xyratex

Advancing Digital Storage Innovation

Map/Reduce on Lustre

Hadoop Performance in HPC Environments

Nathan Rutman, Xyratex

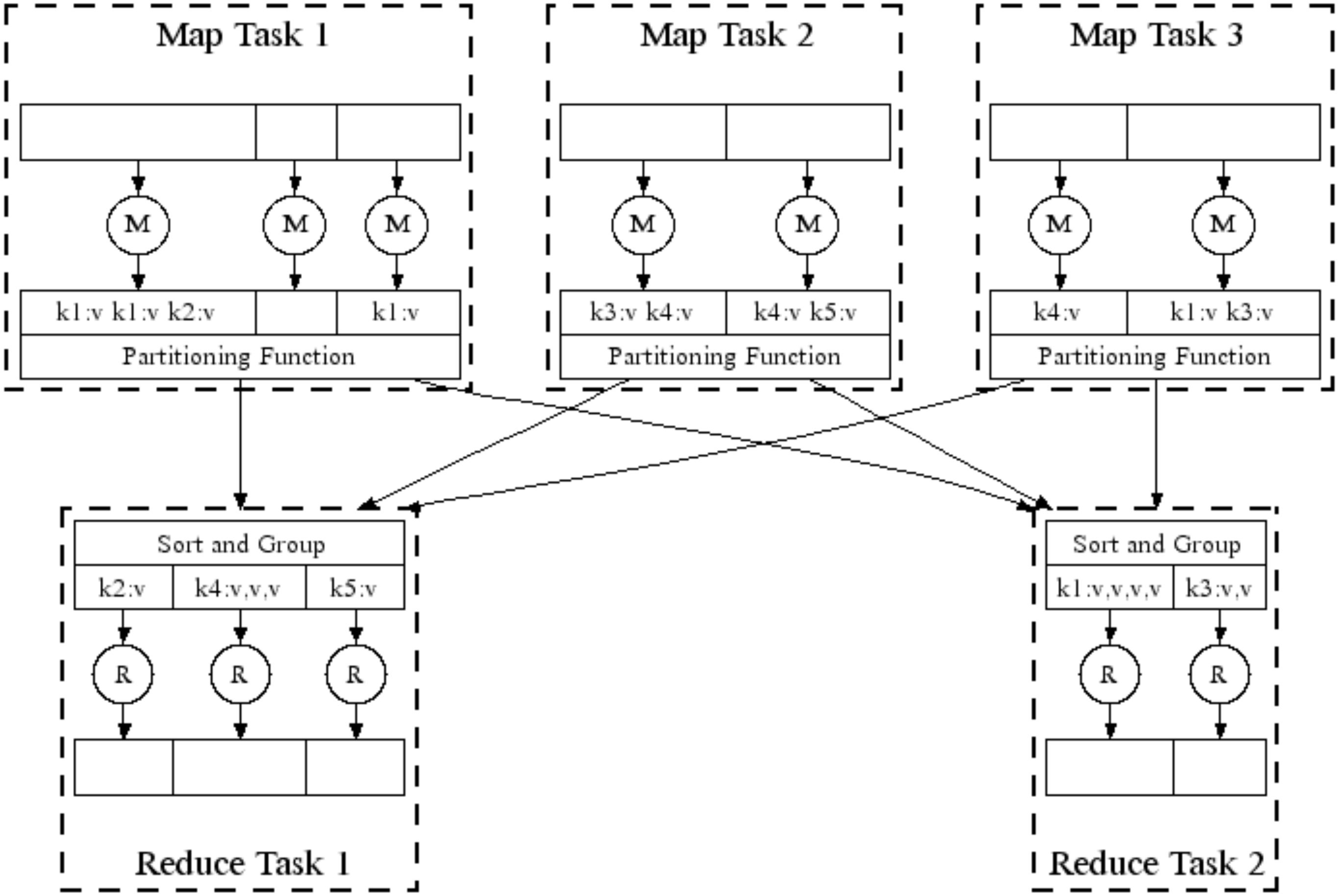
James B. Hofmann, Naval Research Laboratory

Agenda

- Map Reduce Overview
- The Case for Moving Data
- A Combined Lustre / HDFS Cluster
- Theoretical Comparisons
- Benchmark Study
- The Effects of Tuning
- Cost Considerations

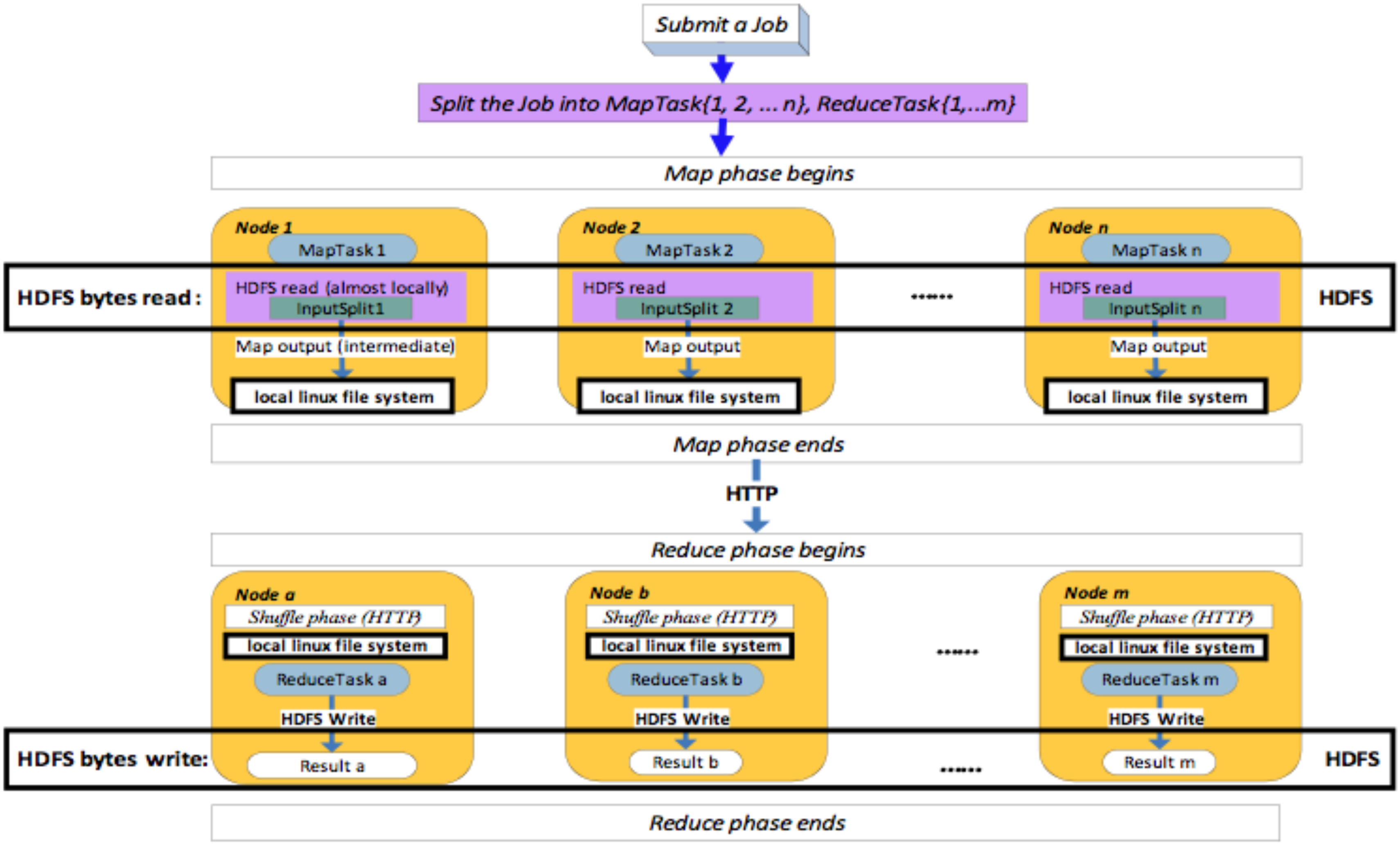
Map Reduce overview

Using Lustre with Apache Hadoop, *Sun Microsystems*



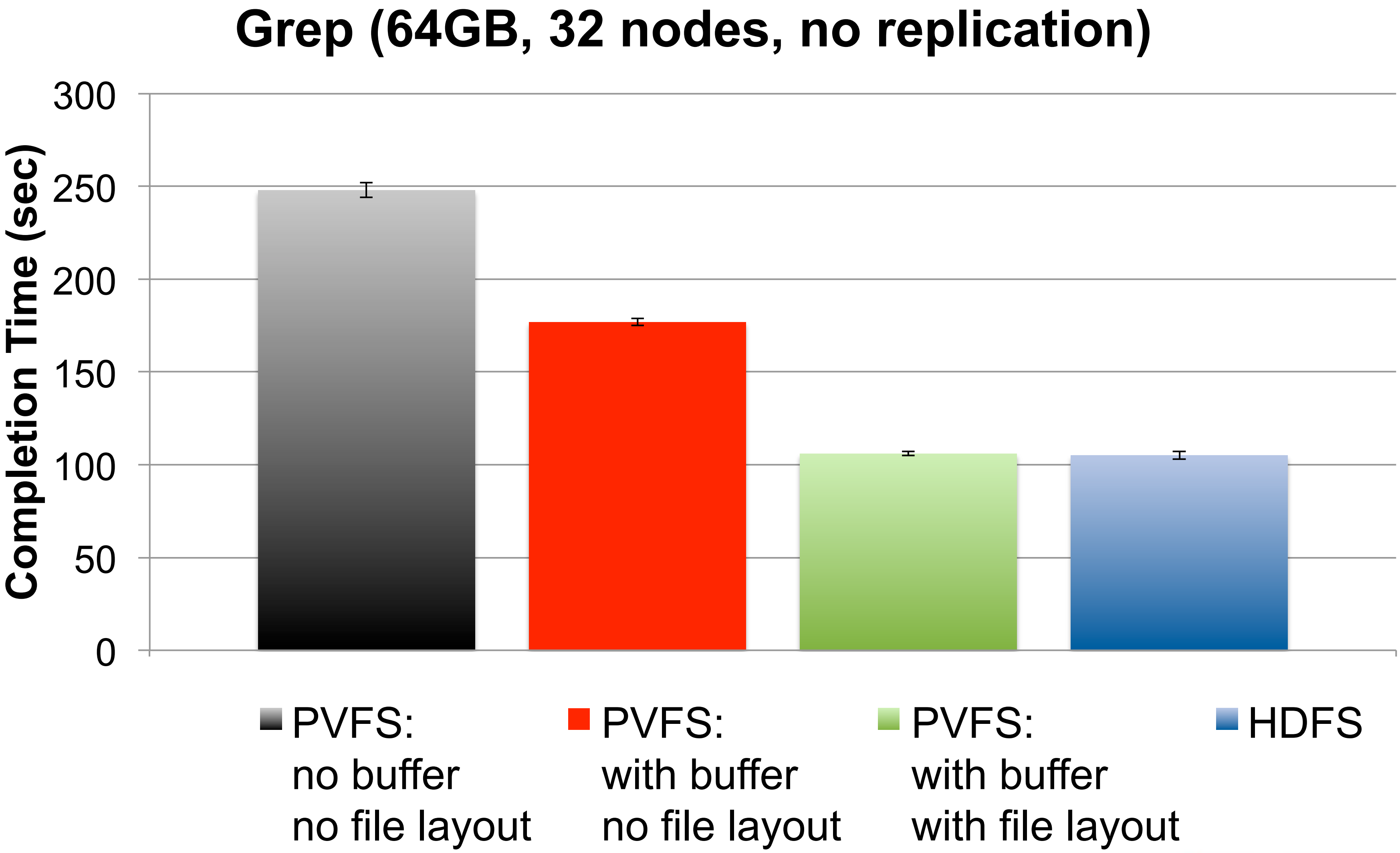
Apache Hadoop disk usage

Using Lustre with Aparche Hadoop, *Sun Microsystems*



Other Studies: Hadoop with PVFS

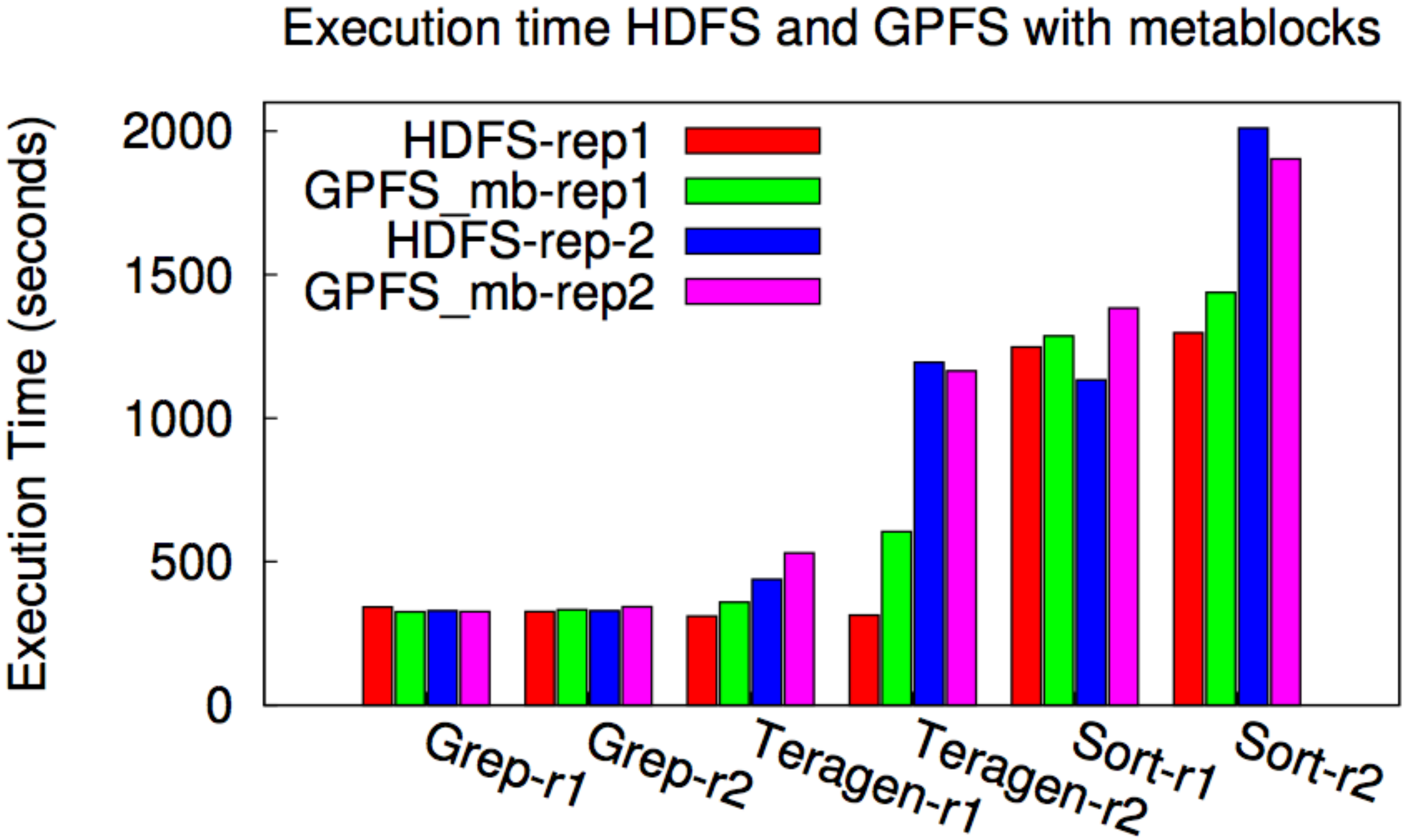
Crossing the Chasm: Sneaking a Parallel File System Into Hadoop , *Carnegie Mellon*



xyratex

Other Studies: Hadoop with GPFS

Cloud analytics: Do we really need to reinvent the storage stack? *IBM Research*



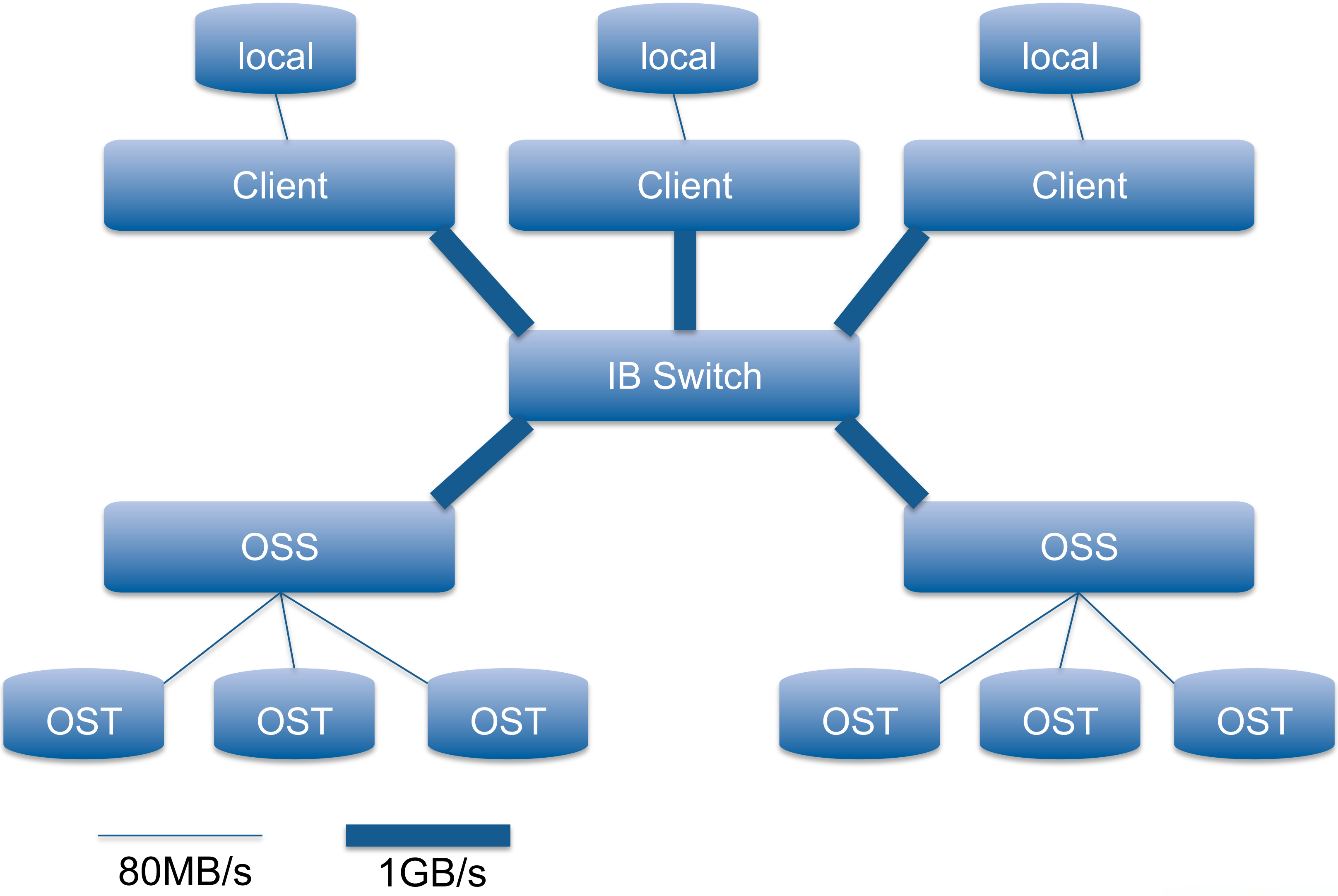
A Critical Oversight

- “Moving Computation is Cheaper Than Moving Data”
- The data ALWAYS has to be moved
  - Either from local disk
  - Or from the network
- And with a good network: the network wins.

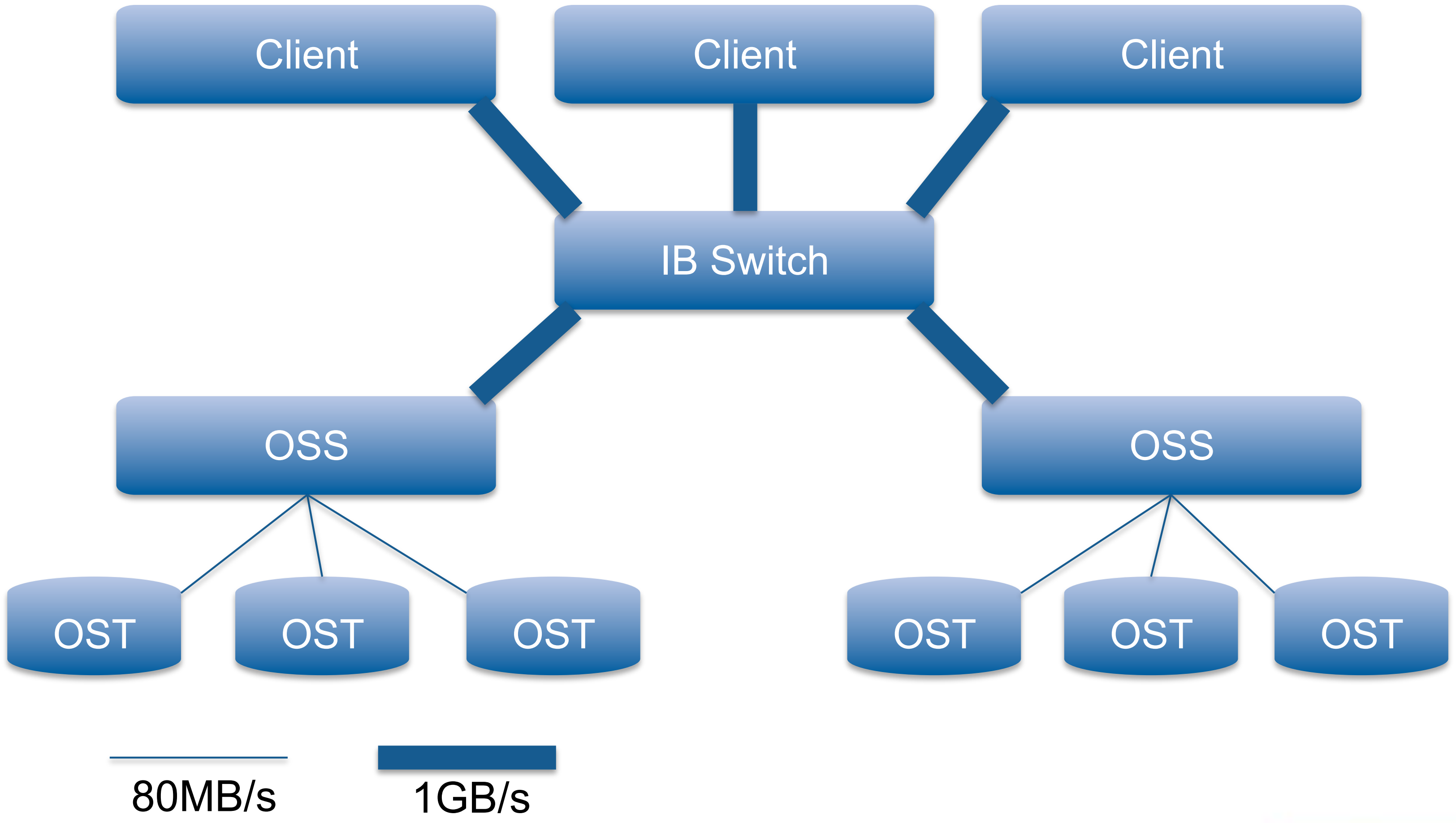
Cluster Setup: HDFS vs Lustre

- 100 clients, 100 disks, Infiniband
- Disks: 1 TB FATSAS drives (Seagate Barracuda)
  - 80 MB/sec bandwidth with cache off
- Network: 4xSDR Infiniband
  - 1GB/s
- HDFS: 1 drive per client
- Lustre: 10 OSSs with 10 OSTs

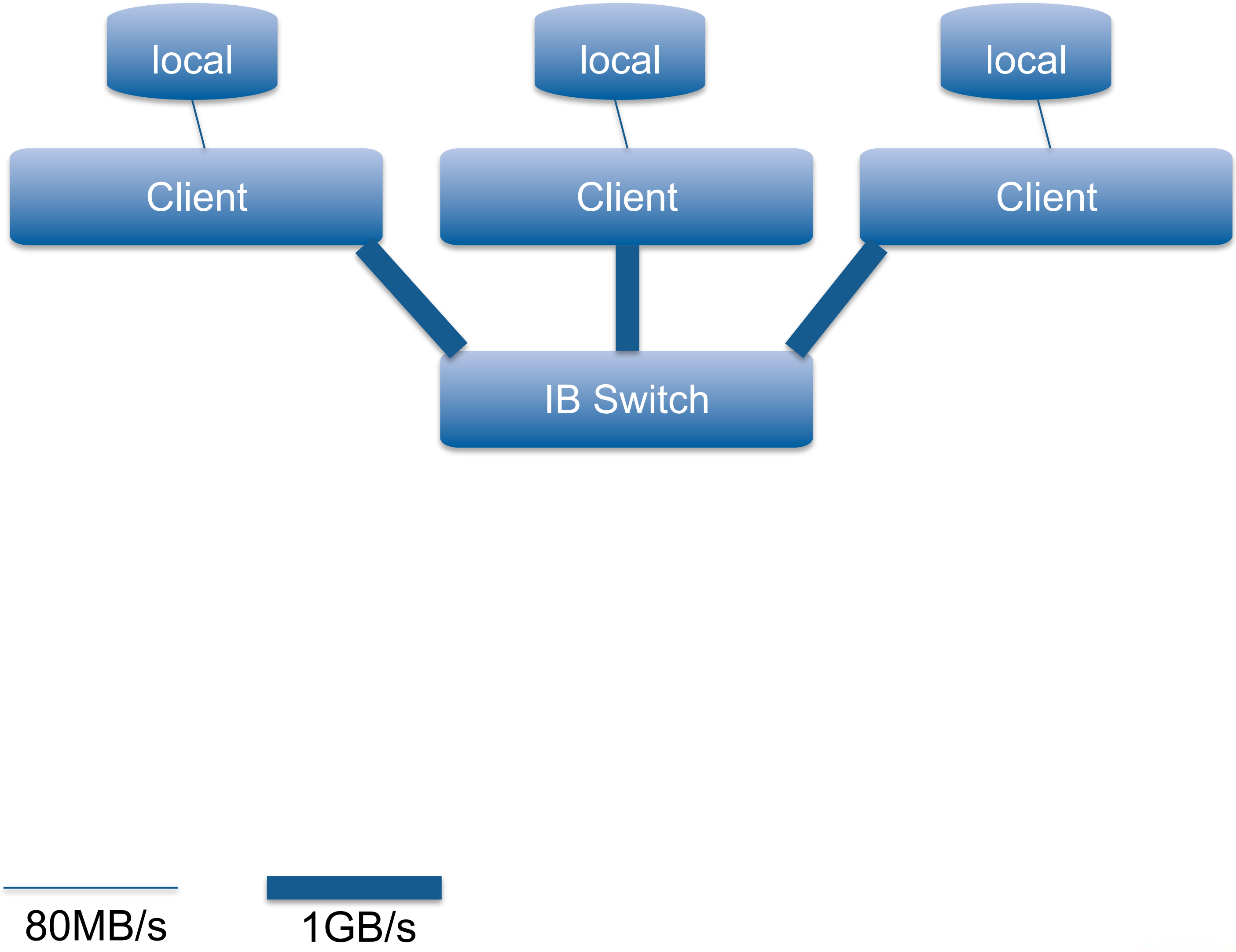
Cluster Setup



Lustre Setup



HDFS Setup


local
local
local
Client
Client
Client
IB Switch
80MB/s
1GB/s

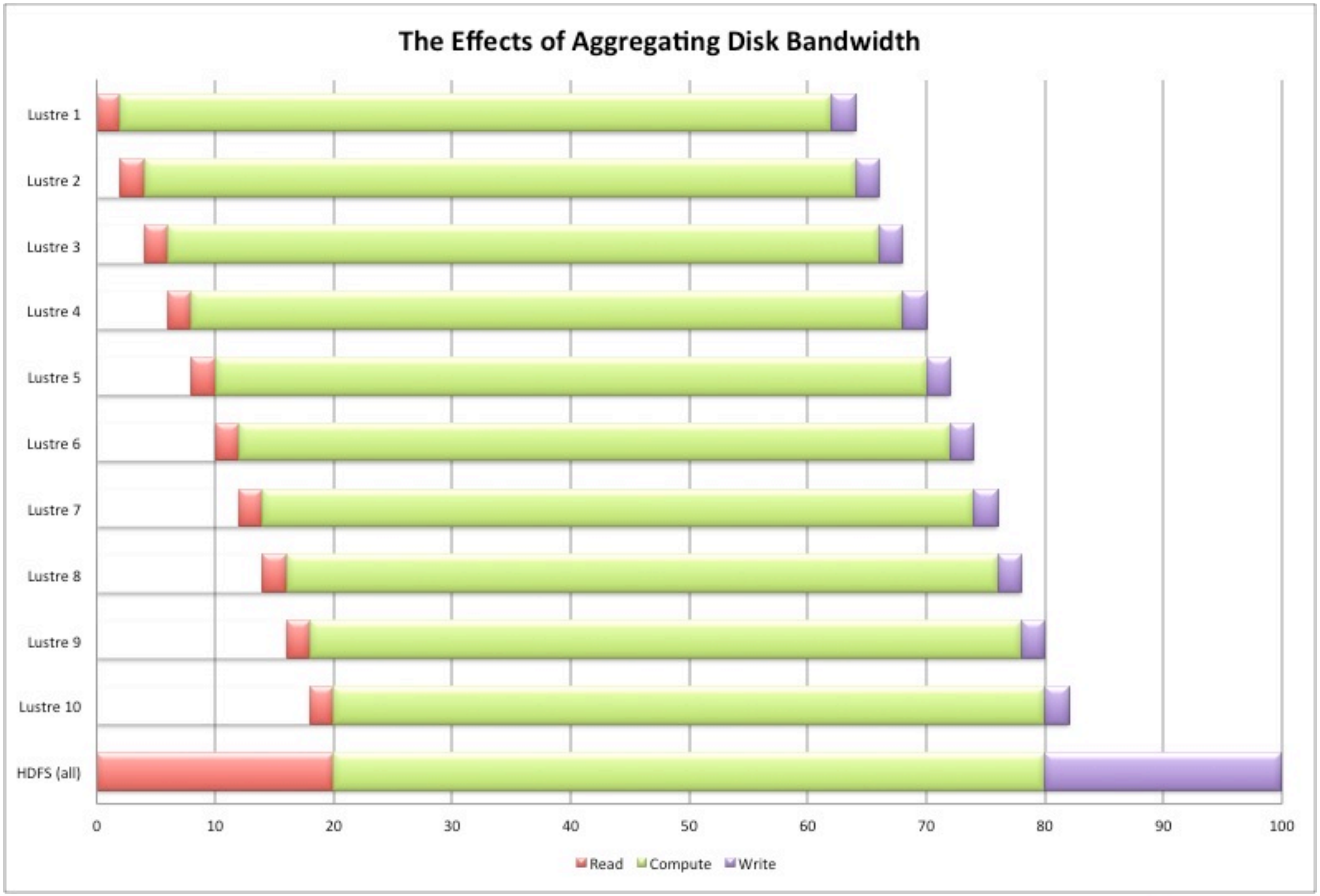
Theoretical Comparison: HDFS vs Lustre

- 100 clients, 100 disks, Infiniband
- HDFS: 1 drive per client
  - Capacity 100 TB
  - Disk bandwidth 8 GB/s aggregate (80MB/s * 100)
- Lustre: Each OSS has
  - Disk bandwidth 800MB/s aggregate (80MB/s * 10)
    - Assuming bus bandwidth to access all drives simultaneously
  - Net bandwidth 1GB/s (IB is point to point)
- With 10 OSSs, we have same the capacity and bandwidth
- Network is not the limiting factor!

Striping

- In terms of raw bandwidth, network does not limit data access rate
- Striping the data for each Hadoop data block, we can focus our bandwidth on delivering a single block
- HDFS limit, for any 1 node: 80MB/s
- Lustre limit, for any 1 node: 800MB/s
    - Assuming striping across 10 OSTs
    - Can deliver that to 10 nodes simultaneously
- Typical MR workload is not simultaneous access (after initial job kickoff)

Striping on MR jobs



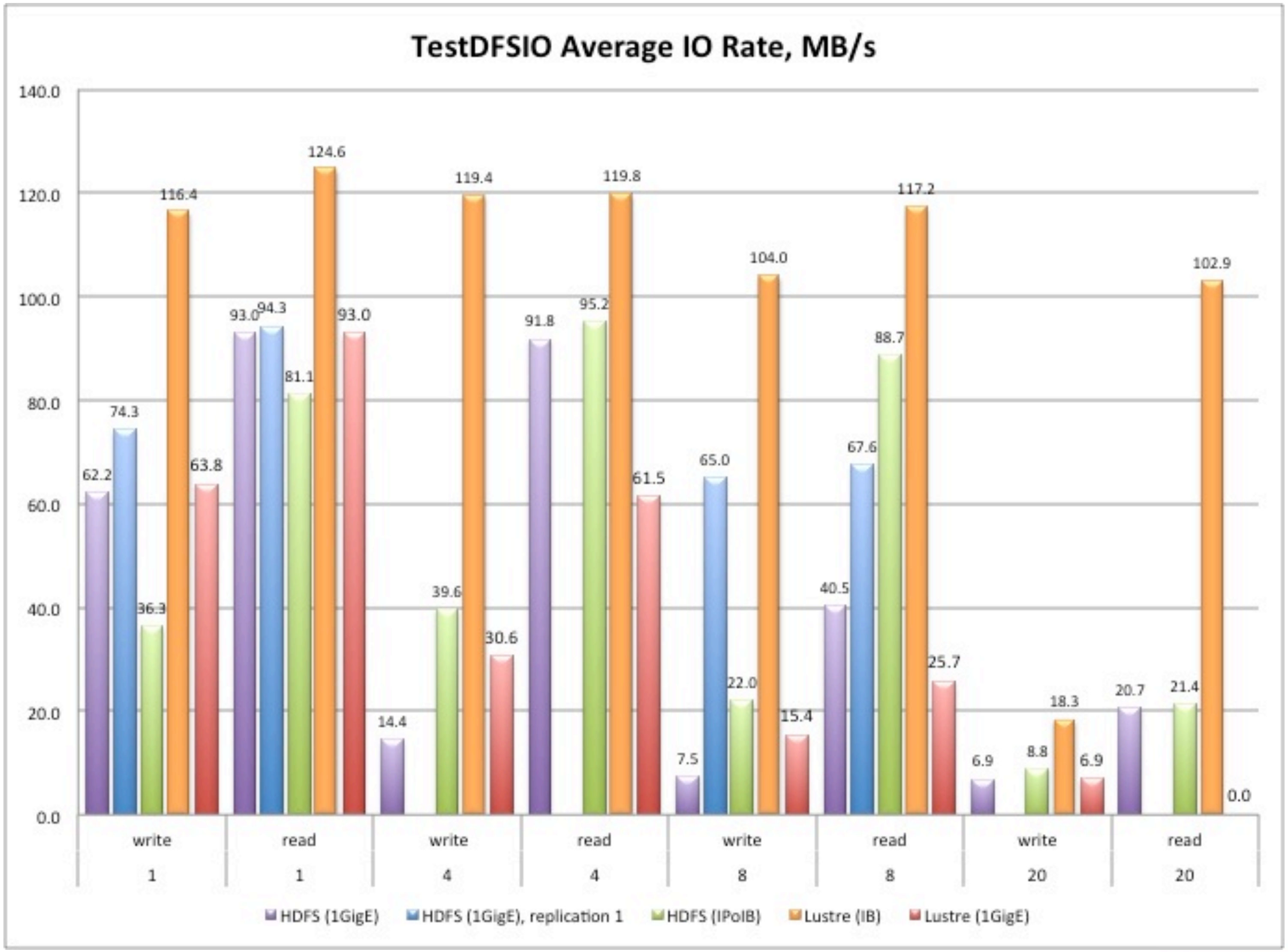
Replication

- HDFS replicates data 3x by default
- Recently Facebook added HDFS-RAID, which effectively trades off some computation (parity) for capacity
  - Can e.g. bring 3x safety for 2.2x storage cost when used
- Replicas should be done “far away”
- Replicas are synchronous
- HDFS writes are VERY expensive
  - 2 network hops, “far”
  - 3x storage
- Can trade off data safety for some performance

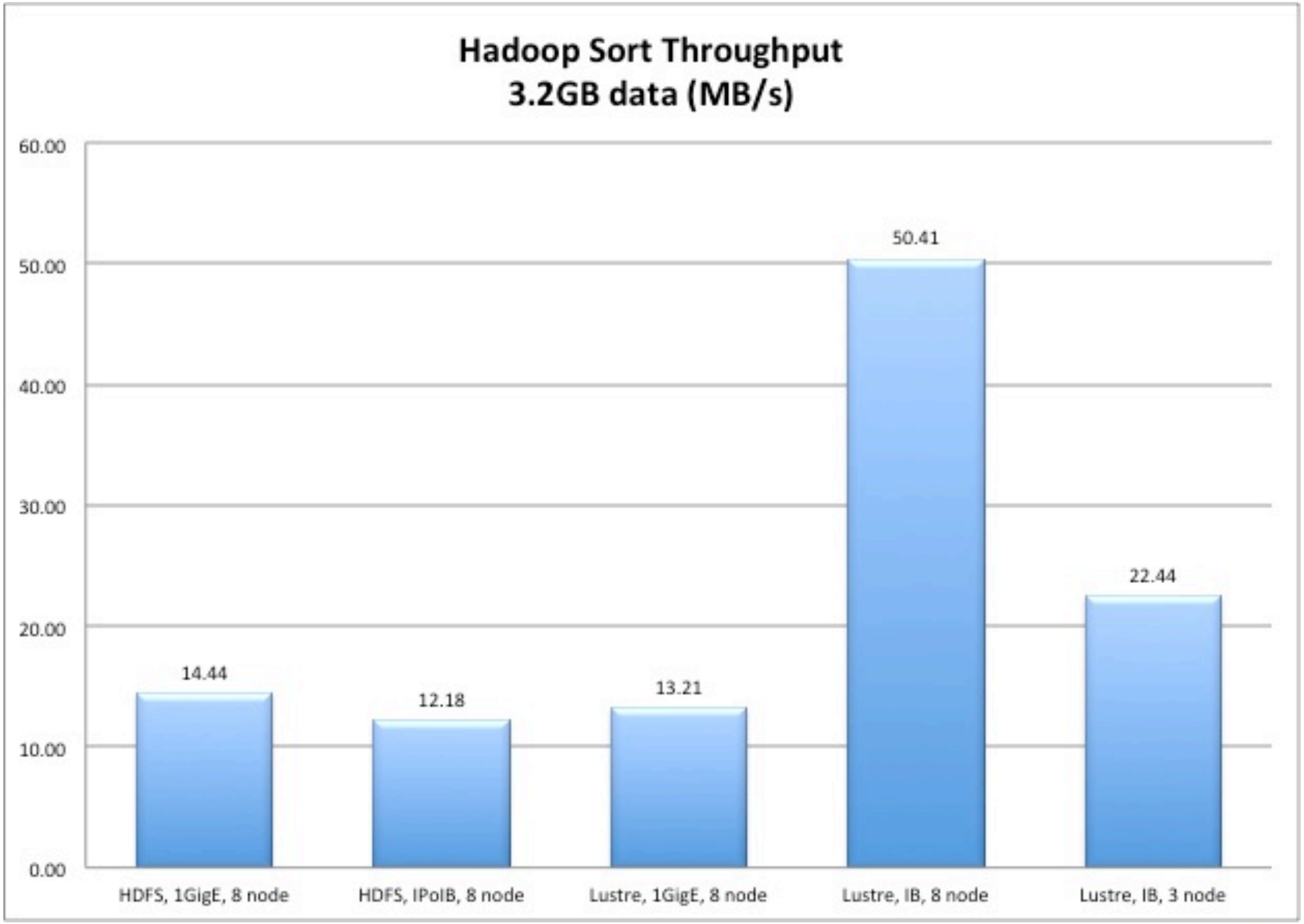
Data Locality

- HDFS reads are efficient ONLY on nodes that store data
  - Not network optimized (HTTP, no DIRECTIO, no DMA)
  - No striping = no aggregating drive bandwidth
  - 1GigE = 100MB/s = quick network saturation for non-local reads
  - Reduced replication = reduced node flexibility
- Lustre reads are equally efficient on any client node
  - Flexible number of map tasks
  - Arbitrary choice of mapper nodes
  - Better cluster utilization
- Lustre reads are fast
  - Striping aggregates disk bandwidth

MR I/O Benchmark

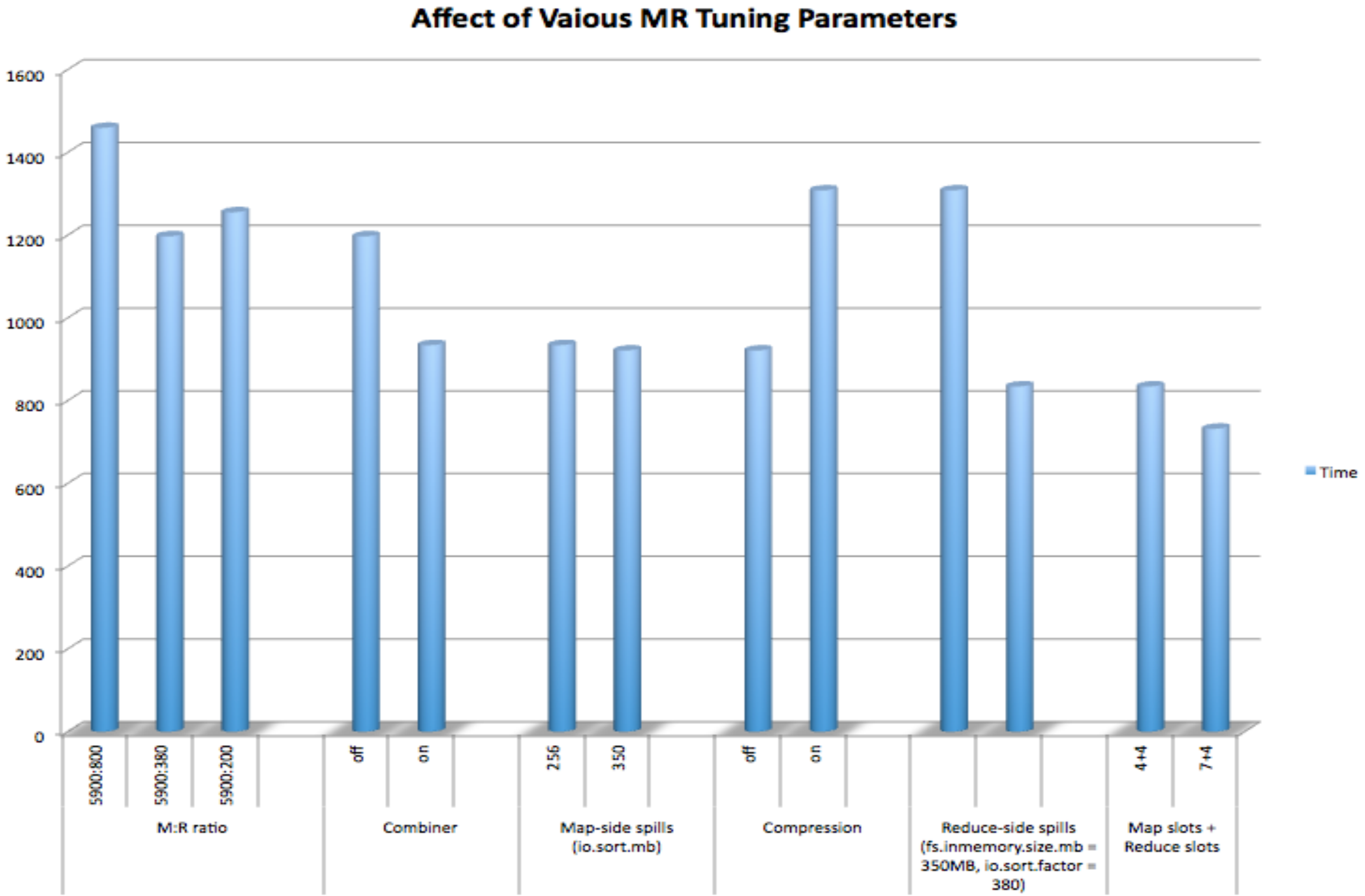


MR Sort Benchmark

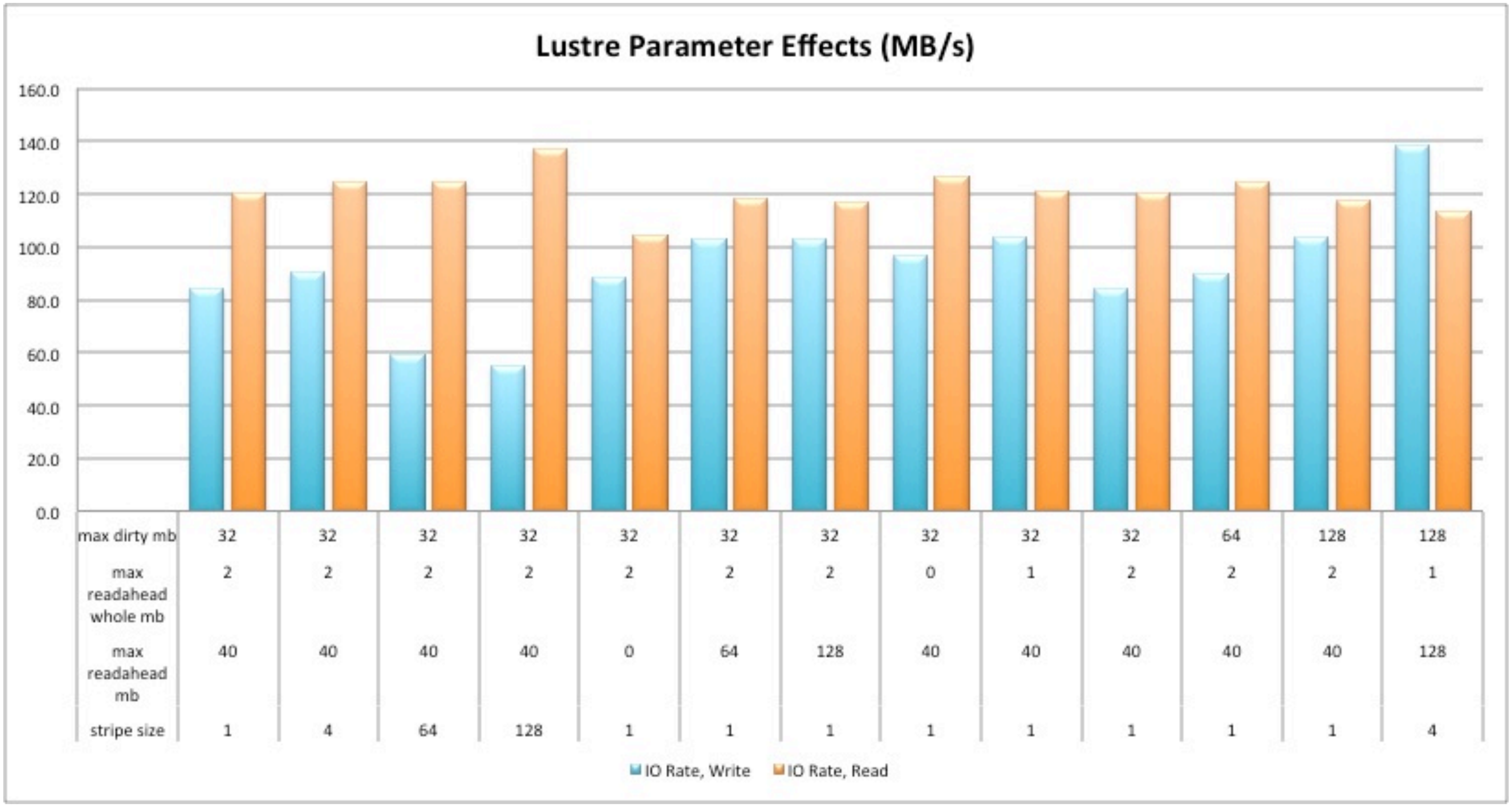


MR tuning

Data from **Hadoop Performance Tuning: A case study** *Berkeley 6/09*

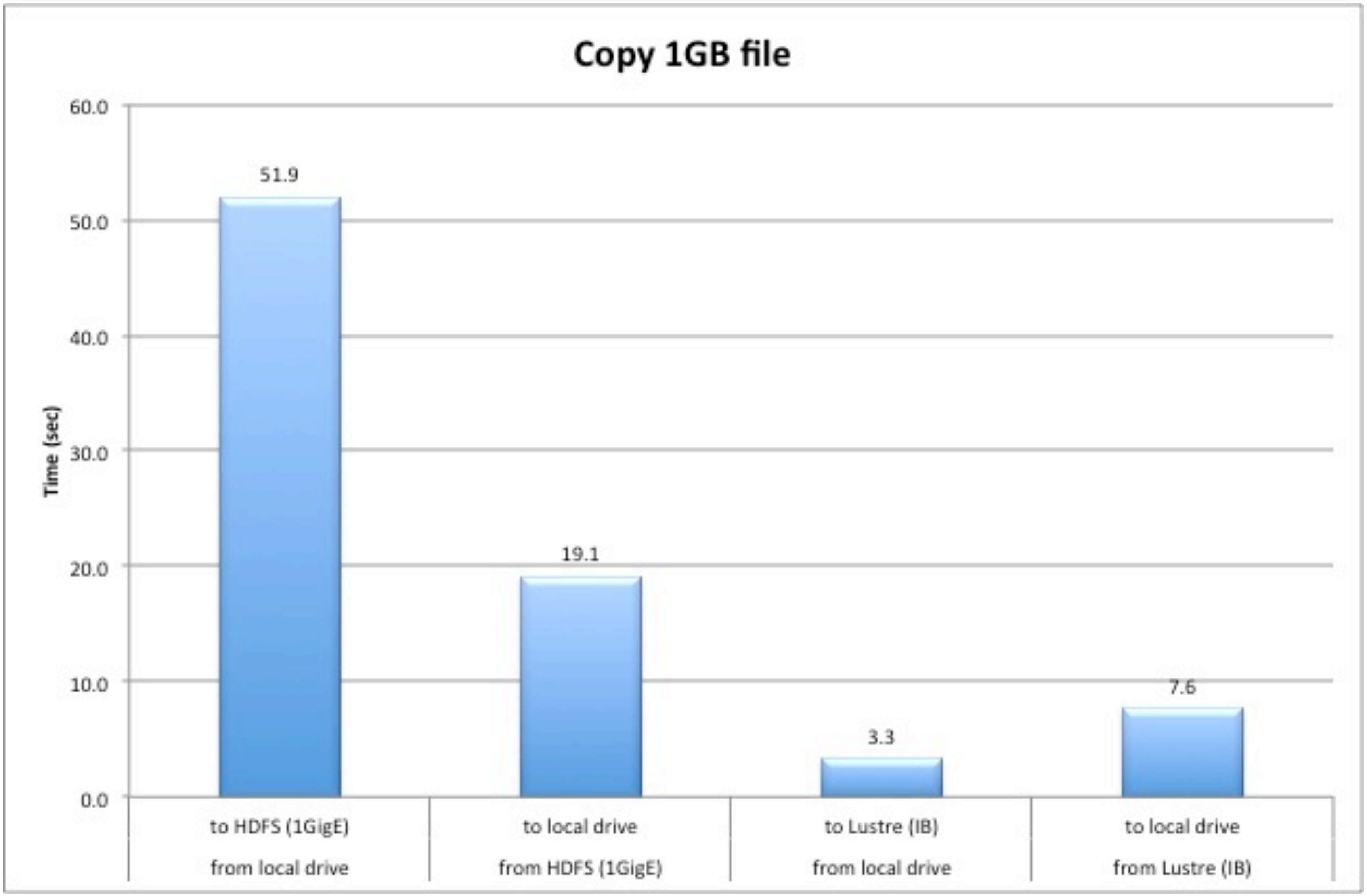


Lustre Tuning: TestDFSIO



max dirty mb	32	32	32	32	32	32	32	32	32	32	64	128	128
max readahead whole mb	2	2	2	2	2	2	2	0	1	2	2	2	1
max readahead mb	40	40	40	40	0	64	128	40	40	40	40	40	128
stripe size	1	4	64	128	1	1	1	1	1	1	1	1	4

Data Staging: Not a Fair Comparison



Hypothetical Cost Comparison

- Assume Lustre IB has 2x performance of HDFS 1GigE
  - 3x for our sort benchmark
  - Top 500 LINPACK efficiency: 1GigE ~45-50%, 4xQDR ~90-95%

	Lustre / IB Cluster			HDFS / 1 GigE Cluster		
	Count	Price	Subtotal	Count	Price	Subtotal
Nodes	100	$7,500	$750,000	200	$7,500	$1,500,000
Switches	9	$6,500	$58,500	12	$4,000	$48,000
Cables	178	$100	$17,800	450	$10	$4,500
OSS	2	$52,000	$104,000	0	---	---
Storage	128TB	---	---	384TB	$100	$38,400
MDS	1	$34,000	$34,000	0	---	---
Racks	4	$8,000	$32,000	6	$8,000	$48,000
Total			**$996,300**			**$1,638,900**

Cost Considerations

- Client node count dominates the overall cost of the cluster
- Doubling size = doubling power, cooling, maintenance costs
- Cluster utilization efficiency
- Data transfer time
- Necessity of maintaining a second cluster

Conclusions

- HPC environments have fast networks
- MR should show theoretical performance gains on an appropriately-designed Lustre cluster
- Test results on a small cluster support these propositions
- Performance effects for a particular job may vary widely
- No reason why Hadoop and Lustre can't live happily together
  - Shared storage
  - Shared compute nodes
  - Better performance

Fini

Thanks!