Optimizing Latency in Beowulf Clusters

Rafael Garabato[1] Andrés More[12] Victor Rosales[1]

Argentina Software Design Center (ASDC - Intel Córdoba)[1]

Instituto Universitario Aeronáutico (IUA)[2]

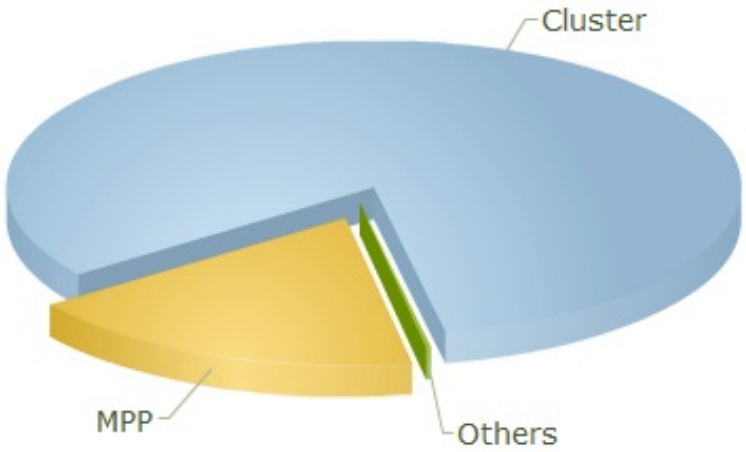
V Latin American Symposium on HPC (HPC LATAM 2012)

Outline

Beowulf Clusters

Cheap Scaling. Open Source Software. Ubiquitous.

- Instead of expensive high-end SMPs, just interconnect commodity systems.
- Key driving factor is cost. Out-of-the-box hardware plus open source software.
- Clusters represent 80% of the systems at the Top500 list.
- Ethernet is included on-board, so it is the preferred network fabric.



Communication Latency

Definitions. Highly Impacts Distributed Application.

- Communication latency is the required time for information to flow from one compute node into another.
- Latency-sensitive applications might justify expensive, special-purpose hardware.

Latency	Technology
30-125 μsec	1Gb Ethernet
5-30 μsec	10Gb Ethernet

Related Work

Component Optimizations. State-of-the-art.

- Communication latency is the required time for information to flow from one compute node into another.
- Latency-sensitive applications might justify expensive, special-purpose hardware.

System	Latency	Description
HP BL280cG65	0.49 μsec	Best Latency
Fujitsu K Computer	6.69 μsec	Top System

Problem Statement

Latency Equation. Zero bytes Latency.

- Latency time is startup plus throughput for n bytes

$$t(n) = \alpha + \beta \times n \tag{1}$$

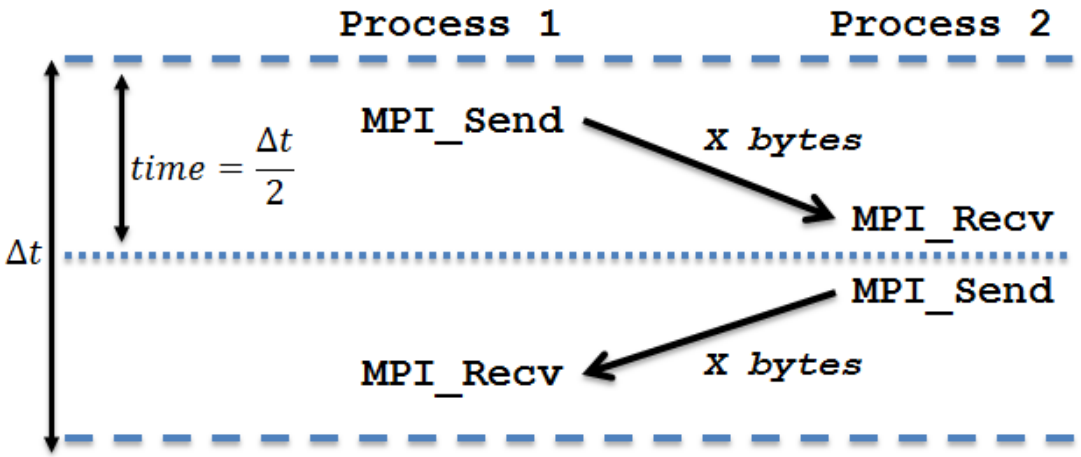
- Zero bytes latency means no payload.

$$t(0) = \alpha \tag{2}$$

Intel MPI Benchmarks

Ping Pong. Basic MPI Primitives.

- IMB Ping Pong performs a round-trip message transfer
- It is run multiple times then averages, uses varying lengths
- Uses only basic MPI routines (`MPI_Send` and `MPI_Recv`)



Other Benchmarks
HPL. HPCC.

- HPL is a pure matrix multiplication benchmark for distributed memory.
- HPCC is a suite made of 7 benchmarks: HPL, DGEMM, STREAM, PTRANS, RandomAccess, FFT, b_eff
- HPL, DGEMM, STREAM and FFT run in parallel. PTRANS, RandomAccess and b_eff run cluster wide. Latency impact is not the same on any kernel.

Latency Optimizations

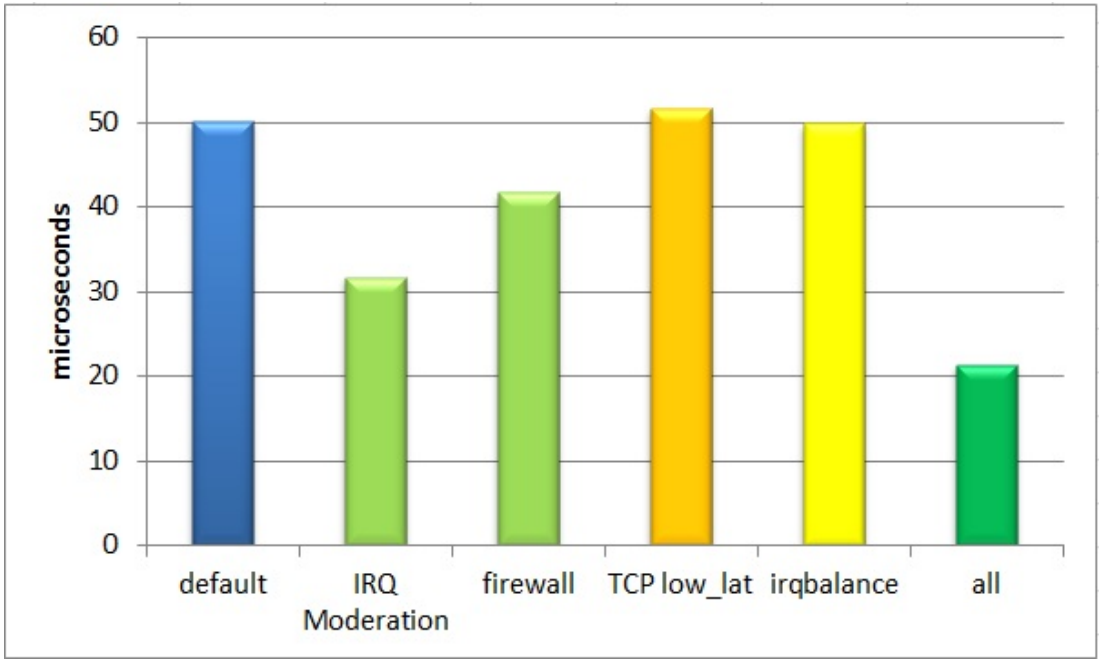
Ethernet Drivers. System Services. Kernel Settings.

Several opportunities were identified after reviewing alternatives

- Drivers: Interrupt Moderation is a technique to reduce CPU load by caching IRQs before service them all together.
- Services: Interrupt Balancing distributes IRQs to balance load and power consumption. Firewalls inspect each exchanged message.
- Kernel: TCP Stack favors higher throughput over low latency by default.

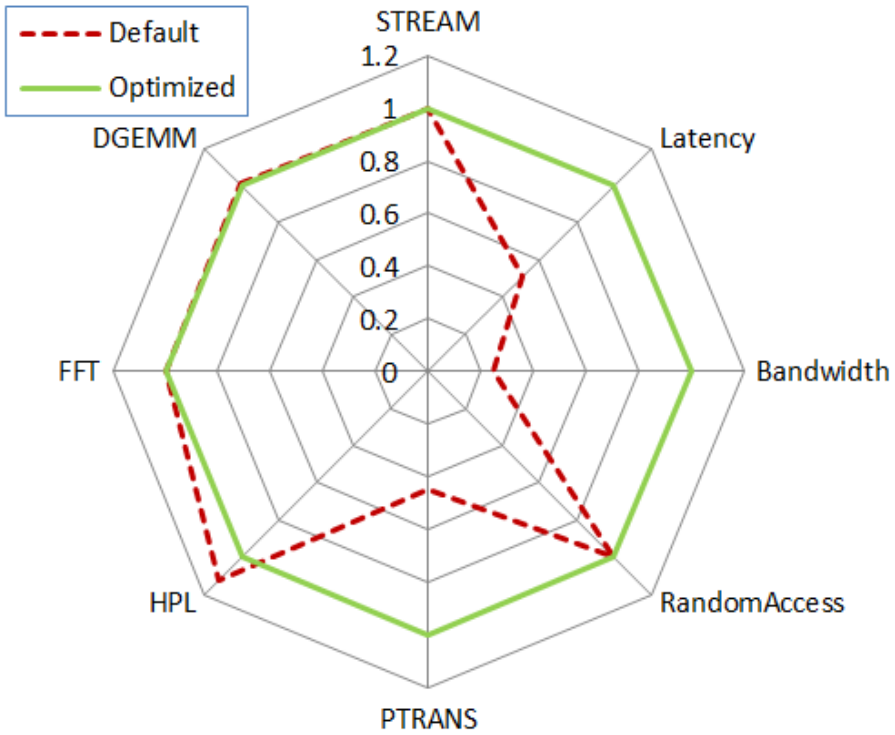
Optimization Impact I

IMB Pingpong.



Optimization	$\tilde{x}$ (σ^2)	Impact
Default	50.03 (4.31)	N/A
IRQ Moderation	31.63 (0.83)	36.79%
Firewall	41.62 (8.90)	16.82 %
TCP LL	51.59 (8.22)	-3.11%
IRQ Balance	49.72 (9.68)	0.62 %
Combined	21.31 (2.09)	57.40 %

Optimization Impact II
HPCC. HPL.



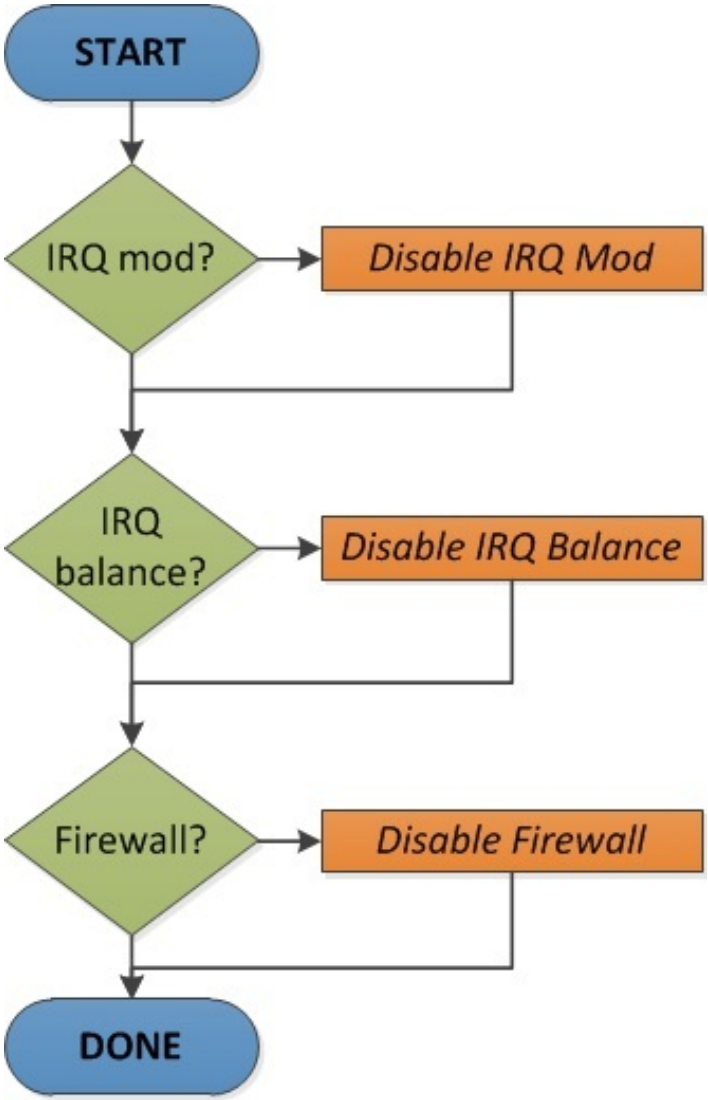
Optimization	Wall-time	Gflops
Default	00:20:46	0.02921
Optimized	00:09:03	0.07216

Optimization Impact III
mpiBLAST.

- mpiBLAST is an open source tool that implements DNA-related algorithms to find regions of similarity between biological sequences.
- For a fixed workload, the execution time was reduced 11%.

Optimization	Wall-time
Default	534.33 seconds
Optimized	475.00 seconds

Optimization Procedure
Detailed Steps.




```
$ pdsh -N -a '/sbin/modinfo -F version e1000e' | uniq
1.2.20-NAPI
```

(c) Are interrupt moderation settings in HPC mode?

```
# pdsh -N -a 'grep "e1000e" /etc/modprobe.conf' | uniq
options e1000e InterruptThrottleRate=0
```

2. System Services

(a) Is the firewall disabled?

```
# pdsh -N -a 'service iptables status' | uniq
Firewall is stopped.
```

(b) Is the firewall disabled at startup?

```
# pdsh -N -a 'chkconfig iptables --list'
irqbalance 0:off 1:off 2:off 3:off 4:off 5:off 6:off
```

(c) Was the system rebooted after stopping firewall services?

```
$ uptime
 15:42:29 up 18:49, 4 users, load average: 0.09, 0.08, 0.09
```

(d) Is the IRQ balancing service disabled?

```
# pdsh -N -a 'service irqbalance status' | uniq
irqbalance is stopped
```

(e) Is IRQ balancing daemon disabled at startup?

```
# pdsh -N -a 'chkconfig irqbalance --list' | uniq
irqbalance 0:off 1:off 2:off 3:off 4:off 5:off 6:off
```

Once gathered all the information required to known if optimizations can be applied, the following list can be used to apply configuration changes. Between each change a complete cycle of measurement should be done. This include contrasting old and new latency average plus their deviation using at least IMB Ping Pong.

Disable IRQ Moderation

```
# pdsh -a 'echo "options e1000e InterruptThrottleRate=0" >> \
/etc/modprobe.conf'
```

Summary

- Only by changing default configuration latency can be highly improved. As a reference, from 50 μs to 20 μs.
- We contrasted different optimization alternatives and their impact, both on benchmarks and a real world application.

- Future Work
  - Impact characterization in computational kernels.
  - Endless optimization opportunities on other components like BIOS, firmware, networking gear.
  - Measure impact on actual research and development processes on academia or industry.