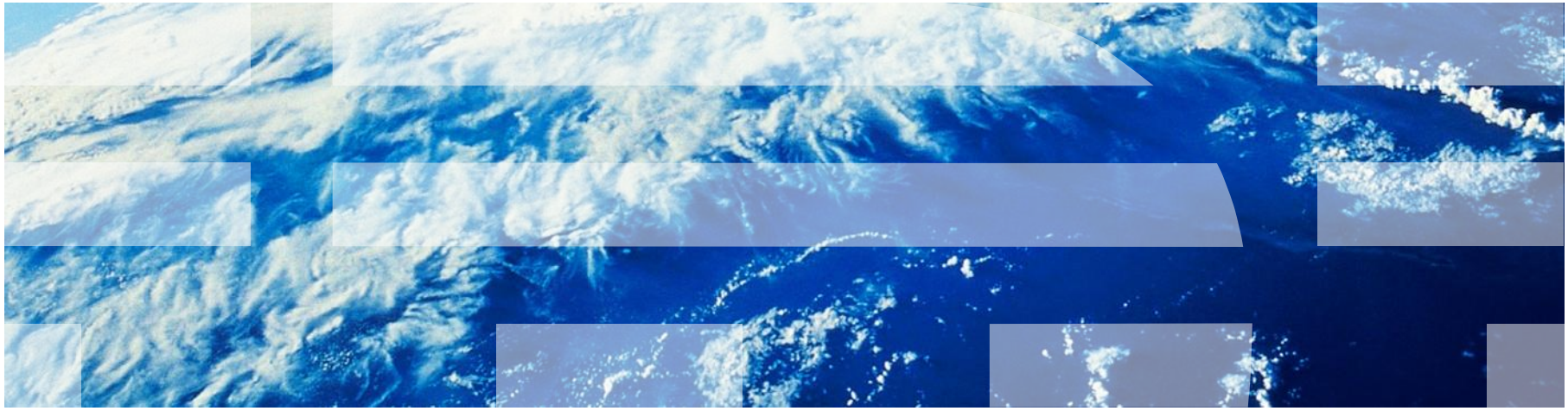
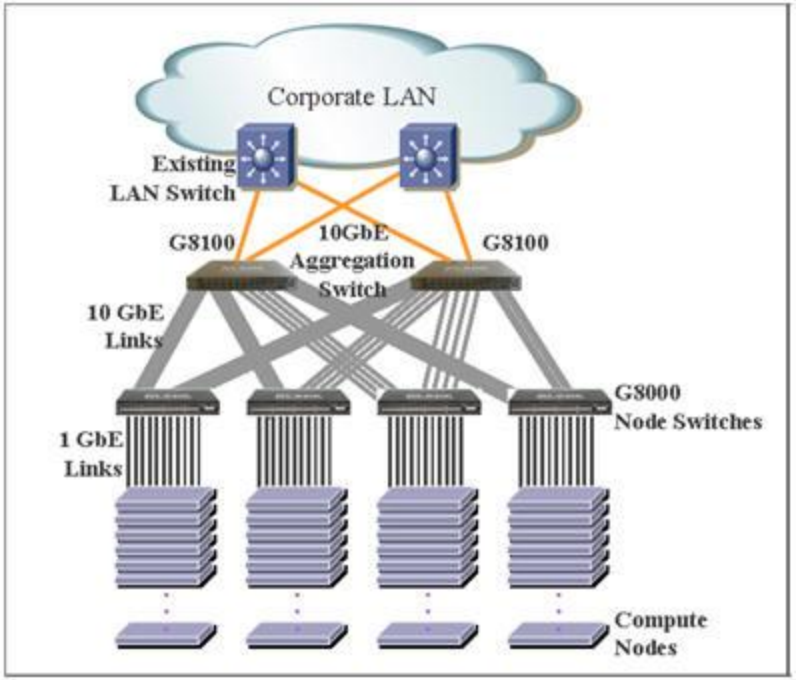
ZoneFS: Stripe Remodeling in Cloud Data Centers

Lanyue Lu - University of Wisconsin-Madison
Dean Hildebrand, Renu Tewari - IBM Almaden Research Lab



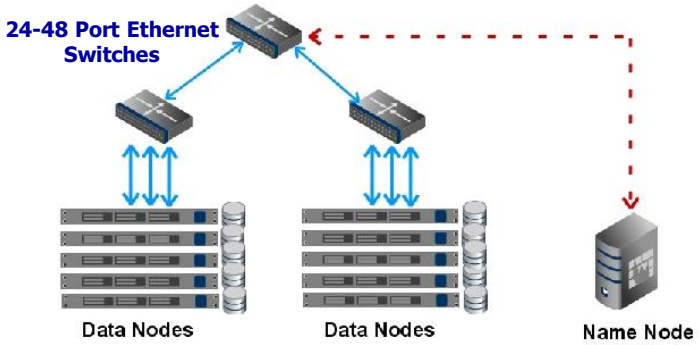
Cloud Data Centers on the Cheap

- Network infrastructure
  - Hierarchical
    - 10GigE at root
    - 1GigE at leaves (racks)
  - Commodity switches
    - Limited switch port buffering
- Oversubscribed 8:1 (125Mbps) or more
- Bandwidth out of each rack is limited
  - Depends on oversubscription at connection level in hierarchy



*from Data Center Design by PTS Data Center Solutions, Inc.

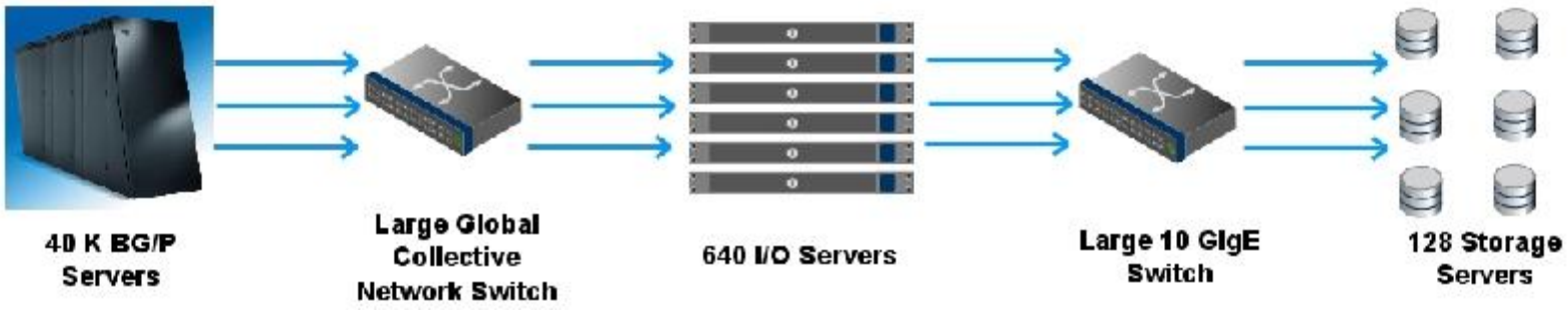
Storage on the cheap (re-use existing local disks)



HDFS Architecture

- Storage software for cheap hardware
- Internet-Scale File Systems (HDFS, GFS, Cloudstore)
  - Applications write all data to local node
    - Reduces parallelism, uneven load balancing
    - Create hotspots
  - Replication for fault tolerance
    - Triplication: 2 copies in 1 rack, 1 copy in another rack
    - Large stripe size (64 MB)
  - Non-standard file system API
    - Not POSIX API or semantics
    - Map/Reduce specific, cannot handle general workloads
  - Fixed 1:1 ratio of storage and compute nodes
  - Limited metadata processing capability
    - Limited data sharing
    - Low small I/O and create performance
    - Space allocation can be slow which space is limited

But Supercomputers do it (with money…)



Intrepid Supercomputer and GPFS Parallel File System

- Storage software for expensive hardware
- Parallel file systems are only as fast as the hardware
- Use large and expensive switches
  - Large port buffers allow a high-level of parallelism
  - $700K for 128-port 10GigE switches (or >100 48-port GigE switches)
- Compute servers realize all available bandwidth

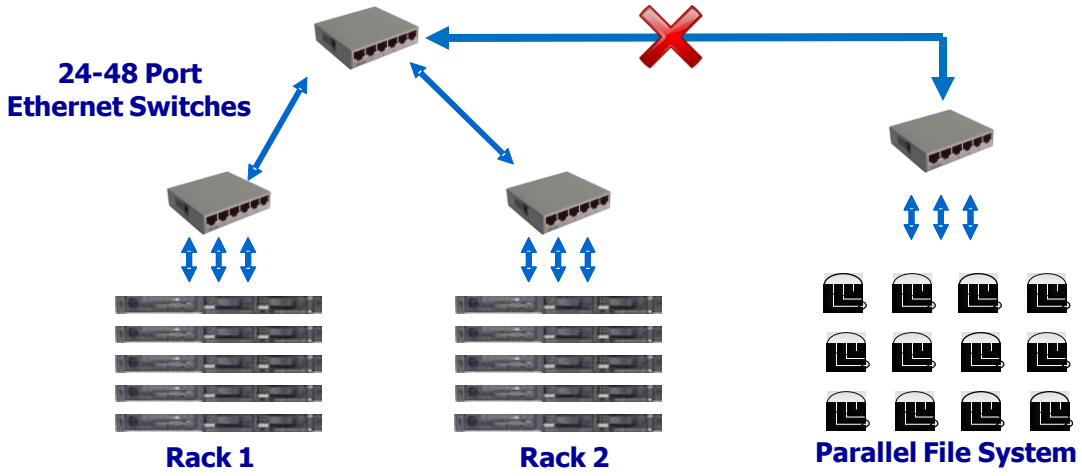
***Number from http://cscads.rice.edu/workshops/summer09/slides/leadership-computing/ALCF_Overview.pdf**

Merge Parallel File Systems and Cheap Hardware?

Goal: Leverage parallel file systems with data center infrastructures

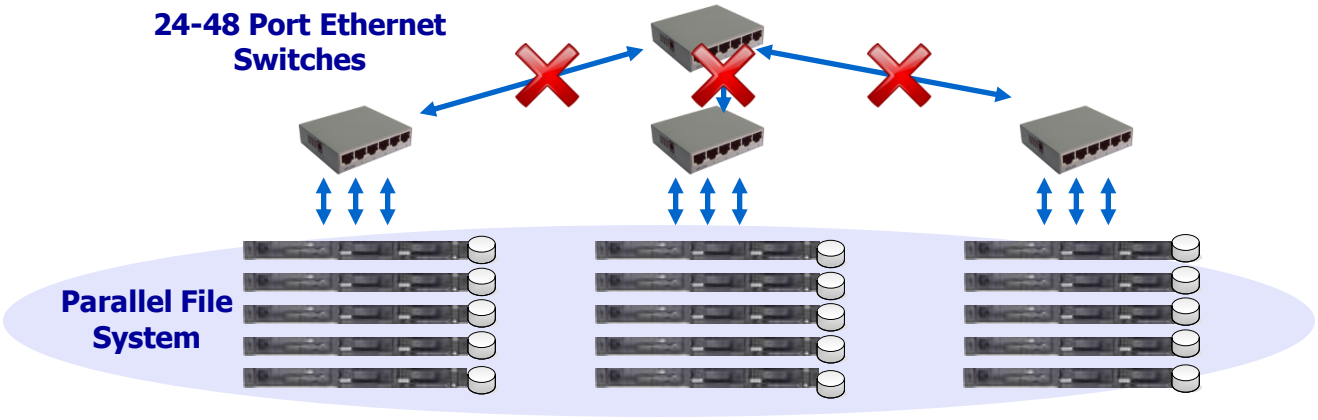
But there are options…….

Data Centers and Parallel File Systems
Option 1: Traditional (separate compute and I/O servers)



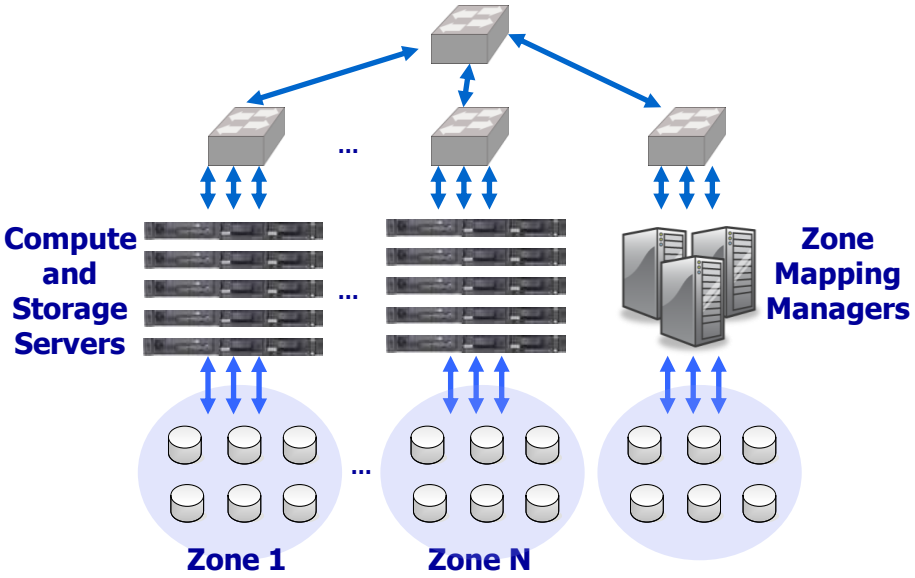
- Oversubscription creates bottlenecks
- Compute servers realize a fraction of desired bandwidth
- Limited port buffering reduces I/O performance even further

Data Centers and Parallel File Systems
Option 2: Overlay compute and storage nodes



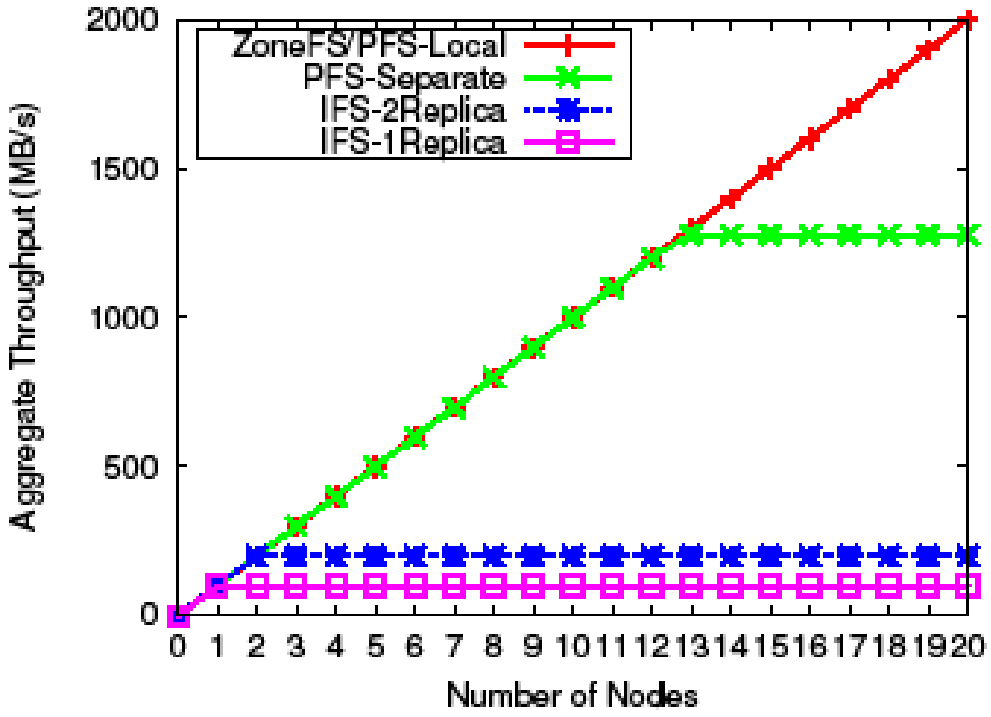
- Network oversubscription defeats purpose of data striping
- Aggregation and leaf switches scale in unison
  - 10GigE switches will emerge at leaf layer
  - 40/100 GigE switches will emerge at aggregation layer

ZoneFS Architecture



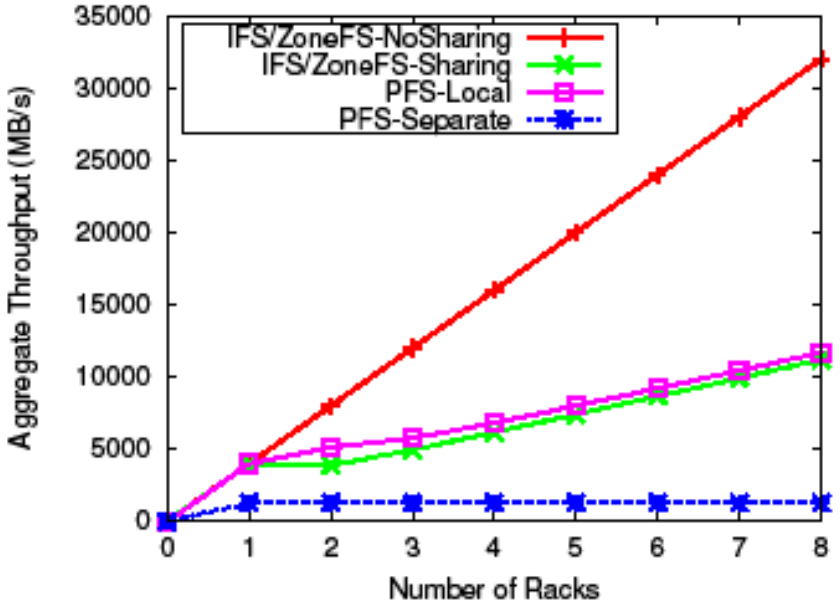
- Re-organize file system into Zones
  - Switch, compute nodes, storage nodes
  - Flexible storage node disk architecture
    - Dedicated or non-dedicated
    - Local disk, SAN
- Stripe data across storage nodes within a Zone
  - Level with the highest bandwidth
- Compute nodes have parallel data access within their Zone

Single Rack Scalability Simulation
Read single data chunk



- Internet scale file systems (IFS) limited by bandwidth to number of replicas
- Parallel File System (PFS-Separate) limited by 10GigE external switch
- ZoneFS and PFS-Local leverage full bandwidth of rack switch

Multi-Rack Scalability Simulation
Access large file test

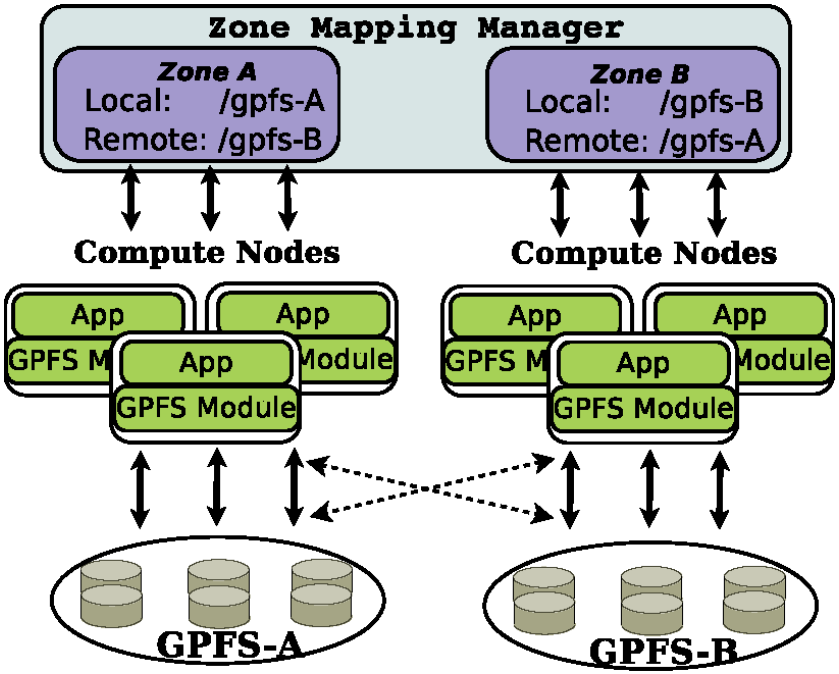


- Parallel File System (PFS-Separate) limited by 10GigE external switch
- {PFS-Local, IFS, ZoneFS}-Sharing all limited by oversubscribed switches
- {IFS, ZoneFS}-NoSharing scale linearly with additional rack switches

Notes:

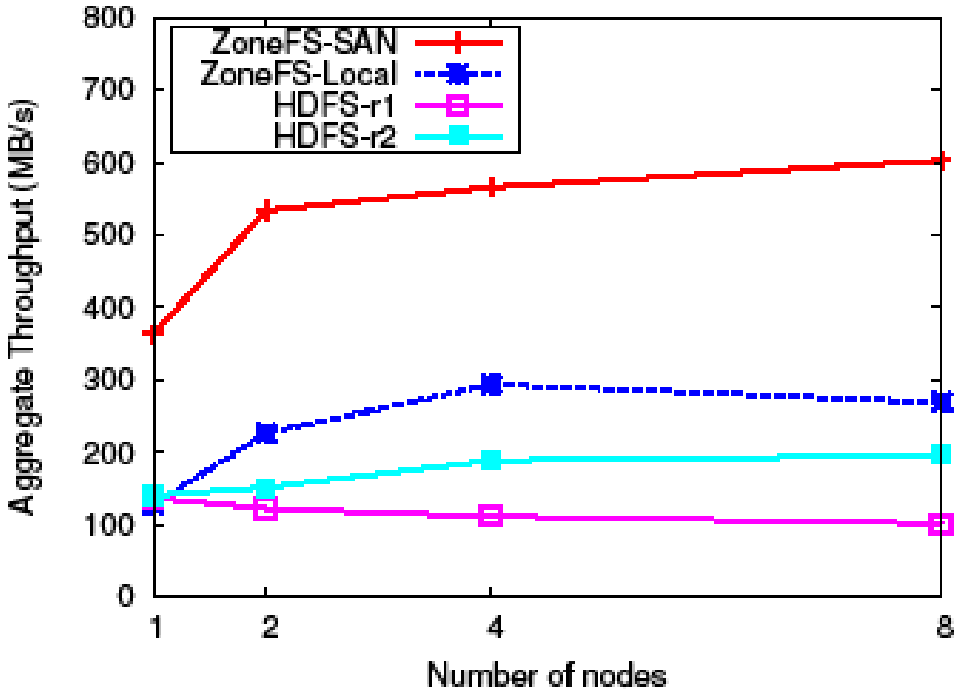
- Sharing – Access data across racks
- NoSharing – Access data within rack
- 40 nodes per rack
- Simulated up to 48 racks

ZoneFS Prototype



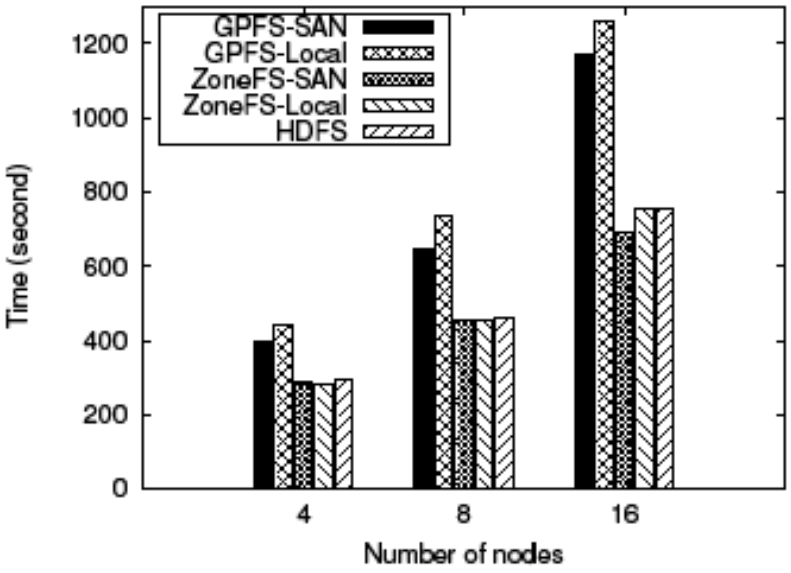
- Layered above GPFS in Linux
  - SAN and DAS
- File system per rack
  - Inter-rack access available through GPFS client to remote file systems
- Job Scheduler
  - Hadoop or MPI
  - GPFS – pick any random node
  - ZoneFS - pick random node in correct zone

Hotspot read performance

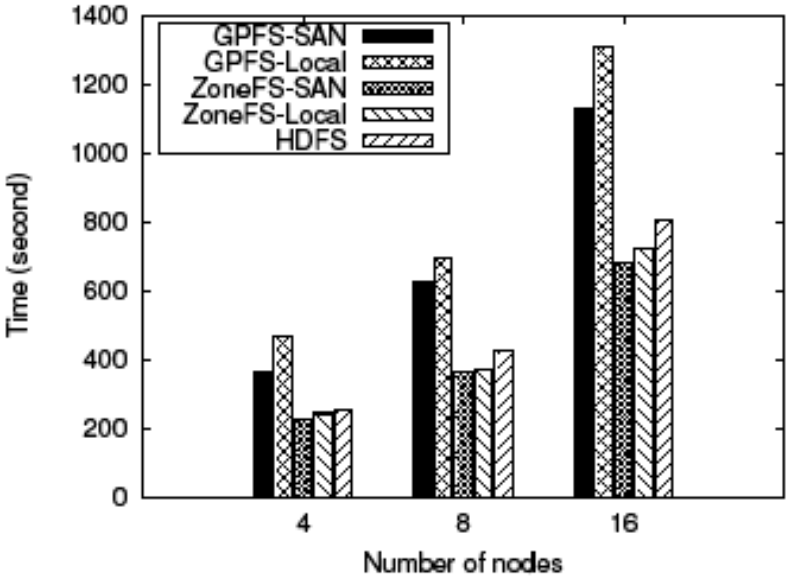


- Setup
  - Single rack
  - All data on a single node
- HDFS constrained by number of replicas
- ZoneFS-Local performance constrained by limited port-buffering in Netgear switch
- ZoneFS-SAN limited by SAN bandwidth

Hadoop Teragen and Terasort



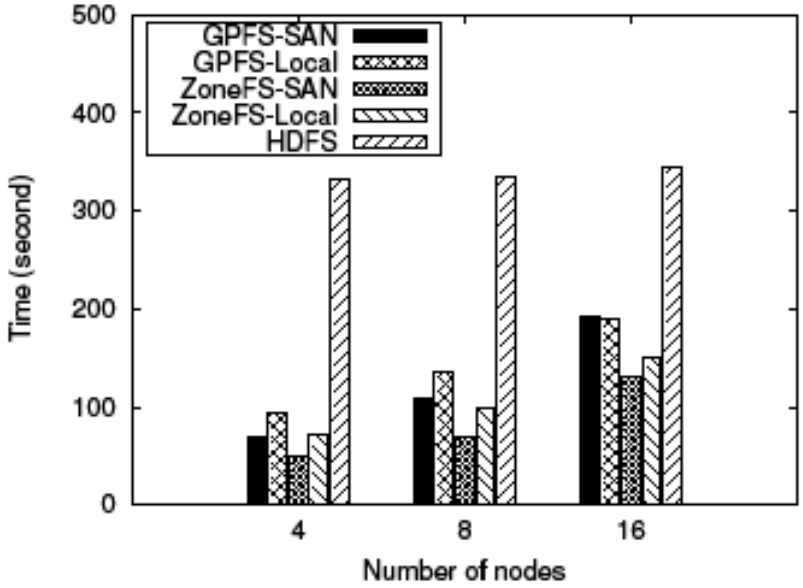
TeraGen
(write intensive)



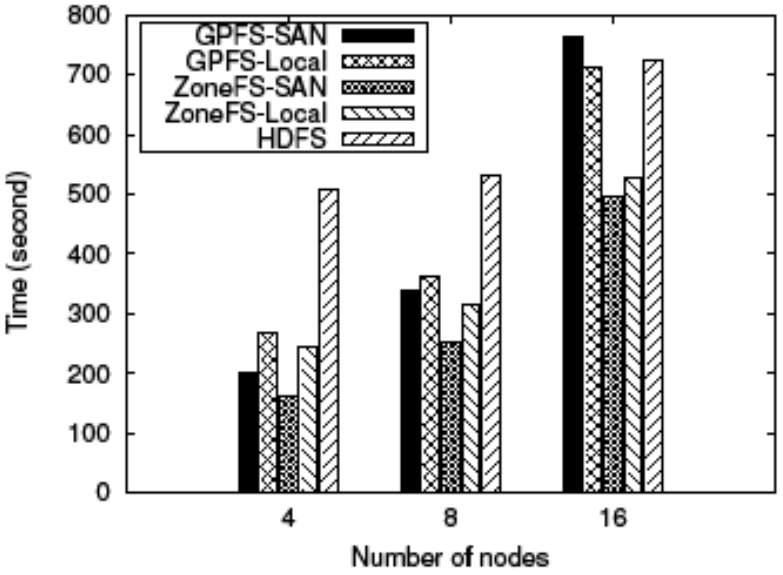
Terasort
(read and write intensive)

- Setup
  - 2 racks
  - Increase nodes in-tandem between racks
  - 1GB file per node
- GPFS limited by inter-rack switch

General workload - Image conversion



1MB JPG to 9MB BMP



4MB JPG to 35MB BMP

- Setup
  - Relatively small files
  - 200 images per node
  - 2 racks
- GPFS constrained by inter-rack switch
- HDFS constrained by single NameNode
- ZoneFS-SAN limited by disk bandwidth
- ZoneFS-local limited by switch port-buffering

Summary

- Enterprise solution
  - Support general applications (POSIX API and semantics)
  - Support Information Lifecycle Management (ILM)

- Scalability and Performance
  - Independent scaling of switch, compute, and storage nodes
  - Add zones as needed
  - Intra-switch parallelism improves load balancing
  - Avoid the inter-rack network bottleneck experience by parallel file systems
  - Good performance for both analytics and general workloads

- Availability and flexibility
  - Avoid ingesting input data and offloading results
    - Single storage system for primary and analytic data
  - Flexible options for compute and storage nodes
    - Local disks or SAN per rack
  - Vary the ratio of compute to storage nodes
    - Independent management of compute and storage node

Thank You !

Questions ?