
ceph days
LONDON 2026

ceph days
LONDON 2026

Cuttle out the middleman

EC Direct Reads

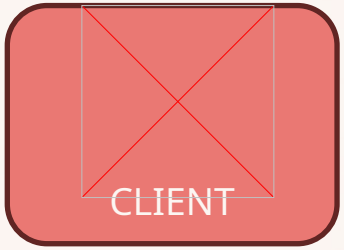
Alex Ainscow

ceph days
LONDON 2026

Small Reads

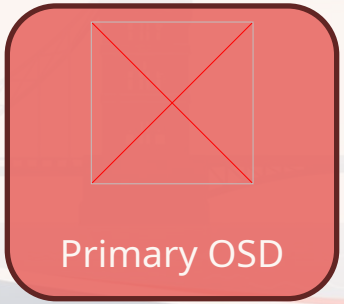
ceph days LONDON 2026

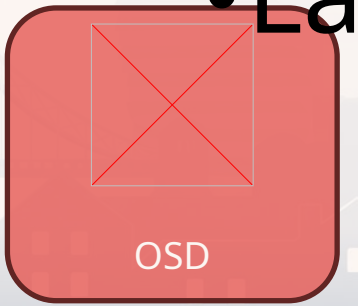
As-is Read Path



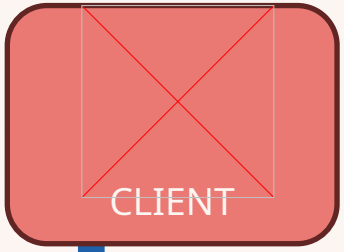
READ: 16KiB
OFFSET = 16K

- Messages: 0
- Data: 0KiB
- Latency: 0t



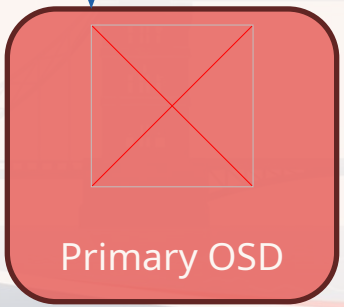


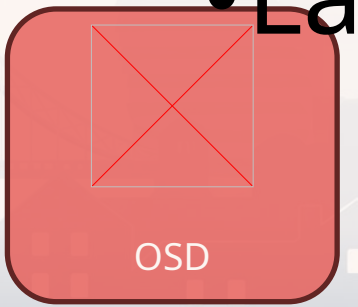
As-is Read Path



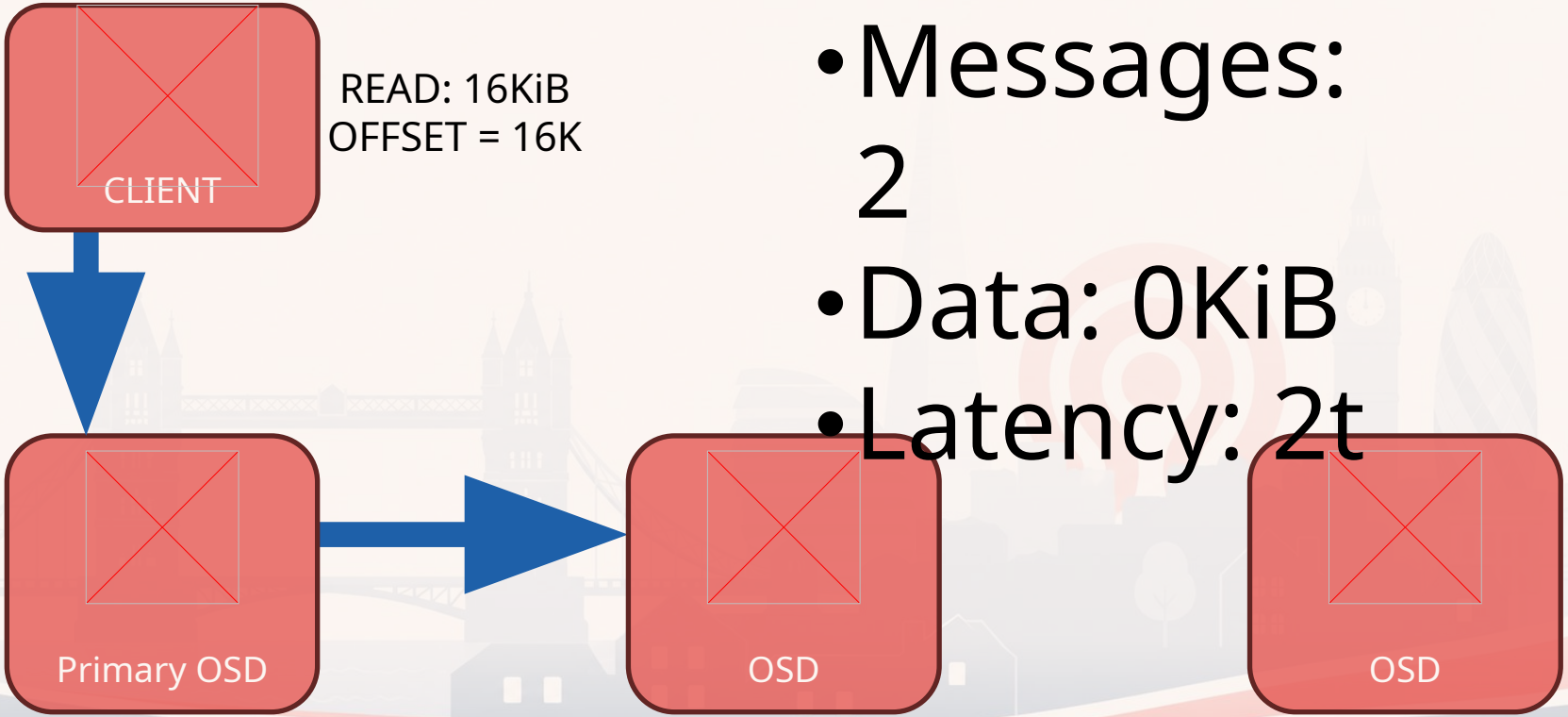
READ: 16KiB
OFFSET = 16K

- Messages: 1
- Data: 0KiB
- Latency: 1t



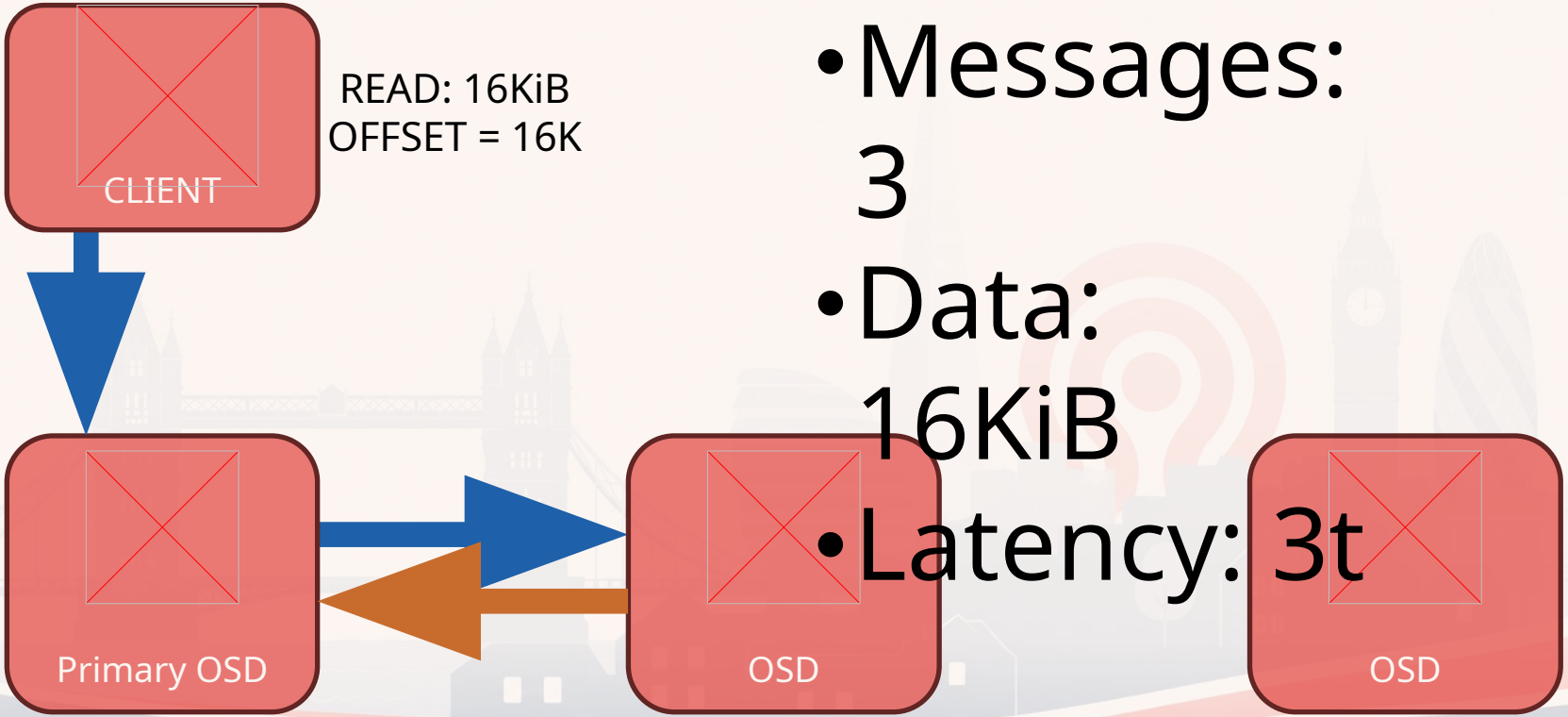


As-is Read Path



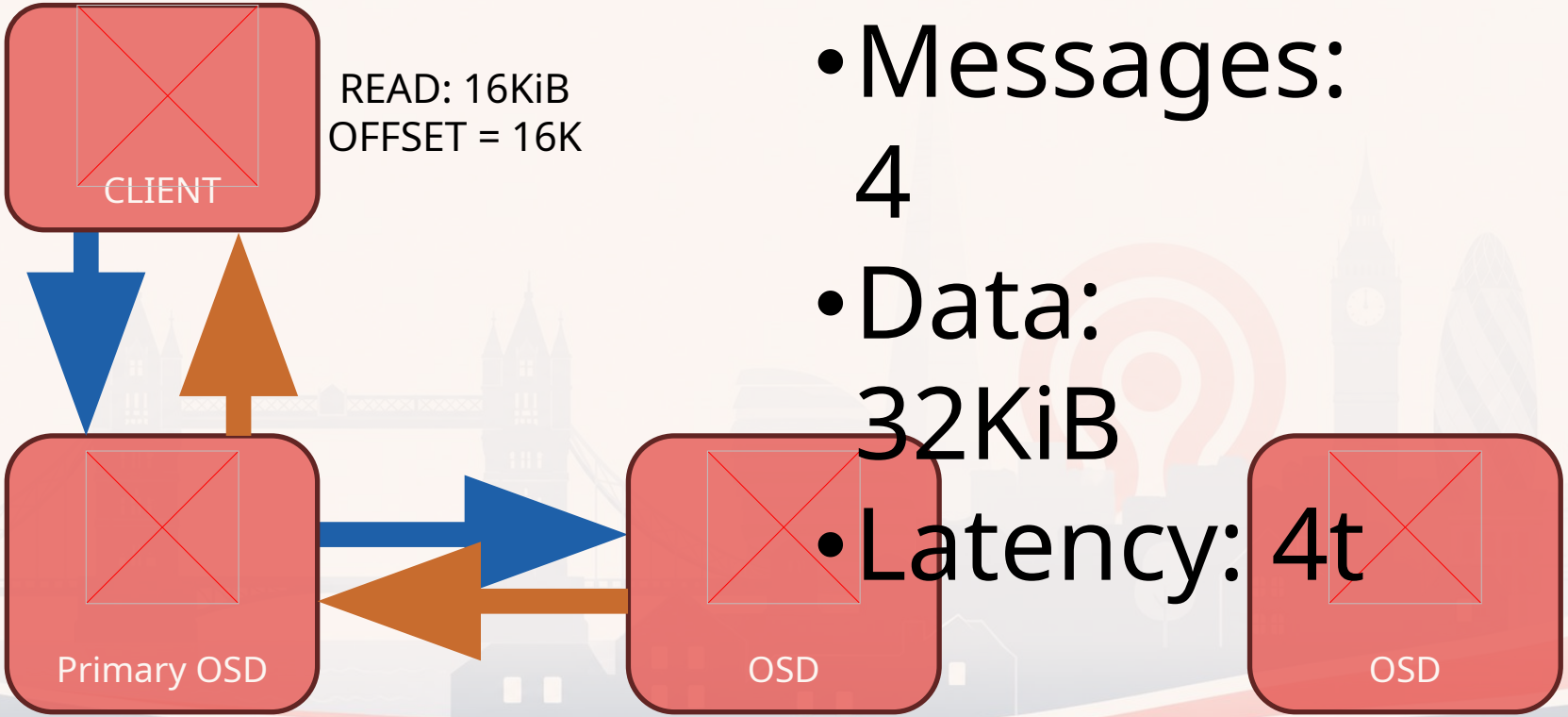
- Messages: 2
- Data: 0KiB
- Latency: 2t

As-is Read Path



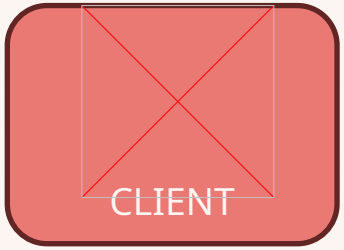
- Messages: 3
- Data: 16KiB
- Latency: 3t

As-is Read Path

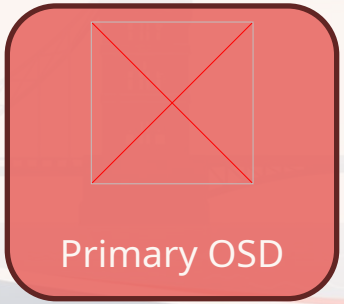


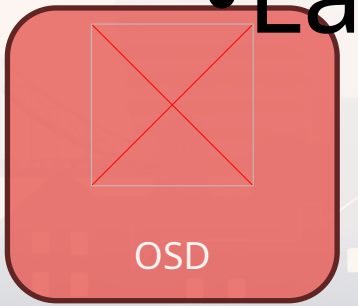
- Messages: 4
- Data: 32KiB
- Latency: 4t

Direct Read Path

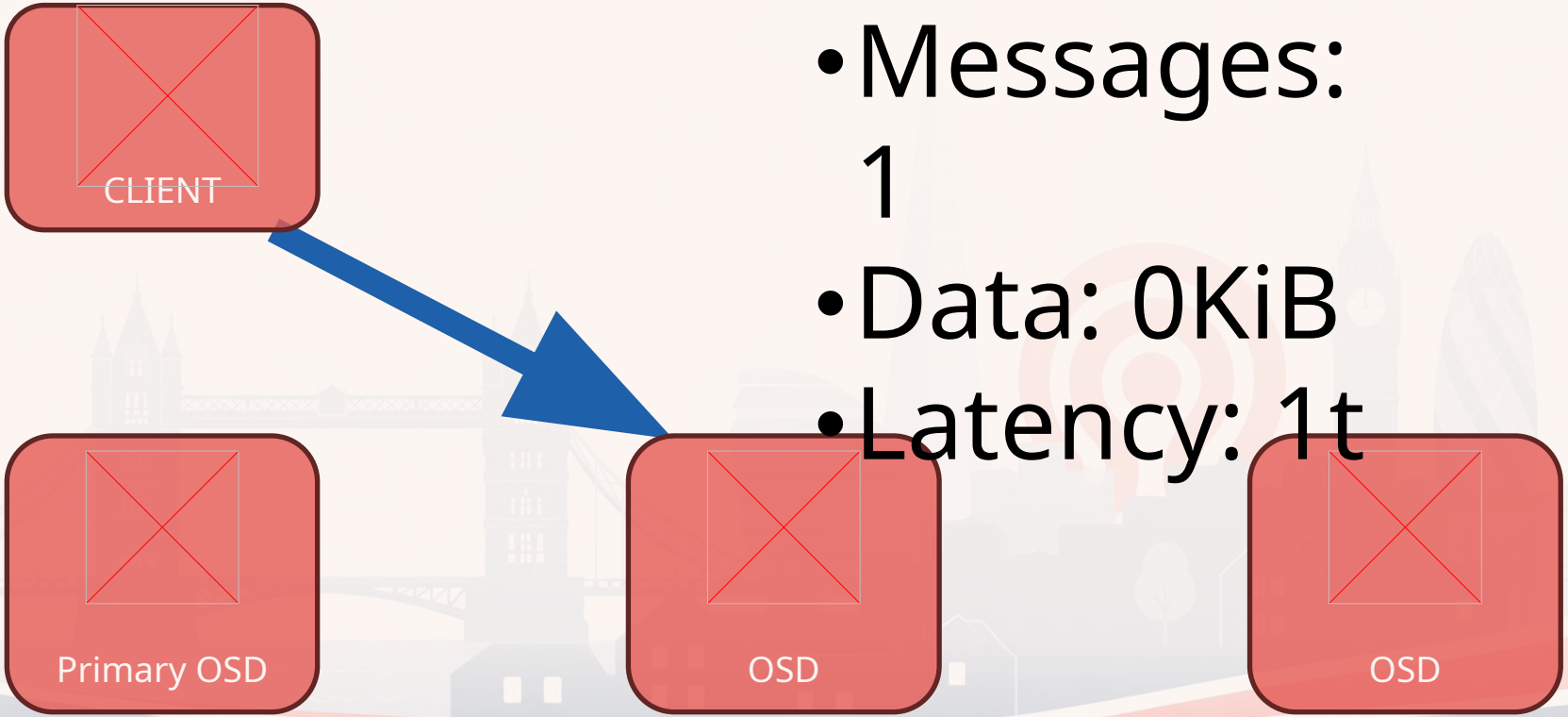


- Messages: 0
- Data: 0KiB
- Latency: 0t



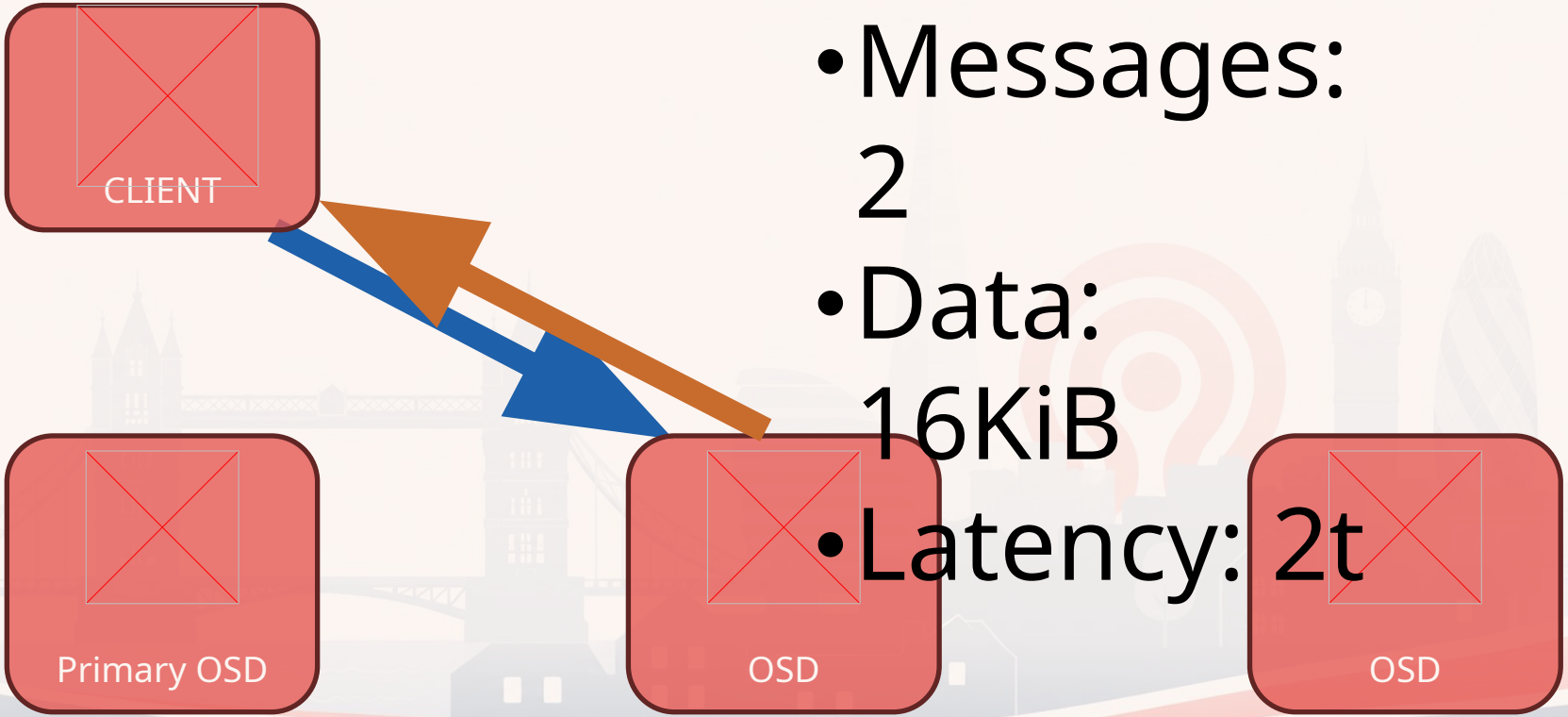


Direct Read Path



- Messages: 1
- Data: 0KiB
- Latency: 1t

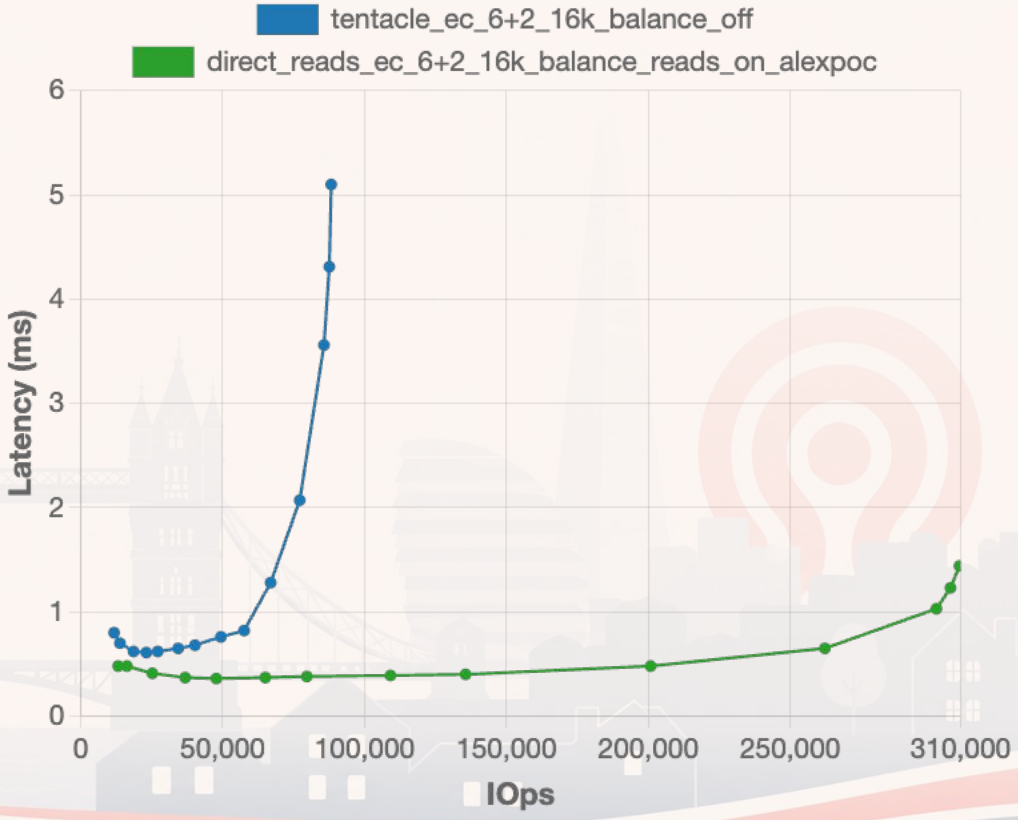
Direct Read Path



- Messages: 2
- Data: 16KiB
- Latency: 2t

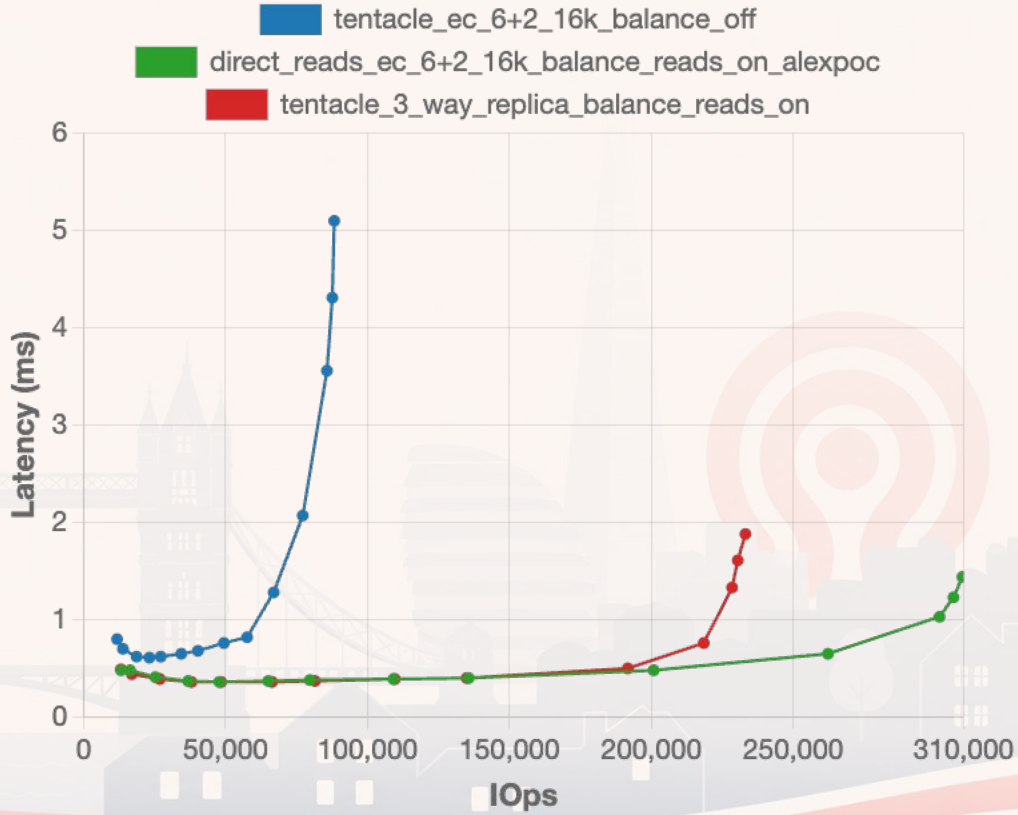
ceph days LONDON 2026

Performance – 8k reads



ceph days LONDON 2026

Performance – 8k reads



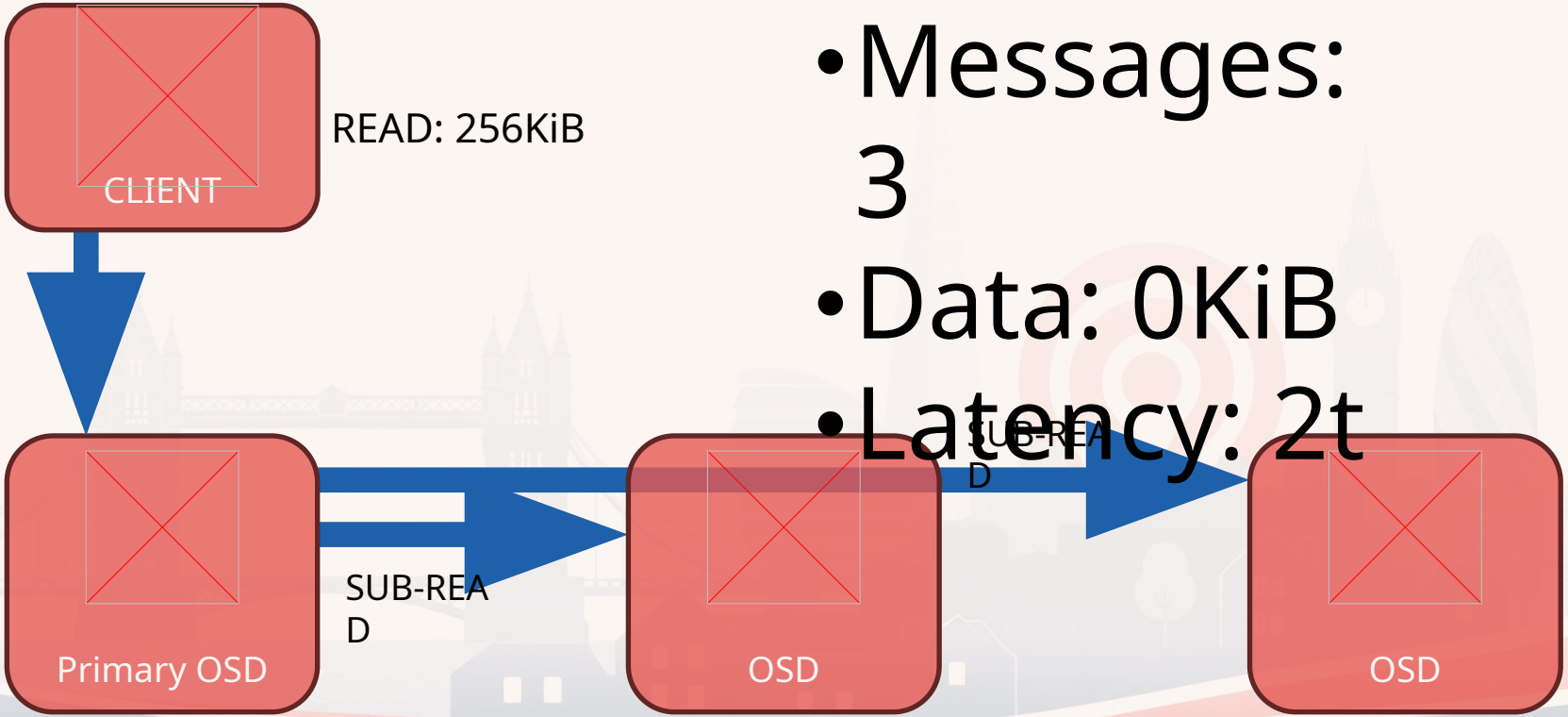
ceph days
LONDON 2026

Large Reads

As-is Read Path

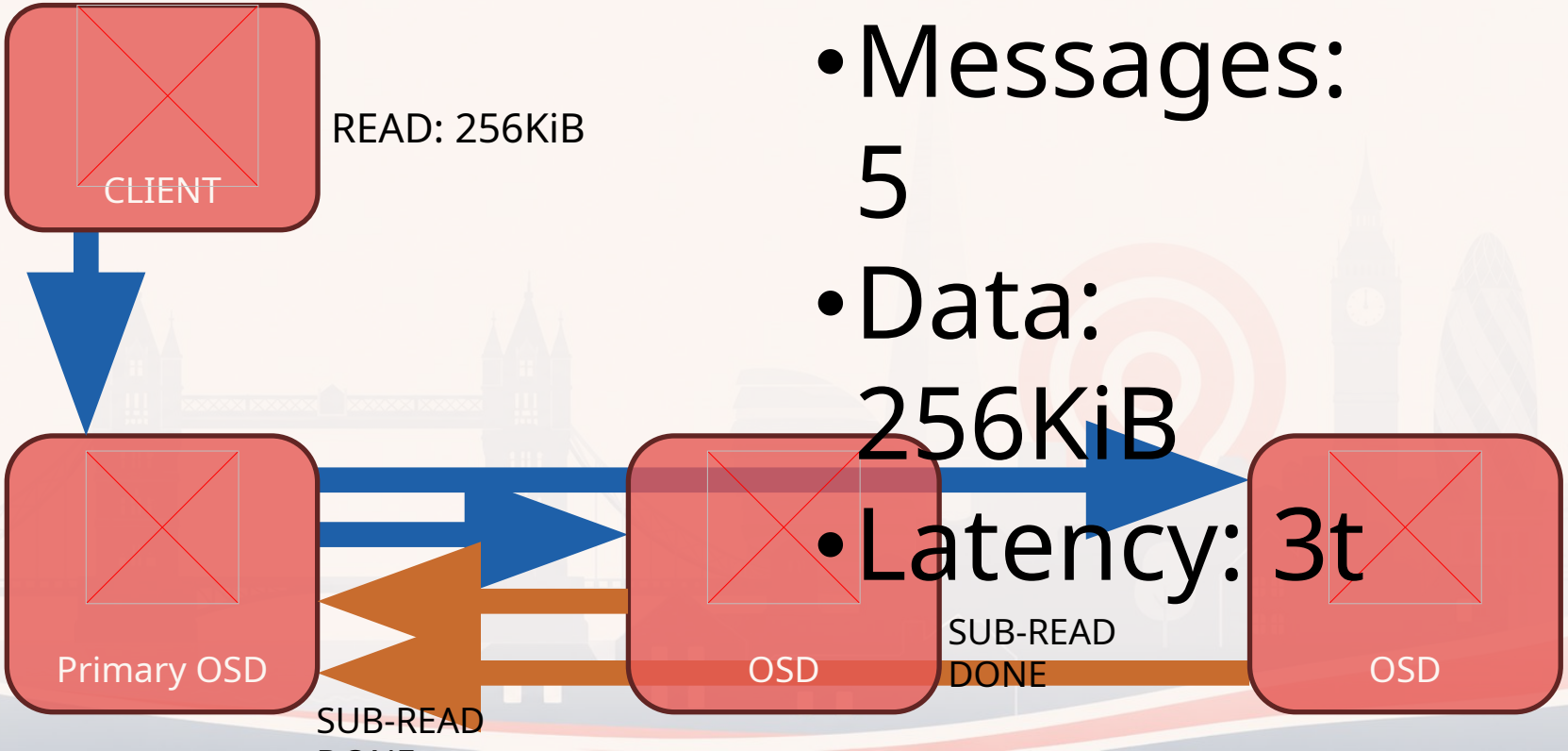
- Messages: 1
- Data: 0KiB
- Latency: 1t

As-is Read Path



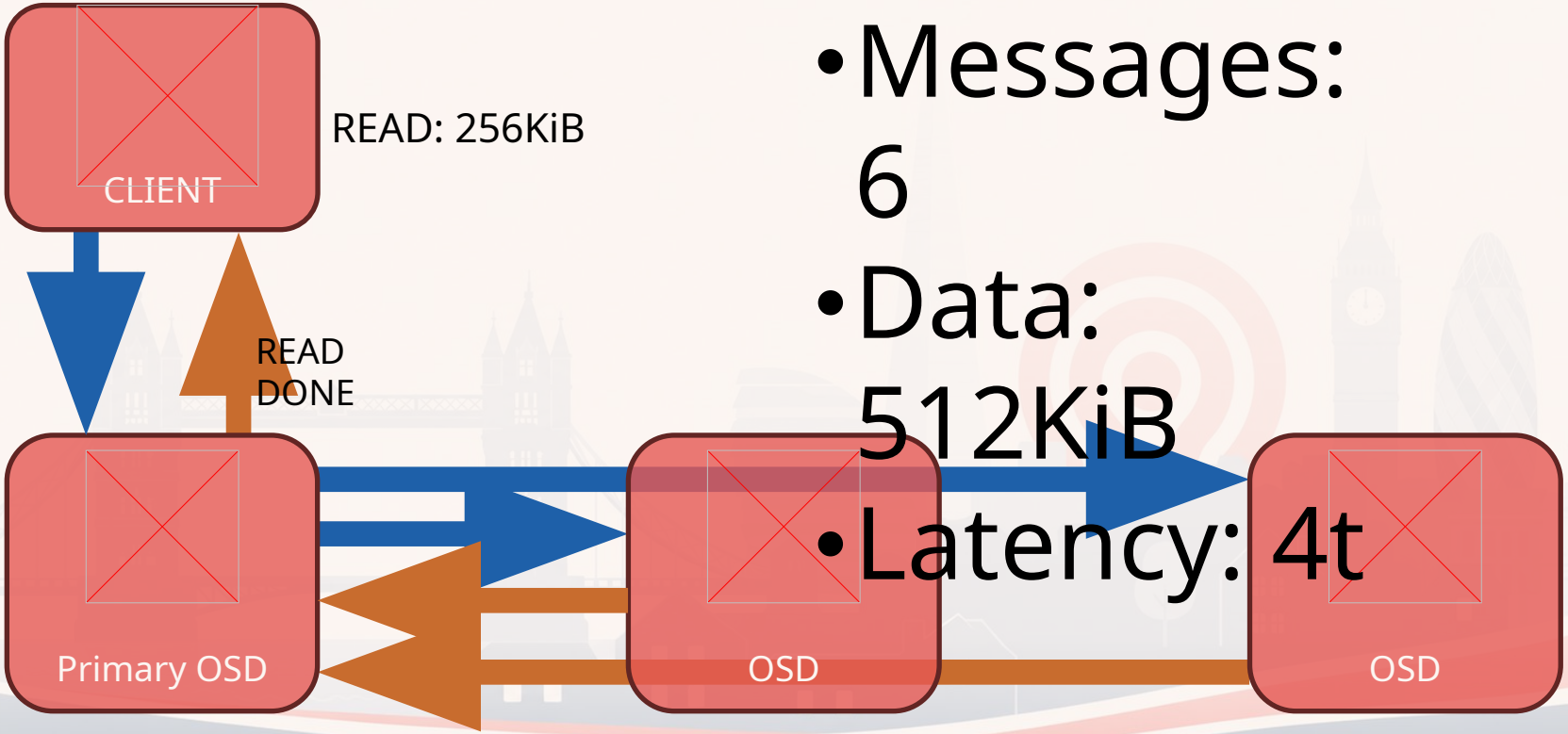
- Messages: 3
- Data: 0KiB
- Latency: 2t

As-is Read Path



- Messages: 5
- Data: 256KiB
- Latency: 3t

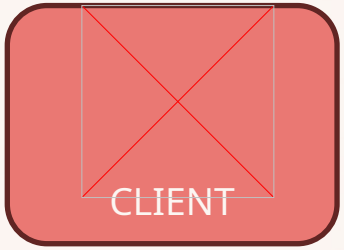
As-is Read Path

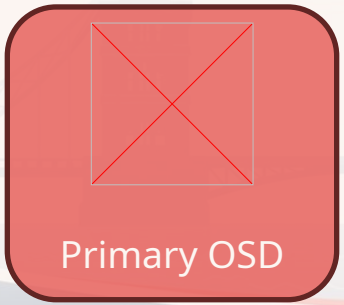


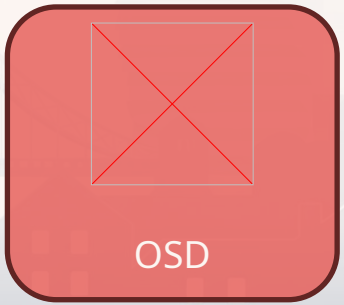
- Messages: 6
- Data: 512KiB
- Latency: 4t

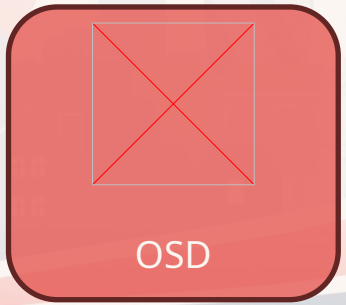
Direct Read Path

- Messages:
- Data:
- Latency:





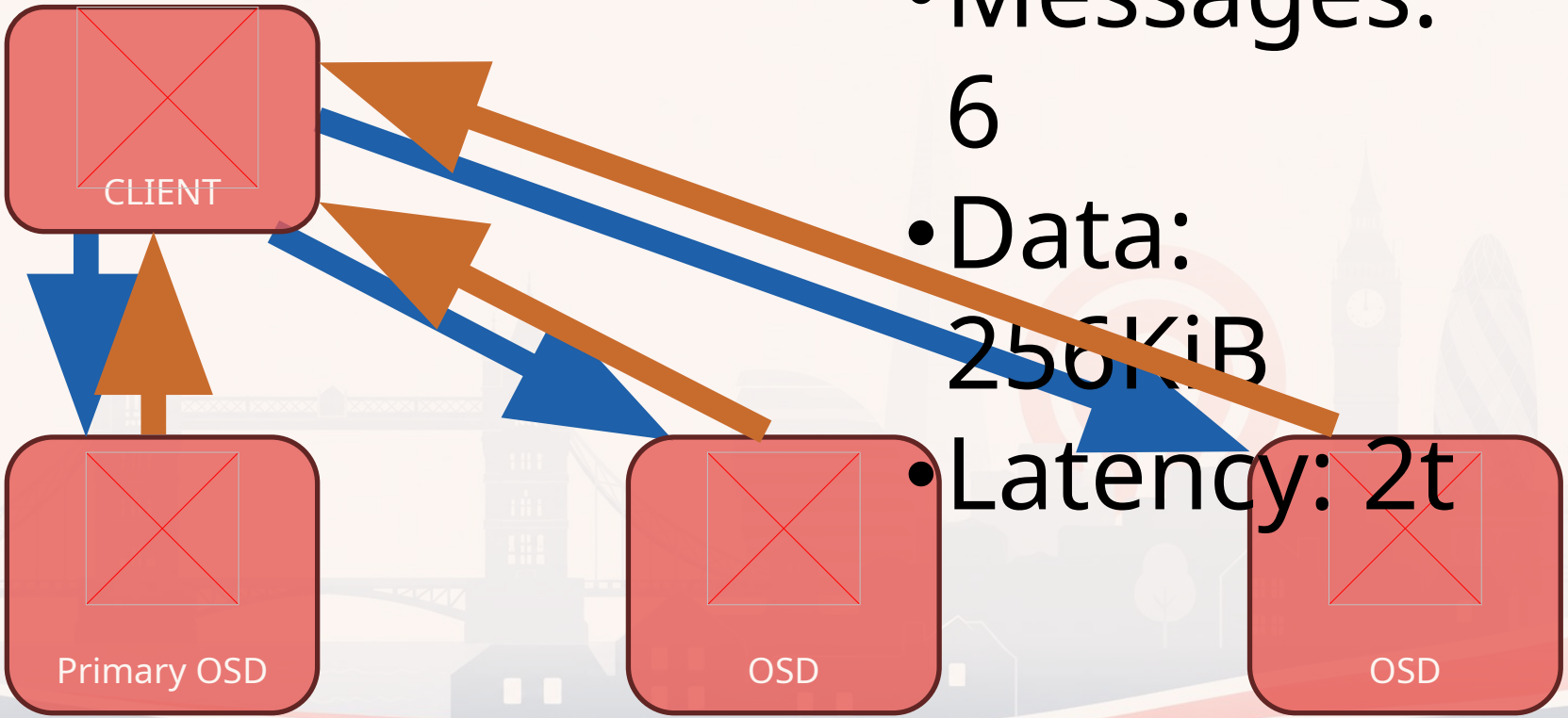




Direct Read Path

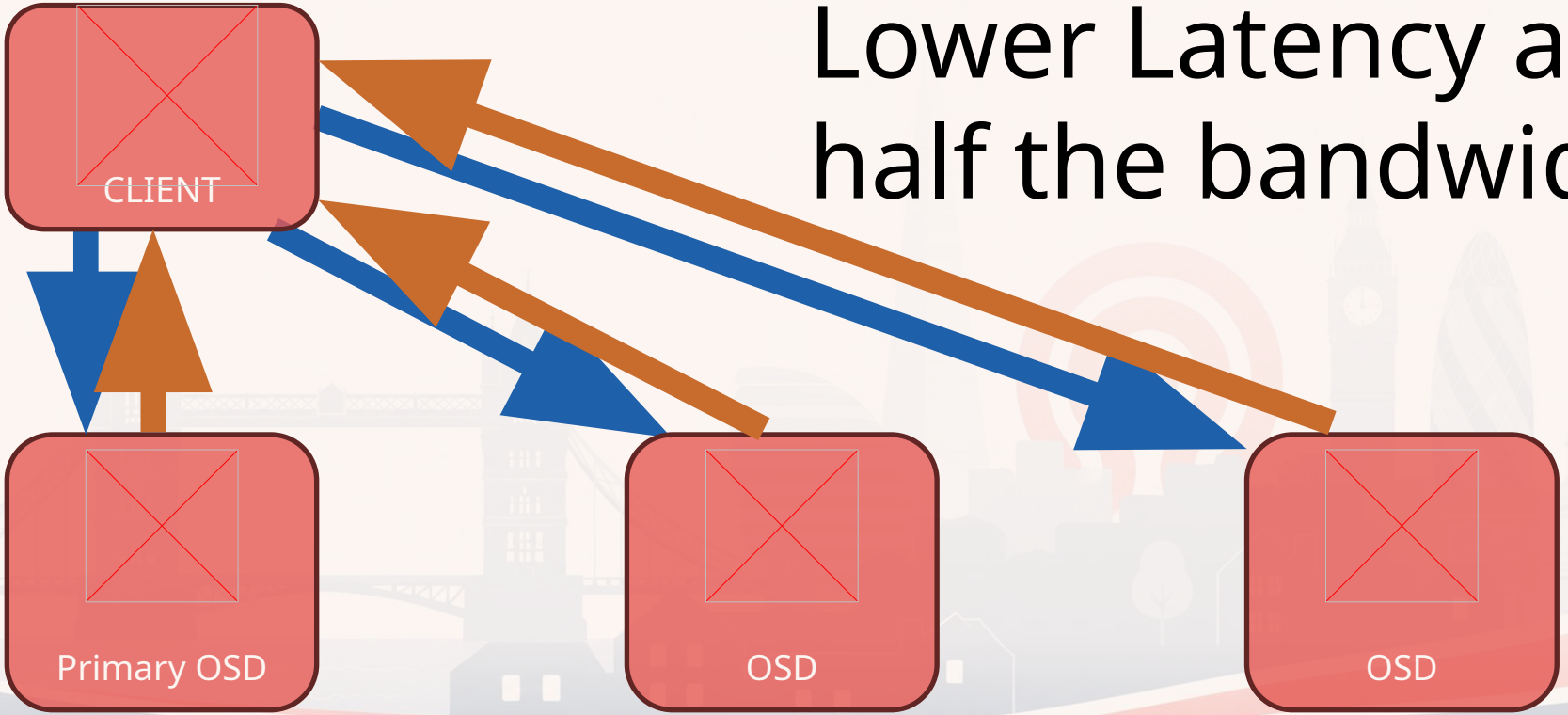
- Messages: 3
- Data: 0KiB
- Latency: 1t

Direct Read Path



- Messages: 6
- Data: 256KiB
- Latency: 2t

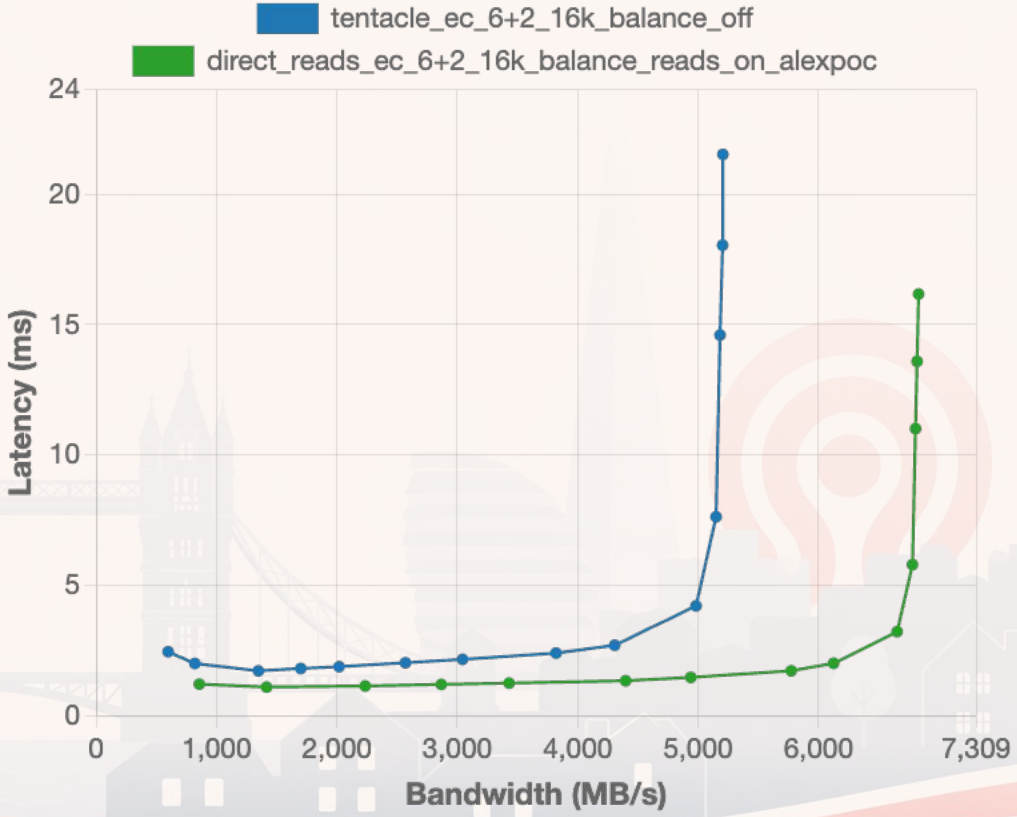
Direct Read Path



Lower Latency at half the bandwidth

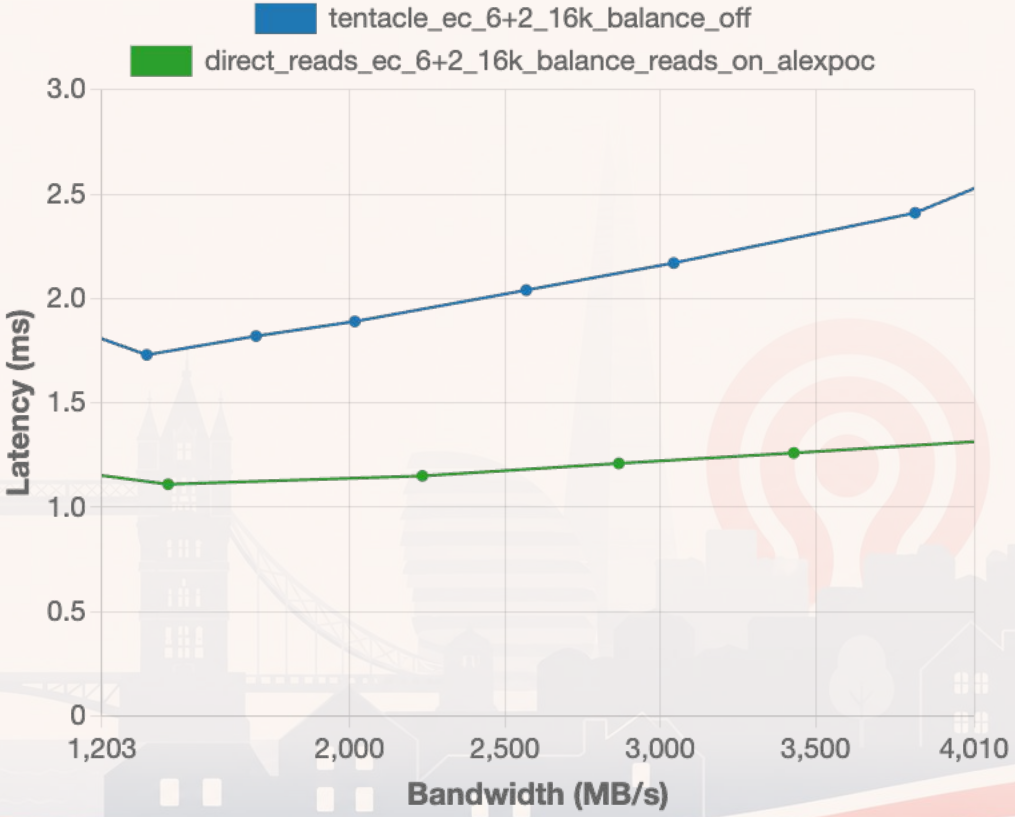
ceph days LONDON 2026

Performance: 256k



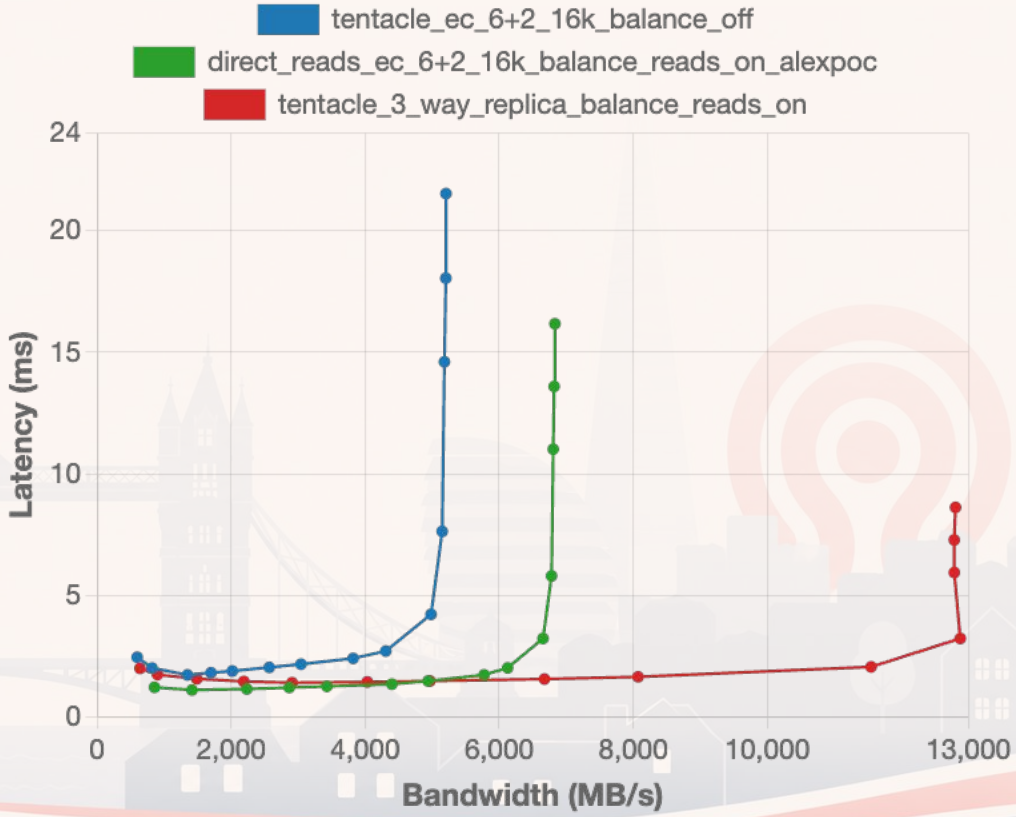
ceph days LONDON 2026

Performance: 256k



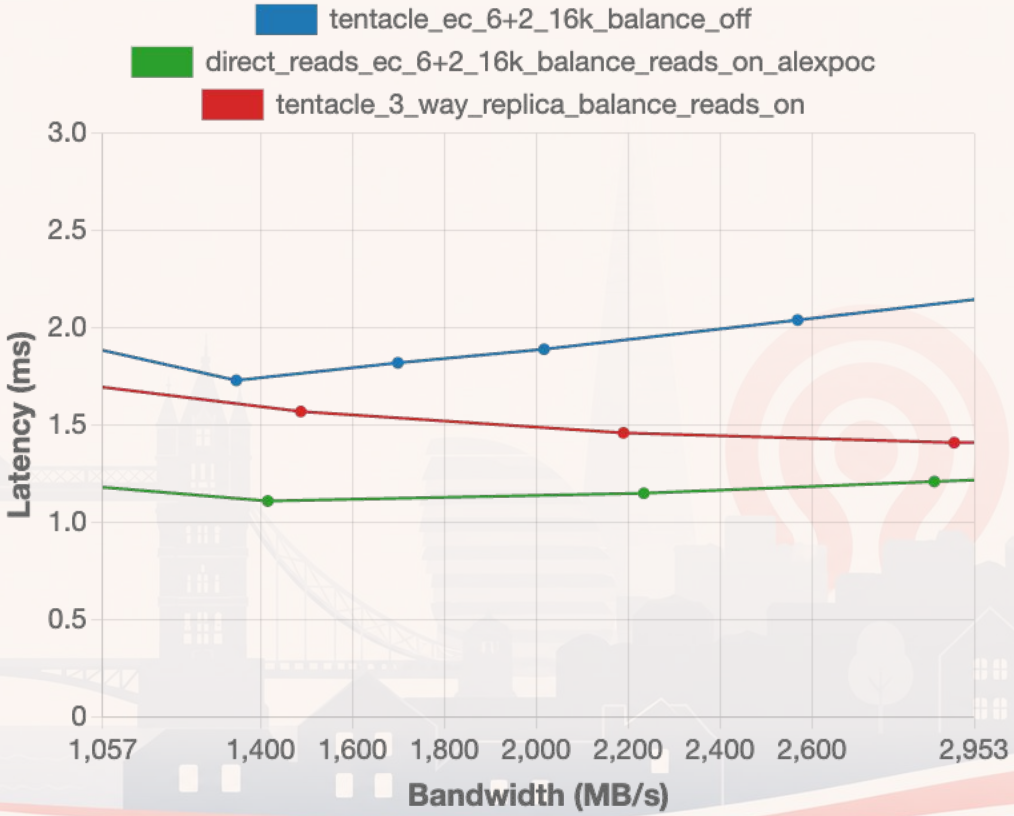
ceph days LONDON 2026

Performance vs Replica



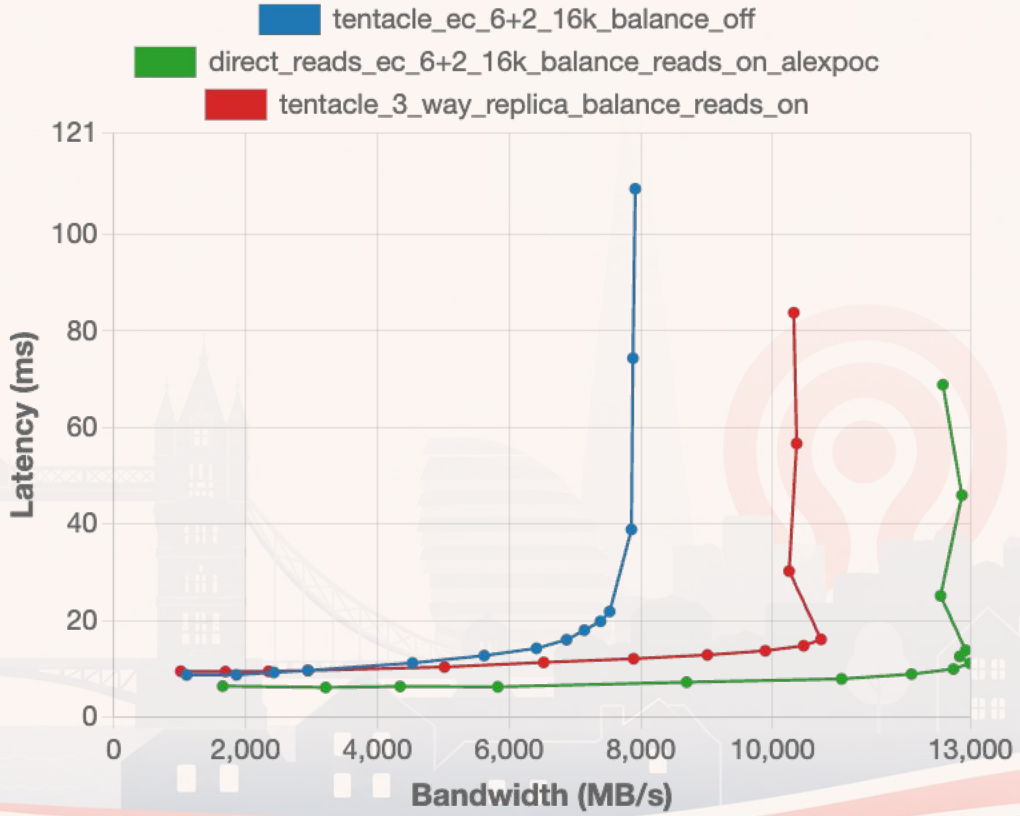
ceph days LONDON 2026

Performance vs Replica



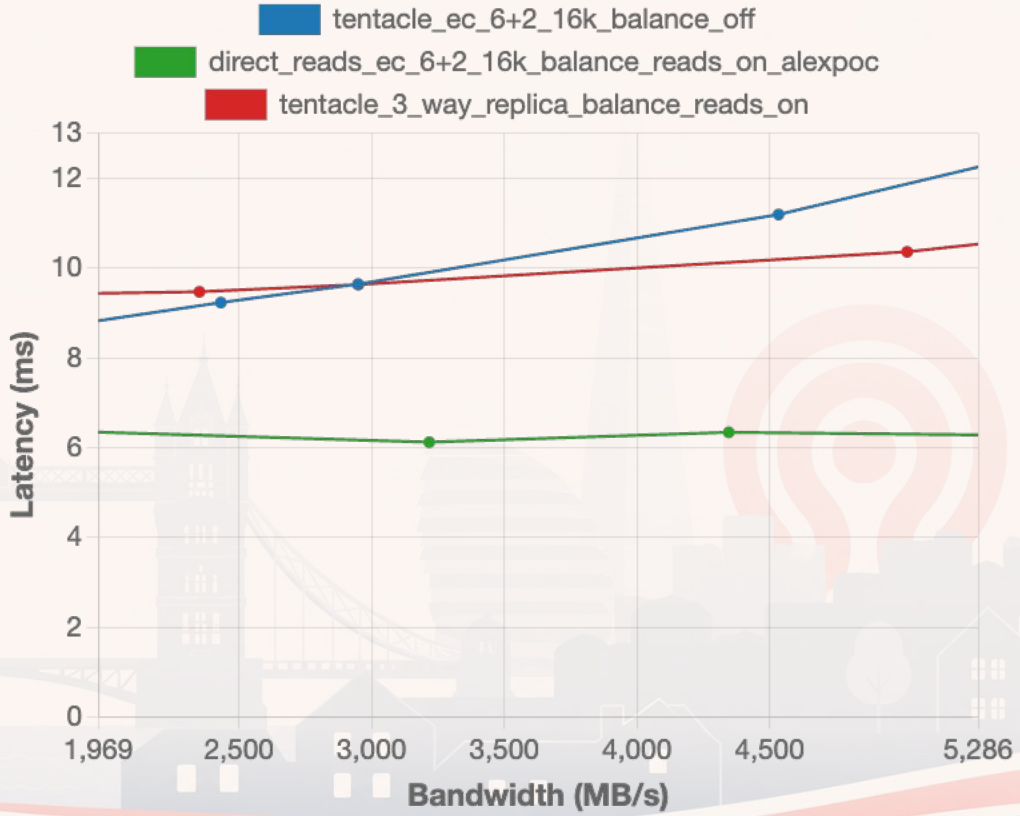
ceph days LONDON 2026

Performance with 4MiB reads!



ceph days LONDON 2026

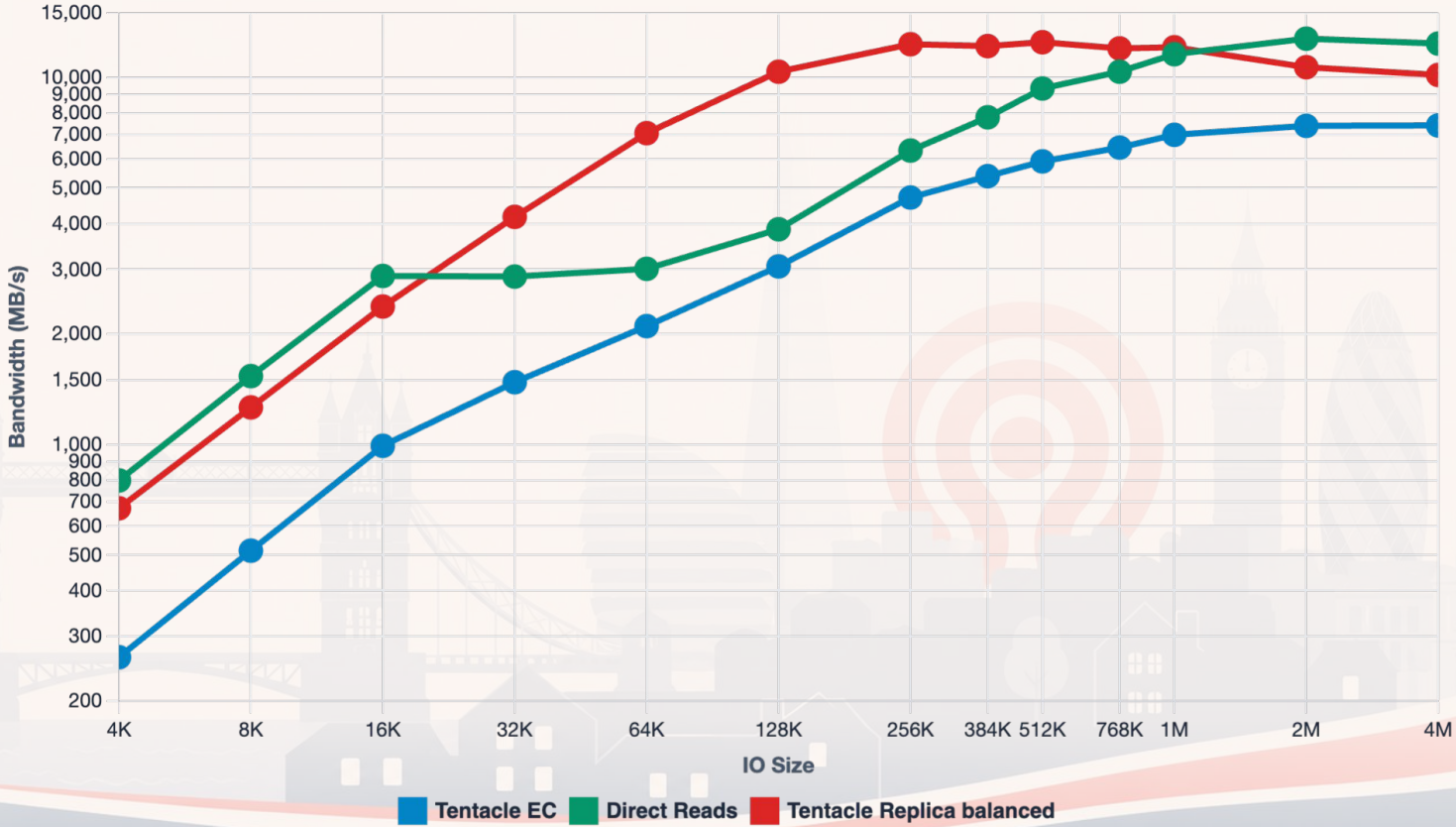
Performance with 4MiB reads!




ceph days
LONDON 2026
Performance Variation with IO Size
4KB Block (High IOPS)
128KB Block (Balanced)
1MB Block (Low IOPS)
Data Archive: Performance Studies
IO Size Manage-ment

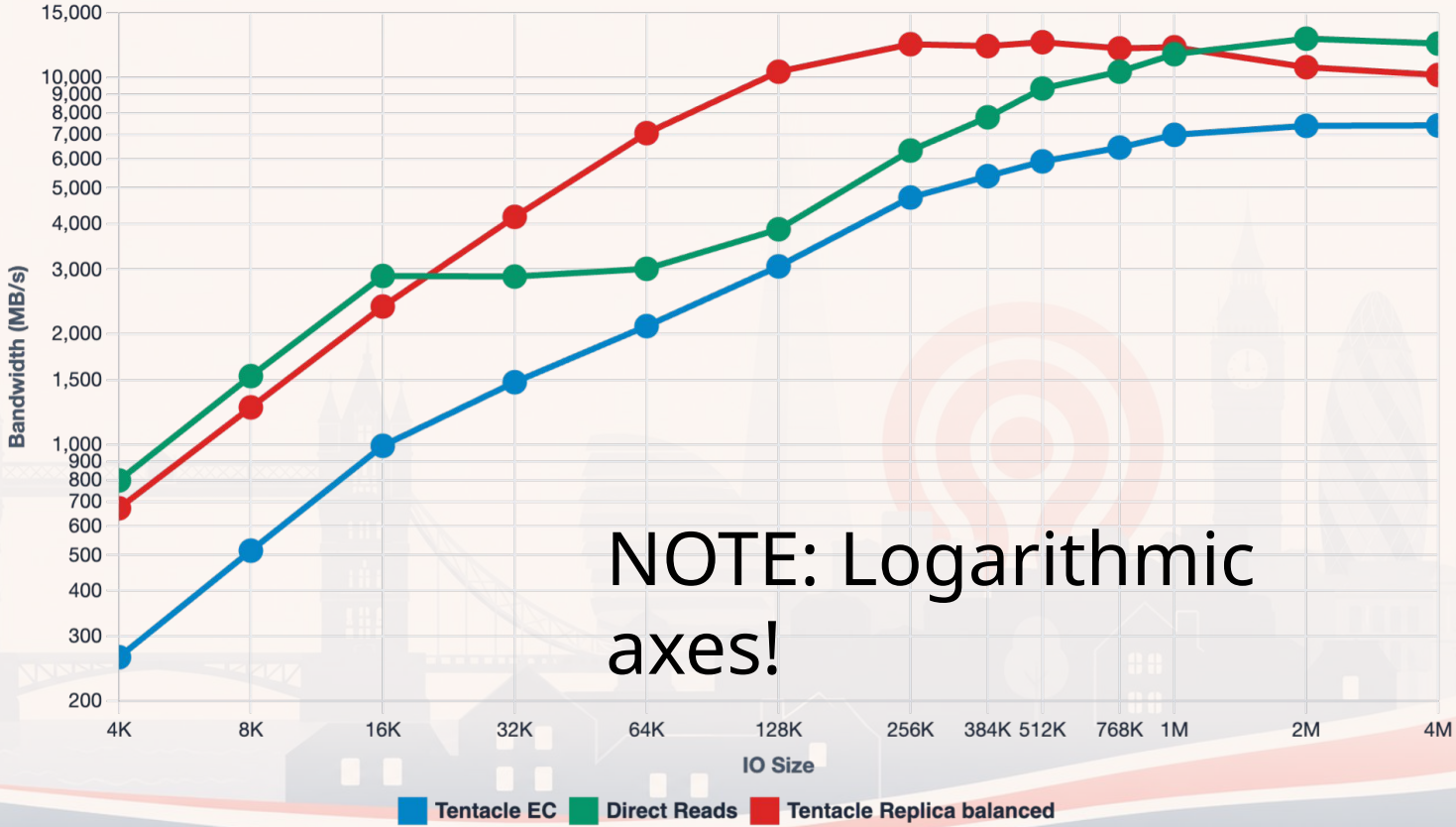
ceph days LONDON 2026

Bandwidth variations vs IO size



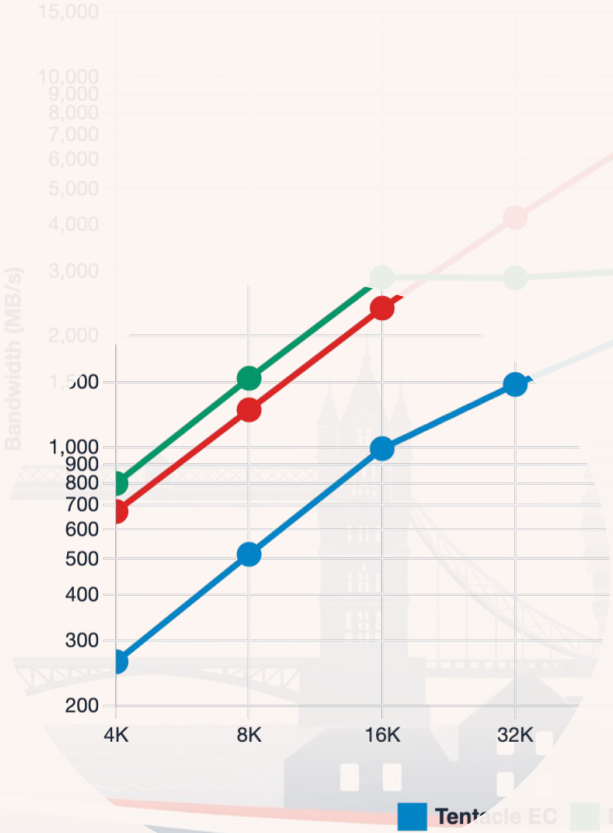
ceph days LONDON 2026

Bandwidth variations vs IO size



NOTE: Logarithmic axes!

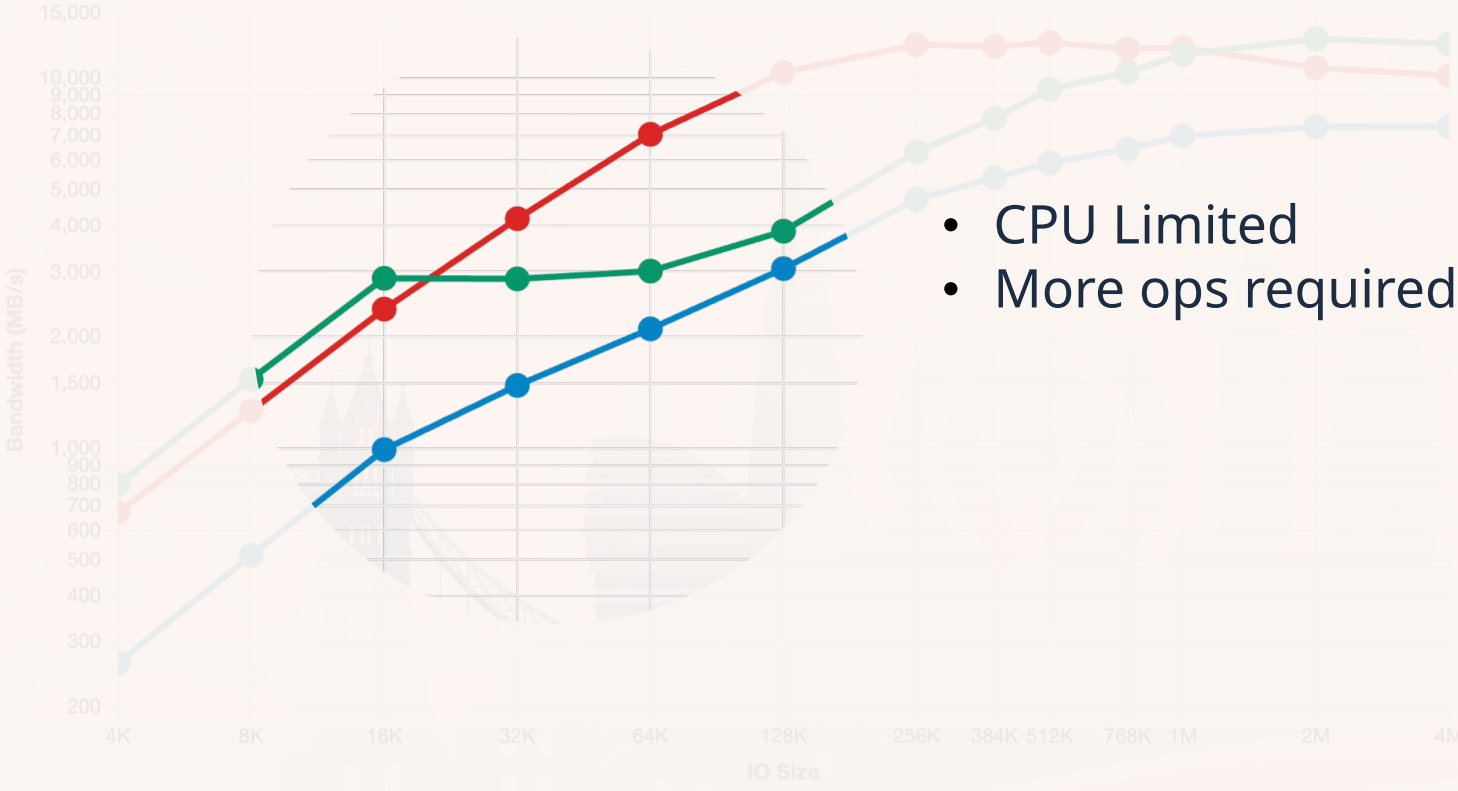
Bandwidth variations vs IO size



Small ops

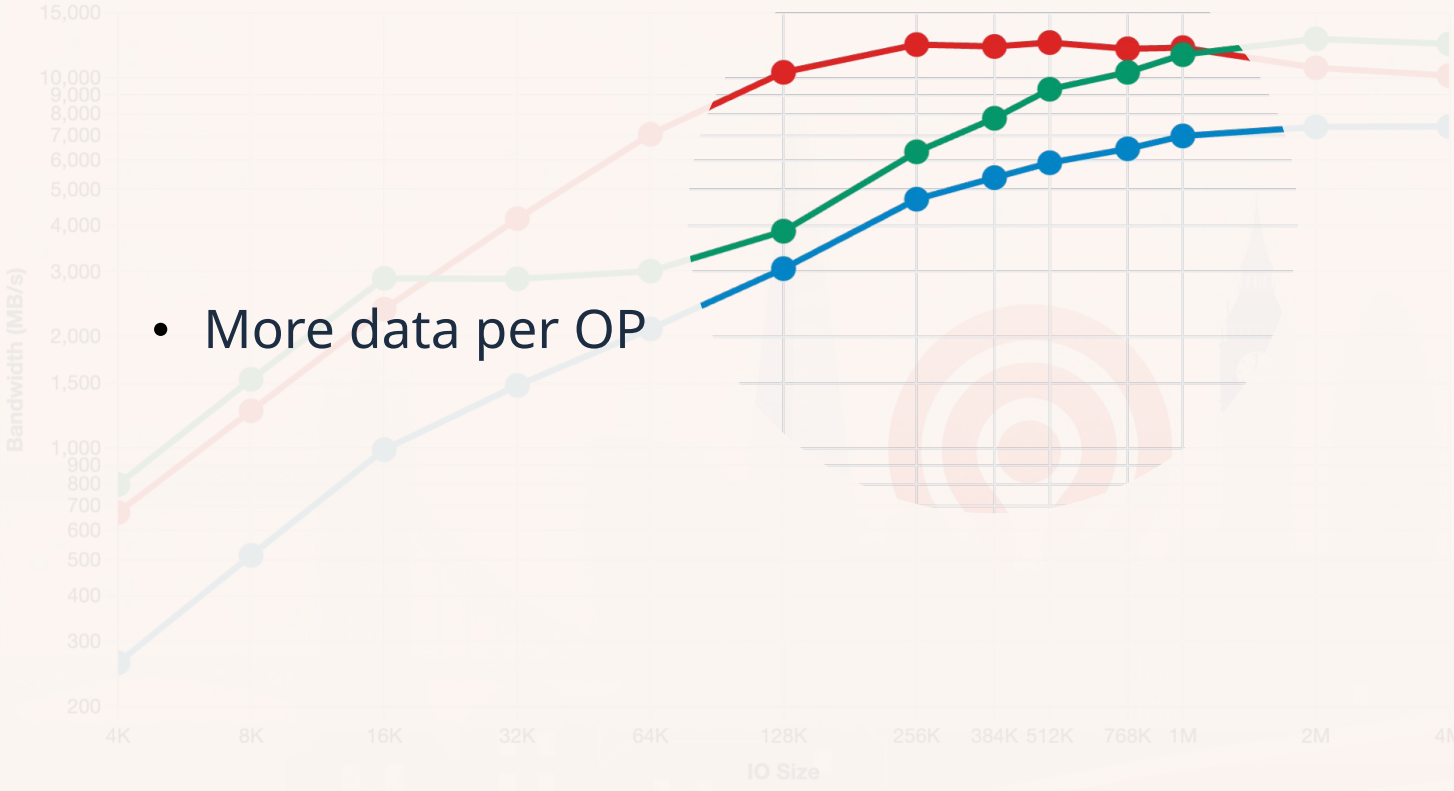
- IO Size < stripe_unit
- Tentacle -> Umbrella improved ?
- Replica IO Path == Tentacle code path.
- CPU Limited

Bandwidth variations vs IO size



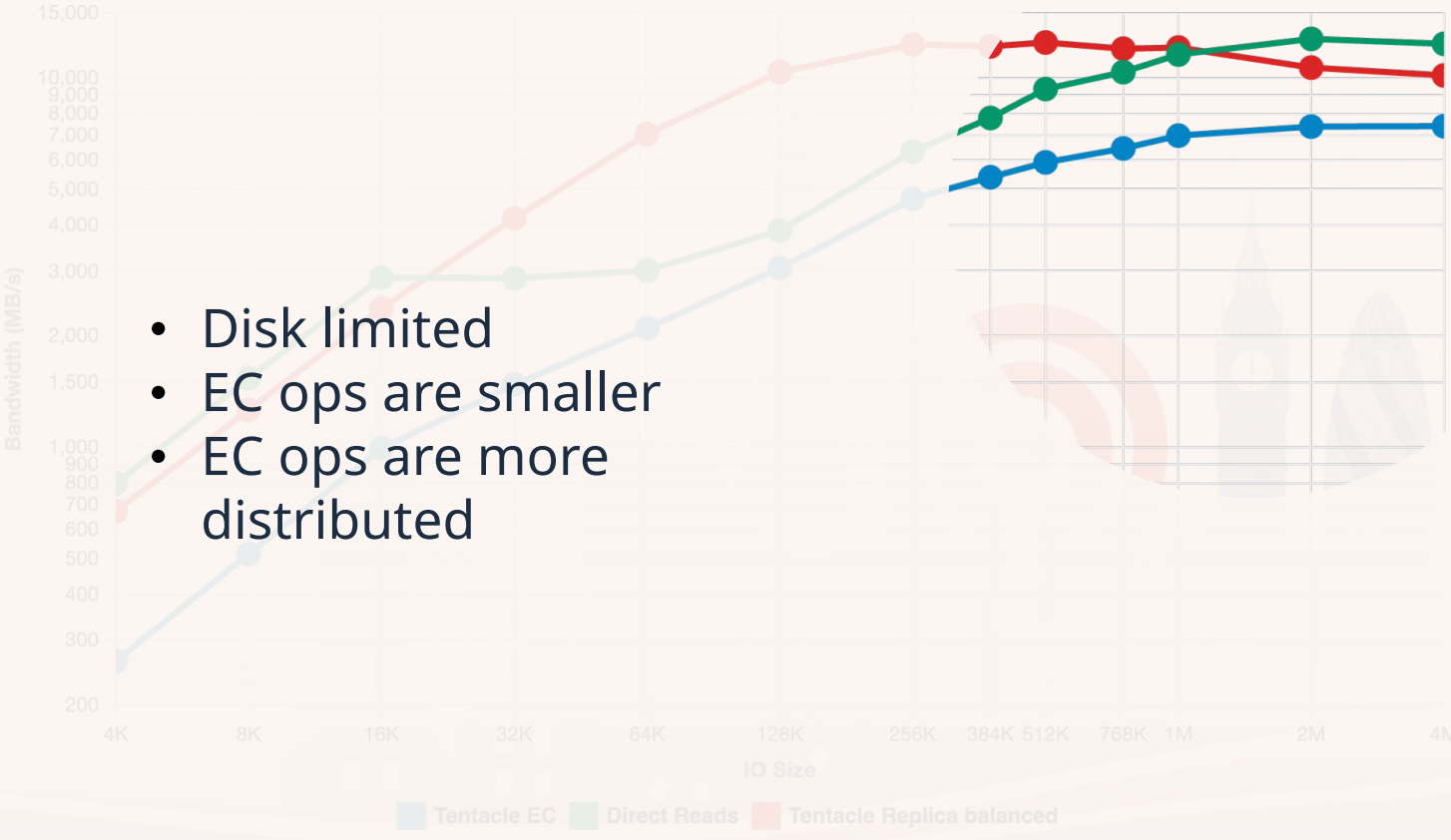
- CPU Limited
- More ops required

Bandwidth variations vs IO size



- More data per OP

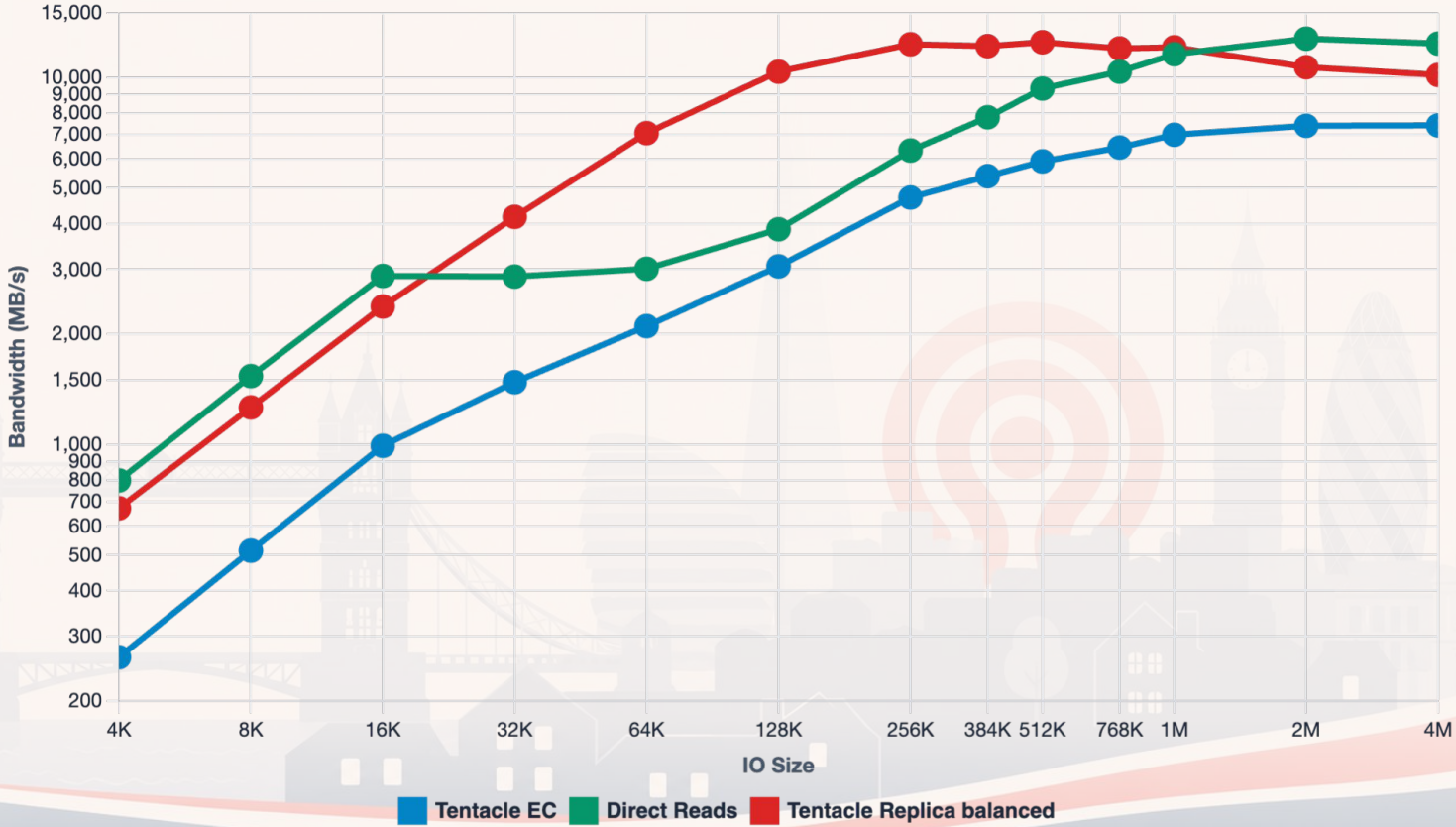
Bandwidth variations vs IO size



- Disk limited
- EC ops are smaller
- EC ops are more distributed

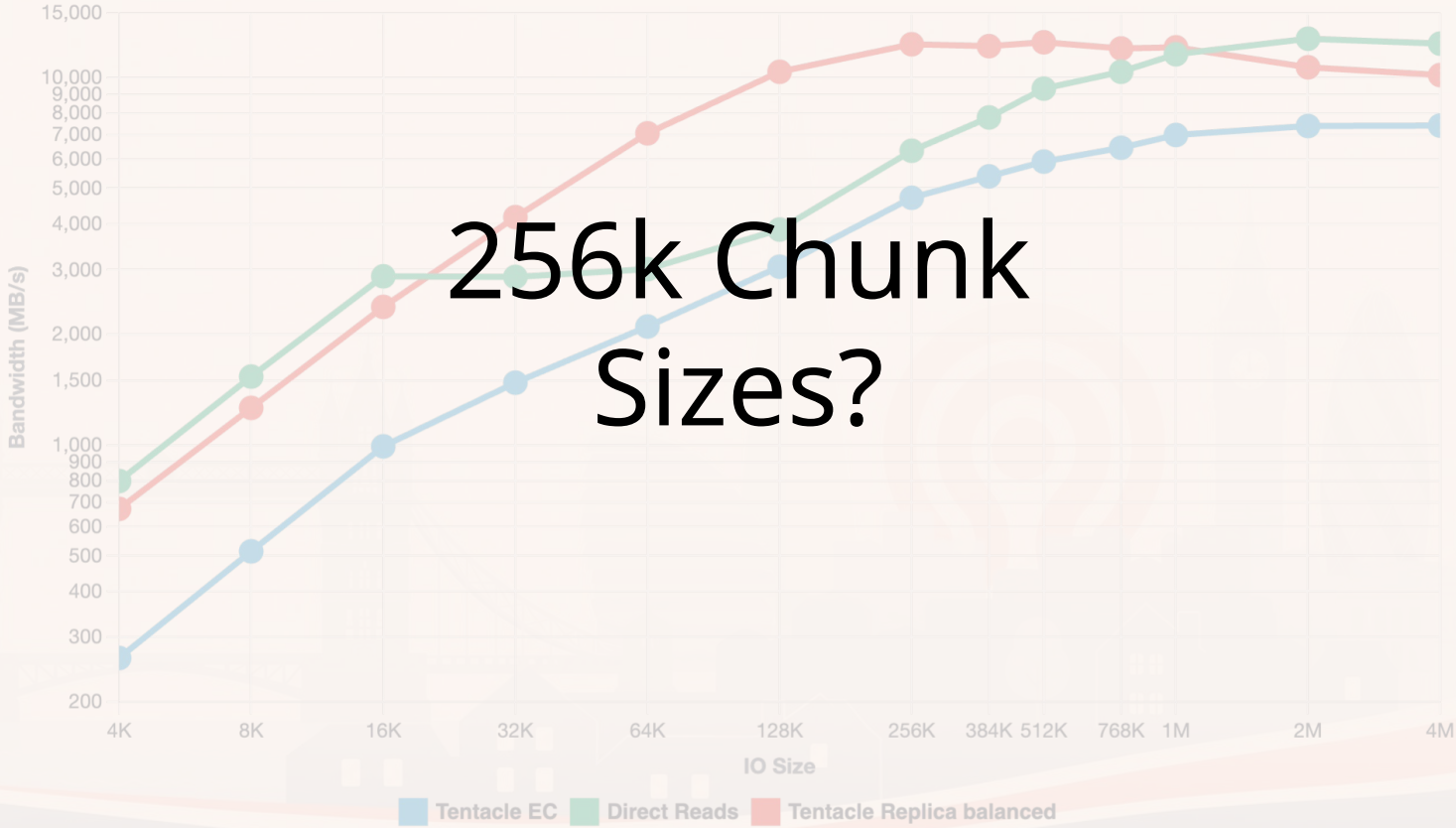
ceph days LONDON 2026

Bandwidth variations vs IO size



ceph days LONDON 2026

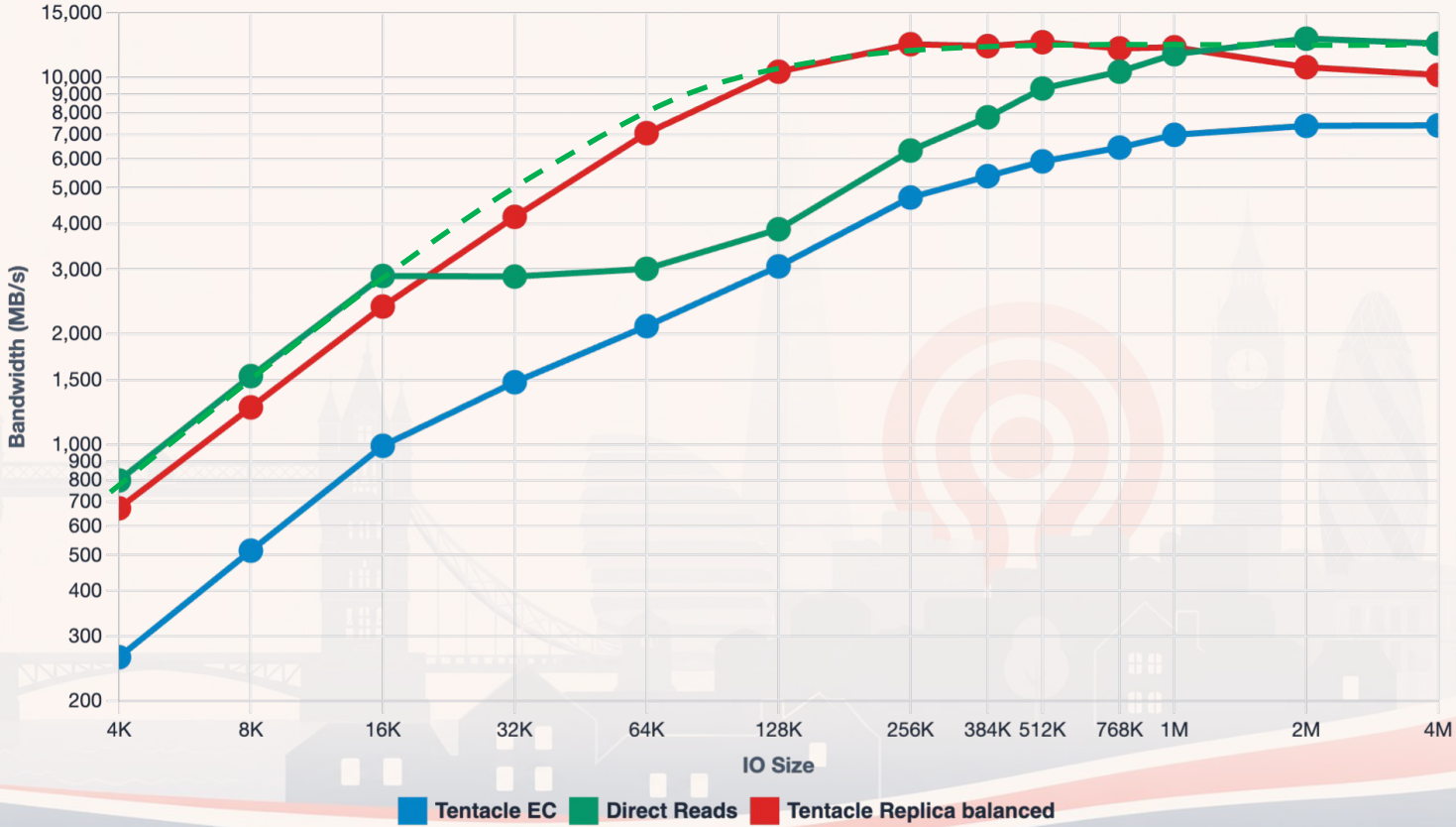
Bandwidth variations vs IO size



256k Chunk Sizes?

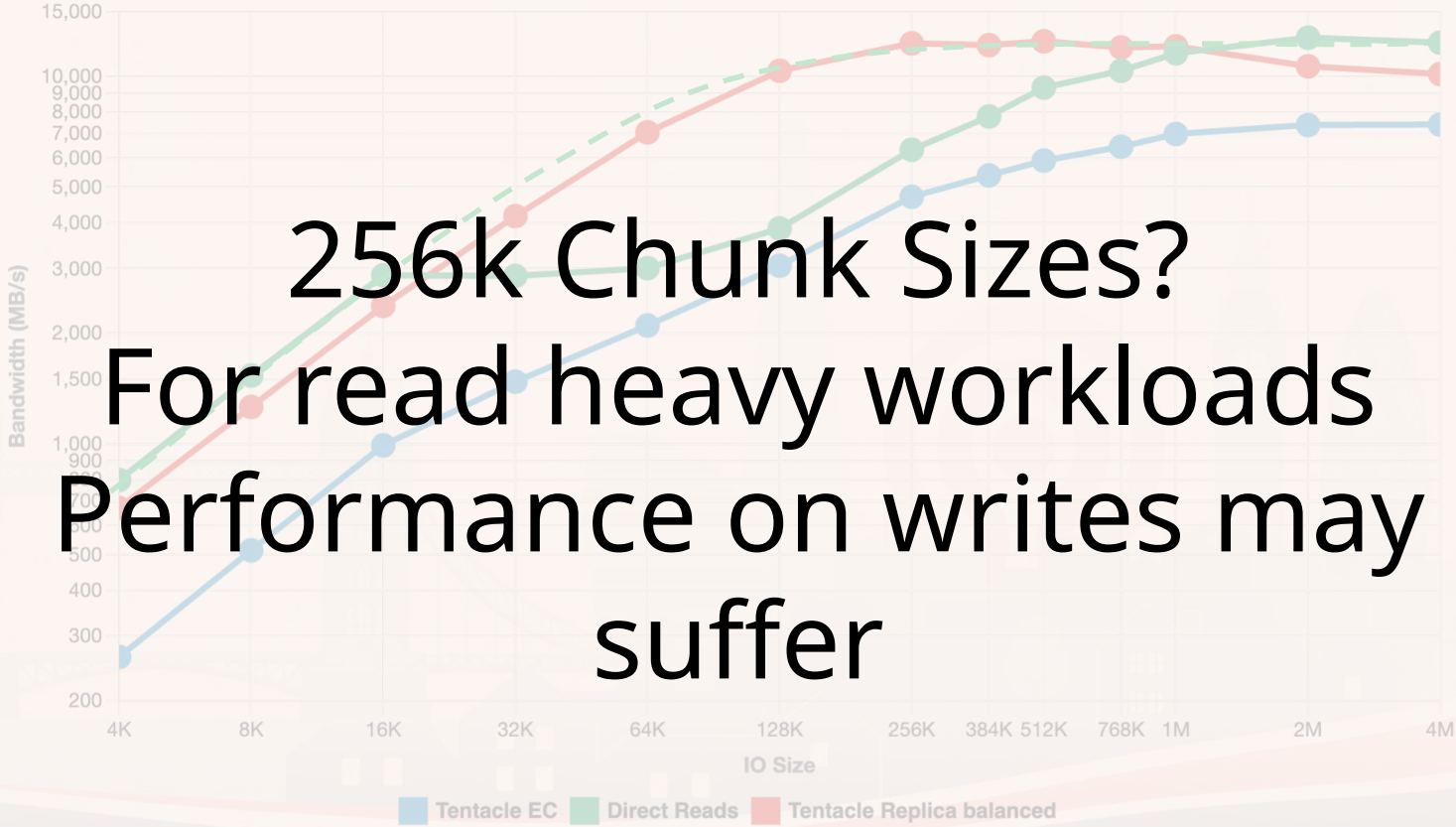
ceph days LONDON 2026

Bandwidth variations vs IO size



ceph days LONDON 2026

Bandwidth variations vs IO size



256k Chunk Sizes?
For read heavy workloads
Performance on writes may suffer


ceph days
LONDON 2026
FAILED
INFRASTRUCTURE
WARNING:
Road Failure ahead -
FAILED OSD SECTOR
OSD
REPAIR
STOP
ceph

Failed OSDs

Failed OSDs

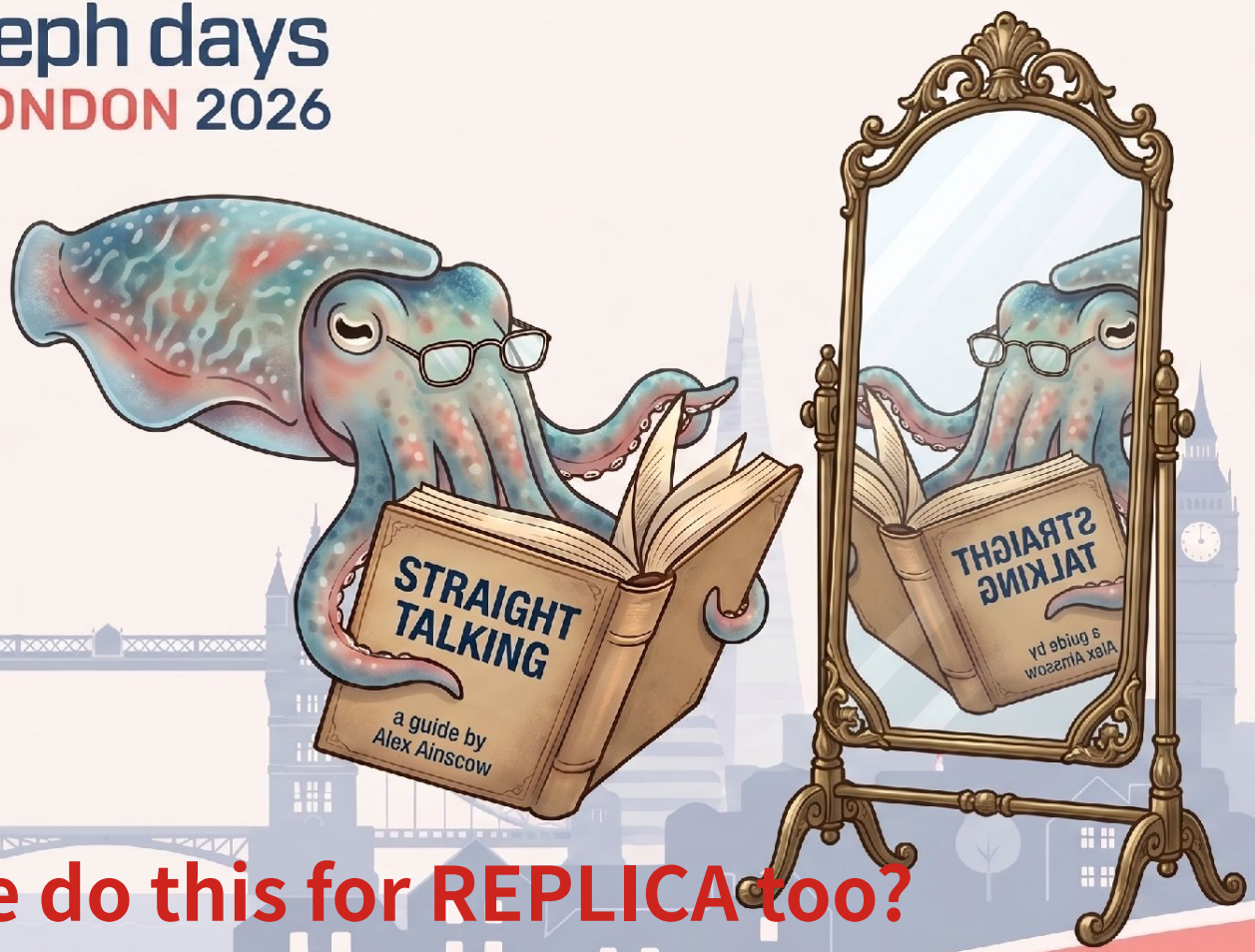
- Client Detects and re-routes ops to primary
- Redrives to primary may occur during transition.

ceph days LONDON 2026

RW Ordering

- Uncommitted writes can cause reads to the same object to be re-driven
- Reads can over-take outstanding writes: All Ceph clients expect and handle this cleanly
- Reads will never be torn (i.e. contain part of a write)

ceph days
LONDON 2026



Can we do this for REPLICA too?

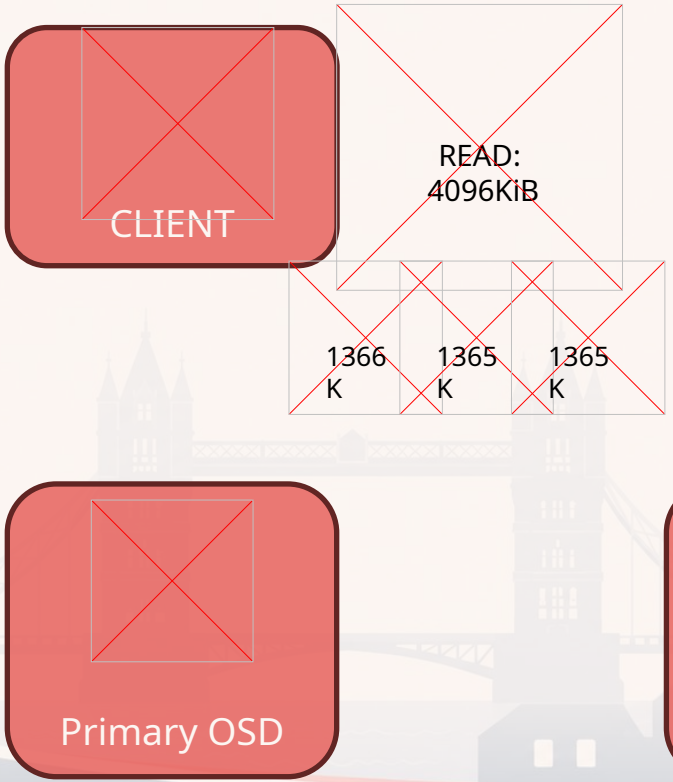
(for large reads)

ceph days
LONDON 2026

Yes!

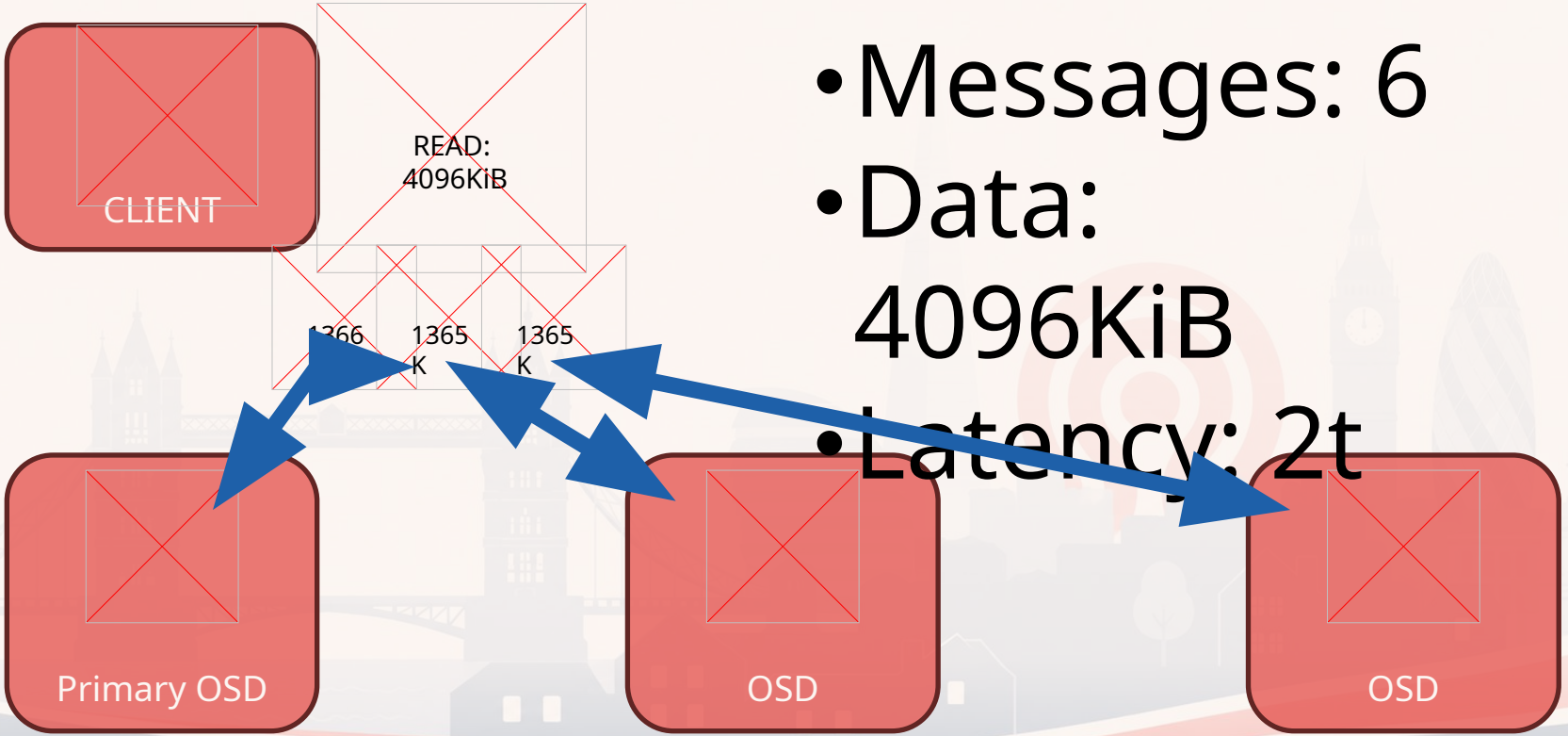
(for large reads)

As-is Read Path



- Messages: 0
- Data: 0KiB
- Latency: 0t

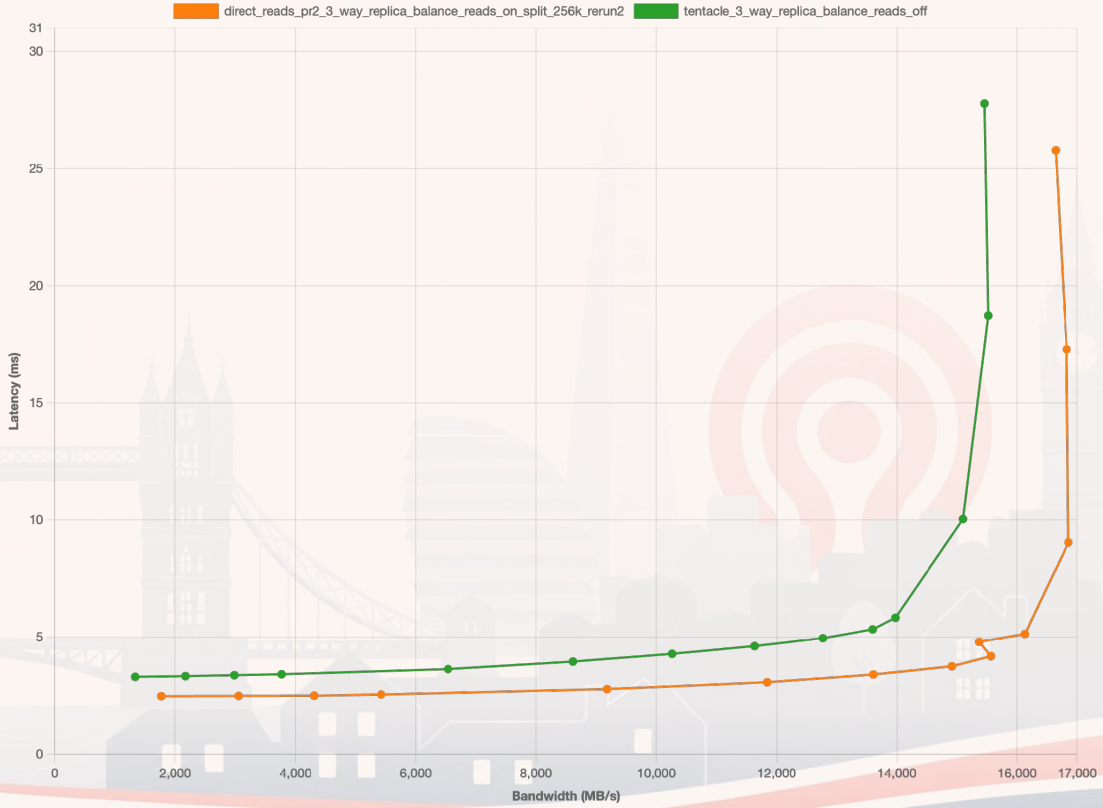
As-is Read Path



- Messages: 6
- Data: 4096KiB
- Latency: 2t

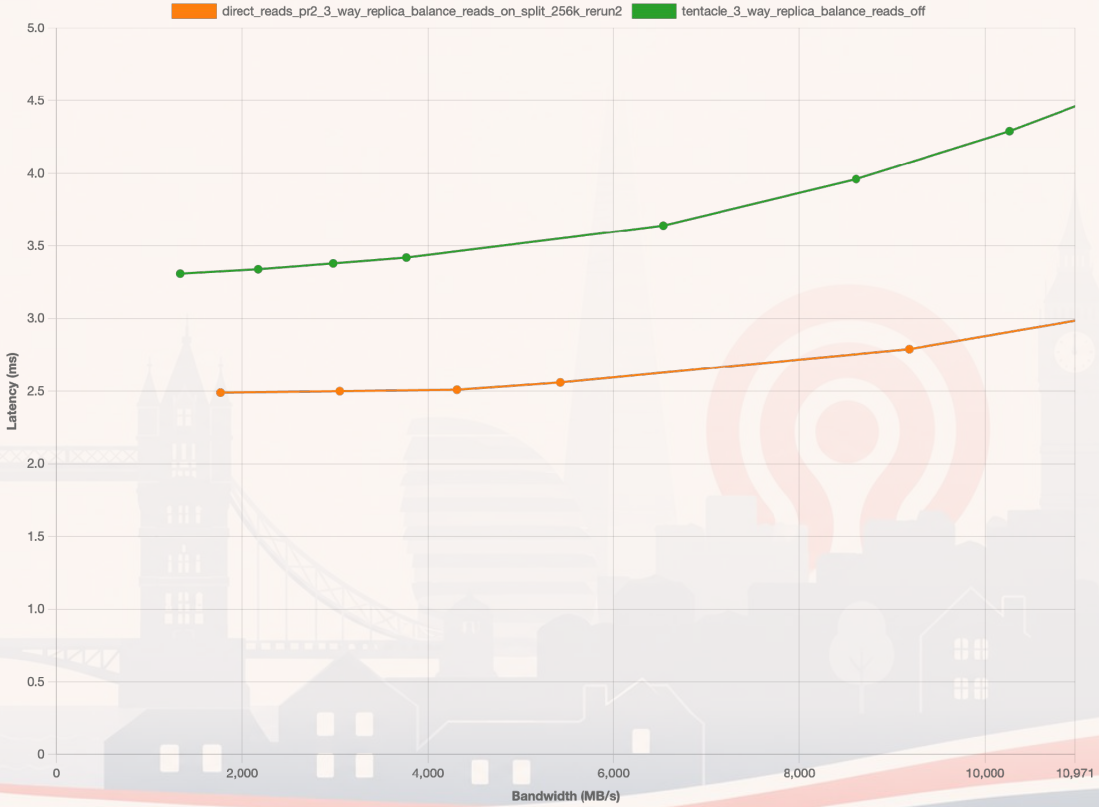
ceph days LONDON 2026

Split Replica Performance (2MiB)



ceph days LONDON 2026

Split Replica Performance (2MiB)



ceph days
LONDON 2026

Turning Direct Reads ON!

ceph days LONDON 2026

Enable in RBD

This is a config-option:

```
[client]
rbd_read_from_replica_policy = balance
```

https://docs.ceph.com/en/latest/rbd/rbd-config-ref

ceph days LONDON 2026

Enable in CephFS

This is a mount option:

```
read_from_replica=balance
```

https://docs.ceph.com/en/latest/man/8/mount.ceph/

Globally (feat RGW)

Balanced reads can be turned on globally using the config option:

```
rados_replica_read_policy = balance
```

NOTE: May not work well with unusual/unsupported features.

ceph days LONDON 2026

Configuring Replica

Enable balanced reads

AND

```
[global]
osd_min_split_replica_read_size = 262144
```

ceph days
LONDON 2026

Any Questions?