A Network-Failure-Tolerant Message-Passing System for Terascale Clusters

Richard L. Graham

Advanced Computing

Los Alamos National Laboratory

LA-MPI team

- David Daniel
- Nehal Desai
- Rich Graham
- Dean Risinger
- Mitch Sukalski

Lampi-support@lanl.gov

Past Contributors

- Ron Minnich
- Sung-Eun Choi
- Craig Rasmussen
- Ling-Ling Chen
- MaryDell Nochumson
- Steve Karmesin
- Peter Beckman

LA-MPI design goals

- Message passing support for terascale clusters
- Fault tolerant (reliable, resilient)
- High performance
- Thread Safe
- Support widely used message passing API (MPI)
- Multi-platform and supportable (Open source)

LA-MPI architecture

- Two component design
  * Run time job control
    - job startup
    - standard I/O
    - job monitoring
  * Message passing library
    - resource management
    - message management

Platforms Supported

OS's

* Linux (i686, alpha)
* TRU64
* 32 and 64 bit versions
* IRIX

"Interconnects"

* Shared Memory
* UDP
* ELAN3
* HIPPI-800

Zero Byte Half Round Trip Latency (uSec)

Platform	LA-MPI	MPT	MPICH
O2K (shared Mem)	7.0	6.5	19.9
O2K (HIPPI-800)	155.3	143.5	N/A
O2K (IP)	526.7	525.6	586.0
i686 (shared Mem)	2.3	N/A	23.5
I686 (IP)	132.8	N/A	123.5
ES45 (1G shared Mem)	2.3	N/A	5.5
ES45 (1G ELAN-3)	13.8	N/A	4.5

Peak Bandwidth (MB/Sec)

Platform	LA-MPI	MPT	MPICH
O2K (shared Mem)	145	135	85.7
O2K (HIPPI-800)	135	73	N/A
O2K (IP)	36.8	34.8	8.7
i686 (shared Mem)	168	N/A	131
I686 (IP)	11.3	N/A	11.0
ES45 (1G shared Mem)	690	N/A	760
ES45 (1G IP)	281(1NIC)	N/A	290

Allgather (uSec/call)

Host x nProcs	LA-MPI		MPT	
	40 bytes	**40000 bytes**	**40 bytes**	**40000 bytes**
1 x 2	24.5	623	35.2	959
1 x 4	39.3	1380	70	3140
1 x 32	595	15500	403	57600
1 x 64	2590	48800	774	153000
1 x 120	8480	129000	Would not run	
2 x 4	329	7150	1660	19900
2 x 32	907	165000	9700	245000
4 x 4	633	18400	2400	51200
4 x 32	1669.2	407891.9	13500	639000