Synchronous Scalability for Real-time AI and HPC.

RealScale™ Chip-to-chip (C2C) Interconnect Technology

Author: Groq Inc.

Date: Q3 2022

Version: Groq RealScale™ Chip-to-chip (C2C) Interconnect Technology v0.1

Legal Statements

Synchronous Scalability for Real-time AI and HPC

Groq Tech Doc

Overview

As AI continues to mature, exponential growth of model sizes and compute requirements is becoming the industry standard rather than the exception. However, legacy "real-time" AI and HPC solutions require throughput, accuracy, or latency tradeoffs, or a mixture of all three, posing a huge stumbling block for most users.

Groq provides predictability because of its deterministic Tensor Streaming Processor (TSP) architecture. With Groq, the total execution time is known at compile time, so you know exactly how much time (clock cycles) it will take to run a workload and the exact order of operations—with zero variance. From a single GroqChip™ processor to a scaled network, workloads always run the exact same way the first time and the millionth time—with no performance variation. Legacy solutions do not provide this predictability, especially at scale.

Our RealScale™ chip-to-chip (C2C) interconnect technology is the driving function behind our GroqChip processor's scalability and an integral element of the TSP architecture. Across the Groq product suite, from GroqChip to GroqCard™ to GroqNode™ to scaled GroqRack™ systems, RealScale provides a communication fabric of point-to-point links, eliminating the need for network interface cards, intermediate layer switches, and overprovisioning, all creating cost savings, as well as performance advantages, for the customer.

Ultimately, RealScale translates to a Groq-based system's ability to effectively act as one large compute core rather than merely a network of chips and cores by enabling multiple devices to synchronously process a machine learning model. The result is near-linear scaling, low latency, no congestion across the entire network, and a deterministic execution model. These results are challenging to achieve with architectures because of their use of external switches and hubs, making execution unpredictable and requiring dynamic profiling and finite tuning to achieve acceptable performance at scale. From a single chip up to complex, distributed systems, RealScale delivers predictable reliability at scale with little to no performance loss or latency hit.

RealScale Implementation Details

Communication among GroqChip processors is comprised of sets of vectors (320 bytes each) that are transmitted at prescheduled times so as to be present at the receiving end when needed. Network links (transmit and receive) are explicitly flow-controlled by software, such that they are **scheduled as first-class functional units, like other on-chip processing elements.** This software-scheduled network provides a synchronous communication fabric with multiple ingress/egress ports on each GroqChip for barrier-free communication, with a single global hop in a 264 GroqChip (four GroqRack) system. This enables communication between any arbitrary pair of GroqChip processors in a 264 GroqChip system with approximately 1.8 microseconds of end-to-end latency.

The RealScale interconnect “software-defined” approach perfectly plans out load balancing at compile time, thus avoiding the complexity of deadlock-avoidance, message reassembly, and tree saturation that is commonplace in conventional Ethernet or Infiniband packet-switched networks. With point-to-point Groq network architecture, direct links are all scheduled at compile time, providing routability from card to card. Additionally, Groq provisions multiple paths for redundancy with multiple external links so data gets to where it needs to go.

Ultimately, point-to-point with Groq determinism means capability and flexibility.

Differentiation from Legacy Providers

Just like GroqChip, RealScale technology is deterministic. The interconnect links are phase-aligned and act like a high bandwidth, fixed latency wire between the chips. This enables deterministic execution of programs across multiple chips at both the node and rack levels. In contrast, the interconnect technologies from legacy providers have inconsistent and indeterminate latency. This can result in congestion because traffic across the link flows in an unplanned and unpredictable manner.

The chip-to-chip links connect GroqChip processors within a GroqNode as well as nodes in a GroqRack to achieve rack-scale computing. Each chip-to-chip link consists of four lanes in each direction running at up to 30 Gbps speed. Thus a single chip-to-chip link provides 15 GB/s of effective bandwidth in each direction. With 11 such links per GroqCard, the total available bandwidth of 330 GB/s is 5.1x the bandwidth of PCIe Gen4 and 2.5x the bandwidth of PCIe Gen5.

Benefits

The GroqChip uses a deterministic architecture with 230MB of on-die memory for parameters and instructions. Groq can scale to handle larger problem sizes by adding more GroqChip processors and interconnecting them using RealScale interconnect links, ultimately combining compute, memory, and network bandwidth to tackle larger models. The net result is a cost-efficient system with **lower system power and higher global bandwidth utilization.** More specifically, the Groq network:

- Provides the same throughput as a fat-tree network with half the number of links, and no additional switches. This makes Groq solutions more efficient from a cost per unit of bandwidth and power ($/Gbps/W) basis [ISCA 2008].
- Deterministic execution allows the use of advanced software scheduling of data transfers to maximize usage of resources and reduce network contention.

- Provides more global bandwidth without the network switches necessary for a conventional fat-tree (folded-Clos) network.

Scale Benefits

Automated partitioning tools

No variability between different nodes in cluster

Scaling efficiency (Total Cost of Ownership - TCO) by eliminating sub-optimal compute, network utilization, and need for massive overprovisioning

Application Benefits

Highly predictable service-level agreements due to no variation in application execution time

Choice between performance and energy efficiency by relying on accurately predictable power analysis

User Benefits

Does not require deep architectural or network knowledge to deploy multi-chip workloads

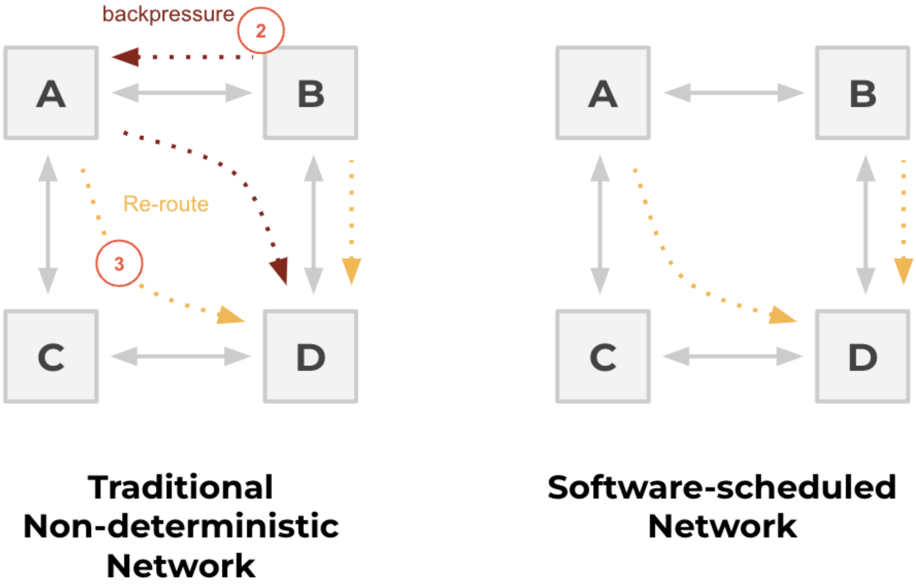
Faster time to production with a compiler flow that does not rely on hand-tuned kernels

Traceability, debuggability, and control deliver predictable execution

Dynamic profiling with no cycle-to-cycle variation

Consider the scenario shown in Figure 1, where both chips A and B want to send a message to D, creating contention for the shared link B to D in the path. In a conventional non-deterministic network scenario, chip A has two minimal route options: two hops via chip B or two hops via chip C. Without awareness of B's downstream congestion, chip A's local algorithm may choose the path via chip B — taking this route would result in chip B asserting flow control to back-pressure A. Once the back-pressure signal has propagated back to chip A, it can then either wait for the route to become free or re-route the data via an alternate path, which adds both complexity to the hardware and latency to the communication.

Instead, the Software-scheduled Network (SSN) moves this decision-making step from hardware to the compiler to schedule data movement across the network. By avoiding contention for the output port on each hop of the path, as mentioned, the RealScale interconnect SSN eliminates complexity and latency, and guarantees deterministic communication.



Traditional Non-deterministic Network

Software-scheduled Network

Figure 1: A traditional non-deterministic network compared to a Software-scheduled Network.

Figure 2 below provides an example of the performance and latency enabled by RealScale as users scale from a single GroqChip to a GroqNode, and a GroqRack.

At rack-scale, three hops deliver end-to-end latency of 1.8µs. Multi-rack systems built from standard eight card GroqNode servers can maintain a fully connected topology with three hops (one external and two internal) up to a maximum of 264 chips. Scaling with more than 264 GroqChip processors is possible by increasing the network diameter.

Scale (# of GroqChip processors)	**Peak INT8/FP16 Performance**	**System SRAM** (GBytes)	**Number of Dimensions**	**Diameter** (hops)	**End-to-end Latency** (µs)
1	750 TeraOps 189 TeraFlops	0.2	N/A	N/A	N/A
8 (1 GroqNode)	6 PetaOps 1.5 PetaFlops	1.76	0 (1 node)	1	0.6
16	12 PetaOps 3 PetaFlops	3.5	1 (2 nodes)	2	1.2
64 (1 GroqRack)	48 PetaOps 12 PetaFlops	14	1 (8 nodes)	3	1.8

Figure 2: Performance and latency enabled by RealScale as users scale from a single GroqChip to a GroqNode, and a GroqRack

By The Numbers

Groq™ Compiler

Two of the biggest advantages of RealScale technology are that it utilizes point-to-point connections and that it enables the ability to statically determine when to schedule transmit and receive units. Both of these advantages allow for a deterministic cycle count regardless of the number of devices in the network. This is different from a traditional non-deterministic network where scheduling and routing is done at runtime, resulting in performance loss and uncertainty.

These capabilities fit well with Groq™ Compiler because the compiler can treat the multiple devices monolithically, scheduling the transmit and receive units as if they were any other functional unit. Furthermore, with partitioning being a fundamental function for scaled compute, it is critical how effectively this partitioning can be balanced across a network of devices. The deterministic Groq architecture extends across multi-chip partitions and enables us to determine exactly how long each operation will take. The benefit of this predictability is that compute and IO communication are balanced, the program achieves higher chip utilization, and the IO network does not introduce ambiguous or asynchronous communication overhead. This leads to a level of workload scalability and agility that delivers performance and TCO improvement.

BERT is a popular natural language processing model. Using BERT-Base as a starting point, as seen in Figure 3, we increase the number of encoder layers as a method for scaling workload intensity as we increase the number of devices being deployed. Even after scaling up to 16 devices, we saw Groq performance grew near-linearly, demonstrating that RealScale can be used to scale out device resources to support growing workloads (e.g. larger, deeper models) without the burden of additional overhead in C2C communications.

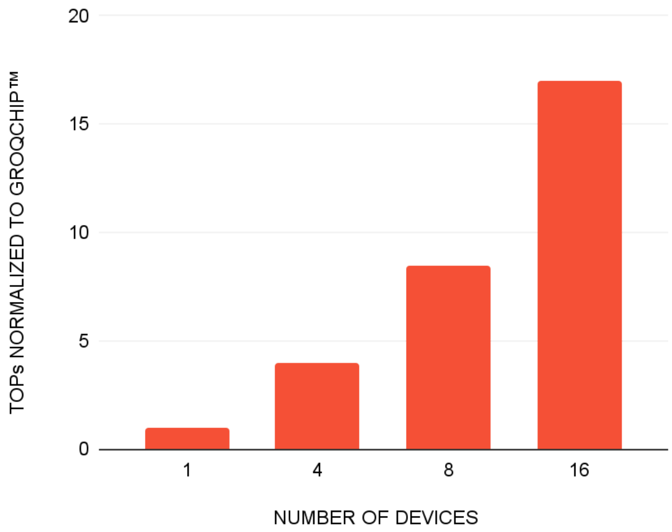


Figure 3: An example using BERT-Base of how RealScale can be used to scale out device resources to support growing workloads.

As seen in Figure 4, we also took a larger version of the BERT model, BERT-Large, and deployed it to various model-parallel topologies. In this use case, a single model can run on multiple devices, up to 16 devices, and act as a single accelerator instance both from the compiler and from the host runtime perspective. **This provides a mechanism to scale up performance with a larger monolithic compute engine but without the burden of host parallelization or multi-device runtime deployment and synchronization.** This demonstrates that RealScale can be used to create an elastic compute fabric that can adapt and scale throughput performance for your existing workloads.

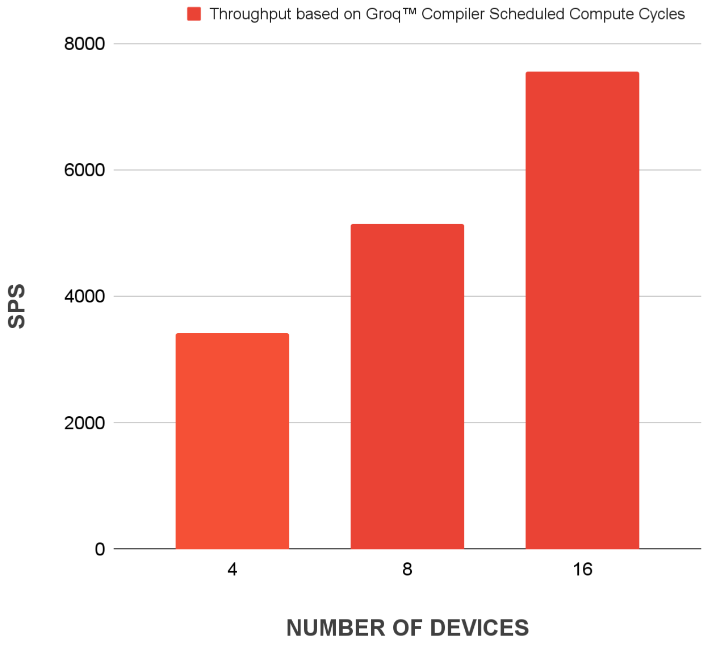


Figure 4: An example using BERT Large of how RealScale can be used to create an elastic compute fabric that can adapt and scale throughput performance for your existing workloads.

Groq API

In addition to compiler use cases, RealScale technology has been used to scale Groq API programs to multi-chip architectures. Fast Fourier Transform (FFT) is a popular signal processing algorithm used to transform an input signal into the frequency domain. Scaling a single-chip GroqChip FFT into a multi-chip program on *N* chips requires an all-to-all data transfer. In other words, each chip needs to pass a chunk of its data to *every* other chip in the network.

The eight card GroqNode has a fully connected RealScale structure, which enables efficient data routing for the FFT algorithm. Each chip can send and receive data on each of its connections in parallel at a fixed cadence without needing a central runtime scheduler to manage the data transfers.

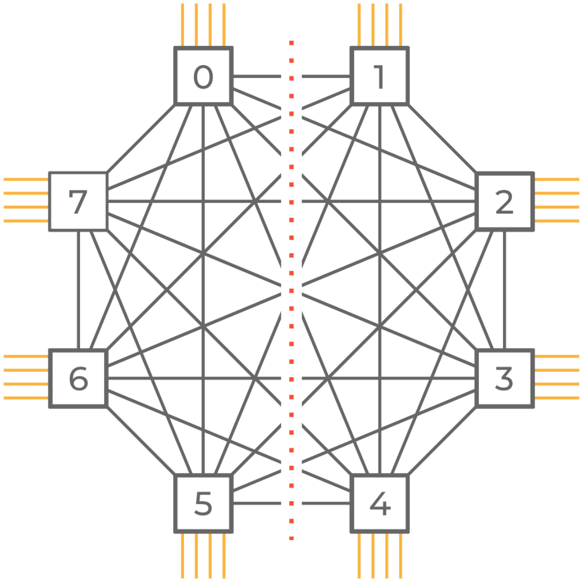


Figure 5: A depiction of the all-to-all connectivity structure within a GroqNode.

In the future, FFT will be implemented on a 64-chip rack consisting of eight GroqNode servers. In the GroqRack architecture, each node is connected to every other node in the architecture. Each node is fully connected, with all-to-all connectivity structure within a GroqNode, as shown in Figure 5. Each chip is able to directly send data to every other chip at an identical rate for all links. The outgoing yellow links on each chip represent the four connections available to connect devices across GroqNode servers in a multi-node topology.

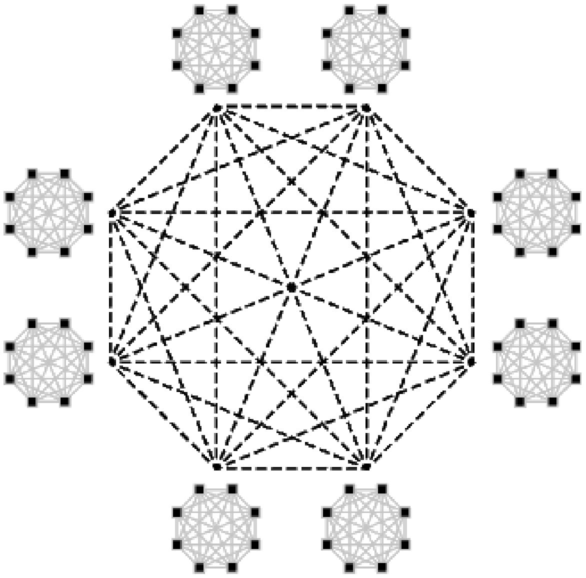


Figure 6: A depiction of how the total latency of a data transfer remains constant across different runs of the same rack-scale program using RealScale.

Despite the increase in network complexity, the total latency of the data transfer remains constant across different runs of the same rack-scale program, as seen in Figure 6. This reliability in execution time is crucial to maintain as the system scales, particularly for applications with sensitive latency requirements. The performance also scales well as the number of chips increases, as shown below in Figure 7.

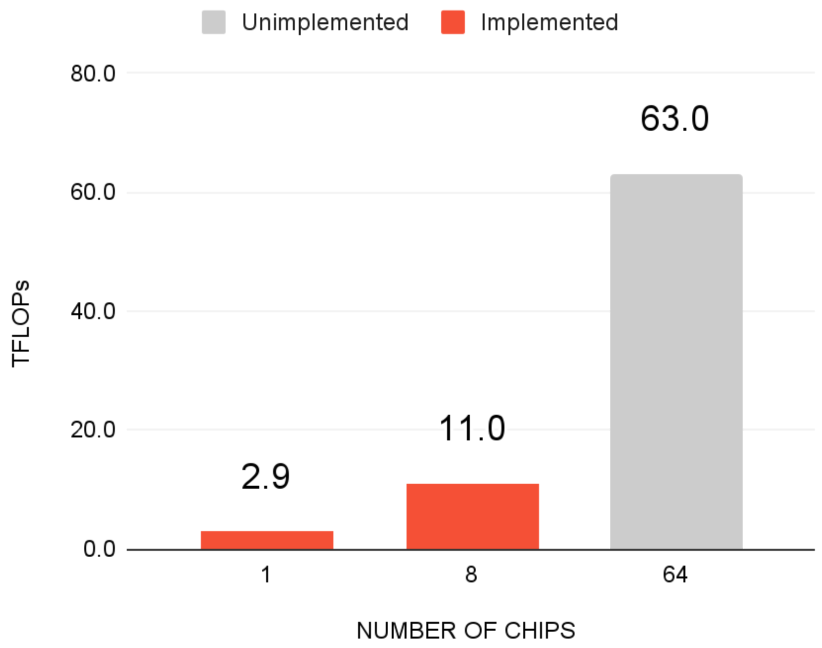


Figure 7: A depiction of how the performance scales as the number of chips increases using RealScale.

Use Cases

Use cases are key to understanding how chips scale and it's important to not only consider the performance of the connectivity solution, but its ease of implementation.

Performance

Using RealScale for large-scale model parallel workloads is already proving its benefit. For example, Groq is working with Argonne National Laboratory on connectomics, a groundbreaking research area where scientists are working to map the human brain, similar to genome sequencing work of the early 2000s. Connectomics involves building a 3D model of the brain based on data-heavy images of the brain, requiring multi-chip solutions that deliver scalability.

Given connectomics is an exascale problem in nature, the ability to accelerate and scale compute, while physically scaling models, is critical. Specifically in the case of our work with Argonne, today this involves decomposing a model and running it across eight GroqChip processors, with the future vision of rack and multi-rack scaling for the workload.

While pipelining isn't new, with RealScale and Groq determinism, at every point in time we know where each point of data will be and how long it will take to travel. Given this, there is no downtime between cycles. Not only does this help meet data transfer deadlines, it saves overhead in the chips, giving way to low latency and performance.

By leveraging RealScale to do this work versus a legacy solution, scientists are able to take advantage of consistent and predictable scaling along with Groq's high global bandwidth–all without the need for overprovisioning or hand coding, thereby accelerating time-to-innovation.

Developer Velocity

RealScale allows customers to scale up the number of chips to meet the exact performance needed. With other solutions, when users need to add more chips, developers must write more software to manage the flow of data across those extra chips. Groq delivers a multi-chip system with the user experience of a single chip, while never requiring developers to touch the code.

This capability is key for automatic speech recognition and natural language processing, sentiment analysis, and other techniques that automatically recognize, transcribe, and translate phone and video conversations in real-time. Over the past 12 months, there have been numerous headlines around increasingly large speech models such as TacoTron and Wav2vec, and the demand for increased capacity makes automating scale incredibly valuable to customers such as telecom providers, video conferencing providers, and call centers.

Interested In Learning More?

For more information on Groq technology and products, contact us at info@groq.com, follow us on Twitter @GroqInc and connect with us on LinkedIn.

Other helpful GroqDocs related to this tech doc.

- Determinism Tech Doc
- Latency Tech Doc
- Accuracy Tech Doc
- Velocity Tech Doc