

Forge · Fabric / Architecture Master Plan

Autonomous Execution Platform · Powered by Agentic AI · Local + Cloud, governed

Forge — Product · Control Plane. The operator experience for managing autonomous production agents across per-vertical workspaces.

Fabric — Core Engine. The proprietary Agentic AI execution engine that observes, reasons, plans, executes & learns across production work — routing a **capability**, never a model, across local and cloud.



Architecture Master Plan · **v3.2** (2026-08)

Aravindh Subramanian
Architect, Forge

6 Forge Agents

7 Verticals

5 Runtimes · incl. engineering

11 Integrations

2,649 Automated Tests

~85 min Saved / Ticket

⌚

<30 sec runtime (was ~1 day)

100% Audited Runs

Zero-touch Unity Forge drives the Editor

Local + Cloud AI capability-routed · min_cap floor

17-state governed engine · stops at draft PR

Repo-free signed package · one FORGE_PROFILE

14 parity areas verified as-built (v3.2)

FORGE — CONTROL PLANE · BROWSER UI · AGENT MANAGEMENT · MONITORING · OPERATIONS

A vanilla-JS single-page app organised into per-vertical **workspaces** (Tech Art · Design · Production · QA · Product · Dev · Art), each with its own agents, pipeline, queue, approvals, history & knowledge — scoped by an X-Workspace-Id header. **Forge Studio** composes agents visually; no code.

Workspace Tabs

Dashboard

Pipeline

Tickets

Queue

Pending Approvals

Pull Requests

Multi-Agent

Completed

Hanging work

History

Operation Rooms — signed fact projections

Incident & Reliability

Health

Settings · Logs

Forge Studio — visual agent composer

drag-drop canvas · node tools · undo/redo

check-flow + repair-flow · certify · replay

AI Draft — ticket → grounded agent → contract auto-repair

version history · rollback · templates · components

Weave — screen-aware assistant

Train Fabric — teach knowledge / skills

Forge Pet — observer-only live-state dock · click → details

▲ HTTP + NDJSON stream · X-Dashboard-Token · X-Workspace-Id ▼

ENTERPRISE SECURITY & ACCESS

Origin / Host validation

CSRF protection

4-mode RBAC — viewer → operator → deployer → master

Canonical Event Spine — server-attested · signed · hash-chained

Private network (Tailscale)

SSO — roadmap (Okta / device enrollment)

secrets in 0600 files · never in agent JSON / RAG / logs

◆ FABRIC — AGENTIC AI EXECUTION ENGINE · ONE RUNTIME, MANY AUTONOMOUS AGENTS

Observe → Reason → Plan → Execute → Learn

continuous improvement — validated outcomes enter durable Fabric knowledge

Execution Runtime

dispatch · scheduling · queue mgmt · parallel execution · retries · checkpointing · NDJSON streaming · SYNC_LOCK serialized writes

Decision Engine

runtime selection · confidence gate (blocks low-confidence deploys) · explainable decisions (Observe→Reason→Compare→Validate→Recommend) · Production Contracts + deterministic Graph Auto-Repair · human-in-the-loop approval · Validator Builder

Knowledge Engine

persistent memory · production context · decision history · confidence prediction · dual-mode RAG over all sources (workspace-scoped) · multi-source grounded drafting (field-level source precedence + provenance · never invents) · learns from decisions

Weave + Vision Engine

context- & screen-aware NL operations · RAG-grounded answers · knowledge search · Fabric Visual Intelligence — art-QA on rendered pixels (local VLM) · feeds confidence

Code Intelligence Engine

ADR-0003 · accepted structural-first understanding that shrinks the LLM search space: impacted symbols · call chain · blast radius · neighbors · hotspots. Graphify is the first provider; derived, trust-tagged evidence — never a gate.

Governed Engineering Engine

the 17-state bug-fix machine (Dev vertical): Bugsnag / JIRA → grounded plan → isolated patch → review → validate → approve → draft PR. Never merges or deploys. → detailed at right

Canonical Forge Event Spine

forge.event.v1 · server-attested · idempotent admission · append-only SQLite · workspace hash chain · read-only Operation Rooms / RAG / audit / live UI projections

5 runtimes, one engine: generic (art) · unity (Zero-touch Action Layer → live Editor over MCP → headless batchmode → handoff; the game repo is never permanently touched) · offer (LiveData → S3) · data (character data/config into the gotm client) · engineering (governed bug-fix). Every new production workflow becomes a new agent on this runtime — not a new tool.

▲ read-only sync · gated writes · live health probes ▼

EXTERNAL SYSTEMS & CONNECTORS · 11 · PROBED LIVE

Perforce (source art)

Jira DC

Confluence DC

Adobe Photoshop

Unity Editor (live MCP bridge)

git → GHE → S3

Google Sheets (Apps Script)

Bugsnag (SmartBear MCP)

Datadog

Slack

Future MCP endpoints

INCIDENT & RELIABILITY — DETECT · CORRELATE · RESPOND · FIX · VERIFY

One governed reliability workflow across Datadog + Bugsnag, with a live monitor correlating signals to incidents by severity.

1 · Detect & assess → 2 · Evidence → 3 · Respond → 4 · Fix → 5 · Verify recovery

SEV 1-4 severity model · impact + action

Live Reliability Monitor · signals every 60s

correlate → own → mitigate → hotfix

feeds the governed bug-fix engine

FABRIC AI — MODEL ROUTING · LOCAL ↔ CLOUD

A capability request asks for a lane, never a model. The router ranks registry cards (data, not code) by the lane's primary axis and applies a min_cap floor — an empty set is an explicit pool_exhausted, not a silent fallback.

fast

reason

code

review

vision

plan

embed

Local provider pool

fabric_providers.select_backend across vLLM + Ollama · lazy load + evict to fit RAM (128 GB) · healthy → resident/warm → backpressure → least-in-flight

Cloud adapter

fabric_ai.chat_with_failover · Anthropic Claude for reason/code · monthly USD budget · degrade to local on exhaustion or health failure

policy split — fast/vision/embed stay local

reason/code burst to cloud when ready

benchmark routing seeds cold picks

learned routing — cloud wins become local teacher signals

◆ DUAL-MODE RAG — FABRIC_KB IS TRUTH, THE INDEX IS DERIVED

Chosen once at deployment, validated by deployment_config (accepts only (local, local) or (hybrid, postgres), else fail-closed). Sources: run history · KB · agents · skills · docs · admin files/folders/URLs.

LOCAL — on-device

JSONL store · nomic-embed-text (768-dim) · brute-force cosine · atomic promotion (temp → swap) · private · \$0 · never auto-syncs

HYBRID — shared

PostgreSQL + pgvector vector(768) · EmbeddingContract (pins model/dims/norm) · QueryScope + row-level security · fail-closed · tombstoned revocation · re-embed audit

one retrieval contract — caller never knows which store answered

evidence returned trust-tagged (rag), never a gate on its own

FABRIC OPTIMIZATION ENGINE — TRANSPARENT, BEFORE EVERY LLM CALL

Never sends a raw prompt: Context → Token Optimizer → Prompt Compiler → Model Router → Reliability → Confidence → Learning. Models are implementation details — users see one reliable Fabric Intelligence.

Context — session memory + rolling summary

Token Optimizer — dedup · compress · fit window

Capability Router — swappable by config · fast draft lane

Reliability — retry + failover + hard wall-clock deadline

Confidence — calibrated per-call

Learning — prompt cache + routing stats

✱ PRODUCTION KNOWLEDGE GRAPH & GRAPHRAG — GOVERNED PRODUCTION INTELLIGENCE

Workspace/game-scoped, read-only projection over Jira · Perforce · Unity · Graphify · runs · artifacts · builds · incidents · owners — federating existing authorities (Graphify AST · workflow DAGs · Event Spine). Hybrid GraphRAG answers bounded typed operations — blast radius, asset consumers, artifact origin, ownership — with trust-classed provenance. The Character Definition-of-Done is one consumer.

trust classes — authoritative · derived · hypothesis

as_of time-travel · signed-event replay · online backup

SQLite/FTS5 + PostgreSQL adapter · migration 014

Definition-of-Done — readiness % + per-requirement evidence

human sign-off stays manual

◆ GOVERNED ENGINEERING ENGINE — 17-STATE FSM OVER TYPED ARTIFACTS · IN DEVELOPMENT

A single bug-fix run advances a finite-state machine over typed, provenance-carrying artifacts and deliberately stops at a draft PR — it never merges or deploys.

observed → grounded → planned → patching → patched → reviewing → reviewed → validating → validated → awaiting_approval → approved → committed → draft_pr_created

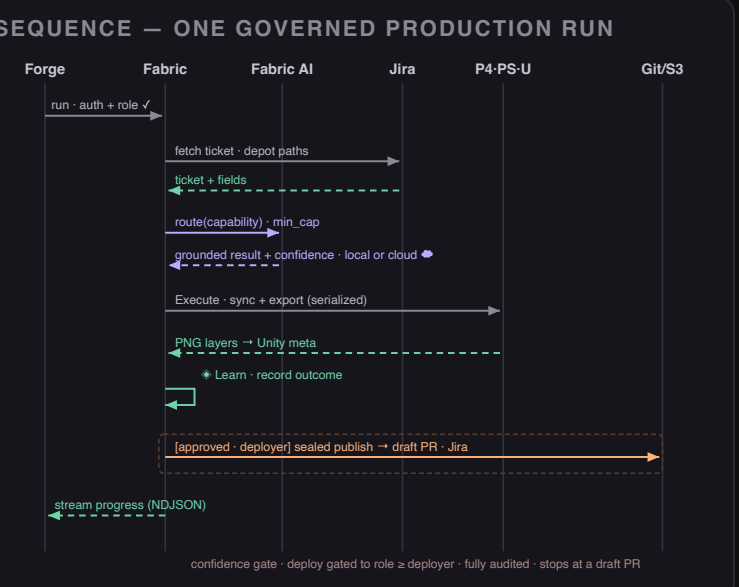
patch under CODE_FLOOR 0.85 · pool-exhausted is terminal

validation blocked if any required check fails

publish = sealed saga (commit + push → draft PR)

off-path: failed · blocked⇒resumed · cancelled · rejected · rolled_back

trust tiers — trusted (our bytes/git) · internal (Jira/Bugsnag) · untrusted (free-form text, quarantined)



✱ FABRIC AGENT HARNESS — VALIDATE · SIMULATE · CERTIFY

Read-only trust layer between Intelligence & Runtime — certifies every agent before it runs (plug-in checks · explainable verdict · advisory to the approval gate, never a hidden blocker).

Readiness certificate + confidence meter

Golden baselines + regression ▲▼

Security — secrets · shell-safety

Replay — predicted vs actual

Contract coverage + AUTO-executor gate

Fleet Quality Gate — studio-wide roll-up · weakest-first

Intelligence Profiles v1.2.0 — resolved caps · sealed skill manifests · tool readiness

◆ SKILLS — KNOWLEDGE VS EXECUTABLE · 48 FROM .CLAUDE/SKILLS

Every enabled skill is usable immediately as RAG-ready playbook guidance (chunked · embedded · validated). Executable skills bind to matching agents at run start and are sealed by version + digest — no silent drift.

36 immediately usable

32 bound to agents

capabilities gated on registry health

real tools (PS · Unity · P4) probed live, gate their caps

DEPLOYMENT — ONE FORGE_PROFILE PICKS THE SHAPE

deployment_profile.py resolves the profile and fails closed on conflict. Both shapes ship as a repo-free, signed, pyc package — a studio stands Forge up without cloning the repo (ADR-0002).

LOCAL

one macOS workstation · launchd backend · :8787 · browser · Ollama pool · JSONL RAG · native tools — all on-device · private · \$0

HYBRID

control-plane (shared backend · Postgres+pgvector · cloud routing · RBAC/SSO — no game tools) + authenticated workers (Unity · PS · P4 · Git) enrolled with one active credential

local · control-plane · worker

control plane holds knowledge + governance

workers hold the tools · horizontal fan-out

SAFETY NETS & CONTROL PLANE

Dry-run default · confidence gate + approval · role-aware deploy

Explainable · zero-invention (unresolved, never guessed)

Approvals · Mission Control · auto-knowledge + provenance

Telemetry · impact · health · baseline · learning · registry

Serialized safe writes — one run / workspace · branch & stash aware

Run journal + post-mortems — partial effects recoverable

1,112 tests

ARCHITECTURE DECISION RECORDS

ADR-0001 RAG deployment modes

Local JSONL vs Hybrid pgvector; EmbeddingContract, QueryScope, RLS, fail-closed. Accepted.

ADR-0002 Distribution & packaging

FORGE_PROFILE local/control-plane/worker; repo-free signed backend; control-plane + worker split. Proposed.

ADR-0003 Code Intelligence Engine

Structural-first intelligence behind an evidence contract; Graphify as first provider. Accepted.

DETAILED DIAGRAM SET — SVG · PNG · PDF

00 · one-pager

01 · layered components

02 · ticket-run sequence

03 · deployment view

04 · AI operating system

05 · model routing

06 · dual-mode RAG

07 · engine FSM

08 · agent harness

◆ FABRIC — THE AI OPERATING SYSTEM

Fabric Intelligence is the only user-facing surface — users never pick a model. A capability contract sits between the product and the models, so the AI layer is replaceable without changing Forge.

Forge → Fabric → capability contract → model plugin

swap any model · nothing above the router moves

PLATFORM EXPANSION — ONE RUNTIME, MANY AUTONOMOUS AGENTS · EACH NEW VERTICAL SHIPS AS NEW AGENTS, NOT A NEW TOOL

Tech Art ✓ → Design × in dev → Dev / Engineering × in dev → Production → QA → Product → Art → LiveOps → Marketing

Tech Art — 3 live in prod · rest rolling out: Ability (generic-13) · Flat Art Import (unity-11) · New Flat Art Setup (unity-10) · Offers-DD + DD Iconic Item (offer-8) · PG Setup

Design — Design Character Setup (data-12) · in development, not yet live

Dev / Engineering — governed bug-fix engine (Bugsnag/JIRA → draft PR) · in development, not yet live

7 verticals stood up · agents rolling out per workspace

✓ ENTERPRISE PARITY & SCALE — V3.2 · 14 WORKFLOW-PLATFORM AREAS VERIFIED AS-BUILT (SEE DIAGRAM 09 · ENTERPRISE ARCHITECTURE) · HOSTED DISTRIBUTED TIER DESIGNED & PROTOTYPED, NOT YET A RELEASE

Workflow DAG — dependency waves · branches · joins · retries · cycle-checked

Callable workflows · minions — digest-bound · typed I/O · versioned publish

Federated dependency migration — sealed multi-copy campaigns · per-draft lifecycle · Harness-bound certification · governed contract remapping · durable notify / reconcile

Triggers — webhook · cron · poll · durable delivery ledger

Typed data mapping — field → field · schema-checked

Debug · replay — checkpoints · execution inspection · pause-anywhere interventions

Agent Architect — natural-language → typed, validated brief

AI evals · regression gates — datasets · baseline floor + drop gate

Approvals · interventions — digest-bound · fenced resume tokens

Sharing · promotion — ACL roles · credential-safe · dev → staging → prod

Runtime queue — claim-fenced leases · bounded concurrency · backpressure · dead-letters

Templates · ecosystem — trust-tiered · digest-pinned installs

Governed credential storage — macOS Keychain · env → Keychain → 0600 facade

Texture · Vision QA — PVRTC · ASTC encode + VLM before/after gate

Hosted M1 (designed) — Postgres leases/fencing/outbox · object-store artifacts · SSO/OIDC · fail-closed hosted startup

AI Operating System — models are plugins, not dependencies · capability contract + scheduler · swap models, nothing above moves

Local + Cloud, governed — capability-routed · min_cap floor · private lanes stay local · budget-bounded cloud

Config-driven agents — new domains = JSON agent defs on one runtime

