FULL-STACK ORCHESTRATION ON HPE PRIVATE CLOUD AI

HPE AI Voice Agent v4.0

Architecture & Engineering Deep Dive

METADATA.JSON

```
// SYSTEM MANIFEST
{
  "version": "4.0.0 (Stable)",
  "architecture": "Event-Driven WebSocket Pipeline",
  "latency_target": "< 1000ms end-to-end",
  "state_management": "PostgreSQL Persistent",
  "deploy_target": "HPE Private Cloud AI"
}
```

REV 4.0 | GENERATED: 2024-05-15T10:00:00Z

A System of Systems: The v4.0 Tech Stack

Orchestrating four powerful engines into a unified agent.

INPUT
(THE EAR)

Whisper ASR

99+ Languages
`ASR_SERVER_ADDRESS`

INTELLIGENCE
(THE BRAIN)

Llama-3.2-1B-Instruct

Orchestrated via API
`llm_model_name`

SYNTHESIS
(THE VOICE)

XTTS v2

Neural Voice Cloning
Latency Optimized

MEMORY
(THE HIPPOCAMPUS)

PostgreSQL

Transactional State
Customers & Tickets

Differentiation: Moving beyond 'Push-to-Talk' to continuous, hands-free 'Conversation Mode'.

Frontend Engineering: The 'Hands-Free' Experience in Inter Tight

Browser-based intelligent voice activity detection.

Client-Side VAD

JavaScript injection prevents server overload.

```
window.vad = {
  thresholdStart: 20,
  thresholdStop: 12, // Hysteresis
  noiseFloor: 0
};
```

Noise Calibration

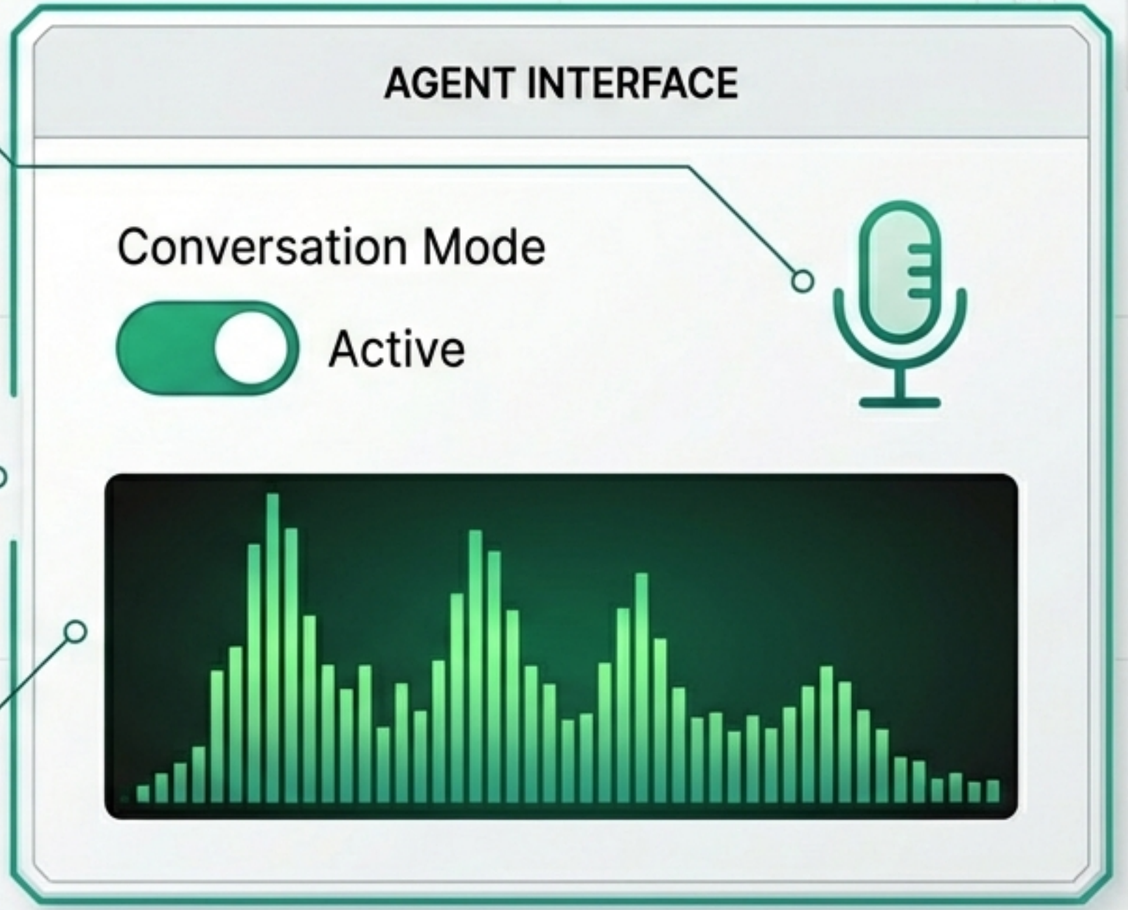
Dynamically calculates noise floor to ignore background hum.

```
calibrateNoise() // Uses 30th percentile of samples
```

Canvas Rendering

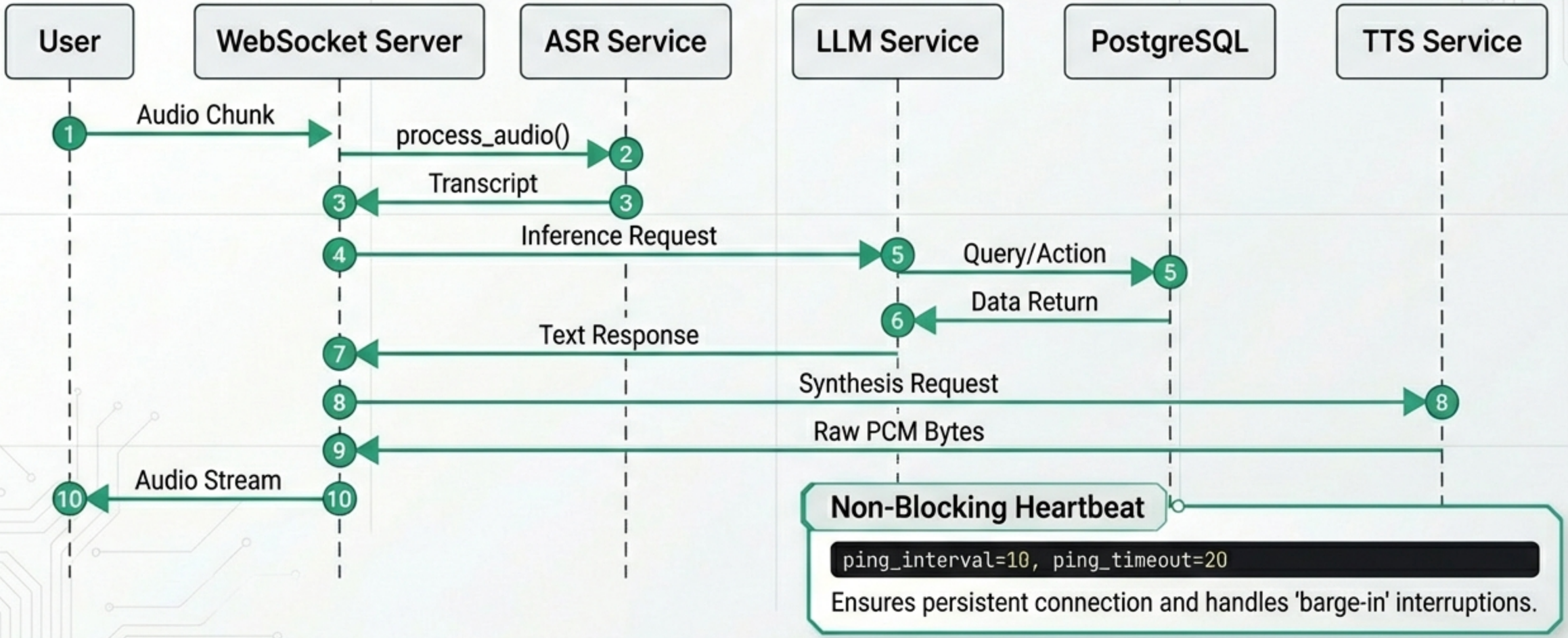
Real-time frequency analysis.

```
analyser.getByteFrequencyData(dataArray);
```



The Intelligent Pipeline: Event-Driven Orchestration

Asynchronous data flow via WebSockets.



Decoupled Multilingual Architecture

Strategy: Understand Global, Speak Local.

INPUT (Understanding)

Supported: 99+

```
ASR_LANGUAGES = {
  'he': 'Hebrew (עברית)',
  'zh': 'Chinese (中文)',
  'ar': 'Arabic (العربية)'
}
```

LLM Context Injection

SYSTEM PROMPT:

"Language Rule: You MUST respond ONLY in [Target Language]. Even if user speaks Hebrew, reply in English."

OUTPUT (Speaking)

Supported: 16
(High Fidelity)

```
TTS_LANGUAGES = {
  'en': 'English',
  'es': 'Spanish',
  'fr': 'French'
}
```

The Brain: Dynamic Persona Injection

Role-playing engine defined by strict prompt engineering.

🤖 Smart Assistant

```
"template": "Keep it SHORT - 1-3 sentences max. Sound like a real person, not a script."
```

📞 Customer Service Pro

```
"template": "Warm and professional... helpful colleague, not corporate robot."
```

💼 Business Advisor

```
"template": "Direct, actionable insights... confident but not arrogant."
```

```
def build_final_prompt(persona, lang):
    return AGENT_PERSONAS [persona]['template'] + \
           TTS_LANGUAGES[lang]['response_instruction']
```

The Voice: High-Fidelity XTTS v2 Server

Engineering for low-latency audio synthesis.

L40S GPU Container

FastAPI Server
(XTTSManager)

Stream
Generation

Waveform Processing

```
def strip_wav_header(data: bytes):
    # Remove 44-byte RIFF header
    return data[44:]
```

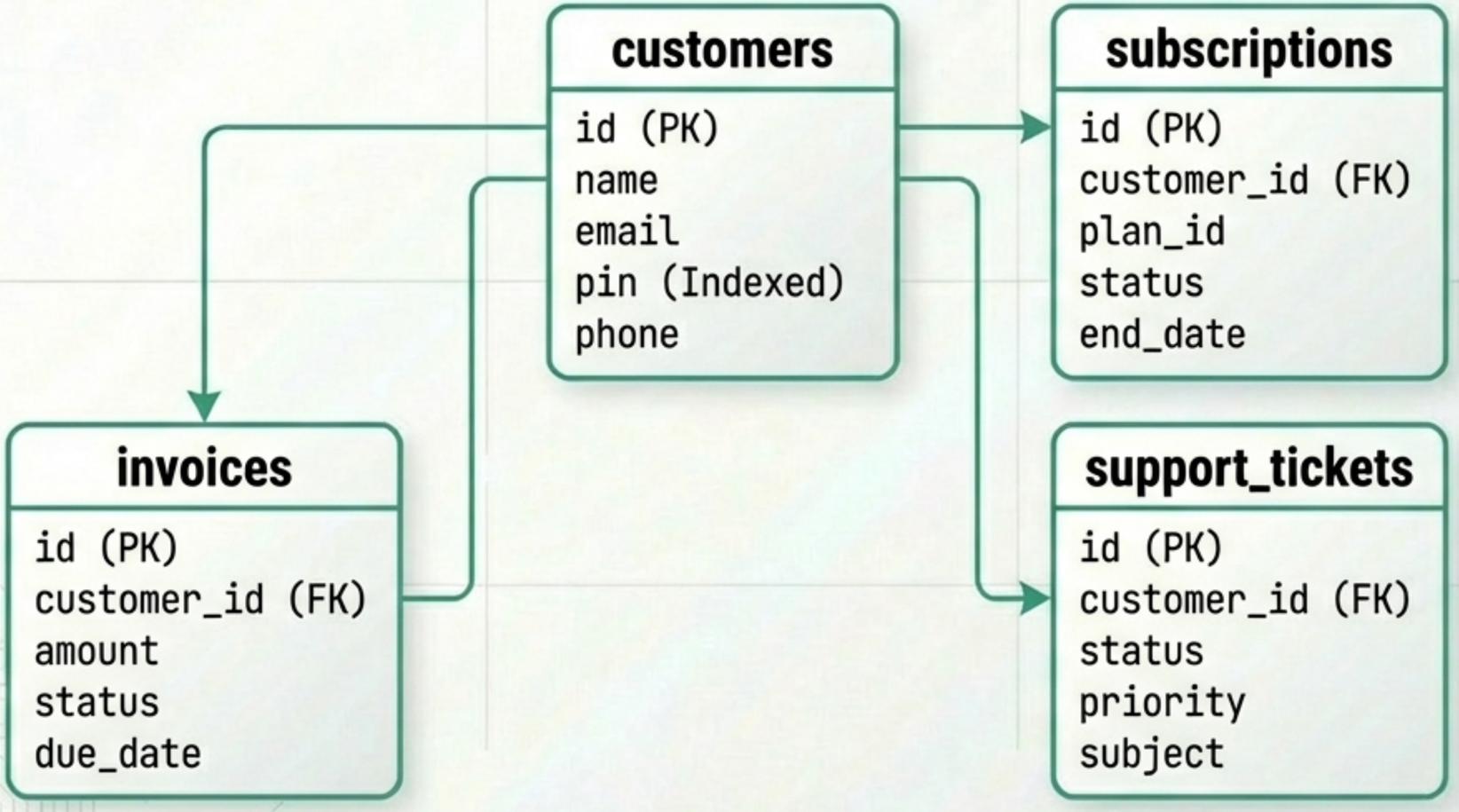
Eliminates file I/O overhead.
Streams raw PCM directly.

Output Specs

- Format: Raw PCM (16-bit signed)
- Sample Rate: 24000Hz
- Channels: Mono
- Voice Cloning: < 6s sample required

The Memory: Stateful Database Integration

PostgreSQL Schema for persistent customer context.



Session Persistence:

Context survives connection drops via `_customer_info_cache` tied to session ID.

Transactional Voice Capabilities

From conversation to execution: The Upgrade Request Workflow.

Step 1:

Voice Command: “Upgrade me to Premium.”

Step 2:

Python Logic

```
if 'upgrade' in intent:
    db.execute('UPDATE subscriptions SET plan_id=3...')
    db.execute('INSERT INTO upgrade_requests...')
```

Step 3:

customer	target_plan	status	priority
J. Smith	Premium	pending	normal

Row Added

Step 4:

Human-in-the-loop: Request appears in “Upgrade Requests” dashboard.

v4.0 Engineering: Stability & Resilience

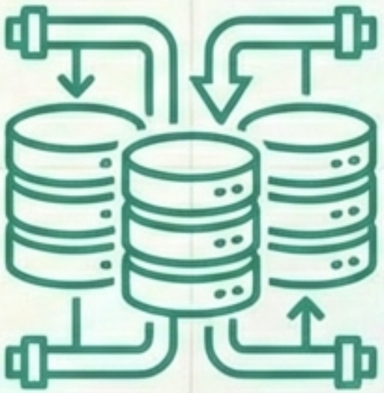
Hardening the system for production loads.

Circuit Breakers

Prevents cascading failures.

```
def check_service_status(url, timeout=10.0):
    # Handle 5xx vs ConnectionRefused
```

Connection Pooling



Prevents socket exhaustion.

```
httpx.AsyncClient(
    limits=httpx.Limits(max_keepalive=20)
)
```

Health Checks

Real-time service monitoring.

```
endpoints = ['/health', '/v1/models']
status = '✅ Online' if response.ok else '❌'
```

Performance Optimization & Caching

Minimizing latency through multi-tiered strategies.

Memory Cache (RAM)

Customer Info & Session Context

```
_customer_info_cache = {} # TTL 3600s
```

Disk Persistence

Metrics & Analytics recovery

```
json.dump(_metrics_cache, f)
```

SQL Query Cache

Frequent read operations

```
SimpleCache(default_ttl=30)
```

Latency Report

ASR Processing	:	**0.42s**
LLM First Token	:	**0.85s**
TTS First Chunk	:	**0.35s**
Total Round-Trip	:	**1.62s**

Security & Authentication

Protecting customer data in voice workflows.

Voice Authentication

```
SELECT name FROM customers
WHERE pin = %s
  AND phone = %s
```

Role-Based Access Control (RBAC)

Admin User

Full DDL Access (Create/Drop DB)

App User

Restricted DML Access (Select/Update)

Network Isolation: All internal traffic routed via Kubernetes DNS `.svc.cluster.local`.

Advanced Analytics: Sentiment & Churn Risk

Turning conversations into business intelligence.

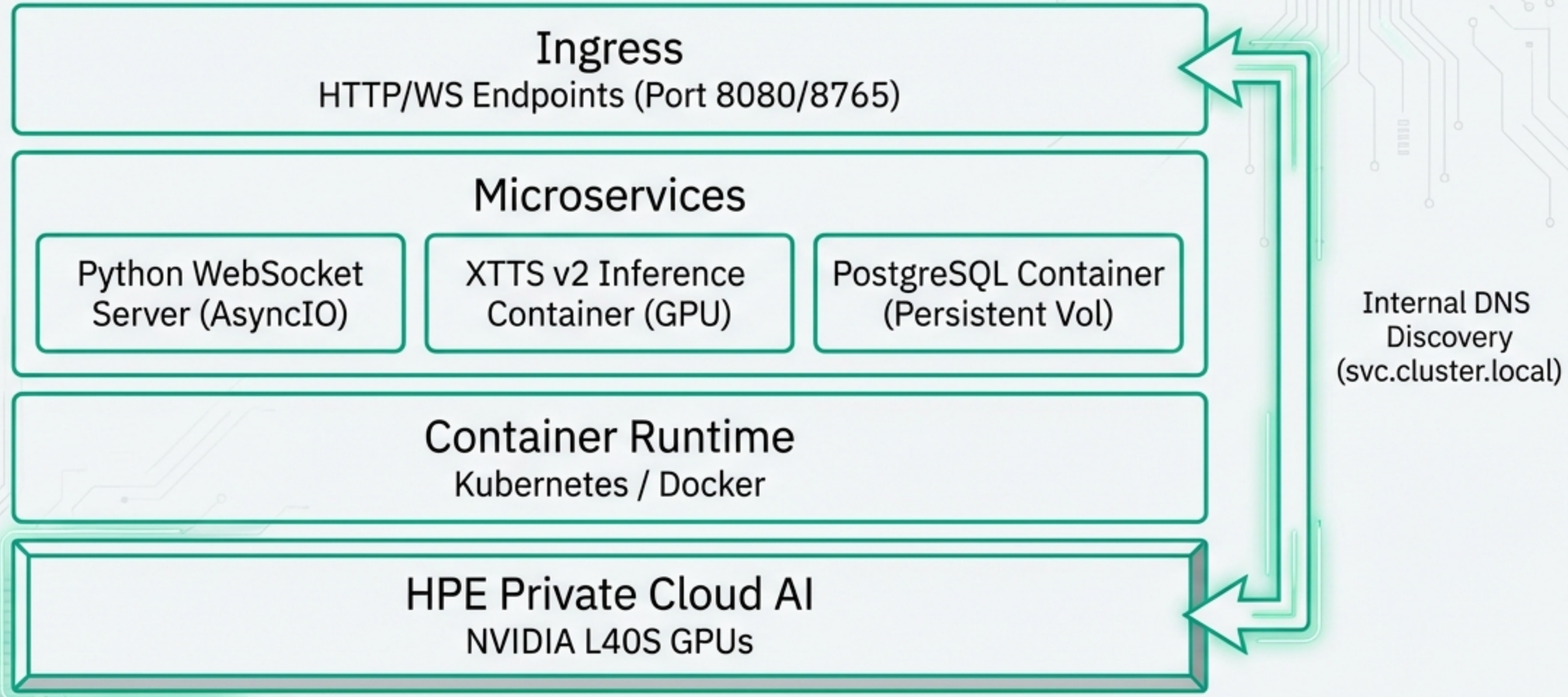
sentiment_alerts

High Alert	**Sentiment Score:** 0.15 (Very Negative)	**Churn Risk:** HIGH 🔴	**Trigger Phrase:** "cancel my subscription"	**Action:** Manager Intervention Required
Medium Alert	**Sentiment Score:** 0.45 (Frustrated)	**Churn Risk:** MEDIUM 🟡	**Trigger Phrase:** "too expensive"	**Action:** Offer Discount

```
CREATE TABLE sentiment_alerts ("
    sentiment_score DECIMAL(3,2),
    churn_risk VARCHAR(20),
    trigger_phrases TEXT
");
```

Infrastructure & Deployment

Orchestrated on HPE Private Cloud AI.



Summary: The “Full-Stack” AI Advantage

The shift from chatting with AI to working with AI.

CONVERSATIONAL

Natural flow, interruptions, multi-lingual.

TRANSACTIONAL

Real SQL execution, stateful memory.

RESILIENT

v4.0 Architecture: Circuit breakers, pooling.

OBSERVABLE

Full latency telemetry and health checks.

HPE AI Voice Agent v4.0: Production Ready