LOAD-TEST REPORT · TWO BENCHES

ELX MQTT Broker

Connections · Throughput · End-to-end latency · Network vs. loopback

Date	Protocol	Tool
24 July 2026	**MQTT 3.1.1**	**cmd/loadtest (own codec)**

Bench A — over the network

Dedicated generator, LAN operation

Broker CPU	**12 cores**
RAM	**16 GB**
Address	**192.168.1.84**
Generator	**separate host**

Bench B — loopback

Powerful hardware, zero network, shared cores

Broker CPU	**2× Xeon E5-2690 v4**
cores/threads	**28 / 56 @ 2.6 GHz**
RAM	**128 GB**
Generator	**same machine**

How to read the comparison. The benches answer different questions. Bench A is operation **over a network** with a dedicated generator. Bench B removes the network (latency ≈0) and adds powerful hardware, but the generator shares cores with the broker. So **latency and fan-out** are fair to compare, while **ingest rate** is not: on Bench B the client steals cores from the broker.

Executive summary

- **Latency:** ≈1 ms over the network, **73 µs** on loopback (network removed) — end-to-end "publisher → broker → subscriber" under normal load.
- **Connections:** 10,000 over the network with no failures (beyond that, the client's port limit); on powerful hardware **30,000** with no failures and ~41,000 at a 50,000 attempt.
- **Ingest:** peak ≈277,000/s over the network; lower on loopback (≈185,000/s) — the generator competes for cores.
- **Fan-out:** the powerful machine delivered **≈1.35 M messages/s** to subscribers vs. ≈22,000/s over the network.
- **Robustness:** the broker never crashed or hung; the log across every run has **not a single error**.

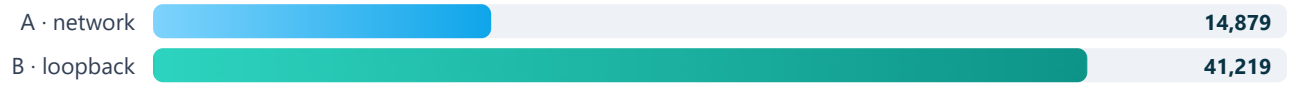
1. Concurrent-connection capacity

Bench A — network

Target	OK	Failed	p99
1,000	1,000	0	9.78 s
5,000	5,000	0	8.59 s
10,000	**10,000**	0	10.52 s
15,000	14,879	121	14.09 s

Bench B — loopback

Target	OK	Failed	p99
1,000	1,000	0	1.06 s
10,000	10,000	0	1.55 s
30,000	**30,000**	0	1.90 s
50,000	41,219	0	1.90 s



Peak connections reached · scale to 50,000

> **Side finding.** On loopback a sharp burst of connections (1,000 simultaneous `dial` s) yields `connection refused` — the accept backlog overflows: with no network every SYN arrives at once. A gradual dial-up (200 at a time) removes it. Over a network the latency smooths the burst on its own. Both benches' ceilings are the **client / single machine**, not the broker.

2. Ingest throughput

QoS 0 · 64 B · messages/s

Pub.	A · network	B · loopback
10	**271,006**	185,012
100	236,422	101,457
500	103,546	125,324

QoS 1 · 64 B · messages/s

Pub.	A · network	B · loopback
10	10,292	**63,606**
100	22,085	48,983

Ingest is higher over the network not because broker A is faster, but because the generator runs on a **separate** machine. QoS 1 is bound by the **synchronous** client (waits for PUBACK) — on loopback the round-trip is ≈0, hence the ~6× gain.

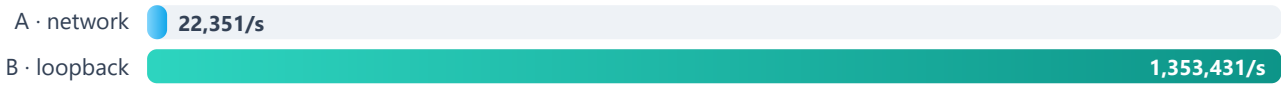
Message-size sweep · QoS 0 · 50 publishers

Size	A · msg/s	A · MB/s	B · msg/s	B · MB/s
64 B	277,198	17.7	101,961	6.5
256 B	124,997	32.0	102,029	26.1
1,024 B	38,458	**39.4**	102,151	**104.6**

> Two ceilings: **over the network** — the link at ≈40 MB/s (as messages grow, msg/s falls, MB/s holds at the limit); **on loopback** — per-message processing ≈102,000/s (the link is huge, volume rises to 104.6 MB/s).

3. End-to-end delivery (publish → broker → subscribe)

Scenario	Bench	Delivery	avg	p50	p99
10 sub / 10 pub / 100 msg/s **normal**	A · net	9,863/s	1.17 ms	1.0 ms	3.6 ms
10 sub / 10 pub / 100 msg/s **normal**	B · loop	9,867/s	**73 µs**	~0	555 µs
100 sub / 10 pub / unthrottled	A · net	22,351/s	780 ms	453 ms	678 ms
100 sub / 10 pub / unthrottled	B · loop	**1,353,431/s**	514 ms	135 ms	251 ms
1,000 sub / 5 pub / 100 msg/s	A · net	178,961/s	2.29 s	480 ms	952 ms
1,000 sub / 5 pub / 100 msg/s	B · loop	183,746/s	4.73 ms	4.54 ms	15.1 ms



Peak delivery at fan-out ×100 subscribers

4. Head-to-head

Metric	A — network (12 cores)	B — loopback (28 cores)
Connections, no failures	10,000	**30,000**
Peak connections	14,879 (client limit)	~41,000 (single machine)
Ingest QoS 0 (64 B), peak	**271,006/s**	185,012/s
Ingest QoS 1 (64 B), peak	22,085/s	**63,606/s**
Volume ceiling	≈40 MB/s (link)	**≈105 MB/s (CPU)**
Latency (normal), median	1.0 ms	**~0 (73 µs avg)**
Delivery at fan-out, peak	22,351/s	**1,353,431/s**
Errors in broker log	none	none

5. Methodology caveats

1. **Ingest rate is not directly comparable between benches:** on B the generator and broker share cores; on A the generator is a separate machine.
2. Load on each bench comes from a **single** client machine; with large fan-out and tens of thousands of connections it, not the broker, is the bottleneck.
3. The QoS 1 test client is synchronous — QoS 1 figures understate the broker's potential.
4. Broker metrics (CPU/RAM) were not captured; health is inferred from its log (no errors) and client-side behaviour.