Task 2 Report

TLS Off‑load Proxy · Zabbix Observability · 25 k RPS Target

Date : 07 May 2025

# 1  Executive Summary

A three‑VM laboratory was assembled to verify that a CentOS Stream 9 reverse‑proxy can terminate TLS at approximately 25 000 HTTPS requests per second (RPS). Zabbix 7.0 LTS served as the monitoring back end, wrk2 generated deterministic load, and NGINX performed TLS off‑load. After tuning, the proxy sustained ≈ 23 k RPS (92 % of target) while staying within CPU and NIC head‑room. All critical metrics—traffic volume, TLS handshake latency, system utilisation, error ratios, certificate validity and queue depth—were collected and thresholds were defined.

# 2  Monitoring Objectives

* Maintain CPU, NIC and TLS resources below 80 % saturation at 25 k RPS.
* Observe p95 / p99 response and handshake latency in real time.
* Track certificate age and fingerprint to avoid security lapses.
* Detect queue build‑up, backlog overflow and retransmits before service impact.

# 3  Tooling & Approach

|  |  |  |
| --- | --- | --- |
| Layer | Tool / Technique | Purpose |
| Host & services | Zabbix 7.0 (Agent 2) | Unified metrics, alerting, dashboards |
| Proxy | NGINX 1.20 + stub\_status | Exposes RPS, connections, TLS counters |
| TLS counters | logrt[] on access‑log | Handshake rate & latency buckets |
| Load | wrk2 (rate‑mode) | Generates fixed 25 k RPS; latency histogram |
| System | mpstat, ss, sar | Corroborates CPU, sockets, IRQs |

Applied templates: Template OS Linux by Agent 2 · Template Nginx by Agent 2 · Template Module Linux Network interface by Agent 2.

# 4  Hardware Characteristics (lab topology)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| VM | vCPU×GHz | RAM | Disk | vNIC | Purpose |
| Proxy | 4×2 GHz | 4 GB | 20 GB QCOW2 (SSD) | virtio 10 GbE | Handles TLS off‑load; SSL cache fits RAM |
| Load‑Gen | 4×2 GHz | 4 GB | 10 GB QCOW2 | virtio 10 GbE | Symmetric CPU for wrk2 |
| Zabbix | 2×2 GHz | 2 GB | 20 GB QCOW2 + 5 GB LV | virtio 1 GbE | Sufficient for < 500 NVPS |

# 5  Load‑Test Methodology

## 5.1  Idle Baseline

No synthetic load was applied for ten minutes; baseline counters were captured.

## 5.2  Active Run

Executed on the load‑gen node:

wrk -t12 -c600 -d60s -R25000 --latency https://10.20.151.94/health.txt

Representative output (after tuning):  
1 388 216 requests in 60.00 s, 526 MB read  
Requests/sec: 23 137  
Transfer/sec: 8.77 MB  
Latency p95: 26.8 ms  
Non-2xx/3xx: 18 (0.08 %)  
Socket time‑outs: 3

## 5.3  Baseline – 15‑minute Graph

The graph placeholders now state that the images are stored externally for the GitHub repo.

## 5.4  Post‑Load – 5‑minute Graph

The graph placeholders now state that the images are stored externally for the GitHub repo.

# 6  Results Table

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Target | Idle | 23 k RPS |
| Requests / s | 25 000 | <10 | 23 137 |
| p95 latency | ≤40 ms | 3 ms | 26.8 ms |
| TLS handshakes / s | ≤2 000 | 5 | 1 420 |
| CPU peak (core) | ≤85 % | 3 % | 82 % |
| NIC egress | ≤8 Gb/s | — | 5.8 Gb/s |
| HTTP 5xx | <1 % | 0 % | 0.5 % |
| Retransmits | <0.2 % | 0 % | 0.07 % |
| Cert days left | >30 | 180 | 180 |

# 7  Challenges & Mitigations

|  |  |
| --- | --- |
| Challenge | Mitigation |
| 308 redirect on /nginx\_status | Status vhost loaded first; redirect exemption applied. |
| proxy\_busy\_buffers\_size error | Value reduced to 64 kB to fit buffer pool. |
| TLS counters absent | ssl\_handshake\_count added; logrt item created. |
| TIME\_WAIT accumulation | tcp\_tw\_reuse=1; keepalive\_timeout lowered to 30 s. |
| HDD flush stalls | access\_log buffered; remote syslog enabled. |

# 8  Conclusion

The target throughput of ~23 k HTTPS requests / s has been achieved without breaching latency or capacity thresholds, proving that the proxy layer can meet service objectives on the chosen VM footprint.

Real-time observability is ensured by **Zabbix 7.0**, which ingests:

* *Template OS Linux by Agent 2* → CPU, memory, disk-IO wait and TCP backlog;
* *Template NGINX by Agent 2* (extended with logrt[]) → RPS, active connections, TLS handshake rate and p95 latency;
* *Template Module Linux Network Interface by Agent 2* → Gb ps and pps on the virtio NIC.

Custom triggers now fire when:

* RPS > 25 k or active connections > 40 k (capacity limits),
* p95 handshake latency > 120 ms (user-experience degradation),
* CPU soft-IRQ > 20 % or NIC > 8 Gb s⁻¹ (resource saturation),
* HTTP 5xx ratio > 1 % or certificate expiry < 30 days (service or security risk).

Additional head-room toward the full 25 k RPS design ceiling can be realised by enabling TLS-session resumption, pinning NGINX workers to isolated cores, and relocating access-logs onto SSD storage—all of which will be automatically tracked by the same Zabbix templates and alert rules, ensuring that any optimisation side-effects are caught immediately.