LA-UR-13-25967

Functional Assessment of Erasure Coded Storage Archive

Computer Systems, Cluster, and Networking Summer Institute

Blair Crossman
NEW MEXICO TECH
SCIENCE • ENGINEERING • RESEARCH UNIVERSITY

Taylor Sanchez
CALIFORNIA STATE UNIVERSITY
SAN BERNARDINO

Josh Sackos
WASHINGTON STATE UNIVERSITY
School of Electrical Engineering and Computer Science

Los Alamos NATIONAL LABORATORY
New Mexico CONSORTIUM
NSF
ISTI INFORMATION SCIENCE & TECHNOLOGY INSTITUTE
PRObE Parallel Reconfigurable Observational Environment

Presentation Overview

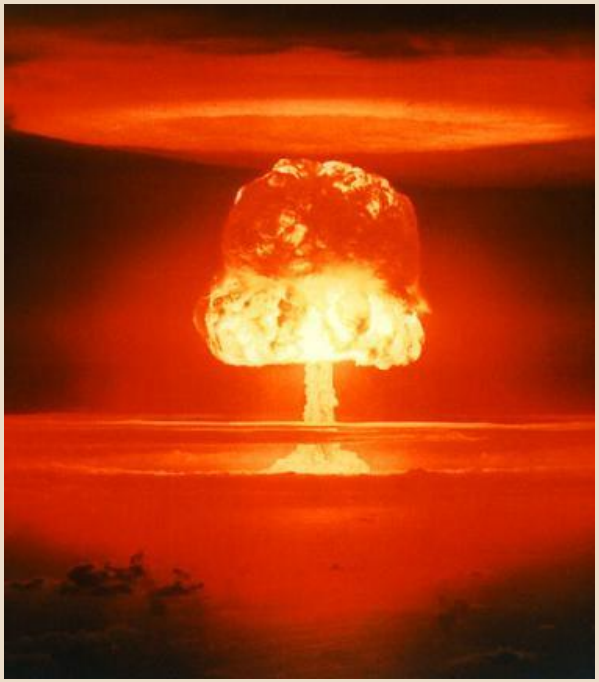
- Introduction
- Caringo Testing
- Scality Testing
- Conclusions

Storage Mediums

- Tape
  - Priced for capacity not bandwidth
- Solid State Drives
  - Priced for bandwidth not capacity
- Hard Disk
  - Bandwidth scales with more drives

Object Storage: Flexible Containers

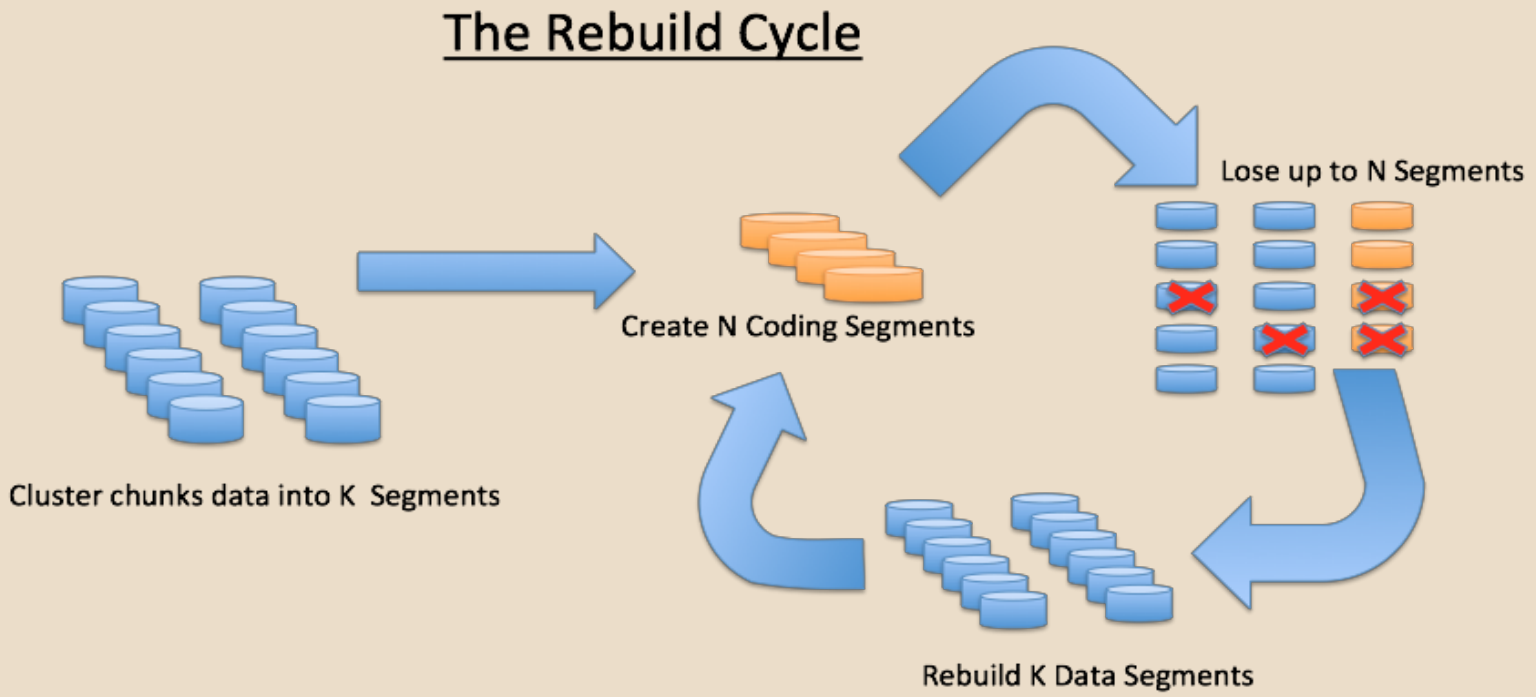
- Files are stored in data containers
- Meta data outside of file system
- Key-value pairs
- File system scales with machines
- METADATA EXPLOSIONS!!



What is the problem?

- RAID, replication, and tape systems were not designed for exascale computing and storage
- Hard disk Capacity continues to grow
- Solution to multiple hard disk failures is needed

Erasure Coding : Reduce Rebuild Recalculate



Reduce! Rebuild! Recalculate!

Project Description

- Erasure coded object storage file system is a potential replacement for LANL's tape archive system
- Installed and configured two prototype archives
  - Scality
  - Caringo
- Verified the functionality of systems

Functionality Not Performance

Caringo

- SuperMicro admin node
- 1GigE interconnect
- 10 IBM System x3755
  - 4 x 1TB HDD
- Erasure coding:
  - n=3
  - k=3

Scality

- SuperMicro admin node
- 1GigE interconnect
- 6 HP Proliant (DL160 G6)
  - 4 x 1TB HDD
- Erasure coding:
  - n=3
  - k=3

Project Testing Requirements

- Data
  - Ingest : Retrieval : Balance : Rebuild
- Metadata
  - Accessibility : Customization : Query
- POSIX Gateway
  - Read : Write : Delete : Performance overhead

How We Broke Data

- Pulled out HDDs (Scality, kill daemon)
- Turned off nodes
- Uploaded files, downloaded files
- Used md5sum to compare originals to downloaded copies

caringo

Caringo: The automated storage system

- Warewulf/Perceus like diskless (RAM) boot
- Reconfigurable, requires reboot
- DHCP PXE boot provisioned
- Little flexibility or customizability
- http://www.caringo.com

caringo

No Node Specialization

- Nodes "bid" for tasks
  - Lowest latency wins
  - Distributes the work
- Each node performs all tasks
  - Administrator : Compute : Storage
- Automated Power management
  - Set a sleep timer
  - Set an interval to check disks
- Limited Administration Options

caringo

Caringo Rebuilds Data As It Is Written

- Balances data as written
  - Primary Access Node
  - Secondary Access Node
- Automated
  - New HDD/Node: auto balanced
  - New drives format automatically
  - Rebuilds Constantly
  - If any node goes down rebuild starts immediately
  - Volumes can go "stale"
  - 14 Day Limit on unused volumes

caringo

What's a POSIX Gateway

- Content File Server
  - Fully Compliant POSIX object
  - Performs system administration tasks
  - Parallel writes
- Was not available for testing

caringo

“Elastic” Metadata

- Accessible
- Query: key values
  - By file size, date, etc.
- Indexing requires “Elastic Search” machine to do indexing
  - Can be the bottleneck in system

Minimum Node Requirements

- Needs a full n + k nodes to:
  - rebuild
  - write
  - balance
- Does not need full n +k to:
  - read
  - query metadata
  - administration

SCALITY

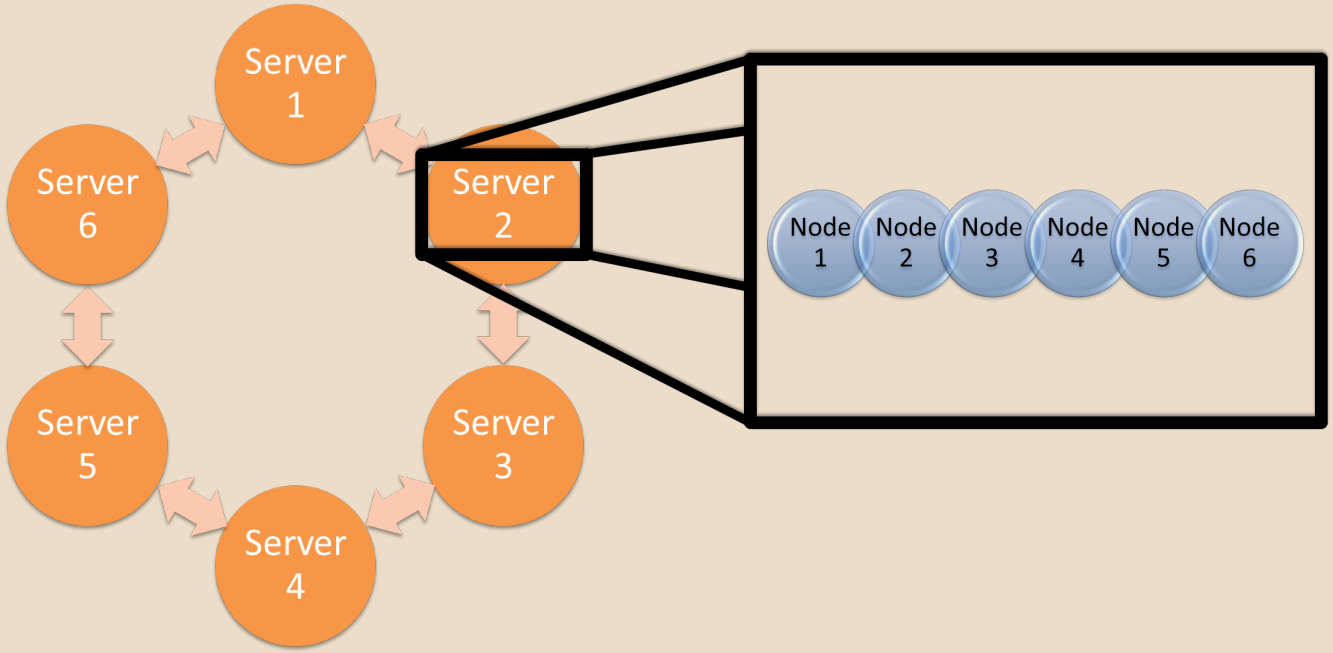
Static Disk Install

- Requires disk install
- Static IP addresses
- Optimizations require deeper knowledge
- http://www.scality.com

SCALITY

Virtual Ring Resilience

- Success until less **virtual** nodes available than n+k erasure configuration.
- Data stored to ‘ring’ via distributed hash table



SCALITY

Manual Rebuilds, But Flexible

- Rebuilds on less than required nodes
  - Lacks full protection
- Populates data back to additional node
- New Node/HDD: Manually add node
- Data is balanced during:
  - Writing
  - Rebuilding

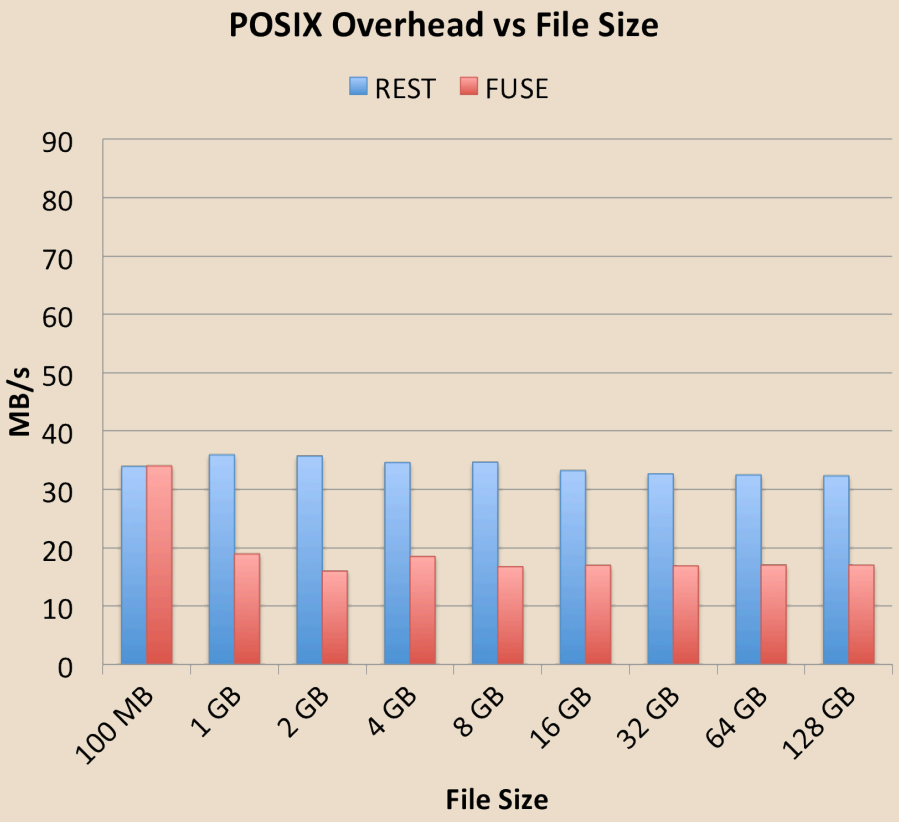
SCALITY

Indexer Sold Separately

- Query all erasure coding metadata per server
- Per item metadata
- User Definable
- Did not test Scality's 'Mesa' indexing service
  - Extra software

SCALITY

Fuse gives 50% Overhead, but scalable



On the right path

- Scality
  - Static installation, flexible erasure coding
  - Helpful
  - Separate indexer
  - 500MB file limit ('Unlimited' update coming)
- Caringo
  - Variable installation, strict erasure coding
  - Good documentation
  - Indexer included
  - 4TB file limit (addressing bits limit)

Very Viable

- Some early limitations
- Changes needed on both products
- Scality seems more ready to make those changes.

Questions?

Acknowledgements

Special Thanks to :

Dane Gardner - NMC Instructor

Matthew Broomfield - NMC Teaching Assistant

HB Chen - HPC-5 - Mentor

Jeff Inman - HPC-1- Mentor

Carolyn Connor - HPC-5, Deputy Director ISTI

Andree Jacobson - Computer & Information Systems Manager NMC

Josephine Olivas - Program Administrator ISTI

Los Alamos National Labs, New Mexico Consortium, and ISTI