Ceph: An Open Source Object Store

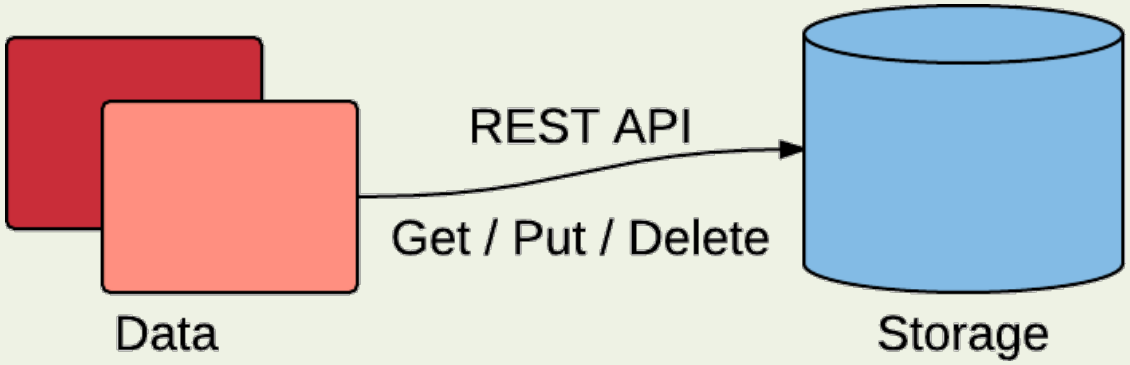
Evan Harvey
Gustavo Rayos
Nick Schuchhardt

Mentors: David Bonnie, Chris Hoffman, Dominic Manno

LA-UR-15-25907

What is an Object Store?

- Manages data as objects
- Offers capabilities that are not supported by other storage systems
- Object Storage vs. Traditional Storage



LA-UR-15-25907

What is Ceph?

- An **object store** and filesystem
- Open source and freely available
- Scalable to the Exabyte level

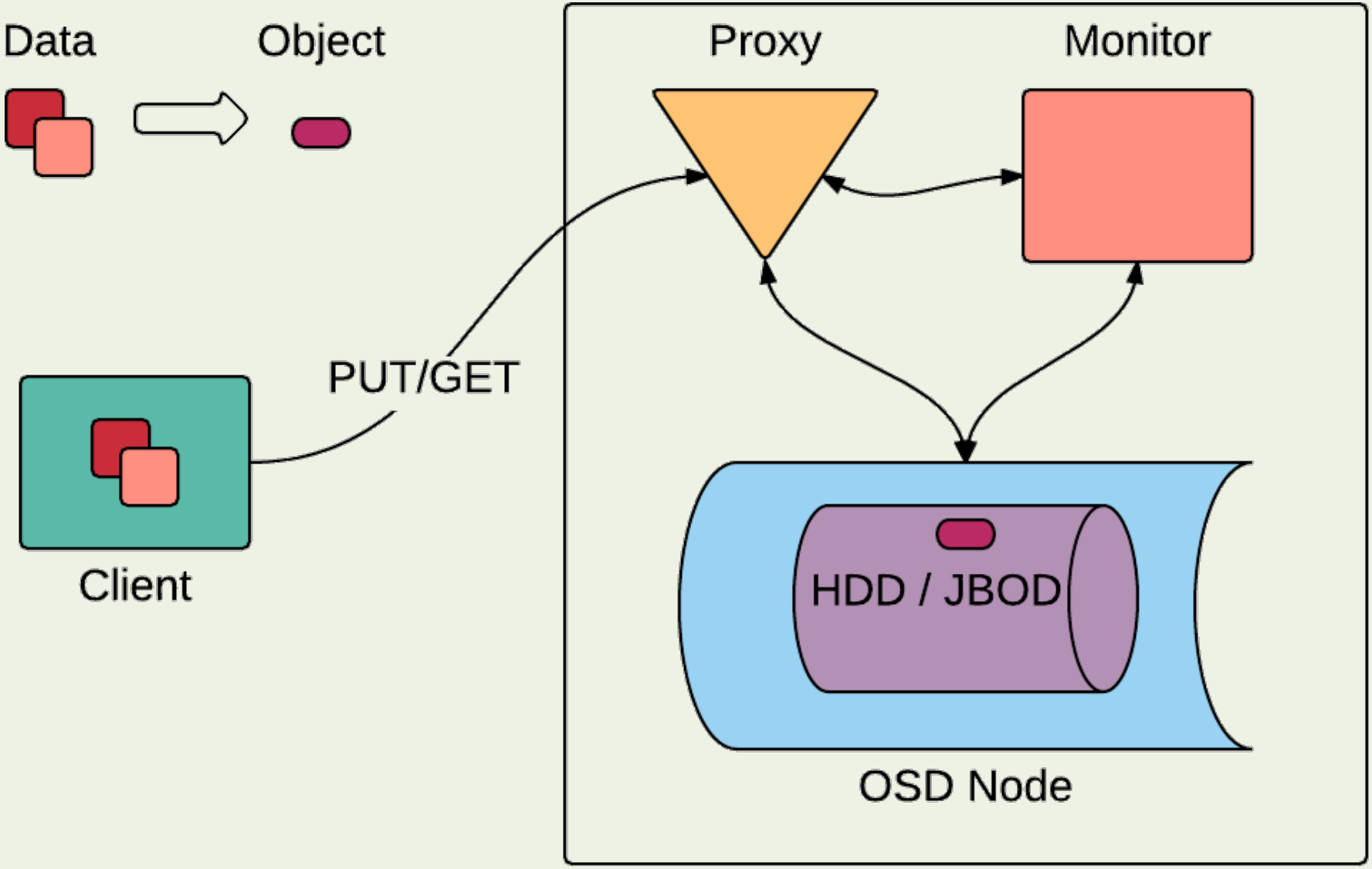
LA-UR-15-25907

Basic Ceph Cluster

- Monitor Node
  - Monitors the health of the Ceph cluster
- OSD Node
  - Runs multiple Object Storage Daemons (One daemon per hard drive)
- Proxy Node
  - Provides an object storage interface
  - Can interact with cluster using PUT/GET operations
  - Provides applications with a RESTful gateway to the Ceph storage cluster

LA-UR-15-25907

Basic Ceph Cluster



But Why?

- Campaign Storage
- More reliable than other file systems
- POSIX compliant
- Scales better than RAID
- Cost efficient

LA-UR-15-25907

Project Goals

- Build a Ceph storage cluster
  - 1 Monitor node
  - 6 OSD nodes (Around 20 OSD daemons each)
  - 3 proxy nodes
- Erasure coding profiles
- Single vs. Multiple proxies

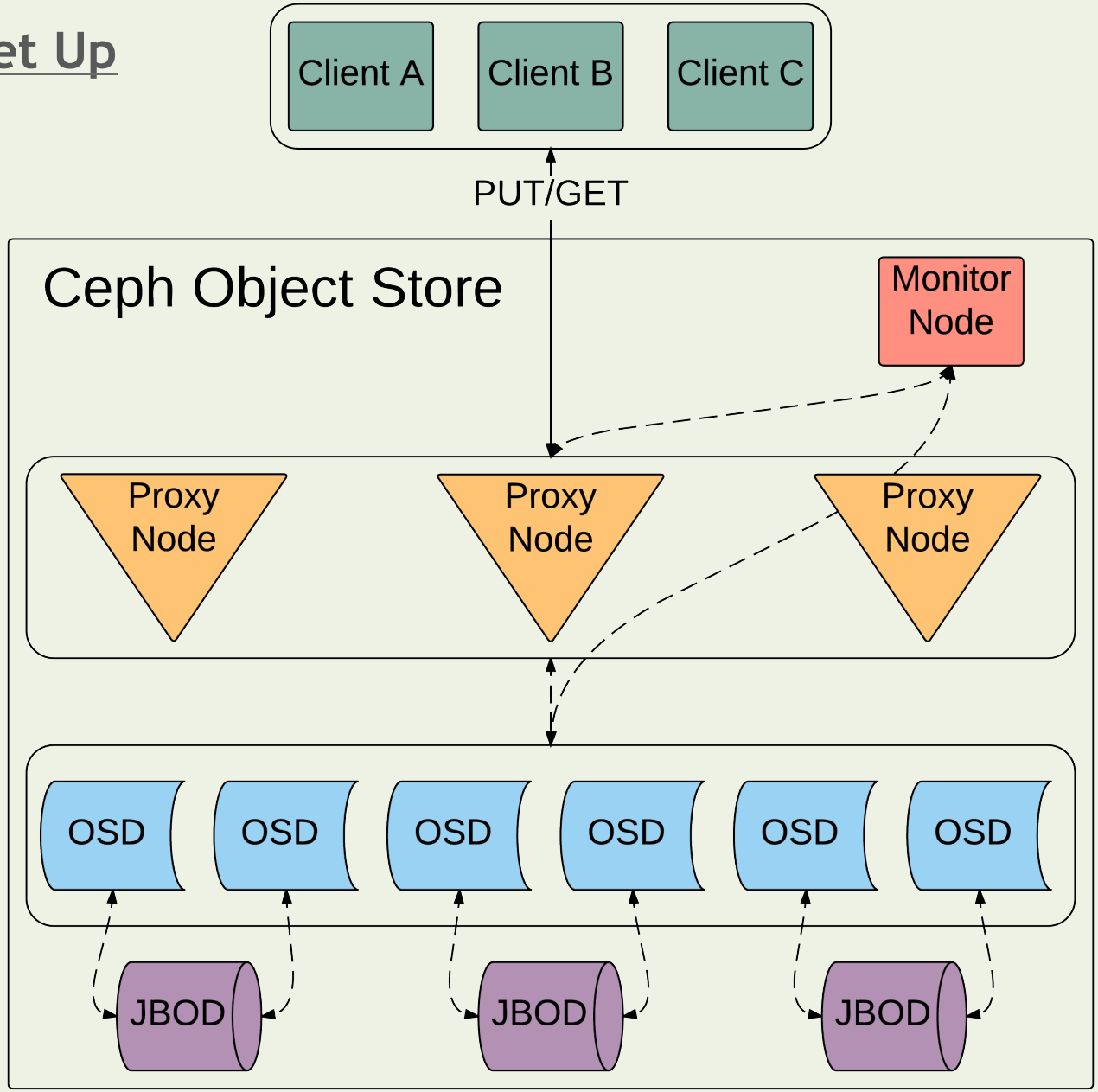
LA-UR-15-25907

Test Environment

- CentOS 6.6
- Ten HP ProLiant D380P Gen8 Servers
- Three Supermicro 847jbod-14 (45 disks each)
- Mellanox Infiniband 56 Gb/s
- Two SAS cards 6 Gb/s
  - 8 ports at 600 MB/s
- Four Raid cards 6 Gb/s
  - 8 PCI Express 3.0 lanes

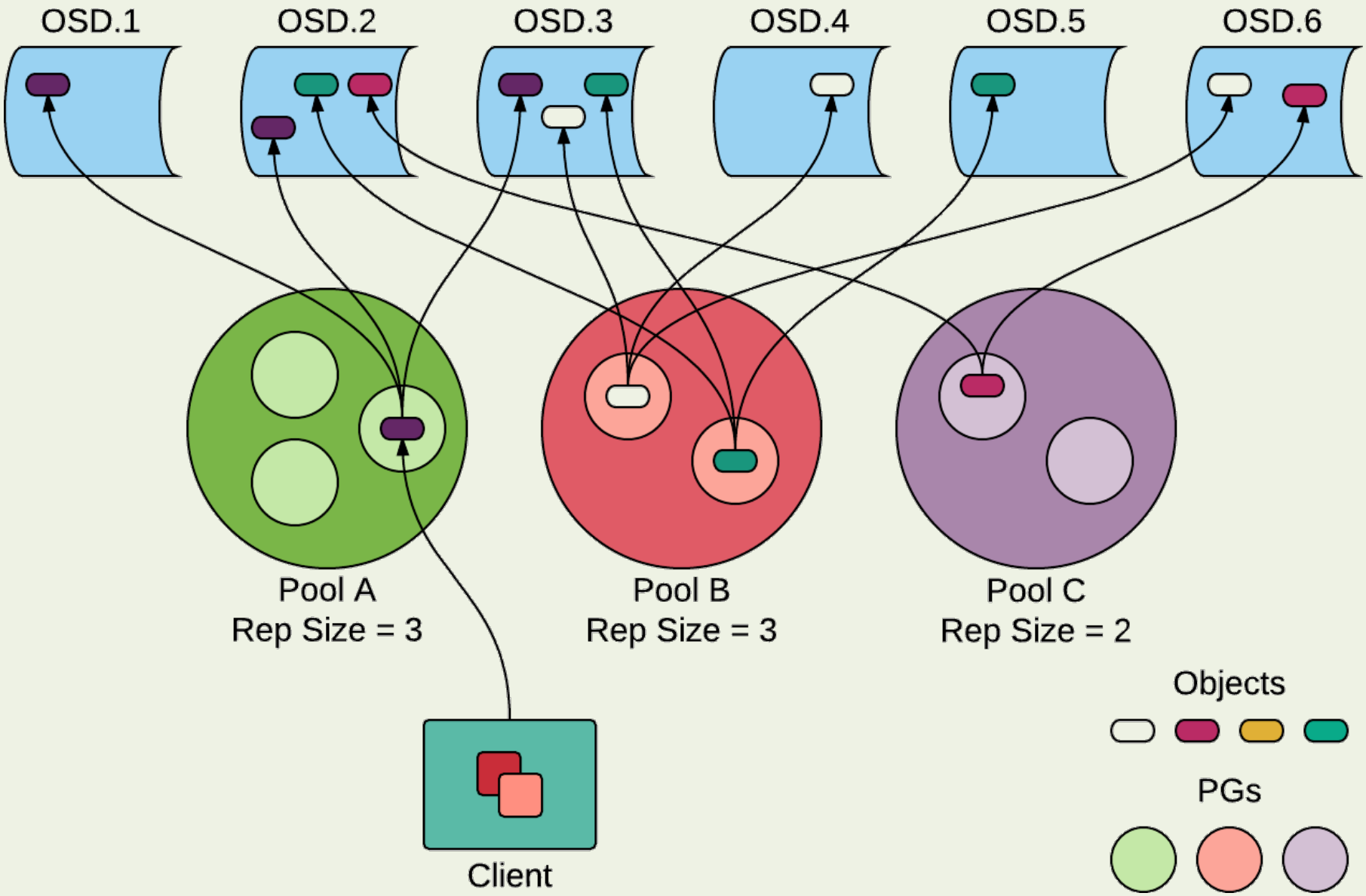
LA-UR-15-25907

Our Set Up



LA-UR-15-25907

Pools and PGs



Pools and Placement Groups

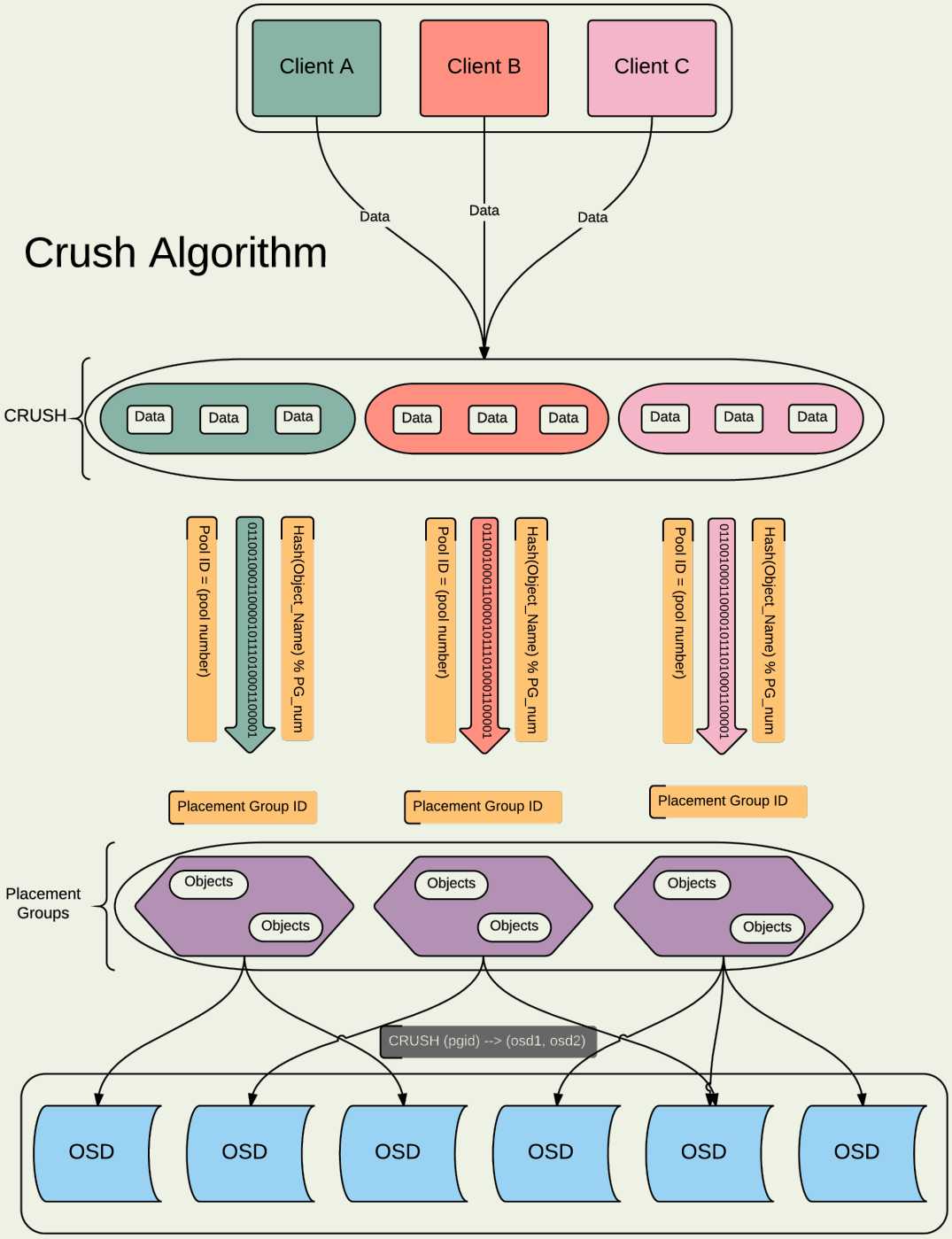
- An object belongs to a single placement group
- Pools group placement groups
- Placement groups belong to multiple OSDs

LA-UR-15-25907

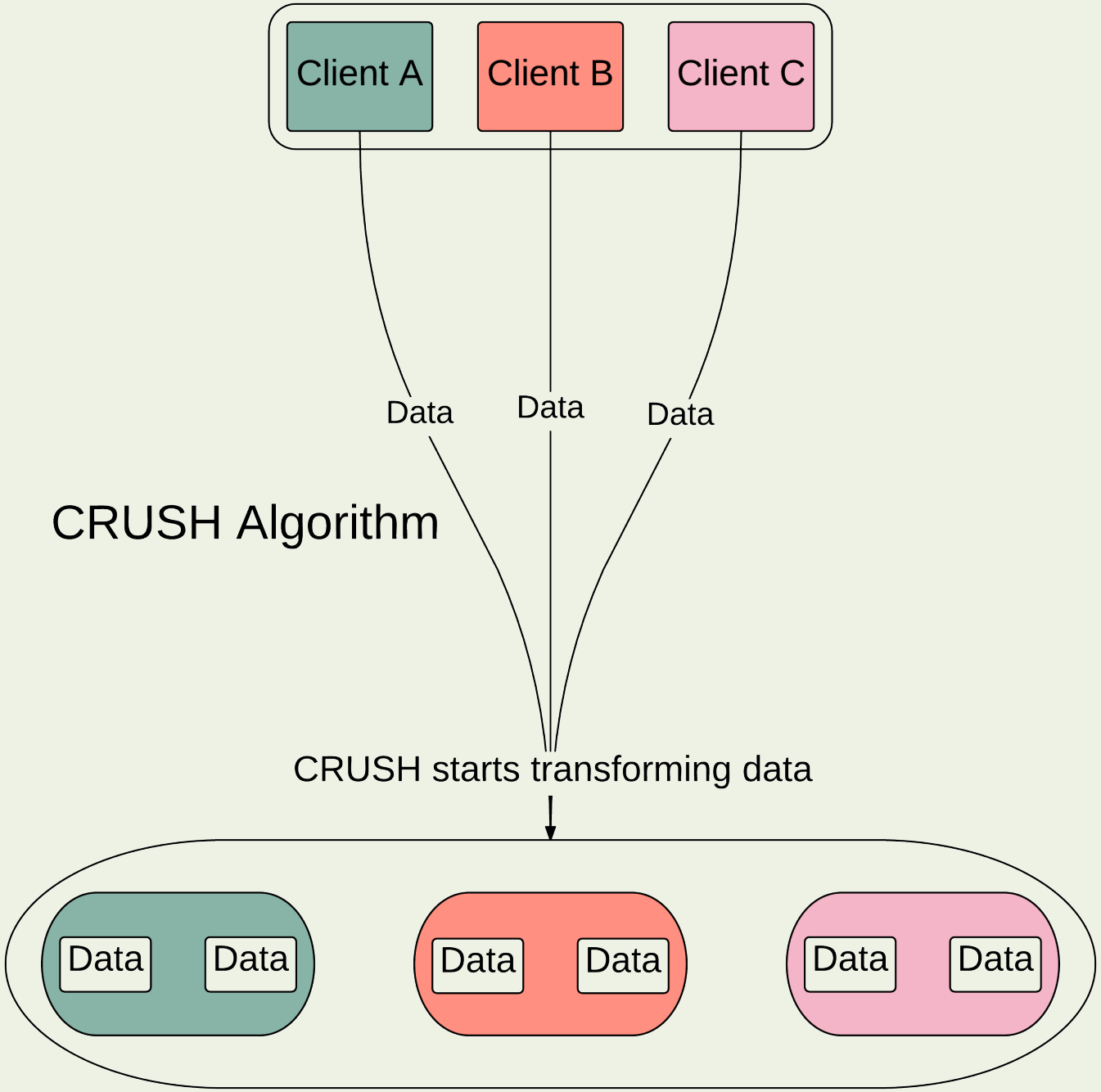
CRUSH!

- Controlled Replication Under Scalable Hashing (CRUSH)
- Algorithm finds optimal location to store objects
- Stripes objects across storage devices
- On the OSDs

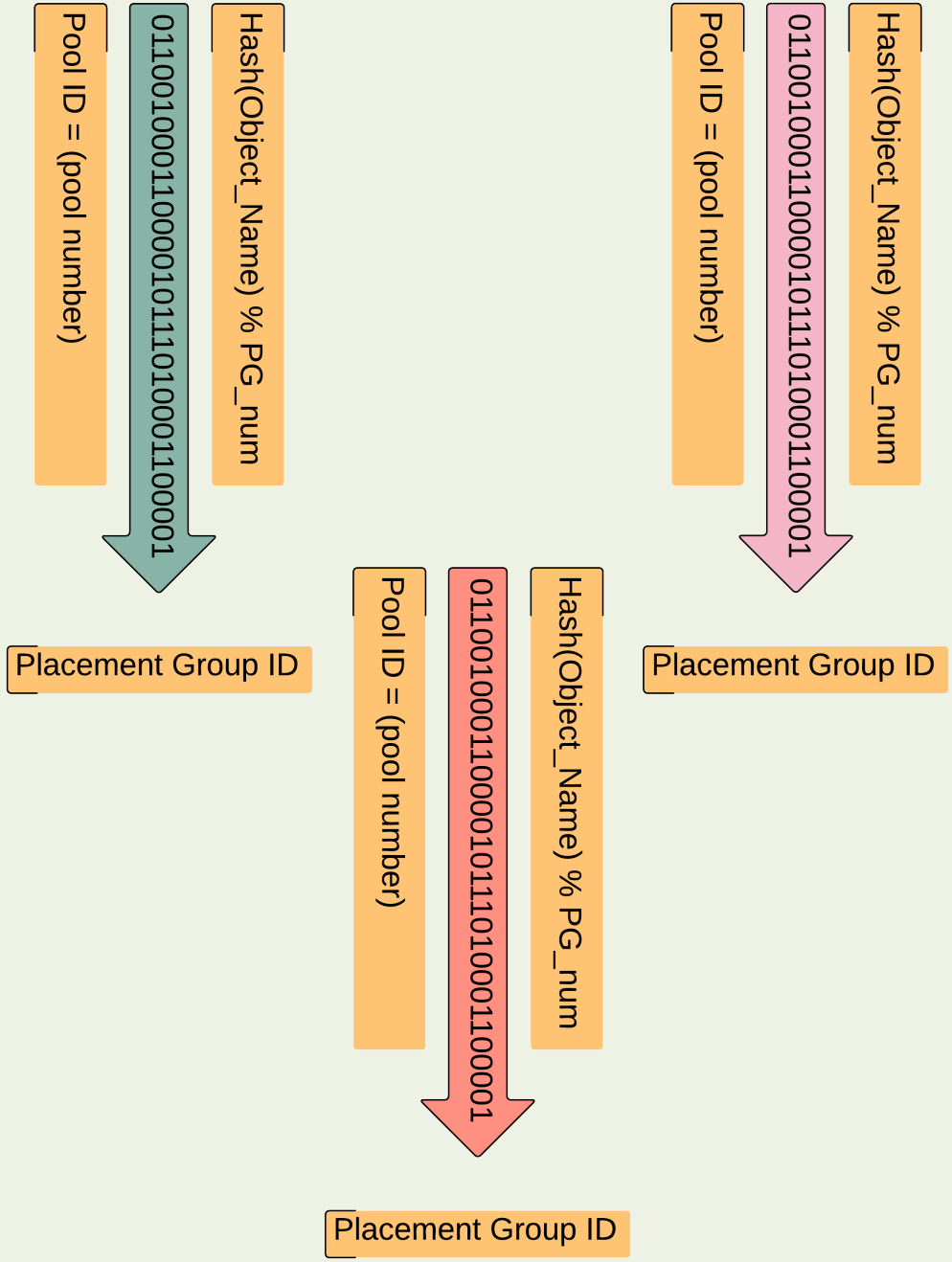
LA-UR-15-25907



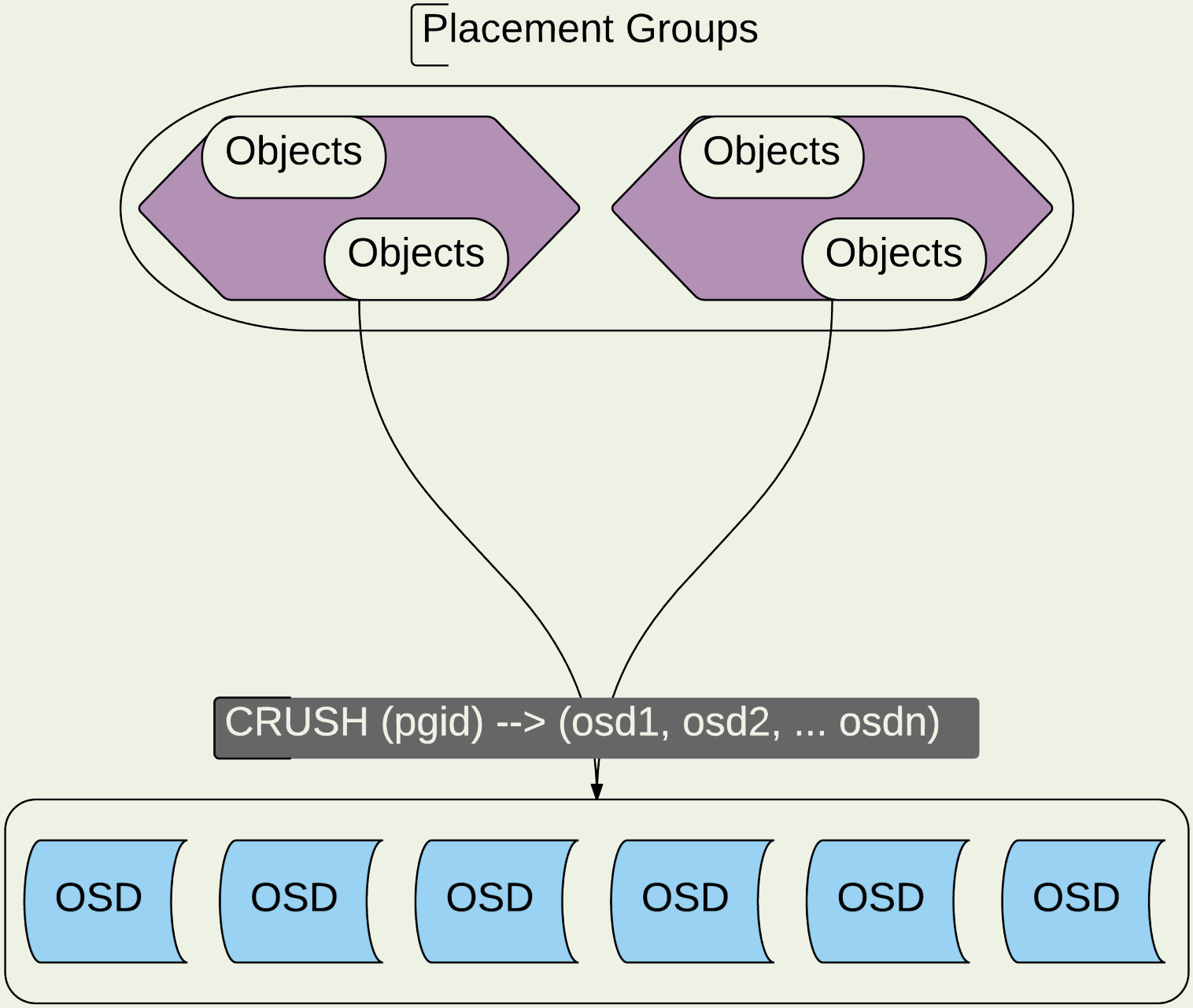
LA-UR-15-25907



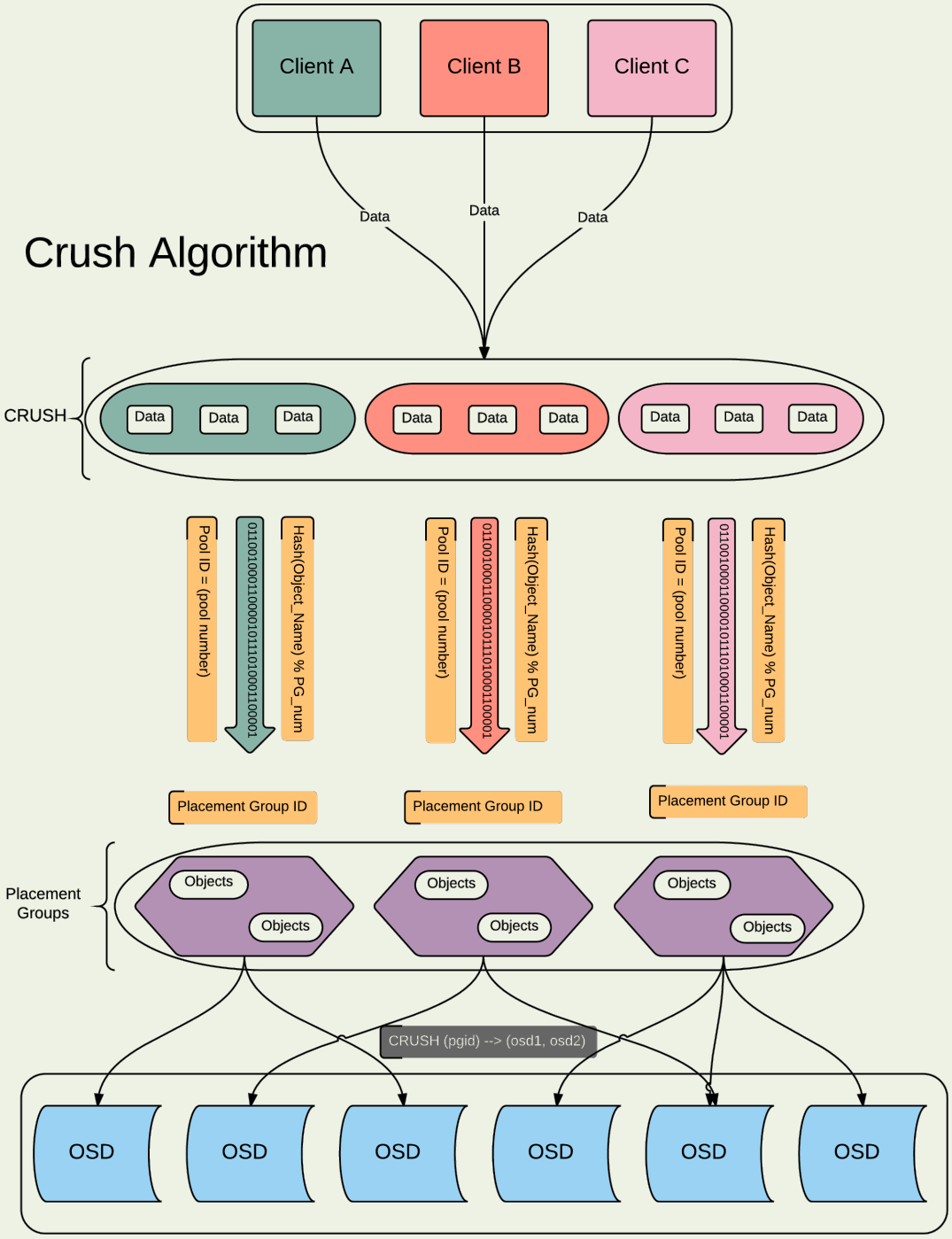
LA-UR-15-25907



LA-UR-15-25907



LA-UR-15-25907



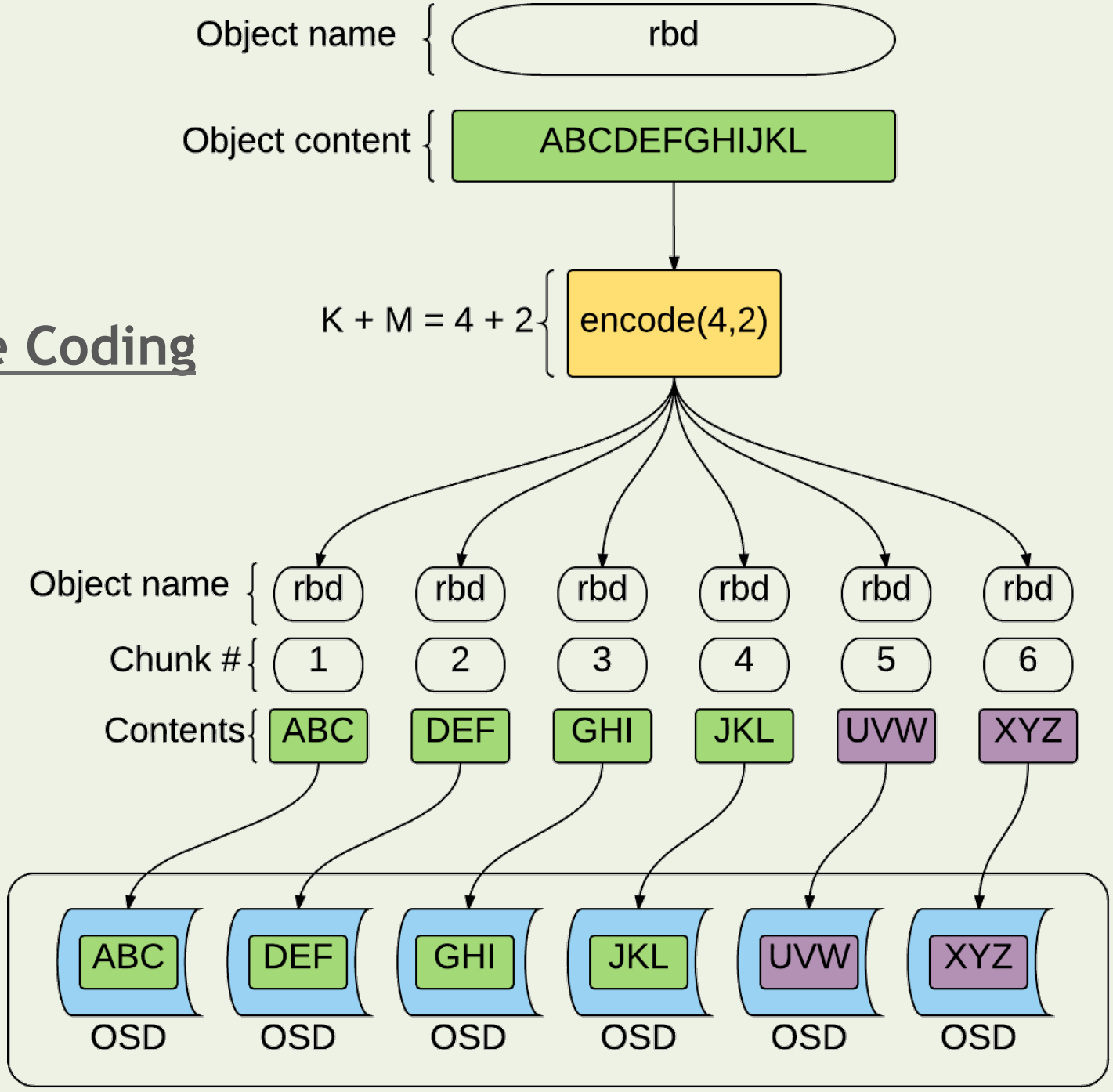
LA-UR-15-25907

Erasure Coding

- High resiliency to data loss
- Smaller storage footprint than RAID
- Data is broken up into object **chunks**
- Striped across many hard drives
- K + M values used to stripe
- Various erasure profiles

LA-UR-15-25907

Erasure Coding

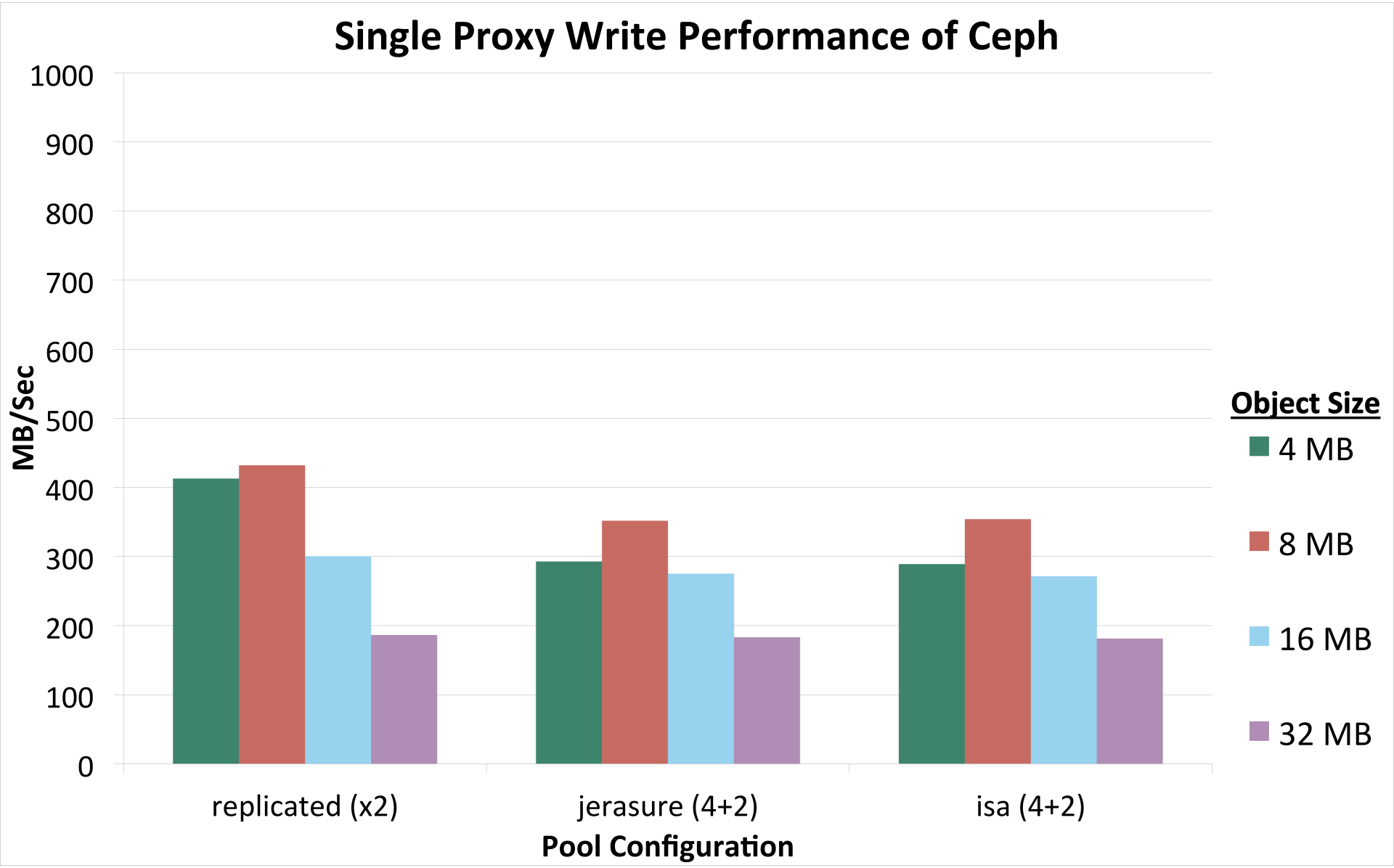


LA-UR-15-25907

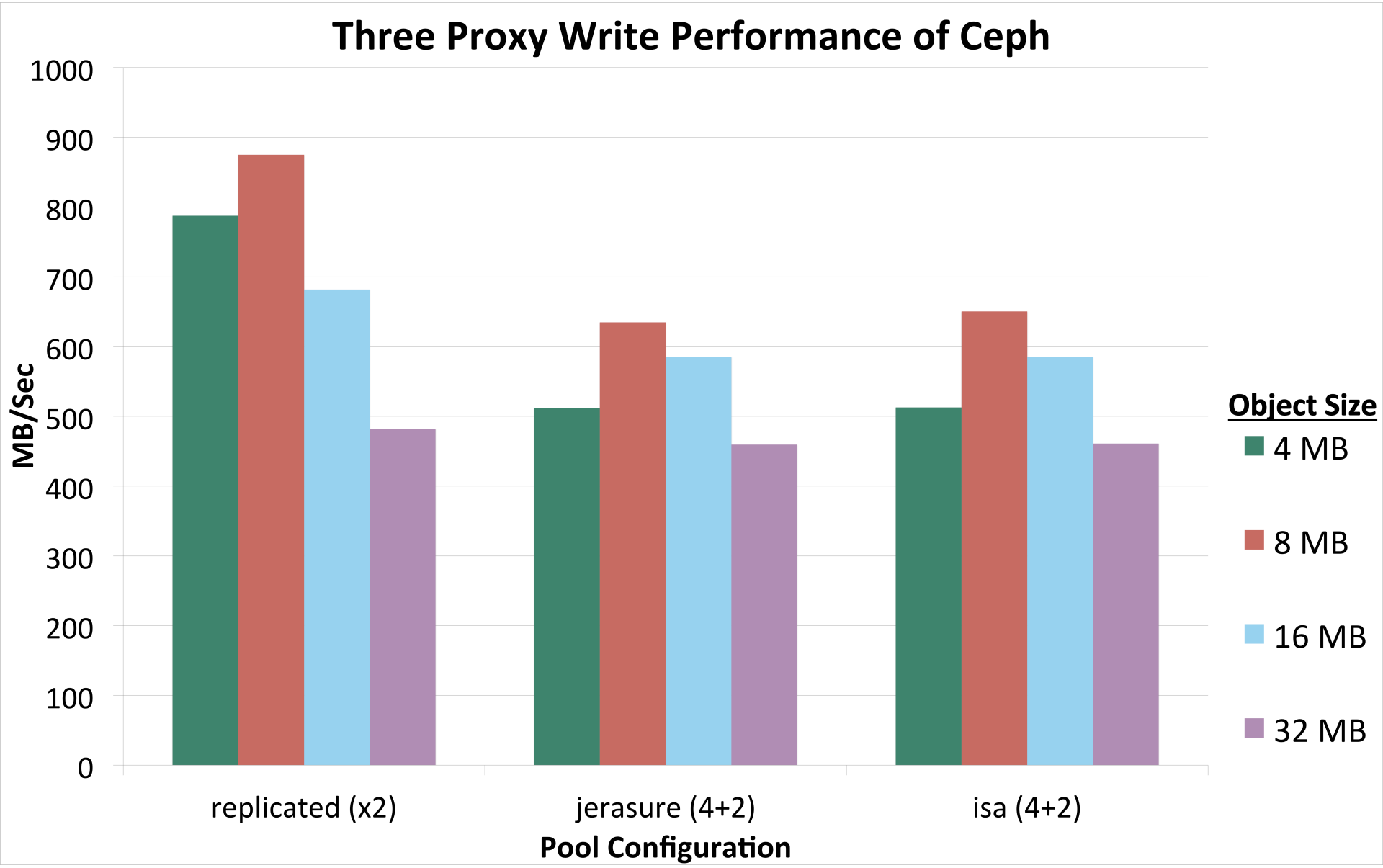
Results

- Difficult to install and configure Ceph on CentOS 6.6
- Multiple proxies write faster than a single proxy
- Replicated profile was faster than the erasure coded profiles
- K + M values did not significantly affect read and write speeds

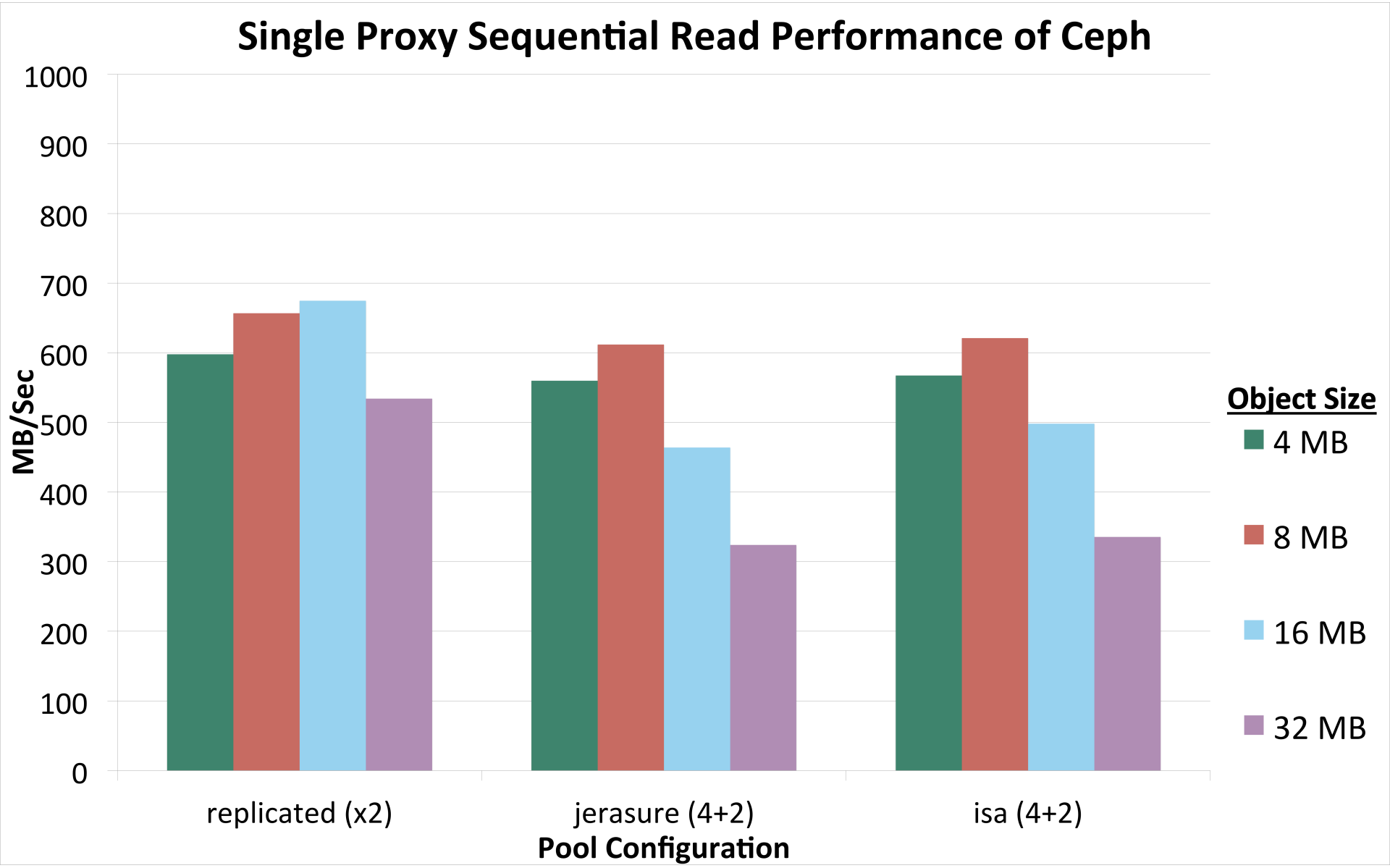
LA-UR-15-25907



LA-UR-15-25907



LA-UR-15-25907



LA-UR-15-25907

Ceph Headaches

- Documentation is inaccurate
- Nodes **must** be configured in specific order
  - Monitor → OSDs → Proxies
- Ceph was unable to recover after hardware failure
- Could only use one out of the four Infiniband lanes
- Unable to read in parallel

LA-UR-15-25907

Conclusion

- Ceph is difficult to install and configure
- Stability of Ceph needs to be improved
- Unable to recover from hardware failures during benchmarking
- Performance was promising

LA-UR-15-25907

Future Work

- Investigate bottleneck of tests
- Further explore pool configurations and PG numbers
- Look into Ceph monitoring solutions
- Test differences between ZFS/BTRFS vs XFS/EXT4

LA-UR-15-25907

Acknowledgements

- Mentors: David Bonnie, Chris Hoffman, Dominic Manno
- Instructors: Matthew Broomfield, assisted by Jarrett Crews
- Administrative Staff: Carolyn Connor, Gary Grider, Josephine Olivas, Andree Jacobson

PRObE Parallel Reconfigurable Observational Environment
New Mexico CONSORTIUM
NSF
ISTI INFORMATION SCIENCE & TECHNOLOGY INSTITUTE
Los Alamos NATIONAL LABORATORY EST. 1943
NM STATE UNIVERSITY
NEW MEXICO TECH SCIENCE · ENGINEERING · RESEARCH · UNIVERSITY
DSU DAKOTA STATE

Questions?

- Objects stores?
- Ceph and our object store?
- Installation and configuration?
- Pools and Placement groups?
- CRUSH?
- Erasure coding?
- K + M?

LA-UR-15-25907