

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)		SZ Single-Zone (ocp-virt-420-cluster)	
Region	us-south (3 AZs)	Region	eu-de (1 AZ)
Workers	3x b2d.metal.96x384 (1 per AZ)	Workers	3x b2d.metal.96x384 (same zone)
OSDs	24 (8 per node, NVMe)	OSDs	24 (8 per node, NVMe)
Raw capacity	~70 TiB	Raw capacity	~70 TiB
Failure domain	zone (cross-AZ replicas)	Failure domain	rack (intra-zone replicas)
ODF / Ceph	4.19.10 / Squid 19.2.1	ODF / Ceph	4.19.10 / Squid 19.2.1
Encryption	Full (data + network + KMS)	Encryption	Full (data + network + KMS)
Read affinity	Enabled	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 wins SZ 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

