Ceph OSD Node Network Tuning Checklist

1. Baseline and Topology Validation

- [ ] Confirm cluster network design
  - [ ] Separate public_network and cluster_network (recommended for production)
- [ ] Validate MTU consistency across:
  - [ ] Switch ports
  - [ ] Bond interfaces
  - [ ] VLAN interfaces
  - [ ] Ceph nodes
- [ ] Verify NIC speed and duplex

```
ethtool <iface>
```

- [ ] Confirm no packet drops

```
ethtool -S <iface> | egrep 'drop|error'
```

---

2. Jumbo Frames (If Enabled)

- [ ] Set MTU 9000 (or consistent value) on:
  - [ ] OS interface
  - [ ] Bond/VLAN
  - [ ] Switch ports
- [ ] Validate MTU

```
ip link show <iface>
```

- [ ] Test jumbo frame path

```
ping -M do -s 8972 <peer-ip>
```

Note: Enable only if the entire network path supports jumbo frames.

---

3. Kernel Network Buffer Tuning

Create or edit:

`/etc/sysctl.d/99-ceph-network.conf`

```
net.core.rmem_max = 268435456
net.core.wmem_max = 268435456
net.core.rmem_default = 67108864
net.core.wmem_default = 67108864

net.ipv4.tcp_rmem = 4096 87380 134217728
net.ipv4.tcp_wmem = 4096 65536 134217728

net.core.netdev_max_backlog = 250000
net.ipv4.tcp_max_syn_backlog = 8192

net.ipv4.tcp_tw_reuse = 1
```

Apply:

```
sysctl --system
```

Checklist:

- ☐ rmem_max and wmem_max set to 256MB or higher
- ☐ netdev_max_backlog set to 250000 or higher
- ☐ tcp_max_syn_backlog increased

---

4. Increase NIC Queue Length

```
ip link set <iface> txqueuelen 10000
```

- [ ] txqueuelen ≥ 5000 for 10G
- [ ] txqueuelen ≥ 10000 for 25G or higher
- [ ] Persist in network configuration

---

5. Ring Buffer Size

Check current values:

```
ethtool -g <iface>
```

Set to maximum supported:

```
ethtool -G <iface> rx 4096 tx 4096
```

- [ ] RX and TX ring buffers maximized
- [ ] No driver errors after change

---

6. Multi-Queue and IRQ Balancing

- [ ] Verify multi-queue support

```
ls -d /sys/class/net/<iface>/queues/tx-*
```

- [ ] Enable irqbalance

```
systemctl enable irqbalance
systemctl start irqbalance
```

- [ ] Verify interrupt distribution

```
cat /proc/interrupts
```

Optional for advanced setups:

- [ ] Manual IRQ pinning aligned with NUMA topology

---

7. Offloading Features

Check current settings:

```
ethtool -k <iface>
```

Recommended:

- [ ] GRO enabled
- [ ] GSO enabled
- [ ] TSO enabled
- [ ] LRO disabled

Disable LRO if required:

```
ethtool -K <iface> lro off
```

---

8. Bonding (If Used)

- [ ] Use mode 4 (802.3ad LACP)
- [ ] Switch-side LACP properly configured
- [ ] Set transmit hash policy to layer3+4

Example:

```
xmit_hash_policy=layer3+4
```

Validate:

```
cat /proc/net/bonding/bond0
```

---

9. Ceph Network Configuration

In ceph.conf:

```
ms_bind_ipv4 = true
ms_bind_ipv6 = false
ms_async_op_threads = 3
```

- [ ] Confirm msgr2 enabled
- [ ] Ensure no unnecessary fallback to msgr1

Check monitor configuration:

```
ceph mon dump
```

---

10. NUMA Awareness (High-Core Systems)

- [ ] Check NUMA topology

```
numactl --hardware
```

- [ ] Ensure NIC and OSD CPU cores are aligned
- [ ] Pin OSD processes to local NUMA node if required

---

11. Monitor for Drops and Saturation

Live monitoring:

```
sar -n DEV 1
iftop
nload
```

- [ ] No RX/TX drops
- [ ] No interface saturation
- [ ] No excessive TCP retransmissions

---

12. Validation After Changes

- [ ] Restart OSDs one by one
- [ ] Check cluster health

```
ceph -s
ceph osd perf
```

- [ ] Run benchmark

```
rados bench -p <pool> 60 write
```

---

13. Optional Advanced Tuning (25G/40G/100G)

- [ ] Tune RPS/XPS if needed
- [ ] Set default qdisc to fq

```
sysctl -w net.core.default_qdisc=fq
```

- [ ] Consider BBR congestion control

```
net.ipv4.tcp_congestion_control=bbr
```

- [ ] Validate switch buffer configuration

Revision #2
Created 2026-02-19 12:36:30 UTC by Mesut Bayrak
Updated 2026-02-19 12:47:06 UTC by Mesut Bayrak