Supreet Sethi

Mobile: +61-488-627-349 | Email: supreet.sethi@gmail.com | Linkedin: https://www.linkedin.com/in/djinn | Greater Melbourne Area

SUMMARY

Experienced Software Engineering Leader building scalable AI and cloud platforms that empower businesses and communities. Delivered infrastructure supporting 2% of Indonesia's GDP, scaled 2000+ GPU workloads, and led cloud-native transformations for FSI and startups. Passionate about mentoring engineers, managing high-performing teams, and partnering with stakeholders to deliver measurable customer value enhancement through GenAI solutions.

SKILLS

Leadership

- People Management: hiring, promoting, and mentoring ICs and managers
- Cross Functional team leadership (teams of 10-50 members)
- Technology roadmap crafting
- Remote-first delivery
- Performance management and career development
- Agile transformation and Scrum implementation

Technical

- AI Infrastructure Nvidia DGX (H100/H200), Orin NX on edge
- Agentic AI with OpenAI OSS-GPT-120B and multi-channel functionality using OpenAI Harmony Prompt Format
- Cloud: Kubernetes, AWS/GCP/Azure, GenAI Pipelines, Golang, Python, GraphQL, REST API
- Customer value enhancement through GenAI solutions
- Modern SDLC: GitOps, Infrastructure as Code (IaC), automated testing frameworks
- CI/CD: Jenkins, GitLab CI, GitHub Actions, automated deployment pipelines
- Agile Methodologies: Scrum, Kanban, SAFe, continuous delivery practices

Business/Strategy

- Cost optimization, risk management, and technology governance
- Product Design Thinking
- Business Objectives alignment
- CSuite Engagement
- Executive & Board Engagement
- Partnerships and P&L management

NOTABLES

- **Speaker** at Melbourne Golang Meetup presenting building native LLMs in Golang
- **Speaking Engagements** at AWS Summits and various Startup Events
- **Advisory** to Sequoia, PeakXV and Openspace portfolio company
- **Authored** a book on Solution Architecture skills

EXPERTISE

- Led strategy and scaling for 2000+ GPU (H100/H200) AI workloads
- SME for vLLM, torch, specifically inference performance

- Deep experience in GenAI (RAGs, MCP, inference), CloudOps, multi-region scale
- Executive stakeholder for ASEAN's top startups and cloud clients
- GPU deployments with k8s using device plugin and GPU operators
- Direct contribution to customer value enhancement through GenAI implementations
- End-to-end DevOps pipeline design and implementation
- Agile coaching and digital transformation leadership

CAREER HIGHLIGHTS

- Designed Tokopedia's search system (2% of Indonesia's GDP)
- Led AWS region launches: Malaysia, Thailand, and supported Jakarta launch
- Owned VC portfolio revenue (45% of ASEAN AWS biz during COVID)
- Principal technical advisor for Prudential (8% of Couchbase global revenue)
- Built agentic AI solutions leveraging multi-channel operating functionality

IMPACT

- AI infra setup time: 65 → 15 days (E2E, 2024-2025)
- AI infra consumption scaling 10x, launched pay-per-use inference on Tir (E2E)
- Cloud cost 40% reduction across ASEAN Startups (AWS, 2023)
- Performance 20–25% improvement with Graviton2 adoption (AWS, 2022)
- Startup Transaction scale 10x with modernisation (StashAway, 2020)
- Development velocity 35% improvement through CI/CD automation (Multiple roles)

EXPERIENCE

VP, AI Solutions — E2ENetworks, *Melbourne* | 2024–Now

- Reduced turnaround time from initial consultation to first AI workload migration by 77%, from 65 days to 15 days, through streamlined processes and automation.
- Spearheaded the expansion of GPU-as-a-Service infrastructure, scaling capacity by 10x to meet growing demand for AI and machine learning workloads.
- Built and led a high-performing team of AI architects and engineers, focusing on scalable, high-performance AI infrastructure with direct contribution to customer value enhancement through GenAI.
- Pioneered the development of a pay-per-use inference service, enabling cost-efficient, on-demand AI model deployment for enterprise clients.
- Built agentic AI solutions using OSS-GPT-120B, specifically leveraging multi-channel operating functionality of OpenAI models for enhanced customer outcomes.
- SME in E2E Networks for Nim, Nemo, Agentic AI, and model-specific scaling using k8s.
- Understanding, implementing and sometimes customising Nvidia Blueprints.
- Deployed Deepseek V3, Llama, Microsoft Phi, and Gemma models for enterprise use cases including Chat, Invoice Processing, RAG Pipelines, and Image Identification.
- Integrated Vector Databases (Milvus, Qdrant) for high-performance similarity search in AI-driven applications.
- Designed and implemented ClickHouse-based analytics solutions for real-time data processing and insights.

Director, Solution Engineering APAC — Couchbase, *Singapore* | 2024

- Shared P&L ownership for APAC region alongside sales leadership, contributing to revenue growth and cost optimization strategies.
- Co-built high frequency trade desk with ANZ bank using Couchbase as Database backend.

- For Telstra, assisted in removing performance bottlenecks which stifled their user behaviour data integration with Contact Centre product.
- IOT usecase implementation with Cochlear.
- Architected and deployed largest global cloud implementation for a major financial services client.
- Provided executive coverage for Prudential migrating to Couchbase Capella, enabling them to leverage Vector capabilities both on edge and on the Cloud.
- Created demos showcasing Vector and Edge AI capabilities of Couchbase using Groovy and Python.
- Developed regional technical strategy focusing on cloud-native database implementations with emphasis on cost optimization and risk management.
- Implemented DevOps best practices and automated deployment pipelines for database migration projects, ensuring zero-downtime transitions.

Specialist Strategy Leader — AWS, *Singapore* | 2022–2023

- Matrix managed 40 specialist solution engineers across ASEAN (Association of Southeast Asian Nations: Indonesia, Thailand, Singapore, Malaysia, Philippines, Vietnam, Myanmar, Cambodia, Laos, Brunei), with most having Seattle-based direct managers while serving as their regional manager.
- Led technical discovery and solution design for enterprise-scale transformation projects with focus on technology governance and cost optimization.
- Created Anti-Gambling Use Case solution for Starhub, ensuring compliance with Singapore regulatory requirements.
- Integrated AWS Redshift and Athena for advanced analytics and reporting capabilities in client solutions.
- Advised Digital Natives like ViSenze on VectorDB solutions for AI-driven applications.
- Provided feedback packaging from customers for GenAI workloads to product teams.
- Delivered Product Requirement Packages to AWS product teams - curated lists of key requirements collated from direct customer requirements, operational gaps identified by Specialist Solution Engineers, and ProServe team insights.
- Specific product advisory on newly conceived Bedrock and Sagemaker Studio.
- GenAI compute procurement consultation for customers like DBS and Petronas.
- Domain-specific fine-tuning of SLMs using AWS Trainium.
- Championed adoption of modern SDLC practices including GitOps, automated testing, and continuous integration across client implementations.

Head, Startup Solution Engineering — AWS, *Singapore* | 2020–2022

- Managed 2 direct report managers and 23 total team members across the ASEAN startup ecosystem.
- Provided strategic direction to the Startup team, ensuring alignment with customer needs and emerging technology trends.
- Enhanced customer satisfaction by 20% through targeted mentorship and best practice sharing.
- Increased revenue by 15% through product and service optimizations with focus on cost management and risk mitigation.
- Drove the adoption of AWS Graviton2, resulting in performance improvements for key clients.
- Advised clients on AWS Redshift, OpenSearch and Athena for data warehousing and analytics.
- Implemented hiring, promotion, and performance management processes for both individual contributors and managers.
- Established Agile coaching programs for startup clients, helping them adopt modern software development practices and improve time-to-market.

- Promoted CI/CD best practices across portfolio companies, leading to measurable improvements in deployment frequency and reliability.

Principal Solution Engineering — AWS, *Singapore*| 2019–2020

- Supported startups in scaling on AWS, providing architectural guidance and strategic direction.
- Played a key role in driving a 40% increase in startup clients through targeted support and guidance.
- Acted as a trusted advisor to CTOs, ensuring smooth scaling and infrastructure optimization with emphasis on cost governance.
- Orchestrated migration from Legacy Cassandra database to hybrid PostgreSQL Aurora and Cassandra architecture for StashAway, enhancing operating performance of the roboadvisory platform operating under MAS (Monetary Authority of Singapore) statutes and SAE (Securities and Commodities Authority) regulations in UAE.
- Advised startups on operating models using AWS Redshift and Athena for scale and cost optimization.
- Implemented DevOps transformation initiatives, introducing automated testing, continuous integration, and deployment pipelines that improved development velocity by 35%.

Head of Engineering — Smartkarma, *Singapore* | 2018–2019

- Led engineering team of 10 members at money markets insight provider serving High Net Worth Individuals through private banks.
- Managed partnerships with Société Générale and UBS to serve their customers through compliant platform.
- Built product to comply with money market regulators' statutes in Europe, Hong Kong, and Singapore.
- Grew user engagement 23% and launched analytics-driven research tools.
- Implemented hiring and mentoring processes for engineering talent.
- Transformed development practices by introducing Scrum methodology and automated CI/CD pipelines, reducing release cycles from monthly to weekly deployments.

SVP Engineering — Ralali, *Jakarta* | 2017–2018

- Delivered 24.3% team productivity lift through strategic leadership and process optimization.
- Scaled platform architecture with focus on cost efficiency and operational excellence.
- Implemented Agile methodologies and modern SDLC practices, introducing automated testing frameworks and continuous deployment processes.

Technical Architect — Tokopedia, *Jakarta* | 2015–2016

- Managed 50 individuals across Payments, Search, Discover, Train Ticketing, Logistics and Mobile app teams through 5 direct report managers.
- Architected core search systems under OJK (Otoritas Jasa Keuangan - Indonesia's Financial Services Authority) regulatory compliance.
- Achieved 18.3% increase in GMV through technical architecture improvements.
- Established performance management and career development frameworks for large engineering organization.
- Led transition from monolithic to microservices architecture, implementing comprehensive CI/CD pipelines and automated deployment strategies across multiple product teams.
- Championed DevOps culture adoption, introducing Infrastructure as Code and automated monitoring solutions that improved system reliability by 40%.

Prior Roles (2000–2015)

Engineer and Lead roles across systems, search, and web products with progressive people management responsibilities. Gained foundational experience in Agile methodologies, continuous integration practices, and modern software development lifecycle management.

EDUCATION

- **MA Critical Media & Cultural Studies**, University of London, 2010
- **BCA**, IGNOU, 2007

LANGUAGES

English (Fluent)

TOOLS

Datadog, Docker, Git, AWS/Azure/GCP, Kubernetes, Jenkins, GitLab CI, GitHub Actions, Terraform, Ansible, Jira, Confluence