DOE/MICS/SciDAC Network Research Program

Title: Enabling Supernova Computations by Integrated Transport and Provisioning Methods Optimized for Dedicated Channels

PI: Nageswara S. V. Rao
PI Institution: Oak Ridge National Laboratory

Project Website: http://www.csm.ornl.gov/netviz

Abstract:
Terascale Supernova Initiative requires unprecedented network capabilities to support collaborative interactive monitoring and steering of computations on supercomputers from distributed sites. We implemented dedicated bandwidth channels between ORNL Cray X1(E) supercomputer and NSCU cluster over CHEETA-UltraScience Net infrastructure. We tuned and configured transport protocols to achieve high throughputs over these dedicated connections and identified performance bottlenecks. We developed a scheme that optimally maps a visualization pipeline onto a network to achieve high end-to-end performance. This system overcomes the sub-optimal performance of the conventional visualization methods that employ monolithic client-server configuration, and also provides computational monitoring and steering capability. These components together represent the building blocks that may be integrated to effectively carry out a TSI computation on a supercomputer while being monitored, visualized and steered by a group of geographically dispersed domain experts.

Introduction

Terascale Supernova Initiative (TSI) is a multidisciplinary collaboration of Oak Ridge National Laboratory (ORNL) and several universities to develop models for core collapse supernova and enabling technologies in radiation transport, radiation hydrodynamics, nuclear structure, linear systems and eigenvalue solution, and collaborative visualization. In general TSI applications involve a wide spectrum of tasks that range from cooperative remote visualization of massive archival data through the distribution of large amounts of simulation data, to the interactive evolution of a supernova computation through computational monitoring and steering. In the general case, the data must be rendered and presented on-line to various participant sites with different end-devices ranging from visualization caves through high-end workstations to personal desktops as shown in Figure 1. The control of this computation and visualization must be coordinated among the end users who might be looking at different data subsets. Another important aspect is the sizes of tera-peta byte datasets generated by TSI computations. Total aggregate data rates for data transfers are of the order of several tens of Gbps.

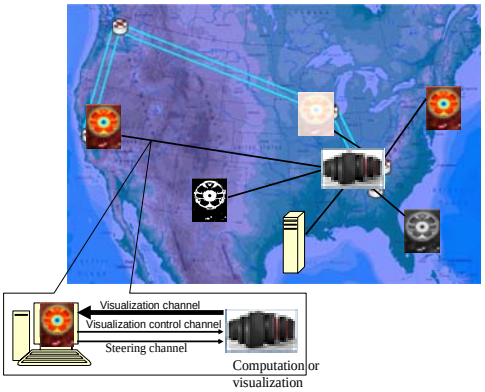

Figure 1. Supernova computation monitored and steered by distributed domain experts.

Currently TSI scientists utilize ORNL supercomputers for computations, and the computed datasets are archived locally. Also,

no interactive visualizations of large datasets across wide-area networks are currently carried out primarily due to the lack of appropriate network and visualization capabilities. While these taskss are motivated primarily by TSI, the underlying technologies are applicable to much a wider class of large-scale science applications.

Dedicated Bandwidth Channels

It is generally believed that networking demands of large-scale science can be effectively addressed by providing on-demand dedicated channels of the required bandwidths directly to end users or applications. The UltraScience Net (USN) provides on-demand dedicated channels: (a) 10Gbps channels for large data transfers, and (b) high-precision channels for fine control operations. Circuit-switched High-speed End-to-End Transport ArcHitecture (CHEETAH) is an NSF project to develop network infrastructure for provisioning dedicated bandwidth channels needed for eScience applications, particularly TSI. The CHEETAH channels are set up using off-the-shelf equipment, such as SONET switches using Generalized Multiple Protocol Label Switching (GMPLS) for dynamic circuit setup/release.

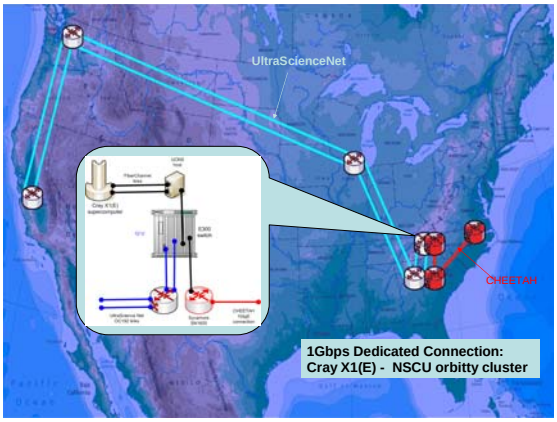

Figure 2. Dedicated 1Gbps channel between ORNL Cray X1 and NCSU cluster.

Supercomputers present networking challenges that are not addressed by the networking community mainly due the complexity of data and execution paths inside those machines and the needed effective and secure interconnections. Data from a Cray node traverses System Port Channel (SPC) channel and then transits to FiberChannel (FC) connection to CNS (Cray Network Subsystem). Then CNS converts FC frames to Ethernet LAN segments and sends them onto GigE NIC, which is connected to wide-area network. Since the current capacity of Cray X1(E)'s NIC is limited to 1Gbps, we developed a new interconnection configuration. We utilized USN-CNS (UCNS) which is a dual Opteron host containing several FC (Emulex 9802DC) cards each with two 2Gbps FC ports, and a Chelsio 10GigE NIC. NCSU cluster is currently connected over a 1Gbps dedicated channels provisioned by CHEETAH-USN. The default TCP over this connection achieved throughputs of the order 5 Mbps. Then bbcp protocol adapted for Cray X1 achieved throughputs in the range 20-30 Mbps depending on the traffic condition. Hurricane protocol tuned for this connection consistently achieved throughputs of the order 400Mps. We diagnosed a performance bottleneck within Cray X1E OS nodes, which is being currently addressed by Cray.

Visualization Mappings

In general, a remote visualization system forms a pipeline consisting of a server at one end holding the data set, and a client at the other end providing rendering and display. In between, zero or more hosts perform a variety of intermediate processing and/or caching and prefetching operations. A wide area network typically connects all the participating nodes. We developed an approach to dynamically decompose and map a visualization pipeline onto wide-area network nodes for achieving fast interactions between users and applications in a distributed remote visualization environment. This scheme is implemented using modules that perform various visualization and networking subtasks to enable the selection and aggregation of nodes with disparate capabilities as well as connections with varying bandwidths. In addition, we developed first versions of server and client modules that enable certain variables of the computation to be remotely monitored, rendered and modified on the fly to steer the computation.