Lustre 1.8.9 - 2.x Client Performance Comparison and Tuning

John Fragalla
HPC Principal Architect
Xyratex, A Seagate Company

Lustre User Group Conference
April 8-10, 2014

xyratex
A Seagate Company

Agenda

- Benchmark Setup
- IOR Parameters and Settings
- Lustre Clients and Methodology
- Single Thread Performance
- Throughput Performance Results and Data
- Summary

Benchmark Setup

- Storage Architecture
  - A Cluster 6000 SSU with GridRaid OSTs
    - Rated Storage Performance ≥ 6 GB/s Read or Write with IOR
  - InfiniBand FDR Interconnect
  - Xyratex Lustre 2.1.0.x4-74
- Client Hardware
  - 8 Clients, each configured with QDR IB, 48GB Memory and 12 Cores
  - Scientific Linux 6.5 with Stock OFED

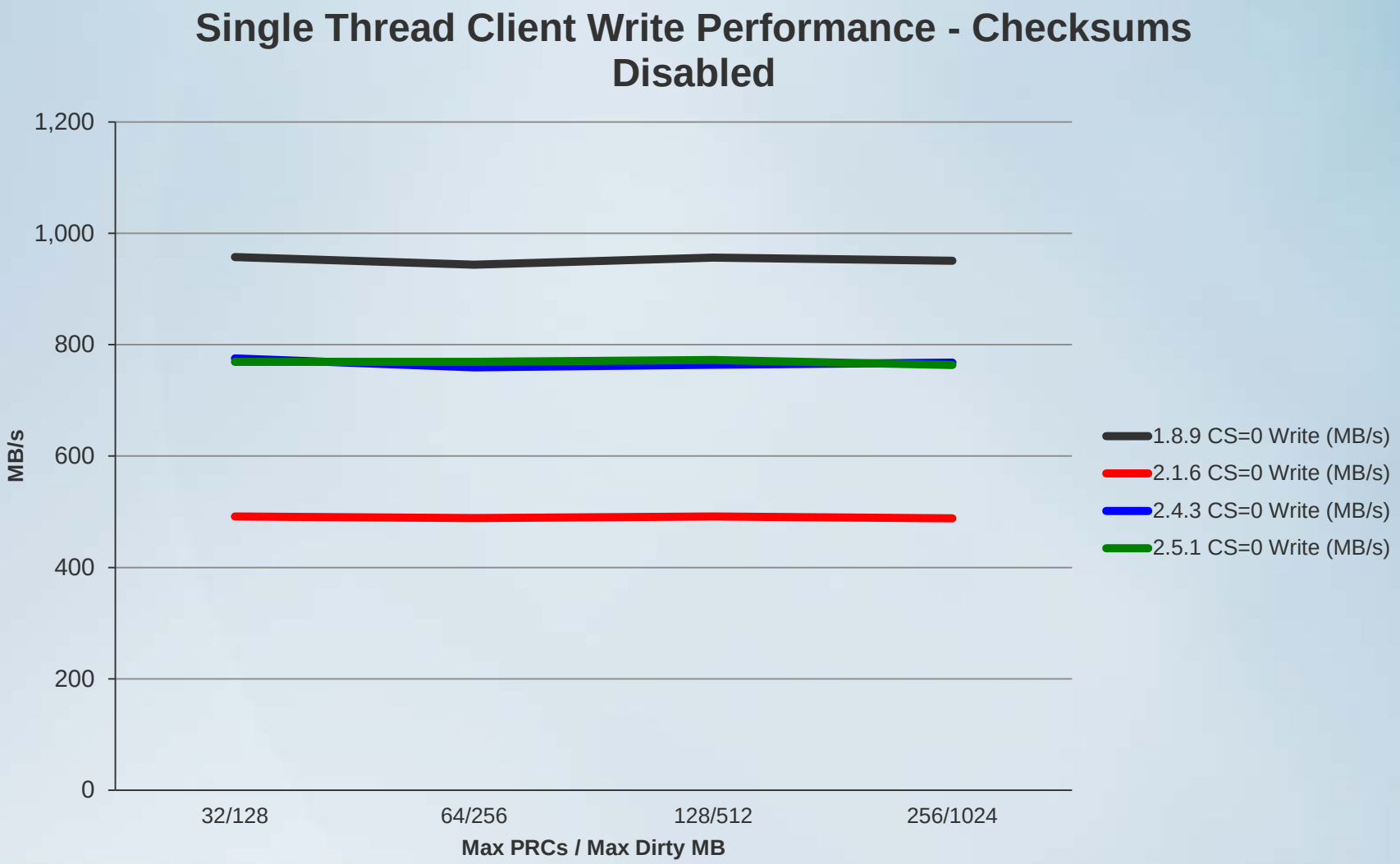
IOR Parameters and Settings

- Used mpirun to execute IOR with --byslot distribution
- IOR Parameters that were constant:
  - -F : File Per Process
  - -B : Direct IO Operation
  - -t 64m: 64m Transfer Size per Task
  - -b: 1024g for Single Thread and 512g for multiple tasks
  - -D: stonewall option, write for 4 minutes, and read for 2 minutes
- Lustre Settings
  - Stripe Count of 1
  - Stripe Size of 1m

Lustre Clients

- Compared the following clients:
  - 1.8.9, 2.1.6, 2.4.3, and 2.5.1
- Collected raw performance data using the following client settings with and without checksums enabled
  - max_rpcs_in flight / max_dirty_mb
    - 32 / 128
    - 64 / 256
    - 128 / 512
    - 256 / 1024

Single Thread Write Performance



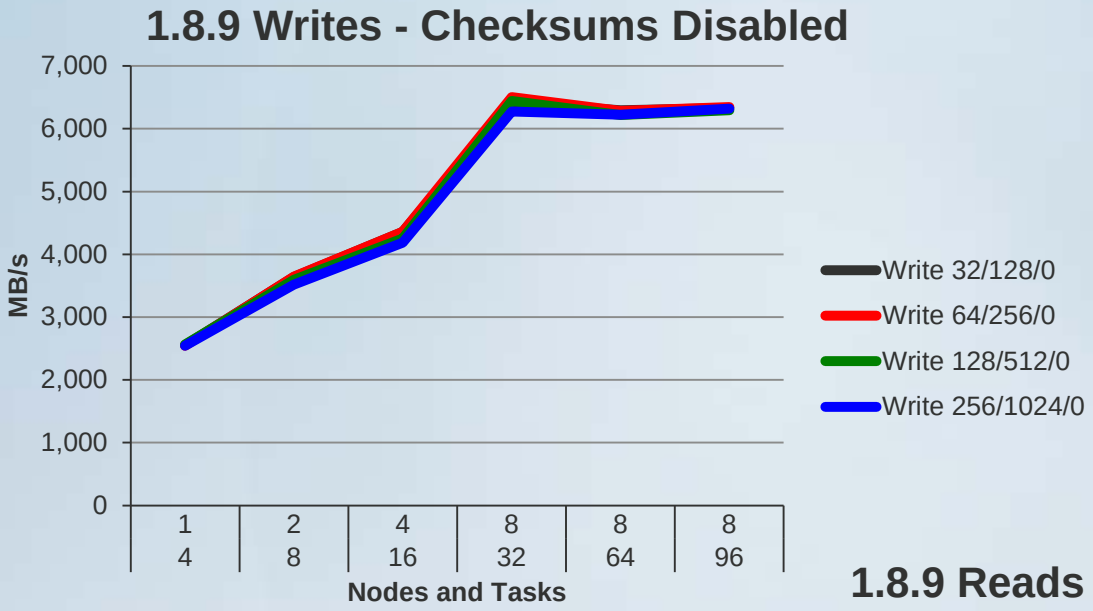
Single Thread Read Performance

xyratex
A Seagate Company

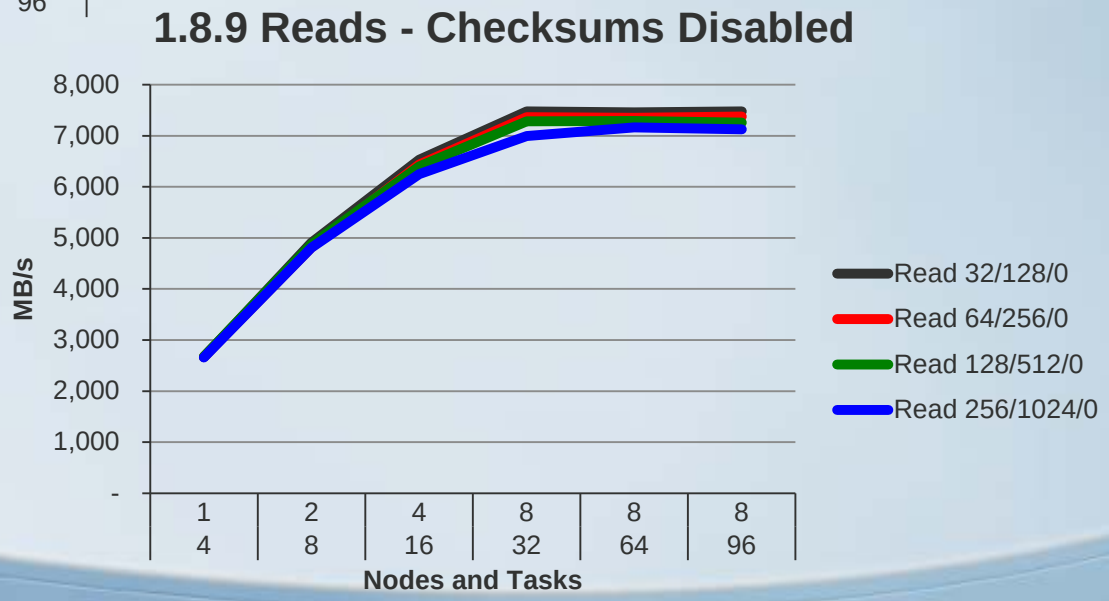
Impact on Single Thread Performance when Checksums were Enabled

	Checksums Impact - Single Thread		
Version		Reads	Writes
1.8.9	Performance Impact Difference (MB/s)	217.80	336.84
	Percentage Impact Reduced	19%	35%
2.1.6	Performance Impact Difference (MB/s)	37.70	27.81
	Percentage Impact Reduced	3%	6%
2.4.3	Performance Impact Difference (MB/s)	45.51	9.14
	Percentage Impact Reduced	6%	1%
2.5.1	Performance Impact Difference (MB/s)	44.91	6.50
	Percentage Impact Reduced	6%	1%

1.8.9 Client Performance Results

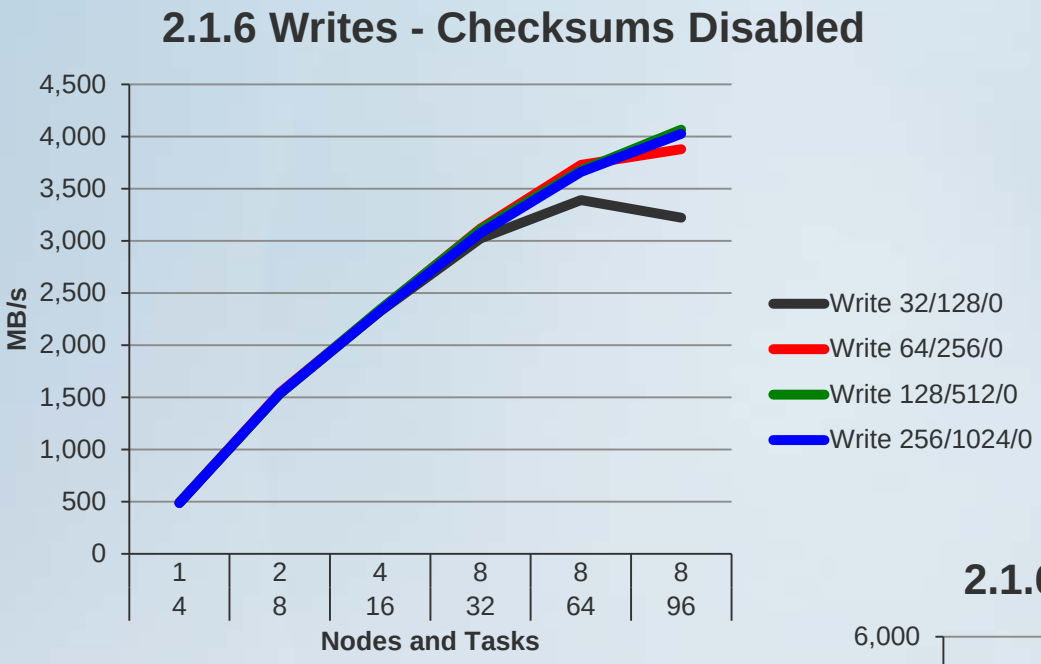


Key:
max_rpcs_in_flight / max_dirty_mb / checksums

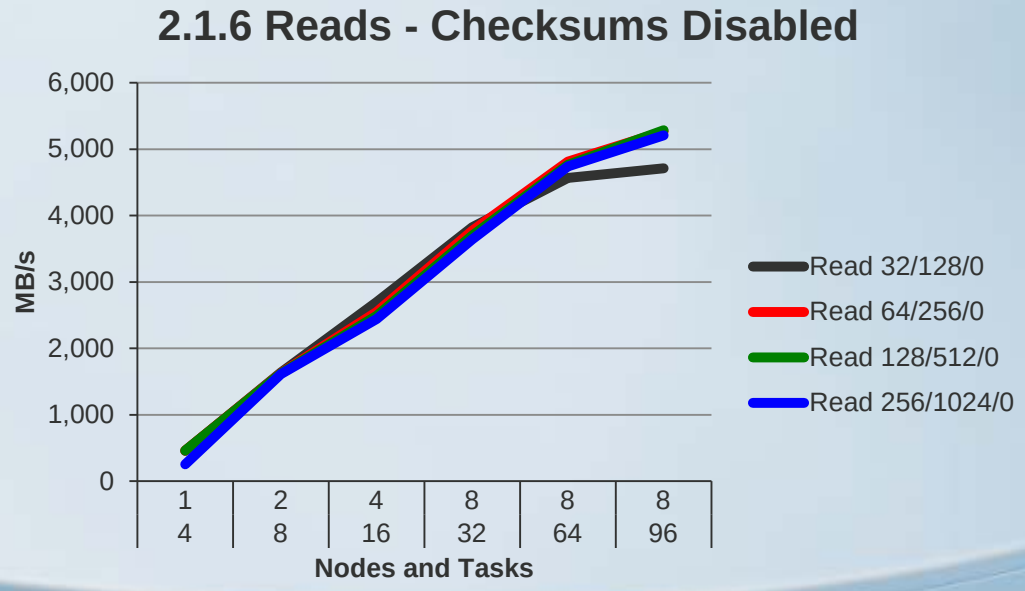


xyratex
A Seagate Company

2.1.6 Client Performance Results

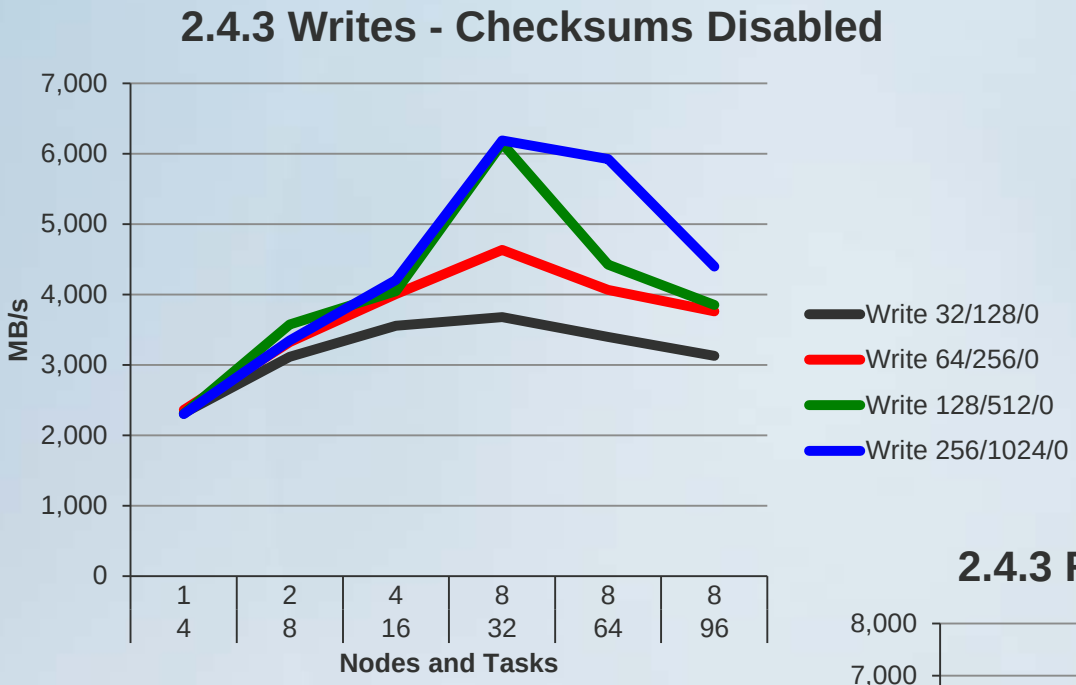


Key:
max_rpcs_in_flight / max_dirty_mb / checksums

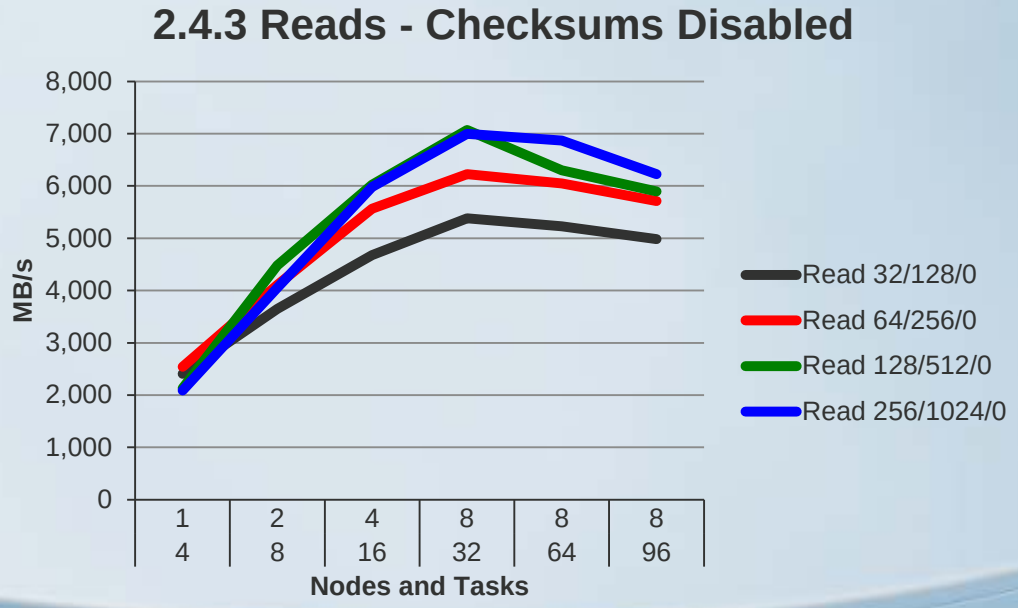


xyratex
A Seagate Company

2.4.3 Client Performance Results

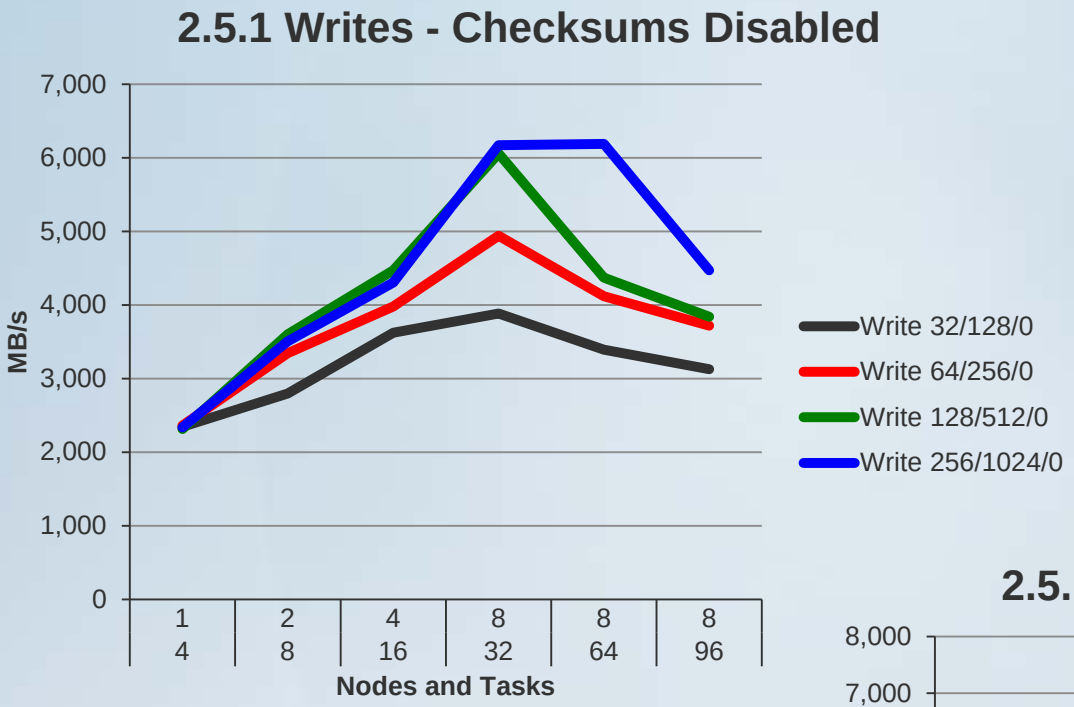


Key:
max_rpcs_in_flight / max_dirty_mb / checksums

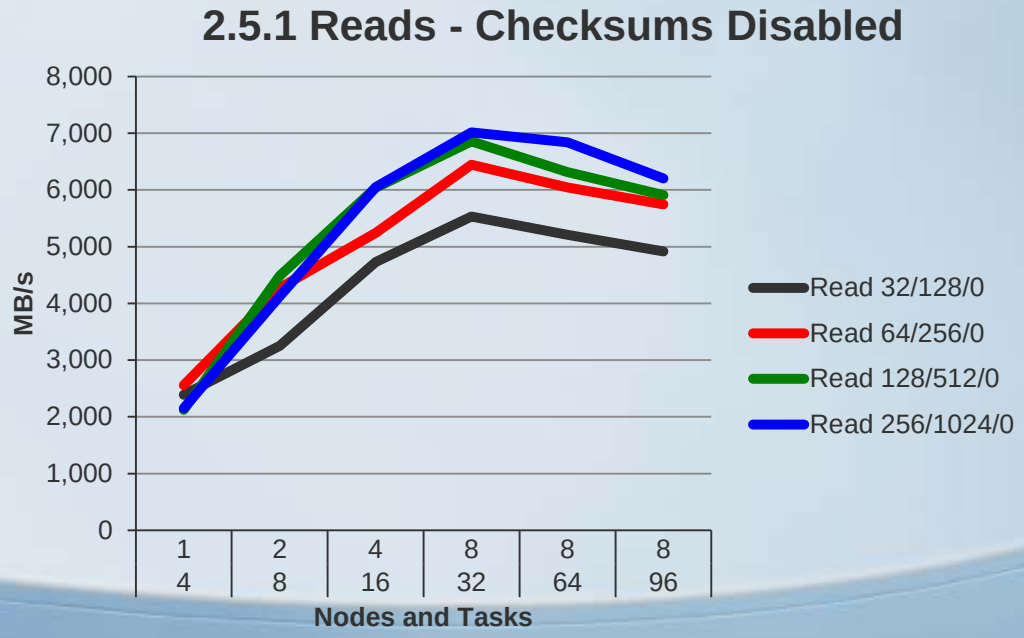


xyratex
A Seagate Company

2.5.1 Client Performance Results



Key:
max_rpcs_in_flight / max_dirty_mb / checksums



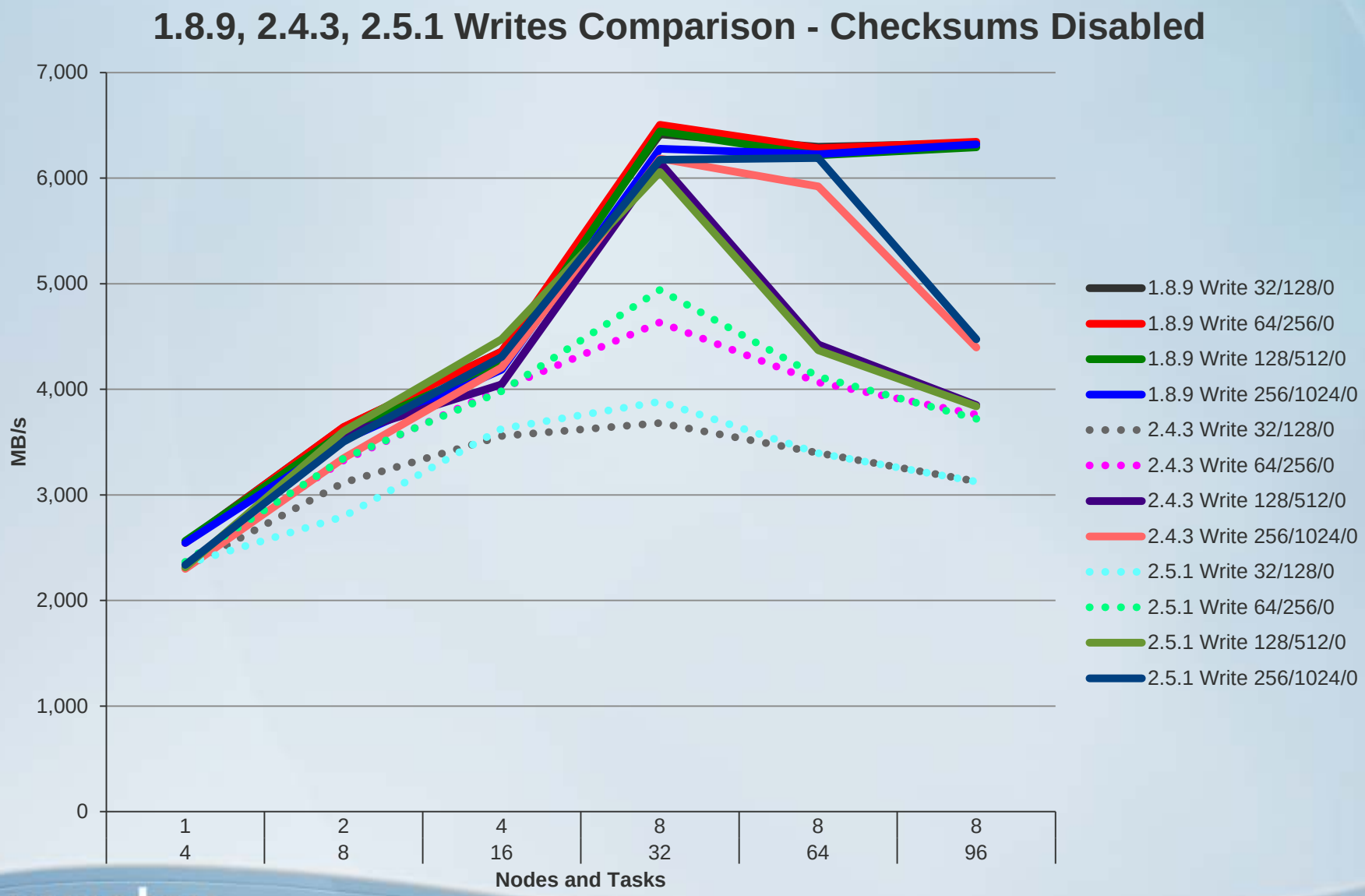
xyratex
A Seagate Company

Average Negative Impact on Performance when Checksums Enabled

Average results using 4-96 Threads across 1-8 Clients

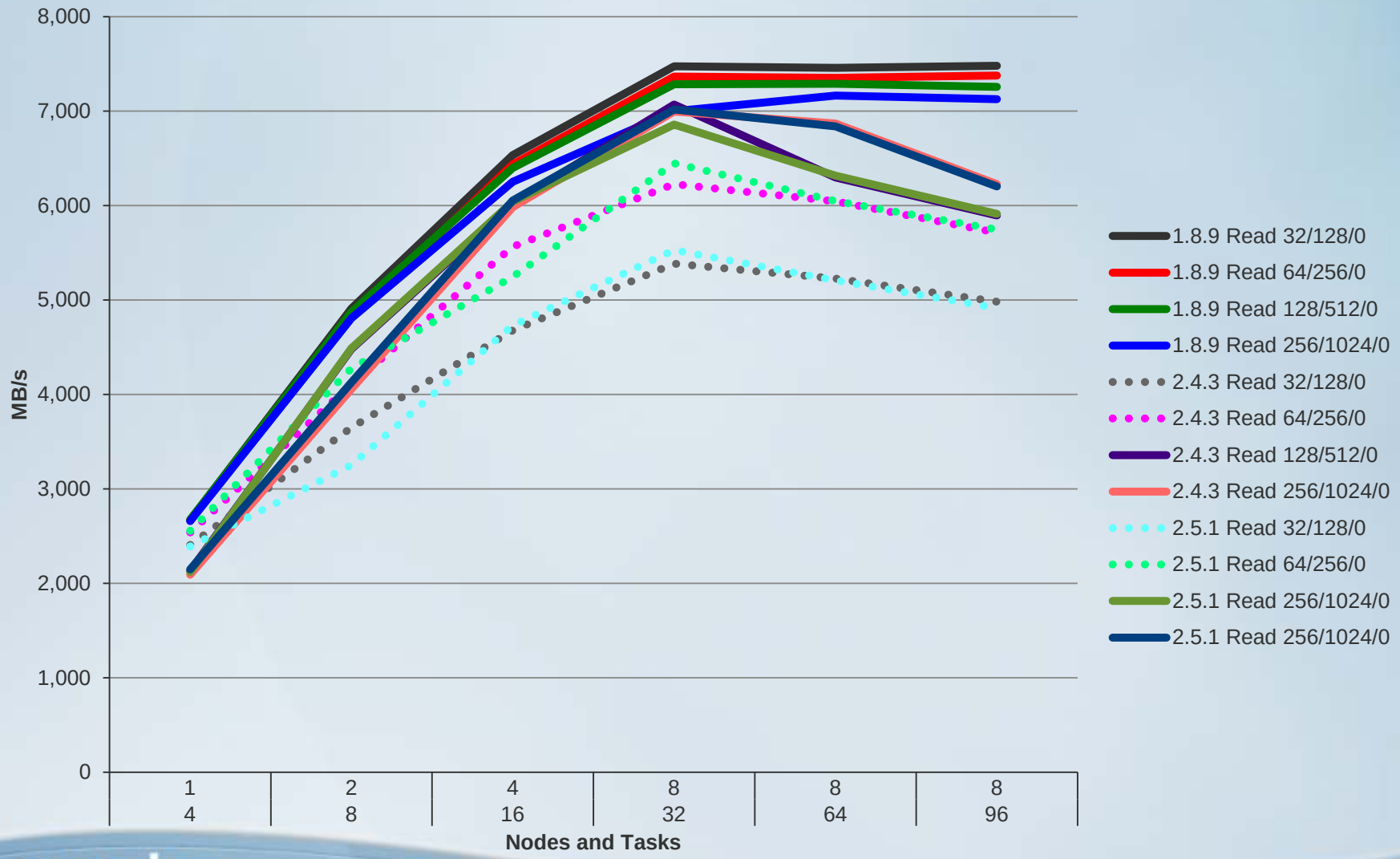
	Checksums Impact (max_rpcs_in_flight/max_dirty_mb)								
		Reads 32/128	Write 32/128	Read 64/256	Write 65/256	Read 128/512	Write 128/512	Read 256/1024	Write 256/1024
1.8.9	Average Impact Difference (MB/s)	203.40	398.09	3.12	434.41	71.03	523.17	31.19	502.31
	Average Impact Percentage	2%	10%	-1%	10%	0%	13%	0%	12%
2.1.6	Average Impact Difference (MB/s)	541.64	269.93	446.17	196.06	445.79	189.33	453.34	180.07
	Average Impact Percentage	22%	12%	19%	9%	18%	8%	9%	8%
2.4.3	Average Impact Difference (MB/s)	333.18	94.53	277.32	77.16	380.47	194.21	326.76	315.71
	Average Impact Percentage	7%	3%	5%	2%	7%	4%	5%	6%
2.5.1	Average Impact Difference (MB/s)	486.11	161.09	245.24	154.50	366.05	256.85	346.78	423.12
	Average Impact Percentage	12%	5%	5%	4%	6%	6%	6%	9%

1.8.9, 2.4.3, 2.5.1 Overall Write Comparison



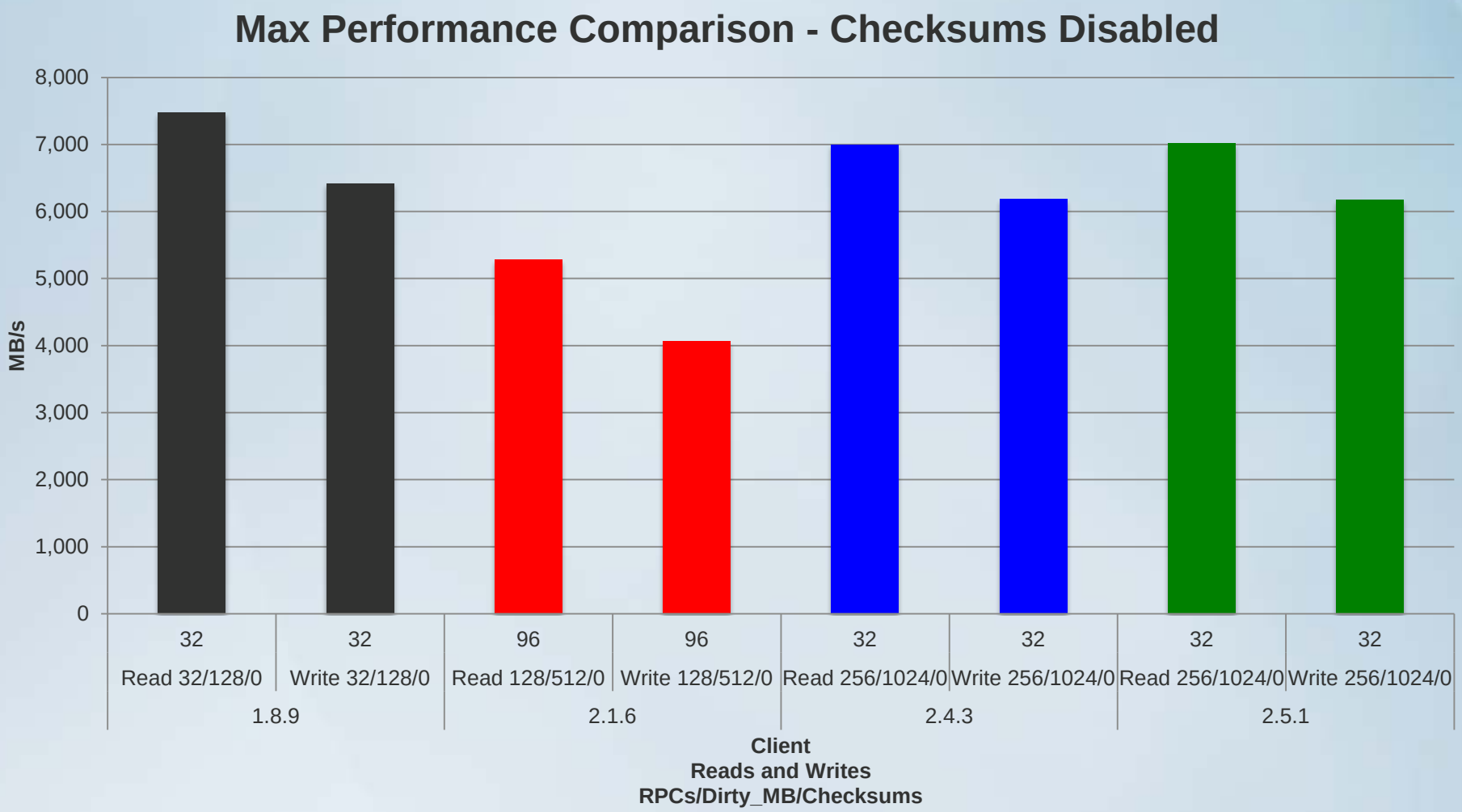
xyratex
A Seagate Company

1.8.9, 2.4.3, 2.5.1 Overall Read Comparison



xyratex
A Seagate Company

Maximizing Performance for Each Client



xyratex
A Seagate Company

Interesting Results Analyzing Raw Data

- In general, 1.8.9 Client performed the same regardless of client settings
- 2.4.3 and 2.5.1 followed the same performance curve
  - Clients settings need to be increased to achieve maximum storage throughput
  - Both max_rpcs_in_flight and max_dirty_mb need to be increased to at lest 256
    - Anything less than 256 will result in less than optimal Storage performance
- The rule of thumb: max_dirty_mb = max_rpcs_in_flight * 4 is not holding true with Client versions 2.4.3 and 2.5.1

Summary

- 1.8.9 single thread performance results are the highest, but 2.4.3 and 2.5.1 improved over previous 2.x client versions
- With the right client settings, 2.4.3 and 2.5.1 client versions can maximize storage throughput, along with 1.8.9 clients
- 2.1.6 Client underperformed, regardless of client tuning
- Checksums impact varied with the number of threads, but, on average, not a “big” performance impact
  - Biggest impact on performance with Checksums enabled was on the Single Thread tests
- 2.4.3 and 2.5.1 performance results were similar across all threads and client tuning parameters

Thank You

John_Fragalla@xyratex.com