Demonstrating the Improvement in the Performance of a Single Lustre* Client from Version 1.8 to Version 2.6

With a Stop Along the Way at Versions 2.1 and 2.4

Andrew Uselton, Gabriele Paciucci, Jinshan Xiong - Intel®

April 8, 2014

Overview

- Experimental Platform
  - Hardware
  - Software
- Single client performance comparison across versions
  - 1.8
  - 2.1
  - 2.4
  - 2.6 pre-release (2.5.57)
- Establishing the peak representative value
  - Data volume
  - Transfer size
  - Saturating the LUN
  - Saturating the server
  - Saturating the network

Experimental Platform: Hardware

- The Grizzly1 cluster at Swindon
  - Clients
    - Dual core Ivy Bridge E5 - 2697 v2 2.70 GHz C1 step (24 CPU/48 threads)
    - 16x R1208GL 1U Node
    - 64 GB RAM
    - Intel True Scale QLE7340 dual homed QDR PCIe x8 gen2 HCA
  - Servers
    - Dual Intel Xeon CPU E5 2637 v2 @ 3.50GHz
    - MDS: 64 GB RAM, OSS: 128 GB RAM
    - 1x Single Port TrueScale 7340 QDR InfiniBand
    - 2 MDSs, each with 4x Intel (fast) SSDs, mds0 with RAID10
    - 4 OSSs, each with 4 RAID5 LUNs: 5x Western Digital WD4001FYYG RE 4TB SAS Hard Drive w/ 7200RPM
  - The mds1 server was repurposed to host one very fast SSD-based OST.
  - Network
    - Intel True Scale 12300 36 port QDR switch

Experimental Platform

Software

Servers:

- RHEL 6.4
- Linux 2.6.32-431
- Lustre* 2.5.1

Clients:

- RHEL 6.4
- Linux 2.6.32-358

With Lustre* versions:

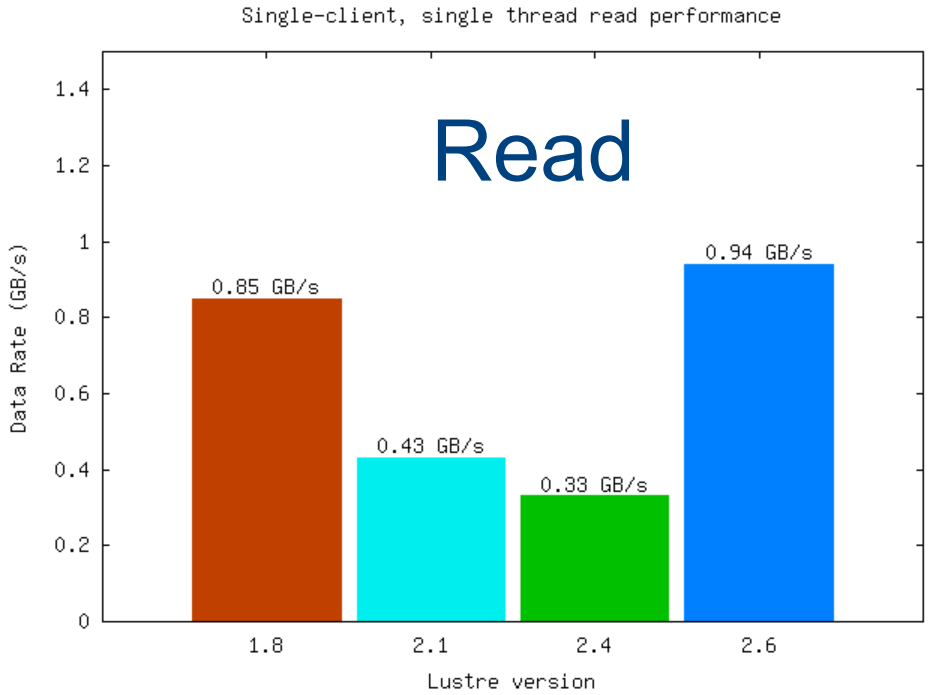
- 1.8.9-wc1
- 2.1.6
- 2.4.3
- 2.6 (2.5.57 release candidate)

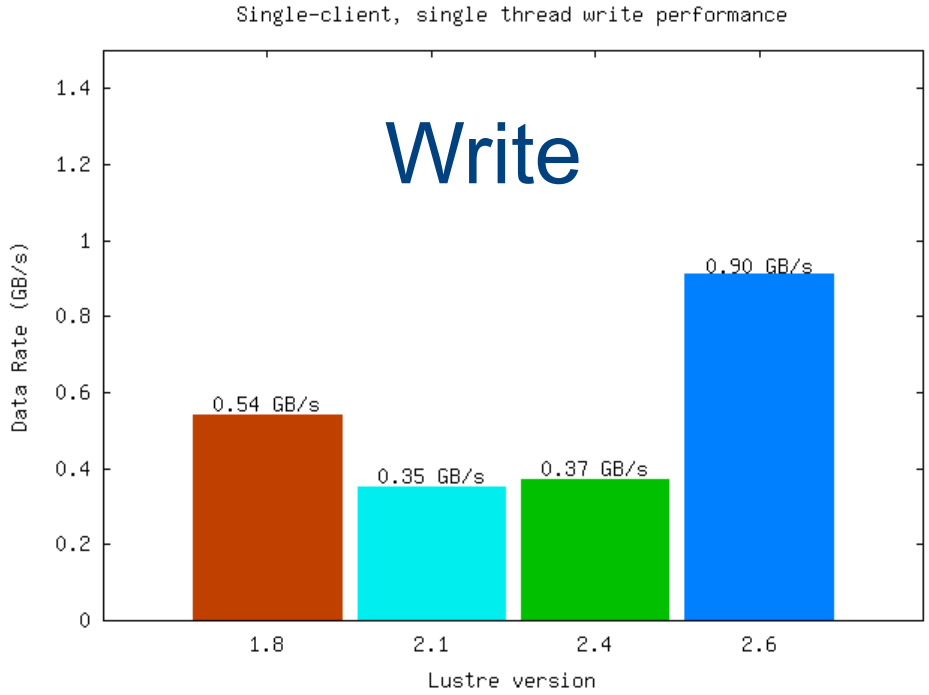
Experimental Design

IOR “Hero” Runs

- File-per-process
- Streaming
- Read-your-neighbor
- Enough data to exceed cache
- 1 MB transfers
- Using all OSTs

Experimental Results: Overview





From Lustre* 1.8 to 2.1 there was a widely recognized regression in the single-client, single-thread performance of Lustre. Over the last year Jinshan Xiong of Intel addressed this issue, and his patches have landed in the 2.6 source.

The rest of this talk will review what the above numbers actually mean and how they were generated.

If we have time, we will look at some additional measures of performance improvements.

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

Experimental Results

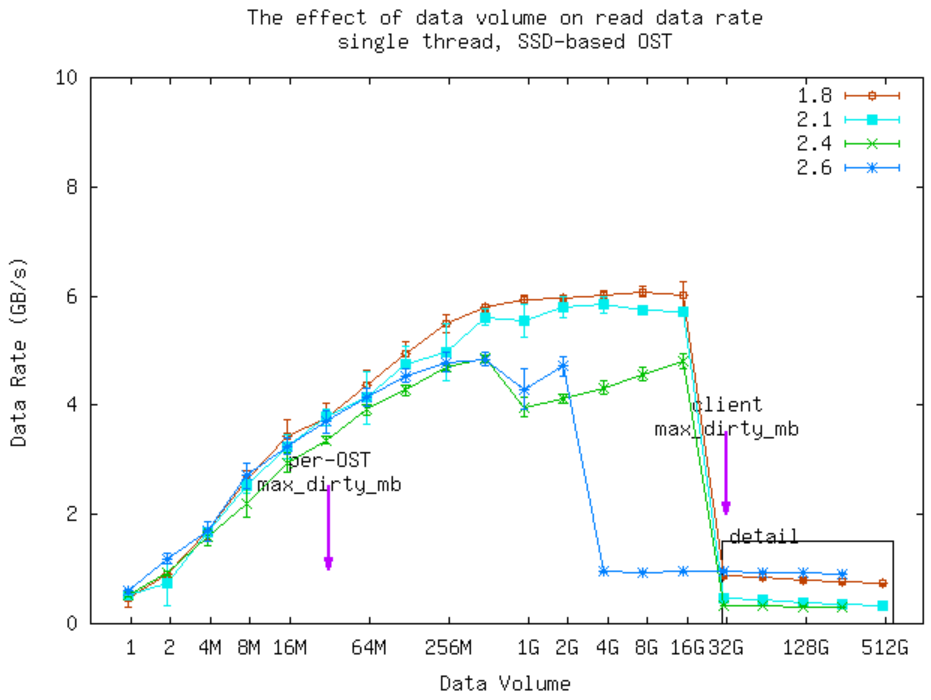
- In order to measure the single-client, single thread performance you need an experimental setup where every stage of the I/O pipeline is as fast as possible.
- We repurposed the four fast SSDs on a secondary MDS to create a very fast single OST. obdfilter-survey measured this OST at 1.6 GB/s.
- With appropriate tunings (map_on_demand=32), the TrueScale IB network sustained 1.6 GB/s point-to-point as well (via lnet selftest).

Establishing the Peak Representative Value

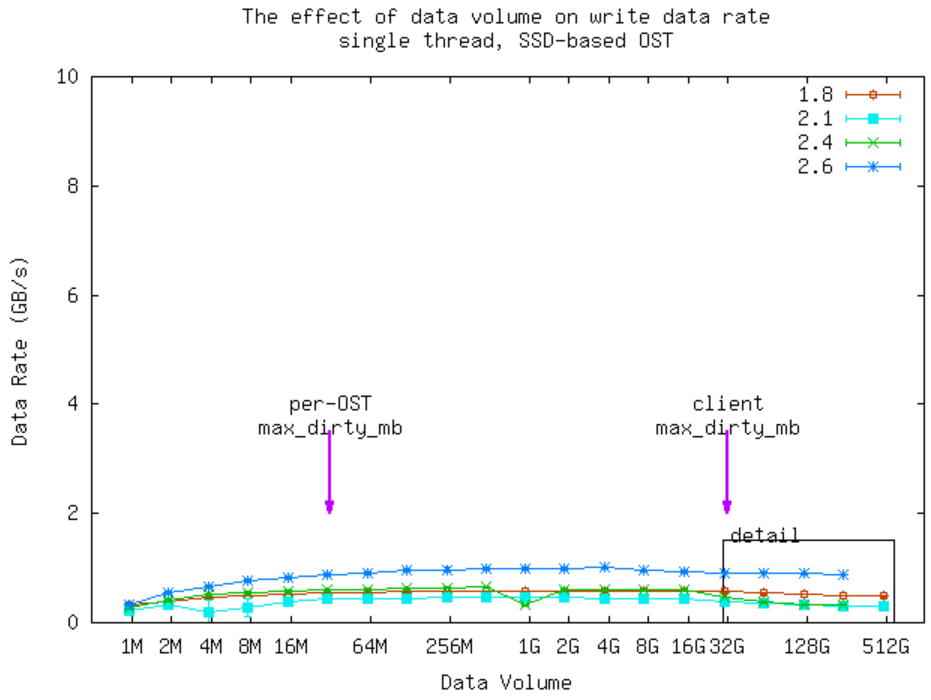
- Did each experiment move enough data to establish the asymptotic limit?
- Was the transfer size large enough to get optimal performance?
- Is the file system as a whole fast enough to measure the peak client performance?
- Is the network fast enough to measure peak client performance?
- Is a single LUN fast enough to measure single task performance?

Single-Client, Single-Task Performance Using an SSD: Data Volume

Comparing Lustre* versions 1.8, 2.1, 2.4, and 2.6

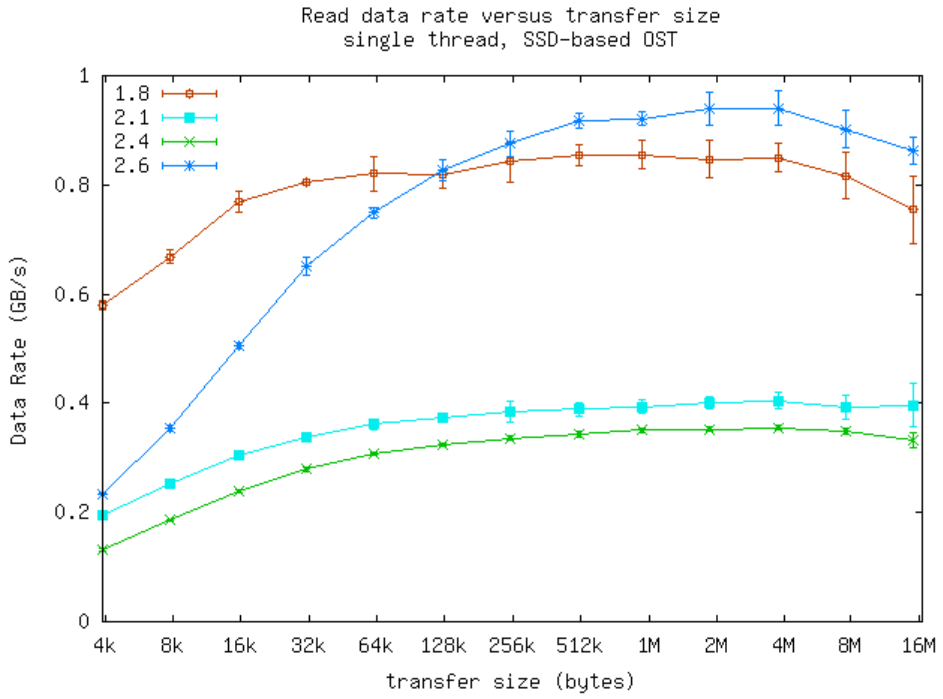


Read

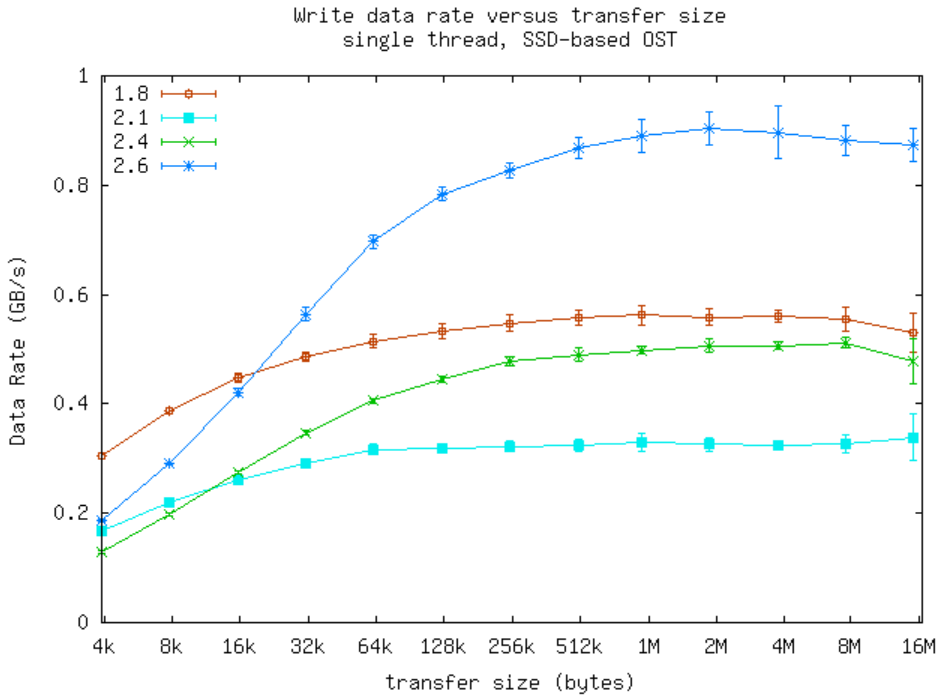


Write

Single-Client, Single-Task Performance Using an SSD: Transfer Size



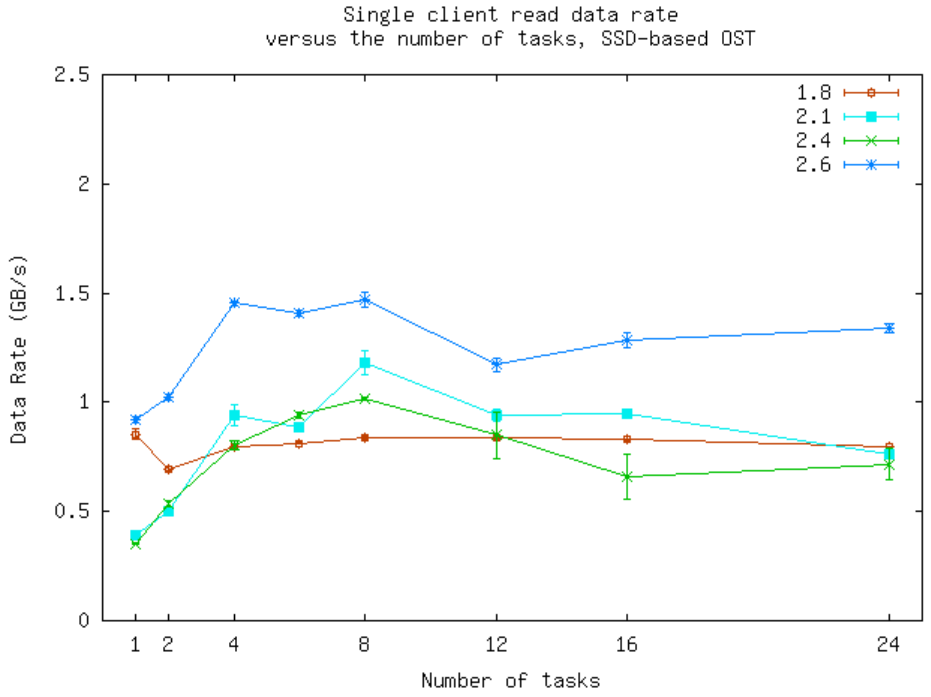
Read

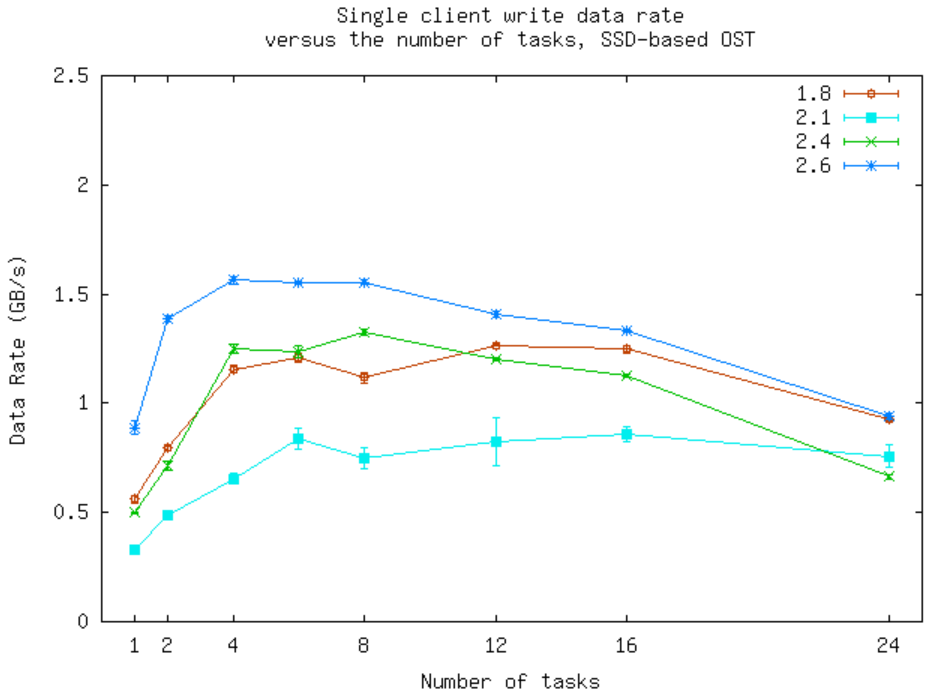


Write

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

Single Client Performance Using an SSD: Number of Tasks





Read

Write

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

Other Performance comparisons

We have established the improvement in single-client, single thread performance to a single target OST, and for that we required a specially configured OST.

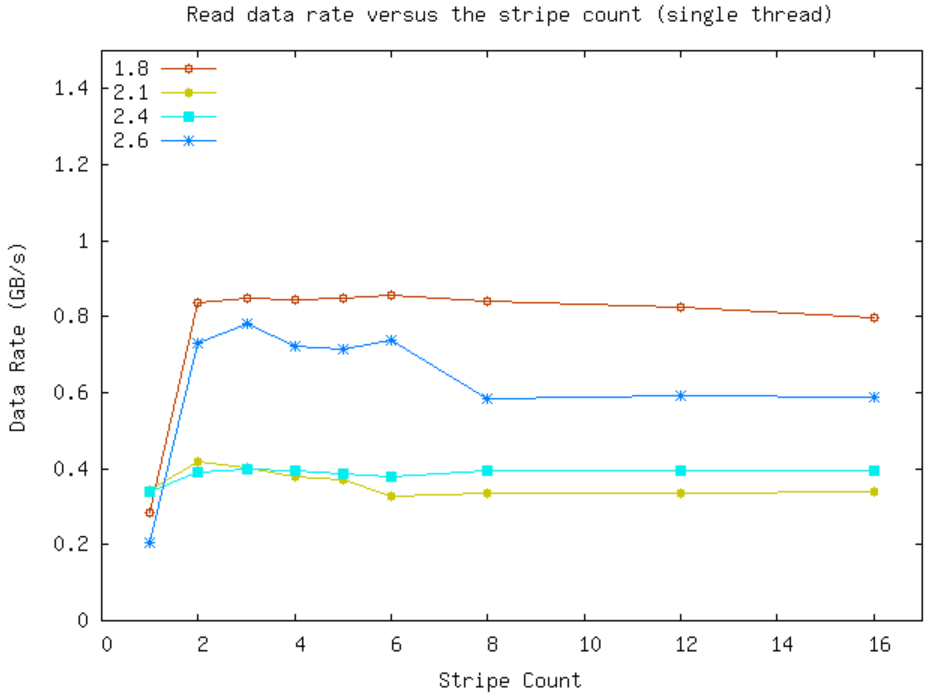
So what about regular spinning-disk-based LUNs (HDDs)?

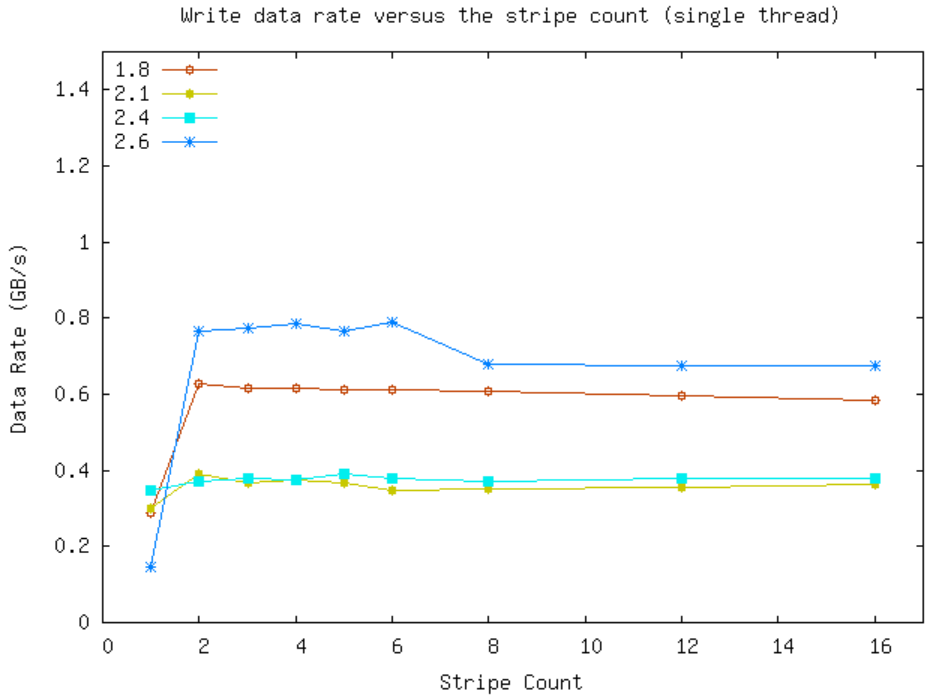
A single LUN is not fast enough to see the peak client performance for 1.8 or 2.6. What about multiple LUNs?

Does the 2.6 client offer benefits to larger scale I/O targeting slower devices?

The HDD-based file system has 16 OSTs on four OSSs, and the 16 clients have 24 cores each.

Single-Client, Single Task Performance Using HDDs: Stripe Count



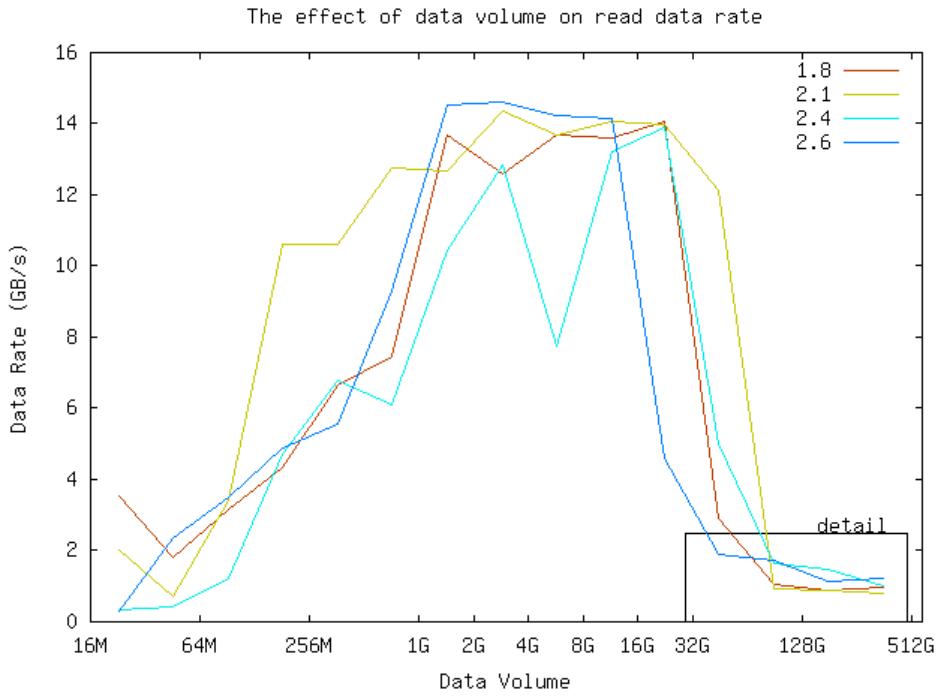


Read

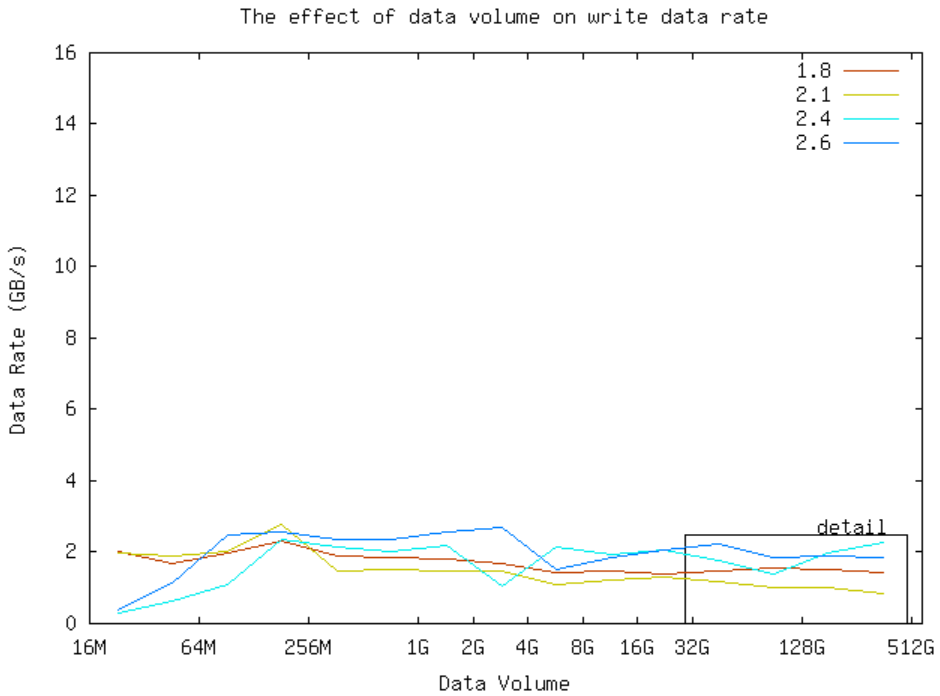
Write

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

Single-Client, 24 Task Performance Using HDDs: Data Volume



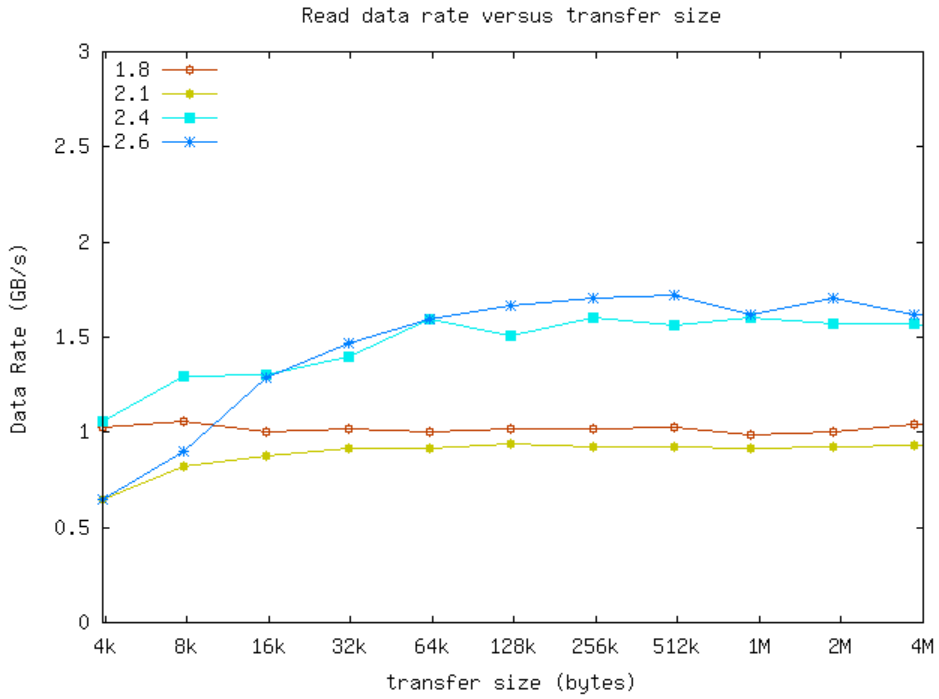
Read



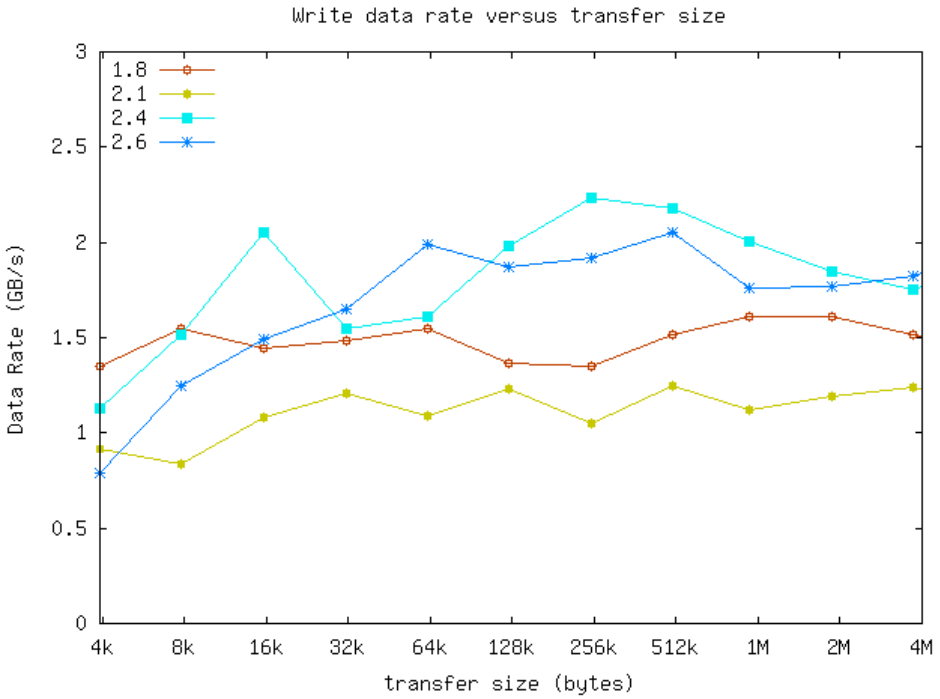
Write

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

Single-Client, 24 Task Performance Using HDDs: Transfer Size



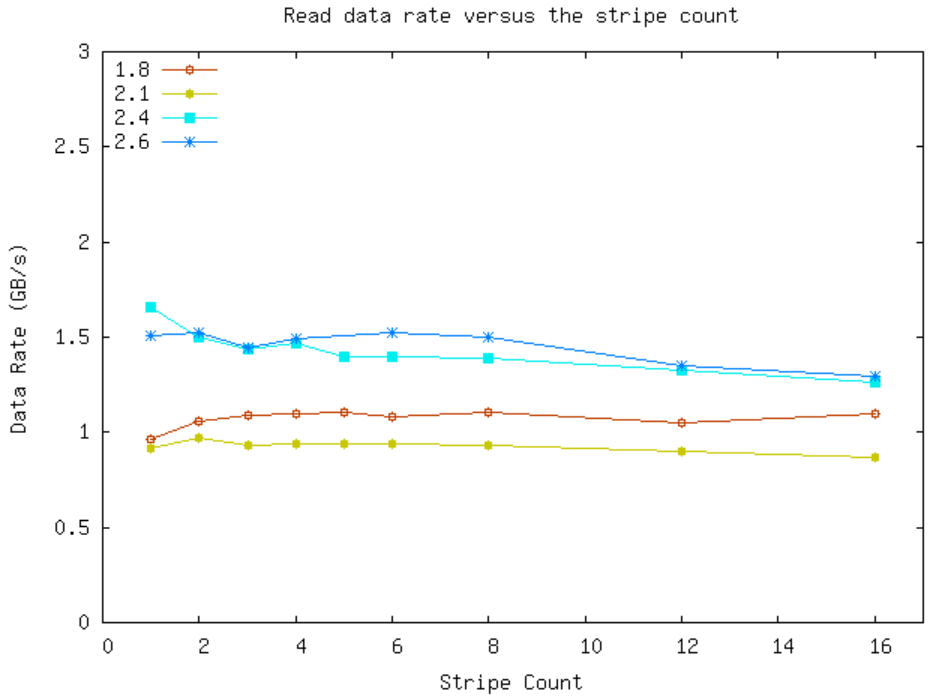
Read

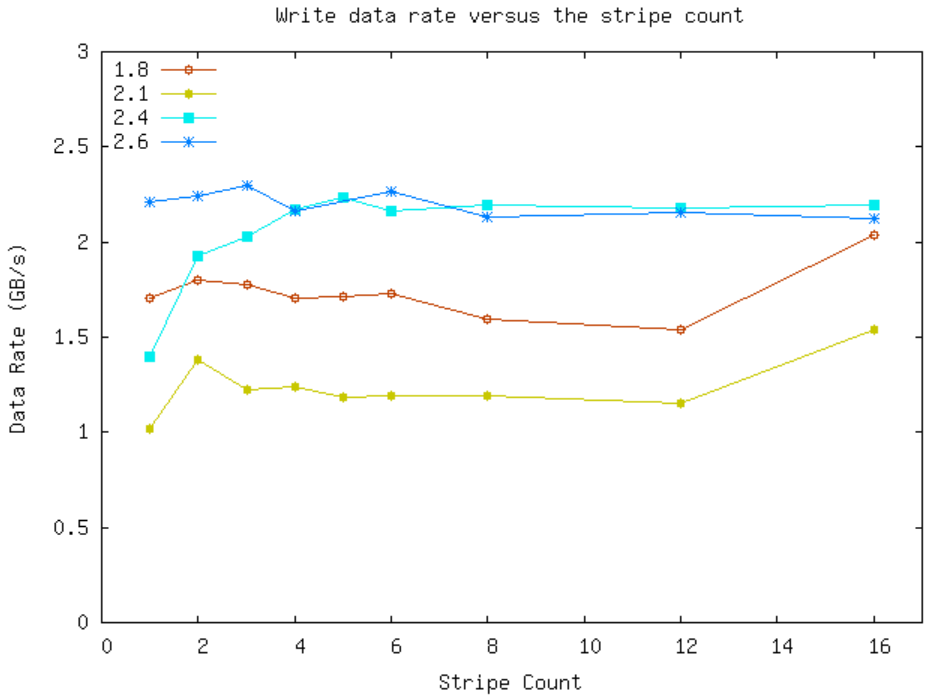


Write

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

Single-Client, 24 Task Performance Using HDDs: Stripe Count



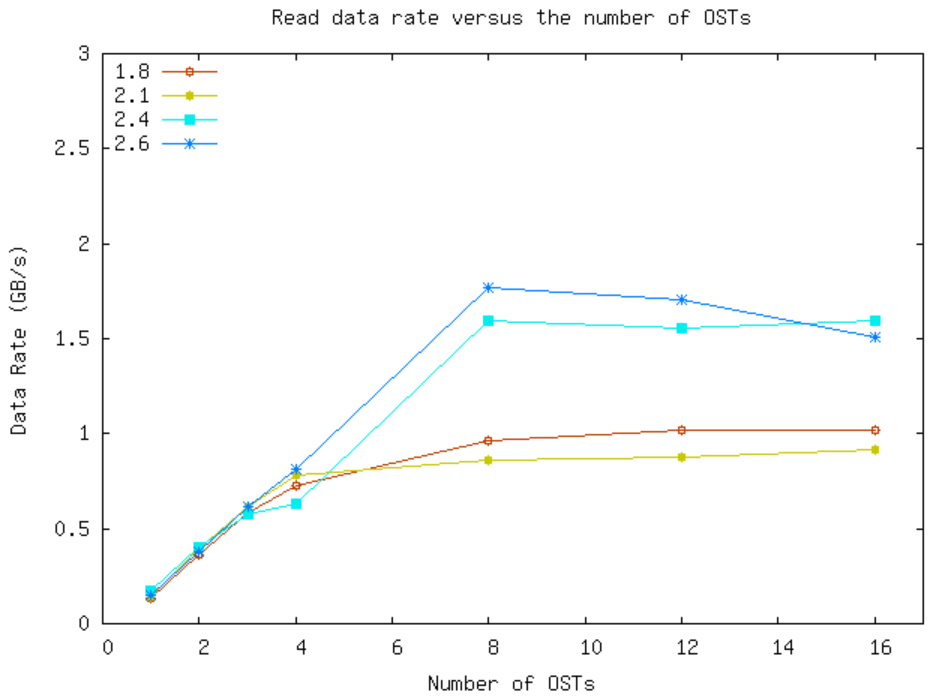


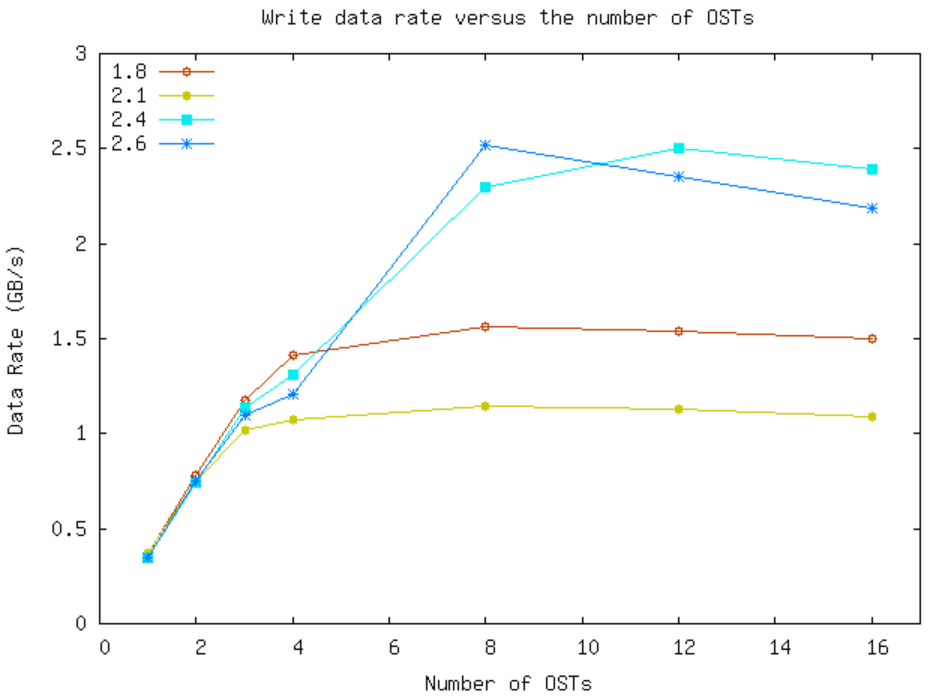
Read

Write

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

Single-Client, 24 Task Performance Using HDDs: OST Count



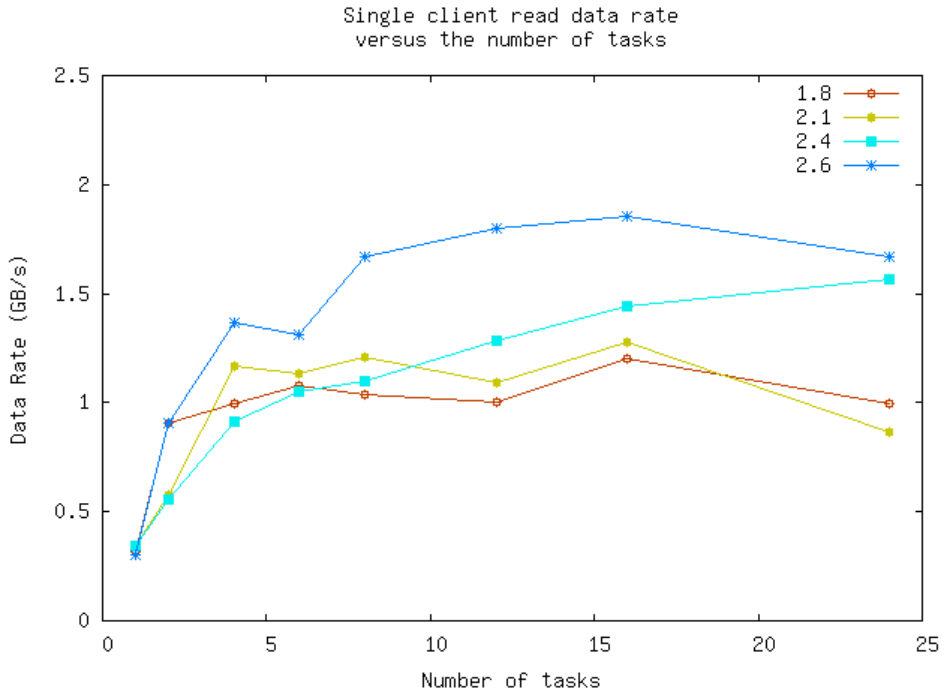


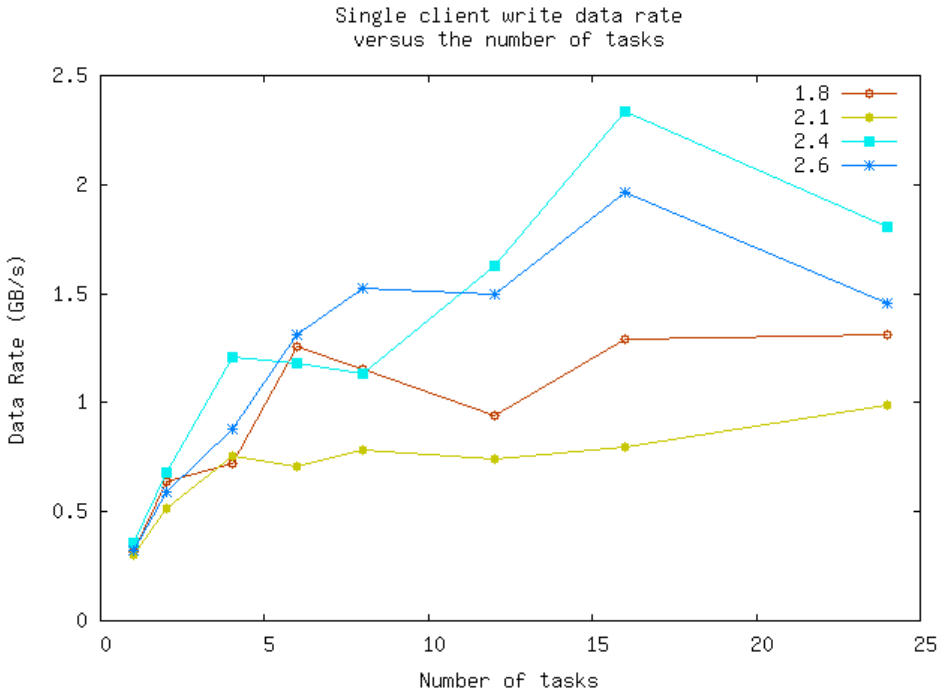
Read

Write

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

Single-Client Performance Using HDDs: Task Count





Read

Write

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

Conclusions

- Single-client, single-task peak performance is much improved in Lustre* 2.6
- This is true for both the single task targeting a single LUN and for the client node as whole targeting the file system
- The 2.6 client also performs well when targeting a collection of spinning-disk OSTs.

To do

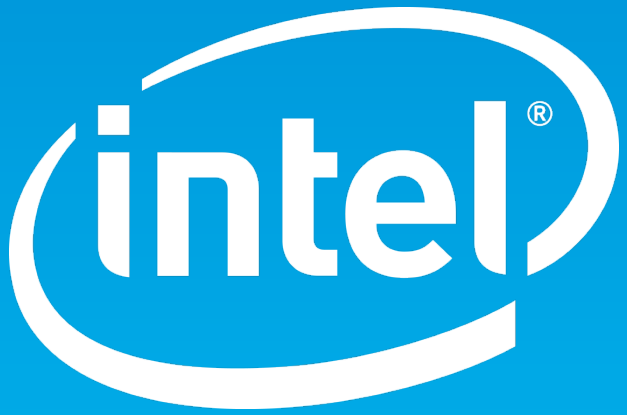
- The per-I/O overhead in 2.x needs to be improved and we are actively pursuing that.
- The read-ahead code will also want some attention, as there may be room for improvement there as well.

Acknowledgements

Jinshan, Gabriele, and I would like to thank the Swindon HPC Lab and Jamie Wilcox, in particular, for their assistance and support for this work.

Thank you

Questions?


intel