SRI HARSHA CHENNU

Director of Engineering · Senior Staff Software Engineer

GPU & LLM Inference Infrastructure · Kubernetes Control Plane · Custom Scheduler Design · Multi-Tenant SDN & DPU Bare-Metal Cloud
Kirkland, WA · (480) 286-7818 · sriharsha.chennu@gmail.com · linkedin.com/in/chennusriharsha · sriharshachennu.dev

SUMMARY

Engineering leader and hands-on architect with 17+ years in large-scale distributed systems, now owning fleet-scale GPU and LLM inference infrastructure at Coupang Intelligence Cloud: 200+ clusters, ~5,000 H200/B200 GPUs, 99.99% vLLM inference SLA, and $8M+ in documented cost savings. Lead 3 teams across the US, Korea, Shanghai, and India and serve as architectural authority for the company's Kubernetes-native AI compute platform, including a patent-pending resource-composition system load-bearing in production. Write the load-bearing designs and stay close enough to the code to debug a production OVS flow pipeline or an NCCL collective on a B200 fabric.

CORE EXPERTISE

GPU & LLM Compute: Fleet-scale GPU inference (vLLM) with capacity-aware routing · Multi-tenant GPU-as-a-Service (H200, B200) · Fractional GPU · VFIO passthrough · InfiniBand/SR-IOV · NCCL topology-aware placement · Bare-metal & VM GPU isolation

Kubernetes Control Plane: Custom CRDs & controllers · Custom scheduler design (gang, fractional, hierarchical-queue, preemption-and-reclaim) · Admission webhooks · Multi-tenant isolation · etcd performance & migration · Drift detection & self-healing reconciliation

SDN, Overlay & DPU: KVM/OVS/OVN · VXLAN · BGP EVPN · OpenFlow pipeline design · DPU (BlueField) host-network offload · NVIDIA DOCA · IPAM for secondary interfaces · Multi-tenant network policy & egress accounting

Leadership: Org design across 4 geos · ARB ownership · Cross-org alignment via written architecture · Promotion sponsorship · Incident command · Partner-engineering with NVIDIA · Roadmap & OKR strategy

Languages & Tooling: Go · Java · Python · gRPC/Protobuf · Linux kernel & networking internals · Prometheus/Grafana · Terraform · Helm · ArgoCD · AWS EKS

PROFESSIONAL EXPERIENCE

Director of Engineering · Senior Staff Software Engineer 2020 – Present
Coupang Intelligence Cloud (CIC) — Seattle, WA

- Own architecture and production operation of CIC's GPU and LLM inference platform — the Kubernetes-native AI compute substrate underneath Coupang's model training and serving. Manage 3 teams across the US, Korea, Shanghai, and India; serve as architectural authority across the org (design reviews, ARB, NVIDIA partner-engineering).
- **LLM inference control plane:** own end-to-end architecture of a Kubernetes-native control plane with declarative CRDs and continuous reconciliation across 200+ clusters, running 10 dedicated GPU clusters (~5,000 H200/B200 GPUs) on vLLM at 99.99% SLA with capacity-aware routing and fractional GPU scheduling.
- **Patent-pending resource composition:** designed the CompositeApplication CRD — a single declarative spec for heterogeneous multi-resource AI/ML applications with controller-enforced lifecycle across compute, storage, networking, and identity; production-load-bearing and the basis for tenant-facing APIs.
- **Custom GPU scheduler:** designer and primary author of CIC's custom Go scheduler (inspired by NVIDIA KAI-Scheduler) — gang scheduling with transactional all-or-nothing allocation, fractional GPU shares, hierarchical fair-share queues, priority preemption-and-reclaim, and an async binding architecture decoupled from API-server latency.
- **DPU-assisted bare-metal cloud:** architected a DPU-offloaded bare-metal platform that moves the entire host network and storage data plane onto DPU hardware (hardware-offloaded OVS) at near-line-rate with negligible host CPU — owning DPU lifecycle (firmware/OS via Redfish, network boot via DOCA SNAP), tenant IP/OS mobility, and active/standby dual-DPU failover.
- **Multi-tenant GPU VM SDN:** defined the production network architecture (KVM/OVS, VXLAN, BGP EVPN) and end-to-end tenant isolation — per-tenant identity, IPAM for InfiniBand/SR-IOV secondary interfaces, MAC scheme, and egress accounting; own the CNI/Calico NetworkPolicy framework and act as incident commander for fleet networking events.
- **B200 enablement:** led B200 SKU enablement as a strategic OKR; validated single-GPU throughput matching the NVIDIA bare-metal baseline, closing the VM-vs-bare-metal performance gap, and drove a multi-SKU abstraction across H200/B200.
- **Vendor decoupling & cost:** removed NVIDIA Base Command Manager dependency from the Kubernetes layer (including etcd migration), eliminating licensing and moving upgrade cadence under the platform team; delivered $8M+ in documented infrastructure savings informing GPU capacity planning.
- **Earlier at Coupang (Catalog ML):** led the team that built a production duplicate-item-matching platform (image + text deep embeddings, FAISS) over 50M+ item catalogs at 3,500 RPS, delivering a 106% recall improvement over the Elasticsearch baseline. Published on the Coupang Engineering Blog (2022).

Senior Software Engineer 2018 – 2020
Amazon Web Services (AWS) — Seattle, WA

- Delivered core components of AWS WAF v2 (WAFV2) and its integration with AWS Firewall Manager — the centralized policy plane that lets enterprise customers manage WAF rule groups across an entire AWS Organization (every account, every Region) from a single administrator account.
- Built the policy-distribution path that propagates AWS Managed Rules (OWASP Top 10 Core Rule Set, SQLi, IP reputation, and Marketplace partner rule sets) to customer web ACLs across child accounts via AWS Config-driven

detection and auto-remediation — eliminating per-account WAF rule deployment toil for customers running hundreds of accounts.

- Operated the service at the multi-billion-request-per-day scale of AWS WAF's evaluation tier; owned production on-call, incident response, and the engineering standards for production readiness — including the count→block rule-deployment safety pattern that protects customers from false-positive outages when enabling new managed rules.

Senior Software Engineer 2015 – 2018
Microsoft — Bellevue, WA

- Led a team of 3 delivering reliability and scale improvements across Dynamics CRM Online, serving enterprise tenants on a high-availability SaaS platform; defined the reliability architecture for the CRM communications platform.
- Designed and shipped a self-healing fault-recovery system that automatically detected and remediated platform failures, reducing manual on-call intervention and improving tenant-facing availability.
- Built an enterprise email dispatch platform that improved delivery SLA and throughput for high-volume customer communications, hardening a previously fragile delivery path under production load.

Lead Software Engineer (SWE → Sr. SWE → Lead) 2009 – 2015
Intel Corporation — Folsom, CA

- Advanced through three levels of scope at Intel Foundry Services, owning platform delivery from inception to production — generated $3M in new revenue and ~$1.2M/year in sustained savings through outage prevention; established reusable integration services and engineering practices.

PATENT

- **Patent-pending:** Dynamic Template-Based Resource Composition System (CompositeApplication CRD), Coupang — Kubernetes-native declarative composition and reconciliation of heterogeneous multi-resource GPU workloads.

EDUCATION

M.S., Information Systems & Statistics · Arizona State University 2008
B.E., Electronics & Communication Engineering · Andhra University 2006

CERTIFICATIONS

Certified ScrumMaster (CSM) · ITIL Foundation · Six Sigma