ASHWIN GUPTA

REAL-TIME MEDIA & AI SYSTEMS ENGINEER

India · +91 79966 01575 · ashwingupta3012@gmail.com · LinkedIn · GitHub · ashwingupta.dev

PROFILE

Real-Time Media & AI Systems Engineer with nearly 4 years of experience. **PJSIP/SIP-RTP**, secure media negotiation, Linux networking, asynchronous concurrency and cloud operations. Independently productionized enterprise voice infrastructure for India/UK; **1,600+ concurrent sessions**, **7× call density**, **~$1.3M projected annualized compute savings**.

SELECTED PRODUCTION IMPACT - HSBC CONVERSATIONAL ANALYTICS | COFORGE

AI Engineer · Real-Time Voice & Media Infrastructure Jun 2024 – Jun 2026

Production stack: PSTN/SBC → Kamailio SIP routing → PJSIP/PJSUA2 media workers (RHEL/GCE) → STT/LLM services (GKE). Case study: ashwingupta.dev/work/hsbc

Media transport, security & interoperability

- **PJSIP/PJSUA2 · SIP/RTP: Led a four-engineer integration team** and Implemented end-to-end media transport; deployed across India/UK banking operations at **1,600+ concurrent sessions.**
- **SDES-SRTP · SDP offer/answer: Independently integrated SRTP with PJSIP**; isolated cipher-suite mismatch and **resolved SDP negotiation failures and encrypted media establishment on SDES.**
- **G.729 · PJMEDIA:** Fixed codec decode/negotiation mismatches in SDP via **codec-priority and media configuration on PJMEDIA**; built **custom multichannel SIPp scenarios**, orchestrated with **tmux**, to simulate upto **2,000 concurrent calls**.

Concurrency architecture & capacity economics

- **CPU profiling · Python GIL: 20 calls per 32-vCPU VM** despite low aggregate CPU; top/htop isolated a single saturated core and thread-bound SIP session processing.
- **asyncio + uvloop · taskset · systemd:** Replaced thread-bound media workers with **8 async processes spawned and CPU-pinned via taskset**; **systemd managed per-process health monitoring and off-hours restarts**; achieved **140–160 calls/VM (~7×)** with **sub-second SIP integration TAT** and **<5% packet loss** under measured load.
- **Kamailio · GCE rightsizing:** Redistributed SIP sessions and shifted projected fleet from **80+ 32-vCPU VMs → ~15 8-vCPU VMs**; estimated **compute savings $118K → $8K/month (~$1.3M+/year)**.

Operational reliability & platform lifecycle

- **GCP Cloud Logging · trace correlation:** Unified SIP/GCE, GKE, network and AI service logs by **conversation ID / agent DNIS**; processed **250K+ lines in 5–6s**; reduced incident resolution **1–2h → ~10 min**.
- **RHEL 8 → 9 · Packer · nftables: Migrated and reproduced the complete SIP/RTP stack**; replaced firewall-cmd with persistent nftables rules and rebuilt machine images; verified **full five-step SIP negotiation**, **15-min session refresh** and **media ingestion in manual and Packer automation builds.**

PROFESSIONAL EXPERIENCE

AI Systems Engineer | Skan.ai Jun 2026 – Present

- **Python · Playwright · Win32/ctypes:** Convert client clickstreams to **SOPs and BPMN**; implement synthetic-data workflows reproducing humanized Windows/web actions.
- **Agentic process validation:** Build supporting web apps; reconcile VA event capture and **Agentic Operating Procedure (AOP)** outputs against defined process behavior.

AI Engineer | Coforge - Enterprise AI delivery Jun 2024 – Jun 2026

- **Azure topology intelligence:** LLM-driven extraction and validation; reduced infrastructure analysis turnaround **2–3 days → 2–3 hours**.

- **Contract intelligence RAG:** Shipped enterprise contract retrieval; **~96% numerical extraction accuracy** for sales and support workflows.

Data Scientist | Gida Technologies Jan 2023 – May 2024

- **Multilingual RAG & developer APIs:** Supported **163+ languages**; RAG service evaluated at **~97% factual accuracy** alongside chatbot/content and API tooling.
- **Graph recommender · NVIDIA T4:** Weighted skill graph; **<50ms inference** in testing; **~30% accuracy improvement**.

INDEPENDENT ENGINEERING & OPEN SOURCE

agent-control | Go-based Linux supervision & execution runtime · GitHub

- **Go · Linux · Unix IPC:** Authenticated supervisor/daemon with task-scoped filesystem operations, policy-gated approvals, resource bounds, durable audit/recovery tests. **Active development**; host integrations/release gates incomplete.

controla | Local-first inference control plane · GitHub

- **CPU/GPU inference control plane:** Provider-agnostic routing, **VRAM-aware model selection**, bounded concurrency, fallback policy and execution telemetry; designed for reproducible local scheduling.

PageIndexOllama | Provider-agnostic document reasoning runtime · GitHub

- **Local-first inference runtime:** Decoupled PageIndex from provider-specific APIs for **Ollama**; normalized output/finish-reason contracts, governed prompts and bounded async execution; expanded integration/e2e tests.

TECHNICAL SKILLS

Real-time media & protocols SIP/SIPp, RTP, SRTP (SDES), SDP offer/answer, PJSIP/PJSUA2, PJMEDIA, G.729, Kamailio, SBC, session refresh, media-path debugging

Systems & programming Python (asyncio, uvloop), Go (independent project), C/C++, Linux/RHEL, Bash, process isolation, CPU affinity, asynchronous I/O, concurrency, performance profiling, nftables

Infrastructure & operations GCP (GCE, GKE, Cloud Logging), Packer, systemd, Docker, Kubernetes, CI/CD, load testing, cross-service observability, incident triage

AI systems Streaming STT integration, LLM orchestration, RAG, model serving, inference routing, evaluation, FastAPI, Redis, FAISS

EDUCATION & PUBLICATION

Executive Diploma - MLOps & Generative AI | IIIT Bangalore, 2025–2027

B.E. - Mechanical Engineering | BMS College of Engineering, 2019–2023

Publication | NCISCT 2022 - Generating MCQs Using Graphs and Language Models