

ABHISHEK SHAH

Sunnyvale, CA | (650) 630-9280 | shahabhishek@gmail.com | linkedin.com/in/shahabhishek | abhishekshah.dev

SUMMARY

Senior engineering leader with 25+ years architecting and operating cloud compute and AI platforms at hyperscale. Currently **Senior Director at Coupang Intelligent AI Cloud (CIC)**, leading three engineering organizations — **API Platform, Kubernetes Platform, and Virtualization** — that together deliver Coupang's internal GPU cloud and broader compute substrate. Lead Software Architect of CIC, founding member of the program, technical owner of CIC's Kubernetes platform (Cortex) and the CompositeApplication CRD primitive (3 patents filed). Prior tenures at **Netflix** (designed and led the OpenConnect CDN control plane), **Facebook/Meta** (led the ad-audience platform processing a trillion updates per day), **Google** (built the original Kubernetes L4 SDN and DNS — code paths still running in every Kubernetes cluster today), and **Roblox** (next-gen pub/sub at 5M msgs/sec). Operate with influence across software, infrastructure, security, and partner-engineering organizations; partner-engineering relationships at the architecture level with NVIDIA, AWS, and Run:AI. Coupang Bar Raiser for senior and principal hiring.

CORE EXPERTISE

GPU & AI Compute

Multi-tenant GPU-as-a-Service (H200, B200), distributed training and inference substrate, fractional GPU allocation, NVIDIA partner-engineering, Run:AI integration and federation, NCCL/InfiniBand topology awareness, bare-metal GPU isolation.

Cloud at Scale

150+ AWS EKS clusters, 5K+ compute nodes, multi-AZ HA/DR, static-stability architecture, 99.99–99.999% availability, Blue-Green deployment via Temporal, capacity planning for peak events.

Leadership

40-engineer organization across US, Korea, India, Shanghai. Three teams (API, Kubernetes, Virtualization). Written-context leadership style. Cross-functional alignment through design authority. Bar Raiser for principal hires.

Kubernetes & Container Platforms

CRD primitives, multi-tenant isolation, container runtime, pod lifecycle, CNI networking, node OS, image distribution. Original author of Kubernetes L4 SDN and DNS at Google.

Distributed Systems Foundations

CDN control planes (Netflix OpenConnect at 1/3 of US peak internet traffic), trillion-updates/day ad-audience platforms (Meta), pub/sub at 5M msgs/sec (Roblox), petabyte-scale live data migrations with zero downtime.

PROFESSIONAL EXPERIENCE

Principal Engineer / Senior Director, **Coupang Intelligent Cloud (CIC)**

Mountain View, CA

Oct 2022 – Present

- **Founding architecture and delivery of CIC GPU Cloud:** Founding member of the CIC program; shipped multi-AZ, multi-SKU GPU provisioning in a single quarter (Q1 2025). CIC now supports H200 and B200 GPU SKUs end-to-end through a unified declarative API surface. Software-defined, API-driven cloud model replaces ticket-based provisioning; resource requests reconcile through controllers with no human handoff in the customer path.
- **NVIDIA partner-engineering at architecture level:** Drove design discussions with NVIDIA's Run:AI team to ship a programmatic federation API for identity-system integration with CIC's core identity infrastructure — replacing

static credentials with token exchange. NVIDIA implemented and shipped the API; CIC's security posture improved meaningfully. Resolved GPU scheduling integration issues directly with NVIDIA architects.

- **Static-stability and Blue-Green for Coupang EKS Compute Platform:** Introduced a static-stability architecture for CSP-tier services on the Coupang EKS Compute Platform, replacing manual zone-failure recovery with auto-healing backed by a single AWS Target Group model, pinning Coupang's 99.99% resiliency goal. Designed and implemented Blue-Green deployment on top of open-source Temporal workflow system with health-based promotion gates.
- **Galaxy → Coupang EKS Compute Platform migration:** Led the migration of Galaxy — Coupang's largest latency-sensitive application set — onto the Coupang EKS Compute Platform as part of company-wide compute centralization. Scaled the platform 2.5x its original compute footprint. Coordinated across S&D, CMG, Catalog, Tech Infrastructure, and Security organizations with zero migration incidents.
- **AWS Idle-CPU root cause analysis:** Led the deep-systems investigation during the Galaxy migration that traced a serious performance regression on newer AWS VMs to a default Idle CPU configuration. Worked directly with AWS engineering leadership to fix it, resulting in doubled performance, enabled migration to newer VM types, and saved ~10% compute costs. Invited to present the work at an AWS conference.
- **Leadership:** Manage three engineering teams (API Platform, Kubernetes Platform, Virtualization) totaling roughly 40 engineers distributed across the US, Korea, India, and Shanghai. Serve as Lead Software Architect and platform-level technical owner. Coupang Bar Raiser for senior and principal-level hiring across engineering and TPM.

Technical Director, Data Storage — Roblox

San Mateo, CA

Nov 2021 – Oct 2022

- Designed a next-gen pub/sub system supporting 50K–100K subscribing processes and 5M messages/sec — the backbone for cache invalidation across Roblox's fleet.
- Root-caused non-uniform load distribution on CockroachDB clusters (unintended binomial query distribution); achieved significant hardware and license cost reduction via deep systems-level debugging in production.
- Revamped backup architecture using Write-Once-Read-Many (WORM) integrity guarantees against ransomware.

Technical Lead, API Systems & Streaming Infrastructure — Netflix

Los Gatos, CA

Jan 2011 – Nov 2012 & Nov 2019 – Nov 2021

- **Netflix OpenConnect CDN control plane (founding architect):** Designed and led the control plane interacting with thousands of edge and origin servers to optimally route Netflix streaming traffic — at peak, approximately one-third of US internet traffic. Predetermined-placement architecture against a predicted demand surface, rather than reactive caching using BGP-aware steering.
- **BFF fanout optimization:** Identified excessive fanout from Node.js BFF systems and reduced via batching, saving \$1M+/year in infrastructure cost.
- **Metacat → Apache Iceberg:** Extended Netflix's metadata platform to support Apache Iceberg; contributions merged at Netflix/metacat. Built the movie-stream URL generation library used for Netflix's entire catalog.

Technical Lead, Audience Infrastructure & Analytics — Facebook (Meta)

Menlo Park, CA

Oct 2016 – Nov 2019

- Led the ad-audience platform processing a trillion updates per day across millions of audiences.
- Led the Audience Size Estimation System read-path: query-time sampling architecture on ZippyDB columnar substrate, aggregator-leaf horizontal scaling (32 partitions), operating at 40K QPS, p50 20ms / p90 100ms, and 2 PB of storage across three regions.
- Directed a petabyte-scale live migration from HBase to RocksDB with added privacy controls and zero downtime across dependent systems.

Technical Lead, Kubernetes — Google

- **Created the software-defined networking layer for Kubernetes:** Designed the L4 SDN giving Kubernetes services stable virtual IPs backed by real endpoints. Code path still running in every Kubernetes cluster in production today.
- **Built the Kubernetes DNS server:** Developed the DNS server for service discovery across microservices, forming a core primitive of the Kubernetes data plane.

Earlier Roles

- **Technical Lead, Smart Allocation — Walmart Labs** (Nov 2012 – Jul 2013): Mixed-Integer-Programming supply-chain allocation models saving millions in shipping costs.
- **Software Developer II, Visual Studio — Microsoft** (Oct 2006 – Sep 2011): Custom virtualized WPF control rendering thousands of shapes in graphical designers (Visual Studio 2012 UML).
- **Software Developer, Ordering Platform — Amazon** (Dec 2004 – Oct 2006): Delivered RAPID, a Tier-1 partition-aware routing service core to Amazon.com ordering with high-availability SLAs.
- **Senior Software Engineer** at Visible Technologies (2010-2011); **Software Developer** at Synogy (2002-2004).

PATENTS & OPEN SOURCE CONTRIBUTIONS

CompositeApplication CRD — Kubernetes custom resource for declarative multi-resource application composition (patent filed, Coupang).

Open-source contributions: Netflix/metacat (Apache Iceberg support, merged); AWS Load Balancer Controller (multi-port, multi-cluster ingress patch, production at Coupang).

EDUCATION

M.S. Computer Engineering — University of Illinois, Chicago

B.Tech — Veermata Jijabai Technological Institute (VJTI), Mumbai University