LS-DYNA® Performance Improvements with Multi-Rail MPI on SGI® Altix® ICE Clusters

Olivier Schreiber, Michael Raymond, Srinivas Kodiyalam
SGI
1140 East Arques Avenue, MS: 275
Sunnyvale, California 94085
(oliviers, mraymond, skodiyal@sgi.com)

Abstract

Multi-Rail networks can improve MPI communication performance by distributing the communication traffic to multiple independent networks (rails). Messages are divided into several chunks and sent out simultaneously using multiple rails. With the dual plane network topology of SGI Altix ICE clusters, MPI communication can hence utilize both the InfiniBand rails, including, ib0 and ib1 fabrics. The performance gains achievable with LS-DYNA for complex crashworthiness simulations through the use of MPT dual-rail over MPT singe-rail on an Altix ICE system are indeed significant.

LS-DYNA MPP Application

LS-DYNA is a general purpose, transient dynamic, non-linear, explicit (with implicit capability), finite element analysis software with the following core competencies:

* Highly nonlinear:
  - o Changing boundary conditions with time such as contacts between parts
  - o Large deformations
  - o Nonlinear (non elastic) materials
* Transient dynamics
* Important inertial forces
* Finite element analysis
* Explicit time integration

John Hallquist originally wrote DYNA-3D for Lawrence Livermore National Laboratory, subsequently released in public domain. LSTC develops LS-DYNA since 1987. LS-DYNA is used in the following industries:

* Automotive
* Aerospace
* Manufacturing and
* Bioengineering
* Consumer

and disciplines:

* Crash
* Metal forming
* Blade containment
* Bird strikes
* Drop testing

* Plastic, glass forming

LS-DYNA sequentially goes through the following phases:

1. Initialization:
    1. reading input file[s],
    2. allocating memory
    3. initializing variables
    4. domain decomposition
2. Element-processing
3. Contacts
4. Rigid bodies

The first MPP capabilities were added in 1993.

SGI Message Passing Toolkit (MPT)

The SGI® Message Passing Toolkit (MPT) provides versions of industry-standard message-passing libraries optimized for SGI® computer systems running the Linux® operating system. These high-performance libraries permit application developers to use standard, portable interfaces for developing applications while obtaining the best possible communications performance. Multi-Rail is a similar concept to Dual-Plane where it means two separate network fabrics. Dual-Plane is the term used when referring to SGI NUMAlink shared memory interconnects. The hardware takes care of managing the traffic between two networks and the MPI (SGI MPT) does not have to do additional work.

Multi-Rail is used for InfiniBand networks. For multi-rail, the MPI library has to manage the network traffic between the available separate network fabrics. The network interface cards do not take care of it. For InfiniBand on SGI Altix platforms MPT includes functionality to spread the communications traffic across the rails. For short and medium sized messages between any pair of processes, MPT routes these messages across through a single rail. These routes though are spread across the rails so that a process will have roughly half of its routes on one rail and the other routes on another rail. For large messages MPT will split the message in half and send each half on a different rail. The distribution of routes for shorter messages spreads the communications load across the available pathways. The splitting of larger messages can almost double the bandwidth.

SGI MPT 1.18 added Multi-Rail InfiniBand support as part of SGI Propack 5 SP 4. MPT 1.19 added splitting of large messages. MPT 1.20 adds further refinements to the intelligent scheduling of InfiniBand traffic.

SGI Altix ICE cluster

The SGI Altix ICE system is a blade-based, scalable, high-density compute server. The building block of this cluster is the Individual Rack Unit (IRU) that provides power, cooling, system control, and network fabric for 16 compute blades. Each compute blade supports up to two quad-

core or dual-core Xeon processor sockets and eight fully-buffered, DDR2 memory DIMMs. One high rack, consisting of up to four IRUs, supports a maximum of 512 processor cores and 2 TB of memory.

The SGI Altix ICE system uses a fully integrated set of 4x DDR InfiniBand switches that are internal to each of its IRUs. In an SGI Altix ICE 8200 cluster, these switches will be interconnected to form a dual-plane, bristled multi-dimensional hypercube topology. One plane is typically dedicated for inter-node communication through the use of the Message Passing Interface (MPI). The other IB plane is typically dedicated for balanced, high-perform I/O.

With an SGI Altix ICE 8200EX cluster, there are four InfiniBand switches integrated into each IRU. Full "dual rail" or "dual plane" MPI is supported on this cluster system with two InfiniBand switch blades are dedicated to each plane. Depending on the size of the cluster, remaining "out-facing" IB ports can be used for high-performance I/O to access a high performance storage sub-system.

Multi-Rail MPI performance benefits for car crash simulations

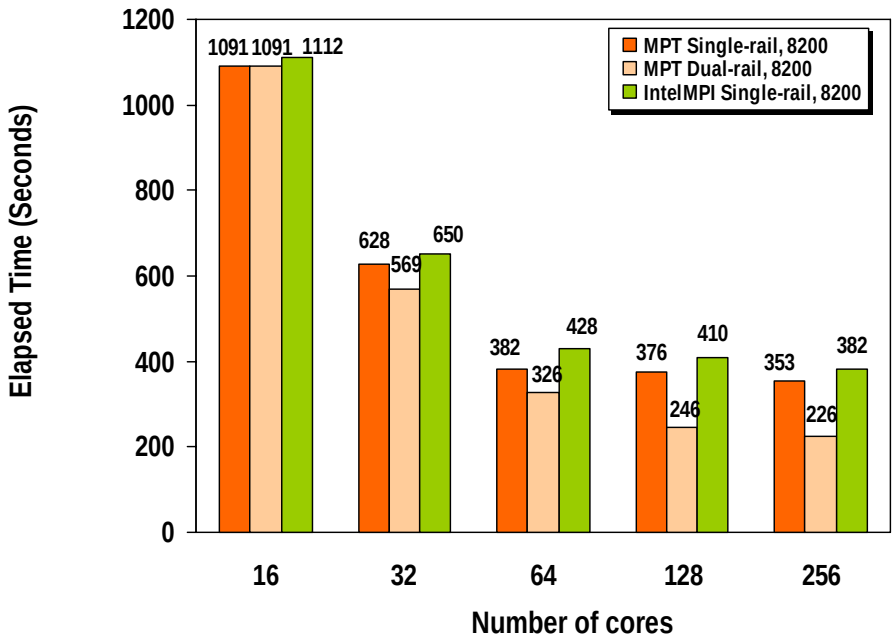
Two industry standard car crash models, neon_refined_revised and 3_vehicle_collison (Source: www.topcrunch.org) are used for evaluating Multi-Rail MPI performance improvements with LS-DYNA crashworthiness simulation.

Case 1: neon_refined_revised model:

This is a frontal crash simulation with initial speed at 31.5 miles/hour. The revised version of the model was developed by LSTC and the original model by NCAC. The model has around 535K elements with a simulation time duration of 30msec.

The crashworthiness simulation was performed using LS-DYNA MPP v971R3.1 on an Altix ICE 8200 cluster with Intel Xeon 5365 3.0GHz quad-core Clovertown processors, 2MB cache/core, 1333MHz FSB, 2GB memory/core and with a hypercube topology that is 2:1 blocking.

The below figure provides a comparison of LS-DYNA performance on Altix ICE (single-rail MPT, dual-rail MPT, and single-rail IntelMPI).

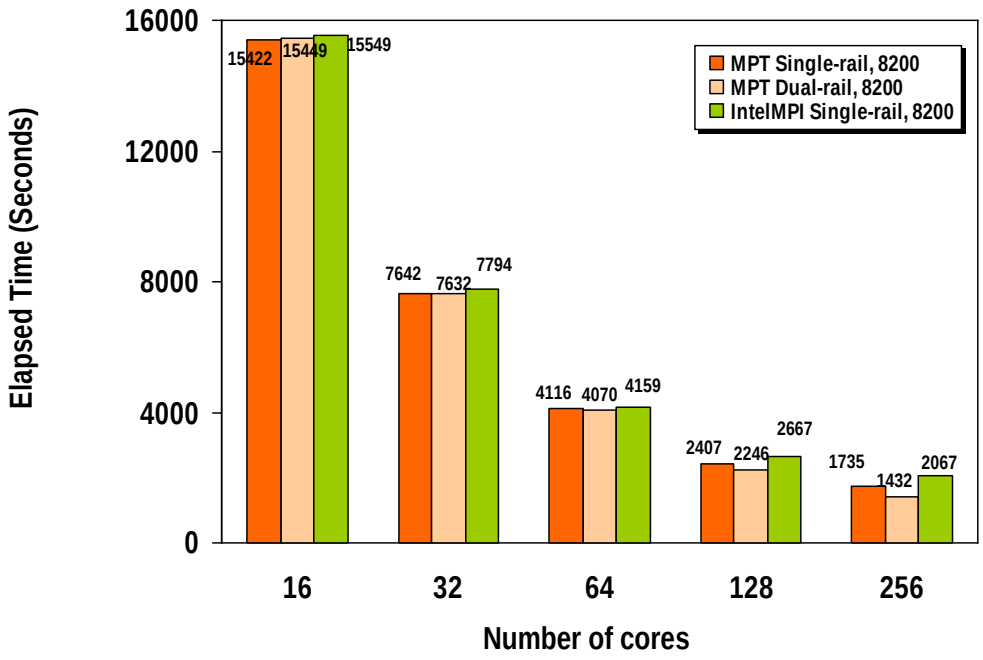


Case 2: 3_vehicle_collision model:

In this simulation, a van crashes into the rear of a compact car, which, in turn, crashes into a midsize car. The model was created by NCAC and includes 791K elements with a simulation time of 150 msec.

The crashworthiness simulation was performed using LS-DYNA MPP v971R3.1 on an Altix ICE 8200 cluster with Intel Xeon 5365 3.0GHz quad-core Clovertown processors, 2MB cache/core, 1333MHz FSB, 2GB memory/core and with a hypercube topology that is 2:1 blocking.

The below figure provides a comparison of LS-DYNA performance on Altix ICE (single-rail MPT, dual-rail MPT, and single-rail IntelMPI).



Discussion of results

On an Altix ICE 8200 system with hypercube topology (2:1 blocking) the performance gains with dual-rail MPT are significant:

- 35% reduction in solution time at 128 cores for the frontal car crash simulation; and,
- 17% reduction in solution time at 256 cores for a more complex 3 vehicle collision simulation.

The performance gain is likely to be even more with a fat-tree, non-blocking topology.

While LS-DYNA performance benefits from Multi-Rail MPI, it may not be suited for all applications. Especially, if an application does a lot of IO and MPI communications, it may be best to keep them separate by running single-rail style on the dual-plane system instead of striping the MPI traffic on both ib0 and ib1 fabrics.