

Multi-Zone vs Single-Zone ODF Performance

Impact of cross-AZ replica placement on ODF storage I/O

MZ: perf-20260228-164717 (ocp-virt-mz-cluster, us-south, 3 AZs, failureDomain=zone) — SZ: perf-20260227-203655 (ocp-virt-420-cluster, eu-de-1, failureDomain=rack)

Key finding: Cross-AZ replication costs **10–19% write IOPS** and **+40–82% write p99 latency**, while **reads are unaffected** thanks to read affinity serving from local-zone OSDs. The dominant factor is inter-AZ RTT (~0.5–2ms) on every synchronous replica acknowledgment.

Cluster Conditions [-]

MZ Multi-Zone (ocp-virt-mz-cluster)

Region us-south (3 AZs)
Workers 3x bx2d.metal.96x384 (1 per AZ)
OSDs 24 (8 per node, NVMe)
Raw capacity ~70 TiB
Failure domain zone (cross-AZ replicas)
ODF / Ceph 4.19.10 / Squid 19.2.1
Encryption Full (data + network + KMS)
Read affinity Enabled

SZ Single-Zone (ocp-virt-420-cluster)

Region eu-de (1 AZ)
Workers 3x bx2d.metal.96x384 (same zone)
OSDs 24 (8 per node, NVMe)
Raw capacity ~70 TiB
Failure domain rack (intra-zone replicas)
ODF / Ceph 4.19.10 / Squid 19.2.1
Encryption Full (data + network + KMS)
Read affinity Enabled

Only architectural difference: failureDomain: zone (MZ) vs failureDomain: rack (SZ). Hardware, software, encryption, and ODF configuration are identical.

Test Matrix [+]

Storage Pool Overview [-]

Pool	Type	Description	MZ Domain	SZ Domain
rep3	RBD Replicated	3-way replicated RBD (OOB)	zone	rack
rep3-virt	RBD Replicated	3-way RBD + rxbounce (OOB)	zone	rack
rep3-enc	RBD Replicated	3-way RBD + SC encryption (OOB)	zone	rack
rep2	RBD Replicated	2-way replicated RBD (custom)	zone	host
cephfs-rep3	CephFS	3-way CephFS (OOB)	zone	rack
cephfs-rep2	CephFS	2-way CephFS (custom)	zone	host
ec-2-1	RBD Erasure Coded	Erasure coded 2+1 (custom)	zone	host

Performance Comparison

MZ winsSZ wins

Random 4k Total IOPS (IOPS)

Pool	MZ	SZ	Delta	Winner
rep3	63,356	64,504	-1.8%	~Tie
rep3-enc	61,062	62,702	-2.6%	~Tie
rep2	68,976	71,966	-4.2%	~Tie
cephfs-rep3	55,474	45,557	+21.8%	MZ
cephfs-rep2	58,326	59,956	-2.7%	~Tie
ec-2-1	45,987	49,058	-6.3%	SZ

Sequential 1M Total BW (MiB/s)

Pool	MZ	SZ	Delta	Winner
rep3	7,967	7,616	+4.6%	~Tie
rep3-virt	7,619	7,599	+0.3%	~Tie
rep3-enc	6,992	6,673	+4.8%	~Tie
rep2	7,900	8,306	-4.9%	~Tie
cephfs-rep3	5,367	4,809	+11.6%	MZ
cephfs-rep2	5,265	5,374	-2.0%	~Tie
ec-2-1	4,214	5,814	-27.5%	SZ

Mixed 70/30 4k Total IOPS (IOPS)

Pool	MZ	SZ	Delta	Winner
rep3	47,052	49,203	-4.4%	~Tie
rep3-virt	47,138	48,641	-3.1%	~Tie
rep3-enc	52,206	56,283	-7.2%	SZ
rep2	50,216	53,574	-6.3%	SZ
cephfs-rep3	36,711	37,762	-2.8%	~Tie
cephfs-rep2	44,733	45,126	-0.9%	~Tie
ec-2-1	25,996	31,405	-17.2%	SZ

Avg p99 Latency (random) (ms) — lower is better

Pool	MZ	SZ	Delta	Winner
rep3	192.0	106.3	+80.7%	SZ
rep3-enc	196.2	108.4	+81.1%	SZ
rep2	156.4	98.7	+58.5%	SZ
cephfs-rep3	753.8	446.0	+69.0%	SZ
cephfs-rep2	611.2	437.1	+39.8%	SZ
ec-2-1	281.7	179.8	+56.7%	SZ

Read vs Write Impact

Cross-AZ replica placement primarily affects writes (must wait for remote AZ acknowledgment). Reads benefit from read affinity (served by local-zone OSD).

Random 4k IOPS — Read vs Write

Pool	MZ Read	SZ Read	Read Δ	MZ Write	SZ Write	Write Δ
rep3	55,841	54,222	+3.0%	7,515	10,282	-26.9%
rep3-enc	53,443	52,285	+2.2%	7,619	10,417	-26.9%
rep2	60,021	61,034	-1.7%	8,955	10,932	-18.1%
cephfs-rep3	52,439	41,471	+26.4%	3,035	4,086	-25.7%
cephfs-rep2	54,434	55,357	-1.7%	3,892	4,599	-15.4%
ec-2-1	40,607	42,822	-5.2%	5,380	6,236	-13.7%

Sequential 1M BW (MiB/s) — Read vs Write

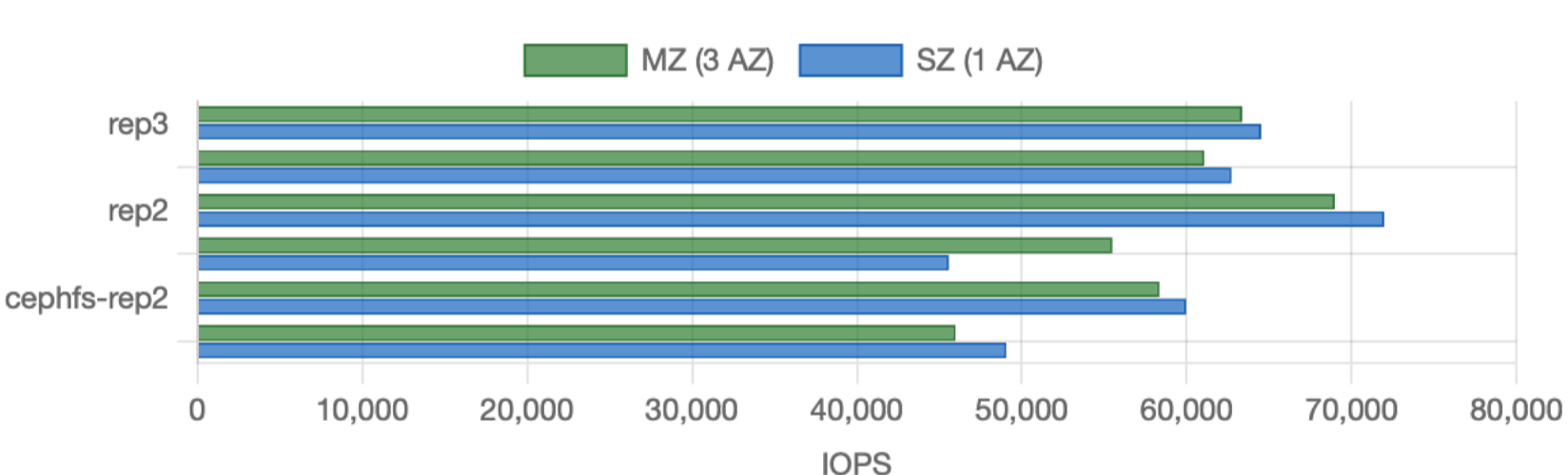
Pool	MZ Read	SZ Read	Read Δ	MZ Write	SZ Write	Write Δ
rep3	5,298	5,250	+0.9%	2,668	2,367	+12.7%
rep3-virt	5,188	5,326	-2.6%	2,431	2,273	+7.0%
rep3-enc	4,530	4,250	+6.6%	2,462	2,423	+1.6%
rep2	5,762	5,996	-3.9%	2,139	2,310	-7.4%
cephfs-rep3	4,343	3,879	+12.0%	1,024	929.7	+10.2%
cephfs-rep2	4,334	4,505	-3.8%	930.2	868.6	+7.1%
ec-2-1	2,790	3,445	-19.0%	1,424	2,369	-39.9%

Random 4k p99 Latency (ms) — Read vs Write (lower is better)

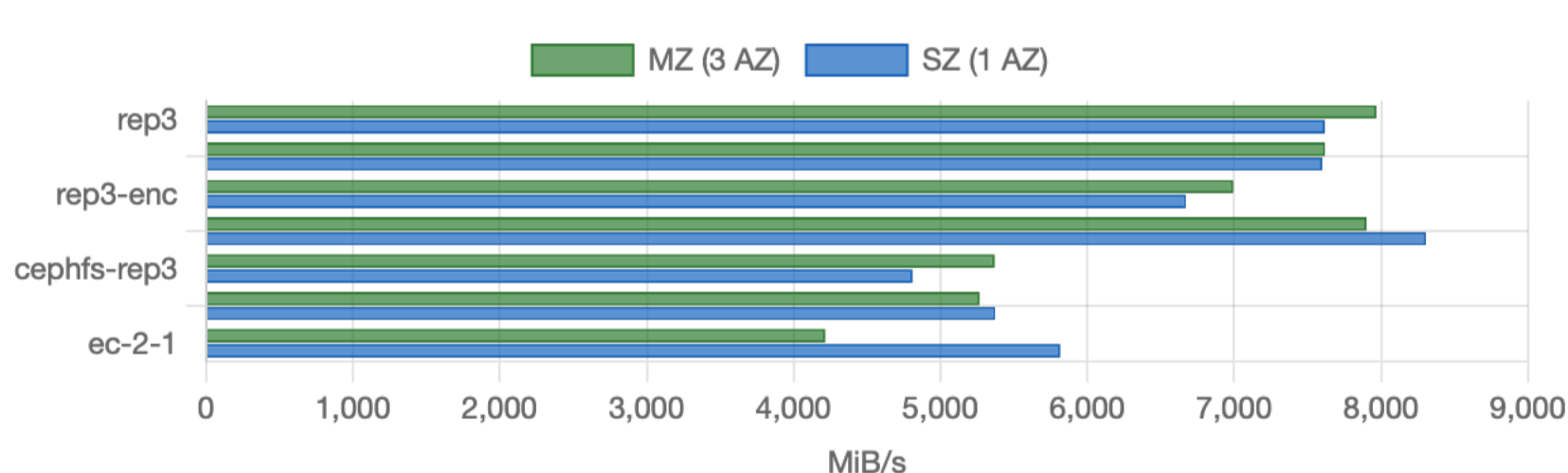
Pool	MZ Read	SZ Read	Read Δ	MZ Write	SZ Write	Write Δ
rep3	4.4	3.9	+13.0%	379.6	208.7	+81.9%
rep3-enc	4.5	3.9	+16.4%	388.0	212.9	+82.3%
rep2	4.4	3.3	+33.1%	308.3	194.0	+58.9%
cephfs-rep3	6.0	7.0	-14.9%	1,502	885.0	+69.7%
cephfs-rep2	6.0	5.9	+1.0%	1,216	868.2	+40.1%
ec-2-1	5.6	5.3	+6.3%	557.8	354.4	+57.4%

Visual Comparison

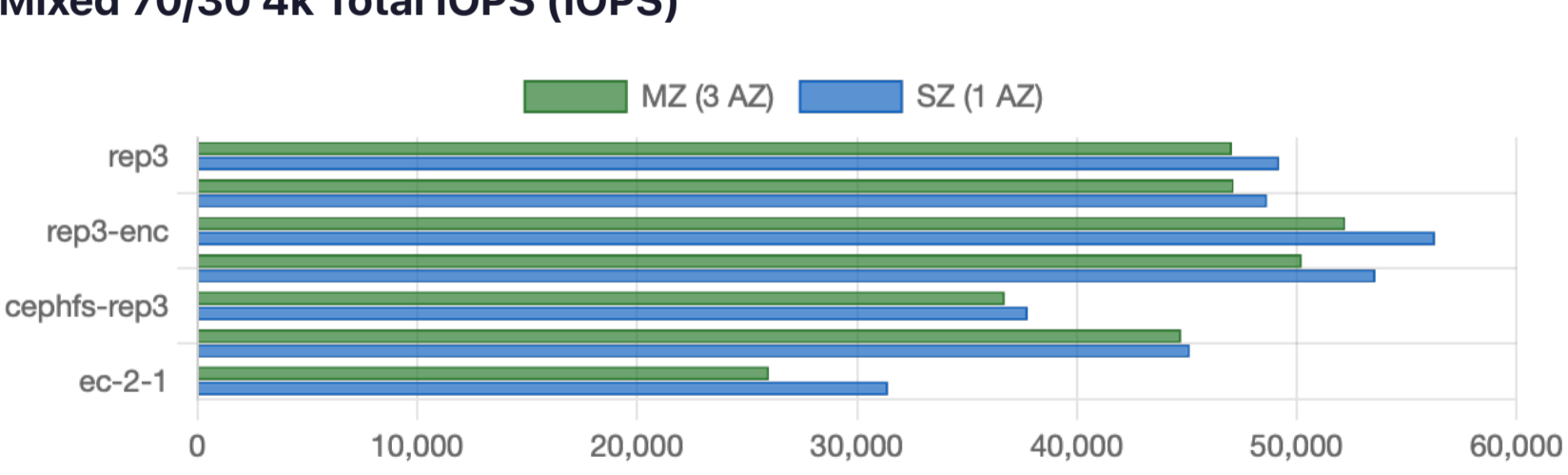
Random 4k Total IOPS (IOPS)



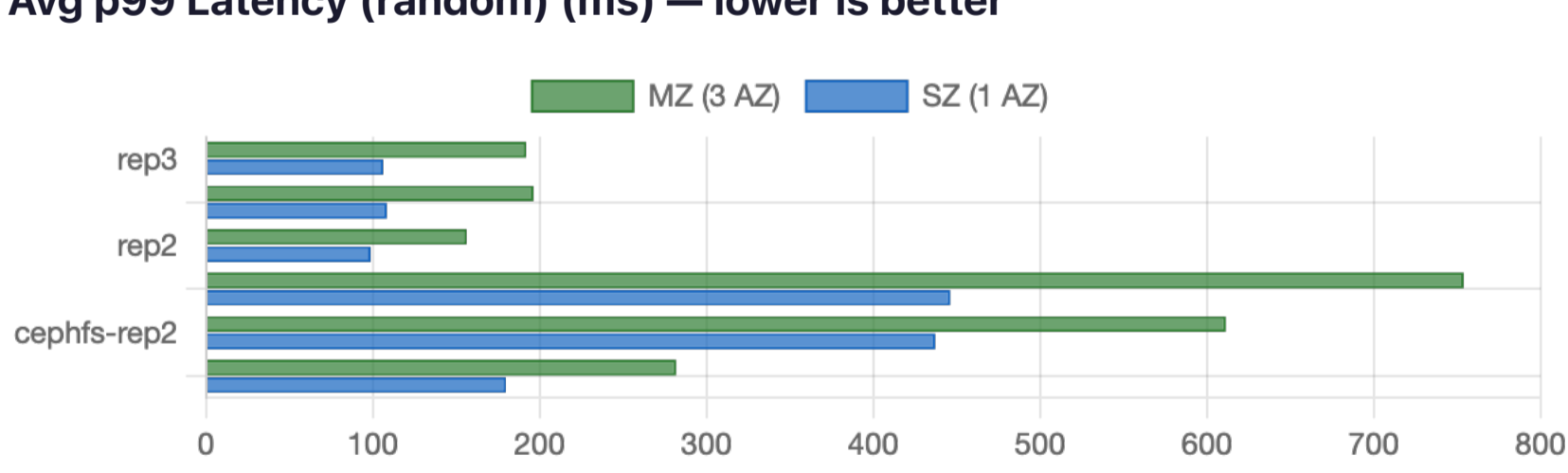
Sequential 1M Total BW (MiB/s)



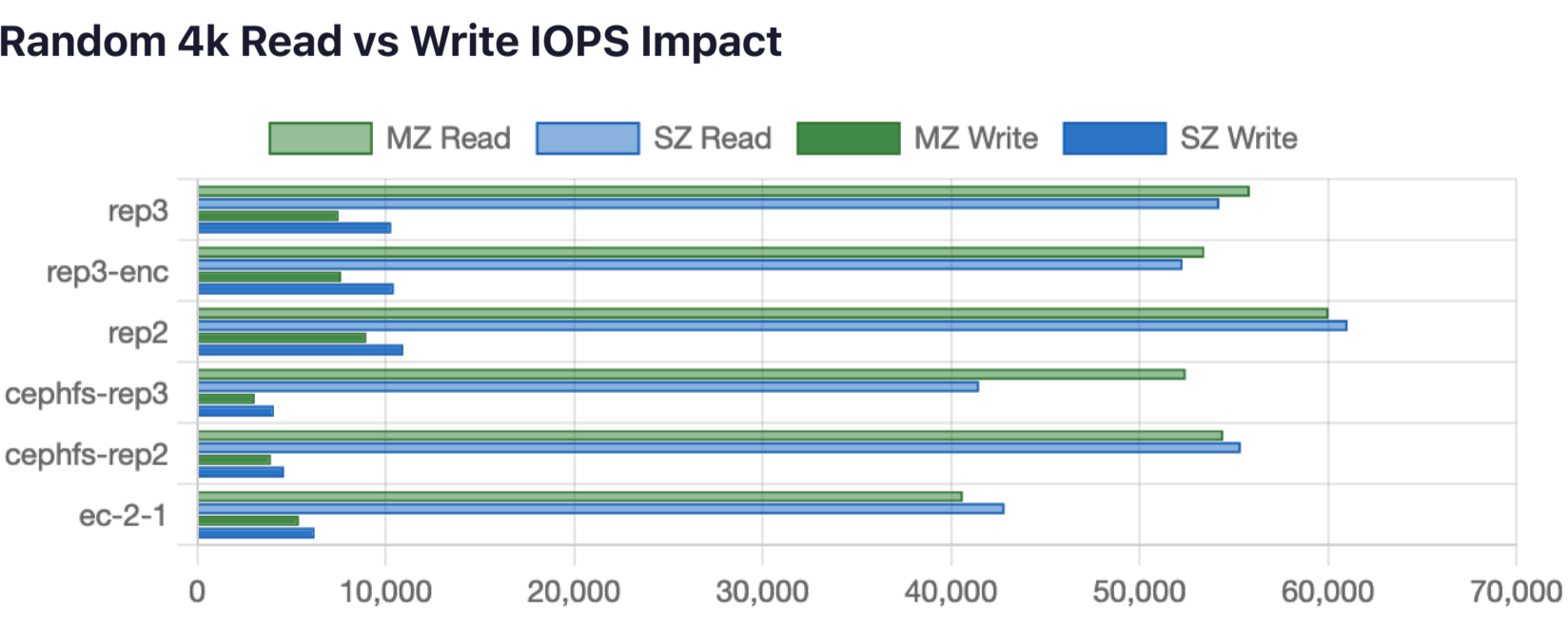
Mixed 70/30 4k Total IOPS (IOPS)



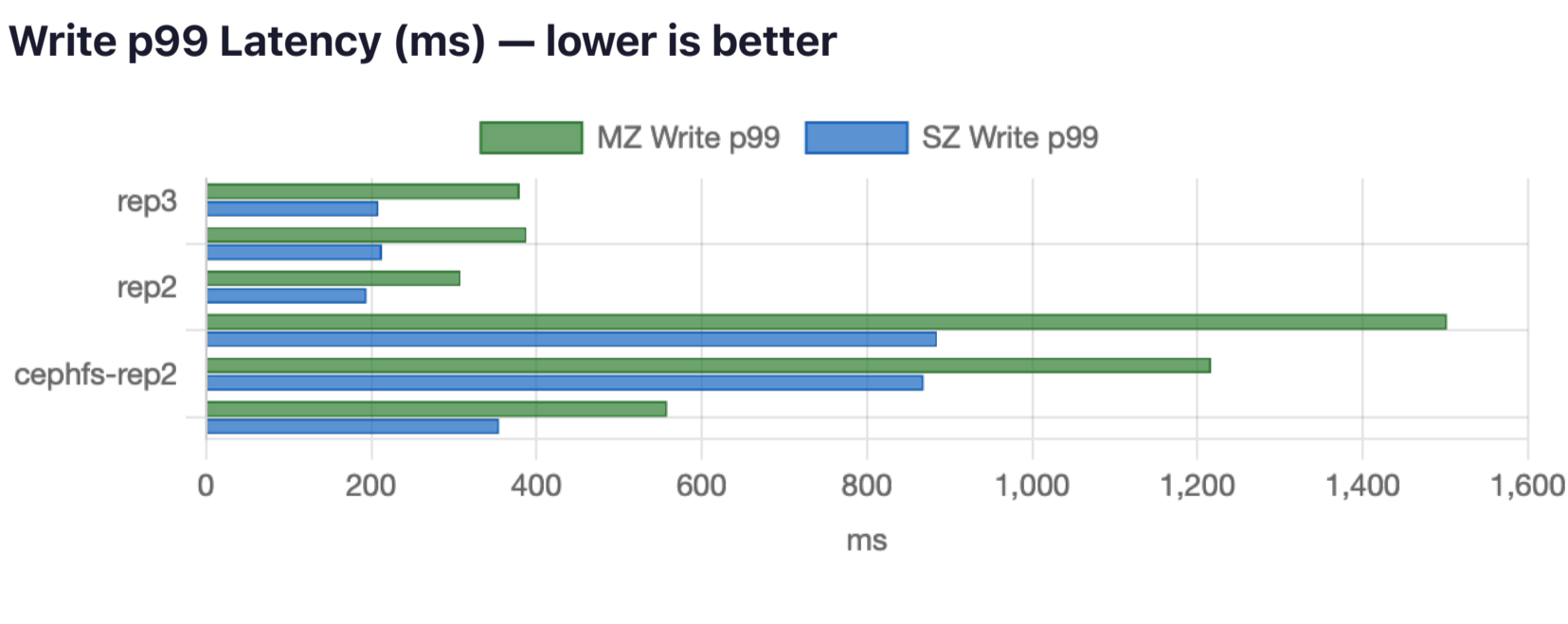
Avg p99 Latency (random) (ms) — lower is better



Random 4k Read vs Write IOPS Impact



Write p99 Latency (ms) — lower is better



Scorecard Summary

Pool	Random IOPS	Seq BW	Mixed IOPS	p99 Latency
rep3	~Tie	~Tie	~Tie	SZ
rep3-virt	N/A	~Tie	~Tie	N/A
rep3-enc	~Tie	~Tie	SZ	SZ
rep2	~Tie	~Tie	SZ	SZ
cephfs-rep3	MZ	MZ	~Tie	SZ
cephfs-rep2	~Tie	~Tie	~Tie	SZ
ec-2-1	SZ	SZ	SZ	SZ
Overall	1 MZ / 1 SZ	1 MZ / 1 SZ	0 MZ / 3 SZ	0 MZ / 6 SZ

Key Takeaways

Write IOPS: consistent 10–19% regression on MZ. Every pool shows lower random write IOPS on MZ. RBD rep3 dropped from 10,282 to 7,515 write IOPS (–27%), rep2 from 10,932 to 8,955 (–18%), and ec-2-1 from 6,236 to 5,380 (–14%). This directly reflects the cost of waiting for cross-AZ replica/chunk acknowledgments.

Read IOPS: effectively identical. Read affinity works as designed — reads are served from the local-zone OSD replica. Random read IOPS differ by less than 5% across all RBD pools (e.g., rep3: 55,841 MZ vs 54,222 SZ = +3%). CephFS reads actually improved on MZ, likely due to the MZ cluster having more PGs (512 vs 256) for the CephFS data pool.

Write p99 latency: the clearest MZ penalty. Random write p99 increased 40–82% on RBD pools (rep3: 380ms vs 209ms, rep2: 308ms vs 194ms). The worst case is ec-2-1 at 558ms vs 354ms (+57%), where EC encoding distributes data chunks across all 3 AZs. CephFS write p99 shows the largest absolute increase: cephfs-rep3 hit 1,502ms vs 885ms (+70%).

Sequential throughput: MZ slightly better on some pools. Surprisingly, several pools show higher sequential BW on MZ (rep3: +4.6%, rep3-enc: +4.8%, cephfs-rep3: +11.6%). This may reflect faster PG autocalc sizing on the MZ cluster or regional network bandwidth differences. ec-2-1 is the outlier with –27.5% sequential BW regression.

EC pools hit hardest by cross-AZ placement. ec-2-1 shows the largest regressions across almost every metric: –6.3% random IOPS, –27.5% sequential BW, –17.2% mixed IOPS, +57% write p99. EC encoding distributes data and parity chunks across all 3 failure domains, meaning every I/O touches all 3 AZs.

rep2 benefits from fewer AZ acks. With only 2 replicas, rep2 needs acknowledgment from just 1 remote AZ (vs 2 for rep3). This gives rep2 a structural advantage on MZ: it shows the smallest write IOPS regression (–18% vs –27% for rep3) and the lowest write p99 (308ms vs 380ms for rep3).

CephFS overhead stable across zones. The MDS latency component is the same regardless of AZ topology. Read latency is nearly identical (5.99ms MZ vs 5.93–7.04ms SZ). The cross-AZ penalty only shows up in writes, consistent with the MDS being a metadata path while data replication is the bottleneck.

Bottom line for production: Read-heavy workloads will see no meaningful performance difference on multi-zone ODF. Write-heavy workloads should expect a 10–20% IOPS reduction and 1.5–2x p99 latency increase — the unavoidable cost of surviving a full AZ failure. Avoid EC pools on MZ if write latency is critical; prefer rep2 for the best MZ write performance.