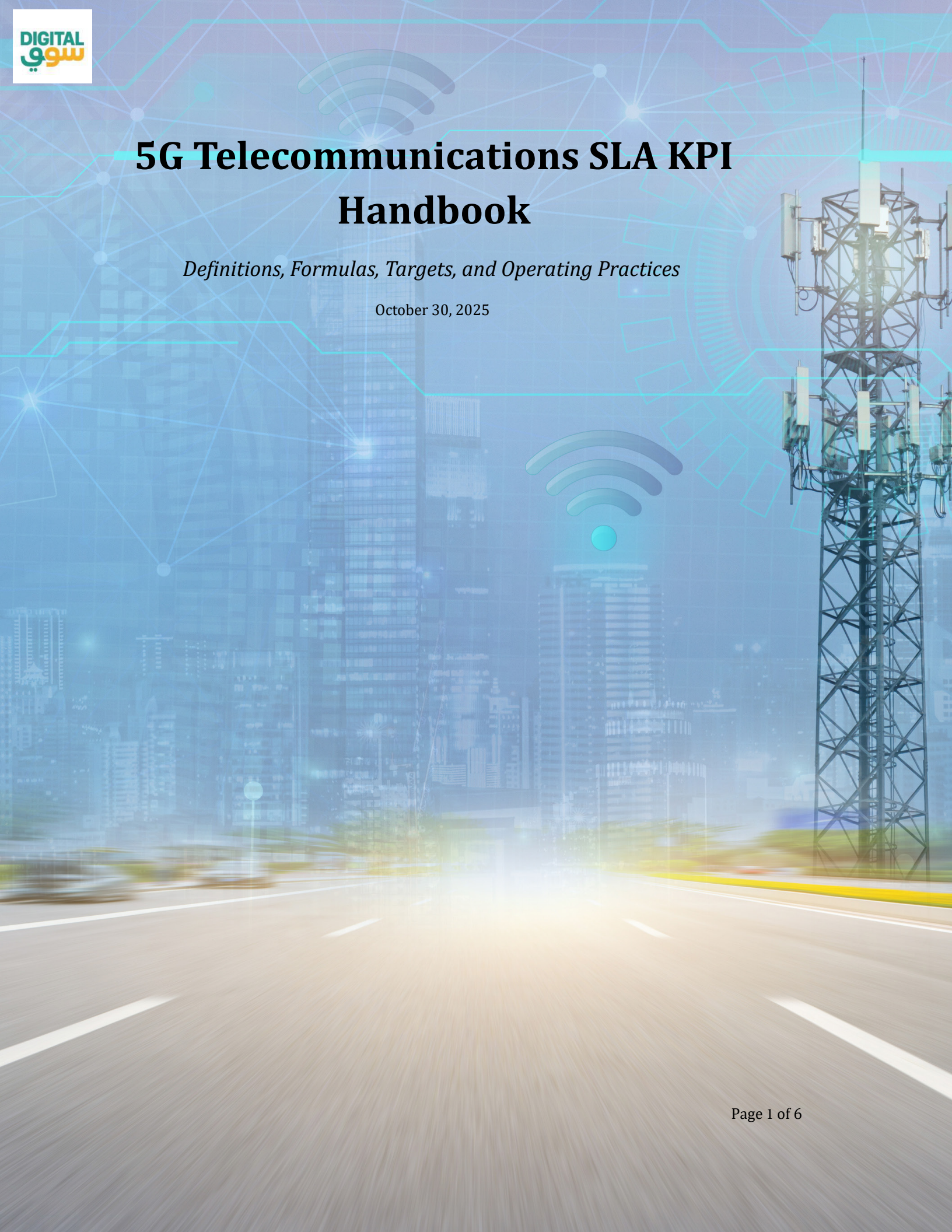

5G Telecommunications SLA KPI Handbook

Definitions, Formulas, Targets, and Operating Practices

October 30, 2025

EXECUTIVE SUMMARY

This handbook provides a practical, vendor-agnostic reference for defining, measuring, and governing Service Level Agreements (SLAs) for 5G networks. It covers end-to-end service KPIs, slice-specific objectives, how to build reliable measurement architectures, and how to turn KPI results into transparent reports and business-ready service credits.

5G introduces cloud-native cores, radio disaggregation, network slicing, and edge computing. These innovations expand what an SLA can cover, but they also complicate demarcation, measurement, and accountability. The guidance here focuses on clarity—precise KPI definitions, reproducible formulas, and methods that are auditable across RAN, transport, core, and MEC domains.

Use this document to align product, engineering, and legal stakeholders on SLA scope, targets, exclusions, and credit calculations. Treat numbers and examples as starting points; tune them to your geography, transport design, spectrum, and application mix.

5G SERVICE LANDSCAPE & SLA CONTEXT

Service types: enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC), and massive Machine-Type Communications (mMTC). Each class drives a different KPI envelope and measurement strategy.

Network slicing allows multiple logical networks—each with distinct QoS and isolation—over shared infrastructure. SLAs may be offered per slice (S-NSSAI) with per-slice telemetry and reporting.

Cloud-native 5G Core (AMF/SMF/UPF) and MEC shift workloads closer to the user, reducing latency but increasing the number of measurement points and clocks that must be synchronized.

End-to-end coverage spans UE → RAN (gNB) → fronthaul/backhaul/transport → 5G Core → Edge/App → external networks. KPI definitions MUST state the exact demarcation and time-budget per domain.

SLA FUNDAMENTALS

Key terms: SLA (contract), SLO (stated objective), KPI (measurable indicator), KQI (customer-perceived quality), OLA (internal agreement), Underpinning Contract (third-party dependency).

Scope and demarcation: spell out what is in scope (e.g., RAN+transport+core) and out of scope (e.g., customer LAN, public Internet). Include maintenance windows, force majeure, and planned work notifications.

Measurement methodology: combine passive counters, active probes (e.g., TWAMP, ICMP/TCP/HTTP synthetic transactions), and per-slice flow telemetry. Define sample sizes, cadence, percentiles (P50/P95/P99), and exclusion rules.

Credits and remedies: define thresholds, grace periods, caps, and the formula converting KPI shortfalls into monthly recurring charge (MRC) credits. Provide worked examples.

KPI TAXONOMY (END-TO-END & PER-SLICE)

- Service Availability (monthly) Page 3 of 6
- Latency—one-way and round-trip; report P50/P95/P99 within stated domain
- Jitter—P95 absolute inter-packet delay variation
- Packet Loss—bi-directional loss over active probes or flow records
- Throughput/Goodput—sustained application-layer rate
- Reliability—probability of success within a time budget (URLLC)
- Registration Success Rate; PDU Session Setup Success Rate
- Handover Success Rate (intra/inter-RAT), RRC success/failure rates
- 5QI Conformance against delay/loss budgets (GBR/non-GBR)
- RAN health: PRB utilization, BLER, scheduling latency
- Core health: AMF/SMF/UPF CPU/memory, PFCP session success, N11/N4 signaling latency
- Transport: link availability, frame delay, delay variation, loss (per CoS)
- Edge/App: container cold-start time, API latency/error rate, autoscaling time
- Security/service protection: DDoS mitigation time, false-positive rate

KPI DEFINITIONS & FORMULAS

Use unambiguous definitions with units and time windows. Prefer one-way latency with clock-sync (PTP) where feasible; otherwise, use calibrated round-trip and state assumptions. Report percentiles as Pxx values over fixed windows.

The table below summarizes standard formulas. For availability, make inclusion/exclusion rules explicit (what counts as downtime, how partial impact is treated, and whether planned maintenance is excluded).

KPI	Definition / Formula	Window
Service Availability	((Total Time − Downtime) / Total Time) × 100%	Monthly
Latency (one-way)	Median / P95 / P99 of timestamped probes (ms)	15 min / Hourly
Jitter	P95(|Δ latency|) or IQR of one-way latency (ms)	15 min / Hourly
Packet Loss	(Lost Packets ÷ Sent Packets) × 100%	15 min / Hourly
Throughput (Goodput)	App-layer bits successfully delivered ÷ time (Mb/s)	15 min / Hourly
Registration Success Rate	(Successful Registrations ÷ Attempts) × 100%	Daily / Monthly
PDU Session Setup SR	(Successful Setups ÷ Requests) × 100%	Daily / Monthly
Handover Success Rate	(Successful Handover ÷	Daily

	Attempts) × 100%	
5QI Conformance	Share of flows meeting 5QI budget (delay/loss)	Hourly / Monthly
URLLC Reliability	Pr{packet ≤ budget & no loss} over 1000 trials	Per slice / Per session

TARGETS BY SERVICE CLASS (ILLUSTRATIVE)

Targets vary by geography, spectrum, RAN design, transport topology, and workload. The values below are sane defaults for discussion—refine them through trials and customer pilots.

Class	Availability	Latency (E2E intra-metro)	Packet Loss	Jitter	Notes
eMBB	≥ 99.90% / month	≤ 30 ms P95	≤ 0.10%	≤ 10 ms	Throughput per contract
URLLC	≥ 99.999% (control path)	≤ 5–10 ms P99	≤ 10^{-5}	≤ 1 ms	Slice-specific, deterministic
mMTC	≥ 99.0% (access)	≤ 100 ms P95	≤ 1.0%	≤ 20 ms	Massive device scale
Private 5G – Industrial	≥ 99.99%	≤ 20 ms P95	≤ 0.10%	≤ 5 ms	On-prem MEC preferred

MEASUREMENT ARCHITECTURE

Deploy both passive telemetry (counters, flow logs) and active measurements (synthetic transactions) at UE, RAN, transport, core, and edge. Ensure per-slice observability—tag metrics with S-NSSAI and 5QI where applicable.

Clock synchronization is critical for one-way latency and jitter. Use PTP (with boundary/transparent clocks) or high-quality NTP. Document time sources, offset limits, and holdover behavior.

Analytics: centralize time-series in a resilient pipeline; compute percentiles correctly (TDigest/CKMS). For anomaly detection, combine rules with ML-based baselines but keep SLA evaluation deterministic and auditable.

DATA QUALITY & TIME SYNC

Validate data completeness, freshness, and clock accuracy before calculating KPIs. Track missing intervals and apply conservative rules (e.g., treat gaps as worst-case for the affected metric unless explained).

For multi-domain paths, partition latency budgets (e.g., RAN 8 ms, transport 7 ms, core+MEC 10 ms within a 25 ms target) and instrument each domain so that root cause is attributable.

REPORTING & VISUALIZATION

Publish monthly SLA reports with executive summaries, KPI tables, percentile charts, outage timelines, and root-cause summaries. Include a per-slice appendix if slices are contracted.

Provide drill-downs for customers: geographic heat maps, per-5QI performance, and device cohorts. Keep raw exports available for audit.

GOVERNANCE & COMPLIANCE

Define roles and RACI across product, NOC, engineering, and legal. Describe incident and change workflows, notification SLAs, and maintenance windows. Clarify exclusions and how exceptions are approved.

Maintain a KPI dictionary with versioning. When definitions change, run both old and new calculations in parallel for one cycle and document the impact.

SERVICE CREDITS & PENALTIES (EXAMPLE)

Credits should be significant enough to matter but capped to protect service continuity. Use a laddered schedule per KPI with a monthly cap across all metrics. The example table applies to Availability, Latency, and Packet Loss.

Example calculation: If the monthly availability target is 99.90% and actual is 99.72% (0.18 percentage points short), the credit is 10% of MRC for that service. If multiple KPIs trigger, cap total credits to the stated maximum (e.g., 20%).

Shortfall vs Target	Credit (% of MRC)	Notes
Meets or exceeds target	0%	No credit
≤ 0.10 percentage points below	5%	Applied per metric
0.10–0.30 percentage points below	10%	Cumulative within month
> 0.30 percentage points below	20% (cap)	Total monthly cap across all KPIs

ANNEX A: AVAILABILITY VS. DOWNTIME

Use this reference to translate percentage targets into the maximum allowed downtime for a 30-day month and a 365-day year.

Availability	Max Downtime / Month (30 days)	Max Downtime / Year (365 days)
99.000%	432.00 minutes	5256.00 minutes
99.500%	216.00 minutes	2628.00 minutes
99.900%	43.20 minutes	525.60 minutes
99.950%	21.60 minutes	262.80 minutes
99.990%	4.32 minutes	52.56 minutes
99.995%	2.16 minutes	26.28 minutes
99.999%	0.43 minutes	5.26 minutes

ANNEX B: GLOSSARY

- 5QI: 5G QoS Identifier defining delay/loss budgets and priority.
- AMF/SMF/UPF: Core functions for access/session management and user plane.
- KPI vs. KQI: Indicator vs. customer-perceived quality (e.g., MOS).
- NWDAF: Network Data Analytics Function for 5G analytics.
- SLA/SLO: Contract vs. objective.
- S-NSSAI: Single Network Slice Selection Assistance Information.
- URLLC/eMBB/mMTC: 5G service families.

DISCLAIMER

This handbook is a technical reference, not legal advice. Final SLA wording and targets should be adapted to your network design, regulatory environment, and customer commitments.