Creating Future Network Architectures

Robert Martinez, Tim Merrigan, Denny Rice, Todd Bowman, CTN-5

Supercomputers were once built with tightly coupled CPUs (processors), data storage, and visualization hardware that communicated using vendor-specific interconnect hardware. Today's high-performance computing (HPC) platforms are architected with thousands of processors organized into clusters that communicate across multiple high-performance networks. The current HPC hardware paradigm is built on large numbers of commodity servers containing multiple processors operating in parallel, each processor operating on a small portion of the overall calculation.

Communications between hundreds or thousands of processors to achieve this parallelism is accomplished on multiple specialized networks optimized for their functions. Communications between processors within clusters to provide intraprocess message passing are interconnect networks that provide low-latency and high-bandwidth performance between processors.

The message-passing networks or interconnect networks provide connections for internal communications between thousands of processors within the cluster as illustrated in Fig. 1.

Interconnect networks are primarily based on niche technologies designed specifically for cluster environments to provide low latency and high data rates, 10 and 20 Gigabits/s. Some interconnects are standards-based such as InfiniBand and Ethernet, while others are based on proprietary implementations such as Myrinet or Quadrics. CTN-5 is developing monitoring and troubleshooting tools to provide technicians with the ability to resolve InfiniBand issues.

A network drawing of an Interconnect network is shown in Fig. 1. Applications that are run or executed on HPC clusters produce data representing computational results that must be saved. The input/output (I/O) network provides high-performance access to the parallel storage file systems. The system administration of HPC clusters are performed through a management network built on commodity Ethernet.

The LANL Network Engineering group, CTN-5, provides network engineering design, installations, and operations support for the HPC clusters at Los Alamos National Laboratory. Specifically, CTN-5 and groups in HPC division work closely together in the parallel scalable back bone (PaScalBB) project to develop

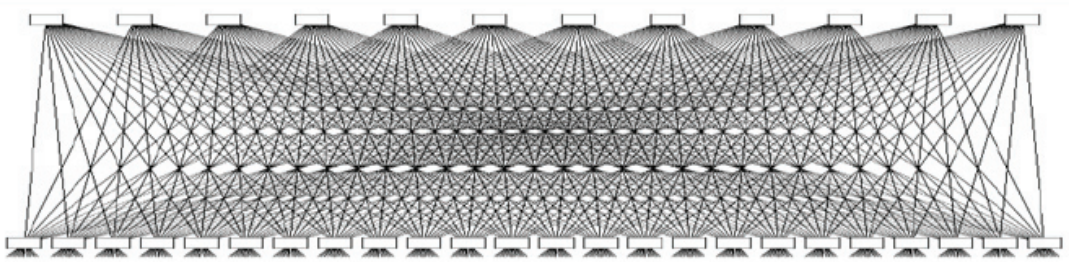

Fig. 1.
The Roadrunner cluster interconnect, used for message passing and within the cluster to move data to the I/O front ends or I/O nodes. Interconnects are built on the niche technologies like InfiniBand that will be used with Roadrunner.

future network architectures that will support the progressively higher performance clusters introduced into the LANL environment. CTN-5 is assuming support responsibilities for the various interconnect networks. During FY07, CTN-5 will be scaling up support for the InfiniBand interconnect technology used in the recently acquired clusters. CTN-5 staff provide 5x10 prime time and 7x24 on-call operations support.

The cluster I/O networks are based on the Gigabit and 10 Gigabit Ethernet technologies that CTN-5 supports on the Laboratory production networks. The Roadrunner cluster I/O network shown in Fig. 2 illustrates the relationship of I/O networks to the cluster and storage file systems. CTN-5 provides depth of knowledge in the design, installation, monitoring, and problem resolution for Ethernet-based networks.

Conclusion

The current supercomputing paradigm is highly dependent on scalable high-performance network technologies and architectures. Future supercomputing platforms will require emerging network technologies to provide the required scalable performance.

The challenge for HPC/CTN network engineers will be to apply these emerging technologies to provide the continually increasing computing demand by LANL researchers.

For more information contact Robert Martinez at ram@lanl.gov.

Funding Acknowledgements

This research was supported by the NNSA tri-Lab Advanced Simulation and Computing Program.

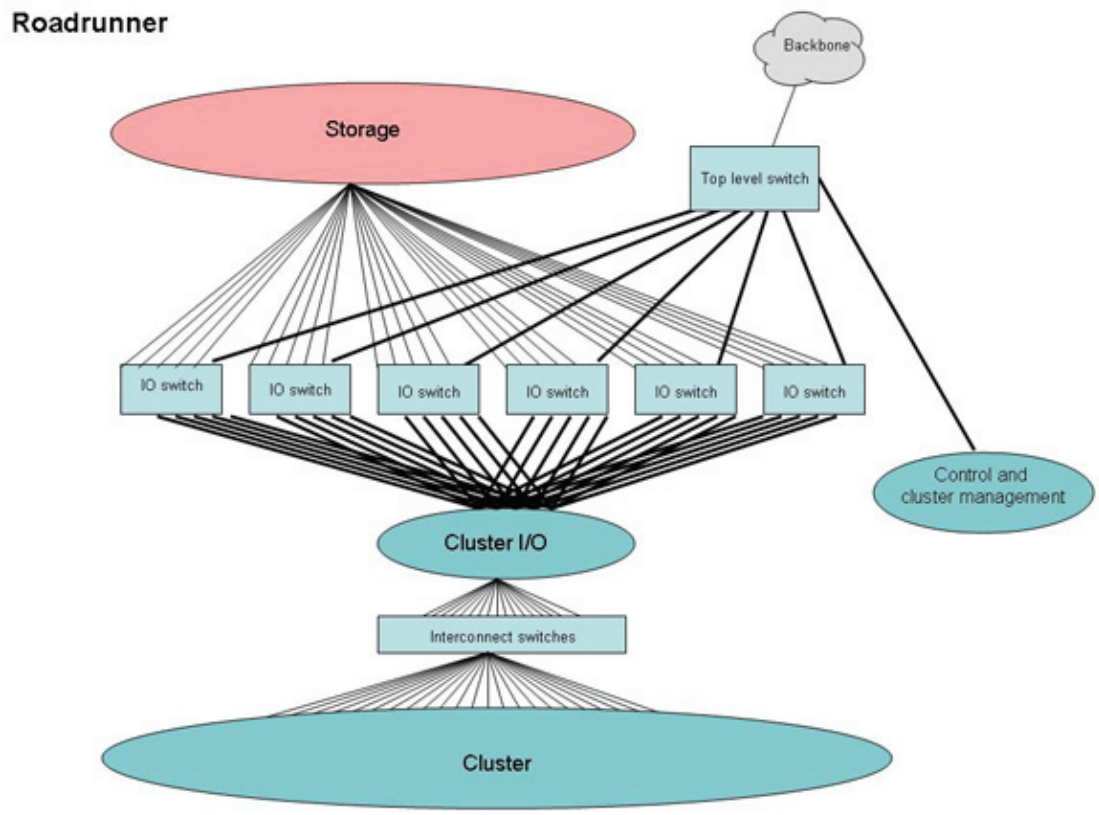

Fig. 2.
Roadrunner I/O network diagram illustrates the relationship between the computational cluster nodes, the interconnect, storage, and I/O switching fabric.