SMSi

Lustre Performance over the Infiniband WAN

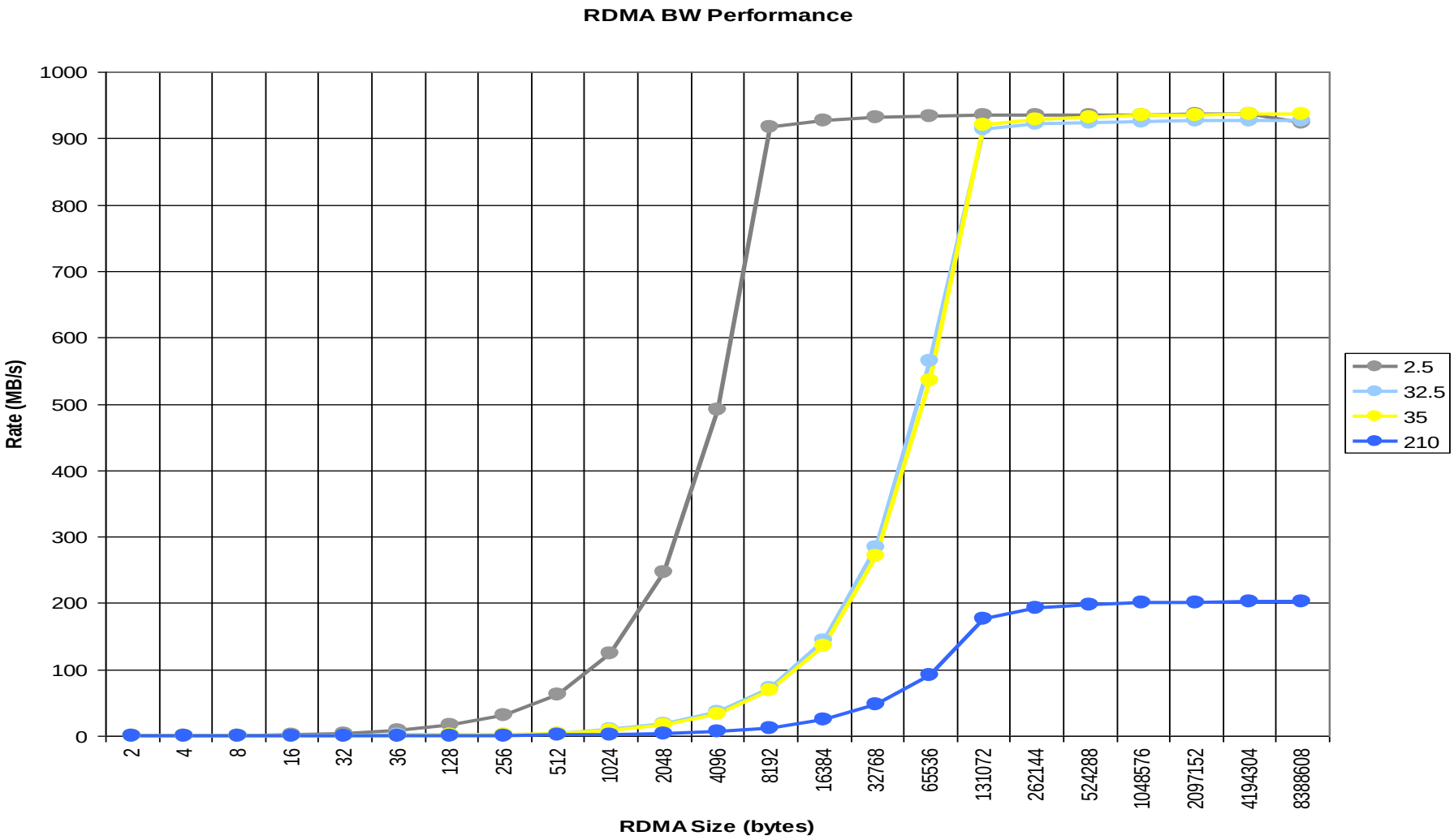
Jeremy Filizetti
jfilizetti@sms-fed.com

Lustre Usage

How we use it

- In the data center, over a MAN and WAN
- Distributed data centers
  - Interconnected at 10 or 2.5 Gbps
  - Utilizes Obsidian Longbow XRs and Bay NX5010s for Infiniband range extension
- Each data center has their own file system
  - o2ib primary interface between OSS's and clients
  - tcp for MDS, secondary interface for OSS's
  - LNet routers used to bridge IB fabrics, convert tcp to o2ib for WAN performance

RDMA Performance



Lustre IO Performance

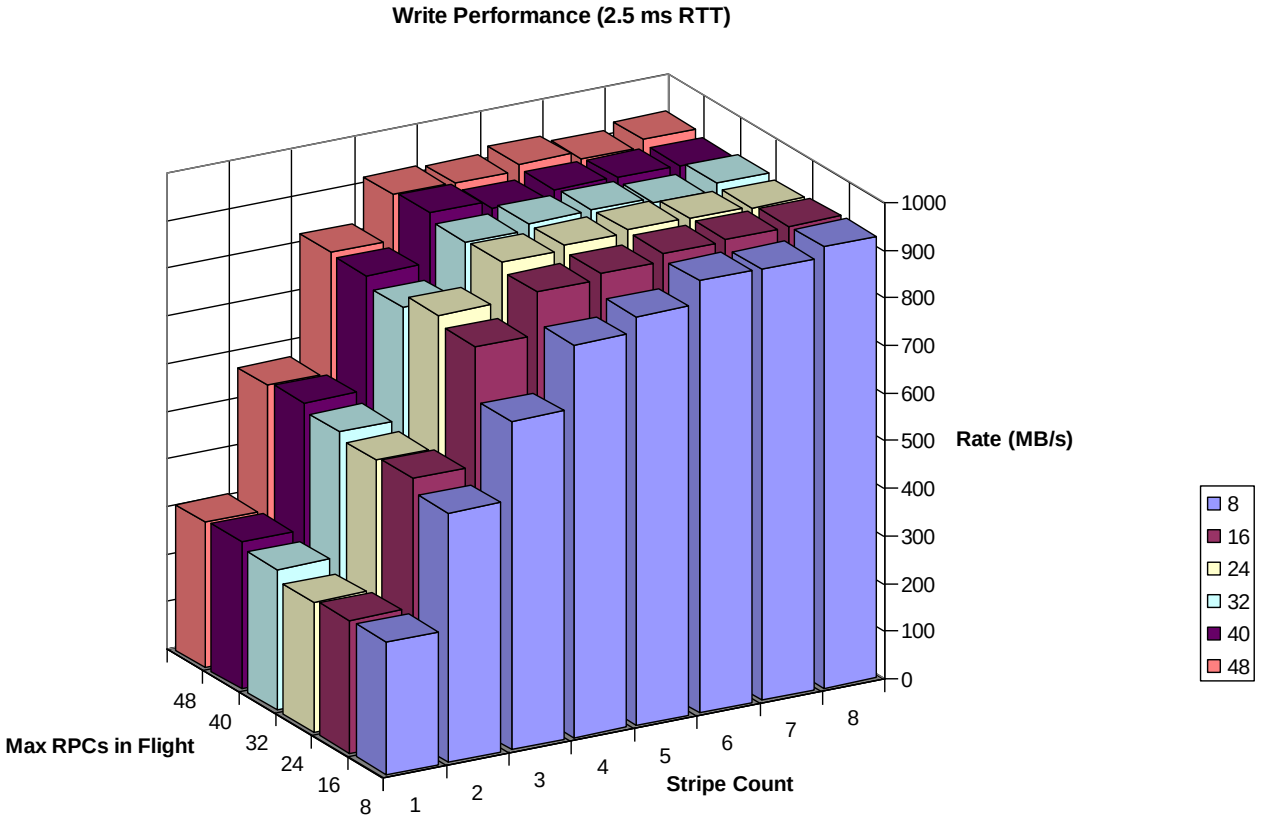
- Infiniband (o2ib) performance over MAN and WAN are excellent
  - Started off very poor (bug 14358)
- Key is to minimize RTT
  - Currently reads take 3 RTTs and writes take 2 RTTs
  - Random read performance will always be poor over WAN but could be improved if RTTs were reduced
- Read ahead is big factor in WAN read performance
- Tuning max_rpcs_in_flight is needed to meet BDP
  - Helps for lower stripe counts
- Problems popped up testing SLES11
  - Kernel CONFIG_SECURITY_FILE_CAPABILITIES option (bug 21439)
  - getxattr being called for every write to OST
  - Across WAN meant 1 RTT per min(write, stripe size)

Lustre IO Tests

Test data on next pages are from iozone

- Included close and fsync in times
- 1 MB stripe size and record length used
- Files of 4 and 8 GB size tested
- Most testing was done on a file system nearly empty
- max_dirty_mb was scaled to 4x max_rpcs_in_flight

IO Performance (2.5 ms RTT)

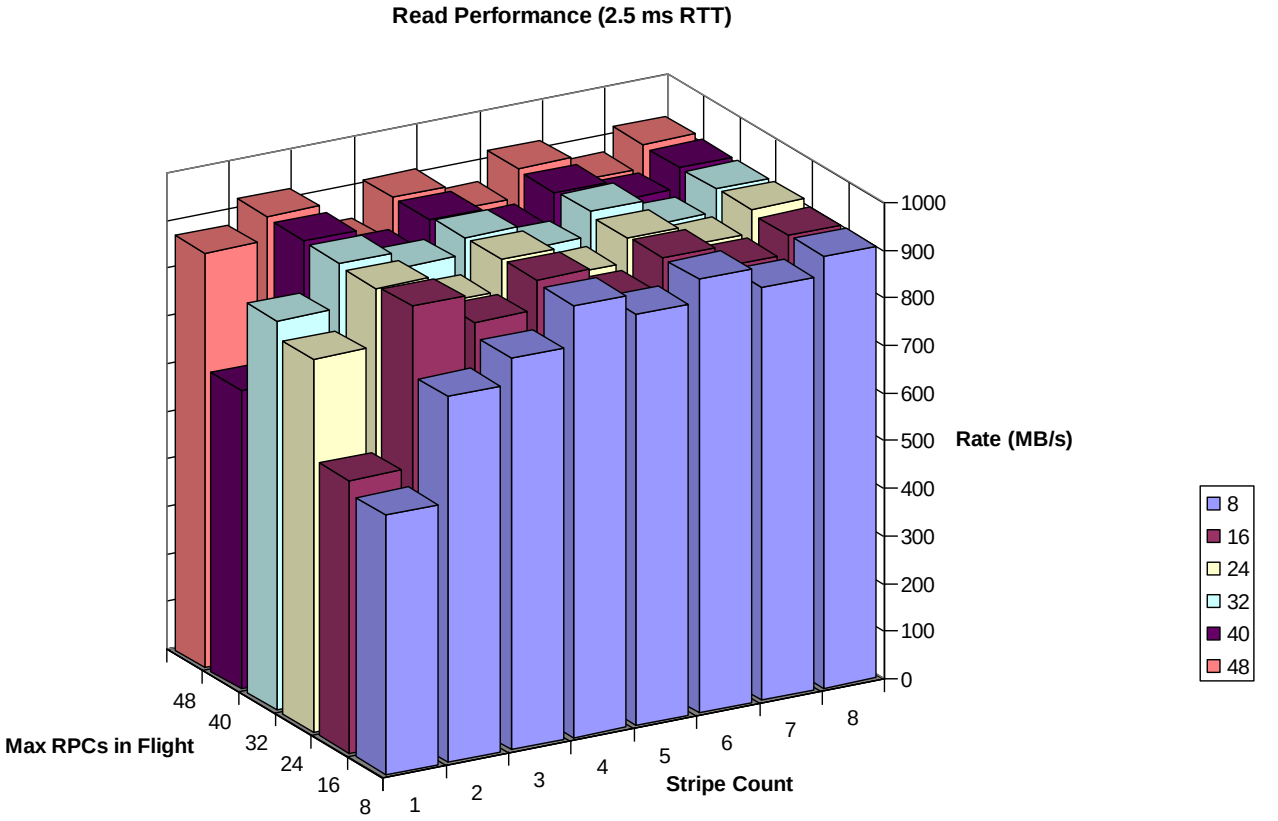


Best case performance Estimate:

- ib_write_bw –q 3 –s 1048576
  - 935.43 MB/s

	1	2	3	4	5	6	7	8
8	279	524	691	825	859	908	906	927
16	280	552	802	891	905	920	924	926
24	273	545	824	909	918	925	922	919
32	296	559	794	906	920	922	911	927
40	310	576	817	923	907	920	919	917
48	305	567	821	916	914	928	913	928

IO Performance (2.5 ms RTT)

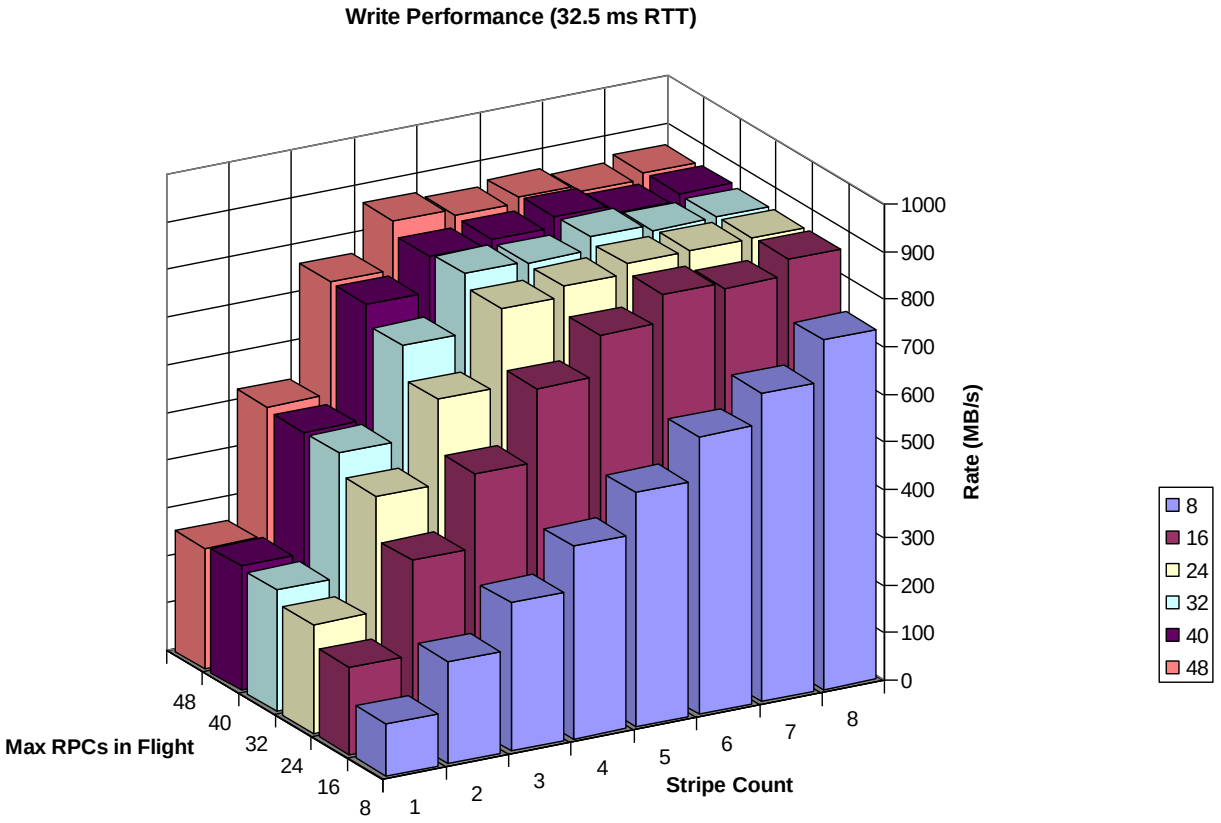


Best case performance Estimate:

- ib_write_bw –q 3 –s 1048576
  - 935.43 MB/s

	1	2	3	4	5	6	7	8
8	545	769	824	908	864	911	868	907
16	573	913	854	917	851	911	868	906
24	784	906	849	915	856	908	866	916
32	819	913	880	916	866	918	865	914
40	628	916	854	909	855	913	870	913
48	871	922	834	910	857	919	868	918

IO Performance (32.5 ms RTT)

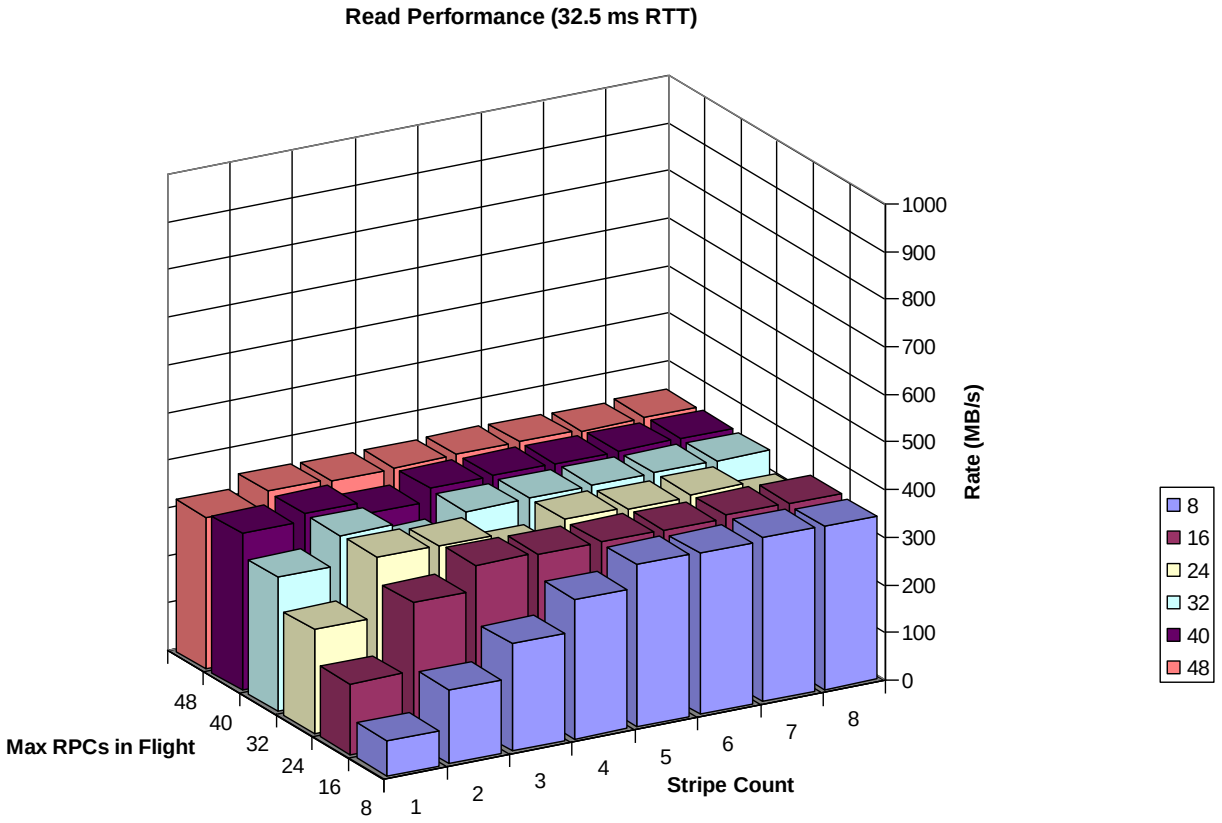


Best case performance Estimate:

- ib_write_bw –q 3 –s 1048576
  - 925.21 MB/s

	1	2	3	4	5	6	7	8
8	110	215	312	405	492	583	648	734
16	184	382	538	691	778	837	824	859
24	229	472	649	815	835	856	858	859
32	256	518	717	843	839	868	856	858
40	261	516	759	834	844	865	848	862
48	254	525	762	863	849	861	847	860

IO Performance (32.5 ms RTT)

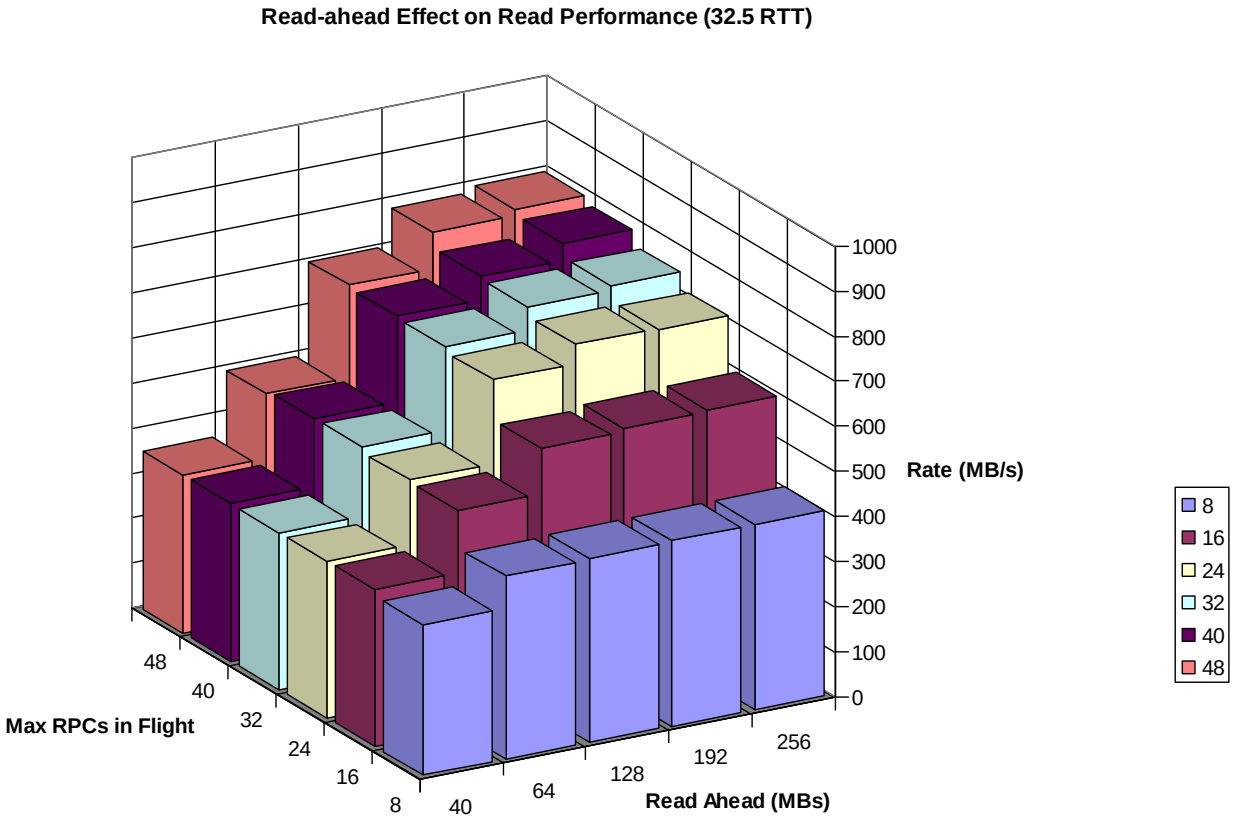


Best case performance Estimate:

- ib_write_bw –q 3 –s 1048576
  - 925.21 MB/s

	1	2	3	4	5	6	7	8
8	74	155	226	293	340	338	346	342
16	149	293	346	343	343	342	347	346
24	218	344	341	301	345	341	346	305
32	283	345	293	343	346	346	346	346
40	330	345	320	347	347	346	348	347
48	318	347	343	345	347	347	344	347

IO Performance (32.5 ms RTT)

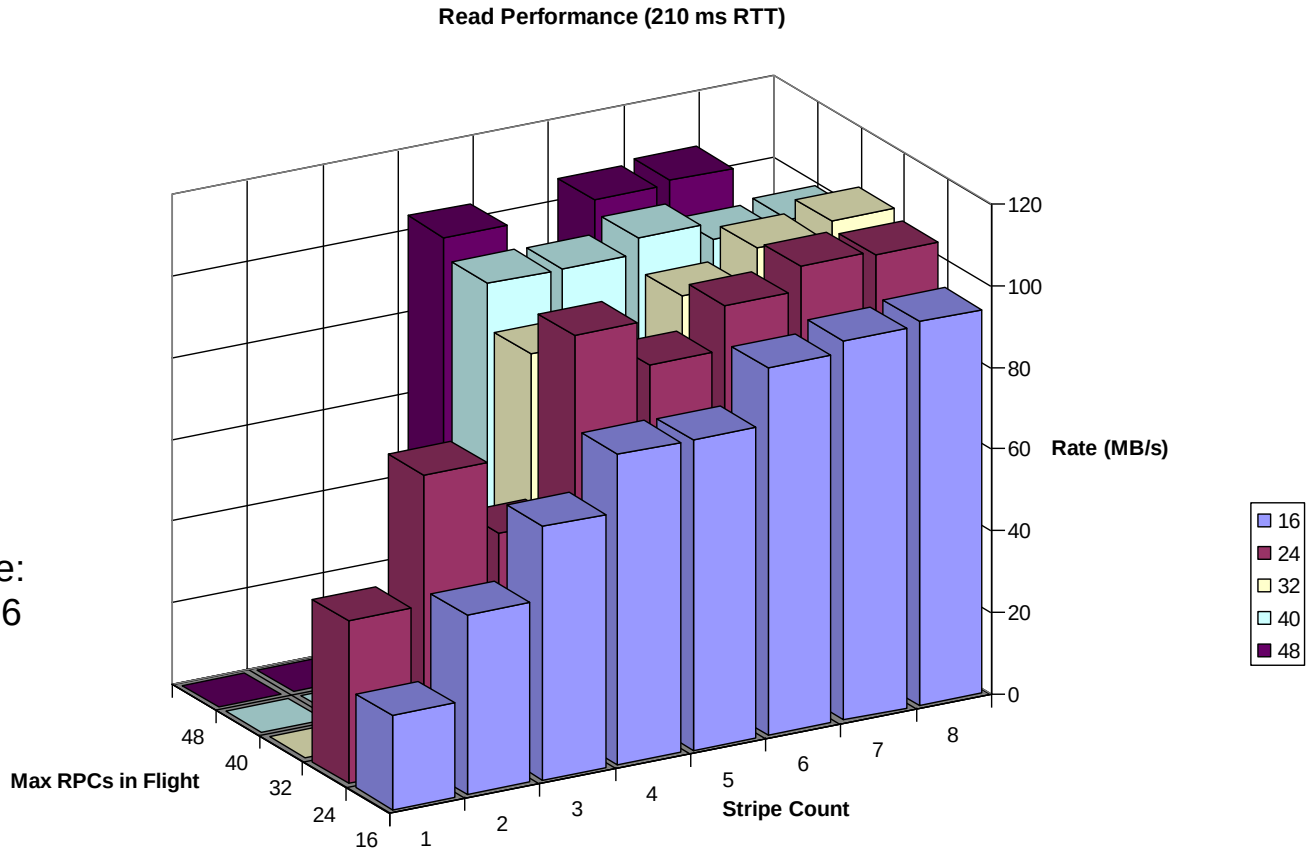


Best case performance Estimate:

- ib_write_bw –q 3 –s 1048576
  - 925.21 MB/s

	40	64	128	192	256
8	334	405	409	411	411
16	349	489	589	598	601
24	348	492	678	720	716
32	349	502	689	741	750
40	352	502	693	745	781
48	351	494	699	781	795

IO Performance (210 ms RTT)

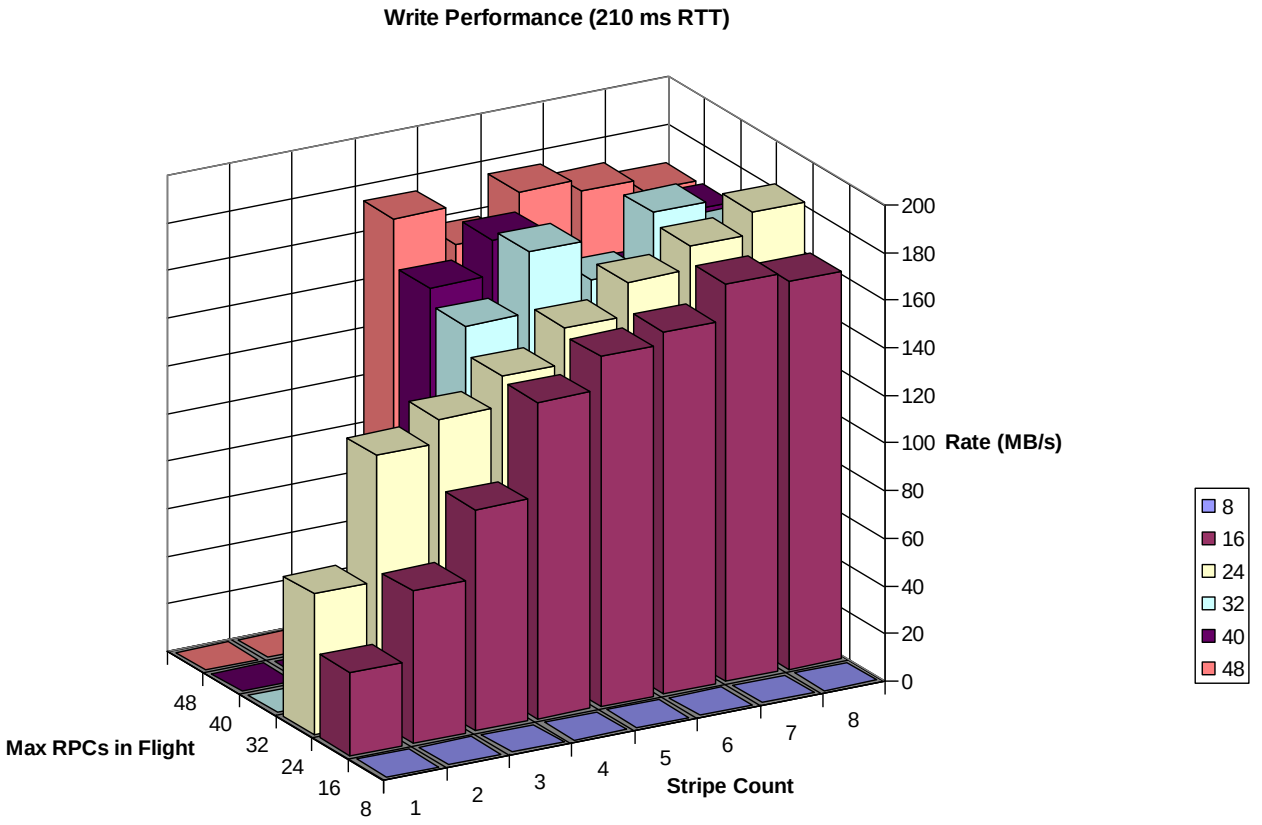


Best case performance Estimate:

- ib_write_bw –q 3 –s 1048576
  - 199.71 MB/s

	1	2	3	4	5	6	7	8
16	23	44	62	76	76	90	93	94
24	40	72	54	99	88	99	105	104
32	0	0	0	88	67	95	103	106
40	0	0	0	99	99	103	99	100
48	0	0	0	104	83	106	107	74

IO Performance (210 ms RTT)



Best case performance Estimate:

- ib_write_bw –q 3 –s 1048576
  - 199.71 MB/s

	1	2	3	4	5	6	7	8
8	0	0	0	0	0	0	0	0
16	35	64	93	133	147	152	167	163
24	59	112	122	135	150	164	174	183
32	0	0	0	147	173	156	179	165
40	0	0	0	154	169	130	143	165
48	0	0	0	174	158	175	170	165

Lustre 1.8 Metadata Performance

Today with Lustre 1.8.1

- Statahead exists to prefetch file info
- Doesn't grab file size from OSTs so statahead doesn't help much over WAN, RTTs for each stat (even if each OST is done in parallel)
- Time for "ls –l" on remote directory with 100k files (1 stripe)
  - Local: 16.7 seconds
  - 2.5 ms Latency: 286 seconds
  - 35 ms Latency: 3528 seconds
  - 210 ms Latency: Untested
    - Expected best case: ~6 hours! (100000 / (1 / .210))

Lustre 1.8 Metadata Performance

Today with Lustre 1.8.1

- “time /bin/ls -U > /dev/null" (no stats, just readdir())
  - 100k files
    - Local: .316 seconds
    - 2.3 ms RTT: 3.717 seconds
    - 32.790 ms RTT: 46.860 seconds
  - 7 million files
    - Local: 40.099 seconds
    - 2.3 ms RTT: 419.82 seconds

Lustre 2.0 Metadata Performance

Testing with Lustre 2.0 Alpha 5 and SOM

- Time for “ls –l” on remote directory with 100k files (1 stripe)
  - Local: 4.5 seconds
    - 1/4th of no SOM
  - 2.5 ms Latency: 38 seconds
    - 1/7th of no SOM
  - 35 ms Latency: 358 seconds
    - 1/10th of no SOM
  - 210 ms Latency: 1699 seconds

Improvements for Lustre

Things we'd like to see improved

- Read-ahead, design doc on bug 20294 is a good start
- Config file or sysctl interface to persistently set file system tuning parameters on a per client basis
  - max_rpcs_in_flight, max_dirty_mb, max_read_ahead_mb
- Root squash NID lists (in 2.0) and all squash with NID list for Lustre 1.8
- Metadata performance over WAN improved (bug 18526)
  - readdirplus(), bulkstat, etc
  - Multiple create RPCs simultaneously from client
  - Large RPCs for readdir() (bug 17833)

SMSi

Questions?