إنتركي
INTERKEY

White paper

Interkey Professional Services for Kong

API Gateway, AI Gateway, Service Mesh and Insomnia

Prepared by
Interkey

Date
September 2026

From Vision to Reality

Executive summary

Interkey delivers Kong as a governed, production-grade connectivity platform in four milestones: Discovery, POC, Production and Scale & Hypercare. Every milestone has a fixed scope, named deliverables and an exit gate the client signs.

Enterprises in the GCC now run three kinds of traffic that each need control: external and partner APIs, service-to-service calls inside Kubernetes, and a fast-growing volume of calls to large language models. Most organisations govern these with separate tools, or not at all. Kong covers all three from one control plane: Kong Gateway for APIs, Kong AI Gateway for LLM traffic, Kong Mesh for east-west service traffic, and Insomnia for API design and testing, all managed through Kong Konnect.

The software is only half the outcome. Interkey supplies the architecture, implementation, security hardening, regulatory mapping (NCA ECC, SAMA, PDPL) and team enablement that turn a Kong licence into a platform the client's teams can run. This paper sets out the service catalog, delivery method, prerequisites, value statements and packaged offerings, using the same milestone-based model Interkey applies to its Datadog observability engagements.

Why now

Four pressures push GCC enterprises to treat connectivity as a governed platform rather than a set of point tools.

- **API sprawl.** Digital channels, open banking, government integrations (Nafath, Absher, ZATCA e-invoicing) and partner ecosystems multiply APIs faster than teams can document, secure and version them.
- **Ungoverned AI traffic.** Teams call OpenAI, Azure OpenAI, Bedrock, Gemini and self-hosted models directly, with keys in code, no token budgets, no PII controls and no view of spend. Agentic AI adds MCP servers and agent-to-agent calls to the same problem.
- **East-west zero trust.** Microservices on OpenShift, EKS, AKS and VMs talk to each other in clear text with implicit trust. NCA ECC expects encryption in transit and least-privilege access inside the network, not only at the edge.
- **Developer productivity.** Without a shared design-first workflow, API contracts drift from what the gateway actually enforces, and testing happens late.

Kong addresses all four from one platform. The gap most clients face is not the product but the delivery capacity to design, deploy, harden and operate it. That is the gap Interkey fills.

The Kong platform

Kong is five products on one control plane; Interkey scopes each one separately so clients can start with one and add the rest.

Product	What it does	Key capabilities (2026)	Where Interkey adds value
Kong Gateway	North-south API gateway for REST, GraphQL, gRPC, WebSocket and event APIs	Plugins for auth (OIDC, mTLS, JWT, key auth), rate limiting, transformations, OpenTelemetry; hybrid or DB-less deployment	Topology, plugin policy design, CI/CD with decK, HA/DR, migration from Apigee, Azure APIM, 3scale, WSO2, NGINX
Kong AI Gateway	Governs traffic from apps and agents to LLMs, MCP servers and other agents	Multi-LLM routing and load balancing, semantic caching, semantic prompt guards, PII sanitisation, token quotas per user/model, MCP auth and auto-generated MCP servers, A2A governance, token spend dashboards; 3.14 adds dynamic model routing, RFC 8693 token exchange, DeepSeek, Databricks and vLLM	LLM provider strategy, guardrail and PII policy, token budget and chargeback model, MCP exposure of existing APIs, self-hosted model routing
Kong Mesh	Enterprise service mesh on Kuma and Envoy for east-west traffic	Automatic mTLS, traffic permissions, L4/L7 routing, circuit breakers, retries, multi-zone across clusters and VMs, RBAC, audit logging, SPIFFE identity	Zero-trust design, zone topology, sidecar onboarding plan, policy-as-code, observability wiring
Insomnia	API design, debugging, testing and mocking client	Design-first OpenAPI editing, collections, mocks, custom linting, native Git sync; Insomnia 13 (June 2026) connects directly to Konnect to test live gateway config	API style guide and lint rules, contract testing in CI with inso CLI, developer enablement

Product	What it does	Key capabilities (2026)	Where Interkey adds value
Kong Konnect	SaaS control plane for all of the above	Gateway and mesh management, service catalog, developer portal, analytics, Dedicated Cloud Gateways; ME geo available	Region and residency design, RBAC and teams, portal setup, analytics and chargeback

Reference architecture

Interkey's standard design separates one control plane from data planes that run where the client's traffic and data must stay.

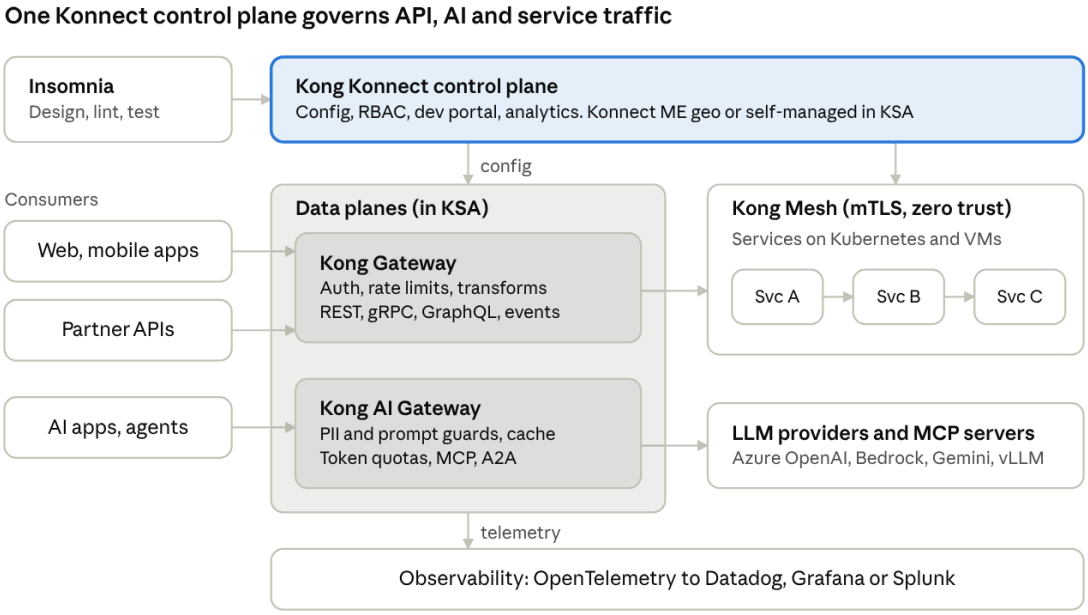


Figure 1. Kong reference architecture · control plane, data planes, mesh

Apps and partners enter through Kong Gateway; AI apps and agents reach LLMs and MCP servers only through Kong AI Gateway; services talk to each other over mTLS inside Kong Mesh. Insomnia tests against the live Konnect configuration, and all telemetry flows to the client's observability platform.

Interkey service catalog

Each service line has a fixed scope and named deliverables, so it can be quoted, delivered and accepted on its own.

Service line	Scope	Deliverables
API platform assessment	Inventory of APIs, gateways and consumers; security posture review; maturity score; target-state design	Current-state report, API inventory, maturity scorecard, target architecture, roadmap
Kong Gateway implementation	Hybrid or self-managed install; control plane and data plane topology; plugin standards (auth, rate limiting, logging); HA/DR across two sites; GitOps with decK and Terraform	High-level and low-level design, installed platform, plugin catalogue, runbooks, CI/CD pipeline
Gateway migration	Move APIs from Apigee, Azure APIM, 3scale, WSO2, Mulesoft or NGINX; policy mapping; cutover and rollback plans	Migration factory, policy mapping matrix, per-wave cutover plans, decommission checklist
Kong AI Gateway enablement	LLM provider onboarding (Azure OpenAI, Bedrock, Gemini, OpenAI, self-hosted vLLM); prompt guard and PII policies; semantic cache; token quotas per team; MCP server exposure of existing APIs	AI traffic policy set, guardrail catalogue, token budget and chargeback model, AI usage dashboard
Kong Mesh zero trust	Mesh topology (single or multi-zone); mTLS rollout; traffic permissions by service; resilience policies; VM and Kubernetes onboarding	Zero-trust design, mesh policy repository, onboarding waves, service dependency map
Insomnia and API-first enablement	API style guide; Spectral lint rules; design-first workflow; contract tests with inso CLI in CI; mock servers	API standards handbook, lint ruleset, CI test templates, trained developers
Developer portal	Konnect Dev Portal setup; API products and documentation; self-service app registration	Live portal, API product catalogue, onboarding guide
Observability integration	Export gateway, AI and mesh telemetry via OpenTelemetry and Prometheus to Datadog, Grafana or Splunk	Dashboards, SLOs, alerts for latency, errors, token spend

Service line	Scope	Deliverables
Managed support	Upgrades, plugin lifecycle, capacity reviews, incident support under SLA	Monthly health report, upgrade plan, quarterly roadmap review

The observability line links this offer to Interkey's Datadog practice: a client running both gets API, AI and mesh telemetry in the same dashboards as infrastructure and APM.

Delivery methodology

Interkey delivers Kong in four milestones, the same model used for its Datadog engagements; no milestone starts until the previous gate is signed.

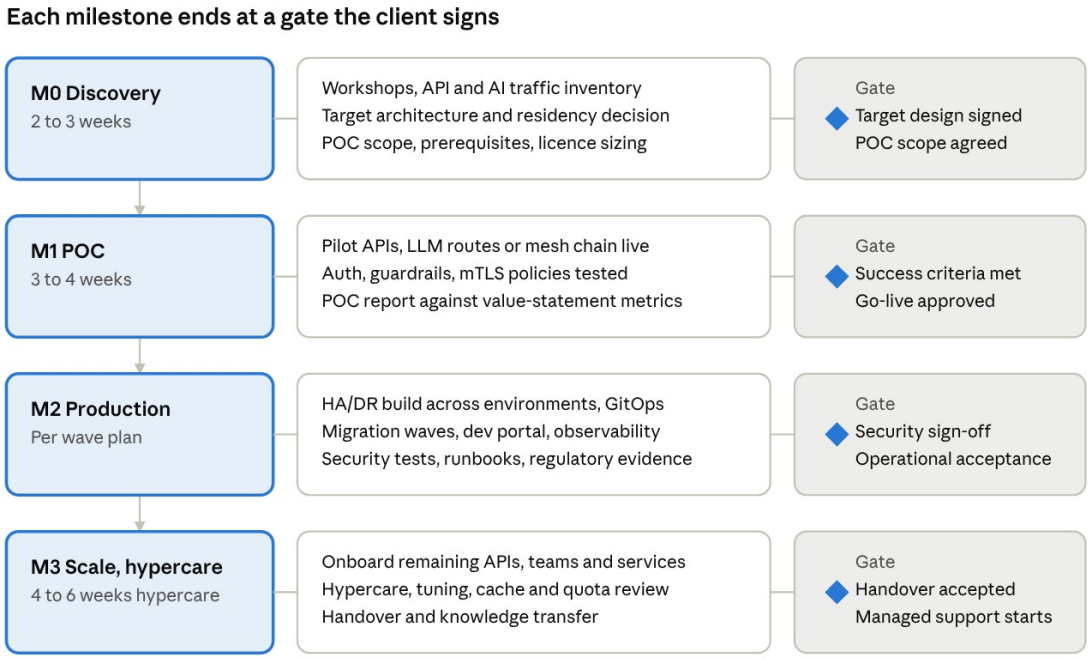


Figure 2. Delivery roadmap · 4 milestones, 4 gates

The POC proves value on a small, representative scope before the client commits to production licensing and rollout. Each gate is measured against the value statements and success metrics below.

Technical prerequisites

The client provides these before the POC starts; missing items move the POC start date, not the scope.

Area	Prerequisite	Needed for	Owner
Licensing	Kong Konnect or Kong Enterprise trial/licence; Insomnia Enterprise seats if SSO or Git sync is in scope	All	Client procurement
Compute	Kubernetes cluster (OpenShift, EKS, AKS, GKE or vanilla, v1.28+) or Linux VMs (RHEL/Ubuntu); 3 nodes minimum for POC data planes	Gateway, Mesh	Client platform team
Database	PostgreSQL 14+ for self-managed control plane (not needed for Konnect or DB-less)	Gateway (self-managed)	Client DBA
Network	Outbound 443 from data planes to Konnect ME geo (hybrid); internal load balancer; DNS entries; firewall rules to upstream services	Gateway, Mesh	Client network team
Certificates	TLS certificates for gateway hostnames; internal CA or Vault for mesh mTLS	Gateway, Mesh	Client security team
Identity	IdP for OIDC (Entra ID, Keycloak, Ping, Okta); test client IDs; SSO for Konnect and Insomnia	Gateway, Insomnia	Client IAM team
LLM access	API keys or endpoints for target LLM providers; vector store (Redis/Valkey) for semantic cache	AI Gateway	Client AI team
Pilot APIs	3 to 5 representative APIs with OpenAPI specs and test data; one internal service chain for mesh	All	Client application teams
Tooling	Git repository and CI/CD runner (GitLab, GitHub, Azure DevOps, Jenkins)	GitOps, Insomnia CI	Client DevOps
Observability	Target platform (Datadog, Grafana, Splunk) and credentials	Observability	Client operations
Access	VPN or jump host access for Interkey engineers; named client contacts per team	All	Client PMO

Value statements and success metrics

Each value statement has a metric the client baselines in Discovery and measures at the POC and Production gates; targets are agreed per client, not promised in advance.

Value statement	Success metric	Measured at
One place to secure and publish every API	% of APIs behind Kong with standard auth and rate limiting	POC, Production
Faster API delivery	Lead time from approved spec to published, secured API	Production
Controlled, visible AI spend	Token spend per team and model visible; % of LLM calls routed through AI Gateway	POC, Production
Lower LLM cost and latency	Semantic cache hit rate; average tokens per request	Scale
No sensitive data sent to external LLMs	PII detections blocked or redacted; prompt guard violations logged	POC, Production
Zero trust inside the cluster	% of service-to-service traffic on mTLS; % of services with explicit traffic permissions	Production, Scale
Fewer outages from cascading failures	Incidents where retries or circuit breakers contained the fault	Scale
API contracts that match reality	% of APIs with linted specs and contract tests in CI	Production
Audit-ready evidence	Controls in the regulatory mapping with evidence produced by the platform	Production

GCC regulatory mapping

Kong produces much of the technical evidence regulators ask for; Interkey maps each control to a Kong capability during design and leaves the client with the evidence trail.

Requirement area	Regulation	Kong capability	Interkey activity
Encryption in transit	NCA ECC, SAMA CSF, PCI DSS	TLS at the gateway; automatic mTLS in Kong Mesh	Certificate and CA design, mTLS rollout plan

Requirement area	Regulation	Kong capability	Interkey activity
Identity and access control	NCA ECC, SAMA CSF	OIDC, OAuth2, mTLS client auth, ACLs; mesh traffic permissions; Konnect RBAC	Access model per API and service, least-privilege policies
Logging and monitoring	NCA ECC, SAMA CSF, PCI DSS	Gateway and mesh access logs, audit logs, OpenTelemetry export	Log forwarding to SIEM, retention settings, alert rules
Personal data protection	PDPL (SDAIA)	AI Gateway PII sanitisation; request/response transformation to mask fields	PII policy for AI and API traffic, data flow register
Data residency	PDPL, NCA CCC, SAMA outsourcing rules	Hybrid mode keeps runtime traffic and data on data planes hosted in KSA; fully self-managed option on-premises	Residency decision record (see below)
Responsible AI	SDAIA AI Ethics Principles	Prompt guards, audit of AI calls, token quotas, model allow-lists	AI usage policy mapped to gateway controls
Open banking	SAMA Open Banking Framework	FAPI-style OAuth2/OIDC, consent-aware routing, developer portal	Open banking API gateway pattern

Data residency note.

The Konnect Middle East geo is hosted in the UAE (AWS me-central-1, Azure UAE North), per Kong's documentation. In hybrid mode, API traffic and payloads stay on data planes the client runs in KSA; only configuration and telemetry metadata reach the control plane. Where a regulator requires the control plane in-Kingdom too, Interkey deploys self-managed Kong Enterprise and Kong Mesh on the client's infrastructure or a KSA cloud region. This is a design decision the client's compliance team signs at the Discovery gate. This mapping is technical guidance, not legal advice.

Engagement models and packaged offerings

Clients can buy a fixed-scope package, a milestone-based implementation, or ongoing managed support; durations below are indicative and confirmed after Discovery.

Package	Scope	Indicative duration	Commercial model
Kong Readiness Assessment	API and AI traffic inventory, maturity score, target architecture, business case	2 to 3 weeks	Fixed price
Kong Gateway POC	One environment, 3 to 5 pilot APIs, standard plugins, OIDC, CI/CD with decK	3 to 4 weeks	Fixed price
AI Gateway Quick Start	2 LLM providers, PII and prompt guard policies, token quotas, semantic cache, usage dashboard	3 to 4 weeks	Fixed price
Mesh Zero-Trust Pilot	One cluster or zone, one service chain on mTLS with traffic permissions	3 to 4 weeks	Fixed price
Platform Implementation	Production rollout of the chosen products across environments, HA/DR, migration waves	Per milestone plan	Milestone-based
API-First Enablement	Style guide, lint rules, Insomnia workflows, training for developers	2 weeks plus coaching	Fixed price
Managed Kong Services	Upgrades, capacity, incident support, quarterly reviews	12-month term	Monthly retainer with SLA

Licensing considerations

- **Konnect vs self-managed.** Konnect (SaaS control plane) removes control plane operations; self-managed Kong Enterprise suits strict in-Kingdom residency. The choice affects licence type and infrastructure cost.
- **Metering.** Kong prices by gateway services, API requests and product add-ons; AI Gateway, Mesh and Insomnia Enterprise are licensed separately. Interkey sizes volumes in Discovery so the licence quote matches real traffic.
- **Environments.** Non-production data planes and mesh zones count toward licensing; confirm dev, test and DR coverage before signing.
- **LLM costs are separate.** AI Gateway governs spend on OpenAI, Azure OpenAI, Bedrock and other providers but does not include it.

- **Open source option.** Kong Gateway OSS and Kuma can serve a low-cost pilot, but lack enterprise plugins, RBAC, the AI Gateway enterprise features and vendor support.

Final licensing terms come from Kong or its authorised distributor; Interkey provides the sizing inputs.

Discovery questionnaire

Interkey sends these questions before the Discovery workshop; the answers size the POC, the licence and the milestone plan.

Area	Question	Client answer
Business	Which outcomes matter most: security, API monetisation, AI governance, developer speed, regulatory evidence?	
Business	Which regulators and frameworks apply (NCA ECC/CCC, SAMA, PDPL, PCI DSS, CST)?	
APIs	How many APIs are in production, internal vs external, and in which protocols (REST, SOAP, GraphQL, gRPC, Kafka)?	
APIs	What peak and average request volumes do you see per month?	
APIs	Which gateway or ESB is in place today, and when does its licence renew?	
AI	Which LLM providers and models are in use or planned, including self-hosted?	
AI	Do you have MCP servers or AI agents in use or planned?	
AI	Is there an AI usage policy covering PII, approved models and budgets?	
Services	How many microservices run on Kubernetes vs VMs, across how many clusters and sites?	
Services	Is service-to-service traffic encrypted today?	
Platform	Which Kubernetes distribution and cloud or on-prem hosting do you use?	

Area	Question	Client answer
Platform	Must the control plane be hosted in KSA, or is the Konnect ME geo acceptable with in-Kingdom data planes?	
Security	Which IdP, certificate authority and secrets store (Vault, CyberArk, cloud KMS) do you use?	
Security	Which SIEM receives security logs?	
Developer	How are API specs written, reviewed and tested today? Is Postman or Insomnia in use?	
Developer	Which CI/CD and Git platforms do you use?	
Operations	Which monitoring platform do you use (Datadog, Grafana, Splunk, Dynatrace)?	
Operations	Who will own the platform after go-live, and how many engineers are in that team?	

Why Interkey

- **Enterprise data and integration depth.** Interkey's architects have designed databases, event-driven platforms and integration layers for GCC banks, insurers, retailers and government entities, so Kong is placed in the context of the whole data flow, not only the gateway.
- **One delivery method across platforms.** The same milestone, POC, prerequisite and value-statement model Interkey uses for Datadog observability, so clients see consistent governance across engagements.
- **Connectivity plus observability.** Kong telemetry lands in the same Datadog dashboards Interkey builds for infrastructure, APM and databases.
- **Local regulatory context.** Designs are mapped to NCA, SAMA and PDPL from day one, with a signed residency decision.
- **Knowledge transfer built in.** Every milestone ends with runbooks and hands-on sessions, so the client's team can run the platform without Interkey.

Next steps

1. Return the discovery questionnaire.
2. Hold a half-day Discovery workshop with architecture, security, AI and platform leads.
3. Interkey issues the Readiness Assessment report, POC scope, prerequisites checklist and licence sizing.
4. Start the POC on the agreed product (Gateway, AI Gateway or Mesh).

Sources

- Kong AI Gateway product page
- Kong AI Gateway 3.14 release (April 2026)
- Kong Gateway 3.13 release
- Kong Mesh documentation
- Insomnia 13 release (June 2026)
- Konnect geographic regions