Omar Sobh

Cloud Innovation and AI Automation Leader

15428 Benedict Ln.
Los Gatos, Ca 95032
(408) 329-0539
om.sobh@gmail.com

Executive Summary

Technical leader with 12+ years designing capacity planning systems for fast-changing computational workloads. Expert in forecasting AI/ML resource allocation, scaling from research environments to hyperscale deployments (Apple, Binance). Built frameworks managing heterogeneous GPU/CPU infrastructure, reducing costs 40% while improving utilization to 85%. Proven ability to navigate ambiguity and influence cross-organizational stakeholders. Currently using Claude Code to build next-generation orchestration achieving 337M ops/second.

EXPERIENCE

Apple, Cupertino, Ca — AI/ML & Infra Intelligence Lead

October 2019 - PRESENT

Multi-Cloud Capacity Strategy
Pioneered capacity planning across AWS/GCP/AliCloud, designing resource allocation frameworks enabling PB-scale processing. Built capacity models balancing security with compute needs, establishing quota management systems for enterprise-wide planning.

AI/ML Workload Optimization
Architected hybrid infrastructure managing 10,000+ concurrent workloads with intelligent scheduling maximizing GPU/CPU utilization. Built forecasting models enabling teams to plan training schedules months ahead. Reduced operational overhead by 80% through AI-driven resource redistribution.

Platform Democratization
Transformed how 1,500+ teams consume resources through self-service provisioning, reducing allocation time from days to minutes. Created capacity transparency enabling data-driven infrastructure investments saving $50M+ annually.

StratoSwarm - AI-Accelerated Platform *(Current Innovation)*

Built a revolutionary capacity management and orchestration system with Claude Code, delivering 60,000 lines of Rust in 3 months.

SKILLS

- Cloud Infrastructure Design & Management
- CI/CD Pipeline Development & Automation
- Artificial Intelligence & Machine Learning Integration
- Scalable Systems Architecture
- DevOps & Agile Methodologies
- Technical Leadership & Team Building
- Geo-distributed Team Management
- Digital Media Production Workflows
- Software Development Lifecycle (SDLC)
- Strategic Planning & Execution
- Innovation & Process Optimization
- Kubernetes & Containerization Technologies
- AWS/GCP/ALI Cloud Services
- Performance Benchmarking & Optimization
- Cross-functional Collaboration
- Project Management & Delivery
- HIPAA and SOX Compliance

Binance, Tokyo, Japan — Senior Infrastructure Manager
July 2018 - Jan 2019

Scaled infrastructure 10x (1M→10M TPS) while reducing costs 40%. Built forecasting models handling 100x traffic spikes. Implemented custom Kubernetes for real-time workloads with deterministic latency.

Guardant Health, Redwood City, CA — Senior Infrastructure Engineer
November 2016 - June 2018

Genomics Capacity Planning

Built the first capacity framework for precision oncology, managing growth to thousands of daily samples. Created predictive models based on sequencer throughput and clinical pipelines. Architected 20-rack/20PB HPC cluster with tiered storage optimizing cost/performance. Partnered with research teams to balance competing priorities, reducing time-to-insight by 65%.

Institute for Genomic Biology (Carl R. Woese), UIUC, Urbana, IL — Lead Infrastructure Engineer

December 2013 - November 2016

Managed capacity for 500+ researchers across 50+ labs within academic budget constraints. Implemented cloud bursting and spot instances (70% cost reduction). Built workload profiles enabling 3x research output within sam

Publications:

Knowledge-guided analysis of "omics" data using the KnowEnG cloud platform:
https://bit.ly/46qEzkt

Invertnet: a new platform for biodiversity research:
https://bit.ly/44MUN6g

LANGUAGES

Arabic, English

EDUCATION

Parkland College, Champaign, IL — *Computer Science*
January 2007 - Jan 2014