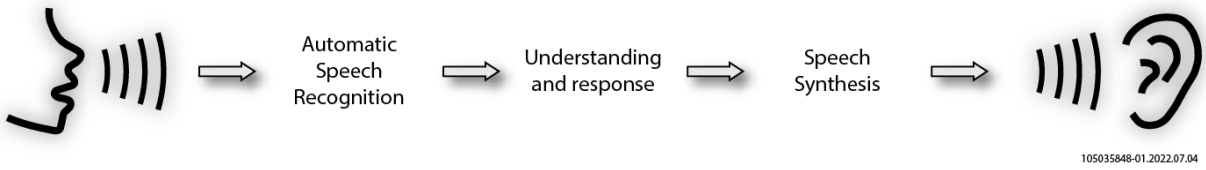

Figure 1: *Conversational AI Building Blocks*

This white paper takes a closer look at automatic speech recognition (ASR) use cases and how Achronix delivers an up to a 90% reduction in costs associated with implementing solutions that use ASR.

Market Segments and Use Cases

With over 110 million virtual assistants in use in the US alone [1], most people are familiar with using CAI services. Primary examples include voice assistants on mobile devices such as Siri from Apple or Alexa from Amazon; voice search assistants on laptops such as Microsoft's Cortana; automated call center agents; and voice-enabled devices such as smart speakers, TVs, and cars.

The deep learning algorithms that power these CAI services are processed on the local electronic device or aggregated in the cloud for remote processing at scale. Large-scale deployments supporting millions of consumer interactions represent an extremely large compute processing challenges which hyperscaler providers have addressed by developing specialized silicon devices to address the processing of these services.

Today, most small enterprises can easily add voice interfaces to their products using cloud APIs provided by the likes of Amazon, IBM, Microsoft and Google. However, when these workloads grow in size (a specific example is described later in this white paper), the costs of using these cloud APIs becomes excessive, driving enterprise companies to seek alternative solutions. Additionally, many enterprise operations have higher data security requirements, necessitating solutions to stay within an organization's data boundary.

Enterprise-scale CAI solutions are needed for:

- Call center automation
- Voice and video communications platforms
- Health and medical services
- Financial and banking services
- Retail and sales kiosks

A Closer Look at ASR Processing

ASR is the first step in the CAI pipeline where speech is transcribed to text. Once text is available, it can be processed in many ways using natural language processing (NLP) algorithms. NLP includes identifying key content, sentiment analysis, indexing, contextualizing content, and analytics. Within an end-to-end conversational AI algorithm, speech synthesis is then used to generate natural sounding voice responses.

State of the art ASR algorithms are implemented with end-to-end deep learning. Recurrent neural networks (RNN), unlike convolutional neural networks (CNNs), are common in speech recognition. As noted in "CNN vs. RNN: How are they different?" by David Petersson from TechTarget[10], RNNs are better suited for processing temporal data, aligning well with ASR applications. RNN-based models require high compute capability and high memory bandwidth to process the neural network model within the strict latency targets required for conversational systems. When real-time or automated responses are too slow, they appear sluggish and unnatural. Often low latency is only achieved at the expense of the processing efficiency which pushes up costs and can become too large for practical deployment.

Achronix has partnered with Myrtle.ai, a technology specialist in FPGA AI inferencing. Myrtle.ai implements performant RNN-based networks on FPGAs using their MAU inferencing acceleration engine. Integrated into the Achronix Speedster®7t AC7t1500 FPGA, the design can leverage key architectural aspects of the Speedster7t architecture (which are explored later in this white paper) to drastically increase the acceleration of real-time ASR neural networks leading to a 2500% increase in number of real-time streams (RTS) that can be processed when compared to a server-class CPU.

Data Accelerators: How to Strike the Right Balance of Resources

Data accelerators, which offload compute, networking, and/or storage processing normally executed by the main CPU, can lead to significant server footprint reductions. This white paper describes replacing up to 25 servers with a single server plus an Achronix ASR-based acceleration card. This architecture dramatically lowers the workload cost, power and latency while increasing workload throughput. However, using data acceleration hardware to achieve high performance and low latency is only useful if the hardware is used efficiently and is cost effective to deploy.

ASR models are challenging for modern data accelerators, often requiring hand tuning to achieve performance greater than single-digit efficiency of the platform’s headline performance specs. Real-time ASR workloads need high memory bandwidth as well as high-performance compute. Data required for these large neural networks is typically stored in DDR memory on accelerator cards. Moving that data from external memory to the compute platform represents the performance bottleneck in this workload, especially when targeting real-time deployments.

GPU architectures are based around data parallel models where smaller batch sizes result in lower utilization of the GPU acceleration hardware, leading to higher costs and lower efficiencies. The performance figures found in data sheets (measured in TOPS) for hardware acceleration solutions are not always a good indication of actual performance as many hardware acceleration devices are under-utilized due to bottlenecks associated with the device architecture. Those figures are referred to as headline TOPS which highlights the processing ability of the accelerator compute engine but leaves out critical factors such as batch size, speed and size of the external memory and any ability to move data between the external memory and the accelerator compute engine. For ASR workloads, focusing on memory bandwidth and moving data efficiently within the accelerator provides a much stronger guide for what accelerator performance and efficiency can be achieved.

The accelerator must have a larger external memory sizes and very high bandwidth. Today's high-end accelerators typically use high-performance external memory capable of 8-16 GB of memory running at speeds up to 4 Tbps. It must also be able to move that data into the compute platform without throttling performance. This data channel, however implemented, between the high-speed memory and the compute engine is, in nearly all cases, the bottleneck in system performance, especially in low-latency applications such as real-time ASR.

FPGAs are designed to provide optimal data routing between memory and compute and so provide an excellent acceleration platform for these workloads.

Achronix Solution Versus Other FPGAs

Competing FPGA architectures in the ML acceleration segment advertise headline TOPS rates for inferencing as high as 150 TOPS. Yet in real-world applications, especially those which are latency sensitive such as ASR, these FPGAs fall well short of their headline TOPS rates due to their inability to efficiently transfer data between the compute and external memory. This loss in performance is due to bottlenecks occurring in the data movement from external memory to the compute engines in the device. The Achronix Speedster7t architecture strikes the right balance of compute engines, high-speed memory interfaces and data transfer, yielding a device that can deliver 64% of the headline TOPS rates for real- time, low-latency ASR workloads.

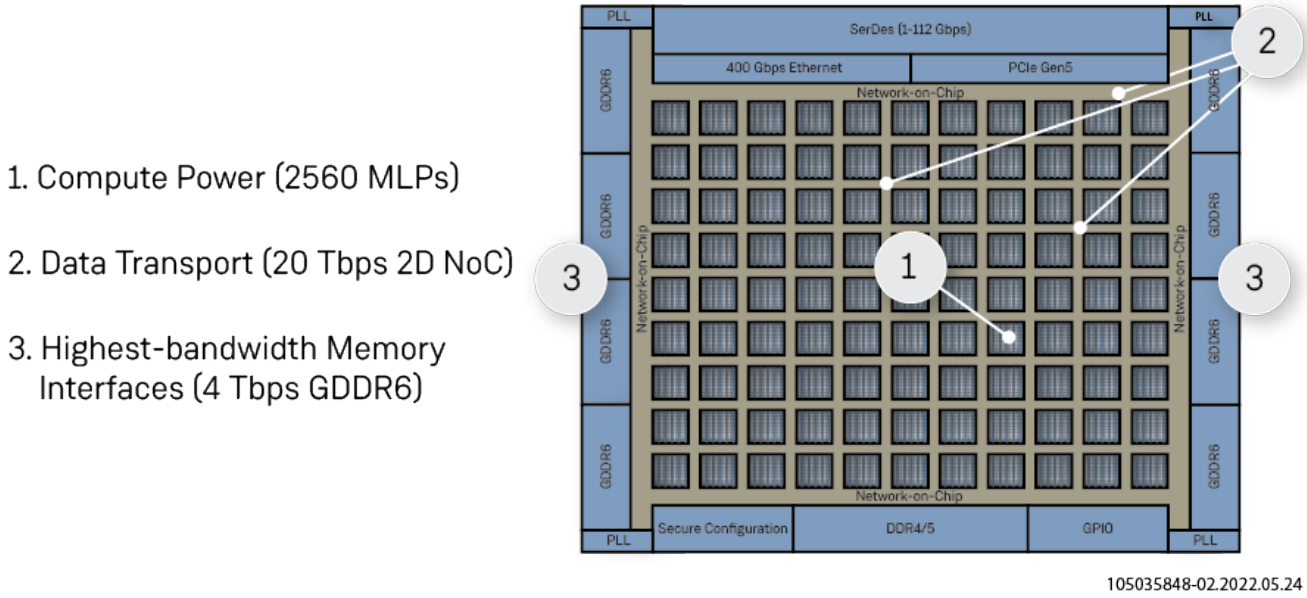

Figure 2: ***Speedster7t Compute, Memory and Data Transfers***

How the Speedster7t Architecture Achieves Higher Compute Efficiency

The machine learning processor (MLP), an optimized matrix/vector multiplication block capable of 32 multiplies and one accumulate in a single clock cycle, is the foundation for the compute engine architecture. Block RAM (BRAM) is co-located with each of the 2560 instances of the MLPs in the AC7t1500, which equates to lower latency and higher throughput.

Myrtle.ai's MAU low latency, high throughput ML inferencing engine has been integrated into the Speedster7t FPGA leveraging key architectural elements

In building the optimal ASR solution, the integration of the previously mentioned MAU inferencing engine from Myrtle.ai leverages 2000 of the 2560 MLPs. Because the MLP is a hard block, it can run at a much higher clock rate than the FPGA fabric itself.

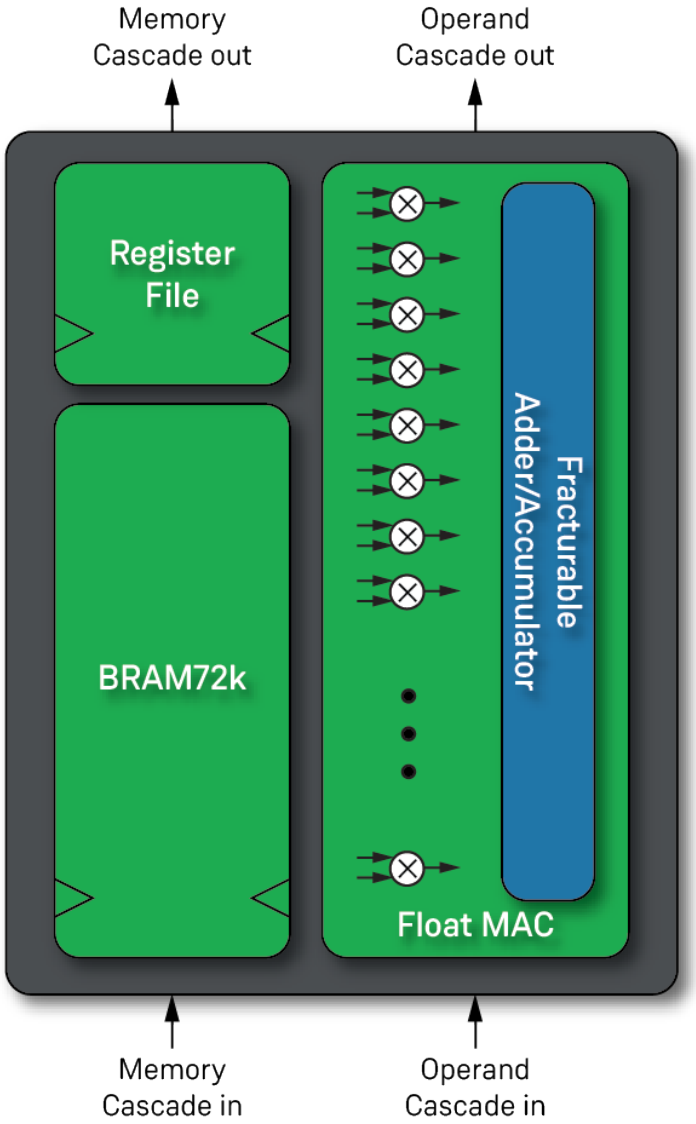

Figure 3: ***Machine Learning Processor***

All eight of the GDDR6 memory controllers are leveraged on the AC7t1500, which delivers up to 4 Tbps of bi-directional bandwidth. As noted above, the strong compute engine and the large, high-bandwidth memory are dependent on high-speed, low-latency, and deterministic data transfers in order to deliver the real-time results needed for low-latency ASR. Enter the Speedster7t two-dimensional NoC (2D NoC). This network is another hard structure in the Speedster7t architecture, with clock rates up to 2 GHz that serves as an interconnect to all I /O, internal hard blocks, and the FPGA fabric itself. With an aggregate bandwidth of 20 Tbps, the 2D NoC provides the highest throughput, and properly implemented, can deliver the most deterministic, low-latency data transfer between external GDDR6 memory and the MLP-based compute engines.

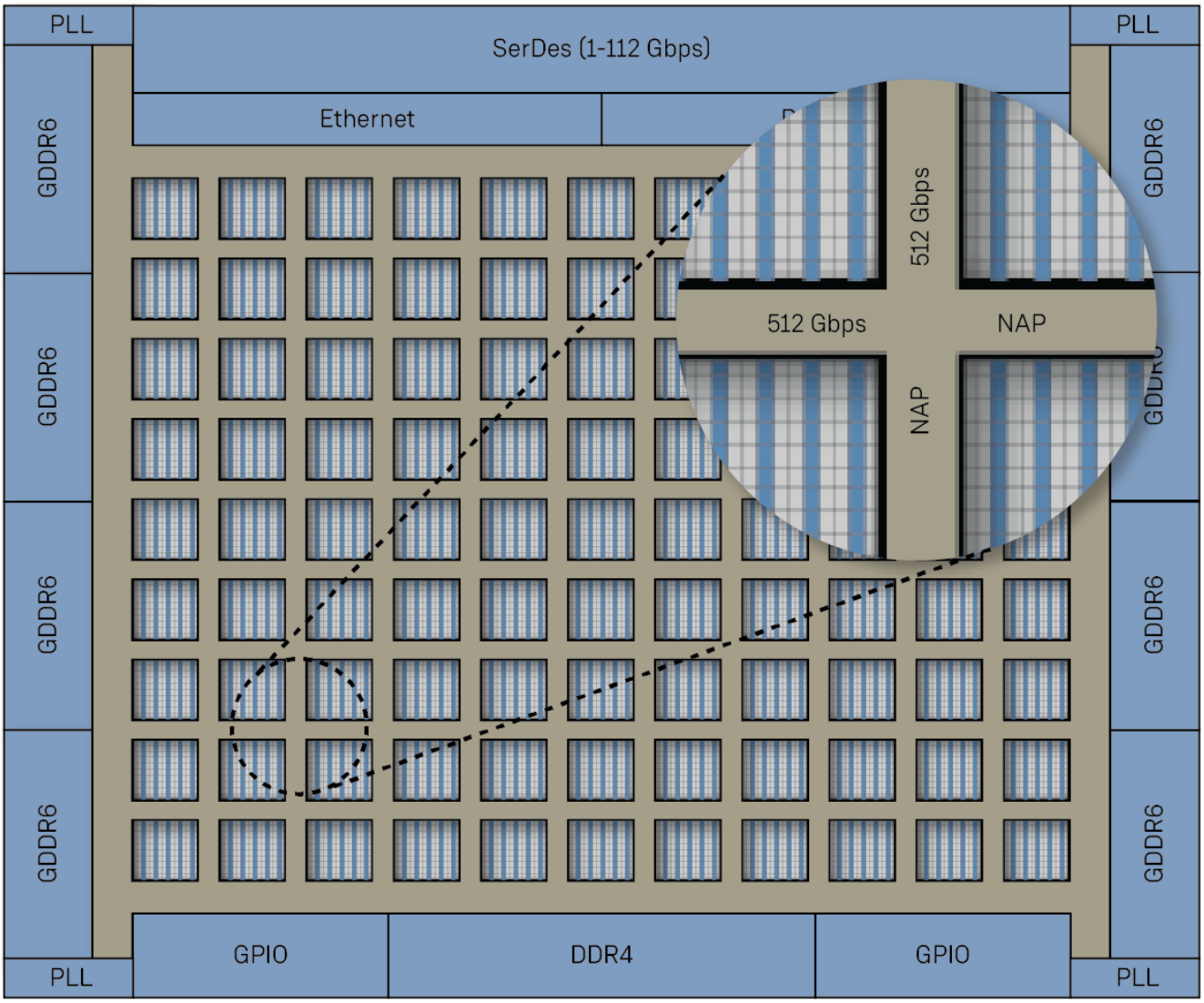

Figure 4: ***20 Tbps 2D NoC***

Unlike the competitors' solutions, the 2D NoC rids the Speedster7t ASR solution of any bottlenecks between the memory and compute engines, yielding a best-in-class utilization rate of the hardware accelerator at these very low batch rates.

Putting all these features together in a roofline diagram clearly illustrates the Achronix Speedster 7t advantage over competing FPGA solutions for low-latency ASR applications. This roofline uses proven TOPS figures published from each manufacturer, showing what these devices can achieve in practice.

The figure below shows a roofline model for effective TOPS using a subset of micro benchmarks (GEMV and MLP) and test builds for Achronix and data published for Company A [4] [5] and Company B (based on architecture). The orange, vertical line indicates the optimal operating point with a batch size of 8 and 80 ms audio block size, for low-latency, real-time ASR streaming applications. At this optimal operating point, Achronix yields a 44% increase in effective TOPS over the Company A and 260% increase over the Company B solution.

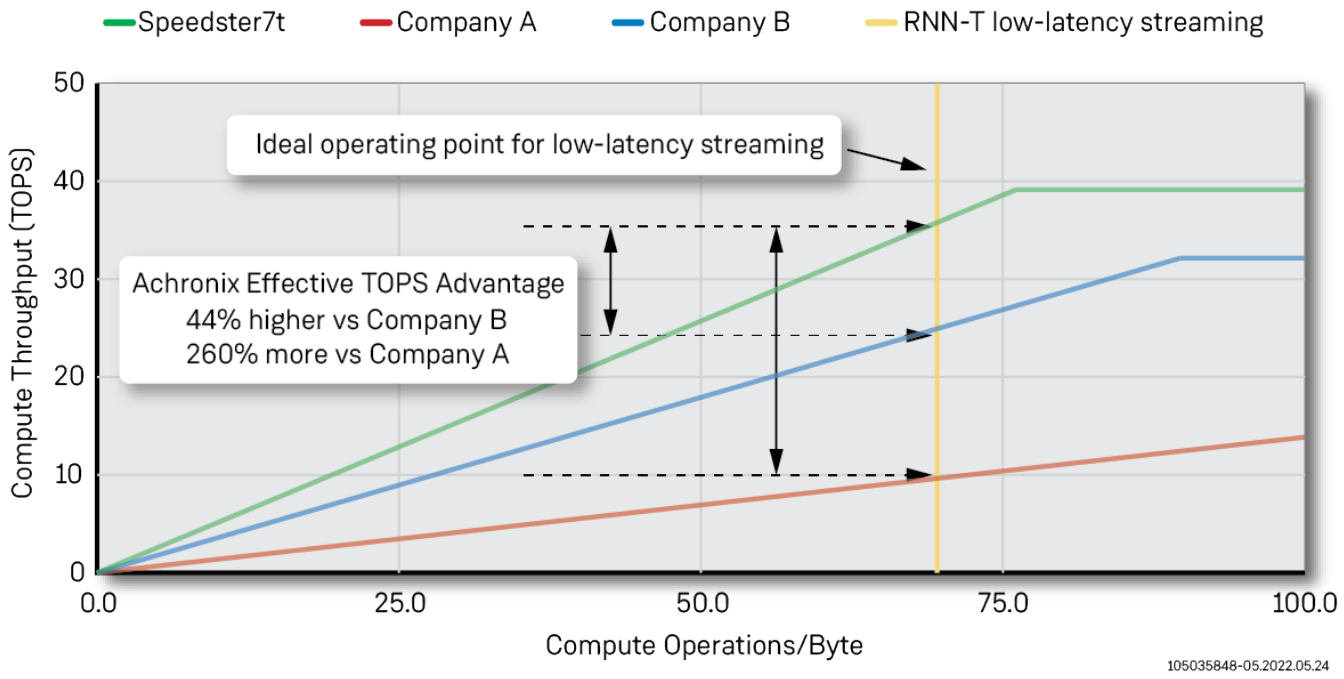

Figure 5: ***Roofline Model for Effective TOPS***

Achieving a 90% Cost Reduction on ASR Processing Within One Year

Most ASR solutions are offered by large-scale cloud service providers such as Google, Amazon, Microsoft Azure, and Oracle. Service providers building products on top of these cloud APIs face increasingly large bills as their operations scale out, and those products achieve success in the market. The publicly advertised cost of the larger ASR providers range from $0.01 to $0.025 per minute [6], [7], [8], [9]. Industry reports suggest that the average call center call is approximately five minutes. Consider a large enterprise data or call center services company fielding 50,000 calls per day at five minutes per call. At the stated rates above, the cost of the ASR processing would range from $1,500 to $6,000 per day or $500,000 to $2,000,000 per year. The Achronix and Myrtle.ai solution can support 4000 RTS on one accelerator card, delivering the capacity to handle over one million calls per day. There are many factors that would dictate the cost of a stand-alone ASR appliance. For this particular example, assume the Achronix ASR acceleration solution delivered on an FPGA-based PCIe card integrated into an x86-based 2U server. Sold from a system integrator, this appliance might be $50,000 and the annual cost of running the server could double that cost. This leads to $100,000 for the first year for an on-premise ASR appliance. Comparing this on-premise solution versus cloud API services, the end user can yield savings of 5× to 20× in the first year.

Table 1: ***Summary of Achronix ASR Solution versus Cloud API***

Item	Speedster7t Solution	Cloud API (Lowest Cost)	Cloud API (Highest Cost)
Cost Per Day (50k RTSs)	$275	$1.5k	$6k
Annual Cost	$100k	$500k	$2,000k
Cost Reduction vs Cloud API	5X to 20X	–	–

The highly compact system enables enterprises to scale as their operations increase without being tied to increasingly expensive ASR cloud APIs and without needing to build huge data center infrastructure to supply on premise solutions.

Summary

The ASR functionality in CAI requires low-latency, high-throughput computation of RNN machine learning algorithms, challenging modern AI accelerators. FPGA hardware accelerators which claim inferencing TOPS rates as high as 150 suffer from bottlenecks associated with the data transfer between the large compute engine and the high-speed memory. Those bottlenecks can result in hardware utilization rates as low as 5%. Achronix and Myrtle.ai are teaming up to deliver an ASR platform consisting of a 200W, x16 PCIe Gen4-based accelerator card and the associated software which together can sustain up to 4000 RTS concurrently, processing up to 1 million five-minute transcriptions per 24-hour period. Comparing this PCIe accelerator card on a single ×86 server to the cost of cloud ASR services, the first year CAPEX and OPEX can be reduced by as much as 90%.

For more information, contact Achronix at www.achronix.com/contact-us.

References:

1. https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/
2. https://www.microsoft.com/en-us/research/project/project-brainwave/
3. Unpublished white paper from Myrtle.ai regarding ASR applications hosted in the Speedster AC7t1500.
4. M. Langhammer, G. Baeckler and S. Gribok, "SpiderWeb - High Performance FPGA NoC," *2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2020, pp. 115-118, doi: 10.1109/IPDPSW50202.2020.00025.
5. https://arxiv.org/pdf/2010.06075.pdf
6. https://aws.amazon.com/transcribe
7. https://cloud.google.com/speech-to-text
8. https://www.ibm.com/cloud/watson-speech-to-text
9. https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service
10. https://www.techtarget.com/searchenterpriseai/feature/CNN-vs-RNN-How-they-differ-and-where-they-overlap

Achronix Semiconductor Corporation

2903 Bunker Hill Lane
Santa Clara, CA 95054
USA

Website: www.achronix.com
E-mail : info@achronix.com

Notice of Disclaimer