

Live Features and Architecture

Evidence-based description of the Enterprise Pilot currently running on 172.16.10.155



Evidence date 1 September 2026	Lifecycle Enterprise Pilot / Gate 10 provisional	Reference artifact 1.0.0-rc1.15 / a7bd086...
-----------------------------------	---	---

Prepared from live cluster inventory, deployed portal/runtime source, signed release artifacts, direct health probes, and current readiness, security and backup evidence.

Executive summary

Nexus One AI is already a broad on-premises AI application and control plane, not a single chatbot. The live pilot combines role-aware AI workspaces, enterprise knowledge, documents, evaluations, automation, App Studio, governed integrations, identity, durable data services, observability, recovery evidence and signed offline release controls.

Most important baseline fact
The latest signed artifact present is RC1.15, but the live cluster also contains later Kubernetes overlays. There is no single immutable version identifier covering everything currently running. Quote both the release artifact and the 1 September 2026 evidence date.

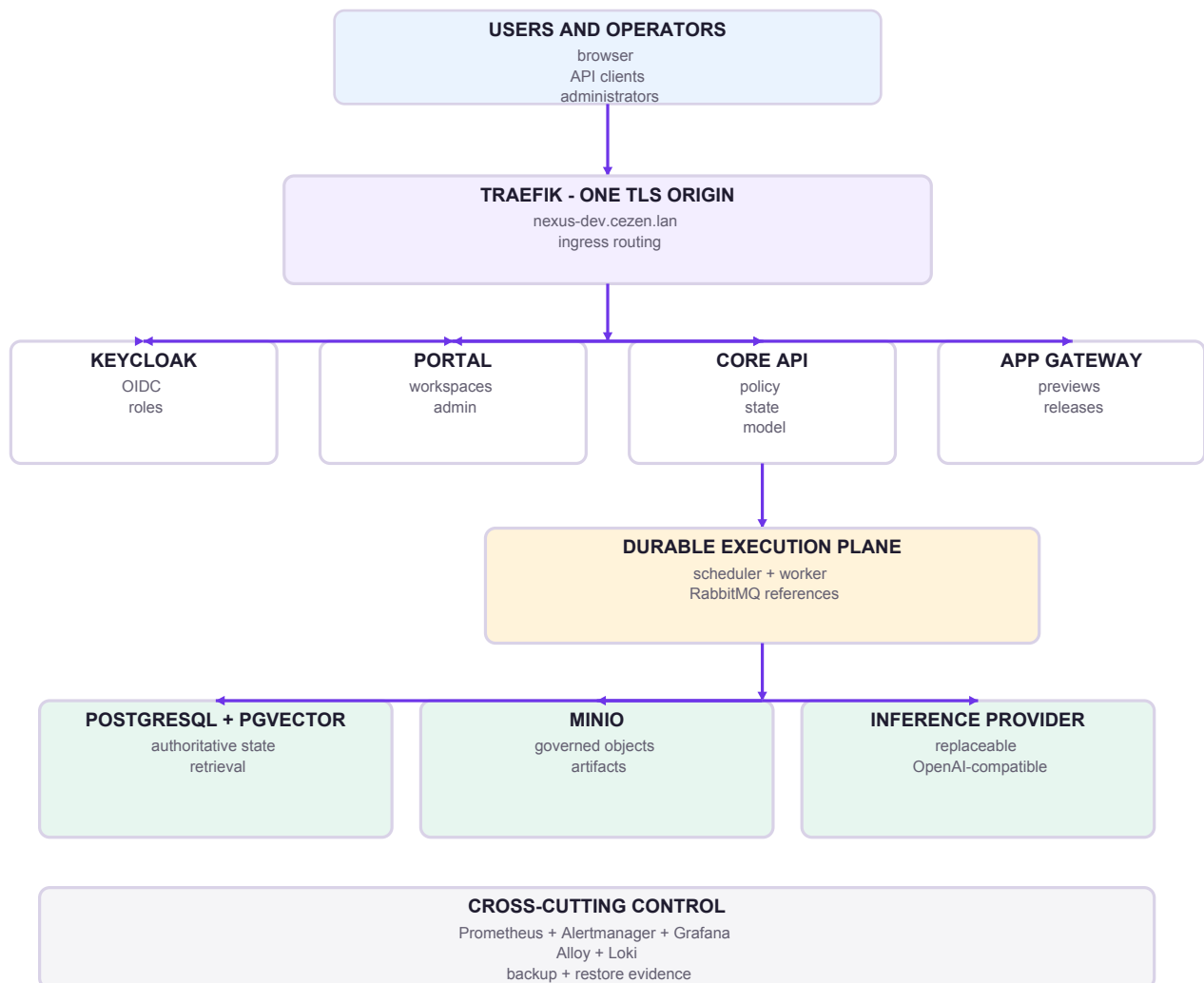
Area	Verified state
Lifecycle	Enterprise Pilot; Gate 10 provisional; not production-promotable
Runtime	Ubuntu 22.04.5; single-node K3s v1.36.2+k3s1
Entry point	Traefik at nexus-dev.cezen.lan with Keycloak OIDC
Core services	Portal, Core API and App Gateway each have two ready replicas
Durable data	PostgreSQL + pgvector, RabbitMQ and MinIO
Observability	Prometheus, Alertmanager, Grafana, Loki and Alloy
Latest evidence	Readiness pass; security pass; backup pass; HTTP 200 core probes
Known issue	One newer document-scanner pod crash-loops while an older replica is ready

What this document does

It separates four states: live, gated, provision-only and known issue. A visible UI screen is not treated as proof of a fully qualified end-to-end capability.

System context and trust boundaries

All browser traffic enters through one TLS origin. Keycloak supplies identity and role claims. The portal exposes user workspaces and administration, while the Core API owns policy and durable state. Background work moves through a scheduler/worker plane. The model provider is replaceable and does not determine control-plane availability.



Architectural invariant

PostgreSQL is authoritative. RabbitMQ carries durable job references, MinIO holds governed objects, and model execution is reached through a logical gateway rather than embedded into every application.

Live deployment topology

The reference environment is a single K3s control-plane node. Product domains are separated by namespaces and service boundaries.

Namespace / layer	Components	Responsibility
nexus-platform	Portal, Core API, Admin Extension, App Gateway, Scheduler, Worker, Open WebUI, document services	Product APIs, workspaces, jobs, document processing and governed application access
nexus-data	PostgreSQL 17 + pgvector, RabbitMQ 4.1, MinIO	Authoritative state, retrieval vectors, durable work transport and governed artifacts
nexus-workspaces	JupyterHub hub and proxy	Persistent authenticated developer workspaces
nexus-app-previews	Isolated preview Deployments and Services	Temporary App Studio validation and preview runtimes
nexus-observability	Prometheus, Alertmanager, Grafana, Loki, Alloy and exporters	Metrics, alerting, dashboards, logs and capacity evidence
kube-system	Traefik, CoreDNS, metrics-server, local-path provisioner	Ingress, discovery, metrics and local storage

Capacity snapshot
8 vCPU, 15 GiB RAM and a 392 GiB root volume at 71% use. The .155 control-plane node has no local GPU; inference and GPU telemetry can be supplied by an associated accelerator provider.

Identity, roles and governance

Keycloak is the central identity provider. Navigation and APIs are filtered by capability and role. Reviewer and read-only roles remain restrictive unless an explicit operational or administrative role is also present.

Role family	Typical access
Application user	AI Workspace, prompts, Notebook, documents, agents, workflows, schedules, meetings and chat rooms
Developer	Build tools, App Studio, model/training surfaces, Jupyter, APIs and credentials
Reviewer / auditor	Assigned reviews, evidence, evaluation and governance/audit surfaces
Operator	Health, logs, jobs, platform operations and GPU monitoring
Pack manager	Signed capability-pack lifecycle
Platform administrator	Identity, models, connectors, policies, packs, quotas, security, operations and settings

Governance boundaries

- Connector secrets remain server-side and the browser receives references only.
- External hosts and capabilities are allowlisted by accepted contracts.
- Generated applications and packs run in isolated namespaces with bounded permissions.
- Consequential lifecycle steps support independent review, acceptance, versioning and rollback.
- Audit ownership, protected secret storage and image pinning are checked by recurring security evidence.

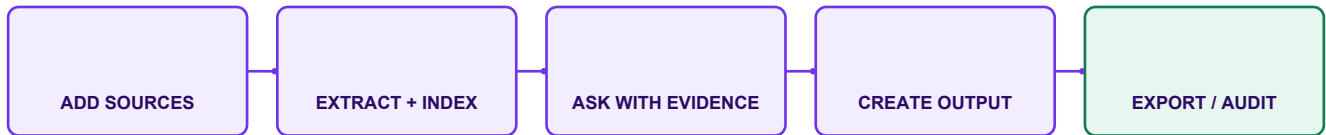
AI Workspace, prompts and quality

Capability	What it does	Current state
AI Workspace	Open WebUI behind Nexus SSO for persistent approved-model conversations and file-assisted chat.	LIVE
Prompt Studio	Variables, expert personas, structured instructions, governed inference runs and evaluation hand-off.	LIVE
Prompt Library	Private/shared versioned prompt assets, categories and controlled changes.	LIVE
Evaluation suites	Repeatable test cases with persisted model/judge quality and safety evidence.	LIVE
Model comparison	Side-by-side response, latency and usage comparison for approved models.	LIVE
RAG quality	Retrieval and grounded-answer assessment against selected collections.	LIVE / DATA-DEPENDENT
Feedback review	User ratings/comments promoted into review and evaluation evidence.	LIVE

Evidence rule
General chat output is not source evidence. When traceability matters, use Nexus Notebook so answers remain tied to governed sources and citations.

Nexus Notebook and grounded knowledge

Notebook is the governed evidence workspace. Collections are private by default, may be shared through access control, and retain sources, conversations, generated outputs and export evidence.



Knowledge processing path

- Ingest approved documents under file, type and policy controls.
- Extract content and metadata; persist original and derived artifacts.
- Create searchable chunks and retrieval vectors in the durable data plane.
- Retrieve and rerank relevant source segments for each question.
- Generate an answer through the logical model gateway.
- Return citations, source links and exportable conversation/output evidence.

Current product boundary

Notebook is a governed retrieval and synthesis system. Answer quality still depends on source quality, extraction, chunking, retrieval configuration, model choice and evaluation coverage.

Document intelligence

Capability	Implemented path	Status
Document Compare	Durable jobs for TXT, Markdown, CSV, JSON, PDF, DOCX, XLSX and PPTX; persisted result, export and grounded AI brief.	LIVE
Document Digitization	Local OCR, assisted field extraction, human review, persistence and Notebook publication; Indian GST/PAN/tender assistance.	DEGRADED
Document Renderer	Isolated rendering service for supported documents and previews.	LIVE
Dataset creation	Generate grounded examples from PDF, DOCX, TXT, Markdown and CSV; review and publish immutable versions.	LIVE
Model training	Request LoRA/full/embedding work from published datasets with checksum lineage.	GATED

Known issue observed
A newer document-scanner pod had repeated CrashLoopBackOff restarts while an older scanner pod remained ready. Digitization should be described as degraded until the rollout is reconciled and a fresh readiness record is preserved.

Agents, workflows and collaboration

Automation uses the durable jobs plane so runs have explicit status and history. Human review can be inserted before consequential actions.

Capability	Current behavior	Boundary
Agent Studio	Governed multistep agents using approved models and tools.	Tool permissions and human-control points must be explicit.
Workflow Designer	Ordered AI, retrieval, transformation and integration steps.	Failures are surfaced; authoritative state remains durable.
Schedules & Runs	Manual or timed execution with run history.	Scheduling does not bypass approval or entitlement.
Meeting Assistant	Provider-neutral record and post-meeting evidence pipeline.	Cloud providers require provisioning, consent and retention decisions.
Chat Rooms	Membership-controlled human and AI rooms with optional Notebook grounding.	Room membership and source ACLs still apply.

Provider provisioning

Microsoft Teams, Zoom and Zoho Meeting are represented as provision-only adapters. Activation requires the customer tenant/account, an approved HTTPS API host, a server-side secret, consent evidence, retention limits and an accepted regional data boundary.

App Studio and application lifecycle

App Studio separates creative work from production acceptance. Projects have durable workspaces and source history; generated code runs in isolated Kubernetes sandboxes; publication is a reviewed lifecycle action.



Stage	Control
Project and workspace	Persistent project record, proposal and source history
Sandbox	Isolated execution with bounded resources and cleanup
Preview	Temporary Deployment/Service reachable through governed preview routing
Review	Independent approval decision and evidence
Publish	Versioned application release through the App Gateway
Operate	Health, usage, logs and lifecycle visibility
Rollback / retire	Controlled predecessor restoration or removal without losing audit history

Extensions, connectors and MCP

Area	What is implemented	Customer work still required
Pack Manager	Signed offline import, validation, acceptance, enable/disable, upgrade, rollback, removal and quarantine.	Approve trust root, pack policy and operational owner.
Extension SDK	Versioned Python and TypeScript clients, templates and offline conformance suite.	Develop, test and sign customer extensions.
Connectors	Governed manifest registry, server-side secrets, allowlisted endpoints and bounded capabilities.	Supply tenant details, credentials and contract tests.
MCP projection	Authenticated JSON-RPC projection of accepted connector capabilities.	Grant exact roles/scopes and monitor use.
Meeting providers	Provider-neutral manifests for Teams, Zoom and Zoho Meeting.	Consent, retention, data boundary and provider permissions.

Fail-closed principle
A configured integration is not automatically active. Provisioning, validation, capability grant and monitoring are distinct lifecycle stages.

Security control plane

The latest security evidence passed all recorded checks. These are environment controls, not an external security certification.

Control	Evidence on 1 September 2026
Kubernetes secret encryption	K3s secretbox enabled; hash consistency verified
Protected secret files	No unexpected group or other access detected
Portal certificate	807 days remaining at evidence time
Workload privilege	No disallowed host/privileged configuration found in running workloads
Image integrity	No unpinned running core containers detected
Audit ownership	Append-only audit owner recorded as nexus_admin
Identity	Keycloak OIDC and role-aware application authorization
Release / packs	Signed inventories, digest pinning and fail-closed validation contracts

Still environment-specific

- Customer production PKI and certificate lifecycle.
- Customer AD/LDAP federation, MFA and identity governance.
- External penetration testing and formal security certification.
- Customer data classification, retention, DLP rules and incident-response integration.

Observability and operations

Signal	Live component / objective
Metrics	Prometheus with Kubernetes, host, PostgreSQL and GPU exporters
Dashboards	Grafana platform, node capacity, inference, GPU and PostgreSQL views
Logs	Alloy collection and Loki retention/query
Alerts	Alertmanager with warning/critical operator response
Portal readiness	Internal objective: 99.5% monthly
Inference latency	Internal objective: 95% under 30 seconds
GPU telemetry	Warning if missing for 5 minutes; critical at 15 minutes
GPU VRAM	Warning at 90%; critical at 97%
Backup freshness	Latest backup should remain under 30 hours
Restore	Daily non-destructive verification should pass

Service-level boundary

These are operational objectives for the current single-node pilot, not contractual customer SLAs.

Backup, recovery and current evidence

Evidence	Latest inspected result
Daily enterprise readiness	PASS - generated 31 Aug 2026 06:06 UTC
Security compliance	PASS - generated 1 Sep 2026 00:00 UTC
Backup evidence	PASS - generated 1 Sep 2026 04:55 UTC
Latest backup age	2.52 hours
Retained backup count	32
Backup storage headroom	28.68% free
Portal probe	HTTP 200
Core API health probe	HTTP 200
Open WebUI probe	HTTP 200

Recovery boundary

Current evidence covers recurring local backup checks and non-destructive restore verification. Independent encrypted off-host recovery and a complete replacement-host restoration remain open release requirements.

Release model and traceability

The server holds signed release bundles and a bootable AMD64 pilot ISO. Runtime images are digest-pinned and the release programme includes source/component inventories, pack signatures, compatibility policy and controlled upgrade/rollback contracts.

Traceability element	Current observation
Latest signed artifact present	nexus-one-ai-1.0.0-rc1.15-a7bd086...tar.zst with checksum and signature
Live source overlays	Portal/jobs/core updates continue after the RC artifact date
Live cluster identity	No single immutable manifest covers all current overlays
Promotion state	Gate 10 open and provisional; promotion not allowed
Clean-machine installation	Gate 11 not started
Required correction	Freeze overlays into source, generate one build manifest and rebuild current signed media

How to cite this environment
Use: "Nexus One AI Enterprise Pilot on 172.16.10.155, verified 1 September 2026; latest signed artifact present RC1.15." Do not call it a final production release.

Known limitations and open release gaps

- Single-node topology does not prove high availability.
- One document-scanner rollout replica is unhealthy.
- Live overlays are not frozen into one signed build identifier.
- Independent off-host and replacement-host recovery are not yet proven.
- Prior-supported-release upgrade, migration compatibility and rollback remain open.
- Customer production PKI and directory federation require site validation.
- External connectors and meeting providers require customer provisioning and acceptance.
- Performance, concurrency and model quality are not certified across commercial hardware tiers.
- Gate 10 remains provisional and Gate 11 clean-machine/offline installation has not started.

Truthful product posture

The platform has substantial implemented capability and recurring evidence, but the correct description today is "functioning Enterprise Pilot release candidate," not "production-certified platform."

Recommended next actions

Priority	Action	Completion evidence
P0	Reconcile the document-scanner ReplicaSet and verify OCR/digitization.	All intended replicas ready; functional digitization test; fresh readiness evidence
P0	Freeze live overlays into a reviewed source revision.	Commit, migration inventory, image digests and one build manifest
P0	Rebuild current signed release media.	Verified bundle/ISO, checksum, signature and clean install result
P1	Prove off-host and replacement-host recovery.	Encrypted export, restore drill and measured RPO/RTO
P1	Qualify supported upgrade and rollback paths.	Migration compatibility and predecessor restoration evidence
P1	Create model/hardware capacity matrices.	TTFT, tokens/sec, concurrency, context and VRAM assumptions
P2	Complete customer integration profiles.	PKI, AD/LDAP, inference, storage, DLP and connector playbooks
P2	Automate this reference from evidence.	Generated inventory and versioned capability-status report

Appendix A - Live component inventory

Domain	Verified components
Ingress and identity	Traefik, Keycloak, canonical TLS origin
Product services	Portal, Core API, Admin Extension, App Gateway
Jobs	Scheduler, Worker, RabbitMQ
AI workspace	Open WebUI, logical model gateway
Knowledge and documents	Notebook, pgvector retrieval, renderer, scanner, compare and digitization workflows
Applications	App Studio previews, application gateway and lifecycle controls
Developer environment	JupyterHub, Python/TypeScript SDKs and templates
Durable storage	PostgreSQL, MinIO and Kubernetes persistent storage
Observability	Prometheus, Alertmanager, Grafana, Loki, Alloy and exporters
Recovery/release	Backup evidence, restore verification, support bundles, signed packs and release artifacts

Evidence sources

- Live K3s node, workload, Service and Ingress inventory.
- Deployed portal and jobs source under /srv/nexus/source/nexus-platform.
- Signed artifacts under /srv/nexus/artifacts/releases.
- RC1.15 Enterprise Pilot README and build-status record.
- Readiness, security and backup JSON evidence.
- Direct canonical-ingress health probes.