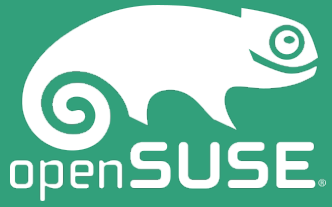


Samba and Ceph

Release the Kraken!

David Disseldorp
ddiss@samba.org

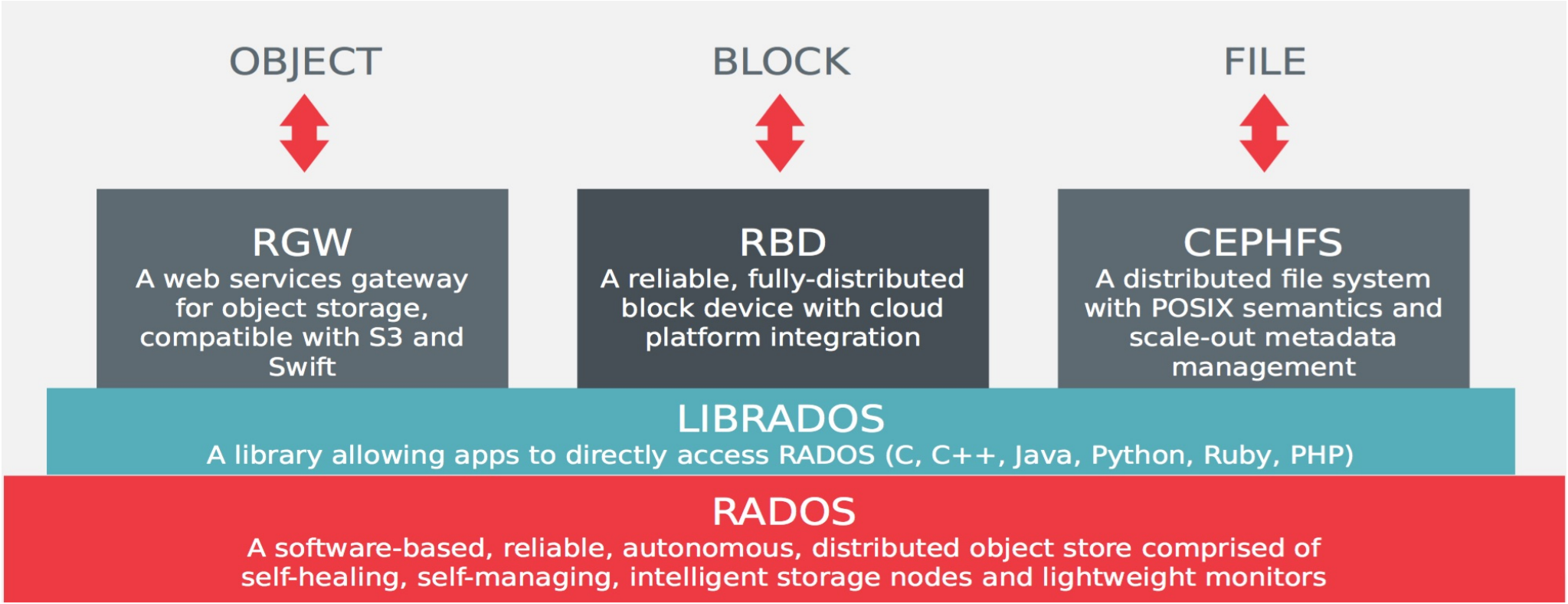
Agenda

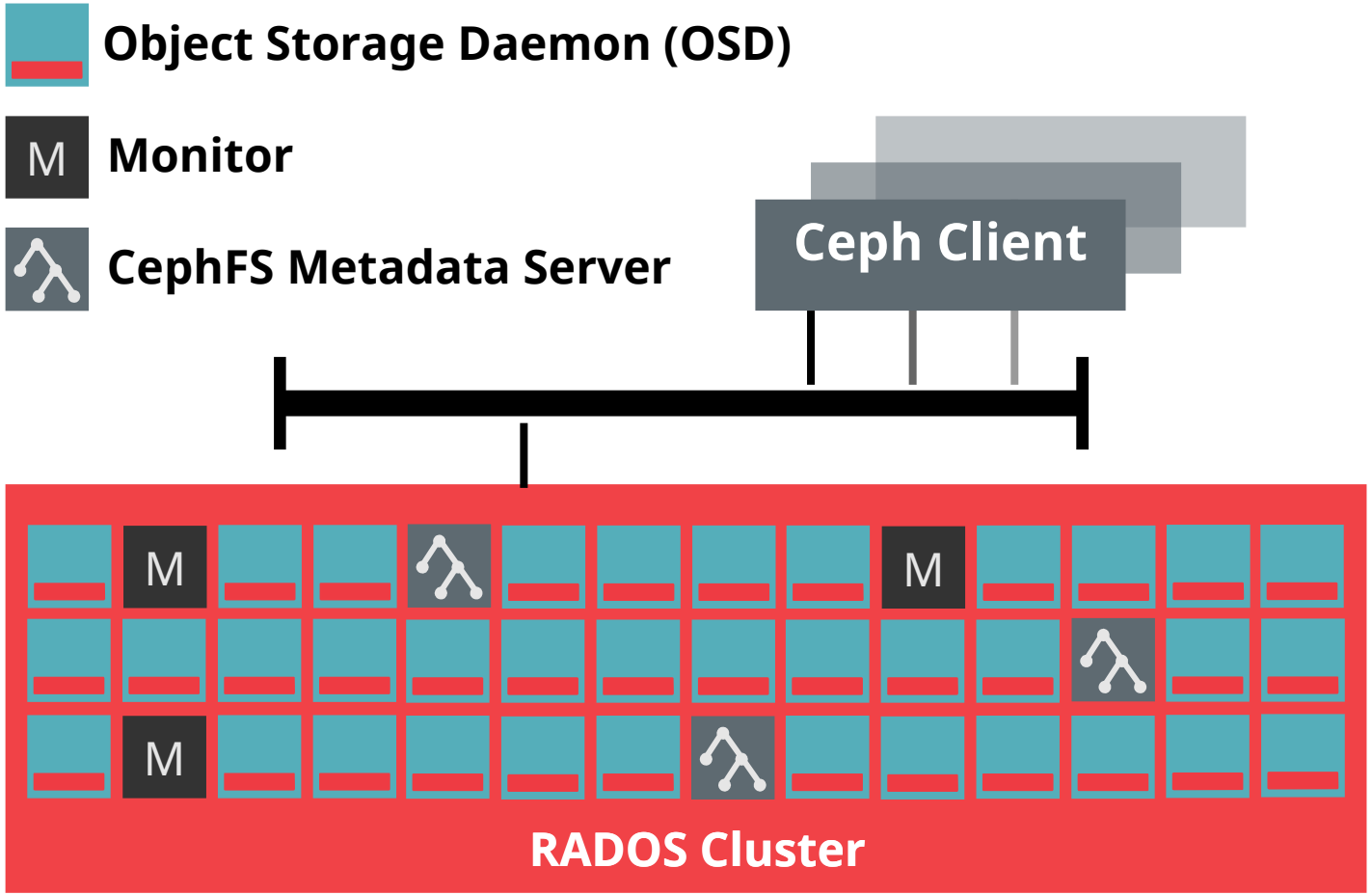
- Ceph Overview
- State of Samba Integration
  - Performance
- Outlook

Ceph

- Distributed storage system
  - Scalable
  - Fault tolerant
  - Performant
  - Self-healing and self-managing
  - Runs on commodity hardware
  - Mature
- Various client access mechanisms
  - All layered atop a Reliable Autonomic Distributed Object Store (RADOS)

Ceph Architecutre

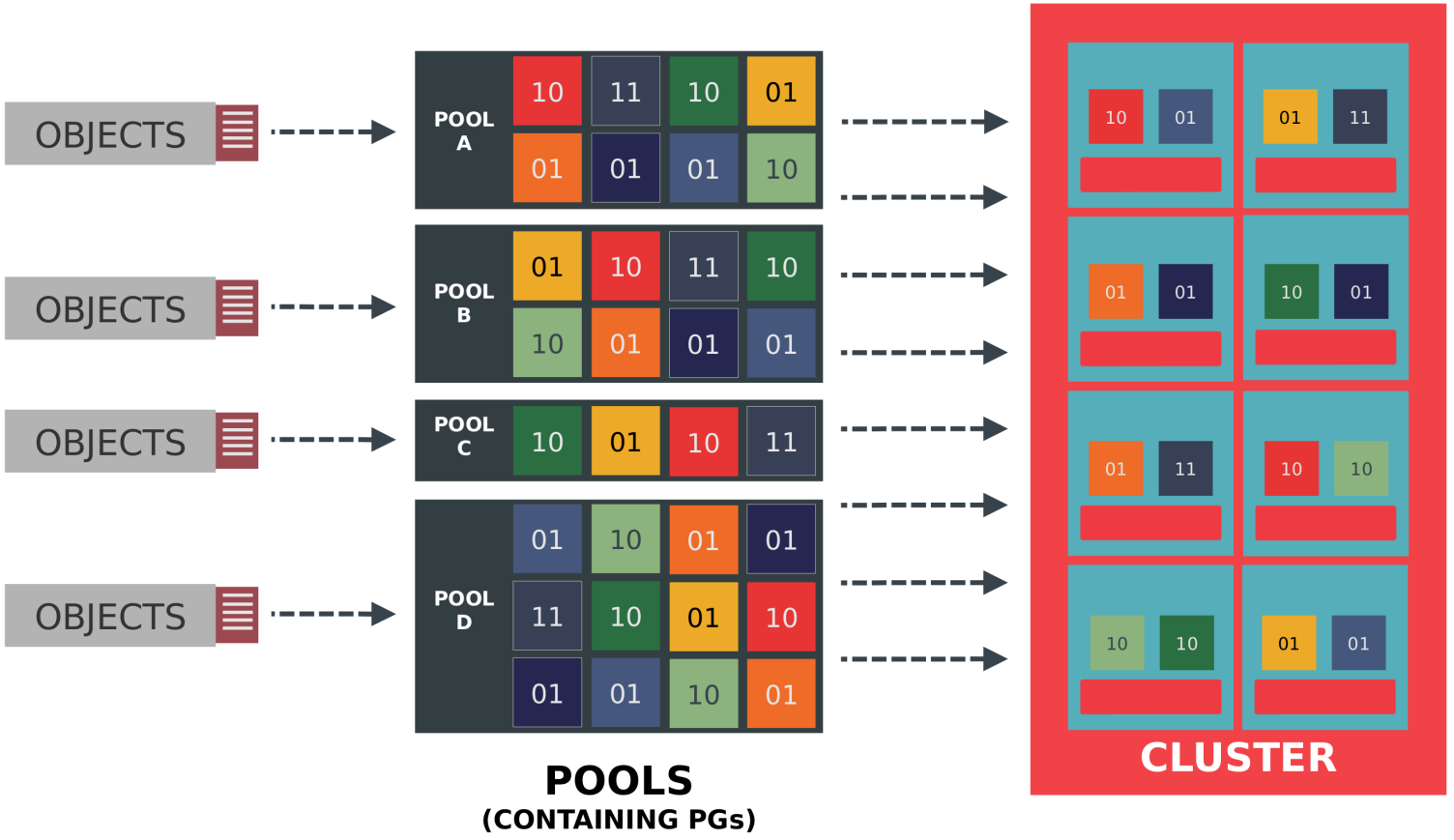



Object Storage Daemon (OSD)
M
Monitor
CephFS Metadata Server
Ceph Client
M
M
M
RADOS Cluster

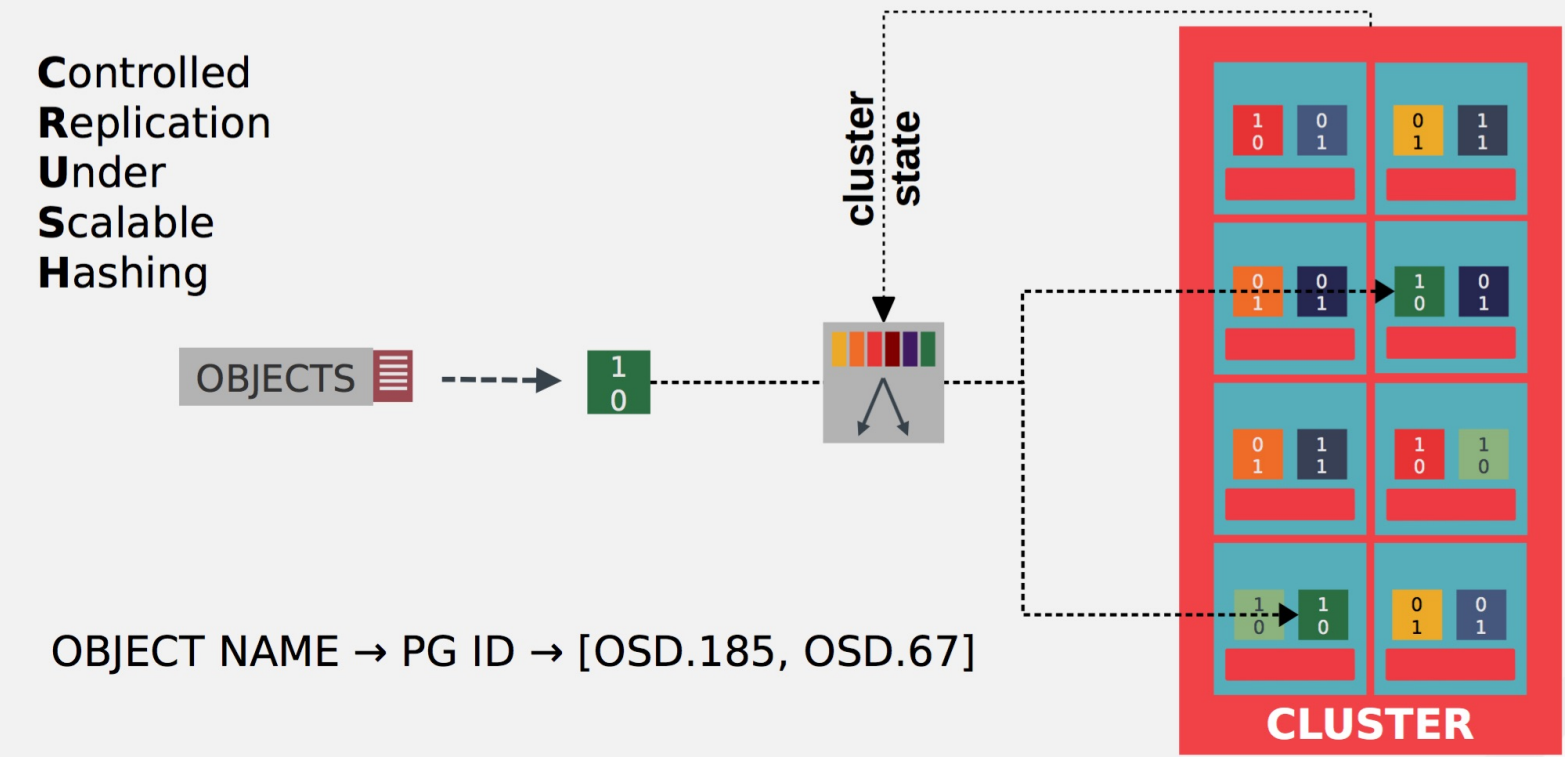
Components

- Object Storage Daemon (OSD)
  - Exposes underlying storage to clients
    - Objects with data and KV metadata
  - One per disk
    - Faster devices can be used for metadata / WAL
  - Handles data replication and recovery
- Monitor
  - Provide consensus on cluster state

Ceph Placement



Ceph Placement



Replication

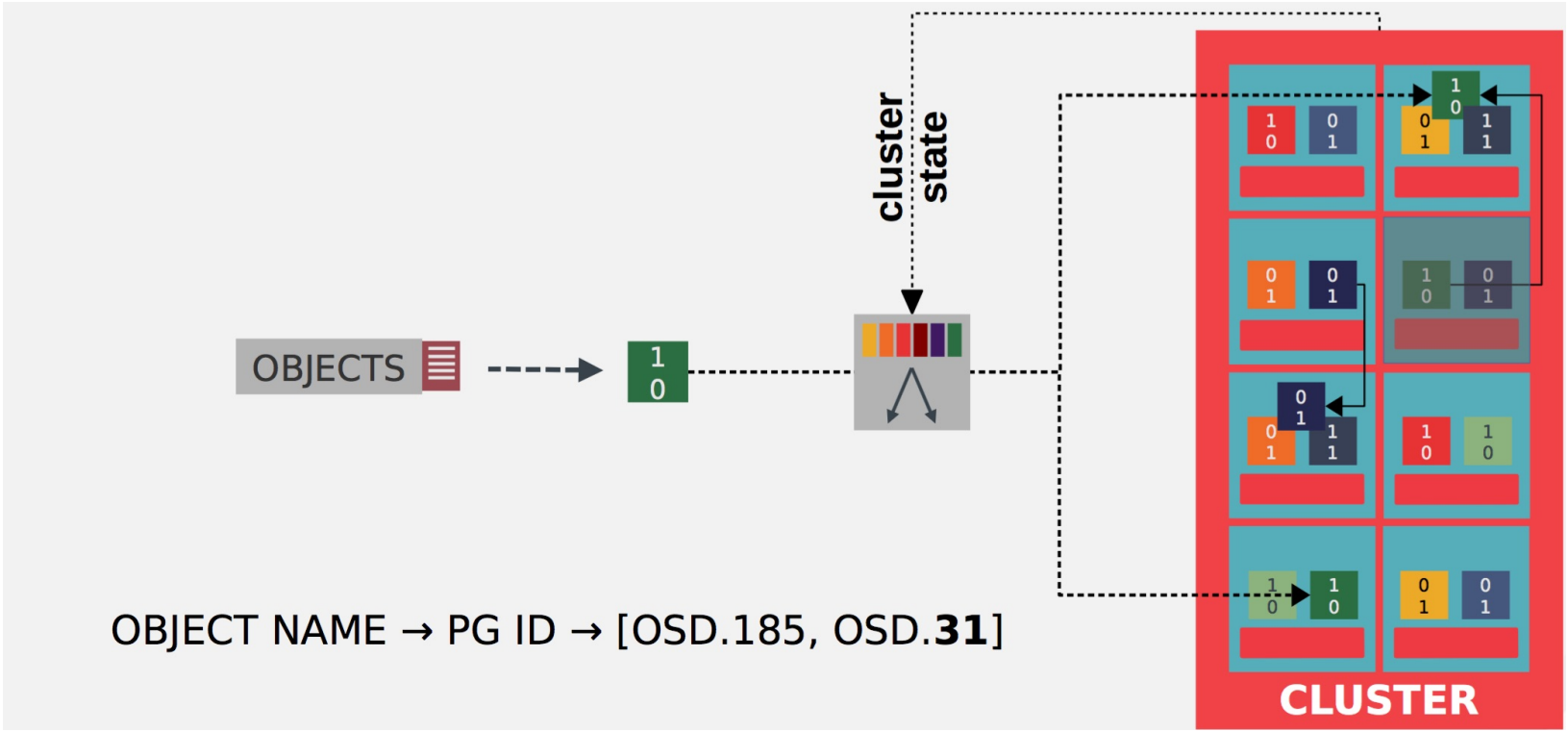
- Client determines PG and corresponding OSDs
  - Sends object I/O to primary OSD
  - Writes acknowledged only after writing to all replicas
- Pools can be replicated or erasure coded
  - User-specified redundancy levels and failure domains
- Private OSD network used for replication traffic

Failure Handling

- Monitors and OSDs check state of other OSDs
  - Following outage, PG is assigned to a new node
  - Backfill from peers

- Periodic scrubbing of data and metadata

Ceph Placement



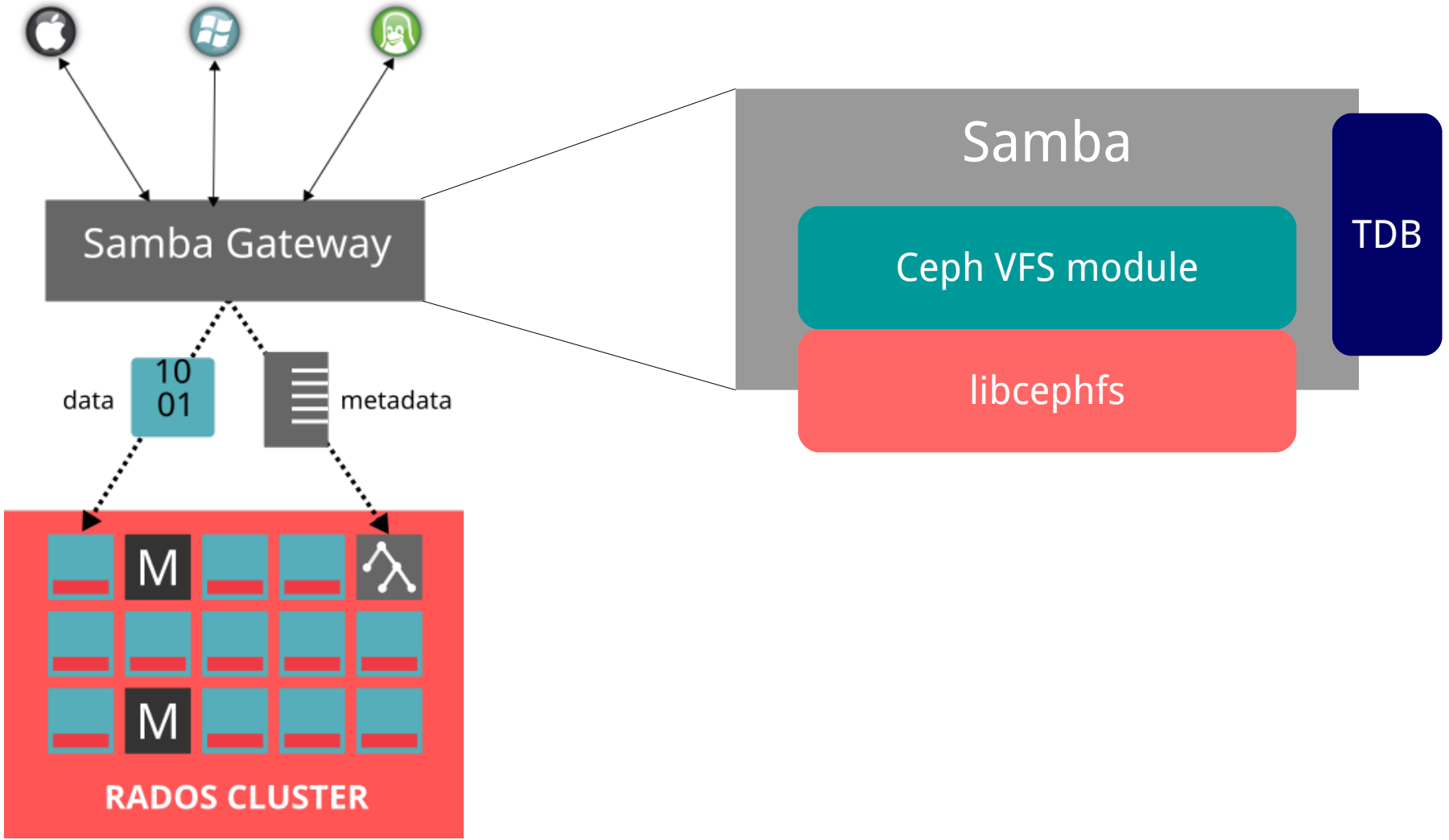
CephFS

- POSIX compatible clustered filesystem atop RADOS
- MDSes manage filesystem namespace
  - Active/Passive or Active/Active redundancy
- Linux kernel and user-space clients
- Snapshots
- Directory to pool mappings

CephFS

- Basic quotas
- Client caching
  - Fine grained
  - Capabilities granted and revoked by MDS

Samba Integration


Samba Gateway
data
10
01
metadata
M
M
RADOS CLUSTER
Samba
Ceph VFS module
libcephfs
TDB

Samba Ceph Integration

- CephFS module for Samba: v*fs_ceph*
  - Added in 2013 by Inktank
  - Maps SMB file and directory I/O to libcephfs API calls
- Static cephx credentials
  - Regardless of Samba authenticated user
  - User configurable via *smb.conf*
- POSIX ACLs

Samba Ceph Integration

- RADOS clustered mutex helper for CTDB
  - Removes recovery lock mount dependency
- Ceph librados service integration *(coming soon)*
  - Register service with manager daemon

Testing

- Ceph *vstart*
  - Deploy mock cluster from source
- Samba smbtorture
- cifs.ko fstests

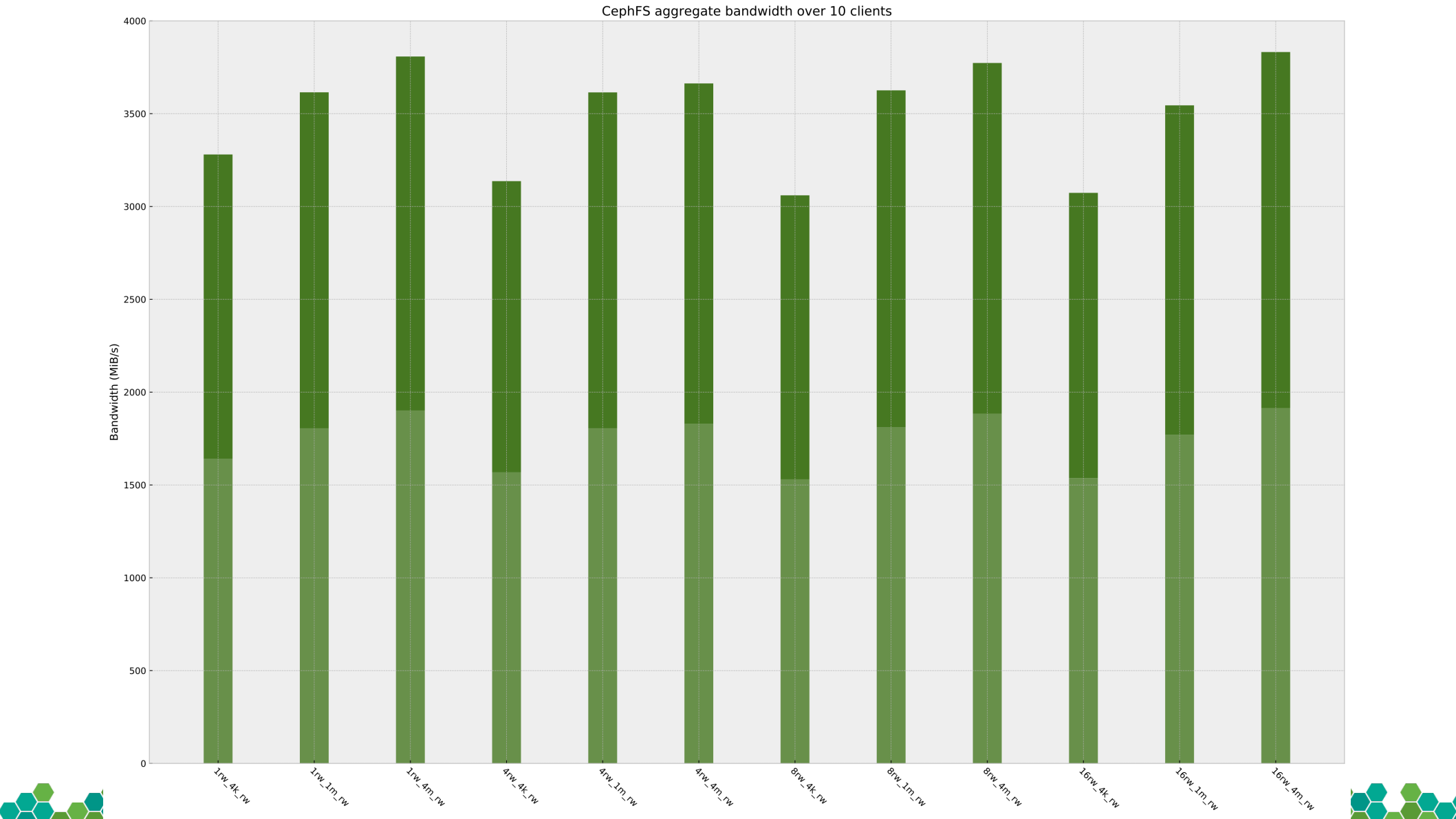
Performance

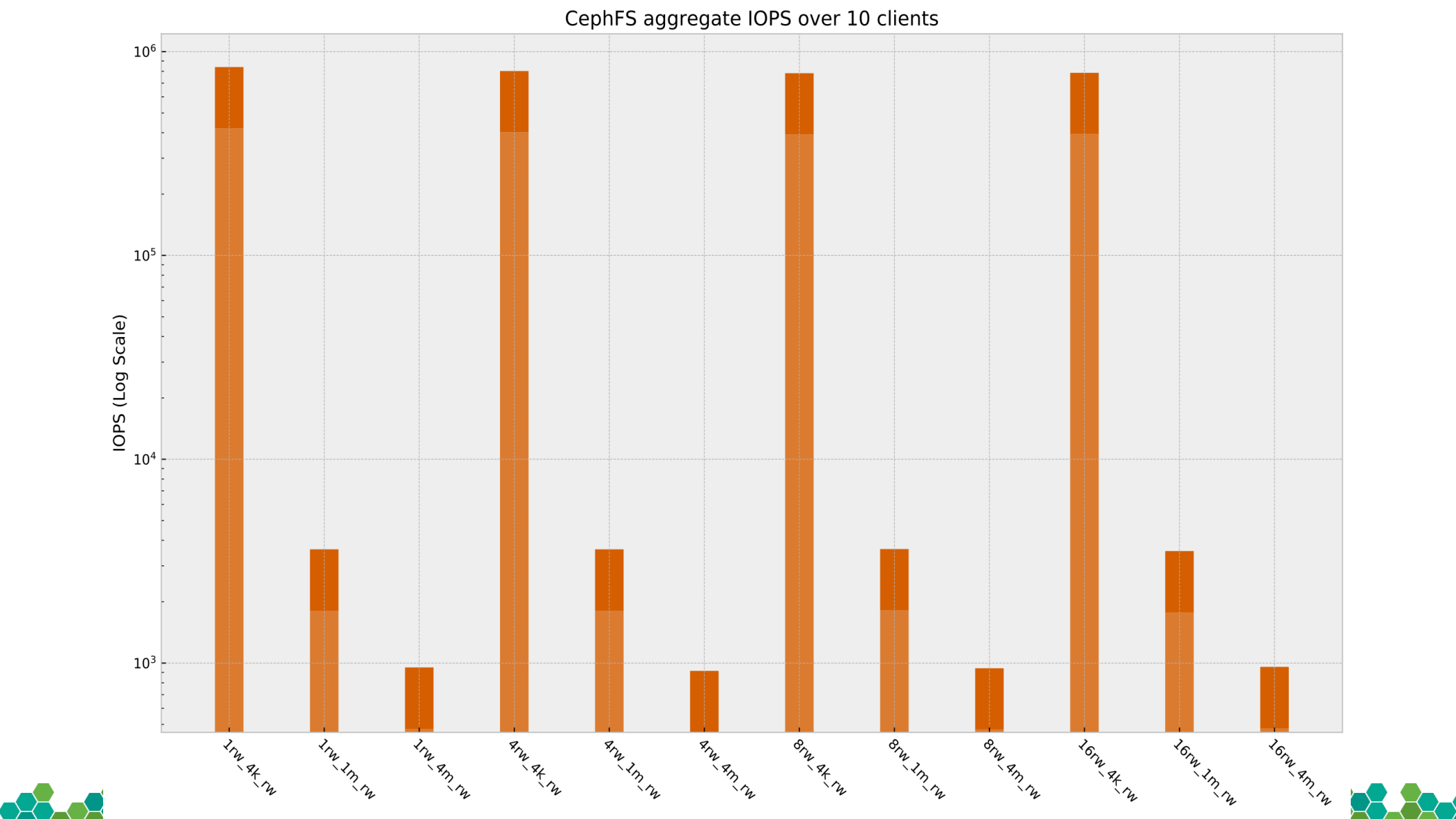
Performance: Samba vs CephFS

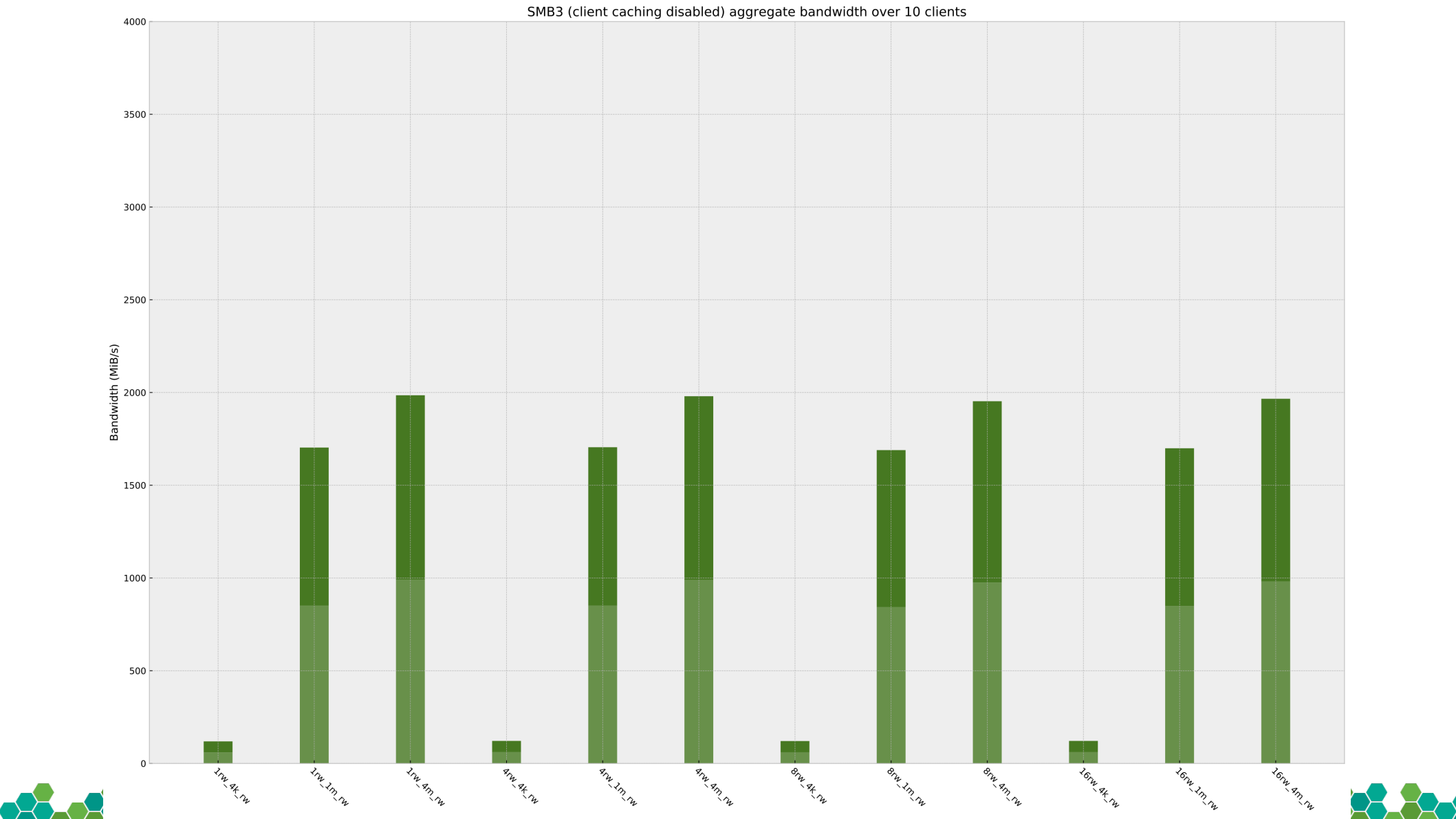
- Preliminary results!
- Environment:
  - Ceph Version 12.2.2
  - Samba 4.6.9
    - Three Samba gateways
    - *vfs_ceph*
    - Non-overlapping share paths
  - Linux cifs.ko client
    - 4.4 kernel with many backports
    - SMB 3.0 mount

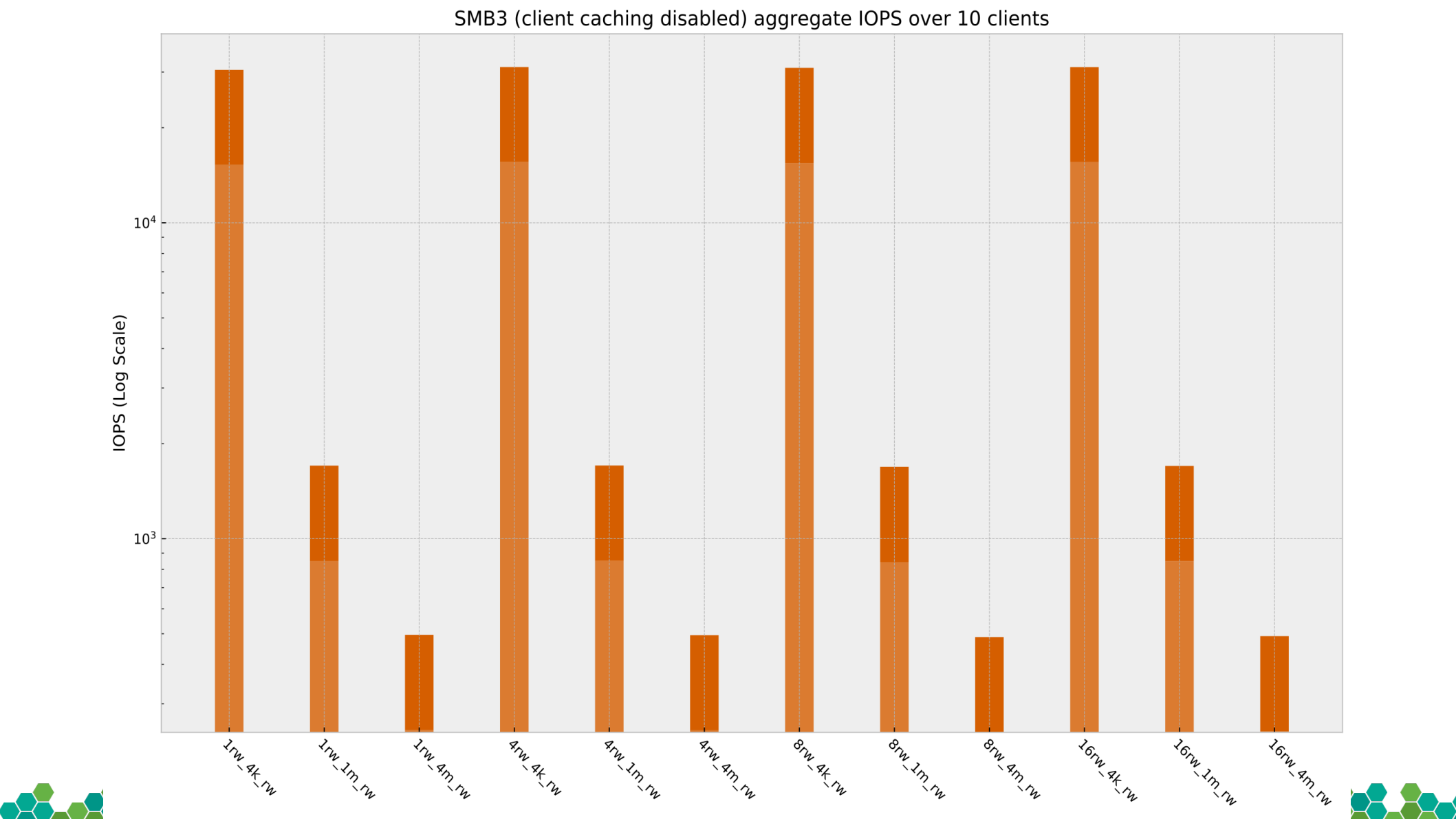
Hardware

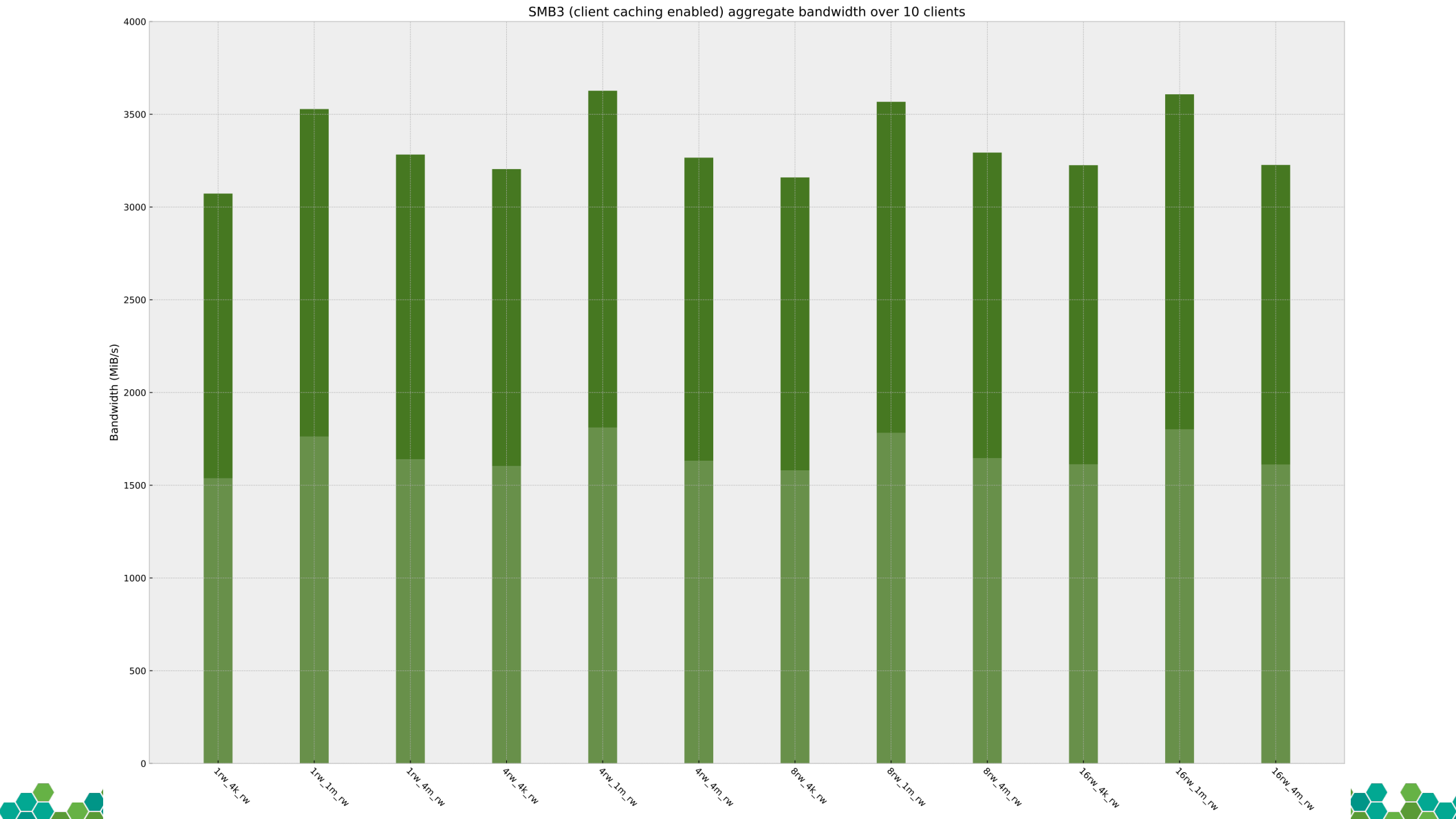
- Ceph setup on 8 nodes
  - 5 OSD nodes – 24 cores – 128 GB RAM
  - 3 MON/MDS nodes – 24 cores – 128 GB RAM
  - 6 OSD daemons per node – Bluestore – SSD/NVME journals
- 10 client nodes
  - 16 cores – 16 GB RAM
- Network interconnect
  - Public network 10Gbit/s
  - Cluster network 100Gbit/s

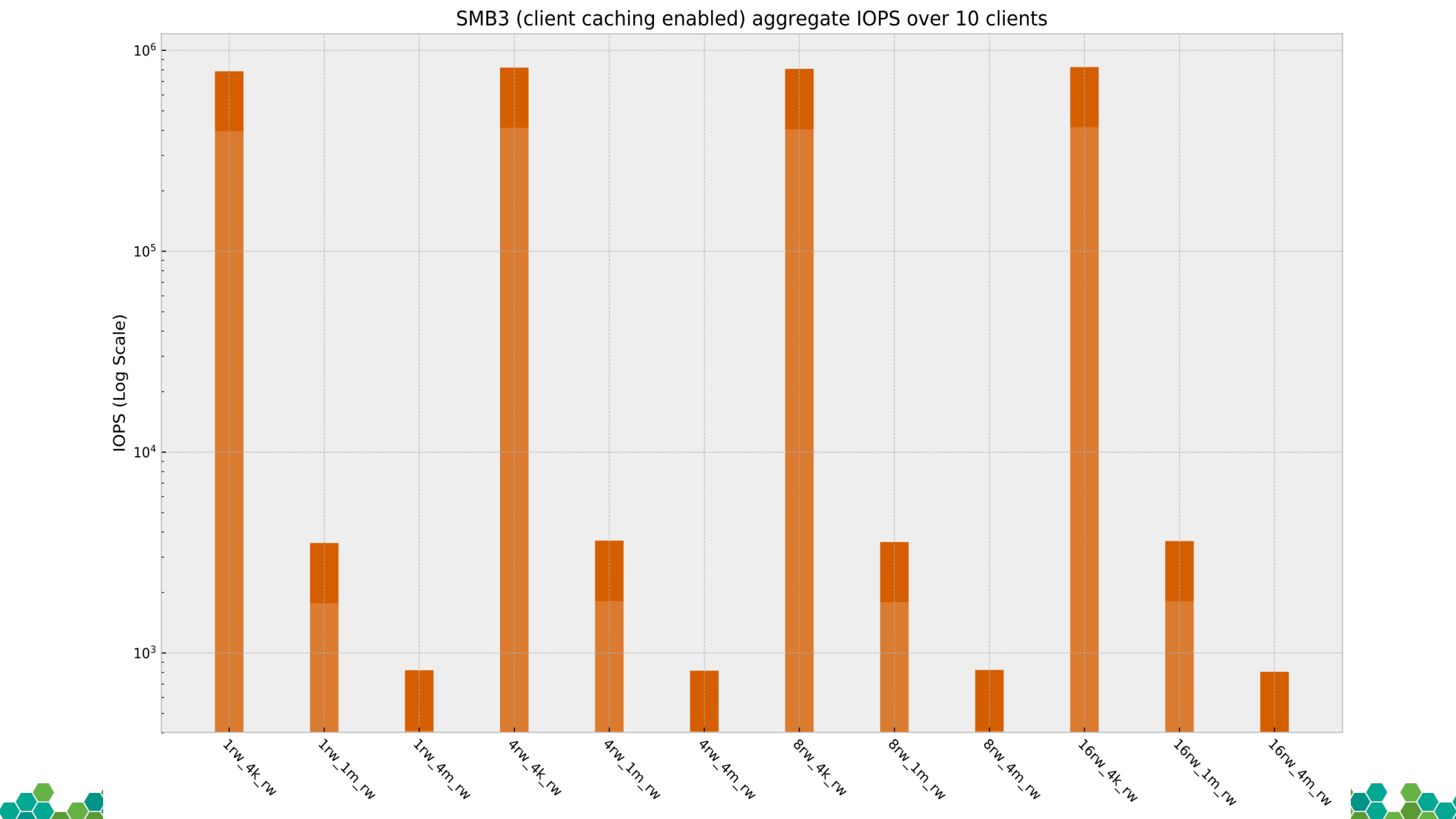

CephFS aggregate bandwidth over 10 clients
Bandwidth (MiB/s)
0
500
1000
1500
2000
2500
3000
3500
4000
1rw_4k_rw
1rw_1m_rw
1rw_4m_rw
4rw_4k_rw
4rw_1m_rw
4rw_4m_rw
8rw_4k_rw
8rw_1m_rw
8rw_4m_rw
16rw_4k_rw
16rw_1m_rw
16rw_4m_rw


CephFS aggregate IOPS over 10 clients
IOPS (Log Scale)
10^6
10^5
10^4
10^3
1rw_4k_rw
1rw_1m_rw
1rw_4m_rw
4rw_4k_rw
4rw_1m_rw
4rw_4m_rw
8rw_4k_rw
8rw_1m_rw
8rw_4m_rw
16rw_4k_rw
16rw_1m_rw
16rw_4m_rw


SMB3 (client caching disabled) aggregate bandwidth over 10 clients
Bandwidth (MiB/s)
4000
3500
3000
2500
2000
1500
1000
500
0
1rw_4k_rw
1rw_1m_rw
1rw_4m_rw
4rw_4k_rw
4rw_1m_rw
4rw_4m_rw
8rw_4k_rw
8rw_1m_rw
8rw_4m_rw
16rw_4k_rw
16rw_1m_rw
16rw_4m_rw


SMB3 (client caching disabled) aggregate IOPS over 10 clients
IOPS (Log Scale)
10^4
10^3
1rw_4k_rw
1rw_1m_rw
1rw_4m_rw
4rw_4k_rw
4rw_1m_rw
4rw_4m_rw
8rw_4k_rw
8rw_1m_rw
8rw_4m_rw
16rw_4k_rw
16rw_1m_rw
16rw_4m_rw


SMB3 (client caching enabled) aggregate bandwidth over 10 clients
Bandwidth (MiB/s)
0
500
1000
1500
2000
2500
3000
3500
4000
1rw_4k_rw
1rw_1m_rw
1rw_4m_rw
4rw_4k_rw
4rw_1m_rw
4rw_4m_rw
8rw_4k_rw
8rw_1m_rw
8rw_4m_rw
16rw_4k_rw
16rw_1m_rw
16rw_4m_rw


SMB3 (client caching enabled) aggregate IOPS over 10 clients
IOPS (Log Scale)
10^6
10^5
10^4
10^3
1rw_4k_rw
1rw_1m_rw
1rw_4m_rw
4rw_4k_rw
4rw_1m_rw
4rw_4m_rw
8rw_4k_rw
8rw_1m_rw
8rw_4m_rw
16rw_4k_rw
16rw_1m_rw
16rw_4m_rw

Challenges and Future

Challenges

- Cross-protocol client support
  - Coherent client caching
    - Map leases to CephFS *FILE* and *AUTH* capabilities
    - New libcephfs delegations API
  - Shared (NFS, CephFS) ACL model

- Unified authentication and user mapping
  - Use Kerberos / AD for Samba gateway and cephx

Challenges

- libcephfs asynchronous I/O
- Multichannel support
  - Experimental in upstream Samba
  - Not integrated with CTDB
- Automated deployment

Challenges

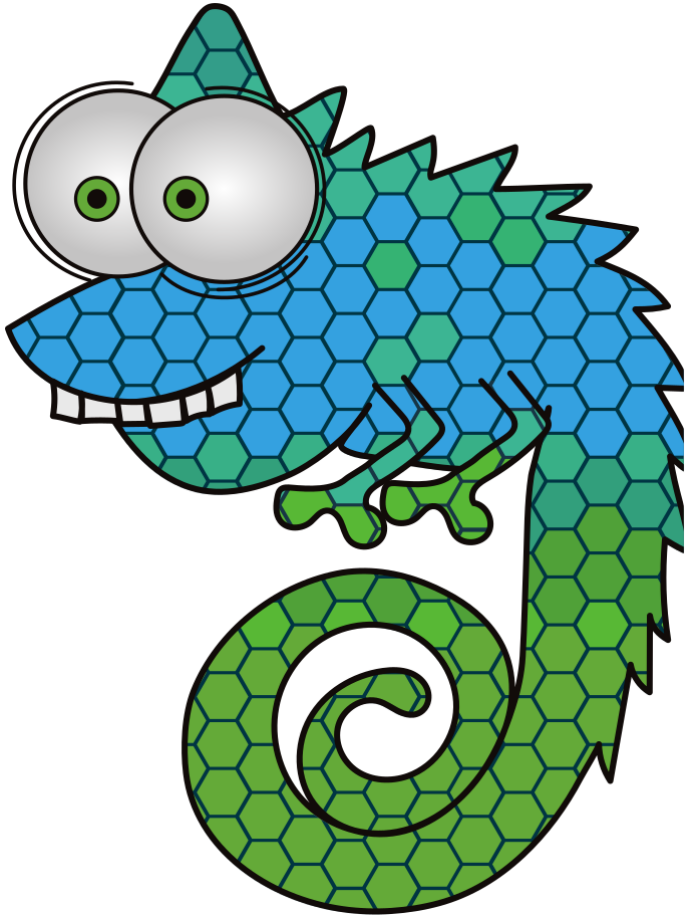
- Witness protocol
  - Continuous availability of SMB shares
  - Advertise Samba cluster state to clients
  - Transparent client failover
  - Load balancing

Samba: Future

- Ceph backed key-value store for Samba
- Replace or modify CTDB
  - Rocksdb?
  - Samba database API demanding
    - Multiple processes and writers
    - Record locking and transactions

References

- Samba: https://samba.org/
- CTDB: https://ctdb.samba.org/
- SMB 3.1.1 encryption: https://technet.microsoft.com/en-us/library/dn551363(v=ws.11).aspx
- Multichannel deployment: https://technet.microsoft.com/en-us/library/dn610980(v=ws.11).aspx
- Witness Protocol: http://www.sambaxp.org/archive_data/SambaXP2015-SLIDES/wed/track1/sambaxp2015-wed-track1-Guenther_Deschner-ImplementingTheWitnessProtocolInSamba.pdf
- Samba Multichannel Blocker Bug: https://bugzilla.samba.org/show_bug.cgi?id=11897
- CephFS cache flags: https://jtlayton.wordpress.com/2016/09/01/cephfs-and-the-nfsv4-change-attribute/
- Greg Farnum: Intro to Ceph, The Distributed Storage System
- Placement diagrams: http://yauuu.me/ride-around-ceph-crush-map.html



Join Us at www.opensuse.org

License

General Disclaimer

Credits

Template
Richard Brown
rbrown@opensuse.org

Design & Inspiration
openSUSE Design Team
http://opensuse.github.io/branding-guidelines/