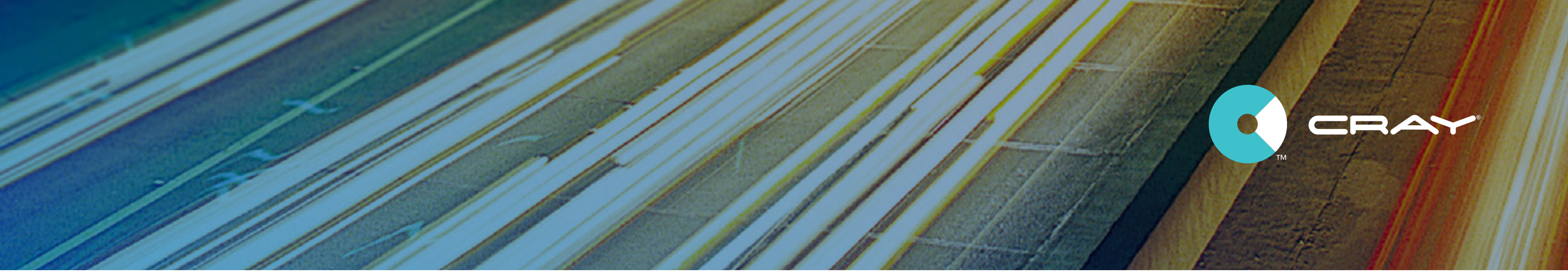
SLINGSHOT: THE INTERCONNECT FOR THE EXASCALE ERA

CRAY®



An Innovative Interconnect for the Next Generation of Supercomputers

Cray's next-generation Shasta supercomputer represents a major advancement in the flexibility and capability of Cray supercomputers. The exascale-class system will be the basis of Cray's converged architecture for simulation, analytics, AI, and data management over the next decade and beyond.

And one key element behind these era-defining capabilities is Slingshot — Cray's latest version of scalable interconnect.

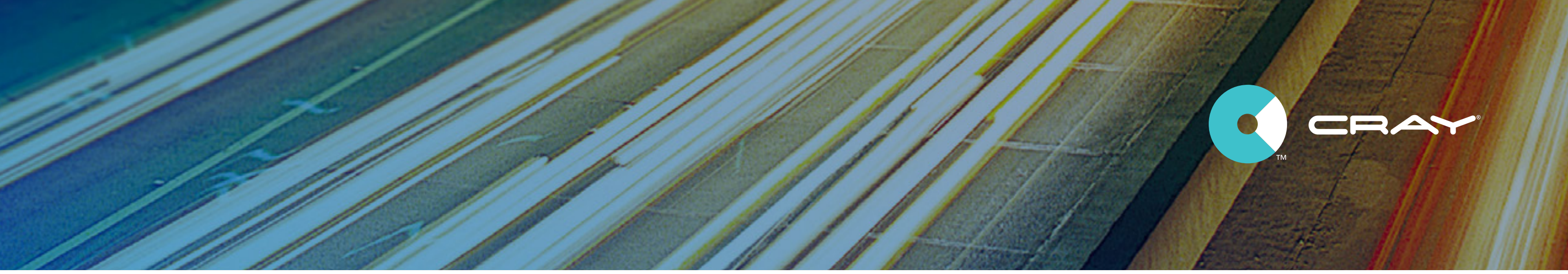
Slingshot was created to act as a suitable network backbone, one that would offer a host of features to allow the Shasta system to comfortably straddle the supercomputing and datacenter worlds. It is the most innovative, cost-optimized, and purpose-built interconnect the supercomputing market has seen to date, designed to provide unprecedented scalability and performance for the most challenging computational and data management workloads, thus aiding users who are tackling tightly coupled problems with big data.

Slingshot Raises the Performance Bar

The next network in a long line

Slingshot is Cray's eighth major generation of scalable high-performance computing network — the latest in a line of important networking milestones.

In 1992 Cray introduced the Cray T3D, the company's first massively parallel processing (MPP) system. That was followed in 1996 by the Cray T3E system, which featured the first-ever implementation of adaptive routing in an HPC network. Cray pioneered the design of high-radix switches in 2005, and Cray's YARC switch for the Cray X2 sytsem implemented 64 ports using a unique tiled architecture, enabling the creation of very low-diameter networks.



The SeaStar network ushered in Cray's XT line of MPP systems. Finally, the Aries network — shipping in the industry-leading Cray® XC™ line of supercomputers — was the first to implement the Dragonfly topology, which is now being adopted by more companies to provide highly efficient, scalable global bandwidth.

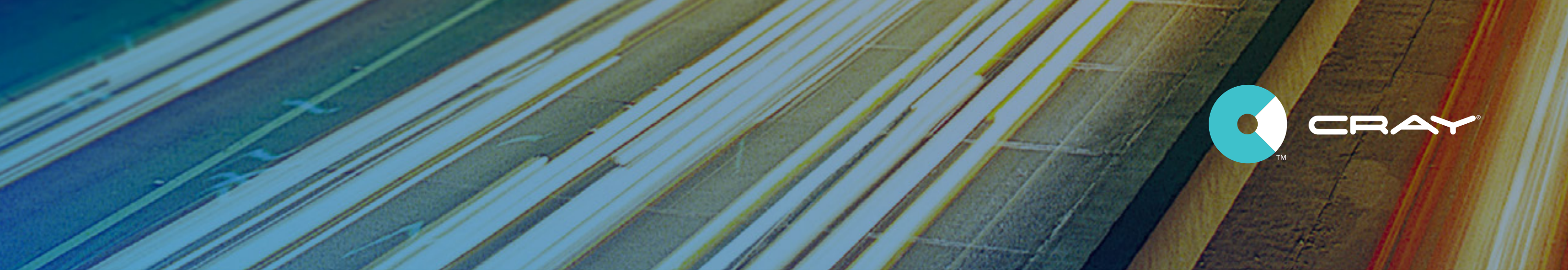
Breaking new ground again

When Cray set out to design Shasta, it started with a simple question: What does the future of supercomputing look like? In exploring that question with customers it was determined that today's HPC users increasingly want to run a mix of workflows (e.g. simulation, analytics, and AI) on one system that can handle them all simultaneously.

The era of addressing different workloads with different systems was rapidly coming to an end, and so Cray's goal quickly became to build a flexible, heterogeneous architecture that could handle all the increasingly datacentric workloads that will need to be run to answer today's largest science, technology, and big data questions.

But Cray also wanted to design a system that would "play well" with standard HPC environments and existing datacenter equipment — an achievement that would bring Shasta's power to a more diverse range of HPC users. A next-gen system of this magnitude needed a high-speed, purpose-built supercomputing interconnect to act as its backbone, and so Slingshot was born.

Slingshot has a crop of new features aimed at datacentric HPC and AI workloads. First and foremost is extremely high bandwidth: 25.6 Tb/s per switch, from 64 200 Gbs ports. Slingshot is the only interconnect currently on the market with 64 ports, which allows Cray to build very large networks of over a quarter-million endpoints with a diameter of just three network hops. Only one of those hops needs to use an expensive optical cable, so it excels at delivering high levels of overall global bandwidth with good price performance.



World-class Adaptive Routing and Congestion Control

Leveraging bandwidth to reduce congestion

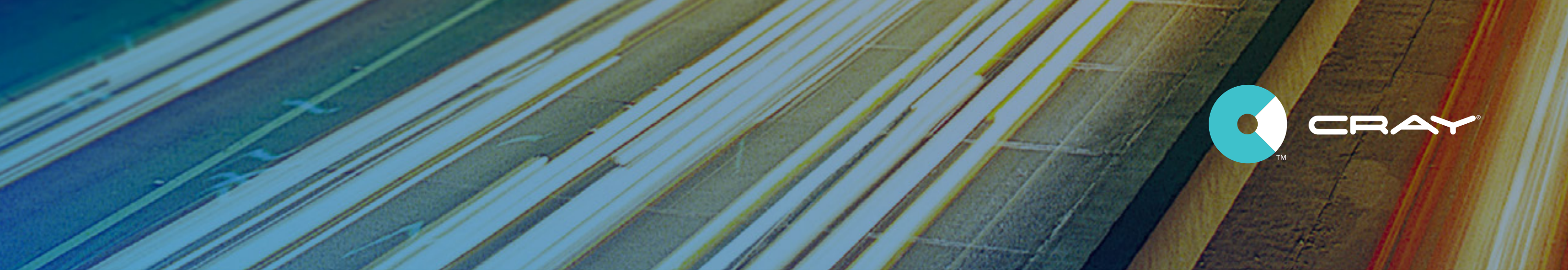
While Slingshot provides great peak performance, what's even more important is how Slingshot leverages that bandwidth. The latency that matters isn't fall-through latency on an idle network; it's latency under load, at scale, which primarily comes from queueing latency. Therefore, the key to achieving low latency is avoiding congestion and queueing in the network.

Adaptive routing

The first way Slingshot avoids congestion is with adaptive routing. This is a technique Cray has refined for over 20 years since introducing the Cray T3E system, which had the first-ever implementation of adaptive routing in an HPC network. Because each switch has a good view of the overall state of the network, it can make fast, well-informed decisions about optimal paths each data packet should follow through the architecture, as well as make automatic adjustments to route around any congestion. The network can optionally preserve packet ordering while adapting routes, or provide even higher performance by allowing each packet to adapt separately.

Congestion control

However, adaptive routing without congestion control can be problematic. As an example, imagine a traffic accident occuring right in the middle of town. Drivers may opt to take side streets to route themselves around the accident, but too many drivers doing this at once only results in congestion on the side streets. The same is true in an HPC system. When a set of nodes sends so much traffic to an endpoint that it exceeds its egress bandwidth, traffic can back up into the network over a branching tree of links targeting that endpoint, an occurrence known as tree saturation. In this case, adaptive routing can actually make matters worse by causing blocked traffic to spread to other links in a futile attempt to route around the congestion.



This is where Slingshot's most important feature comes in: an innovative congestion control mechanism that's extremely responsive, requires no tuning, and is stable across a wide range of dynamic HPC workloads.

First, the mechanism instructs any source(s) trying to send more data into the network than can be delivered to "back off" to avoid wasting buffer space in the network. More importantly, the backpressure affects *only* the offending data sources, not the rest of the "innocent" traffic (e.g. data headed for an endpoint that can accept it) that may be sharing some links with the congestion-causing traffic. As the congestion clears, the participating sources are allowed to ramp up their transmission, at just the right bandwidth to keep the bottleneck link(s) fully utilized but without causing bubbles in the bandwidth.

Best of all, this happens automatically through the hardware with zero software setup, using sophisticated machinery that tracks everything going on in the network on a per-packet basis.

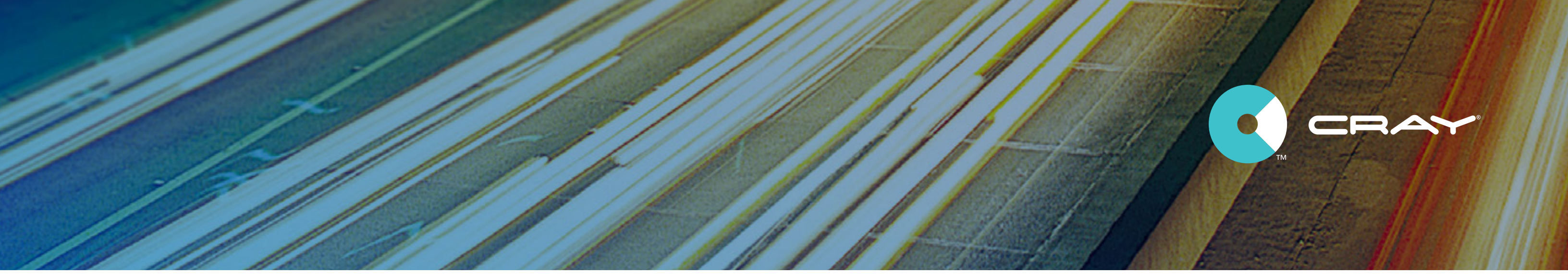
Isolated workloads

The net result is that Slingshot provides highly effective performance isolation between workloads. No longer does a poorly written application cause congestion that interferes with other workloads on the system. This means that latency variation is dramatically reduced in the network, which enables the consistently low network latency that is key to making latency-sensitive and synchronization-heavy applications perform well. It's a capability sure to have a dramatic impact on the performance of the increasingly heterogeneous and datacentric workloads expected to run on future systems.

A Better Datacenter Citizen

Ethernet compatibility enhances interoperability

To enhance interoperability with storage architectures and datacenters, Cray made Slingshot Ethernet compatible. This means Slingshot switches can connect directly to third-party Ethernet-based storage devices as well as to datacenter Ethernet networks. Applications running on Shasta compute nodes



can directly exchange IP/Ethernet traffic with the outside world, making it easier and more efficient to ingest data from external sources — an increasingly important consideration in this highly networked and data-driven world.

However, standard Ethernet can have high overheads and be poorly suited for HPC workloads. To overcome these issues, Cray developed an optimized, high-performance Ethernet protocol that has smaller headers, support for smaller packet sizes, credit-based flow control, reliable hardware delivery, and a full suite of HPC synchronization primitives. Slingshot uses this optimized protocol for internal communication, but can also intermix standard Ethernet traffic on all ports at packet-level granularity. This best-of-both-worlds approach allows the Shasta system to comfortably straddle the supercomputing and datacenter worlds.

Decoupling from software

Slingshot is also innovative in that, unlike previous interconnects, it has been decoupled from software. This is helpful for users who want a cluster-style cabinet, because Slingshot can be used in this environment with commodity software, allowing users to get the supercomputing performance they need without having to rip and replace their standard software stack. The physical packaging of Shasta further helps it fit into existing or standard HPC environments. Shasta is the first Cray architecture to support multiple cabinet types, including a standard 19-inch datacenter rack, providing easy flexibility and upgradeability over time.