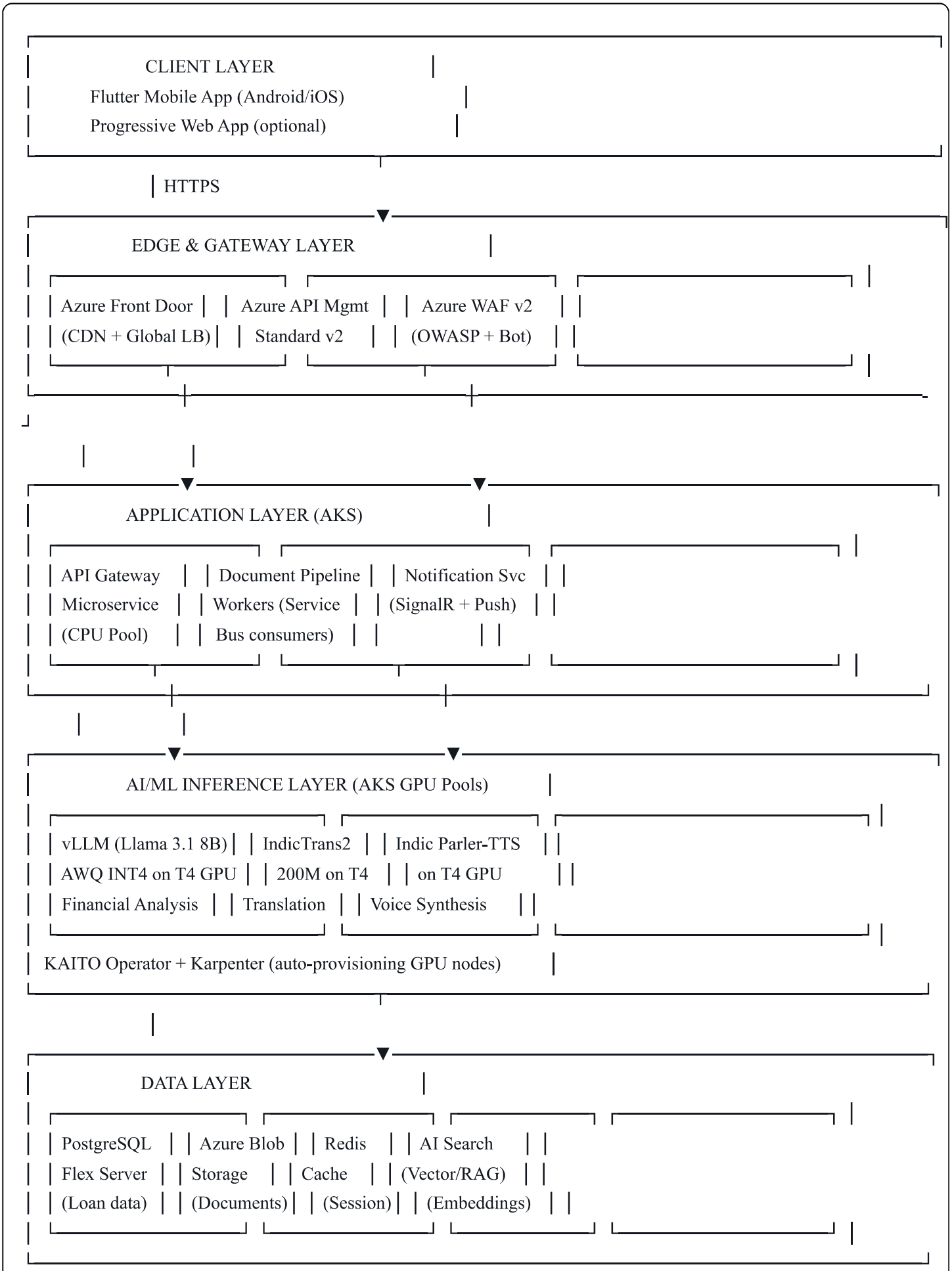
Production Azure architecture for an Indian loan document analyzer

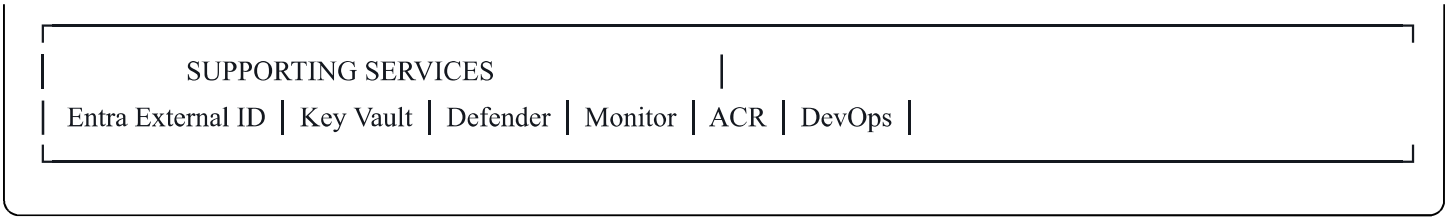
The optimal architecture centers on AKS with NCas_T4_v3 GPU node pools in Azure Central India, running Llama 3.1 8B via vLLM for financial analysis, paired with Azure Document Intelligence for OCR and AI4Bharat's open-source stack for Indian language TTS and translation. This design delivers sub-second inference latency at **$1,000–$1,500/month for MVP**, scaling to **$10,000–$25,000/month for 100K+ users**, while maintaining strict RBI and DPDPA 2023 data residency compliance. The architecture follows Azure Well-Architected Framework principles (Microsoft Learn) and evolves across three stages without re-platforming. What follows is the complete blueprint — every service, SKU, price, and configuration decision.

---

1. Architecture overview and system design

The system processes loan documents through a five-stage pipeline: **ingest → OCR/extract → AI analyze → translate → voice synthesize**. Each stage maps to a distinct infrastructure layer with independent scaling characteristics.


CLIENT LAYER
Flutter Mobile App (Android/iOS)
Progressive Web App (optional)
HTTPS
EDGE & GATEWAY LAYER
Azure Front Door
(CDN + Global LB)
Azure API Mgmt
Standard v2
Azure WAF v2
(OWASP + Bot)
APPLICATION LAYER (AKS)
API Gateway
Microservice
(CPU Pool)
Document Pipeline
Workers (Service
Bus consumers)
Notification Svc
(SignalR + Push)
AI/ML INFERENCE LAYER (AKS GPU Pools)
vLLM (Llama 3.1 8B)
AWQ INT4 on T4 GPU
Financial Analysis
IndicTrans2
200M on T4
Translation
Indic Parler-TTS
on T4 GPU
Voice Synthesis
KAITO Operator + Karpenter (auto-provisioning GPU nodes)
DATA LAYER
PostgreSQL
Flex Server
(Loan data)
Azure Blob
Storage
(Documents)
Redis
Cache
(Session)
AI Search
(Vector/RAG)
(Embeddings)



The critical architectural decision is **self-hosting all AI models** (LLM, translation, TTS) on AKS GPU node pools rather than using Azure's managed AI services. This eliminates per-token and per-character costs that would otherwise dominate at scale, provides full data residency control for RBI compliance, and avoids Azure OpenAI's undocumented data processing location concerns. Microsoft Learn

---

2. Compute layer: AKS as the unified orchestration platform

Azure Kubernetes Service (AKS) Standard tier is the recommended compute platform, providing unified orchestration for both CPU application workloads and GPU inference workloads through separate node pools. AKS was chosen over Azure Container Apps (limited GPU availability in Central India, less control) and Azure App Service (no GPU support, scaling ceiling of 30 instances).

AKS cluster configuration

The AKS cluster uses **three distinct node pool types**: a system pool for Kubernetes infrastructure, CPU application pools for the API backend and document processing workers, and GPU pools for AI inference.

Node Pool	VM Series	Purpose	Scaling
System	Standard_D2s_v5 (2 vCPU, 8 GB)	K8s system pods, CoreDNS, metrics	Fixed 2 nodes (HA)
App	Standard_D4s_v5 (4 vCPU, 16 GB)	API services, workers, queue consumers	Autoscale 2–10 nodes
GPU-LLM	Standard_NC4as_T4_v3 (4 vCPU, 28 GB, 1×T4)	LLM inference via vLLM	Autoscale 1–4 nodes
GPU-AI	Standard_NC4as_T4_v3	TTS + Translation models	Autoscale 0–2 nodes

NCas_T4_v3 is the only confirmed GPU VM series in Azure Central India. The NVIDIA T4 GPU provides **16 GB GDDR6 VRAM**, 2,560 CUDA cores, and 65 TFLOPS FP16 — ideal for inference workloads. Microsoft Learn Azure Docs NCv3 (V100) was retired September 2025. Azure Docs +3 NC A100 v4 and

NCadsH100_v5 availability in Central India remains unconfirmed and should be verified via `az vm list-skus --location centralindia`.

Pricing for NC4as_T4_v3 in Central India:

Pricing Model	Hourly Rate	Monthly (730h)	Annual
Pay-as-you-go	~$0.58	~$423	~$5,076
1-Year Reserved	~$0.37	~$270	~$3,240
3-Year Reserved	~$0.25	~$183	~$2,190
Spot (variable)	~$0.25–0.35	~$183–255	~$2,190–3,060

The recommended cost strategy is **3-year Reserved Instances for baseline GPU capacity** (the always-on LLM node) plus **Spot node pools for burst capacity** on TTS and translation workloads. The AKS Standard tier control plane costs **$0.10/hr (~$73/month)** and provides a **99.95% SLA** with availability zones. Cloudchipr

GPU auto-scaling with KEDA and Karpenter

GPU auto-scaling uses a two-layer approach. The **Kubernetes Cluster Autoscaler** manages GPU node provisioning (5–10 minute cold start for new GPU nodes), while **KEDA (Kubernetes Event-Driven Autoscaling)** triggers pod scaling based on Azure Service Bus queue depth and NVIDIA DCGM GPU utilization metrics exposed through Azure Managed Prometheus. Microsoft Learn

To mitigate GPU cold start latency, the architecture employs **Deallocate scale-down mode** (preserves container images on deallocated nodes for faster restart), **PersistentVolumeClaims on Azure Files** for pre-cached model weights (avoiding 5–15 minute model downloads on scale-up), and a **minimum node count of 1** for the LLM GPU pool to ensure instant availability. The **KAITO (Kubernetes AI Toolchain Operator)** add-on, now a CNCF Sandbox project, simplifies GPU node provisioning through Karpenter integration and provides preset deployment manifests for popular models including Llama and Phi families. Azure Docs

---

3. Self-hosted LLM infrastructure for financial analysis

Model selection strategy

The architecture uses a **tiered model approach**: a primary LLM for financial document analysis and explanation, with specialized Indic language models for translation and synthesis.

Model	Parameters	Quantization	VRAM	Role	T4 Fit?
Llama 3.1 8B	8B	AWQ INT4	~5 GB	Primary financial analysis, Hindi output	✅ Excellent
Phi-4	14B	AWQ INT4	~7 GB	Numerical reasoning, EMI calculations	✅ Good
Navarasa 2.0 (Gemma-based)	7B	INT4	~4 GB	Broad Indic language generation (15 languages)	✅ Good
IndicTrans2	200M (distilled)	FP16	~2–4 GB	Translation (22 Indian languages)	✅ Excellent
Indic Parler-TTS	~2–4 GB	FP32	~4–8 GB	Voice synthesis (21 languages)	✅ Good

Llama 3.1 8B is the primary recommendation because it officially supports Hindi (one of 8 supported languages), has 128K context length for processing long loan documents, and delivers strong financial reasoning. At AWQ INT4 quantization, it consumes only ~5 GB VRAM on the T4's 16 GB, IntuitionLabs leaving ample headroom for KV cache to serve **20–50 concurrent requests** via vLLM's PagedAttention mechanism.

For broader Indic language coverage (Telugu, Tamil, Bengali, Marathi, Kannada, Malayalam, Gujarati), **Navarasa 2.0** — a Gemma-based model fine-tuned on 15 Indian languages — fills the gap that Llama 3.1's limited official Indic support cannot. For pure numerical loan optimization tasks, **Phi-4** (14B, AWQ INT4 at ~7 GB) provides superior mathematical reasoning.

AWQ (Activation-Aware Weight Quantization) is the recommended quantization format for production GPU serving. With the Marlin kernel in vLLM, AWQ achieves **741 tokens/sec** decode throughput while maintaining output quality close to FP16 (perplexity 6.84 vs 6.74 for GGUF Q4_K_M). GPTQ with Marlin is a close alternative at 712 tokens/sec.

vLLM as the model serving framework

vLLM is the recommended inference engine, deployed on AKS GPU node pools with an OpenAI-compatible API endpoint. Key advantages over alternatives:

Framework	Throughput	Concurrent Users	Production Readiness
vLLM	120–160 req/s; 2,300+ tok/s	Excellent (1,000+)	✅ Best
HF TGI	100–140 req/s	Good	✅ Good
NVIDIA Triton	Varies	Enterprise-grade	✅ Good (complex setup)
Ollama	1–3 req/s	Poor (single-user)	❌ Dev only

vLLM's **PagedAttention** reduces VRAM waste by 60–80% versus static KV cache allocation, and **continuous batching** maximizes GPU utilization by dynamically grouping incoming requests. (iTechs Online) On a single T4 with Llama 3.1 8B INT4, expect **30–50 tokens/sec per request** with multiple concurrent requests handled efficiently.

A Kubernetes deployment for vLLM on AKS:

yaml

```yaml
apiVersion: apps/v1
kind: Deployment
metadata:
  name: vllm-llama
spec:
  replicas: 1
  selector:
    matchLabels:
      app: vllm-llama
  template:
    spec:
      nodeSelector:
        kubernetes.azure.com/accelerator: nvidia
      tolerations:
      - key: "sku"
        operator: "Equal"
        value: "gpu"
        effect: "NoSchedule"
      containers:
      - name: vllm
        image: vllm/vllm-openai:latest
        args:
        - "--model"
        - "TheBloke/Llama-3.1-8B-AWQ"
        - "--quantization"
        - "awq"
        - "--max-model-len"
        - "8192"
        - "--gpu-memory-utilization"
        - "0.90"
        resources:
          limits:
            nvidia.com/gpu: 1
        volumeMounts:
        - name: model-cache
          mountPath: /root/.cache/huggingface
      volumes:
      - name: model-cache
        persistentVolumeClaim:
          claimName: model-storage-pvc
```

Model weights are stored on an **Azure Files Premium PVC** (ReadWriteMany) shared across pods, so scale-up events don't require re-downloading multi-gigabyte model files. The Azure Container Registry (ACR) Standard

tier (**~$20/month**) hosts custom vLLM container images with pre-baked configurations.

---

4. Document processing pipeline

The document processing pipeline is the app's core workflow, orchestrating OCR extraction, AI analysis, translation, and voice synthesis through an asynchronous, event-driven architecture.

Azure AI Document Intelligence for OCR

Azure AI Document Intelligence S0 tier handles all OCR and table extraction. The service is confirmed available in Central India with data residency guarantees (Microsoft Learn) — critical for RBI compliance.

Feature	Price per 1,000 Pages	Use Case
Layout	$1.50	Table extraction from EMI schedules
Prebuilt (Invoice/Receipt)	$10.00	Structured sanction letter parsing
Custom Neural Model	$30.00	Indian loan-specific field extraction
Read (OCR only)	$1.50	Basic text extraction

The recommended approach is a **two-phase extraction**: first, the **Layout model** ($1.50/1K pages) extracts tables and text structure from EMI schedules and amortization tables. Then, a **Custom Neural Model** ($30/1K pages) trained on labeled Indian loan documents extracts domain-specific fields — loan amount, interest rate (fixed/floating), EMI amount, tenure, processing fees, prepayment charges, and borrower details. Custom Neural models support up to 50,000 training pages and handle variable-layout documents well. (Microsoft Learn)

Training is done through Document Intelligence Studio: upload 50–200 labeled sample documents per document type (sanction letters, EMI schedules, loan agreements, Key Fact Statements), label fields, train, and deploy. **First 10 hours of training are free**, then $3/hour. (Azure Docs)

Async pipeline architecture

```
Document Upload → Blob Storage (Hot) → Service Bus Queue →
  Processing Worker → Document Intelligence (OCR) →
  LLM Analysis (vLLM) → Translation (IndicTrans2) →
  TTS (Indic Parler-TTS) → Results → PostgreSQL + Blob (audio) →
  SignalR Notification → Client
```

Azure Service Bus Standard tier (~$10/month base + $0.05/million operations) provides the async messaging backbone (Microsoft Learn) (GitHub) with dead-letter queues for failed processing, sessions for ordered document

processing, and topic subscriptions for fan-out to multiple consumers (analysis, translation, TTS can run in parallel after OCR completes).

Azure Blob Storage Hot tier stores uploaded documents and generated audio files. At Central India LRS pricing: **$0.018/GB/month** for storage, $0.055/10K write operations. Documents older than 30 days auto-transition to **Cool tier** ($0.01/GB/month) via lifecycle management policies. Generated audio files for common explanations are cached in Blob with CDN distribution through Azure Front Door.

---

5. Indian language TTS and translation infrastructure

Text-to-speech: hybrid Azure + self-hosted approach

Azure AI Speech Service supports **13 Indian languages** with Neural TTS voices (Microsoft Community Hub) at **$16 per million characters**: (SaaSworthy)

Language	Neural Voices	Quality
Hindi (hi-IN)	SwaraNeural (F), MadhurNeural (M) + 5 more	Excellent
Telugu (te-IN)	ShrutiNeural (F), MohanNeural (M)	Good
Tamil (ta-IN)	PallaviNeural (F), ValluvarNeural (M)	Good
Bengali (bn-IN)	TanishaaNeural (F), BashkarNeural (M)	Good
Marathi (mr-IN)	AarohiNeural (F), ManoharNeural (M)	Good
Kannada (kn-IN)	SapnaNeural (F), GaganNeural (M)	Good
Malayalam (ml-IN)	SobhanaNeural (F), MidhunNeural (M)	Good
Gujarati (gu-IN)	DhwaniNeural (F), NiranjanNeural (M)	Good

For MVP stage, use Azure Neural TTS — the quality is enterprise-grade, there's no infrastructure to manage, and costs are predictable at $16/1M characters. At **10K users generating ~500 characters per explanation**, that's roughly 5M characters/month = **~$80/month**.

At Growth/Scale stage, transition to self-hosted AI4Bharat Indic Parler-TTS. This open-source model (permissive license) covers **21 Indian languages** (AI Models) — 8 more than Azure — with 69 unique voices (X) and 12 controllable emotions. It runs on a single T4 GPU (~4–8 GB VRAM) and eliminates per-character costs entirely. At 100K users, the self-hosted approach saves **$800+/month** versus Azure TTS. Quality is competitive with Azure Neural TTS for Indian languages based on MOS evaluations. (Hugging Face)

Translation: IndicTrans2 self-hosted from day one

IndicTrans2 (distilled 200M parameters) is recommended over Azure Translator from the start. The MIT-licensed model from AI4Bharat/IIT Madras (AI4Bharat) supports **all 22 scheduled Indian languages** (GitHub) (GitHub) and fits easily on a T4 GPU at ~2–4 GB VRAM (FP16). It performs on par with or better than commercial translation systems for Indian language pairs. (arXiv)

Azure Translator charges **$10 per million characters** and supports 22+ Indian languages. At scale (100K users, ~50M characters/month), that's **$500/month** in translation costs alone. IndicTrans2 self-hosted on a shared GPU pod eliminates this entirely. The distilled 200M model provides excellent speed for real-time translation while the 1.1B base model offers higher quality for batch processing.

Both IndicTrans2 and Indic Parler-TTS can share the GPU-AI node pool (Standard_NC4as_T4_v3). With IndicTrans2 at ~3 GB and Parler-TTS at ~5 GB, both models fit within the T4's 16 GB VRAM, (Hetzner Cloud) enabling a **single GPU node to serve both translation and TTS —** a significant cost optimization.

---

6. Database, caching, and search layer

PostgreSQL Flexible Server for structured loan data

Azure Database for PostgreSQL Flexible Server stores all structured data: user profiles, loan details, extracted document fields, analysis results, and audit logs. PostgreSQL's ACID compliance, complex query support, and native JSON capabilities make it ideal for financial data.

Stage	SKU	vCores	RAM	Storage	Monthly Cost
MVP	Burstable B2s	2	4 GB	64 GB	~$65
Growth	General Purpose D4ds_v5	4	16 GB	256 GB	~$250
Scale	General Purpose D8ds_v5 + HA	8	32 GB	512 GB	~$700

Zone-redundant high availability (doubles compute cost) is recommended at Scale stage. **3-year reserved pricing saves up to 60%** on compute. Built-in PgBouncer handles connection pooling. The (pgvector) extension enables storing document embeddings directly in PostgreSQL, potentially eliminating the need for Azure AI Search at MVP stage.

Azure Cache for Redis (migrating to Azure Managed Redis)

Redis serves three caching roles: **LLM response cache** (semantic caching of similar loan explanations), **session state** (user authentication tokens, processing status), and **rate limiting** (API throttling per user). Note that Azure

Cache for Redis Basic/Standard/Premium tiers retire September 2028 (Microsoft Azure) — plan for **Azure Managed Redis** migration.

Stage	Tier	Capacity	Monthly Cost
MVP	Basic C0	250 MB	~$16
Growth	Standard C1	1 GB	~$42
Scale	Premium P1	6 GB (clustering)	~$225

Azure AI Search for RAG vector search

Azure AI Search Basic tier (~$75/month, 15 GB storage) provides vector search for a Retrieval-Augmented Generation pipeline. The RAG pipeline indexes RBI guidelines, loan regulation documents, and financial terminology databases. When the LLM generates explanations, it retrieves relevant regulatory context to ensure accuracy.

Vector search is supported on all billable tiers (Microsoft Azure) with HNSW indexing. Semantic ranker (1,000 free queries/month) improves relevance for natural language queries. At Scale stage, upgrade to **Standard S1** (~$245/month) for higher throughput and 160 GB storage.

---

7. Networking, security, and identity

Network architecture

The VNet design uses **four subnets** with NSG-enforced isolation:

Subnet	CIDR	Resources	Access
aks-subnet	/16 or /21	AKS nodes (Azure CNI)	Outbound only
appgw-subnet	/24	Application Gateway / WAF	Internet-facing
data-subnet	/24	Private Endpoints (PostgreSQL, Redis, Storage, ACR)	AKS-only ingress
mgmt-subnet	/27	Azure Bastion, jump boxes	Admin-only

All data services connect through **Azure Private Endpoints** (~$7.30/month each). This ensures database traffic never traverses the public internet — a strict requirement for RBI compliance. Private endpoints are configured for PostgreSQL, Redis, Blob Storage, Key Vault, ACR, and Service Bus.

Azure Front Door Premium (~~$330/month) serves as the global entry point with built-in WAF (OWASP Core Rule Set + bot protection), CDN for static assets and cached audio files, and SSL termination.~~ **~~Azure API Management Standard v2~~** (~~~~$670/month) provides API gateway functionality with rate limiting, request transformation, JWT validation, and developer portal. For MVP, APIM can be deferred — use AKS ingress controller (NGINX) with Azure Front Door directly.

Identity and authentication

Microsoft Entra External ID (successor to Azure AD B2C, which stopped accepting new customers May 2025) handles user authentication with **phone OTP sign-in** — the dominant authentication pattern in India. First **50,000 monthly active users are free**. Beyond that, P1 tier costs ~$0.00325/MAU. SMS OTP verification costs **~$0.03 per attempt**. Custom policies enable integration with Aadhaar eKYC providers for enhanced identity verification.

Azure Key Vault Standard tier manages all secrets (database connection strings, API keys, model access tokens) at $0.03/10K operations. All AKS workloads authenticate to Key Vault using **Managed Identities** (no credential management, zero additional cost). The AKS cluster uses a system-assigned managed identity; individual pods use **Workload Identity Federation** for fine-grained access control.

DPDPA 2023 and RBI compliance

Data residency is the paramount compliance requirement. The RBI's April 2018 directive mandates that **all payment system data must be stored exclusively in India** (M2pfintech) — this extends to loan-related payment data, borrower financial details, and NACH mandate information. Azure Central India (Pune) satisfies this requirement. (Azure)

Key compliance implementations:

- **Azure Policy "Allowed Locations"** restricts all resource deployment to Central India and South India regions only (Microsoft Azure)
- **Consent management** module in the application layer captures DPDPA-compliant consent (free, specific, informed, unambiguous) before processing any personal data
- **Breach notification pipeline** — Azure Defender for Cloud alerts trigger automated incident response workflow reporting to the Data Protection Board (mandatory for ALL breaches under DPDPA, regardless of severity) (CookieYes)
- **Data Principal rights API** — endpoints for access, correction, and erasure requests with 90-day response SLA (Securiti)
- **Encryption**: AES-256 at rest (Azure default for all services), TLS 1.3 in transit, customer-managed keys via Key Vault for sensitive loan data
- **Microsoft Defender for Cloud** at $7/vCore/month (Containers plan) + $15/instance/month (Servers Plan 2) provides runtime threat protection and compliance scoring

8. Monitoring, observability, and DevOps

Observability stack

Azure Monitor with Application Insights provides end-to-end observability. Log Analytics workspace pricing is **~$2.50/GB ingested** (pay-as-you-go) (Azure Docs) with 5 GB/month free. (Medium) For GPU workloads, **NVIDIA DCGM Exporter** feeds GPU utilization, memory, and temperature metrics into Azure Managed Prometheus, enabling KEDA-based auto-scaling decisions.

Custom dashboards track four critical metrics: **document processing latency** (P95 target: <30 seconds end-to-end), **LLM inference throughput** (tokens/sec per GPU), **TTS generation time** (target: <3 seconds per explanation), and **GPU utilization** (target: 60–80% to balance cost and headroom). Alerts fire when GPU utilization exceeds 85% sustained or document queue depth exceeds 100 messages.

CI/CD pipeline

GitHub Actions is recommended over Azure DevOps for its modern workflow syntax, superior marketplace ecosystem, and free tier (2,000 minutes/month). The deployment pipeline follows a GitOps model:

```
Developer → Git push → GitHub Actions (CI) → Build + test → Push image to ACR →
  Update K8s manifests in config repo → Flux/ArgoCD (CD) → Deploy to AKS →
  Argo Rollouts (canary analysis for model updates)
```

Infrastructure as Code uses Terraform (not Bicep) because the architecture includes self-hosted open-source AI models from HuggingFace and AI4Bharat that may eventually expand to multi-cloud. Terraform's large ecosystem and state management are advantages for complex AKS + GPU configurations. However, for Azure-only teams, Bicep offers day-zero resource support and no state file management.

Key Terraform resources for the GPU infrastructure:

```hcl
```

```
resource "azurerm_kubernetes_cluster_node_pool" "gpu_llm" {
  name                  = "gpullm"
  kubernetes_cluster_id = azurerm_kubernetes_cluster.main.id
  vm_size               = "Standard_NC4as_T4_v3"
  min_count             = 1
  max_count             = 4
  enable_auto_scaling   = true
  priority              = "Regular"  # Use "Spot" for burst pool
  node_taints           = ["sku=gpu:NoSchedule"]
  node_labels           = { "workload" = "llm-inference" }
  os_disk_size_gb       = 128

  tags = {
    environment = "production"
    workload    = "gpu-inference"
  }
}
```

Blue-green deployments for zero-downtime LLM model updates use **Argo Rollouts** with canary analysis: deploy the new model version to 10% of traffic, validate accuracy metrics and latency via Prometheus queries, then progressively shift to 30% → 50% → 100% with automated rollback on quality regression.

---

9. Mobile backend and real-time services

Azure SignalR Service Standard (~$49/unit/month) provides real-time document processing status updates to the mobile app. (Microsoft Azure) When a user uploads a loan document, SignalR pushes progress events (OCR started → extraction complete → analysis running → translation done → audio ready) without client polling. Each unit supports 1,000 concurrent connections and 1M messages/day. (Ably)

Azure Notification Hubs Basic (~$10/month) delivers push notifications (Microsoft Learn) for completed analyses, new feature announcements, and EMI payment reminders. The Basic tier supports 10M pushes/month and 200K registered devices — sufficient through Growth stage.

Azure Front Door Standard/Premium replaces the deprecated standalone Azure CDN for serving static mobile app assets (JavaScript bundles, images) and cached TTS audio files. Frequently generated audio explanations (e.g., "What is compound interest" in Hindi) are pre-generated, stored in Blob, and served via Front Door's edge POPs with **<50ms latency** across India.

---

10. High availability and disaster recovery

Multi-region design

Component	Primary (Central India)	DR (South India)	Replication
AKS Cluster	Full production	Warm standby (reduced capacity)	ACR geo-replication
PostgreSQL	Read-write primary	Read replica	Async replication (RPO ~minutes)
Blob Storage	RA-GRS	Read-only secondary	Async (RPO ~15 min)
Redis	Premium with geo-replication	Passive secondary	Active geo-replication
GPU VMs	Full capacity	Cold standby (ASR protected)	Azure Site Recovery
AI Search	Primary index	Rebuild from source on failover	No native geo-replication

RTO/RPO targets: RTO < 4 hours, RPO < 30 minutes for standard operations. The application layer (AKS) can failover within minutes via Azure Front Door health probe–based routing. Database failover (PostgreSQL replica promotion) takes 5–15 minutes with manual intervention.

Azure Front Door handles failover routing automatically. Health probes test **actual application dependencies** (database connectivity, GPU availability, queue health) via a deep health endpoint — not just a superficial `/health` 200 response. South India (Chennai) is Central India's **official paired region**, ensuring that Azure's automated geo-failover for GRS storage works correctly.

Blob Storage uses RA-GRS (Read-Access Geo-Redundant Storage) at ~2.2× LRS cost, enabling read access from the secondary region during a primary region outage. This ensures uploaded documents and cached audio files remain accessible during failover.

---

11. Detailed cost estimation across all stages

MVP stage (1K–10K users): ~$1,200–$1,800/month

Component	Specification	Monthly Cost
AKS Control Plane	Free tier	$0
AKS System Pool	2× Standard_D2s_v5	~$140
AKS App Pool	2× Standard_D4s_v5	~$280

Component	Specification	Monthly Cost
GPU Pool (LLM)	1× NC4as_T4_v3 (3yr RI)	~$183
PostgreSQL	Burstable B2s, 64 GB	~$65
Redis	Basic C0 (250 MB)	~$16
Blob Storage (LRS)	100 GB Hot	~$5
Document Intelligence	S0, ~5K pages/month (Layout + Prebuilt)	~$60
Azure TTS	Neural, ~2M chars/month	~$32
Service Bus	Standard	~$10
Key Vault	Standard	~$3
Monitoring	App Insights, ~5 GB/month	~$15
ACR	Basic	~$5
Networking	Standard LB + basic egress	~$30
Entra External ID	<50K MAU (free)	$0
Total MVP		**~$844–$1,000**
+ SMS OTP costs (~$0.03/auth × active users)		*+$30–300*

At MVP, Azure TTS is used instead of self-hosted TTS to minimize GPU cost. The single GPU node runs Llama 3.1 8B (AWQ) via vLLM. Translation can use Azure Translator ($10/1M chars) or share the GPU with IndicTrans2.

Growth stage (10K–100K users): ~$3,500–$6,000/month

Component	Specification	Monthly Cost
AKS Control Plane	Standard tier	~$73
AKS System Pool	2× Standard_D2s_v5	~$140
AKS App Pool	3–5× Standard_D4s_v5	~$420–700
GPU Pool (LLM)	2× NC4as_T4_v3 (3yr RI)	~$366
GPU Pool (AI)	1× NC4as_T4_v3 (TTS + Translation, Spot)	~$183–255

Component	Specification	Monthly Cost
PostgreSQL	GP D4ds_v5, 256 GB	~$250
Redis	Standard C1 (1 GB)	~$42
AI Search	Basic	~$75
Blob Storage (ZRS)	500 GB Hot	~$12
Document Intelligence	S0, ~50K pages/month	~$500
Service Bus	Standard	~$15
Front Door	Standard	~$35 + transfer
APIM	Standard v2	~$670
Key Vault	Standard	~$5
Monitoring	App Insights, ~20 GB/month	~$50
ACR	Standard	~$20
SignalR	Standard (1 unit)	~$49
Notification Hubs	Basic	~$10
Defender	Containers + Servers P1	~$100
Total Growth		**~$3,000–$3,500**
With APIM + Front Door		**~$3,700–$4,200**

At Growth stage, self-hosted IndicTrans2 and Indic Parler-TTS replace Azure Translator and Azure TTS, running on the shared GPU-AI Spot node. Document Intelligence becomes the largest AI service cost at ~$500/month for 50K pages.

Scale stage (100K+ users): ~$10,000–$20,000/month

Component	Specification	Monthly Cost
AKS Control Plane	Standard tier	~$73
AKS System + App Pools	2 system + 8–15 app nodes (D4s_v5)	~$1,540–2,240
GPU Pool (LLM)	3–4× NC4as_T4_v3 (3yr RI)	~$550–730

Component	Specification	Monthly Cost
GPU Pool (AI)	2× NC4as_T4_v3 (1 RI + 1 Spot)	~$366–440
PostgreSQL	GP D8ds_v5 + HA, 512 GB	~$700
Redis	Premium P1 (6 GB, clustering)	~$225
AI Search	Standard S1	~$245
Blob Storage (RA-GRS)	2 TB	~$80
Document Intelligence	S0, ~200K pages/month (commitment tier)	~$1,600
Service Bus	Premium (1 MU)	~$668
Front Door	Premium	~$330 + transfer
APIM	Standard v2 (2 units)	~$1,340
Monitoring	Full stack, ~50 GB/month	~$130
DR Standby (South India)	Reduced capacity	~$1,500–3,000
Security Stack	Defender + auditing	~$200
Other services	ACR Premium, SignalR, Notification Hubs	~$300
Total Scale		**~$9,850–$12,300**
With full DR + reserved capacity		**~$12,000–$18,000**

Cost optimization levers across all stages

- **GPU Reserved Instances (3-year)** save ~57% — applying this to 2 baseline GPU VMs saves ~$480/month versus PAYG
- **Spot VMs for GPU burst** save ~50% — use for TTS/translation pods that tolerate eviction
- **Document Intelligence commitment tiers** reduce per-page costs by 15–30% at high volumes
- **Blob lifecycle management** auto-tiers documents to Cool/Cold, saving 45–80% on storage
- **LLM response caching** in Redis reduces GPU inference calls by an estimated 30–40% for common loan explanations
- **Scheduled GPU scaling** — shut down non-critical GPU nodes during 11 PM – 7 AM IST (save ~33% on non-reserved nodes)
- **Azure Advisor** provides automated right-sizing recommendations for underutilized VMs

12. Why this architecture works for the Indian market

Three design decisions make this architecture specifically suited for India. First, **complete data sovereignty**: every byte of user data stays within Azure's India geography (Central India primary, South India DR), satisfying both DPDPA 2023's flexible requirements and RBI's strict payment data localization mandate. Self-hosted LLMs eliminate the ambiguity around Azure OpenAI's data processing location.

Second, **language-first design**: the AI4Bharat stack (IndicTrans2 for 22 languages, Indic Parler-TTS for 21 languages, IndicBERT for NLU) was built specifically for Indian languages by IIT Madras researchers. These models outperform generic multilingual models on Indian language pairs and cover all 8 priority languages: Hindi (528M speakers), Bengali (97M), Marathi (83M), Telugu (81M), Tamil (69M — highest internet adoption at 42%), Gujarati (55M), Kannada (44M), and Malayalam (35M).

Third, **mobile-first authentication**: Entra External ID with phone OTP sign-in mirrors the authentication pattern 99% of Indian language internet users expect. The architecture supports future Aadhaar eKYC integration (OTP-based) for enhanced identity verification required by financial services, and eSign for digital loan agreement signing.

Conclusion

This architecture makes two bets that differentiate it from a conventional Azure AI deployment. The first is that **self-hosted open-source models on T4 GPUs deliver better unit economics than managed AI services** at scale — the break-even point occurs around 50K monthly active users, after which the GPU infrastructure cost grows sub-linearly while per-call API costs grow linearly. The second bet is that **India-specific AI models (AI4Bharat stack) outperform generic multilingual models** for Indian language tasks — IndicTrans2's BLEU scores on Indian language pairs exceed Google Translate and Azure Translator, and Indic Parler-TTS covers 8 more Indian languages than Azure Neural TTS.

The architecture's most important property is its **evolutionary design**: it starts at ~$1,000/month for MVP using managed Azure services (TTS, Translator) alongside a single GPU node, then progressively self-hosts more capabilities as user volume justifies the infrastructure investment. No re-platforming is required — AKS node pools simply scale, new model pods deploy alongside existing ones, and Terraform modules add services incrementally. Every component has a clear upgrade path from MVP through Scale without architectural changes.