



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Purple L1 Milestone Review Panel - MPI

Terry Jones

December 11, 2006

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Purple L1 Milestone Review Panel

MPI

Terry Jones, 12/1/2006

Deliverable

The MPI deliverables for the Purple system were designed to ensure that applications which depend on MPI benefit from a robust, functionally complete, and high performance MPI. We specifically targeted three categories of MPI validation: robustness, functionally complete, and high performance. These three categories were intended to address the following needs:

- Robustness: It doesn't matter how fast you arrive at an answer if the answer is wrong. Since any new flagship machine for the DOE complex will have pushed the envelope for scale, tests were designed to investigate behavior at scale.
- Functionally complete: MPI functionality concerns usually deal more with coverage than concerns over correctness (no doubt a result of the maturity of the specification). We validated the desired interfaces are present and their operation proceeds as expected.
- High performance: For a software stack to be considered "high performance" it must efficiently deliver the capabilities of the underlying hardware and provide levels of performance in keeping with the leading machines of the time.

Criteria

LLNL established separate items for each of the three component areas of robustness, functionally complete, and high performance. Included in *functionality* was a demonstration of scaling to 8192 tasks, a demonstration of scalable memory usage, acceptable documentation, and full MPI-2 minus dynamic tasking. The *robustness* element for MPI was addressed separately via full MPI application MTBF in the Synthetic Workload (SWL).

Results

In November of 2005, a series of tests were performed on Purple in which all MPI performance and functionality Statement of Work items were passed, save one item (discussed below). The following table outlines the performance measurements:

November 2005 Performance Highlights		
Description	Target	Actual
1tpn Interconnect link bandwidth (any number sources to sinks via striping)	4.46 GB/sec bi (57% of 8 GB/sec) 3.23 GB/sec uni (84% of 4 GB/sec)	5.68 GB/sec bi-directional 3.31 GB/sec uni-directional
8tpn Interconnect link bandwidth (8 sources to 8 sinks)	4.8 GB/sec bi (60% of 8 GB/sec) 3.2 GB/sec uni (85% of 4 GB/sec)	5.85 GB/sec bi-directional 3.77 GB/sec uni-directional
1tpn Interconnect link latency (1 source to 1 sink)	5.5 us ping-poning (msg + ack)	5.01 us
8tpn Interconnect link latency (8 sources to 8 sinks)	8.0 us ping-poning (msg + ack)	6.00 us
Collective Operation Scaling (10,000 allreduces with "crunch cycle")	Verify scales as Log2(ntask).	passed with co-scheduler

Table 1.

The final performance metric, bi-section bandwidth, was achieved in January 2006 (see Table 2). LLNL and IBM undertook an effort to understand the extent of impact for various levels of shortfall on ASC applications while other efforts continued in parallel to bring up the metric up to the target of 45% efficiency for worse case pairings.

January 2006 Performance Highlights		
Description	Target	Actual
Aggregate machine bi-section BW, worse case pairing (all communication across 3 rd stage)	45% efficient (NumNodes * .45 * 8 GB/sec)	47.19% random-random ~70% typical 98.8% nearest neighbor

Table 2.

By using environment tuning, we were able to achieve 47% efficiency for worse case (see Figure 1). Most pairings actually perform much higher.

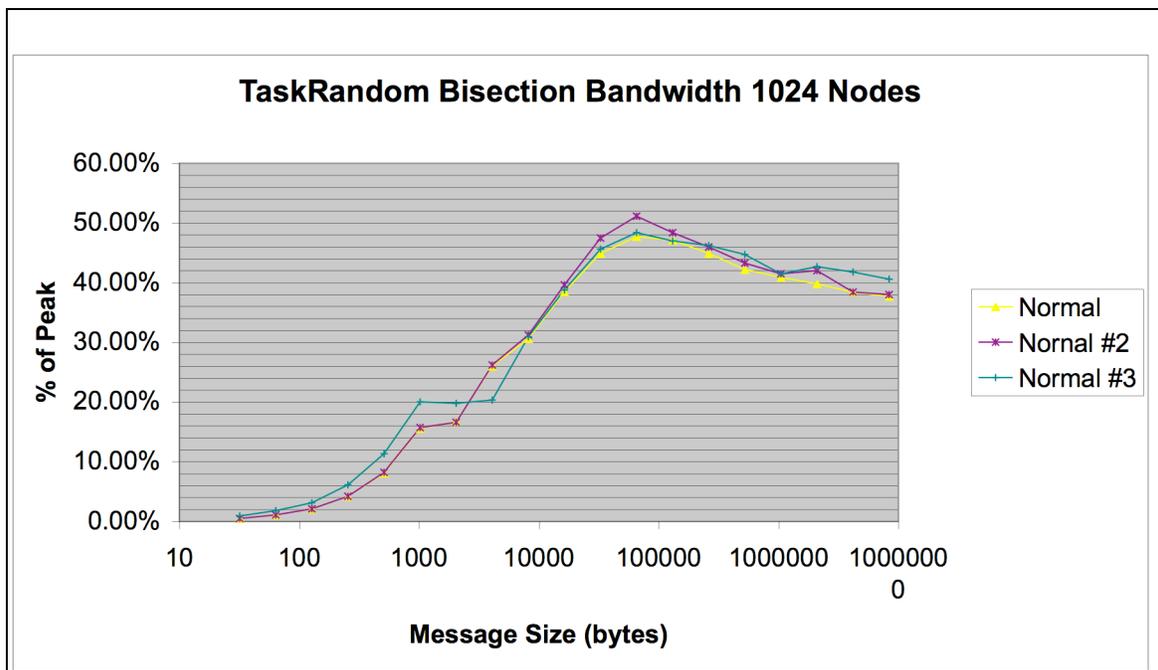


Figure 1.

The Robustness category of MPI was demonstrated by the Synthetic Workload application load Stability Test, or SWL-ST. The test results that were documented and archived will be included in the L1 Milestone completion documentation.

Conclusion

All MPI related Statement of Work (SOW) target performance objectives have been met. Both MPI-only and Hybrid-MPI codes have successfully met scaling expectations on Purple (including ale3d, yf3d, and other classified applications).