For Amazon EBS gp3 volumes with IOPS of 3000 and throughput configurations of 125 MiB/s or 250 MiB/s, here are some typical performance benchmarks and guidance on interpreting the values for the specified metrics:

1. VolumeQueueLength

Expected Range: Ideally, VolumeQueueLength should be close to 0 most of the time, indicating that there’s minimal waiting for I/O operations.

Typical Values:

A value between 0 to 1 is typical for moderate workloads on gp3 volumes.

Threshold for Concern: If VolumeQueueLength is consistently above 5, it could indicate that the volume is under heavy load or bottlenecked by IOPS or throughput limits.

Guidance: Spikes in VolumeQueueLength during peak periods are normal, but sustained high values suggest the need for tuning or re-evaluation of the instance’s workload.

2. NetworkIn / NetworkOut

Expected Range: The exact values will depend on the volume of data transferred to and from the instance.

Typical Throughput: With gp3 volumes’ throughput limits:

125 MiB/s throughput: Expect network traffic rates up to 125 MiB/s when the volume reaches its throughput cap.

250 MiB/s throughput: For volumes configured with 250 MiB/s, network traffic rates can reach up to this value during peak write/read operations.

Guidance:

High NetworkIn/Out values could indicate large data transfers or significant application traffic.

Track this metric alongside VolumeQueueLength and Write/Read Latency to identify potential bottlenecks.

3. NetworkPacketsIn / NetworkPacketsOut

Expected Range: Network packets depend on the data payload size and number of requests. For typical EC2 use cases with gp3 volumes:

High packet rates are common in applications with small but frequent read/write operations (e.g., transactional databases).

Threshold for Concern: Extremely high or spiking packet rates could indicate a high-volume, low-latency workload, which may require network tuning.

Guidance: Track NetworkPacketsIn/Out alongside NetworkIn/Out to ensure there’s sufficient bandwidth and that latency is within acceptable levels.

4. Average Write Latencies

Expected Range: For gp3 volumes, write latencies are typically single-digit milliseconds (1–5 ms) due to their SSD nature.

Typical Values:

0–2 ms is ideal for moderate workloads.

Threshold for Concern: Sustained latencies above 10 ms may indicate IOPS or throughput limits, or high queue lengths.

Guidance: If average write latencies rise, check VolumeQueueLength and EBSWriteOps to confirm whether the volume can handle the current workload.

5. EBSWriteOps

Expected Range: With 3000 IOPS configured for gp3 volumes:

Average Write Ops can reach up to 3000 writes per second, depending on workload demands.

Guidance:

Track EBSWriteOps along with VolumeQueueLength and Write Latency. If EBSWriteOps is consistently near the 3000 IOPS limit, it could indicate that your workload is fully utilizing the IOPS provisioned for the volume.

If VolumeQueueLength also increases in these situations, consider increasing the volume’s IOPS or optimizing the workload.

Summary of Expected Value Ranges for Monitoring

| Metric | Expected Range | Threshold for Concern |
| --- | --- | --- |
| VolumeQueueLength | 0–1 (ideal), spikes <5 | Sustained >5 |
| NetworkIn/Out | Up to 125 MiB/s or 250 MiB/s | N/A (Monitor for trends) |
| NetworkPacketsIn/Out | Variable based on workload | Spikes or consistently high values |
| Average Write Latency | 1–5 ms (ideal) | Sustained >10 ms |
| EBSWriteOps | Up to 3000 per second | Consistently near 3000 + queue buildup |

Monitoring these metrics together helps you get a comprehensive view of instance and EBS volume health, enabling you to address any potential performance bottlenecks proactively.