

DECISIONFLOW

Technical Capacity, Reliability and Cost Expansion Plan

Qualifying higher production volume without treating architecture as infinitely scalable

CONTROLLED SCALE PHASE | Version 1.0 | Illustrative 12-month bounded-growth program

EVIDENCE & PUBLISHING INTEGRITY

This public portfolio edition distinguishes actual evidence, assumed continuity steps, simulated case-study outcomes, targets, and recommendations. Simulated customers, commercial results, benchmarks, incidents, quotations, and later-lifecycle outcomes are illustrative and must not be interpreted as results from a live commercial product. Evidence labels are retained throughout the package to support transparent and responsible interpretation.

Document Control

Control	Definition
Document code	CS-05
Version / status	2.0 — Final Public Portfolio Edition — Simulated Controlled-Scale Scenario
Baseline / author	14 July 2026 - Min Thu Kyaw
Scenario horizon	Illustrative 12-month period after Deliver; outcome data is simulated.
Lifecycle continuity	Actual MVP -> assumed productionization -> simulated Launch -> Qualify -> Deliver -> Controlled Scale.
Scale decision	Grow the individual workflow within tested customer, AI, commercial, support, and capacity boundaries.
Scope boundary	No broad collaboration, integrations, richer ingestion, multilingual promise, SSO, or enterprise-platform claim.
Evidence rule	Actual, assumed, simulated, target, and recommendation statements are separated.

Actual Project Artifacts

Artifact	Reference
Low-Fidelity Prototype	View prototype
High-Fidelity Prototype	View prototype
MVP Frontend Repository	Open repository
MVP Backend Repository	Open repository

Contents

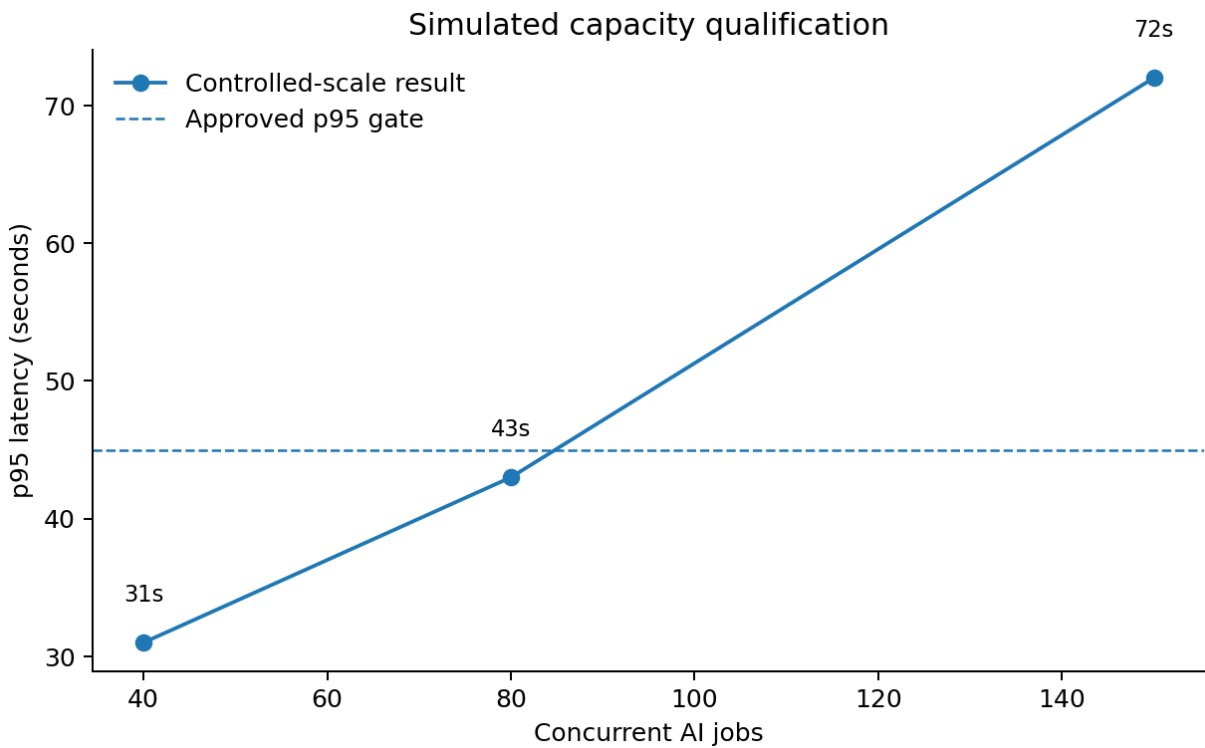
01	Technical scale objective
02	Capacity model and boundaries
03	Reliability and error budgets
04	Cost and resource controls
05	Resilience and recovery
06	Security and data isolation at scale
07	Technical decision

Technical Scale Objective

Technical scale expands the known operating range of the released service. It does not claim an unlimited cloud architecture. The approved range is based on load tests, production telemetry, recovery drills, provider constraints, security controls, and support readiness.

Approved boundary - 80 concurrent AI jobs are qualified at the defined p95 latency and failure gates. The 150-job profile remains outside the supported boundary and is a future investment decision.

Capacity Model and Boundaries



Capacity qualification

Profile	p95 latency	Failure rate	Decision
40 concurrent AI jobs	31s	0.4%	Supported with healthy headroom.
80 concurrent AI jobs	43s	0.8%	Approved controlled-scale ceiling.
150 concurrent AI jobs	72s	1.9%	Not qualified.
Capacity input		Operating control	
Queue depth / age		Alert, admission control, and customer-visible status.	
Worker concurrency		Configured by tested provider/database limits.	
Provider quota		Tracked as a hard dependency; commercial growth cannot assume unlimited quota.	
Database connections		Pool limits, query telemetry, and migration review.	
Retry behavior		Bounded attempts, idempotency, and dead-letter handling.	
Input size		Character/upload limits and actionable rejection before provider cost.	

Reliability and Error Budgets

Indicator	Objective	Scaled result	Status
-----------	-----------	---------------	--------

Indicator	Objective	Scaled result	Status
Availability	>=99.9%	99.97%	Pass
p95 non-AI API	<500 ms	315 ms	Pass
p95 normal AI generation	<=35s	24s	Pass
80-job p95 AI latency	<=45s	43s	Pass
80-job AI failure	<1%	0.8%	Pass
Duplicate generation	0 from one user action	0 in sampled window	Pass
Critical unresolved security issue	0	0	Pass
Restore RTO / RPO	<=45m / <=5m	32m / 3m	Pass
Budget state		Growth action	
Healthy		Continue bounded acquisition and normal releases.	
Half consumed early		Reduce release risk and examine capacity/support demand.	
Exhausted		Pause non-essential growth and focus reliability.	
Security / isolation failure		Immediate containment; not subject to ordinary error-budget trade-off.	

Cost and Resource Controls

Cost driver	Scaled control	
AI tokens / generation	Input limits, prompt efficiency, qualified routing, regeneration visibility.	
Worker compute	Autoscaling within tested bounds and queue-age alerts.	
Database	Query/index review, retention policy, backup lifecycle, and storage monitoring.	
Observability	Useful structured telemetry without raw content; retention tiering.	
Support	Self-service diagnostics and correlation IDs reduce engineering escalation.	
Unused capacity	Do not pre-build 150-job architecture without observed demand.	
Measure	Deliver	Controlled scale
Variable AI cost / finalized decision	\$0.29	\$0.23
Support contacts / activation	0.24	0.16
Gross margin	84.0%	85.5%
Support-adjusted contribution	58%	65%

Resilience and Recovery

Failure scenario	Required behavior
Primary AI provider outage	Pause/queue or use approved alternate; expose provenance and status.
Queue backlog	Admission control, progress state, retry cap, and no duplicate decision.
Database degradation	Protect writes, alert, restore/rollback plan, and ownership validation after recovery.
Deployment defect	Staged release, smoke test, rollback, and migration compatibility.
Billing/entitlement delay	Reconcile from provider event IDs; avoid accidental loss of customer access.
Region/platform outage	Document recovery boundary; do not market untested multi-region resilience.

- Quarterly restore drills and provider-failure game days are required.
- Recovery validation includes authorization, finalized-record integrity, and AI provenance—not only service availability.
- A rollback cannot silently switch users to a behaviorally different model or prompt.

Security and Data Isolation at Scale

Control	Scale requirement
Authorization	Owner-scoped access at every query and negative isolation tests in CI.
Session	Secure cookie/session, expiry, revocation, password recovery, rate limits.
Secrets	Managed storage, rotation, no production defaults, access review.
Sensitive content	Retention/deletion policy, provider disclosure, no raw content in analytics.
Abuse	Auth and AI rate limits, size controls, anomaly review.
Dependency risk	Patch/vulnerability process and provider/vendor inventory.
New multi-user surface	Separate tenancy, permission, audit, and threat-model qualification before delivery.

Technical Decision

Decision - THE SERVICE IS QUALIFIED FOR CONTROLLED GROWTH UP TO THE TESTED 80-JOB BOUNDARY. Investment in a 150-job or multi-tenant enterprise architecture is not approved without observed demand and a separate business case.

- Continue monitoring queue age, provider limits, database capacity, and support load.
- Review the growth cap before campaigns or partner commitments.
- Preserve architecture and runbook documentation for future transfer, integration, maintenance, or retirement decisions.

Evidence discipline: replace all simulated controlled-scale values with observed, reproducible results before presenting them as actual product performance.