



Optimizing Linux Cluster Performance by Exploring the Correlation between Application Characteristics and Gigabit Ethernet Device Parameters

May 2004

Onur Celebioglu
Dell Inc.



Agenda

- Introduction
- Methods to reduce CPU overhead in communications
- Testing environment
- Baselineing
- Benchmark results
- Conclusion

Introduction

- Cluster interconnect is one of the key components in High Performance Computing Cluster (HPCC)
- Several important factors in choosing the interconnect are latency, performance, price per port and communication characteristics of the application
- The impact of each of these parameters, i.e., latency, bandwidth, CPU utilization, on the overall system performance depends on the applications' computation/communication mix
- Gigabit Ethernet is one of the most popular cluster interconnects
- Ethernet has higher CPU overhead compared to specialized cluster interconnects such as Myrinet, InfiniBand, etc

Methods to reduce CPU utilization

- Jumbo Frames – Reduce the number of frames
- Interrupt Coalescing – Reduce number of interrupts
- Checksum offload – Offload checksum calculation
- TCP Segmentation offload – Offload packet segmentation
- TOE – TCP Offload Engine

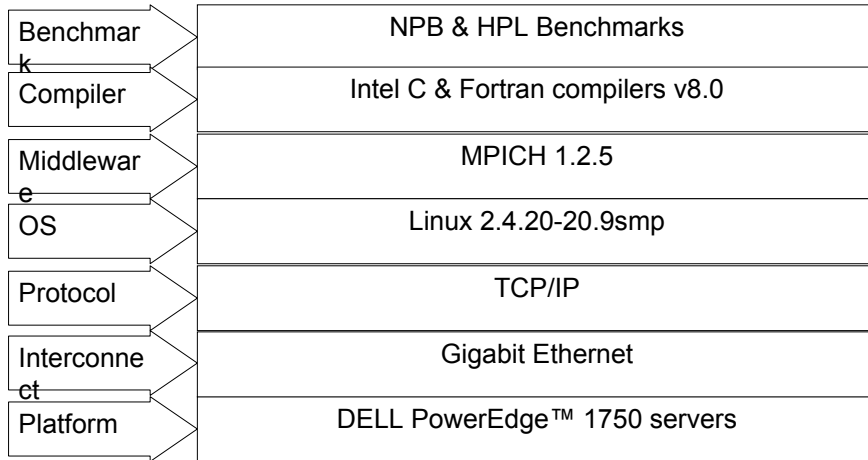
Interrupt Coalescing

- Gigabit Ethernet without interrupt coalescing may produce significant number of interrupts
- Reduce the CPU overhead by grouping interrupts
- At the receiving end, the CPU gets notified about the departure of a group of packets through a single interrupt
- Decision to generate interrupts may be based on number of packets or a fixed timeout
- This mechanism was already being used in networks before Gigabit Ethernet

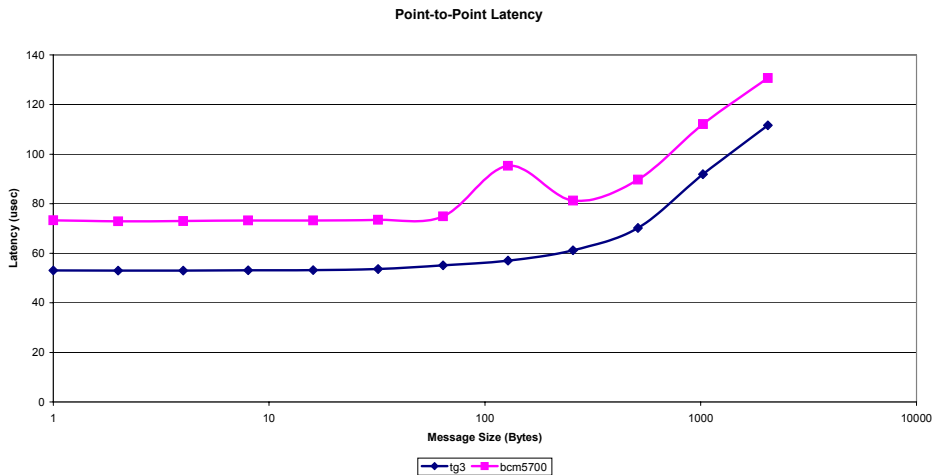
Question?

- What are the driver parameters we can modify to tune application performance?
- How are different applications effected by changing these parameters?

Experimental Environment

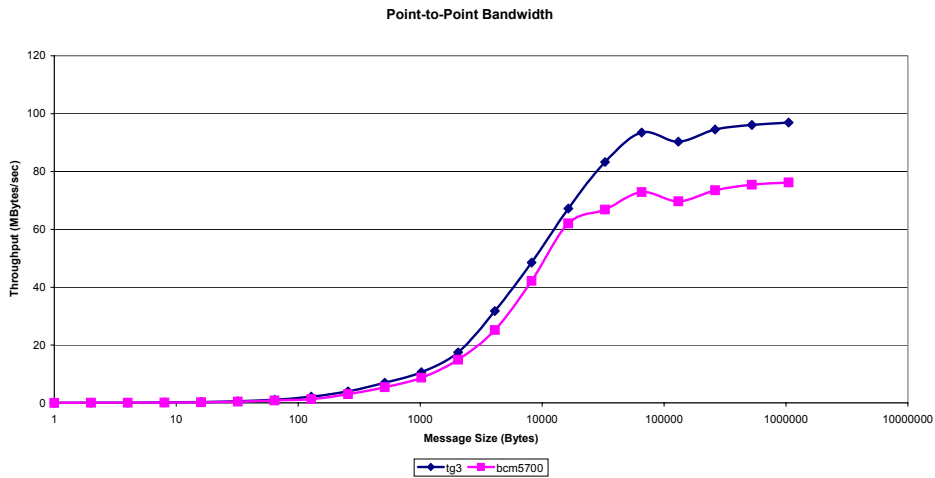


Point-to-point half round-trip time with varying message size



Based on Pallas MPI benchmark test performed by Dell Labs in Jan, 2004 on two Power Edge 1750 with 3.06GHz processors, 4GB RAM and running RedHat® Linux®. Actual performance will vary based on configuration, usage and manufacturing variability.

Point-to-point bandwidth with varying message size



Based on Pallas MPI benchmark test performed by Dell Labs in Jan, 2004 on two Power Edge 1750 with 3.06GHz processors, 4GB RAM and running RedHat® Linux®. Actual performance will vary based on configuration, usage and manufacturing variability.

Dell Enterprise Solutions Engineering

CPU utilization with different drivers using netperf

TG3

Socket Size	Message Size	Send Throughput	Recv local CPU %T	Send remote CPU %T
bytes	bytes	Mbits/s		
262142	4096	940.79	61.98	83.67
262142	8192	940.84	55.69	87.05
262142	32768	940.9	49.38	87.47

BCM5700

Socket Size	Message Size	Send Throughput	Recv local CPU %T	Send remote CPU %T
bytes	bytes	Mbits/s		
262142	4096	940.64	51.8	19.05
262142	8192	940.94	44.52	19.53
262142	32768	940.72	41.75	19.42

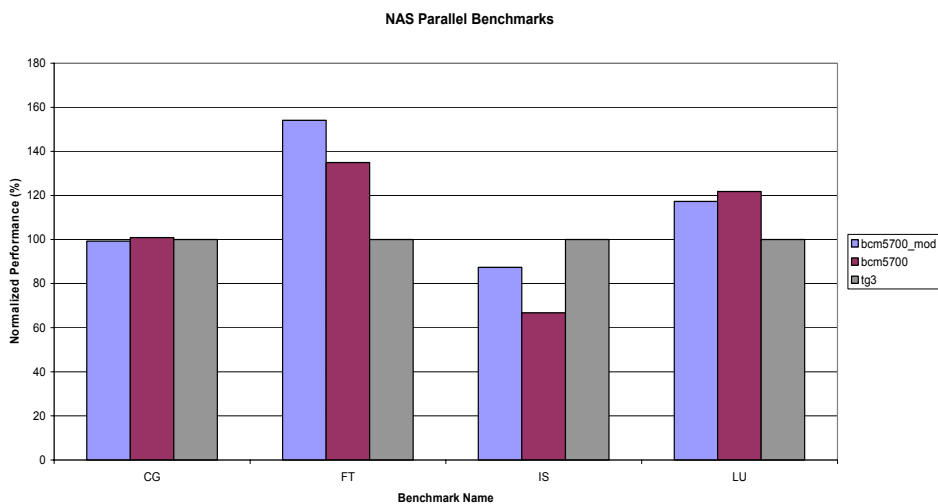
Based on Netperf test performed by Dell Labs in Jan, 2004 on two Power Edge 1750 with 3.06GHz processors, 4GB RAM and running RedHat® Linux®. Actual performance will vary based on configuration, usage and manufacturing variability.

Dell Enterprise Solutions Engineering

Interrupt Coalescing Parameters Available in bcm5700 Drivers

- rx_coalesce_ticks
 - Configures the number of 1 usec ticks before the NIC generates receive interrupt after receiving a frame.
- rx_max_coalesce_frames
 - Configures the number of received frames before the NIC generates receive interrupt.
- tx_coalesce_ticks
 - Configures the number of 1 usec ticks before the NIC generates transmit interrupt after transmitting a frame.
- tx_max_coalesce_frames
 - Configures the number of transmitted frames before the NIC generates transmit interrupt.

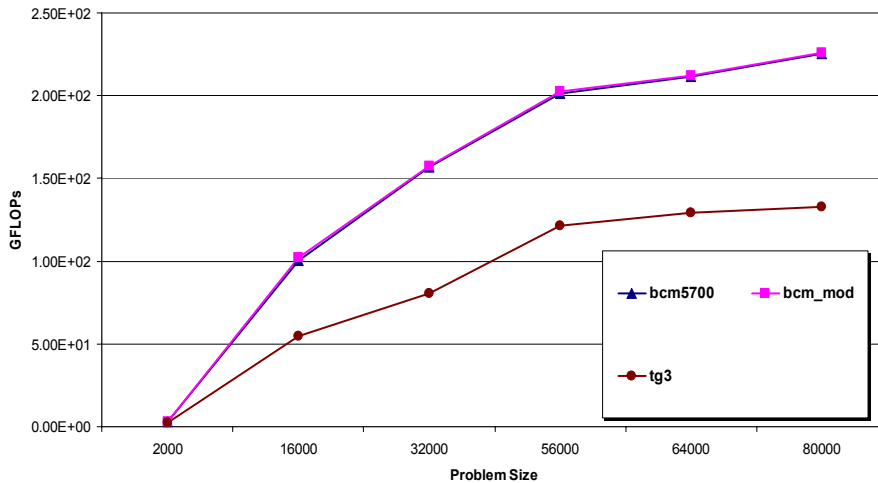
NAS Parallel Benchmark results



Based on NAS Parallel benchmark tests performed by Dell Labs in Jan, 2004 on 32 Power Edge 1750 with 3.06GHz processors, 4GB RAM and running RedHat® Linux®. Actual performance will vary based on configuration, usage and manufacturing variability.

High Performance Linpack (HPL) performance

HPL Results



Based on HPLinpack benchmark tests performed by Dell Labs in Jan, 2004 on 32 Power Edge 1750 with 3.06GHz processors, 4GB RAM and running RedHat® Linux®. Actual performance will vary based on configuration, usage and

Dell Enterprise Solutions Engineering

Conclusions

- It is not possible to correlate the application performance only to point-to-point message passing bandwidth and latency. CPU overhead in communications is a very important factor for the performance of some applications
- Ethernet device driver is a major component that affects system performance. Even for a standardized interconnect such as Gigabit Ethernet, the choice of the most efficient device driver is critical
- The choice and configuration of the Ethernet device driver depends not only on applications' communication pattern but also the applications' computation characteristics
 - Applications that have low communication and high computation needs tend to perform better with drivers optimized for low CPU utilization
 - Applications that are more communication intensive tend to perform better with bandwidth and latency optimized Ethernet device drivers even if the CPU utilization in communications is high.
- Network device drivers should provide means to adjust the parameters that affect latency, throughput and CPU utilization.

Dell Enterprise Solutions Engineering