SDC
STORAGE DEVELOPER CONFERENCE
SNIA ■ SANTA CLARA, 2012

Accelerating Hadoop with Data Optimization

Hank Cohen
Altior Inc.

ALTIOR™
elevated performance

Outline

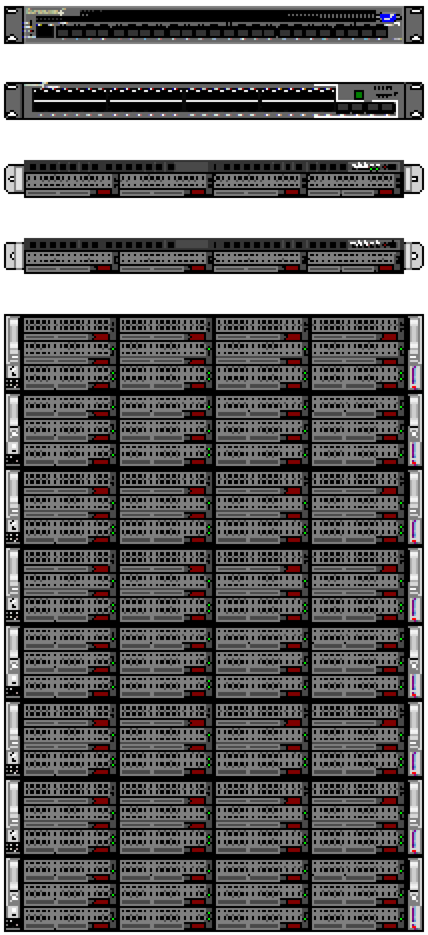
1. Introduction to Hadoop
2. Introduction to CeDeFS
3. Acceleration via Data Optimization
4. Benchmarks
5. Results

What is hadoop?

- A processing system for reliable scalable distributed computing
  - Map/Reduce Framework
  - Hadoop Distributed File System
- Hadoop is designed to manage petabyte processing tasks
- Clusters from 10 to 2000 servers
  - Facebook Data Warehouse
    - 21 Petabytes
    - 2000 nodes, 12 TB/node

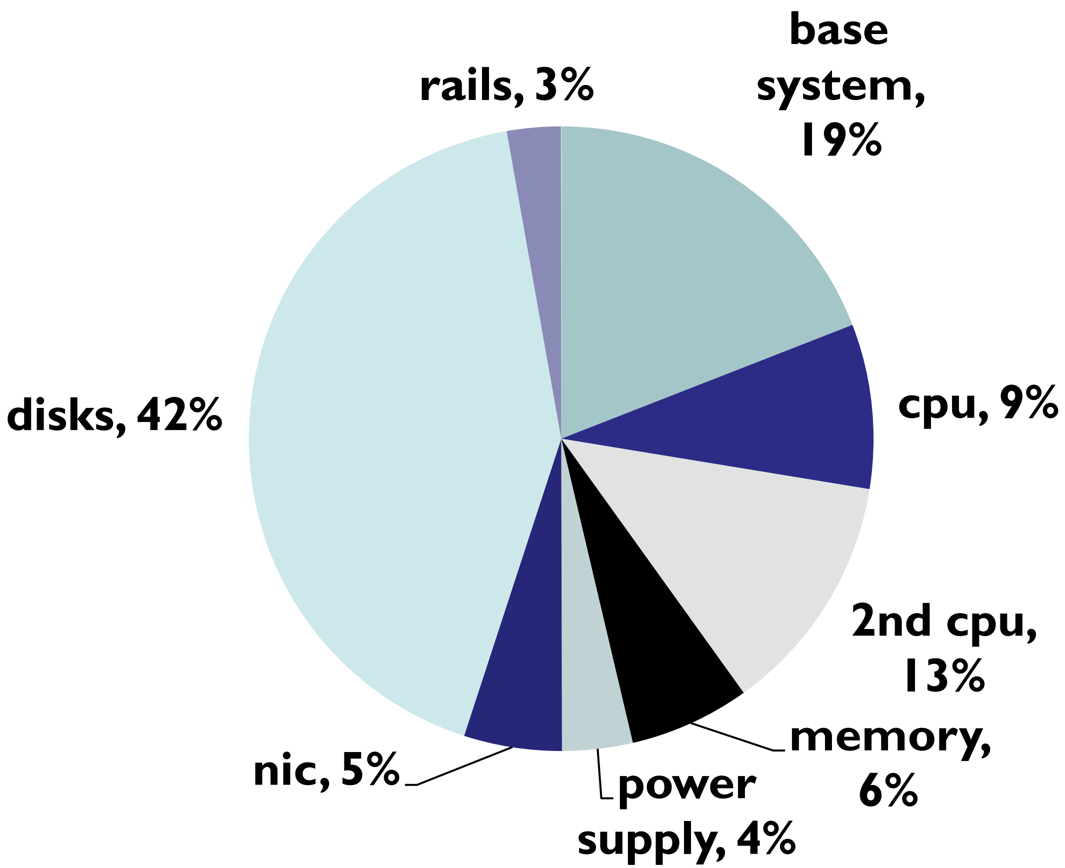
A small Hadoop cluster

- 8 Data Nodes
- 2 Name Nodes
- Each Data Node
  - 2 sockets Xeon E5640
  - 6 core w. hyperthreads
  - 48 GB memory
  - 12 x 1 TB disks
- Total 96 cores 192 threads
- Benchmark system courtesy of Supermicro

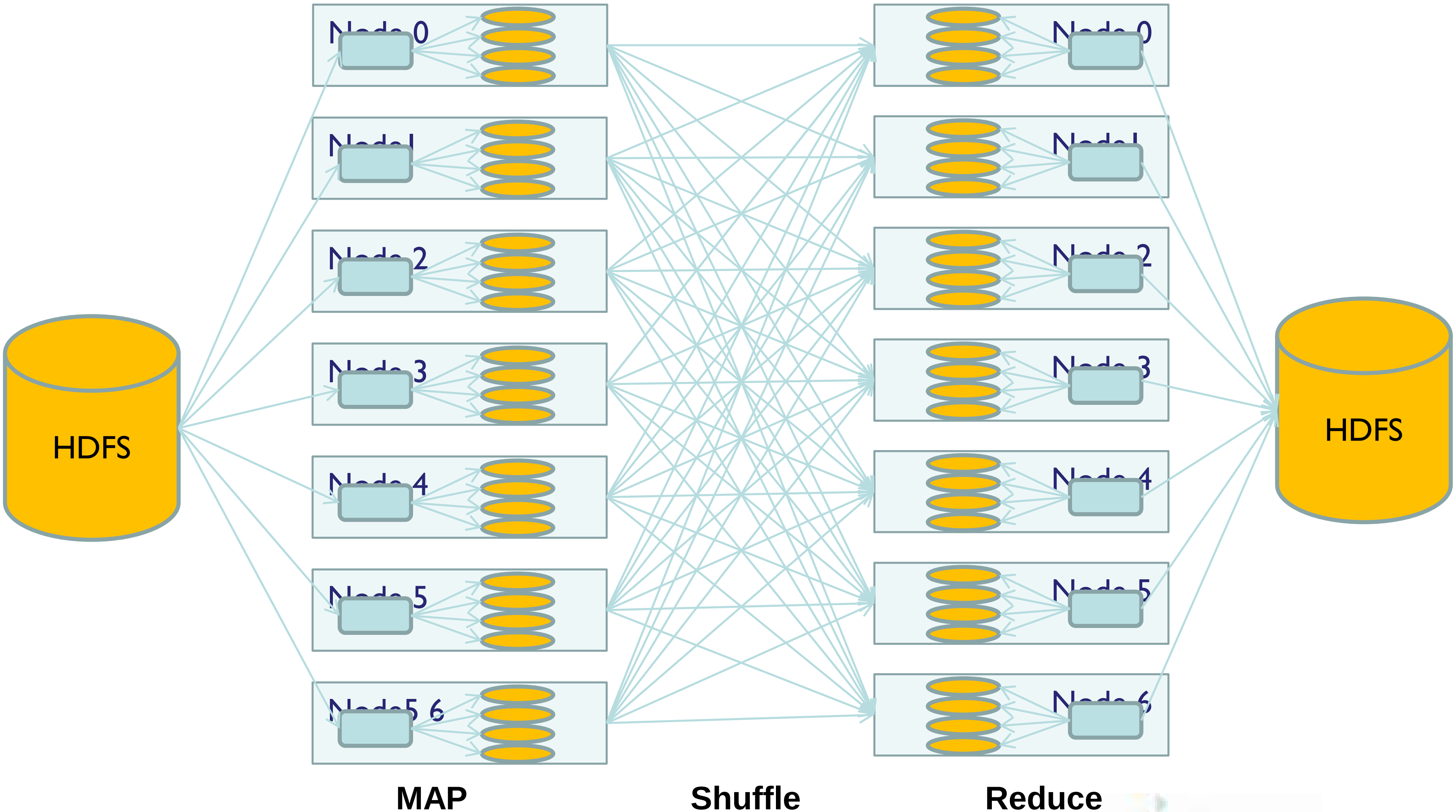


Hadoop cluster cost

Data Optimization can reduce system cost



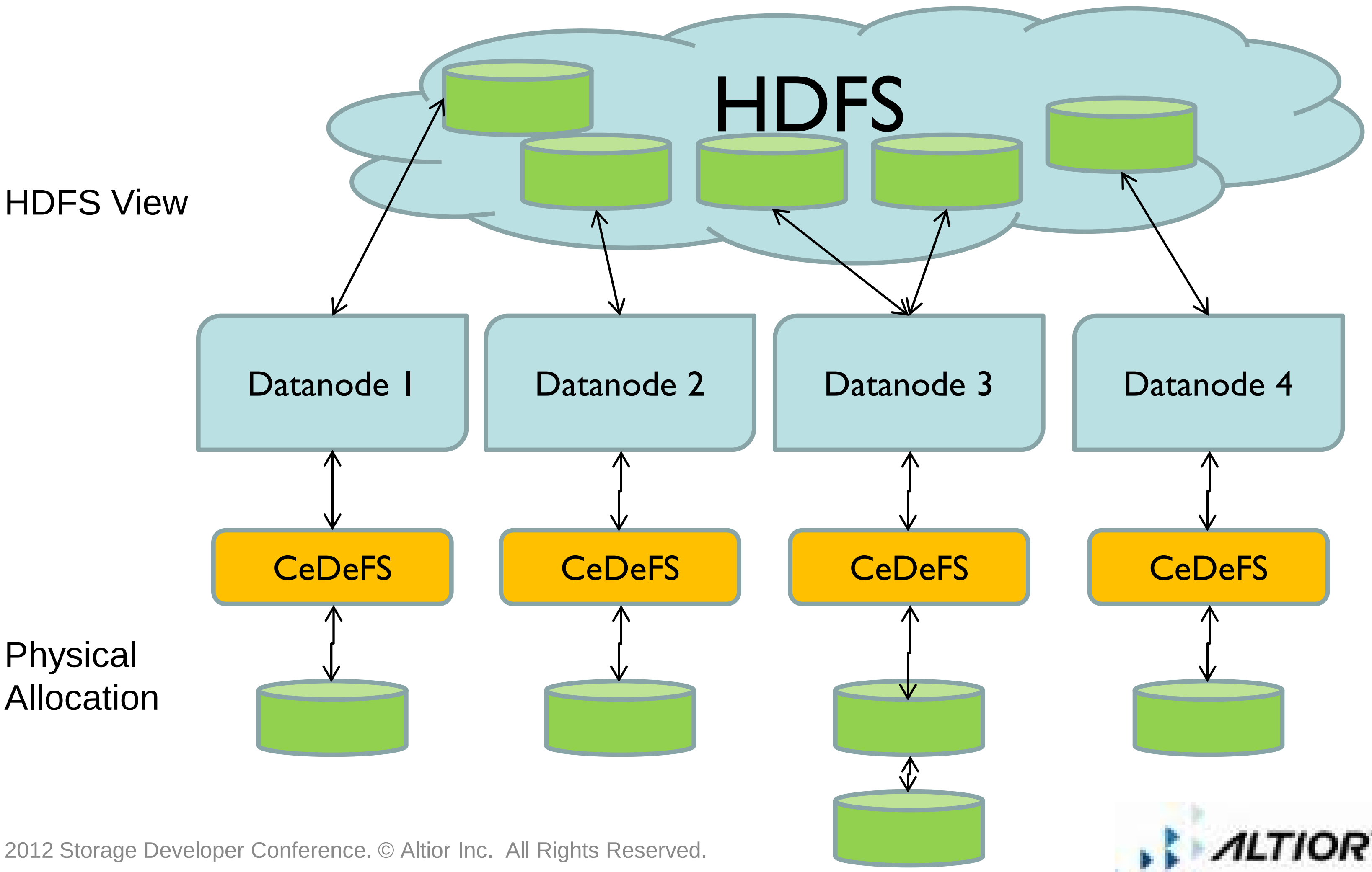
Hadoop – Map Reduce Dataflow



Hadoop File System

- Files are distributed across the cluster
- Blocks are allocated as files on the local file system on each data node
- The Namenode keeps track of all metadata
  - Where are the blocks – rack awareness
  - Replication

HDFS with Data Optimization



Why does data optimization speed execution?

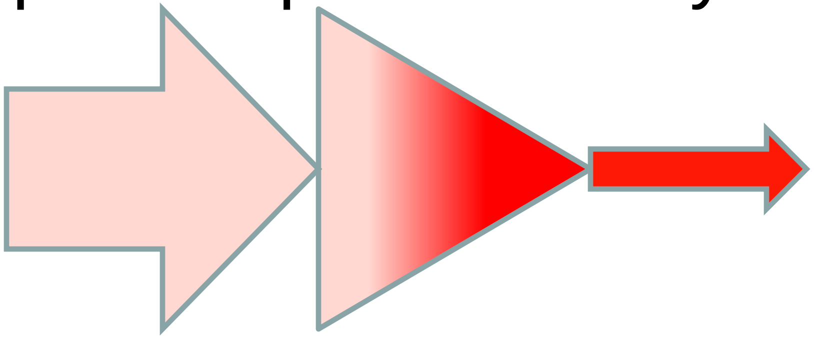
- Hardware accelerated Compression
  - No CPU overhead
  - High compression ratio
  - Asynchronous I/O doesn't stall processing threads
  - I/O system is unburdened
  - Compression multiplies read throughput
    - Less I/O wait time for I/O bound processes
- HADOOP – fatter data nodes
  - Increased capacity of data nodes means fewer are required. Less shuffle traffic.

Your Mileage may vary!

- Compression is data dependent
  - Text can compress very well ~ 6:1
  - Encrypted or random data will not compress at all
  - Compressed data will compress little or none
  - Multi-media files are already compressed
- Hadoop data is usually very compressible
  - ASCII text compresses well
  - “http://www.” Might compress to 4 bits

Software compression helps a little

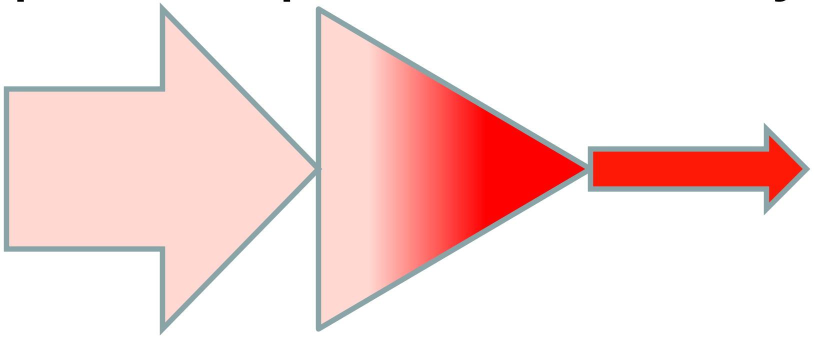
- LZ0 or Snappy in software
  - Compression ratio 2:1 for Calgary Corpus
  - Throughput at input --100MByte/Second/Core



  - Throughput to disk – 50 MB/S/Core
  - 100% CPU utilization for dedicated cores

Hardware acceleration helps a lot

- AltraFlex hardware accelerated gzip
  - Compression ratio 3.3:1 for Calgary Corpus
  - Throughput at input -- 1000MByte/Second



  - Throughput to disk -- 270MB/S
  - Increased information density on the disk
  - Fewer I/O ops, less I/O overhead
  - Less than 5% CPU overhead

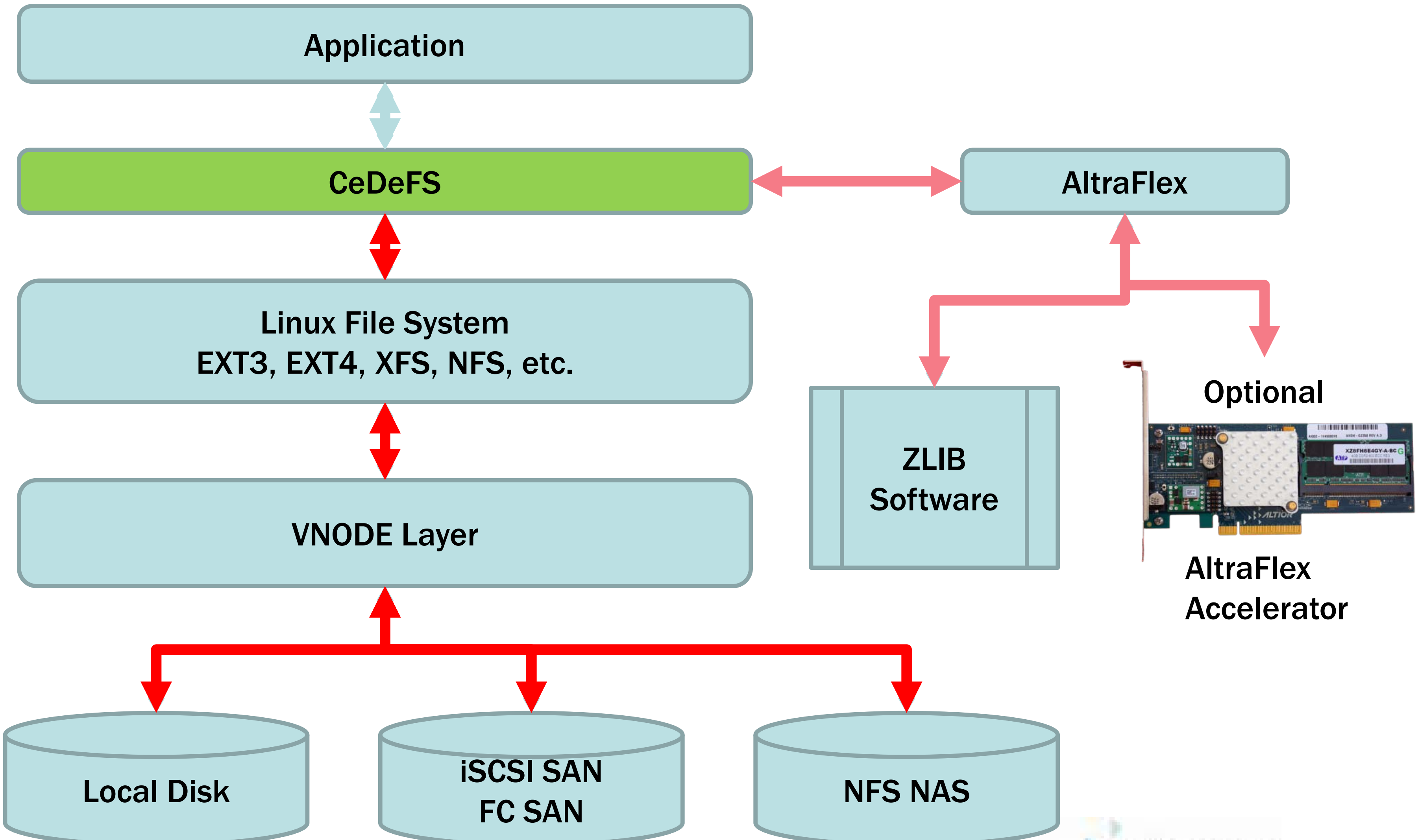
Decompression Delivers I/O Accleration

- Data rate = disk throughput x compression ratio
- A single disk ~100 MB/S will deliver 330MB/S to the CPU
- GZIP and LZO both deliver about 100MB/S per core but GZIP has 2x better compression so it uses ½ as much disk throughput
- This is faster than an SSD!
- I/O bound tasks spend less time in I/O wait state

CeDeFS - Compression/Decompression Filter Software

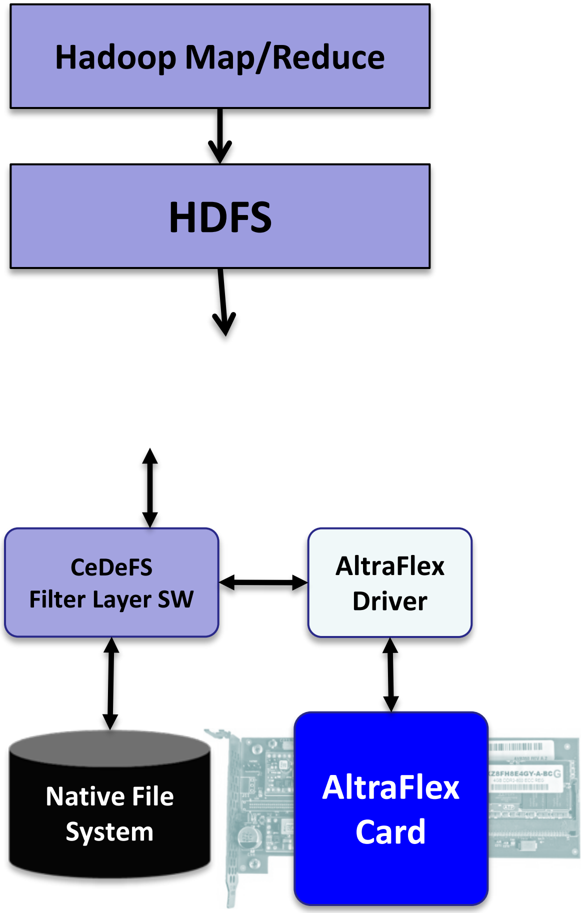
- A file system filter layer to manage compression and decompression.
- Increase storage capacity up to 6x
- Compress all Linux applications
- No modifications to applications or workflows
- Preserves native file system semantics
- Use any Linux file system: EXT3, EXT4, XFS
- Disk and I/O optimization for Primary, Secondary and Archival storage

CeDeFS Block Diagram



Accelerated Hardware Configuration

- A Hadoop Cluster with CeDeFS enabled Nodes consists of
  - Altior CeDeFS Filter SW
  - AltraFlex hardware accelerator
- CeDeFS is transparent to Hadoop. No code changes required and workflow remains the same
- 3x-6x increase in storage capacity in each node
- Enhanced CPU utilization and reduced runtime through I/O reduction and optimization
- Significantly benefits I/O bound tasks.
- Increased data density reduces the shuffle traffic
- Reduction in Power – Per Node, Per Cluster



Terasort Benchmark Results

SUPERMICRO

Hardware Configuration	
Cluster Size	8 Data Nodes; 2 Name Nodes
CPU	E5640; Dual Socket 6 Core CPU; 96 Cores Total
Memory	48 GB
Storage	12 * 1 TB
Network Link	1 10G Link; 1 1G Link
Switch	TBD
Altior HW	AltraFlex PCIe Card based on GZ350 FPGA

Software Configuration	
Hadoop Version	CDH3
Operating System	RHEL 6.2

Normalized Terasort Test Results 512GB			
	Elapsed time		
	12 Disks	6 Disks	8 Disks
Native	100%	207%	141%
LZO	49%	60%	53%
CeDeFS	36%	42%	37%

Additional Datapoints

- A 8 TB terasort test case completed on a 6 disk per node cluster using CeDeFS and AltraFlex accelerators.
- The same 8TB sort using software LZO failed running out of space.

Conclusions

- Hardware accelerated compression provides meaningful acceleration as well as added capacity
- Acceleration plus added capacity means bigger jobs executed in less time
- Very significant savings in both CAPEX and OPEX