VA Guidelines for Establishing Non-Functional Service Level Objectives (SLO)

Introduction

This Enterprise Technology Guideline (ETG) provides the official Office of Information and Technology (OIT) capability guidance and establishes the technology standard for Service Level Objectives (SLO) in VA. The intended audience for this document series ranges from Product Line Management to Technical Subject Matter Experts (SME). These documents will be revised periodically to reflect the latest SLO-related information available to IT stakeholders and their business partners, and it will be updated to reflect pertinent VA and/or OIT policy changes.

SLO Business Case

Service Level Objectives are measurable targets that define how reliable and performant a system should be. They help organizations align technical operations with business goals. A key concept to understand is that a well-designed SLO not only indicates when a service is degraded, it also quantifies the extent of the degradation and determines whether the level of degradation is acceptable, or if action and remediation are required.

Leading Information Technology organizations such as Google, Netflix, IBM, and Microsoft have reported significant improvements in system stability, customer satisfaction, and operational efficiency by adopting SLOs. Industry best practices recommend maintaining a balanced SLO generation approach by creating only necessary, high-quality SLOs for each service, based on real customer expectations. Error budgets are used to balance innovation with reliability, and continuous measurement and refinement of SLOs are achieved through monitoring and observability.

To implement SLOs effectively, organizations should:

- ✓ Identify critical user journeys and define corresponding Service Level Indicators (SLIs).
- ✓ Use monitoring tools to track reliability metrics in real time.
- ✓ Integrate SLO reviews into governance processes and sprint ceremonies.

IT organizations have found that adopting SLOs leads to improved reliability, fewer customer-reported incidents, more data-driven prioritization of engineering efforts, and a stronger alignment between IT operations and business objectives.

SLO Document Series Overview

Enterprise SLO tech guidelines are intended to be both at-a-glance *intuitive* to non-technical product line personnel, and *sufficiently comprehensive* to technical product line SMEs. Non-Functional SLO Enterprise Technical Guidelines (NF SLO ETG) aim to provide fundamental Site Reliability Engineering (SRE) best-practice guidance and support for high quality, customer-centric, VA SLO documentation. This ETG introduces an approach that utilizes two categories of related documents:

1. The IT-Service-Definition-Checklist; and,
2. The Non-Functional-SLO-Brief and Non-Functional-SLO-Enterprise-Tech-Guidelines.

This document defines a consistent, globally applicable set of Non-Functional Service Level Objective (SLO) guidelines across all production services — including APIs, databases, Kubernetes workloads, and data pipelines. Guidelines are organized by service tier (reflecting criticality and user impact) and service type (reflecting the nature of the workload).

SLOs use a 28-day rolling window as the operational measurement period unless otherwise noted. A calendar month measurement period is used for Leadership reporting. Error budgets are derived as the complement of the SLO target.

Example: A 99.9% availability SLO yields a 0.1% error budget — approximately 40 minutes of allowable downtime per month.

Maintenance Windows and SLO Accounting

Maintenance windows follow established Change Control processes: VA Directive 6004, VA OIT Change Control Standard Operating Procedure. Maintenance window durations are not automatically excluded from SLO calculations. A maintenance window is a period of time where required maintenance may cause a disruption of service. It represents the business' agreement that if a disruption must occur, during the

maintenance window is the preferred time for service disruption. Disruption during a maintenance window counts against the SLO unless the service's ATO or equivalent governing documentation states that the service is expected to be unavailable during that window. Another exception accounts for the use case scenario where 24x7 service is not explicitly required in system or service governance documentation. For example, if the system or service is expected to be available M-F, 8 AM EST – 8 PM EST, then this baseline is utilized to determine availability calculations.

The key test: *Would it be acceptable to service consumers if the service were taken fully offline for the entire duration of every maintenance window?* If the answer is no, then consumers have an expectation of availability during that window and service disruption can count against the SLO.

Teams should review their ATO and relevant availability commitments to determine whether maintenance window exclusions are permissible. If it is not clearly stated, assume maintenance windows count in error budget burn rate calculations.

Service Classification

Tier Classification

Tiers are determined by a scoring matrix across four operational risk dimensions. This approach makes classification consistent and defensible regardless of user count or service type. For example, a critical internal auth service with 500 users should be classified identically to a high-traffic external service with equivalent risk characteristics.

Score each dimension 1 – 4 using the criteria below. Next, sum the scores to determine the service's tier.

Scoring Dimensions

Dimension	Score 1	Score 2	Score 3	Score 4
Failure Impact	If the service/system were to fail entirely (outage) or experience service degradation, there is no measurable Veteran services impact (fully redundant service).	If the service/system were to fail entirely (outage) or experience service degradation, there is low, measurable Veteran services impact.	If the service/system were to fail entirely (outage) or experience service degradation, there is measurable Veteran services impact.	If the service/system were to fail entirely (outage) or experience service degradation, there is high, measurable Veteran services impact.
Structural Criticality	Standalone service; no downstream or upstream (C-100) critical dependencies.	One downstream or upstream (C-100) critical dependency.	Two downstream or upstream (C-100) critical dependencies.	Three or more downstream or upstream (C-100) critical dependencies.
Recovery Complexity	Fully automated recovery. Restoration typically in minutes, no downstream and/or upstream C-100 systems outage or services disruptions during Disaster Recovery (DR).	Automated recovery with SME last-mile config. One downstream and/or upstream C-100 systems outage or services degradation during Disaster Recovery (DR).	Semi-automated with manual steps required; restoration typically in hours, two downstream and/or upstream C-100 systems outages or services degradation during Disaster Recovery (DR).	Complex or manual recovery. Data loss risk. Three or more downstream and/or upstream C-100 systems outages or services degradation during Disaster Recovery (DR).

Tier Assignment

Total Score	Tier	Label
4 – 6	T3	Principal
7 – 9	T2	Critical
10 – 12	T1	Bedrock

Override Rules:

Score alone may not capture every risk scenario. The following overrides may apply:

- Any service scoring 3 in Failure Impact is automatically elevated to at least T2.
- Any service that is a direct dependency of a T1 service (i.e., a T1 service cannot function without it) is automatically elevated to at least T2.

Service Type Classification

Type	Description	Examples
Request/ Response	Synchronous services responding to user or system requests	REST APIs, GraphQL endpoints, gRPC services
Storage	Persistence layers storing and retrieving data	Relational databases, object stores, caches
Pipeline	Asynchronous batch or streaming data processing	ETL jobs, Kafka consumers, data sync jobs
Platform	Infrastructure workloads hosting the above service types	Kubernetes pods, node pools, ingress controllers

Global SLO Guidelines

The following SLOs can apply to all production services regardless of type. Service Tier helps determine which of the specific target values would be appropriate for your service. Some services may not have monitoring in place to provide Service Level Indicators (SLI) that drive the data behind recommended industry-standard SLO benchmarks. This type of a use-case scenario should not discourage a Service Owner from establishing an SLO; rather, available SLIs should be utilized for nascent SLO document creation, and as observability increases for your Service, additional SLOs can be added based on SLI (observability) increases.

Availability

Availability measures the proportion of requests that are successfully served. For clustered and distributed systems, aggregate availability is the required calculation method — time-based availability (measuring contiguous downtime windows) applies only when a service is completely offline and is not appropriate for clustered services where partial degradation (e.g., 1 of 4 nodes failing) causes a proportion of requests to fail rather than total outage.

Time-Based SLI Definition: (uptime / (uptime + downtime))
Aggregate SLI Definition: (successful_requests / total_requests) × 100

For non-request-driven services: percentage of measurement intervals where the service is reachable and healthy.

Tier	SLO Target	Degradation Duration (30-day)
T1	99.9%	~43 minutes
T2	99.5%	~3.6 hours
T3	99.0%	~7.2 hours

The "Monthly Error Budget" column is expressed as a total-failure equivalent for planning purposes only. In practice, because availability uses an aggregate calculation, budget is consumed proportionally — a 50% degradation for 86 minutes consumes the same budget as a full outage for 43 minutes.

Notes:

- Time-based availability is a useful concept when discussing aggregate availability systems because it is intuitive and easy to interpret. However, it is important to recognize that time-based availability reflects system performance under average load conditions. In practice, many systems seldom operate at true average load due to seasonal or business-driven fluctuations in demand.
- Maintenance window exclusions require explicit ATO or governing documentation authorization and a Change Request in the ITSM tool of record. Please reference VA Directive 6004 and VA OIT Change Control Standard Operating Procedure for further information.
- Degraded responses (e.g., partial data, fallback behavior) must be classified explicitly as either "good" or "bad" events and documented in the SLO definition.
- Avoid using "downtime" to describe partial degradation. A clustered service is rarely "down" in the sense of 100% unavailable. Availability SLO breaches may occur due to degraded — not solely absent — service delivery.

Latency

Latency measures the time from request initiation to response delivery. Percentile-based targets are required — averages alone are insufficient as they mask tail latency.
SLI Definition: Proportion of requests completing within defined latency thresholds.

Request/Response Services (APIs)

Tier	p50	p90	p99
T1	< 100 ms	< 300 ms	< 1,000 ms
T2	< 250 ms	< 750 ms	< 2,000 ms
T3	< 500 ms	< 2,000 ms	< 5,000 ms

Storage Services (Databases)

Tier	Read p90	Write p90	Read p99	Write p99
T1	< 20 ms	< 50 ms	< 100 ms	< 200 ms
T2	< 50 ms	< 100 ms	< 250 ms	< 500 ms
T3	< 200 ms	< 500 ms	< 1,000 ms	< 2,000 ms

Notes:

- Latency is measured at the service boundary closest to the end user or consumer, not at the infrastructure layer. For example, a web server using an Application Load Balancer or other Layer 7 proxy should measure latency (and availability and errors) at the customer facing interface of the load balancer.
- Latency SLOs apply to read and write paths independently for storage services.
- Background/administrative queries (e.g., vacuum, garbage collection, reindex) are excluded from SLI calculations.

Error Rate

Error rate measures the proportion of requests that result in an error, timeout, or invalid response. Note that service-impacting errors are already accounted for in the Availability SLO. Error rate as a standalone SLO is most valuable for isolating error signal from overall availability — particularly for services where partial errors do not fully degrade availability but signal a reliability risk. Critically, errors need to be relevant to the operation of the service. Errors that are not indicative of service health are not the kinds of error messages that should be utilized to calculate Error Rate.

SLI Definition: (error_responses / total_requests) × 100

Tier	Max Error Rate
T1	< 0.1%
T2	< 1.0%
T3	< 5.0%

Error classifications:

- ✓ 5xx responses (server errors): In general, always counted as errors.
- ✓ 4xx responses (client errors): Generally excluded from SLO. Include only when the service is explicitly responsible for input validation at a boundary (e.g., an API gateway performing authentication), and this is documented in the service's SLO definition.
- ✓ Timeouts: Always counted as errors, using the service's defined timeout threshold.
- ✓ Partial failures: Must be explicitly defined per service — document whether degraded responses count as errors, fail open and safe.

Notes:

- If there is a reason where a given 5xx error in a specific scenario (perhaps on an API) is not considered a failure, teams should adjust accordingly.

Throughput

Throughput defines the minimum request handling capacity a service must sustain without SLO degradation.

SLI Definition: Requests per second (RPS) or transactions per second (TPS) sustained without availability or latency SLO breach.

Tier	Minimum Sustained Throughput
T1	Service team must define a baseline from p95 peak traffic × 1.5× headroom
T2	Service team must define a baseline from p95 peak traffic × 1.25× headroom
T3	No formal throughput SLO required; document expected load

Notes:

- Throughput targets must be re-evaluated after any architecture change or when observed peak traffic exceeds 80% of the defined target. Throughput headroom (spare capacity above observed peak) is a Saturation signal per the Four Golden Signals framework (Section 8.1). It is a leading indicator that informs Availability and Latency SLOs but does not require its own SLO unless customer commitments exist.

Mean Time to Detect (MTTD)

MTTD measures the time elapsed between when an SLO-violating condition begins and when an alert reaches the on-call responder, human or AI, that is capable of taking action to mitigate the alerting condition. MTTD must be calibrated relative to the service's error budget — a long MTTD on a service with a small budget can consume most of that budget before the team is even aware of an incident.

SLI Definition: Time from the first SLO-violating data point to first alert notification delivered to on-call (not from when the probe runs or the alert fires in the monitoring system — from when the responder is notified).

MTTD targets should be set as a fraction of the tier's monthly error budget, ensuring that detection alone cannot consume more than a defined proportion of the budget:

Tier	MTTD Target
T1	< 5 minutes
T2	< 15 minutes
T3	< 60 minutes

Notes:

- MTTD is measured from the start of the violating condition, not from when the alert fires.
- Probe interval is not MTTD. A synthetic probe running every 10 minutes has a *worst-case* detection time of approximately 10 minutes (failure occurring just after a probe) and an *expected* MTTD of ~5 minutes. For planning, assume worst-case detection = probe interval.
- The full detection-to-notification chain (probe → monitoring platform → event management → alerting system → on-call notification) must be measured end-to-end. Each stage adds latency. MTTD clocks until the on-call responder receives the notification, not until the monitoring platform detects the probe failure.
- Synthetic probe intervals must be set such that the worst-case probe detection time (= probe interval) leaves adequate time for the full notification chain to complete within the MTTD target.
- Alert burn rate thresholds must be calibrated so that a fast-burning incident triggers within the MTTD window.
- Applies to all service types.

Mean Time to Recovery (MTTR)

MTTR measures the average time from the start of a user-impacting incident to full service restoration within SLO targets. MTTD + MTTR represents total error budget consumed per incident.

SLI Definition: Time from incident declaration to SLO metrics returning within target thresholds.

Tier	MTTR Target
T1	< 1 hour
T2	< 4 hours
T3	< 24 hours

Notes:

- MTTR is tracked as a rolling 30-day average across all incidents, not per-incident.
- Applies to all service types.
- Runbooks must exist for all T1 services, Absence of a runbook is itself an SLO risk.
- Teams should track MTTD and MTTR trends monthly alongside error budget burn rate.

Service Type Specific SLO Requirements

Storage Services (Databases)

In addition to the global requirements above, all production databases must meet or track the following. Note that several storage metrics below are internal operational metrics -supporting risk awareness and Recovery Point Objectives (RPO)/ Recovery Time Objectives (RTO) planning, rather than customer-centric SLOs.

Replication Lag (where applicable)

Replication lag is a consumer-facing SLO only when service consumers directly access the replica (e.g., a reporting database, a read replica serving application reads). If the replica's sole purpose is backup, failover, or DR, replication lag is an internal operational metric that informs RPO — not a consumer SLO.

When consumers access the replica, RPO strongly informs the SLO target.

Tier	Max Replication Lag
T1	< 1 second (p99)
T2	< 10 seconds (p99)

T3	< 60 seconds (p99)

Connection Pool Saturation (internal operational metric)

Connection pool saturation is a leading indicator and risk assessment metric, not a consumer SLO. Breaching connection pool thresholds does not by itself indicate consumer impact — but it predicts imminent Availability and Latency SLO risk.

Tier	Max Connection Pool Utilization
T1	< 70% of max connections at p95
T2	< 80% of max connections at p95
T3	< 90% of max connections at p95

Recovery Point Objective (RPO)

RPO defines the maximum acceptable data loss expressed as a time window. RPO is a disaster recovery planning constraint (not a consumer SLO), but it informs and bounds storage SLO definitions — particularly replication and backup requirements.

SLI Definition (planning input): Age of the most recent successful backup or replication checkpoint at time of failure.

Tier	RPO Target	Required Backup Strategy
T1	< 1 hour	Continuous replication (WAL shipping / binlog) or ≤ 1-hour incremental backups; PITR required
T2	< 4 hours	≤ 4-hour incremental backups; PITR recommended
T3	< 24 hours	Daily backups acceptable

Notes:

- RPO must be validated through restore testing, not just backup job success monitoring.
- For T1 databases with continuous replication, RPO approaches near-zero — this is the expected standard.
- Does not apply to: Request/Response services (stateless) or Platform services (inherit from hosted storage).

Durability

Durability measures the probability that written data is not lost.

Tier	Durability Target
T1	99.999% (5 nines)
T2	99.99% (4 nines)
T3	99.9% (3 nines)

Pipeline Services (Batch / Streaming)

Pipelines use different SLI categories from request-driven services.

Freshness

Freshness is a consumer-facing SLO when there is an explicit commitment to service consumers about how current the data will be. When no such commitment exists, freshness should still be monitored as an operational signal. VA has many systems with replication (e.g., VistA and EHRM services) where freshness commitments are likely appropriate.

SLI Definition: Proportion of measurement intervals where output data age is within the threshold.

Tier	Max Acceptable Data Age
T1	< 5 minutes
T2	< 30 minutes
T3	< 4 hours

Correctness

Correctness measures the proportion of records processed without data loss or corruption.

Tier	Min Correctness Rate
T1	**99.99%** of records processed correctly

T2	**99.9%** of records processed correctly
T3	**99.0%** of records processed correctly

Coverage

Coverage measures the proportion of expected input records that were successfully processed.

Tier	Min Coverage Rate
T1	**99.9%** of expected records processed
T2	**99.0%** of expected records processed
T3	**95.0%** of expected records processed

Consumer Lag (queue-backed pipelines only)

Consumer lag measures how far behind a message consumer is from the head of its queue or topic partition. High lag is a leading indicator of an imminent Freshness SLO breach and, if queue retention expires, data loss.

SLI Definition: Number of unprocessed messages behind the queue head, or equivalent time-based lag.

Tier	Max Consumer Lag
T1	< 1,000 messages or < 5 minutes of lag (whichever is more restrictive)
T2	< 10,000 messages or < 30 minutes of lag
T3	< 100,000 messages or < 4 hours of lag

Notes:

- Time-based lag: now - timestamp_of_oldest_unconsumed_message
- Consumer lag targets align with Freshness targets — lag exceeding the Freshness threshold means the Freshness SLO is already breached.
- Does not apply to: Request/Response, Storage, or Platform services.

Queue Age (queue-backed pipelines only)

Queue age measures the age of the oldest unprocessed message currently in the queue. While consumer lag counts messages, queue age reveals how long the oldest work item has been waiting — critical for detecting stalled consumers even when message count appears low.

SLI Definition: now - timestamp_of_oldest_unconsumed_message

Tier	Max Queue Age
T1	< 5 minutes
T2	< 30 minutes
T3	< 4 hours

Dead Letter Queue (DLQ) Rate (queue-backed pipelines only)

DLQ rate measures the proportion of messages routed to a dead letter queue after exhausting all retry attempts.

SLI Definition: (messages_routed_to_DLQ / total_messages_attempted) × 100

Tier	Max DLQ Rate	DLQ Review SLA
T1	< 0.01%	Alert on any DLQ entry; review within 1 hour
T2	< 0.1%	Review within 4 hours
T3	< 1.0%	Review within 24 hours

Notes:

- T1 pipelines must alert on the first DLQ entry — a single poison-pill message silently blocking processing is unacceptable.
- DLQ growth rate (not just total count) should be tracked; a sudden spike indicates a systematic failure distinct from isolated bad messages.
- Does not apply to: Request/Response, Storage, or Platform services.

Recovery Point Objective (RPO) (pipelines)

For pipelines, RPO defines the maximum acceptable window of events that must be replayed following a pipeline failure. This is a DR planning constraint that informs pipeline design, not a consumer SLO.

Tier	RPO Target	Required Capability
T1	< 1 hour	Checkpointing or offset management with ≤ 1-hour granularity; replay capability required
T2	< 4 hours	Checkpointing with ≤ 4-hour granularity
T3	< 24 hours	Daily checkpointing acceptable

Platform Services (Kubernetes)

Kubernetes workloads hosting T1 or T2 services must meet platform-level SLOs to support the application SLOs above.

Pod Availability

Tier	Min Healthy Pod Ratio
T1	≥ 99.9% of replicas healthy at any given time
T2	≥ 99.0% of replicas healthy at any given time
T3	≥ 95.0% of replicas healthy at any given time

Restart Rate

Excessive pod restarts indicate instability and are tracked as a reliability signal.

Tier	Max Restart Rate
T1	< 1 restart per pod per 24 hours (p99 across fleet)
T2	< 3 restarts per pod per 24 hours (p99 across fleet)
T3	No formal target; restarts must be investigated if they cause availability SLO breach

Resource Headroom (internal operational metric)

Resource headroom is a leading indicator (Saturation signal per the Four Golden Signals). Breaching these thresholds triggers scaling or capacity review — it does not by itself constitute a consumer SLO breach.

Tier	Max CPU Utilization	Max Memory Utilization
T1	< 60% of request limit (p95)	< 70% of request limit (p95)

T2	< 70% of request limit (p95)	< 80% of request limit (p95)
T3	< 80% of request limit (p95)	< 85% of request limit (p95)

Notes:

- Workloads must define resource requests and limits; pods without these set do not qualify for T1 classification.

Vertical Auto-Scaling (Operational Guidance — Not an SLO)

Vertical scaling configurations must be driven by the resource headroom thresholds in Section 3.3.3. From an SLO perspective, scaling policy itself is not an SLO — only the consumer-visible outcome matters. These formulas provide consistent, SLO-derived scaling targets so that the scaling policy and the SLO remain in sync. If the service remains within its Availability and Latency SLOs, the specific CPU/memory allocation is an internal operational concern.

Right-Sizing Formula

target_cpu_limit = observed_p95_cpu ÷ tier_cpu_ceiling
target_memory_limit = observed_p99_memory ÷ tier_memory_ceiling

Example (T1 service, CPU ceiling = 60%):

Observed p95 CPU	Tier Ceiling	Calculated Limit	Resulting Utilization
300m	60%	300m ÷ 0.60 = 500m	300/500 = 60% ✓
600m	60%	600m ÷ 0.60 = 1,000m	600/1000 = 60% ✓
150m	60%	150m ÷ 0.60 = 250m	150/250 = 60% ✓

Scale-Up/Scale-Down Triggers

Scale-up and scale-down is triggered when utilization breaches the ceiling or floor thresholds and remains there for a sustained observation window, confirming it is not a transient spike:

Tier	CPU Ceiling	Memory Ceiling	Observation Window	CPU Floor	Memory Floor	Cooldown Period

T1	60%	70%	5 minutes	30%	35%	24 hours
T2	70%	80%	10 minutes	35%	40%	12 hours
T3	80%	85%	15 minutes	40%	42%	6 hours

Step Size Limits

To prevent runaway allocation or aggressive de-provisioning from a single scaling event:

Direction	Max Change Per Step
Scale-up	+100% of current limit (one doubling per event)
Scale-down	−25% of current limit

If the right-sizing formula produces a target more than 2× the current limit, apply the change in sequential steps with one observation window between each.

Minimum Resource Floors

No workload should be right-sized below these minimums regardless of observed usage:

Resource	Minimum
CPU request	100m (0.1 cores)
Memory request	128Mi
Memory limit	256Mi

Security-Related Metrics

The metrics in this section are security and compliance operational metrics. They are not consumer-facing SLOs in the traditional sense — certificate expiry and patching cadence do not directly measure service delivery quality. However, they represent commitments and obligations (some contractual via SLA, some regulatory via ATO/FedRAMP) that must be tracked and enforced.

TLS Certificate Expiry (operational/SLA metric)

Certificate expiry causes immediate and total availability failure. Pre-expiry monitoring is a preventive operational obligation. If an SLA explicitly commits to certificate renewal timelines, this becomes a contractual metric.

Tier	Target	Renewal Method
T1	100% of certs renewed ≥ 30 days before expiry	Automated renewal required (Venafi, cert-manager, AWS ACM, or equivalent); manual renewal is prohibited
T2	100% of certs renewed ≥ 30 days before expiry	Automated or manual with ≥ 60-day calendar alert
T3	100% of certs renewed ≥ 14 days before expiry	Manual acceptable with ≥ 30-day alert

Alert thresholds (all tiers): Warning at 60 days remaining; Critical at 30 days remaining.

Vulnerability Patching Cadence (compliance/ATO metric)

Patching timelines are an ATO and compliance obligation, not a consumer SLO. Unpatched vulnerabilities may eventually cause availability or security events and those impacts can be captured by Availability and other SLOs when they occur.

CVSS Score	Severity	T1 Max Days	T2 Max Days	T3 Max Days
9.0 – 10.0	Critical	15	30	30
7.0 – 8.9	High	30	60	90
4.0 – 6.9	Medium	60	90	180
< 4.0	Low	Best effort	Best effort	Best effort

Notes:

- Patch window begins from the earlier of: NVD publication date or vendor advisory date.
- "Remediation" includes: patch applied, dependency updated, compensating control deployed and documented, or formal risk acceptance signed by the service security owner.
- Automated dependency and container image scanning is required for T1 and T2 services.

Authentication Anomaly Rate (Request/Response services only)

Auth anomaly rate detects abnormal spikes in authentication failures that may indicate credential stuffing, an authentication dependency outage, or misconfiguration.

Tier	Alert Threshold	Observation Window	Response
T1	> 3× 7-day baseline auth failure rate	5 minutes	P1 — immediate investigation
T2	> 5× 7-day baseline auth failure rate	15 minutes	P2 — investigation within 1 hour
T3	> 10× 7-day baseline auth failure rate	30 minutes	Ticket — investigation within 24 hours

SLI Definition: Alert when the auth failure rate exceeds a multiple of the 7-day baseline, sustained over the observation window.

Observability Metrics

Observability metrics ensure the telemetry infrastructure supporting SLO measurement is itself reliable. Note that some metrics in this section are operational health metrics rather than consumer-facing SLOs.

Metric Ingestion Lag (SLO — directly enables MTTD)

Metric ingestion lag is a legitimate SLO-class metric because stale metrics directly inflate MTTD past its target — they degrade the reliability of the monitoring system that all other SLOs depend on.

SLI Definition: Time between a metric being recorded at the source and being queryable in the monitoring platform (p99).

Note: Two distinct problems can appear as ingestion lag: (1) actual delay from event to ingestion; (2) time synchronization errors — system clocks out of sync or logging tools misinterpreting time zones can cause events to appear hours in the future or past. Both

must be monitored. Clock drift and time zone issues are not ingestion lag but may be misreported as such.

Tier	Max Ingestion Lag
T1	< 60 seconds
T2	< 5 minutes
T3	< 15 minutes

Synthetic Probe Success Rate (operational/observability metric — not a consumer SLO)

Synthetic probes continuously test service endpoints from outside the service's network boundary. Probe success rate is not a consumer-facing SLO — probe failures are captured by the Availability SLO (all synthetic probes feed into Availability measurement). It is, however, a valuable early-warning and confirmation signal and should be monitored.

SLI Definition (internal metric): Percentage of synthetic probe executions receiving a valid response within the service's latency SLO.

Tier	Min Success Rate	Probe Interval	Required
T1	≥ 99.9%	Every 60 seconds	Required
T2	≥ 99.5%	Every 5 minutes	Recommended
T3	—	—	Not required

Log Ingestion Completeness (compliance/observability metric)

Log ingestion completeness measures whether operational and security-relevant logs are arriving at the centralized logging system within the required window. Missing logs may constitute a compliance violation under FedRAMP or HIPAA audit requirements.

SLI Definition: Percentage of expected log volume (based on 7-day trailing baseline) received within the ingestion window.

Tier	Min Completeness	Max Ingestion Lag	Retention Period
T1	≥ 99.9% of expected volume	within 5 minutes	≥ 1 year (FedRAMP requirement)
T2	≥ 99.5% of expected volume	within 15 minutes	≥ 90 days
T3	≥ 99.0% of expected volume	within 60 minutes	≥ 30 days

Error Budget Policy

Error budgets translate SLO targets into actionable operational guidance. Error budgets are calculated on a rolling 28-day window (aligned with standard SRE practice and Google's recommended error budget period) unless the service's SLO definition explicitly specifies otherwise. Leadership reporting uses a calendar window.

Condition	Required Action
Error budget > 50% remaining	Normal development velocity; feature work and deployments proceed as planned
Error budget 25–50% remaining	Elevated caution; increased deployment review; consider freeze on risky changes
Error budget 10–25% remaining	Reliability work is prioritized over new features; all deployments require SRE review
Error budget < 10% remaining	Reliability freeze: No new feature deployments; all capacity directed at SLO recovery
Error budget exhausted	Formal incident review required; return to normal velocity only after root cause resolved and remediation plan approved

Measurement and Reporting Standards

1. Measurement window: Rolling 28-day window for all SLOs (see Section 4).
2. Instrumentation proximity: SLIs must be measured as close to the user or consumer as possible (e.g., load balancer or ingress for APIs; application query layer for databases — not OS-level disk metrics).
3. Maintenance window exclusions: Planned maintenance windows are not automatically excluded from SLI calculation. Exclusions require explicit authorization in the service's

ATO or equivalent governing documentation. When uncertain, count maintenance windows. See Overview for full guidance.
4. Review cadence: SLO attainment is reviewed monthly by service owners. Annual review of targets is required to tighten or adjust based on observed performance and user feedback.
5. SLA headroom: External SLAs (contractual commitments) must be set at least one tier below internal SLO targets to preserve error budget buffer.

SLI Applicability by Service Type

Not every SLI category applies to every service type. This table defines which SLIs are applicable, required, or not relevant — and notes which are consumer SLOs vs. internal operational metrics.

Global SLOs

SLI Category	Request/Response	Storage	Pipeline	Platform	Notes
Availability	X	X	X	X	Universal — applies to all service types
Latency	X	X	—	—	Pipelines are asynchronous; latency is not user-facing
Error Rate	X	X	Required*	—	*Pipeline error rate = % of records that fail processing
Throughput	X	X	X	—	Unit varies: RPS for APIs, TPS for DBs, records/sec for pipelines
MTTD	X	X	X	X	Universal — every service must have alerting meeting MTTD targets
MTTR	X	X	X	X	Universal — tracked as 30-day rolling average across incidents
RTO	X	X	X	X	Every service must validate recovery time against tier target
Change Failure Rate	X	X	X	X	Universal — tracks deployment reliability

					across all change types

Storage-Specific

SLI Category	Request/ Response	Storage	Pipeline	Platform	Notes
Durability	—	X	—	—	Only meaningful for persistence layers
RPO	—	X	Required†	—	†Pipeline RPO = max replayable event window
Replication Lag	—	Conditional‡	—	—	‡Required only for databases with read replicas
Connection Pool Saturation	—	X	—	—	

Pipeline-Specific

SLI Category	Request/Response	Storage	Pipeline	Platform	Notes
Data Currency	—	Required§	Required¶	—	§Replication Lag for DBs / Cache Staleness for caches; ¶Freshness for pipelines
Correctness	—	—	X	—	Record-level accuracy; not applicable to synchronous services
Coverage	—	—	X	—	% of expected records processed
Consumer Lag	—	—	Conditional*	—	*Required for queue-backed pipelines only
Queue Age	—	—	Conditional*	—	*Required for queue-backed pipelines only
DLQ Rate	—	—	Conditional*	—	*Required for queue-backed pipelines only

Platform-Specific

SLI Category	Request/Response	Storage	Pipeline	Platform	Notes
Pod Health	—	—	—	X	% of healthy replicas
Restart Rate	—	—	—	X	
Resource Headroom	—	—	—	X	CPU/memory margin; leading indicator
Vertical Auto-Scaling	—	—	—	X	Formula derived from resource headroom SLOs

Security/Compliance Metrics

SLI Category	Request/ Response	Storage	Pipeline	Platform	Notes
TLS Cert Expiry	x	Conditional*	Conditional*	x	*Required if TLS in transit; N/A if mTLS fully managed by service mesh
Vulnerability Patching	x	x	x	x	Universal — all services have a dependency tree
Auth Anomaly Rate	x	—	—	—	Auth signals only meaningful at the request-serving boundary

Observability Metrics

SLI Category	Request/Response	Storage	Pipeline	Platform	Notes
Metric Ingestion Lag	x	x	x	x	Universal — all services must have current metrics
Synthetic Probe Success	Required (T1) / Rec (T2)	—	—	—	Not externally probeable for Storage/Pipeline/Platform
Log Completeness	x	x	x	x	Universal — all services produce logs that must be retained

SLO Applicability Matrix

The matrix below shows the applicable consumer-facing SLOs for each service type, organized by tier. Operational metrics are tracked separately from SLOs.

Score key: Impact · Criticality · Data Classification · Recovery (each 1 – 3; total 4 – 12)

Tier 1 — Critical (Score 10 – 12)

Service SLOs (customer-centric)

Service	Type	Score	Availability	Latency (p99)	Error Rate	Durability	Data Currency
High-transaction DB (veteran records)	Storage	3+3+3+2 = 11	99.9%	Read < 100ms / Write < 200ms	< 0.1% query errors	99.999% †	Replication lag < 1s (if replica-served)
Public-facing APIs	Request/Response	3+3+3+1 = 10	99.9%	< 1,000 ms	< 0.1%	—	—
K8s pods (T1 workloads)	Platform	(inherits)	≥ 99.9% healthy	—	—	—	—
Critical data pipeline	Pipeline	3+2+3+2 = 10	99.9%	—	< 0.01 % recor	—	Freshness < 5 min ¶

					d failur es		

Freshness SLO applies only when consumer freshness commitment exists.

Storage: RPO < 1h (DR constraint) · Connection pool < 70% (internal metric)
K8s: < 1 restart/pod/24h · CPU < 60% (internal leading indicator) · Memory < 70% (internal leading indicator)
Pipeline: Correctness ≥ 99.99% · Coverage ≥ 99.9% · Consumer lag < 1,000 msg / 5 min · Queue age < 5 min · DLQ rate < 0.01% · Pipeline RPO < 1h

Operational Metrics (T1)

Service	MTTD	MTTR	RTO	CFR	Cert Expiry	Vuln Patch (Critical)	Auth Anomaly	Metric Lag
High-transaction DB	< 5 min	< 1 hr	< 1 hr	< 5% (internal)	Auto-renew ≥ 30d	≤ 15 days	—	< 60 sec
Public-facing APIs	< 5 min	< 1 hr	< 1 hr	< 5% (internal)	Auto-renew ≥ 30d	≤ 15 days	> 3× baseline / 5 min **	< 60 sec
K8s pods (T1)	< 5 min	< 1 hr	< 1 hr	< 5% (internal)	Auto-renew ≥ 30d	≤ 15 days	—	< 60 sec
Critical pipeline	< 5 min	< 1 hr	< 1 hr	< 5% (internal)	≥ 30d	≤ 15 days	—	< 60 sec

Auth Anomaly column retained pending [REQUIRES MANUAL INTERVENTION] resolution from Section 3.5.3

Tier 2 — Standard (Score 7 – 9)

Service SLOs (customer-centric)

Service	Type	Score	Availability	Latency (p99)	Error Rate	Durability	Data Currency
Internal employee DB	Storage	2+2+2+2 = 8	99.5%	Read < 250ms / Write < 500ms	< 1.0% query errors	99.99% ‡	Replication lag < 10s (if replica-served)
Internal / employee-facing APIs	Request/Response	2+2+2+1 = 7	99.5%	< 2,000 ms	< 1.0%	—	—
K8s pods (T2 workloads)	Platform	(inherits)	≥ 99.0% healthy	—	—	—	—
Standard data pipeline	Pipeline	2+1+2+2 = 7	99.5%	—	< 0.1% record failures	—	Freshness < 30 min ¶

Storage: RPO < 4h (DR constraint) · Connection pool < 80% (internal metric)
K8s: < 3 restarts/pod/24h · CPU < 70% (internal) · Memory < 80% (internal)

Pipeline: Correctness ≥ 99.9% · Coverage ≥ 99.0% · Consumer lag < 10,000 msg / 30 min · Queue age < 30 min · DLQ rate < 0.1% · Pipeline RPO < 4h

Operational Metrics (T2)

Service	MTTD	MTTR	RTO	CFR	Cert Expiry	Vuln Patch (Critical)	Auth Anomaly	Metric Lag
Internal employee DB	< 15 min	< 4 hr	< 4 hr	< 10% (internal)	≥ 30d	≤ 30 days	—	< 5 min
Internal APIs	< 15 min	< 4 hr	< 4 hr	< 10% (internal)	≥ 30d	≤ 30 days	> 5× baseline / 15 min **	< 5 min
K8s pods (T2)	< 15 min	< 4 hr	< 4 hr	< 10% (internal)	≥ 30d	≤ 30 days	—	< 5 min
Standard pipeline	< 15 min	< 4 hr	< 4 hr	< 10% (internal)	≥ 30d	≤ 30 days	—	< 5 min

Tier 3 — Non-Critical (Score 4 – 6)

Service SLOs (customer-centric)

Service	Type	Score	Availability	Latency (p99)	Error Rate	Durability	Data Currency
Analytics / reporting DB	Storage	1+1+1+2 = 5	99.0%	Read < 1,000 ms / Write <	< 5.0% query errors	99.9% †	Replication lag < 60s (if replica-served)

				2,000 ms			
Internal tooling / staging APIs	Request/Response	1+1+1+1 = 4	99.0%	< 5,000 ms	< 5.0%	—	—
K8s pods (T3 workloads)	Platform	(inherits)	≥ 95.0% healthy	—	—	—	—
Reporting / ETL pipeline	Pipeline	1+1+1+2 = 5	99.0%	—	< 5.0% record failures	—	Freshness < 4 hrs ¶

Storage: RPO < 24h (DR constraint) · Connection pool < 90% (internal metric)
K8s: No formal restart target (investigate if causing breach) · CPU < 80% (internal) · Memory < 85% (internal)
Pipeline: Correctness ≥ 99.0% · Coverage ≥ 95.0% · Consumer lag < 100,000 msg / 4 hr · Queue age < 4 hr · DLQ rate < 1.0% · Pipeline RPO < 24h

Operational Metrics (T3)

Service	**MTTD**	**MTTR**	**RTO**	**CFR**	**Cert Expiry**	**Vuln Patch (Critical)**	**Auth Anomaly**	**Metric Lag**
Analytics DB	< 60 min	< 24 hr	< 24 hr	< 15% (internal)	≥ 14d	≤ 30 days	—	< 15 min

Internal tooling APIs	< 60 min	< 24 hr	< 24 hr	< 15% (internal)	≥ 14d	≤ 30 days	> 10× baseline / 30 min **	< 15 min
K8s pods (T3)	< 60 min	< 24 hr	< 24 hr	< 15% (internal)	≥ 14d	≤ 30 days	—	< 15 min
Reporting pipeline	< 60 min	< 24 hr	< 24 hr	< 15% (internal)	≥ 14d	≤ 30 days	—	< 15 min

Measurement of all four signals is required for T1 services and recommended for T2. Traffic and Saturation function as leading indicators — they don't directly define the SLO but determine when Latency and Error SLOs are at risk.

Applying the Framework to Services and Subsystems

The Four Golden Signals

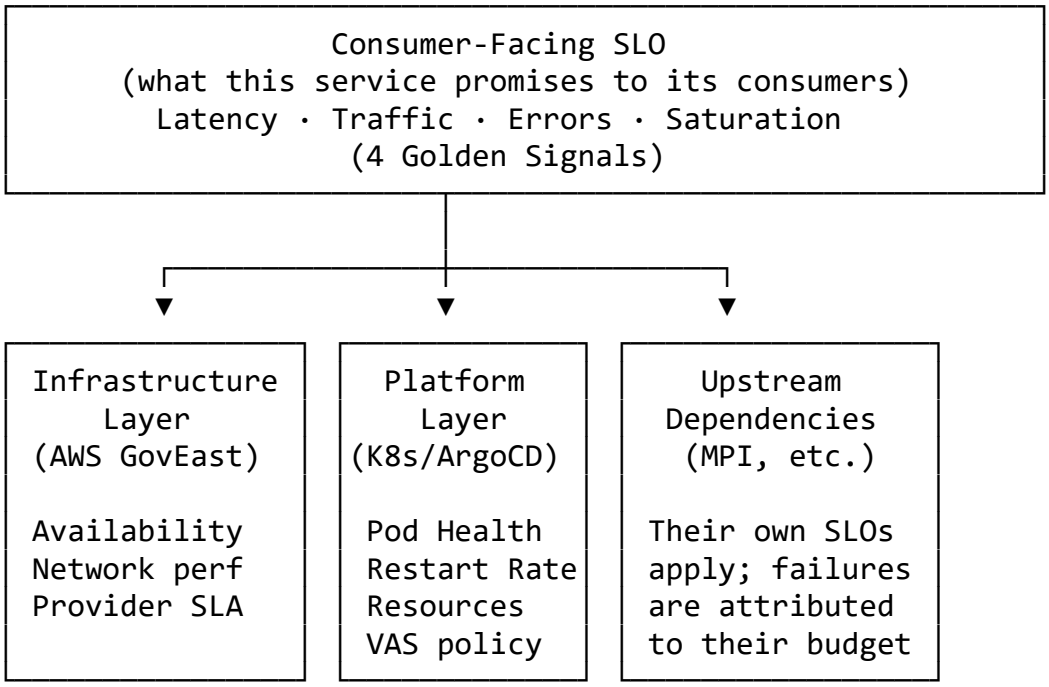
Google SRE's four golden signals provide the observability foundation every service must have. All SLI categories in this framework are derived from them. When defining SLOs for any service, applicable signals should have at least one SLI/SLO pair.

Golden Signal	Definition	SLI Categories in This Framework
Latency	Time to service a request — measured for both successful and failed requests	Page 10
Traffic	Volume of demand placed on the system (RPS, TPS, records/sec)	Page 12
Errors	Rate of requests that fail — explicitly, implicitly, or by policy	Page 11
Saturation	How "full" the service is; utilization approaching capacity limits	Page 17

Measurement of all four signals is required for T1 services and recommended for T2. Traffic and Saturation function as leading indicators — they don't directly define the SLO but determine when Latency and Error SLOs are at risk.

Service SLO Hierarchy

Every production service is composed of layers. Each layer has its own SLO responsibility, and the consumer-facing SLO at the top is only achievable if each underlying layer meets its targets.



Each layer is classified and measured independently. A platform layer hosting a T1 service is itself classified as T1 for the purposes of resource headroom, restart targets, and MTTD.

Composite Availability and Dependency Attribution

When a service depends on upstream services, its theoretical maximum availability is bounded by those dependencies:

$$\text{Composite\_Availability} \leq \text{Own\_SLO} \times \text{Dep}_1\text{\_SLO} \times \text{Dep}_2\text{\_SLO} \times \ldots$$

Example: VA.gov API (own SLO: 99.9%) depends on MPI (SLO: 99.9%) and VA Profile (SLO: 99.9%):

Composite_Availability ≤ 99.9% × 99.9% × 99.9% ≈ 99.7%

This means VA.gov API cannot achieve 99.9% if its dependencies fail — unless it implements resilience patterns (circuit breakers, cached fallbacks, graceful degradation, failing open and safe, etc., et. Al.) that isolate it from upstream failures.

Dependency Attribution Policy

Upstream failures must not consume the downstream service's own error budget. All services must:

1. Declare dependencies — all upstream services are listed in the service SLO definition.
2. Tag failures at resolution — incidents caused by upstream failures are attributed to the upstream service's error budget.
3. Report budgets separately — each service's monthly report distinguishes: (a) budget consumed by own failures, (b) budget consumed by attributed dependency failures.

Dependency Budget Reservation

To account for expected upstream reliability, reserve a portion of the own error budget proportional to dependency risk:

Reserved_Budget = Σ (upstream_error_budget × risk_factor)

Upstream Tier	Risk Factor
T1	0.5 — well-managed, low failure rate
T2	0.75 — moderate failure rate
T3	1.0 — assume full budget may be consumed

If the total reserved budget approaches or exceeds the own error budget, the service must implement resilience patterns to reduce dependency exposure — not simply accept a lower effective SLO.

Scaling to the C-100 Services List

Each service can maintain a standalone SLO definition document using the companion service-slo-template.md. Together, the 100 service definitions form a service reliability inventory with the following properties:

- Every service has a justified tier (scored, not assumed).
- Every service declares its upstream dependencies, enabling a full dependency graph.
- No T1 service may depend on an unclassified or T3 service without a formal risk acceptance record.
- The composite availability of any T1 service and its full dependency chain is calculated and reviewed quarterly.

Glossary

API – Application Programming Interface. An API is a set of rules and protocols that enable different software applications to communicate and interact with each other. It acts as a bridge between systems, allowing them to exchange data or perform actions without needing to understand each other's internal workings.

DR – Disaster Recovery. Disaster recovery is the way in which you resume regular operations after a disaster. This is typically accomplished through the resumption of essential activities and the processes and systems used to support them. Disaster recovery must go according to a disaster recovery plan, which is a detailed, documented set of procedures designed to minimize the amount of time it takes for the organization to recover. Disaster Recovery focuses on the IT infrastructure and systems needed by the organization to resume operation after an interruption occurs.

PITR - Point-in-Time Recovery is a data recovery mechanism that allows a system to restore to a specific point in time, typically by replaying transaction logs or write-ahead logs. It is crucial for recovering from data corruption, accidental deletions, or system failures. PITR is implemented by combining periodic backups with continuous change records, enabling granular recovery options. It is particularly useful in cloud computing and database management, allowing organizations to recover data to any specific point in time, minimizing data loss in case of unexpected events.

SLI - Service Level Indicator. An SLI is a service level indicator—a carefully defined quantitative measure of some aspect of the level of service that is provided.

SLO – Service Level Objective. A target value or range of values for a service level that is measured by an SLI. A natural structure for SLOs is thus SLI ≤ target, or lower bound ≤ SLI ≤ upper bound.

SLA – Service Level Agreement. A contract between a service provider and a customer that defines the service to be provided and the level of performance to be expected. An SLA also describes how performance will be measured and approved, and what happens if performance levels are not met.

VAS **–** Vertical Autoscaler, Kubernetes Vertical Pod Autoscaler – autoscaling for Kubernetes pod that acts before SLO saturation is breached.

References

Datadog - SLOs: How to establish and define service level objectives

Fortinet - Disaster Recovery: Definition, Types & Planning

Grafana - Databases and SLOs: How to apply service level objectives to your databases with synthetic monitoring

Google Cloud Blog - SRE fundamentals 2021: SLIs vs SLAs vs SLOs

Google SRE Workbook, Chapter 2 – Implementing SLOs

Google SRE Workbook, Chapter 3 - SLO Engineering Case Studies

Google SRE Workbook, Chapter 4 – Service Level Objectives

Google SRE Workbook, Chapter 6 - Monitoring Distributed Systems

IBM - What is an SLA (service level agreement)?

The Friday Deploy — Measuring Freshness with Synthetics

VA DIRECTIVE 6004

VA OIT Change Control Standard Operating Procedure

Document Version Control

Date	Version	Change Details	By
4/14/2026	1.0	Initial version developed by the EERT and coordinated with stakeholders and the Transformation Planning Team (TPT)	OCTO/TAS

Disclaimer: This document serves both internal and external customers. Links displayed throughout this document may not be viewable to all users outside the VA domain. This document may also include links to websites outside VA control and jurisdiction. VA is not responsible for the privacy practices or the content of non-VA websites. We encourage you to review the privacy policy or terms and conditions of those sites to fully understand what information is collected and how it is used.

Statement of Endorsement: Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, and shall not be used for advertising or product endorsement purposes.