Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Center for Adaptive Supercomputing Software-Multithreaded Architectures (CASS-MT)

Lightweight User Communication Environment (LUCE)

At a glance

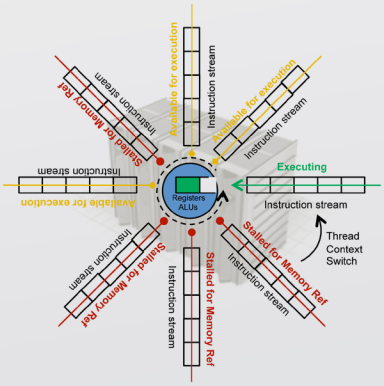
The goal of this research is to advance the current state of the art in data intensive supercomputing by developing an efficient communication environment that enables easy integration of heterogeneous high performance systems in a hybrid computing environment. The initial focus is on supporting integrated applications running in an environment consisting of a Cray XMT, a Netezza TwinFin Data Warehouse, and a commodity MPI cluster. An initial LUCE API that supports basic get and put operations has recently been released.

What we do

Different stages of running applications may require distinct computing resources – not just to maximize performance and efficiency, but to enable computations that would not be possible on traditional high performance computing clusters. By integrating novel systems with traditional computing clusters, new scientific and analytical results can be achieved. The Cray XMT and Netezza systems are examples novel architectures that enable transformational applications for deep analytics.

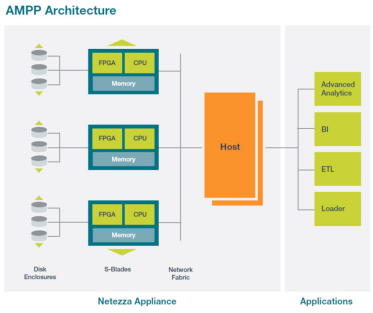
1. Cray XMT

The Cray XMT is a shared-memory multithreaded architecture that is designed to support irregular applications (i.e. applications such as large-scale graph analytics that cannot exploit locality or deep cache structures seen in traditional HPC architectures). The system is divided into custom-designed multithreaded threadstorm "compute" nodes and dual-socket Opteron AMD "service/IO" nodes. The nodes are connected via a Cray Seastar-2.2 high speed interconnect. Each Threadstorm processor supports 128 concurrent HW threads that can context switch in a single cycle thus hiding latency of memory accesses. The memory among Threadstorm processors is shared; however communication between the AMD Opteron service nodes and the Threadstorm nodes is accomplished through remote procedure calls utilizing RDMA. A key goal of this research activity is to develop efficient communication interfaces between the AMD Opteron nodes and the Threadstorm nodes within the XMT as well as optimize the communication between the Opteron nodes and external system nodes such as the Netezza TwinFin or an MPI based cluster.



2. Netezza TwinFin

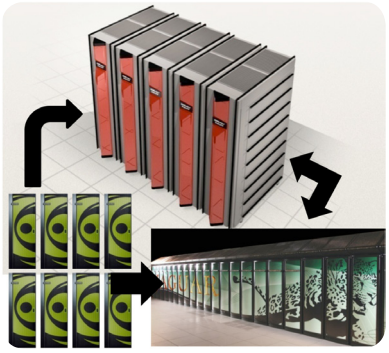
The Netezza TwinFin is a data warehouse appliance designed for complex analytics on massive amounts of data. It integrates a database, processing capability, and storage into a single unit. It employs a unique hardware architecture, called a snippet blade that fuses multi-core Nehalem CPUs with FPGAs and gigabytes of RAM connected directly to disk via a high speed interconnect so that data is streamed directly to the snippet blades. The data is highly compressed on disk and uncompressed in real time at line speed, thus increasing amount of data that can reside on the machine. This design is ideal for supporting complex queries against tens to hundreds of terabytes to extract complex data structures such as large scale graphs in the hundreds of gigabytes to a few terabytes that can be sent to systems such as the Cray XMT for deep analytic exploration.n.



How we do it

There are three models for hybrid applications running on the Cray XMT: 1) applications running on the service nodes of the XMT that communicate with both the compute nodes and external systems; 2) applications running on the compute nodes that communicate with the service nodes and through the service nodes to external systems; and 3) applications are running external to the XMT and remotely call routines on the XMT. This research activity is focused on optimizing the hybrid communication interfaces in the first and third models Data is passed

U.S. DEPARTMENT OF ENERGY



through the systems using combinations of shared memory, sockets, and remote procedure calls in a way that is transparent to the user.

An initial LUCE API to support basic get and put operations has recently been released. The LUCE API is designed to be similar to the Aggregate Remote Memory Copy Interface (ARMCI) so that ARMCI developers will find it easy to adopt LUCE.

Applications

Current Applications

LUCE has been used in two demonstrations. In the first, an application runs on the service nodes of the Cray XMT and pulling subsets of data from a multi-TB database on a Netezza TwinFin containing billions of enterprise network flow records—approximately one month of data.

From this data, a graph is extracted by the service nodes where the nodes of the graph are IP addresses and the edges represent communication between two systems (IP addresses). The graph is passed to the Cray XMT compute nodes and a a census of the possible communication patterns between any three nodes in the graph is calculated. The entire database is streamed through the compute nodes in this fashion to obtain a dynamic view of how triads evolve in the network over time. The other demonstration is an MPI application running on a commodity cluster that blocks to send data in parallel to the Cray XMT service nodes which just pass through the data in parallel to the XMT service nodes. The service nodes assemble the distributed data into a single data structure, perform a calculation on that data structure and then send results in parallel back through the service nodes to the blocked MPI application.

Future Applications

Future applications include integration with real-time analysis of the power grid and analysis and decision support for cyber security. For this application, LUCE will be enhanced to support a streaming computing model for these applications.

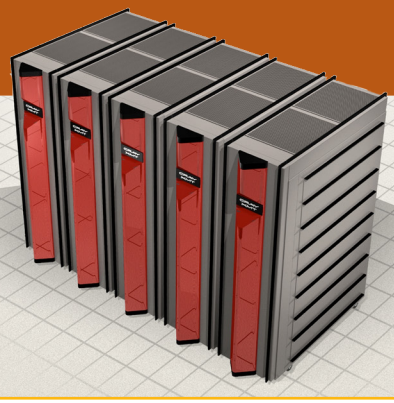
CASS-MT is dedicated to research on systems software, programming environments, and applications in a High-Performance Computing (HPC) multithreaded architecture environment.

We offer the only Open Science Cray XMT system, a one-of-a-kind supercomputer consisting of 128 multithreaded processors, 1 TB RAM, and a 7.7 TB Lustre parallel filesystem.

The Cray XMT supercomputer has the potential to substantially accelerate data analysis and predictive analytics beyond the limitations of traditional computing. Multithreaded processors allow multiple, simultaneous processing, helping researchers find solutions to the world's most complex challenges faster. The XMT can process irregular, data-intensive applications that have random memory access patterns. Unlike many applications where data delivery is dependent on memory speed, the Cray XMT's multi-threaded architecture tolerates memory access latencies by switching context between multiple threads that work continuously, overlapping the memory latency and preventing the processor from being held up while it waits for data to arrive.

The multithreaded technology powering our Cray XMT is ideally suited to perform pattern matching, scenario development, behavioral prediction, anomaly identification, and graph analysis.

Try it for yourself. We seek to create collaborations and provide expertise for porting and optimizing applications. The opportunity to use our Cray XMT system is available to internal and external research partners.



John Feo,
CASS-MT Director
(509) 375-3768
John.feo@pnl.gov
cass-mt.pnl.gov/

John R. Johnson
Task Lead
High-Performance Computing
509-375-2651
john.johnson@pnl.gov

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*