



AETOS
TECHLABS

LATTICENET
VOICE AI INFRASTRUCTURE

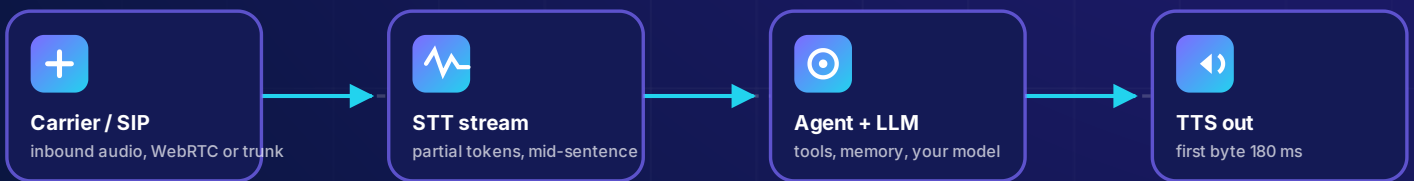


LATTICENET

Voice infrastructure you actually own.

A modular voice AI stack you run in our cloud or on your own hardware — and rebrand as your own product. Swap any provider, keep every millisecond and every recording where you want it.

THE PATH OF A TURN · FOUR HOPS, UNDER HALF A SECOND



live-session · agt_36d20

LIVE

stt.partial	"book a site survey for..."
agent.tool	calendar.find_slot()
tts.first_byte	180 ms
turn.total	450 ms

EVERY LAYER IS A SERVICE

STT · LLM · TTS — independent, versioned,
swappable without touching the rest.

No black box between your caller and your model.

450 ms

turn latency —
streaming end to end

12k

concurrent sessions

28

provider plugins,
typed contracts

99.98 %

uptime

Infrastructure-first, not another API key.

Every layer is a service you can inspect, replace or pin. The same product in our cloud, your private cloud or on-premises — same console, same APIs, same agents.

ARCHITECTURE

Modular by design

STT, TTS and LLM are independent, versioned services. Swap any one without touching the rest of the stack.

LATENCY

Sub-500 ms turns

Streaming end to end. Partial transcripts feed the model before the caller finishes speaking.

DEPLOYMENT

Runs on your metal

Kubernetes, Compose or a single box. No egress to a vendor, no per-minute meter running.

STABILITY

N and N-1 support

Never a forced upgrade. Current and previous majors run side by side during migration.

EXTENSIBILITY

Plugin contracts

Typed interfaces for custom providers and agent logic. Ship your own without forking core.

SECURITY

Private by default

Audio and transcripts stay inside your perimeter. HIPAA and GDPR deployments supported.

FASTEST TO LAUNCH

LatticeNet Cloud

We operate the whole stack in our regions. Console, APIs and numbers on day one — no servers, no upgrade windows, no pager.

- Provisioned in minutes
- Managed upgrades and backups
- Regional data residency
- Usage billed monthly

YOUR ACCOUNT

Private cloud

The same artefacts deployed into your own AWS, Azure or GCP account. You hold the keys and the bill; we operate it or hand over the runbook.

- Your cloud account and VPC
- Your KMS keys and buckets
- Terraform and Helm supplied
- Managed or self-operated

NOTHING LEAVES SITE

On-premises

Installed in your own data centre or on a single box on the factory floor. Runs with no internet at all — licence checks, models and telephony stay local.

- Your hardware, your racks
- Fully air-gapped installs
- Local GPU inference supported
- Offline licence verification

Start hosted and move on-premises later, or the reverse. The licence follows you, and there is no re-implementation either way — same console, same APIs, same agents in all three.

Run your own STT, LLM and TTS — or don't.

Every model is a swappable sidecar behind one contract. Run them on your GPUs for zero egress and flat cost, call a cloud provider, or mix the two — local by default with cloud fallback when the GPU is saturated.

Speech to text

SELF-HOSTED

faster-whisper (large-v3, distil)

NVIDIA Parakeet

Vosk — tiny footprints

OR CLOUD

Deepgram

AssemblyAI

Azure Speech

Language model

SELF-HOSTED

vLLM — Qwen, Llama, Mistral

Ollama for small boxes

Any OpenAI-compatible server

OR CLOUD

OpenAI

Anthropic

Google

Groq — via one gateway

Text to speech

SELF-HOSTED

Piper — CPU, very fast

Kokoro

XTTS for voice cloning

OR CLOUD

ElevenLabs

Cartesia

Azure Neural

WHAT A GPU BOX GIVES YOU

One RTX 4090.

Sessions	Roughly 6–10 concurrent, fully local
Typical stack	Parakeet or faster-whisper · Qwen on vLLM · Piper or Kokoro
Cost shape	Fixed hardware cost instead of per-minute metering
Egress	None — audio and transcripts never leave the machine

HOW SWAPPING WORKS

One contract, any provider.

Contract	Providers speak provider-api v1 over a local socket
Independence	Upgrade one model without touching Core or the others
Per agent	Each agent picks its own STT, LLM and TTS
Fallback	Route to cloud automatically when local capacity runs out

Any target

Docker Compose on one box, Kubernetes with Helm, or bare metal under systemd. Same artefacts, same versions.

Your storage

Recordings, transcripts and vectors land in your own S3, Postgres and Qdrant. Nothing is mirrored to us.

Offline licensing

A signed licence file verifies locally with a grace period. No phone-home, no kill switch, no metering callback.

You choose upgrades

Pinned, signed releases from a private registry. Core supports N and N-1 so nothing upgrades behind your back.

Operable by your team

Prometheus metrics, OpenTelemetry traces, structured logs and a documented runbook for every service.

Egress optional

Point at cloud providers or run everything on your GPUs. Fully air-gapped installs are supported.

REFERENCE FOOTPRINT

Single node	8 vCPU / 16 GB — up to ~40 concurrent sessions
With local GPU	1× RTX 4090 — ~6–10 fully-local sessions
Datastores	Postgres, Redis, Qdrant, S3-compatible object storage
Ingress	SIP trunk or WebRTC; Caddy or nginx terminating TLS

WHAT LICENSING COVERS

Model	Commercial subscription licence, per deployment
Source	Source-available under NDA; escrow on Enterprise
Verification	Signed licence file, checked offline, with grace period
Not included	No Apache, MIT or other open-source grant applies

Resell it under your own brand. We stay invisible.

Put your name on it and sell it to your own customers, hosted by us or by you. No co-branding, no attribution, no contact with your accounts.

Wholesale pricing

Partner rates with margin built in. You set your own retail price and keep the difference.

Provision in a click

One tenant per customer from your partner console — own brand profile, subdomain, users and limits.

One bill to you

We invoice you once for aggregate usage. Your customers are billed by you, in your name, on your terms.

You own the relationship

You run tier 1 under your brand. We back you on tier 2 and 3, privately, never in front of your customer.

WHAT WE GUARANTEE

Four promises, in the contract.

Invisibility

Our name appears nowhere your customers can see.

No poaching

We will not solicit or sell directly to your accounts.

Escalation

Named engineer for tier 2 and 3, under your brand.

Portability

Move tenants between our cloud and yours without re-licensing.

Brand tokens, not a fork

Logo, wordmark, favicon, colour ramps, radii and typography are theme variables. Your look survives every upgrade.

Your domain, your TLS

Console, API and docs served from your hostnames. No LatticeNet domain appears anywhere in the product.

No trace of us

Product name, marks, footer credits, page titles and error pages are all yours. No attribution required.

Mail and notifications

Transactional email, invite flows and alert templates carry your sender identity and copy.

Per-tenant themes

Ship one install and give each customer their own brand profile, subdomain and login screen.

Docs, CLI and SDKs

Generated API reference, CLI name and client SDK package names renamed to match your product.

MOST PARTNERS START HERE

Reseller — sell it as your own product

Full white-label included · zero attribution · per-customer tenant provisioning · wholesale pricing, you set retail · one aggregate invoice · named solutions engineer. Available cloud-hosted, self-hosted or on-premises.

ENTERPRISE

Run it yourself, at your own scale

Full white-label included · unlimited internal tenants · source-available under NDA, escrow on request · custom provider development · security and compliance review · contracted SLA and named escalation · migration and onboarding included.



See it running on your own stack.

We walk through a deployment on your infrastructure, under your brand, and scope the licence to your concurrency. Pricing is quoted, not listed — a real figure the same week.



CALL

+91 99666 12678
+91 90309 03030



WEB

aetostechlabs.com



EMAIL

krishna@aetostechlabs.com
aetostechlabs@yahoo.com



OFFICE

Aetos Techlabs LLP
Hyderabad, Telangana