Internal / appendix-ready

Goodland Cloud Data Service

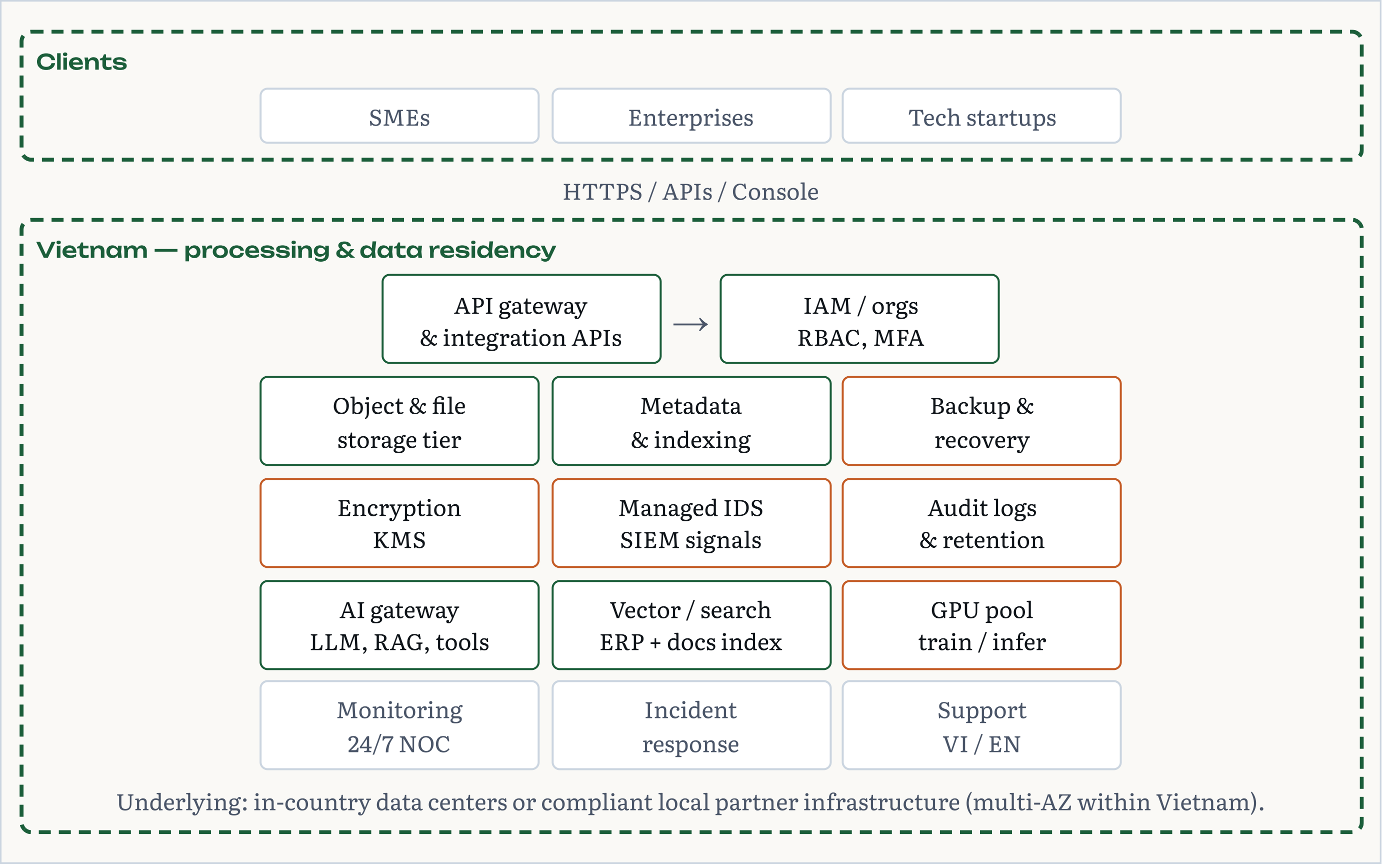
Technical architecture overview & indicative pricing
Vietnam market entry · Data residency · Storage, backup, IDS, compliance · ERP AI · GPU rental

This document supports the business plan appendix. Diagrams describe a target-state reference architecture. Pricing is illustrative until vendor and tax positions are finalized.

1. Executive alignment

- **Objective:** Secure, scalable cloud data services for Vietnamese businesses.
- **Offerings:** Cloud storage, managed backup & recovery, security monitoring (IDS), APIs, 24/7 support (Vietnamese & English), **ERP-oriented AI services** (assistant, OCR, analytics, automation, voice, knowledge), and **on-demand GPU rental** for training and inference.
- **Differentiation:** Local compliance posture, in-country data residency, transparent tiers, integrated security operations, optional colocation of AI pipelines and ERP integrations within the same trust boundary.

2. High-level architecture (Vietnam boundary)



Clients

SMEs | Enterprises | Tech startups

HTTPS / APIs / Console

Vietnam — processing & data residency

API gateway & integration APIs → IAM / orgs RBAC, MFA

Object & file storage tier | Metadata & indexing | Backup & recovery

Encryption KMS | Managed IDS SIEM signals | Audit logs & retention

AI gateway LLM, RAG, tools | Vector / search ERP + docs index | GPU pool train / infer

Monitoring 24/7 NOC | Incident response | Support VI / EN

Underlying: in-country data centers or compliant local partner infrastructure (multi-AZ within Vietnam).

3. Layered logical architecture

Layer	Components	Role
Integration	Web console, SDKs/CLI, REST & event APIs	Customer access & automation
Control plane	IAM, billing/quotas, policy & compliance configuration	Governance & commercial metering
Data plane	Distributed object storage, snapshots/versioning, in-VN replication	Durable, scalable data services
Security plane	Edge protections (WAF/DDoS as applicable), IDS, hardening baselines	Detective & preventive controls
AI, automation & compute	Managed LLM endpoints (e.g. Azure OpenAI, GPT-4o/4.1), RAG pipelines, OCR/Form Recognizer, analytics sandboxes (Python/AutoML), RPA connectors (UiPath, Power Automate), speech (Whisper, Azure Speech), recommendation services, **tenant GPU instances** (configurable vCPU, RAM, disk, GPU SKU)	ERP intelligence, document capture, forecasting, workflows, voice ERP, GPU rental for ML

4. Backup & recovery flow (conceptual)

1. Customer workload or agent triggers scheduled/immediate backup via API.
2. Keys obtained from KMS; data encrypted in transit (TLS 1.2+) and at rest.
3. Chunks + metadata written to storage tier; immutable / retention policies applied by backup service.
4. IDS and access analytics consume telemetry; anomalies routed to monitoring & IR workflows.
5. Restore path: verified read, policy checks, decrypt, delivery to customer environment (paths remain in Vietnam).

5. Compliance & security mapping

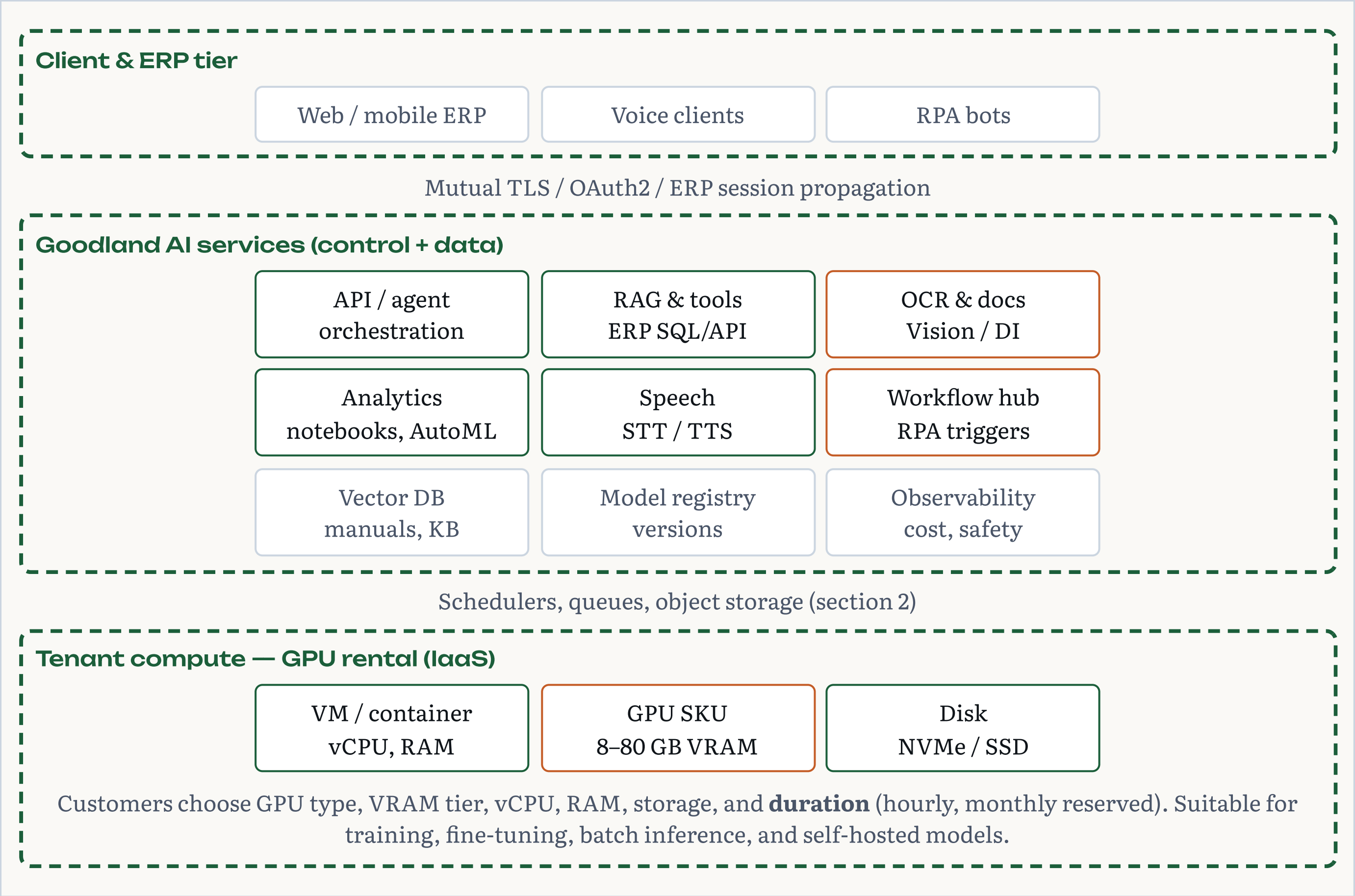
- **Data residency:** Customer data stored and processed within Vietnam for regulated workloads.
- **Regulations:** Alignment with Vietnam Cybersecurity Law (2018) and sector guidance; legal review for customer verticals.
- **Standards:** ISO 27001-style ISMS as a certification target; SLAs/DPA templates for uptime, IR, and subprocessors.
- **Cryptography:** Strong encryption at rest (e.g. AES-256) and modern TLS; key separation and rotation.
- **AI & subprocessors:** Where third-party models/APIs are used (OpenAI, Azure, Google Vision, etc.), contracts and data-flow diagrams must reflect customer consent, DPA terms, and whether prompts/embeddings leave Vietnam — offer VN-hosted or sovereign options when required.

6. AI applications for internal ERP systems (reference catalog)

Below maps common ERP AI patterns to technology options. Actual stack is selected per tenant (latency, cost, data sovereignty). RAG and knowledge services connect to approved ERP APIs, warehouses, and document stores inside the residency boundary where possible.

#	Capability	Primary use cases	Technology options	Architecture notes
1	**Internal AI chatbot (ERP assistant)**	Natural-language data queries ("this month's revenue?"), procedure help, instant report Q&A	OpenAI GPT-4o / GPT-4.1; Azure OpenAI Service; **RAG** over ERP DB views, cubes, and curated APIs	Orchestration layer enforces RBAC: queries execute as the user's ERP identity; audit log of questions & tool calls.
2	**AI OCR & document processing**	Invoice → bookkeeping lines; automated data entry from scans/PDFs	Google Vision API; Azure AI Document Intelligence / Form Recognizer; custom models on tenant GPUs	Pipeline: ingest → classify → extract → validation rules → ERP posting API; human-in-the-loop queue for low confidence.
3	**AI data analytics (AI BI)**	Revenue & inventory forecasting; anomaly detection (e.g. expense spikes); financial commentary	Python + ML (scikit-learn); Power BI + AI features; AutoML (Google Vertex / Azure ML)	Feature store or warehouse connection; scheduled jobs on CPU or GPU nodes; exports to dashboards and alerts.
4	**AI-driven process automation (AI + RPA)**	Order generation, approval routing, intelligent workflows	UiPath (AI Center, Document Understanding); Microsoft Power Automate + AI Builder	Event bus from ERP ↔ automation runners; secrets in vault; IDS monitors bot service accounts.
5	**AI voice (voice-enabled ERP)**	Hands-free KPI answers ("today's revenue?"); mobile & field scenarios	Whisper (STT); Azure Speech (STT/TTS); LLM backend as in (1)	Low-latency path: edge device → STT → intent → secured data API → TTS response; PII minimization in audio retention.
6	**AI personalization (recommendation)**	Purchase suggestions, pricing hints, lead/customer recommendations	Collaborative filtering + gradient boosting; embeddings from product/customer history; optional deep models on GPU	Offline training on GPU rental; online serving API with A/B flags and explainability hooks.
7	**AI for support**	Internal L1/L2 bot; ticketing integration; troubleshooting ("why can't I run AR report?")	Same LLM stack as (1); integration with ITSM/ticket APIs; runbooks as RAG sources	Escalation to human with full transcript; link to known-error DB.
8	**AI knowledge base**	Centralized ERP manuals, implementation guides, tribal knowledge	Document ingestion, chunking, embeddings, vector DB; optional multilingual models	Versioning tied to ERP release; access aligned to module licensing.
9	**AI help (contextual process guidance)**	Step-by-step "create warehouse receipt" with deep links to screens	LLM + structured procedure graph; role-aware (accountant vs warehouse); UI context from client shell	Differentiator: combines RBAC, current module/screen, and org policies — not static PDFs only.

7. AI / ML platform architecture (conceptual)


Client & ERP tier
Web / mobile ERP
Voice clients
RPA bots
Mutual TLS / OAuth2 / ERP session propagation
Goodland AI services (control + data)
API / agent orchestration
RAG & tools ERP SQL/API
OCR & docs Vision / DI
Analytics notebooks, AutoML
Speech STT / TTS
Workflow hub RPA triggers
Vector DB manuals, KB
Model registry versions
Observability cost, safety
Schedulers, queues, object storage (section 2)
Tenant compute — GPU rental (IaaS)
VM / container vCPU, RAM
GPU SKU 8–80 GB VRAM
Disk NVMe / SSD

Customers choose GPU type, VRAM tier, vCPU, RAM, storage, and **duration** (hourly, monthly reserved). Suitable for training, fine-tuning, batch inference, and self-hosted models.

- **Isolation:** Dedicated VPC/tenant projects; network policies between ERP data plane and GPU workers.
- **Scaling:** Auto-shutdown policies for dev GPUs; queues for batch inference; spot/preemptible classes optional (lower price, best-effort).

8. GPU rental — SKU reference (illustrative)

SKUs align to NVIDIA-class accelerators commonly requested for ML. Exact chip generation depends on datacenter supply; VRAM columns reflect customer-facing **usable VRAM tiers** (8, 16, 24, 32, 40, 48, 80 GB). **H100** offered as flagship training/inference tier (typically 80 GB).

Tier	Typical GPU families (examples)	VRAM focus	Typical workloads
Entry	T4-class, L4-class	8–16 GB	Light inference, dev/test, small CV/NLP
Professional	RTX / A-series workstation class	24 GB	Fine-tuning small LLMs, batch OCR post-processing
High-memory	A30 / L40S-class (examples)	32–48 GB	Larger models, multi-GPU optional
Datacenter	A100-class	40 / 80 GB	Training, LLM serving at scale
Flagship	**NVIDIA H100**	80 GB	Frontier training, large-scale inference, HPC-style jobs

Add vCPU and system RAM in standard blocks (e.g. 8–32 vCPU, 32–256 GB RAM) and network-attached or local NVMe volumes per performance tier.

9. Indicative pricing — core data platform (VND / month)

VAT may apply. Annual prepay typically 15% below 12× monthly on Starter–Business tiers. Enterprise is custom.

Tier	Target segment	Included storage	Backup	IDS / monitoring	Support	Indicative monthly (VND)
Starter	Small SME, pilots	500 GB	Daily, 14-day retention	Basic alerting	Business hours, ticket	2,500,000 – 4,000,000
Growth	Growing SME, startups	2 TB	Daily + weekly, 30-day	Managed IDS bundle, monthly report	Extended + chat	8,000,000 – 12,000,000
Business	Mid-market	10 TB	Policy-based RPO/RTO options	Dedicated correlation, SOC integration option	24/7, optional CSM	28,000,000 – 45,000,000
Enterprise	Regulated / large orgs	Custom (50 TB+)	In-VN geo-redundancy, legal hold	Custom playbooks, IR retainer	24/7 + on-site (major cities)	Custom (from ~80,000,000 + commit)

Add-ons — core platform (illustrative)

Add-on	Notes	Typical range (VND / month)
Extra storage	Per TB, volume discounts	400,000 – 900,000 / TB
Extended retention	e.g. 90d → 365d	+15–35% on backup component
API & integration pack	Higher limits, dedicated endpoints	1,500,000 – 5,000,000
Premium SLA	e.g. 99.9% target	+20–40% platform fee

10. Indicative pricing — ERP AI services (VND)

Bundles assume Goodland-hosted orchestration + metering. **Third-party model usage** (OpenAI, Azure OpenAI, Google Cloud) is often passed through at cost + margin or requires customer bring-your-own-key (BYOK). Numbers are order-of-magnitude for planning.

Package	What's included	Indicative monthly (VND)	Notes
AI ERP Lite	Internal assistant + knowledge base (1), (8); up to ~5k RAG queries/mo; 1 connector	6,000,000 – 12,000,000	Add seats or queries ala carte
AI ERP Standard	Lite + OCR pipeline (2) up to 2k pages/mo + support bot (7)	15,000,000 – 28,000,000	Extra pages 800 – 2,500 VND/page by volume
AI ERP Plus	Standard + voice channel (5) + workflow hooks for RPA (4) — 2 flows	28,000,000 – 48,000,000	STT/TTS minutes billed separately (below)
AI Analytics add-on	Forecasting & anomaly jobs (3); 1 dashboard workspace; scheduled scoring	8,000,000 – 22,000,000	Excludes Power BI / cloud AutoML licenses if customer-owned

Package	What's included	Indicative monthly (VND)	Notes
Recommendations add-on	Reco API (6); retrain monthly; A/B hooks	5,000,000 – 15,000,000	Heavy training uses GPU hours (section 11)
Contextual AI Help	Role-aware guidance (9); procedure graph build + 40h professional services once	4,000,000 – 10,000,000 / mo + one-time 40,000,000 – 120,000,000	PS for screen-map & content curation

AI usage meters (illustrative)

Meter	Unit	Typical range (VND)
LLM / RAG query (after bundle)	per 1,000 calls	600,000 – 2,500,000
STT (Whisper / Azure Speech)	per audio hour	180,000 – 550,000
TTS	per 1M characters	400,000 – 1,200,000
OCR / document AI	per page	800 – 2,500
RPA production bot	per bot / month	3,000,000 – 9,000,000

11. Indicative pricing — GPU rental (VND / hour)

Linux VM with selected GPU. **vCPU** and **RAM** priced additively (examples: 2,000 – 8,000 VND / vCPU-hour; 1,500 – 5,000 VND / GB-RAM-hour). **Disk:** ~400 – 1,200 VND / GB-month (NVMe premium). **Committed use:** 1-month commit ~8–12% off; 12-month ~20–35% off headline hourly. **Spot / interruptible** (if offered): ~40–70% below on-demand. VAT may apply.

GPU tier	VRAM (customer-facing)	Indicative VND / GPU-hour	Comment
Entry (e.g. T4 / L4 class)	8 GB	12,000 – 28,000	Dev / light inference
Entry+	16 GB	18,000 – 38,000	Small fine-tunes
Workstation-class	24 GB	35,000 – 75,000	Mid-size models
High-memory	32 GB	55,000 – 110,000	Training / batch
High-memory	40–48 GB	85,000 – 190,000	Larger batches, multi-worker
Datacenter (A100-class)	40 GB	150,000 – 320,000	Production training
Datacenter (A100-class)	80 GB	220,000 – 450,000	Large-model training
Flagship (H100-class)	**80 GB**	**480,000 – 980,000**	Frontier training & heavy inference

Packaging: Publish **bundled SKUs** (GPU + fixed vCPU/RAM/disk) for common sizes to simplify quoting; use component meters above for custom builds.

Goodland Cloud — Architecture & Pricing appendix (core platform, ERP AI catalog, GPU rental). Generated for planning; not a contractual offer. GPU and model prices fluctuate with supply and FX.