**MCPX‑KendoBridge — Non‑Functional Requirements (NFR)**

**Document:** docs/06\_nfr.docx  
**Project Code:** MCPX‑KendoBridge  
**Version:** 2.0.0 (Kendo Migration)  
**Last Updated:** 2025‑09‑27  
**Owner:** DoSE (Accountable) — SRE Lead (Responsible) — DocFactory (Author) — SecLead, T‑Arch, QA, DBA (Consulted)

**Purpose.** Define measurable **quality attributes and budgets** for performance, availability, scalability, security, operability, maintainability, accessibility, and compliance—aligned to Technijian’s GitHub‑first SDLC and evidence model. This NFR set drives monitoring SLOs, CI/CD gates, test plans, and runbooks for **Alpha → Beta → RTM → Prod**, with **RTM validating on the Prod DB (read‑only)** to catch drift before production.

**DB COMPLIANCE (applies to all NFRs):** Schema changes are **add‑only**; the application has **Stored‑procedure‑only** DAL (**EXECUTE‑only** on SPs); **No‑Hard‑Coding** of dynamic values. Runtime settings (child command/args/cwd, request timeouts, SSE keep‑alive cadence, Origin allow‑list, feature flags) are **DB‑sourced** via AppConfig/FeatureFlag through **sp\_Config\_GetValue**, **sp\_Config\_GetAll**, **sp\_Feature\_IsEnabled**, **sp\_Lookup\_Get**. **Secrets** (SQL connection strings, Telerik license) are **never** stored in code/docs/DB—configure them in **GitHub Environments** or vendor portals only.

**1) Scope & Environments**

* **Components in scope:** .NET 8 MCP proxy (Streamable‑HTTP + SSE), SQL Server (non‑secret config/flags), KendoReact Admin Portal (read‑only ops).
* **Environments:** **Alpha → Beta → RTM → Prod**; **RTM** points to **Prod DB (read‑only)** and blocks promotion on parity drift.

**2) SLO Summary (targets) & SLIs (how measured)**

| **Domain** | **SLO Target (per env; Prod enforced)** | **Primary SLI** | **Source / Notes** |
| --- | --- | --- | --- |
| **Availability** | **≥ 99.9% monthly** | Availability per endpoint (/mcp, /ready, /healthz) | Error budget ≈ 43.2 min/mo; failures are 5xx. |
| **Latency (JSON)** | **p50 ≤ 300 ms; p95 ≤ 800 ms** | /mcp non‑streaming | Intra‑VPC. Budget verified in Beta & Prod. |
| **Streaming TTFB** | **p95 ≤ 200 ms** | First byte for POST /mcp with Accept: text/event-stream | Heartbeats required. |
| **Restart‑to‑Ready** | **≤ 30 s** | Time from container start to /ready=ok | Shown on Dashboard; alert if breached. |
| **Concurrency** | **≥ 200 concurrent sessions/replica** | session\_count per pod | CPU‑bound before memory per design. |

**Verification:** SLOs are enforced by CI perf smoke (Beta), RTM parity checks (Config/Ready), and 24‑hour post‑release monitoring snapshots (Prod). Evidence retained **≥ 1 year**.

**3) Performance & Throughput**

**NFR‑P1 — Non‑streaming latency.**

* **Target:** p50 ≤ 300 ms, p95 ≤ 800 ms for /mcp when responding with JSON.
* **Measure:** Histogram buckets http\_server\_duration\_ms\_bucket{path="/mcp",mode="json"}.
* **Validation:** CI k6 smoke (Beta/Prod) and dashboard p50/p95.
* **Notes:** Timeouts are **DB‑driven** via Network:RequestTimeoutSeconds (range 30–600).

**NFR‑P2 — Streaming TTFB.**

* **Target:** p95 ≤ 200 ms for first byte after POST /mcp with Accept: text/event-stream.
* **Measure:** sse\_ttfb\_ms\_bucket{path="/mcp"}; browser synthetic optional.
* **Validation:** E2E streamed test (02\_streamed\_tool\_call), perf smoke, monitoring panel.

**NFR‑P3 — Heartbeat cadence.**

* **Target:** gap between SSE heartbeats ≈ Network:SseKeepAliveSeconds (±1 s).
* **Measure:** sse\_heartbeat\_gap\_ms\_bucket.
* **Validation:** E2E, UI display of heartbeat “age,” streaming dashboard.

**NFR‑P4 — Child spawn overhead.**

* **Target:** Child spawn & warmup ≤ 2 s p95 on Alpha/Beta.
* **Measure:** timed segment in readiness probe (optional) and logs.
* **Validation:** Readiness gating; incident runbook if exceeded.

**4) Availability, Reliability & Resilience**

**NFR‑A1 — Availability.**

* **Target:** ≥ 99.9% monthly across primary endpoints.
* **Measure:** 1 − (5xx / total), per path.
* **Controls:** Readiness gates during rollout; PDB and graceful drain for scale‑in; sticky sessions by Mcp‑Session‑Id.

**NFR‑A2 — Graceful shutdown.**

* **Target:** No mid‑stream truncation on rolling updates.
* **Measure:** Streams complete within terminationGracePeriodSeconds; “drain” logs observed.
* **Controls:** flip readiness → drain SSE → terminate child → exit. (See runbooks.)

**NFR‑A3 — Error handling.**

* **Target:** All HTTP errors use the **stable envelope** {code,message,requestId?}; **no secrets** or payload bodies in messages/logs.
* **Measure:** Contract tests; log scrapers.
* **Codes:** origin\_forbidden, missing\_session\_id, feature\_disabled, timeout, not\_ready, spawn\_failed, internal\_error. (See Error Catalog.)

**5) Scalability & Capacity**

**NFR‑S1 — Horizontal scale.**

* **Target:** Support **≥ 200 concurrent sessions per replica**; HPA scales on CPU (primary) and optional session\_count.
* **Measure:** session\_count per pod, CPU utilization, TTFB under load.
* **Controls:** Sticky routing by Mcp‑Session‑Id; **PDB** to prevent mass eviction; ingress **no buffering** for SSE.

**NFR‑S2 — Resource bounds.**

* **Target:** CPU becomes the limiting factor before memory; no OOM at target concurrency.
* **Measure:** Node/pod CPU & RSS; GC stats if enabled.
* **Controls:** Set requests/limits; cap maxReplicas to avoid runaway autoscale.

**NFR‑S3 — Config path latency.**

* **Target:** p95 config\_fetch\_duration\_ms ≤ 200 ms.
* **Measure:** Instrumented around calls to **SPs** (sp\_Config\_\*, sp\_Feature\_IsEnabled).
* **Notes:** DB is control‑plane for **non‑secret** settings; **no inline SQL**.

**6) Security & Privacy**

**NFR‑Sec1 — Origin allow‑list.**

* **Target:** Requests with Origin not in **DB** Security:AllowedOrigins return 403 origin\_forbidden.
* **Measure:** Error counters by code; integration tests.
* **Notes:** Values are visible through /config/effective (**non‑secret**).

**NFR‑Sec2 — Secret handling.**

* **Target:** **Zero** secret leakage in code, DB, logs, artifacts.
* **Measure:** Secret Scanning; log redaction checks.
* **Notes:** SQL connection strings & **Telerik license** live only in **GitHub Environments**.

**NFR‑Sec3 — Transport & headers.**

* **Target:** TLS everywhere; ingress **not buffering** text/event-stream; CSP default‑deny for UI; security headers present.
* **Measure:** Synthetic checks; header verifiers in CI (optional).
* **Notes:** See Compliance & UI/UX docs for CSP baseline.

**7) Operability, Observability & Supportability**

**NFR‑O1 — Health & readiness.**

* **Target:** /healthz 200 during normal ops; /ready reflects DB/SP reachability and (optionally) child probe.
* **Measure:** Probes every 30s; readiness flips pre‑drain on rollout.
* **Notes:** RTM /ready uses **Prod DB (read‑only)**.

**NFR‑O2 — Telemetry.**

* **Target:** Structured JSON logs with requestId, sessionId, childPid, status, latency\_ms, mode=json|sse; **no payload bodies**.
* **Measure:** Sampling of logs; automated schema checks.

**NFR‑O3 — Metrics.**

* **Target:** Expose SLIs: availability, latency, **TTFB**, heartbeat gaps, session/child counts, readiness, config fetch duration.
* **Measure:** Prometheus/Otel metrics.
* **Notes:** Alerting thresholds defined in Monitoring doc.

**NFR‑O4 — Evidence & retention.**

* **Target:** Attach CI, contract, a11y, SBOM, OpenAPI, monitoring snapshots, approvals to Release; retain **≥ 1 year**.
* **Measure:** Release artifacts present per checklist.

**8) Maintainability & Change Management**

**NFR‑M1 — GitHub‑first gates.**

* **Target:** Merge queue with required checks: Build/Tests, OpenAPI lint/diff, CodeQL, Dependency Review (fail on **High**), Secret Scanning (setting), SBOM.
* **Measure:** Required checks green on PRs.

**NFR‑M2 — Add‑only & SP‑only.**

* **Target:** No destructive DB DDL; app obtains only **EXECUTE** on SPs; **no table rights**.
* **Measure:** Migration review; grants script in Evidence.

**NFR‑M3 — Theming pipeline.**

* **Target:** Figma Make → ThemeBuilder → Kendo Fluent v12; import base theme then overrides; parity checklist passes.
* **Measure:** Theme assets in Evidence; axe smoke passes.

**9) Accessibility (UI)**

**NFR‑A11y1 — Conformance.**

* **Target:** **WCAG 2.2 AA**, zero critical axe violations on /, /sessions, /config, /access.
* **Measure:** CI a11y smoke; manual keyboard/focus review.

**NFR‑A11y2 — Streaming announcements.**

* **Target:** ARIA live feedback for stream state (Connected / Reconnecting / Disconnected); visible focus for controls.
* **Measure:** UI tests & manual review.

**10) Portability & Deployability**

**NFR‑D1 — Environments.**

* **Target:** Reproducible deploy to Alpha/Beta/RTM/Prod via GitHub Environments and scripts; RTM uses **Prod DB (read‑only)**.
* **Measure:** Deploy workflow logs & approvals in Evidence.

**NFR‑D2 — Image hygiene & SBOM.**

* **Target:** SBOM generated per build; no known **High** vulnerabilities admitted.
* **Measure:** Dependency Review & SBOM scan results.

**11) Compliance & Audit**

**NFR‑C1 — Evidence retention.**

* **Target:** **≥ 1 year** retention; canonical structure with index.yaml and artifact hashes.
* **Measure:** Release assets present; retention policy enforced.

**NFR‑C2 — Secrets policy.**

* **Target:** All secrets in **GitHub Environments**; not present in code, DB, or evidence.
* **Measure:** Secret Scanning; reviewer checklist.

**12) Verification & Acceptance**

For each NFR above, verification occurs via one or more of:

* **Automated tests** (unit/integration/E2E/axe/perf),
* **OpenAPI governance** (lint/diff),
* **Monitoring panels & alerts** (TTFB/latency/availability/heartbeat),
* **Release Evidence Pack** attachments,
* **Runbook drills** (graceful drain, rollback, license rotation).

Promotion to **Prod** requires: CI gates green; Beta perf smoke within budgets; RTM parity ok; no **High** vulnerabilities; Evidence Pack assembled.

**13) Assumptions & Out‑of‑Scope**

* Ingress supports SSE **without buffering**; read/idle timeouts align to DB config.
* UI is **read‑only** (no writes to DB); Admin Portal uses same origin as API.
* Authentication handled by platform/gateway bearer; fine‑grained AuthZ is out of scope.
* Rate limiting is deferred (future ADR).

**14) Cross‑References**

* **FR:** docs/05\_fr.docx (transport, sessioning, config surface, legacy flag).
* **UI/UX:** docs/08\_ui\_ux.docx (Kendo Fluent v12 + ThemeBuilder, streaming UX).
* **Monitoring:** docs/11\_monitoring.docx (SLIs/SLOs, alerts, 24‑h checks).
* **Test Strategy:** docs/09\_test\_strategy.docx (axe/perf harness, coverage targets).
* **Compliance:** docs/13\_compliance.docx (CSP/egress, secrets policy).
* **Evidence Pack:** docs/12\_evidence\_pack.docx.

**15) Next Steps**

1. Wire final metrics (sse\_ttfb\_ms, sse\_heartbeat\_gap\_ms, config\_fetch\_duration\_ms) and dashboards.
2. Add Beta perf smoke to deploy workflow; attach results to Evidence Pack.
3. Schedule quarterly **graceful‑drain drills** and RTM parity checks; record results in Evidence.

**Footer (optional for Word header/footer):**  
*MCPX‑KendoBridge • Non‑Functional Requirements • v2.0.0 • 2025‑09‑27 • Confidential — Technijian Internal*