Vivik: Deterministic Telephony AI

Architectural Determinism in the Non-Deterministic Media Plane

TECHNICAL OVERVIEW

Bajpai Labs Engineering Team

May 2026

Section 1

Executive Summary

The fundamental challenge in modern telephony artificial intelligence lies not in the sophistication of the language model, but in the physical reality of the signal path. Traditional voice-over-IP (VoIP) and conversational AI systems operate as a collection of bolt-on services — discrete modules for transcription, reasoning, and synthesis connected by non-deterministic network hops.

This architectural fragmentation introduces cumulative jitter and latency that fundamentally break the human conversational loop. Vivik represents a radical departure from this paradigm, grounded in first-principles engineering and the strict application of the Two-Worlds Principle. By bifurcating the system into a hard real-time media plane and a high-throughput control plane, Vivik achieves a level of determinism previously reserved for specialised digital signal processing hardware.

> *“Not just fast — deterministic. Every audio frame processed within a fixed 20ms budget, regardless of system load.”*

Section 2

The Two-Worlds Principle: Bifurcated Architecture

At the core of the Vivik architecture is the Two-Worlds Principle: a total separation between the deterministic media processing path and the non-deterministic orchestration layer. In telephony AI, “determinism” means processing every audio frame within a fixed 20-millisecond budget — without exception.

The Media Plane handles raw pulse-code modulation (PCM) data — resampling, echo cancellation, noise suppression, and voice activity detection. Because these operations are sensitive to even microsecond-level delays, the Media Plane is implemented in Rust with SIMD-accelerated numeric kernels, providing total control over memory management and execution timing, completely free from garbage-collector interference.

The Control Plane manages the high-level logic of the conversation: enforcing security policies, managing session state, routing calls to specific AI agents, and orchestrating the flow between the LLM and the media stream. It operates in the world of high-concurrency asynchronous I/O (Go + NATS), maximising throughput and managing distributed state across a cluster.

Feature	Media Plane (Deterministic)	Control Plane (Orchestration)
Primary Unit	20ms Audio Frame	Conversation Turn / Session
Language	Rust (No GC)	Go (Managed GC)
Critical Metric	P99.999 Latency (Hard Real-Time)	P99 Latency / Requests/s
Key Operations	Resampling, VAD, Sinc Interpolation	Routing, Auth, Tool Calling
Concurrency	Lock-Free / Wait-Free Queues	Goroutines / Channels

Table 1: Comparison of Media and Control Plane Architectures

Section 3

Signal Processing: From Analog Voice to AI-Ready Data

The transition from analog speech to AI-ready data is governed by the Nyquist-Shannon sampling theorem — the fundamental bridge between the continuous signals produced by the human vocal tract and the discrete sequences required by neural networks.

Traditional telephony (G.711) samples audio at 8 kHz, sufficient for intelligibility but lacking the spectral richness required by modern AI models. Gemini Live, for example, requires inputs at 16 kHz or 24 kHz, making real-time upsampling a mathematically demanding engineering requirement that must occur entirely within the deterministic media plane.

Vivik implements truncated sinc interpolation with Kaiser windowing, achieving a stopband attenuation of -90 dB. This keeps the noise floor introduced by resampling far below the sensitivity threshold of any speech-to-text model — preserving the full spectral quality that separates natural-sounding AI from robotic transcription artifacts.

Parameter	8 kHz (PSTN)	16 kHz (AI Input)	24 kHz (AI Output)
Max Frequency	4 kHz	8 kHz	12 kHz
Bit Depth	8-bit (A-law/μ-law)	16-bit PCM	16-bit PCM
Spectral Detail	Low (Intelligibility)	Moderate (Accuracy)	High (Affective)

Table 2: Sampling Rates and Spectral Dynamics in the Vivik Signal Chain

Section 4

The Rust Advantage: Predictable Performance at Scale

The requirements of the Media Plane — hard real-time execution, zero jitter, and massive numeric throughput — make the choice of programming language a critical engineering decision. While the broader AI industry has gravitated toward Python or Go, Vivik's Media Plane is built entirely in Rust.

4.1 Eliminating Garbage Collection Pauses

In managed languages, the runtime periodically performs garbage collection to reclaim memory. Even sub-millisecond GC pauses are fundamentally non-deterministic. In a high-concurrency gateway handling 5,000 requests per second, a 2 ms pause creates catastrophic tail latency spikes.

Empirical benchmarks show that under 25,000 requests per second, a Go implementation produced a P99 latency of 1,550 ms, while Rust with the Tokio runtime maintained a P99 of 310 ms. For a telephony AI system with a sub-500 ms response SLA, that difference is the line between a natural conversation and a broken one.

4.2 SIMD Acceleration

Rust's first-class SIMD support is essential for FFT-based synchronous resampling. Using AVX-512 or ARM NEON vector instructions, the Media Plane processes multiple audio samples in a single CPU cycle — dramatically increasing the density of concurrent audio streams a single server can handle.

Section 5

Concurrency at Scale: Go and the Control Plane

While the Media Plane demands Rust's uncompromising control, the Control Plane is built in Go — optimised not for signal processing but for orchestrating tens of thousands of concurrent I/O-bound tasks.

Go's goroutines are extremely lightweight threads managed by the runtime rather than the OS. A single Go process handles hundreds of thousands of goroutines with ease. When a call arrives, a goroutine spawns to handle that session — coordinating authentication, agent configuration lookup, and WebSocket connection to the LLM — all without blocking the rest of the system.

The strict boundary between the Rust Media Plane and the Go Control Plane ensures that no GC event in the orchestration layer can ever introduce jitter into the audio stream. The two worlds communicate exclusively through NATS — a high-performance message bus that routes events at line speed without a central broker bottleneck.

Section 6

Voice Activity Detection: Speaking the Language of Speech

Voice Activity Detection (VAD) is arguably the most consequential component of the media pipeline. It acts as a binary gatekeeper: is the caller speaking, or not? In the Two-Worlds architecture, VAD must reside entirely in the Media Plane — detection must occur at line speed, with no latency from an orchestration boundary.

Vivik implements a dual-gate VAD combining two first-principles signal metrics: **Root Mean Square (RMS) energy**, which correlates to perceived loudness, and **Zero-Crossing Rate (ZCR)**, which distinguishes voiced sounds (low ZCR, ~100 Hz) from unvoiced fricatives like "sh" (high ZCR, ~3,000 crossings per second).

A frame is classified as speech when it satisfies either: high energy with low ZCR (a voiced vowel) or high energy with high ZCR (an unvoiced consonant). This approach runs in O(n) time, adds only a few milliseconds of processing, and adapts automatically to the caller's specific acoustic environment by calibrating the energy threshold against the first few hundred milliseconds of the call.

In complex noise environments, Vivik augments VAD with Spectral Centroid analysis (which filters background hiss and tonal noise) and Linear Predictive Coding (which models the human vocal tract to distinguish speech from sirens or hum).

> *"By detecting end-of-speech the instant the caller stops — not 200ms later — Vivik's VAD shaves up to 300ms off every response latency."*

Section 7

System Architecture

The following diagram illustrates the Vivik two-worlds architecture: the deterministic media plane (dashed boundary) containing the Rust Telephony Bridge and the telephony endpoint, connected via the NATS Event Bus to the Go API Server and the Integration Runtime in the orchestration layer below.

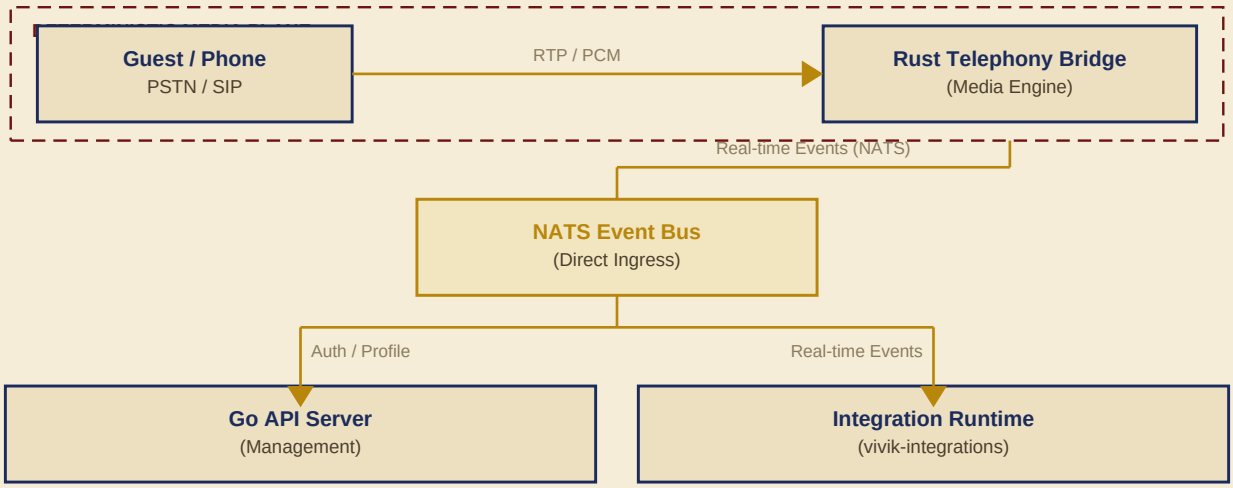

Figure 1: Vivik Architecture: Deterministic Media Plane and NATS Event Pipeline

Section 8

Engineering for Sub-500ms Response Latency

In human psychoacoustics, a delay of 250 ms is perceived as instantaneous. At 500 ms it is noticeable but acceptable. At 800 ms the conversation begins to feel strained; at 1.5 seconds the conversational illusion is completely shattered.

Achieving a sub-500 ms end-to-end response demands uncompromising focus on every millisecond in the pipeline. The five stages of the AI voice pipeline, and how Vivik optimises each:

ASR Processing (100–300 ms): Accelerated by deterministic VAD, which signals end-of-speech the instant the caller stops — not after a 200ms silence timeout.

LLM Inference (200–800 ms): Optimised via prefix caching and smaller quantised models for simple turns, reserving reasoning models for complex logic.

TTS Synthesis (100–400 ms): Streaming TTS begins playing audio as soon as the first words are synthesised, rather than waiting for the complete response.

Network Round Trip (50–200 ms): Minimised by co-locating the Media Engine in the same data centres as the telephony core (Telnyx / Twilio) and AI inference clusters.

Orchestration Overhead (0–100 ms): Effectively eliminated by the Two-Worlds Principle and lock-free Rust/Go communication over NATS.

Section 9

Conclusions: The Future of Native Telephony AI

Vivik represents the convergence of traditional digital signal processing and modern generative artificial intelligence. By rejecting the bolt-on approach and adhering to first-principles engineering, the platform addresses the fundamental constraints of the telephony medium. The Two-Worlds Principle provides the architectural rigour to isolate the non-deterministic Control Plane from the deterministic Media Plane.

The engineering choices — Rust for the Media Plane, Go for the Control Plane, NATS for distributed orchestration, and Kaiser-windowed sinc resampling for signal fidelity — are not preferences but mathematical necessities. Together they enable a conversational experience that is not merely “fast,” but truly natural.

As AI models continue to evolve toward native speech-to-speech architectures, the only constraint that remains is the network and the platform that governs it. Vivik’s deterministic architecture ensures that as the intelligence of the AI grows more powerful, the nervous system beneath it can deliver that intelligence with the speed and reliability of a human voice.

Learn more at bajpailabs.com · hello@bajpailabs.com