Kyūdō

SOVEREIGNTY-GRADE AI · GRC

Customer-Hosted GRC — Architecture Guide

Vigilance with Purpose. Security with Control.

PRESENTED BY

Kyūdō — a KMicro Technologies platform

3525-265 Hyland Avenue

Costa Mesa, CA 92626 · kyudo.ai

hello@kyudo.ai

kyudo.ai · April 2026

CLASSIFICATION: PUBLIC

00 · WHY ARCHITECTURE IS THE DIFFERENTIATOR

Sovereignty as topology, not policy.

This document is for the security architect, the cloud architect, the network engineer, the platform engineer, and the procurement reviewer. It is the architectural depth piece for Kyūdō's client-hosted Azure deployment topology. If you have read the Kyūdō Platform Overview, you have the four-vector frame; this document goes underneath it. If you are running a vendor-risk assessment, this document is what your committee will ask you to produce.

The single architectural commitment that defines Kyūdō's client-hosted topology is this: every component of the platform that processes governance data — the application services, the data plane, the AI inference, the secrets material, the operational telemetry that is not anonymized — runs inside the customer's Azure subscription. There is no cross-tenant data plane. There is no vendor-side database that aggregates customer compliance posture. There is no path by which Kyūdō operations personnel can reach customer compliance data, because no such path exists in the architecture.

This is sovereignty as topology, not as policy. A SaaS GRC platform that holds your data in its cloud and asserts 'we don't look at it' is making a contractual promise that depends on vendor governance. A client-hosted deployment where the data physically resides in your tenant and the vendor has no production credentials makes the same claim architecturally — and the architecture is what your auditor, your regulator, and your customer's vendor-risk team will examine.

What this guide will and will not cover

Will cover	Will not cover
Hub-and-spoke topology and the rationale for it. Compute, networking, data, identity, and AI layers in detail. Private endpoint patterns. Customer-managed keys. Sovereignty mechanisms. Azure Policy enforcement. Servicing model. RACI between KMicro and the client.	Specific deployment template internals (those are in the Reference Runbook distributed to engaged customers). Detailed monitoring runbooks. Disaster recovery procedures. Specific OAuth scope definitions per integration (those are in the Quick Start Guide). Operational training material.

Source authority

This guide is derived from the Kyūdō Client-Hosted Azure Reference Architecture & Servicing Runbook v1.0 (November 2025), the Kyūdō Software Requirements Specification v7, and the Kyūdō Architecture Brief. Where this guide simplifies for readability, the Reference Runbook is the authoritative document for engaged customers.

01 · THE REFERENCE ARCHITECTURE

Hub-and-spoke. Three planes. One tenant.

Kyūdō's client-hosted topology follows a hub-and-spoke pattern with strict private endpoints. The hub holds shared services (identity, networking, security perimeter); the spokes hold the application workloads (Kyūdō services, data plane, AI services). Every Azure service in the deployment exposes private endpoints only — traffic stays on the Microsoft backbone, never traversing the public internet for service-to-service communication.

The architecture decomposes into three planes, each with its own boundary, its own access controls, and its own audit trail.

Plane	What it holds	Boundary
Control plane	Kyūdō services on Azure compute: Controls Hub, Evidence Hub, Policy Center, Risk Management, VRM, Trust Center. The application logic layer.	Reachable only via Application Gateway through Front Door + WAF. Private addressing only.
Data plane	Azure storage services for relational, document, and artifact state, plus Azure Key Vault. The persistent state layer.	Private endpoints only. No public network exposure. Managed-identity-bound access.
AI plane	Azure OpenAI Service plus the supporting retrieval and inference workloads. The reasoning layer.	Private Link for Azure OpenAI. Tenant-scoped model deployments. Inference context never leaves the tenant.

The procurement filter

Sovereignty is becoming a procurement filter, not a preference. Three regulatory forces are converging:

- EU GDPR, the EU AI Act, and EU DORA each contain data-residency or processor-control provisions that vendor-controlled SaaS GRC platforms struggle to satisfy. The EU AI Act treats high-risk AI systems as subject to data-localization obligations that align poorly with cross-border SaaS architectures.
- U.S. CMMC Phase 2 enforcement (beginning November 10, 2026) explicitly contemplates the controlled environment within which Controlled Unclassified Information (CUI) may be processed. Vendor-controlled SaaS environments aggregating CUI across defense contractors raise material boundary questions.
- HIPAA Security Rule, including the pending NPRM, treats the boundary of ePHI as a regulatory artifact — the BAA chain breaks when ePHI traverses an environment the covered entity does not control.

Each of these regulatory regimes is satisfiable in a SaaS architecture with sufficient contractual instrumentation. None is satisfiable as cleanly as in a client-hosted topology where the regulated data physically does not leave the customer’s Azure subscription. The architecture is the answer to the regulatory question.

02 · THE COMPUTE LAYER

Azure compute. Managed identities. Service isolation.

Kyūdō services run on Azure compute infrastructure inside the customer's subscription. The compute layer is provisioned and scaled to provide the namespace isolation, network policy enforcement, and workload-level identity binding required for the multi-service topology Kyūdō uses, with the operational maturity required for production GRC workloads.

Compute topology

Compute resources are organized into workload classes, each sized for the operations it performs. The platform admin can adjust sizing at deployment time and scale capacity up or down based on observed load.

Workload class	What it runs	Default sizing
Platform workloads	System-level workloads, ingress, DNS, observability agents.	Production baseline; horizontally scaled
Application workloads	Kyūdō services: Controls Hub, Evidence Hub, Policy Center, Risk Management, VRM, Trust Center.	Production baseline; horizontal autoscaling enabled
AI workloads	Light-RAG inference, Knowledge Graph query services, vector retrieval workloads.	Higher compute density; scaled per peak compliance cycles

The compute control plane is private — management endpoints are not publicly resolvable. Operations are performed via Azure Bastion, a jump host in the customer's management subnet, or Microsoft Entra ID-authenticated tooling through a network-restricted path.

Service isolation

Each Kyūdō service runs in its own isolated namespace with namespace-level network policy. The blast radius of any single service fault is engineered to be less than 1% of the deployment, consistent with the SRS's 99.99% uptime target. Services communicate via an internal service mesh; cross-namespace traffic is governed by network policy and service-to-service authentication uses managed identities, not shared secrets.

Container registry

Azure Container Registry holds the Kyūdō service images. The registry is accessed via private endpoint; image pulls authenticate via managed identity bound to the compute layer. Image

signing uses Notary v2 / Cosign; signature verification is enforced at admission via Open Policy Agent / Gatekeeper or the Azure Policy add-on.

Service-package deployment

Kyūdō services deploy via versioned service packages alongside the application code. Package rollback capability is the per-service unit of recovery; combined with blue/green deployment per service, this produces zero-downtime updates and a predictable rollback path. Package versioning is semantic (major.minor.patch) and the package manifest is part of the audit trail for any deployment event.

> ### Compute layer commitments
>
> All Azure compute resources, the container registry, and the service packages run inside the customer’s Azure subscription. KMicro has no production credentials; compute operations performed via the Reference Runbook are executed by the customer’s platform admin or, where contracted, by KMicro engineers acting under a time-bound delegated identity that the customer can revoke at any time.

03 · THE NETWORKING LAYER

Private endpoints by default. Public network access denied.

Kyūdō's networking model is private-first by default. The Azure Policy initiative described in Section 08 enforces this at deployment time: any service deployed into the Kyūdō resource group with a public endpoint is denied. The default state of every Azure service in the deployment is private endpoint only.

Network topology

Hub-and-spoke virtual network topology, deployed via parameterized templates and adjusted for the customer's IP address ranges. The hub virtual network holds the perimeter services; spoke virtual networks hold the workload services.

VNet role	Subnets	Services hosted
Hub VNet	Firewall, Gateway, Bastion, Management.	Azure Firewall Premium, Application Gateway, Azure Bastion, jump hosts. Outbound internet egress flows through the firewall.
Spoke VNet — Application	Compute, application, AI workload subnets; private endpoint subnet.	Kyūdō services, private endpoints for storage and AI services, Application Insights agents.
Spoke VNet — Data (optional)	Database, storage, key vault private endpoint subnets.	Storage service private endpoints, Key Vault private endpoint. Optional separation when customer policy requires.

VNet peering connects the hub to each spoke. Inter-spoke traffic is denied by default; transit between spokes flows through the hub firewall when required.

Perimeter services

Azure Front Door + WAF

Azure Front Door provides global HTTP entry for the Kyūdō web UI and API surface. The Front Door WAF (Web Application Firewall) is configured with the Microsoft-managed Default Rule Set (DRS) plus the Bot Manager ruleset. Custom rules can be configured per customer (rate limiting, geo-restriction, IP allowlists). Front Door terminates TLS at the edge and re-encrypts to the origin via HTTPS.

Application Gateway

Application Gateway sits between Front Door and the application ingress. It provides Layer 7 load balancing, mutual TLS termination for backend service authentication, and the second WAF layer (defense in depth).

Azure Firewall Premium

Azure Firewall Premium is the egress and east-west firewall. All outbound internet traffic from Kyūdō services (e.g., Microsoft Graph API calls for evidence ingestion) flows through the firewall. The Premium tier provides TLS inspection, IDPS (intrusion detection and prevention), URL filtering, and threat intelligence-based filtering. Outbound rules are Microsoft service-tag-based where possible (allow `MicrosoftContainerRegistry`, `AzureKeyVault`, `AzureActiveDirectory`, etc.) and FQDN-based for non-Microsoft services.

Private DNS

Private DNS zones resolve every Azure service used by Kyūdō to its private endpoint IP within the customer’s VNet. Standard zones cover the storage, secrets, and AI services in the deployment. Public Microsoft DNS resolution is bypassed.

Private endpoint summary

Every Azure service in the Kyūdō deployment exposes private endpoints only. The Azure Policy initiative `Kyūdō Baseline Compliance` denies any deployment that creates a public endpoint on any of the following service categories.

Service category	Private endpoint requirement
Relational data services	Private endpoint on the data service. Public network access set to Disabled.
Document and graph data services	Private endpoint on each account. Public access disabled.
Object and artifact storage	Private endpoint on the storage account. Public network access disabled.
Azure Key Vault	Private endpoint on the vault. Public network access disabled. Network ACL set to deny.
Azure OpenAI Service	Private endpoint on the cognitive services resource. Public network access set to Disabled. Inference traffic stays on the Microsoft backbone.
Azure Container Registry	Private endpoint on the registry. Public network access set to Disabled. Anonymous pull disabled.
Compute control plane	Private mode. Management endpoints not publicly resolvable.
Application Insights	Private Link Scope (AMPLS) binding all monitored resources to the workspace. Telemetry ingestion via private endpoint.

04 · THE DATA LAYER

Azure storage. Customer-managed keys. Residency under your control.

Kyūdō's data plane is composed of Azure storage services chosen for specific workload classes, each accessible only via private endpoint, each encrypted at rest with customer-managed keys (CMK) when the customer provisions them. Specific service selections are documented in the Reference Runbook; the architectural commitments below are constant across the data layer.

Relational state

Azure storage holds the relational state for the Knowledge Graph (where relational semantics are appropriate), the user and tenant tables, the audit log, and the workflow engine state. The relational store supports point-in-time restore (PITR) with extended retention to support governance audit requirements.

- Encryption at rest: AES-256 with customer-managed keys via Azure Key Vault. Transparent Data Encryption is enabled.
- Encryption in transit: TLS 1.3 minimum, enforced at the data service firewall.
- Backup: PITR with extended retention; geo-redundant backup storage in the paired region.
- High availability: Active geo-replication to the paired region. RPO ≤ 5 seconds; RTO ≤ 30 seconds.
- Access: Private endpoint only. Authentication via managed identity for service-to-service calls; legacy authentication disabled.

Document and graph state

Azure storage holds the Knowledge Graph traversal cache, the document-shaped artifacts (policies, evidence metadata, vendor questionnaire responses), and the high-velocity write-heavy workloads (audit log streams, integration ingestion buffers). Configured as a multi-region service with strong consistency for governance-critical writes.

- Encryption at rest: AES-256 with customer-managed keys.
- Encryption in transit: TLS 1.2 minimum; TLS 1.3 where supported.
- Replication: Multi-region writes enabled for the production region pair (e.g., Canada Central + Canada East).
- Backup: Continuous backup with PITR.
- Access: Private endpoint only. Authentication via managed identity.

Object and artifact storage

Azure storage holds the immutable evidence artifacts: configuration exports, log files, screenshots, attestation records, generated PDF reports. The hierarchical namespace and access control list semantics support per-control-folder access patterns and the regulatory immutability requirements (HIPAA Audit Controls, SOC 2 CC4, etc.).

- Encryption at rest: AES-256 with customer-managed keys.
- Encryption in transit: TLS 1.3 minimum.
- Replication: Read-Access Geo-Redundant Storage. Six copies across two regions.
- Versioning: Enabled. Soft-delete: 30 days. Immutable storage policies: configurable per container for regulatory holds.
- Access: Private endpoint only. Storage account key authentication disabled. Managed identity required.

Azure Key Vault — the secrets and key store

Key Vault holds three classes of material: (1) the customer-managed keys (CMKs) that encrypt the Azure storage services above, (2) the application secrets used by Kyūdō services (encrypted connection strings, integration tokens), (3) the certificate material for TLS termination.

- HSM-backed: Premium SKU with FIPS 140-2 Level 2 hardware security modules for key material.
- Access: Private endpoint only. Public network access disabled.
- Authentication: Managed identity for Kyūdō services; named principals for human operators with Conditional Access requiring phishing-resistant MFA.
- Audit: Every key access logged to Log Analytics; logs retained per the customer’s retention policy.
- Soft-delete and purge protection: Enabled. Secrets, keys, and certificates are recoverable for 90 days.

Customer-managed key topology

Customer-managed keys (CMK) are optional but recommended for regulated deployments. When CMK is provisioned, the encryption hierarchy is:

- Customer Key Vault holds the customer-managed key (RSA 2048 or larger; HSM-backed in Premium tier).
- Each Azure storage service is configured to use the CMK as the encryption key for the service’s data encryption keys.
- Key rotation is customer-controlled. Annual rotation is recommended baseline; quarterly is feasible.
- Key revocation is customer-controlled. If the CMK is revoked, the storage services become inaccessible — the customer holds the kill switch.

05 · THE IDENTITY LAYER

Microsoft Entra ID. Conditional Access. No shared secrets.

Identity is the architecture's most consequential plane. Kyūdō's identity model has two surfaces: human identity (administrators and users) and service identity (service-to-service and service-to-data-plane authentication). Both surfaces are anchored in Microsoft Entra ID; neither uses shared secrets in production.

Human identity

All human authentication to Kyūdō flows through the customer's Microsoft Entra ID tenant via OpenID Connect. The OIDC flow is the standard Authorization Code with PKCE; tokens are signed and validated against the customer tenant's discovery document. Kyūdō never holds user passwords; password material remains in the customer's Entra ID and never reaches the application layer.

Conditional Access

The customer's Conditional Access policies apply to Kyūdō access in the same way they apply to any other Entra-integrated application. Recommended baseline policies for Kyūdō administrative access:

- Require phishing-resistant MFA. FIDO2 security keys, Windows Hello for Business, certificate-based authentication, or platform passkeys. Push-notification MFA is not recommended for administrative roles.
- Require compliant device. The accessing device must be enrolled in Intune (or the customer's mobile device management) and must meet the security baseline.
- Geographic restrictions per the customer's policy. Block sign-ins from countries the customer does not operate in.
- Risk-based policies. Sign-in risk and user risk policies should treat any high-risk signal as requiring re-authentication or block.
- Privileged Identity Management (PIM) for the Kyūdō Tenant Admin role. Standing access to administrative scope is replaced with eligible-and-activate; activation is logged and time-bound.

RBAC mapping

Kyūdō's role-based access control maps to Entra ID security groups in the customer's tenant. Eight Kyūdō roles map to eight (or more, if customer-defined) security groups. The mapping is one-time configuration during onboarding; group membership changes propagate automatically on next sign-in.

Kyūdō role	Conventional Entra group name	Privilege scope
Tenant Admin	Kyudo_TenantAdmin	Tenant-wide; full administrative scope
Compliance Officer	Kyudo_ComplianceOfficer	Read/write controls + evidence; report generation
Policy Manager	Kyudo_PolicyManager	Read/write Policy Center; AI-author + approve
Risk Manager	Kyudo_RiskManager	Read/write Risk Management; risk treatment workflow
Vendor Risk Analyst	Kyudo_VRMAnalyst	Read/write VRM; vendor inventory + questionnaire
Auditor	Kyudo_Auditor	Read everything; annotation only; no modification
General User	Kyudo_User	Limited to assigned workflows and queues

Service identity

Service-to-service authentication uses Azure managed identities exclusively. There are no shared secrets in production deployments — no static API keys, no service account passwords, no bearer tokens with long lifetimes. The pattern:

- Compute-level system-assigned managed identity. Used for container registry pulls and platform-level Azure operations.
- Per-service user-assigned managed identities. Each Kyūdō service has its own identity bound at workload-level. Service X cannot impersonate service Y.
- Data plane access. Each service's managed identity is granted least-privilege access to only the storage services it needs (e.g., the Evidence Hub service has read/write on the evidence container but read-only on the Knowledge Graph relational state; the Policy Center service has read/write on the policy container but no access to the audit log).
- Key Vault access. Managed identity-based; access policies are bound to specific keys/secrets, not vault-wide.
- Azure OpenAI access. Managed identity-based; tokens are short-lived and per-call.

Identity boundary

The Entra ID tenant in question is the customer’s tenant, not KMicro’s tenant. Kyūdō operations personnel do not have standing access to the customer’s Entra ID. Where KMicro engineers require operational access (e.g., during a P1 incident under contract), access is granted via the customer’s Entra ID guest invitation flow with Conditional Access enforced and PIM eligibility revocable by the customer at any time.

Identity layer commitments

No KMicro standing credentials to the customer environment. No shared secrets. No service accounts with non-rotating passwords. Every authentication path — human or service — anchors in the customer’s Entra ID, with the customer holding the kill switch on every identity that can reach the platform.

06 · THE AI LAYER

Tenant-scoped inference. Cited outputs. Two-layer architecture.

Kyūdō's AI architecture is the most heavily scrutinized layer in any vendor risk assessment. Two architectural commitments together produce an AI substrate that is auditable, defensible, and — most importantly — sovereign.

Architectural commitment 1: Tenant-scoped inference

Every AI inference operation that processes customer compliance data executes against an Azure OpenAI Service deployment in the customer's Azure subscription. There is no centralized vendor-side LLM that aggregates governance posture across customers. There is no path for one customer's evidence to influence another customer's model behavior, because each customer has their own model deployment with their own context, their own audit logs, and their own content filtering.

The Azure OpenAI deployment uses Private Link; inference traffic stays on the Microsoft backbone. The model deployments support fine-tuning for the GRC domain; when fine-tuning is performed, the training data and resulting model weights remain in the customer tenant or in a tenant-isolated Azure-hosted environment. Inference context (the prompt, the retrieved Knowledge Graph subgraph, the conversation history) never leaves the customer's Azure subscription.

Architectural commitment 2: Two-layer reasoning

Kyūdō separates deterministic operations from advisory AI operations at the architectural layer. The split is not a UX convention; it is enforced in the data flow.

Layer	What it does	How it is auditable
Layer 1 — Deterministic	STRM mathematical relationships per NIST IR 8477; CMCAE control-completeness scoring; capability maturity calculation; RBAC decisions; Knowledge Graph traversal logic.	Rule-based and reproducible. Same inputs produce same outputs. Auditor sees a graph operation, not a model inference.
Layer 2 — Advisory AI	AI explanations of control state; policy authoring; questionnaire pre-fill; risk narrative generation; Sensei persona dialogue.	Confidence-scored on every output. Source-cited (every claim has a source in the Knowledge Graph). Logged with

Layer	What it does	How it is auditable
		prompt + retrieved context + output for replay. HITL threshold (0.7 default) for human review.

This separation is what makes Kyūdō auditor-defensible. The decisions that drive scoring, control state, and access control are deterministic and reproducible. The AI is advisory only; its outputs are confidence-scored, source-cited, and below 0.7 confidence are flagged for human review before they propagate.

Model Context Protocol (MCP)

Kyūdō uses Model Context Protocol to govern how the AI layer accesses tools, data, and external systems. MCP provides a structured, auditable interface between the LLM and the systems it can act on (or read from). Every MCP call is logged with the calling model, the tool invoked, the arguments passed, and the response received — the audit trail for AI behavior is itself a Knowledge Graph entity.

MCP is what allows Kyūdō's 27 Sensei AI personas to act with role-bounded scope: a Sensei mapped to the Vendor Risk Analyst RBAC role can invoke MCP tools for vendor questionnaire actions but cannot invoke MCP tools that would modify control state — the role boundary is enforced at the MCP layer, not just at the UI layer.

AI layer summary

Property	Implementation
Where inference runs	Customer's Azure subscription, via Private Link to Azure OpenAI Service
Model isolation	Per-customer model deployment; no cross-customer training or context
Cited outputs	Every AI claim grounded in Knowledge Graph; source citation on every claim
Confidence scoring	Every AI output carries a confidence score 0–1; below 0.7 flagged for HITL review
Auditability	Every prompt + retrieved context + output logged for replay
Tool boundary	MCP enforces role-bounded tool access; Sensei personas operate within RBAC scope

07 · SOVEREIGNTY PATTERNS

Four mechanisms. No exceptions.

The sovereignty claim Kyūdō makes is enforced through four architectural mechanisms. Each is verifiable independently. Together, they answer the question every regulated procurement reviewer asks: ‘what does sovereignty actually mean in this deployment.’

Mechanism 1: No cross-tenant data plane

There is no path in the architecture by which compliance data leaves the customer’s Azure subscription. The control plane (Kyūdō services), the data plane (Azure storage), the AI plane (Azure OpenAI), the secrets material (Key Vault), and the operational state (workflows, audit logs) all run inside the customer’s tenant boundary.

This is verifiable in two ways: (a) by inspecting the deployment template, which provisions every resource in the customer’s subscription with no cross-subscription references; (b) by examining the network egress rules at the Azure Firewall Premium, which deny outbound traffic to any KMicro-controlled endpoint as a baseline policy. Operational telemetry (uptime, latency, anonymized error counts) is the only data that can leave the tenant; even that is anonymized at source and is opt-out by configuration.

Mechanism 2: Customer-managed keys

Encryption keys for the data plane are held in the customer’s Key Vault. The customer controls key creation, rotation, and revocation. Revocation is a kill switch: revoking the CMK renders the encrypted Azure storage services inaccessible — to KMicro, to the customer, to anyone. The kill switch exists by design; it is the cryptographic enforcement of the sovereignty boundary.

Customer-managed keys are recommended baseline for regulated deployments. Where the customer chooses Microsoft-managed keys (the default Azure pattern), the encryption is still in place; the customer simply delegates key management to Microsoft’s platform-level key infrastructure rather than to their own Key Vault. The architectural boundary is the same; the trust boundary differs.

Mechanism 3: Anonymized operational telemetry

KMicro receives operational telemetry from Kyūdō deployments to support the platform: uptime, latency, error counts, AI confidence trend, model performance baselines. This telemetry is anonymized at source — stripped of customer identifiers, tenant IDs, control identifiers, and any data that could reveal a specific customer’s posture. The anonymization is performed in the customer tenant before any data leaves it.

The customer can opt out of operational telemetry entirely. Opt-out reduces KMicro’s ability to support the platform proactively (e.g., to detect a specific service trending toward SLO breach)

but does not affect product functionality. For Defense and Federal customers, opt-out is the recommended baseline.

Mechanism 4: Tenant-scoped AI inference

Every AI inference operation that processes customer compliance data runs in the customer's Azure subscription against the customer's Azure OpenAI Service deployment. There is no centralized vendor-side LLM. There is no cross-customer training. There is no inference context that aggregates governance posture across multiple customers. Each customer is, from the AI layer's perspective, the only customer.

The verification path

Each mechanism above is verifiable by the customer's security architect or auditor. The verification path is documented:

Mechanism	How to verify
No cross-tenant data plane	Inspect deployment template for cross-subscription references (none should exist). Inspect Azure Firewall egress rules for KMicro endpoint allowlist (none should exist for production data). Inspect network policies.
Customer-managed keys	Inspect Key Vault for the customer-managed key. Inspect each storage service's encryption configuration to confirm CMK reference. Test key revocation in a non-production environment.
Anonymized telemetry	Inspect the telemetry export destination (App Insights to Log Analytics; optional secondary export to KMicro).
Tenant-scoped AI	Inspect the Azure OpenAI deployment in the customer subscription. Inspect Private Link configuration. Inspect network policy for outbound paths to OpenAI endpoints.

Azure Policy initiative. Defender for Cloud baselines.

The architectural commitments above are not aspirational. They are enforced via the Azure Policy initiative `Kyūdō Baseline Compliance`, applied at the resource group or subscription scope where Kyūdō is deployed. The initiative's policies are deny-mode where deviation would compromise the architecture; audit-mode where the deviation is not architecturally fatal but is operationally significant.

The Kyūdō Baseline Compliance initiative

Sample policies in the initiative (the full initiative definition is in the Reference Runbook):

Policy	Mode	Rationale
Deny public endpoints on relational storage services	Deny	Sovereignty boundary: no access path that bypasses private endpoint
Deny public endpoints on object storage accounts	Deny	Same as above for artifact storage
Deny public endpoints on Azure OpenAI Service	Deny	AI inference path must remain on Microsoft backbone
Require Private Link for document and graph storage	Deny	Knowledge Graph cache and document state must not be publicly addressable
Deny non-managed-identity authentication on storage services	Deny	No shared secrets or static credentials
Require diagnostic settings to Log Analytics on every Azure resource	Audit	Operational visibility and audit trail
Require Defender for Cloud Standard tier	Audit	Continuous workload protection and posture management
Require compute control plane in private mode	Deny	Management endpoints not publicly resolvable
Require Key Vault soft-delete and purge protection	Deny	Secrets recoverability for incident response

Policy	Mode	Rationale
Require Customer-Managed Keys (where customer-elected)	Audit	Encryption sovereignty enforcement

Defender for Cloud baselines

Microsoft Defender for Cloud is enabled at the subscription level (Standard tier minimum) and covers the Kyūdō resource group. Defender for Cloud's built-in regulatory compliance assessments — NIST SP 800-53, ISO 27001, CIS Microsoft Azure Foundations, NIST SP 800-171 — produce the baseline against which the Kyūdō Controls Hub measures Azure infrastructure compliance. The integration with Kyūdō's Evidence Hub is described in the Quick Start Guide and the Microsoft integration documentation.

Policy enforcement at deployment time

The deployment template assigns the Kyūdō Baseline Compliance initiative to the deployed resource group as part of the deployment itself. Any subsequent resource creation that violates a deny-mode policy is rejected at the Resource Manager layer, before the resource exists. Audit-mode policy violations are surfaced in Defender for Cloud and in the Kyūdō Trust Center.

09 · SERVICING & LIFECYCLE

Release rings. Blue/green. Customer-controlled cadence.

Kyūdō's servicing model balances rapid security response with the change-management cycles regulated organizations require. The model has four mechanisms: release rings, blue/green deployment per service, monthly maintenance windows, and emergency CVE fast-track.

Release rings

Three rings, customer-elected at deployment time and changeable on request.

Ring	When updates arrive	When to choose this ring
Canary	Within 24 hours of KMicro release. First customer environments to receive each update.	Customers with high tolerance for change; technology-aligned organizations; non-production environments.
Pilot	Within 7–10 days of KMicro release. After Canary has validated.	Production environments where rapid feature adoption is valuable; mid-market customers.
Broad	Within 21–30 days of KMicro release. After Canary and Pilot have validated.	Regulated production environments; customers with strict change-management cycles.

Customers can hold updates indefinitely. The deployment will not auto-update past a customer-set hold; the customer's change-management process is sovereign over the deployment cadence. Held updates accumulate in a release queue and are applied when the customer schedules them.

Blue/green deployment per service

Each Kyūdō service deploys with a blue/green pattern: the new version is deployed alongside the running version, traffic is gradually shifted, and the previous version is retained for rapid rollback. This produces zero-downtime updates for any single service. Service-package revision history is the rollback mechanism; rollback is a single command and completes in seconds.

Monthly maintenance windows

Standard maintenance occurs in customer-elected monthly windows. Typical patterns: second Tuesday of the month for Pilot ring; fourth Tuesday for Broad ring. The window is 2–4 hours; updates that exceed the window are deferred to the next cycle. The customer can change the

window at any time; emergency updates (CVE fast-track) operate independently of the maintenance window cadence.

Emergency CVE fast-track

Critical security CVEs that affect Kyūdō services, the Azure storage services, or the compute layer itself trigger the emergency fast-track process. The target SLA is 48 hours from KMicro's receipt of the CVE notification to deployment in customer environments — dependent on the customer's ring election and any holds in place.

CVE fast-track deployments are validated through the same blue/green mechanism as standard updates, with automated pre-deployment validation (test suite, security scan, regression check) and automated rollback if any validation step fails. The customer is notified of every CVE fast-track deployment with the CVE identifier, the affected component, the fix description, and the deployment timestamp.

CI/CD pipeline

Kyūdō's build and deployment pipeline runs in KMicro's subscription (not the customer's). The pipeline produces signed container images and service packages; signing keys are HSM-backed. The customer's deployment pulls signed artifacts from the customer's container registry (which mirrors from KMicro's public registry under contract); image signature validation at admission rejects any unsigned or tampered image.

The CI/CD pipeline produces a Software Bill of Materials (SBOM) for every release, signed and attached to the release artifact. The SBOM is the customer's artifact for vulnerability tracking, license compliance, and supply-chain attestation.

10 · THE RACI MATRIX

KMicro versus the customer. Layer by layer.

The client-hosted topology requires a clear allocation of responsibility between KMicro (the platform vendor) and the customer (the deployment owner). The RACI matrix below is the canonical allocation; specific contracts may adjust it, but this is the default.

Capability	KMicro	Customer	Notes
Landing zone & policy guardrails	Consult	Accountable + Responsible	Customer owns the Azure subscription and tenant-level governance
Kyūdō app deployment (Managed App / template)	Responsible	Accountable	KMicro provides the deployment template; customer triggers and approves
Identity (Entra, Conditional Access)	Consult	Accountable + Responsible	Customer's Entra ID; customer's Conditional Access policies
Networking (VNets, Firewall, Private Link)	Consult	Accountable + Responsible	Customer owns the network topology; KMicro consults on Kyūdō-specific requirements
Azure storage services	Consult	Accountable + Responsible	All storage services in customer subscription
Secrets (Key Vault)	Consult	Accountable + Responsible	Customer's Key Vault; customer-managed keys when elected
AI endpoints (Azure OpenAI)	Consult	Accountable + Responsible	Azure OpenAI deployment in customer subscription
App servicing (images, packages, config)	Responsible	Accountable	KMicro builds; customer deploys per ring election
Security patching (Kyūdō services)	Responsible	Accountable	KMicro produces the patches; customer applies
Platform patching	Consult	Accountable + Responsible	Standard Azure shared-responsibility model for platform OS

Capability	KMicro	Customer	Notes
Monitoring & alerts	Responsible	Accountable	KMicro operates the Kyūdō monitoring stack; customer owns the alerting destinations
Incident management (P1–P3)	Responsible	Accountable	KMicro investigates Kyūdō-side; customer drives resolution per their incident process
Compliance evidence & attestations	Consult	Accountable + Responsible	KMicro produces SOC 2 etc. for the platform itself; customer produces evidence for their environment

Source — Kyūdō Client-Hosted Azure Reference Architecture & Servicing Runbook v1.0, Section 8.

What this means in practice

In every regulated organization's vendor risk assessment, the question of 'where does the line sit' between vendor and customer responsibility is the question that determines whether the deployment is auditor-defensible. The matrix above resolves the question explicitly. The customer is accountable for the Azure substrate (subscription, tenant, network, identity, storage services, keys); KMicro is responsible for the Kyūdō application layer that runs on top of it. The boundary is bright; the responsibility is allocated; the audit trail follows.

Template-based deployment. Post-deployment binding.

The deployment template is the artifact for the entire Kyūdō client-hosted topology. A single deployment invocation provisions every component above with the architectural commitments enforced. This section summarizes the deployment shape; the full template and parameter reference are in the Reference Runbook.

Deployment quick-start

The five-step deployment sequence:

1. Authenticate to Azure CLI and set the target subscription. The customer's platform admin executes this step; KMicro never holds the customer's subscription credentials.
2. Create the resource group in the customer-elected primary region. Region selection drives data residency; primary + DR region pair defaults to a Microsoft-paired region set per customer policy.
3. Deploy the Kyūdō baseline template, parameterized for region, environment, compute sizing, and network ranges. Deployment completes in 30–60 minutes.
4. Retrieve outputs from the deployment. Outputs include managed-identity resource IDs, private DNS zone bindings, and the platform connection metadata.
5. Post-deployment binding: bind Private Endpoints if not auto-created; configure Conditional Access policies for Kyūdō administrative access; register Entra ID groups to Kyūdō RBAC; onboard Sentinel analytics and dashboards.

Source: Kyūdō Reference Runbook v1.0, Appendix D.

Post-deployment validation

After deployment, the platform admin validates the deployment against the SLO targets in Appendix A and against the Azure Policy initiative compliance state. Both validations are scripted in the Reference Runbook; both should pass at zero violations before the platform is declared production-ready.

Operations day-to-day

Once operational, day-to-day operations follow the standard Azure operations model. Application Insights and Log Analytics surface platform telemetry. Defender for Cloud surfaces posture and threat findings. Microsoft Sentinel correlates security events. Action Groups route alerts to the customer's notification channels (Teams, email, ticketing). KMicro Customer Success provides escalation paths for any issue that exceeds the customer's in-house diagnostic capability.

Backup, disaster recovery, business continuity

The data plane services have native Azure backup and DR mechanisms; the deployment template configures them per the architectural targets:

- Relational state: PITR with extended retention; active geo-replication to paired region; RPO ≤ 5 seconds; RTO ≤ 30 seconds.
- Document and graph state: Continuous backup with PITR; multi-region writes; RPO ≈ 0; RTO seconds.
- Object and artifact storage: Versioning + soft-delete + Read-Access Geo-Redundant Storage; recoverable artifacts within 30 days; six replicas across two regions.
- Service packages: Per-service revision history allows rollback to any previous release.
- Azure Backup: Platform-level backup for persistent volumes when the workload pattern requires it.

Disaster recovery runbooks are in the Reference Runbook. The architectural targets are 99.99% uptime and less than 1% blast radius per service fault — both of which are achievable within the Azure SLA envelope when the deployment follows the reference architecture.

APPENDIX A · SLOS AND OPERATIONAL TARGETS

Service-level objectives. Platform performance baselines.

The SLOs below are the Kyūdō platform's production performance baselines. Customer deployments are validated against these targets at deployment time and are continuously monitored against them through Application Insights and Log Analytics. Deviations from baseline are surfaced in the Kyūdō Trust Center.

SLO	Target	Measurement
API P95 latency	≤ 500 ms	Application Insights, server-side, all API endpoints
Evidence ingest latency	≤ 2 seconds	Time from integration signal to evidence-registered event
Light-RAG P95 latency	≤ 500 ms	AI inference endpoint, end-to-end including retrieval
HITL threshold	0.7 confidence	AI confidence below threshold flags for human review
Uptime	99.99%	Per-month, measured at API surface
Blast radius per service fault	< 1% of deployment	Service isolation boundary
RPO (Recovery Point Objective)	≤ 5 seconds (relational) / ≈ 0 (document/graph)	Cross-region replication lag
RTO (Recovery Time Objective)	≤ 30 seconds (relational) / seconds (document/graph)	Failover invocation to availability

Source — Kyūdō Reference Runbook v1.0, Appendix A; SRS v7.

How SLOs translate to user experience

API P95 latency ≤ 500 ms means the user interface feels responsive: clicking into a control, opening an evidence artifact, switching between modules complete in under half a second for 95% of operations. Evidence ingest latency ≤ 2 seconds means a Microsoft signal (e.g., a new Defender for Cloud finding) is registered as evidence in the Knowledge Graph within 2 seconds of the source signal.

Light-RAG P95 latency ≤ 500 ms is the AI-side performance commitment: AI explanations, policy authoring, questionnaire pre-fill, and Sensei dialogues all complete within half a second on average. The HITL threshold of 0.7 means any AI output below 70% confidence is flagged for human review before propagating — the safety net auditors most fear in AI-assisted compliance evidence.

APPENDIX B · AZURE POLICY BASELINE

Sample policies in the Kyūdō Baseline Compliance initiative.

The sample policies below are illustrative of the Kyūdō Baseline Compliance Azure Policy initiative. The full initiative definition (with all policies, parameters, and remediation tasks) is in the Reference Runbook (Appendix B). Policies are deny-mode where deviation would compromise the architecture; audit-mode where the deviation is operationally significant but not architecturally fatal.

Network and exposure

- `Deny-PublicEndpoints-Relational` (Deny). Blocks creation of relational storage resources with public network access enabled.
- `Deny-PublicEndpoints-Object` (Deny). Blocks creation of object storage accounts with public network access enabled.
- `Deny-PublicEndpoints-OpenAI` (Deny). Blocks creation of Azure OpenAI Service resources with public network access enabled.
- `Deny-PublicEndpoints-Document` (Deny). Blocks creation of document and graph storage accounts with public network access enabled.
- `Deny-PublicEndpoints-KeyVault` (Deny). Blocks creation of Key Vaults with public network access enabled.
- `Require-Compute-PrivateMode` (Deny). Blocks creation of compute resources where the management endpoint is not in private mode.

Identity and authentication

- `Deny-LocalAuth-Storage` (Deny). Blocks storage configurations that allow local authentication; Entra ID-only authentication required.
- `Deny-StorageKeyAuth` (Deny). Blocks storage account configurations that allow shared key authentication; managed identity required.
- `Require-Document-ManagedIdentity` (Deny). Requires document storage authentication via managed identity; legacy keys disabled.
- `Require-KeyVaultRBAC` (Audit). Audits Key Vaults using legacy access policies rather than Azure RBAC; remediation task migrates to RBAC.

Encryption and data protection

- `Require-CMK-Relational` (Audit). Audits relational storage resources not using customer-managed keys; recommendation generated for remediation.
- `Require-CMK-Object` (Audit). Audits object storage accounts not using customer-managed keys.

- `Require-CMK-Document` (Audit). Audits document storage accounts not using customer-managed keys.
- `Require-TLS-Min12` (Deny). Blocks resources accepting TLS 1.1 or earlier.

Observability and audit

- `Require-DiagnosticSettings-LogAnalytics` (Audit). Audits resources without diagnostic settings to Log Analytics; remediation task creates the diagnostic setting.
- `Require-Defender-Standard` (Audit). Audits subscriptions where Defender for Cloud Standard tier is not enabled.
- `Require-Sentinel-Connection` (Audit). Audits Log Analytics workspaces not connected to Microsoft Sentinel.

Backup and recovery

- `Require-Storage-PITR` (Audit). Audits storage resources without point-in-time restore enabled.
- `Require-Storage-Versioning` (Audit). Audits storage accounts without blob versioning enabled.
- `Require-KeyVault-PurgeProtection` (Deny). Blocks Key Vaults without soft-delete and purge protection enabled.

Source — Kyūdō Reference Runbook v1.0, Appendix B; full initiative definition available on engagement.

APPENDIX C · GLOSSARY

Architectural terms used in this guide.

Term	Definition
AMPLS	Azure Monitor Private Link Scope. The mechanism that ensures Application Insights and Log Analytics telemetry traverses Private Link rather than the public internet.
Application Gateway	Azure's Layer 7 load balancer with WAF; sits between Front Door and the application ingress in the Kyūdō topology.
Azure Firewall Premium	The egress and east-west firewall in the Kyūdō hub VNet. Premium tier provides TLS inspection, IDPS, and URL filtering.
Azure storage	The Azure-native storage services that hold Kyūdō state. Specific services are documented in the Reference Runbook; all are accessible only via private endpoint and encrypted with customer-managed keys when elected.
Blue/green deployment	A deployment pattern where two environments (blue = current, green = new) run simultaneously and traffic is shifted between them. Used per Kyūdō service for zero-downtime updates.
CMCAE	Continuous Multi-Framework Control Assessment Engine. The Kyūdō engine that recalculates control completeness, capability maturity levels, and residual risk on every signal change.
CMK	Customer-Managed Key. Encryption key held in the customer's Key Vault rather than in Microsoft-platform-managed key infrastructure. The cryptographic enforcement of the sovereignty boundary.
Conditional Access	Microsoft Entra ID policies that govern when and how users can sign in (device compliance, MFA method, geographic restrictions).
Confidence score	AI-computed score on every Kyūdō output. Indicates how strongly the source signals support the conclusion. Below the human-in-the-loop threshold of 0.7, outputs are flagged for human review.
Front Door	Azure's global HTTP load balancer with WAF; the global entry point for Kyūdō web UI and API surface.
HITL threshold	Human-in-the-Loop. The confidence threshold (0.7 default) below which AI-produced outputs are flagged for human review before propagating.

Term	Definition
Hub-and-spoke	Network topology pattern where a hub VNet hosts shared services (firewall, gateway, identity) and spoke VNets host workload services. The Kyūdō deployment topology.
MCP	Model Context Protocol. The structured, auditable interface between LLMs and external tools/data. The mechanism by which Kyūdō AI personas operate within RBAC scope.
NIST IR 8477	Mapping Relationships Between Documentary Standards, Regulations, Frameworks, and Guidelines. The NIST Internal Report that formalizes Set Theory Relationship Mapping (STRM).
PIM	Privileged Identity Management. Azure feature that replaces standing privileged access with eligible-and-activate; activation is logged and time-bound.
Private endpoint	Azure mechanism for service-to-service connectivity via private IP addresses on the Microsoft backbone, never traversing the public internet. The default Kyūdō networking pattern for every Azure service.
Private Link	The Azure feature family that includes Private Endpoint and Private Link Service. Underlies the private-by-default networking in Kyūdō.
Release ring	Servicing model that groups customers into update-cadence tiers (Canary, Pilot, Broad). Customers elect their ring; updates propagate through rings in sequence.
RPO / RTO	Recovery Point Objective (data loss tolerance) and Recovery Time Objective (downtime tolerance). The Kyūdō architectural targets are RPO ≤ 5 seconds and RTO ≤ 30 seconds for relational state.
SBOM	Software Bill of Materials. The signed inventory of every component in a Kyūdō release. Produced for every release and attached to the release artifact.
SCF	Secure Controls Framework. The meta-framework substrate that anchors the Kyūdō Knowledge Graph; over 1,470 controls across 80+ frameworks.
SLO	Service Level Objective. The Kyūdō platform SLOs: API P95 ≤ 500 ms; evidence ingest ≤ 2 seconds; Light-RAG P95 ≤ 500 ms; HITL threshold 0.7; 99.99% uptime; <1% blast radius.
STRM	Set Theory Relationship Mapping (per NIST IR 8477). The mathematical substrate that makes one Kyūdō control set satisfy 70+ frameworks simultaneously.

Term	Definition
Workload identity	Azure feature that binds workloads to Azure managed identities. Used in Kyūdō for per-service identity binding without shared secrets.

© 2026 KMicro Technologies, Inc. Kyūdō and the Kyūdō logo are trademarks of KMicro Technologies, Inc. Microsoft, Azure, Microsoft 365, Entra ID, Defender, Sentinel, and Purview are trademarks of Microsoft Corporation. This Architecture Guide is published by Kyūdō, kyudo.ai, for the use of security architects, cloud architects, network engineers, platform engineers, and procurement reviewers evaluating Kyūdō's client-hosted Azure deployment topology. Architectural specifics are accurate as of April 2026 and align with the Kyūdō Client-Hosted Azure Reference Architecture & Servicing Runbook v1.0; the Reference Runbook is the authoritative document for engaged customers and is provided on engagement. Contact hello@kyudo.ai or your Customer Success engineer for the current state of any specific architectural commitment.