Contact

www.linkedin.com/in/swapnilghike (LinkedIn)

Top Skills

GPU
Large Language Models (LLM)
Generative AI

Languages

Hindi
Marathi
English

Honors-Awards

Siebel Scholar, Class of 2012

Publications

A Distributed Algorithm for Pattern Formation by Autonomous Robots with No Agreement on Coordinate Compass

Directive-Based Compilers for GPUs

Pattern Formation for Asynchronous Robots without Agreement in Chirality

Comparing the power and performance of Intel's SCC to state-of-the-art CPUs and GPUs

Swapnil Ghike

Principal Staff Software Engineer at LinkedIn (AI products and platform / Generative AI / RecSys)
San Francisco Bay Area

Summary

I love building intelligent products and the systems necessary to realize them. I have led teams to solve hard technical challenges in distributed systems at the scale of hundreds of millions of users and coached engineers and managers to help them grow and succeed. I thrive in high velocity environments, particularly relishing challenges where the industry standards and product requirements evolve fast (such as Generative AI), and negotiations need to be done between delivering near term business value and building a solid long term foundation. I have also been used as a "fixer" multiple times (such as unblocking major backend migrations, turning deadlocks into problem solving with appropriate tools including escalations). Experienced in working with non-engineering partners.

Experience

LinkedIn
12 years 6 months

Principal Staff Software Engineer
July 2020 - Present (4 years 7 months)
Mountain View, California, United States

One of the tech leads for the company's AI Platform that supports 1000+ AI engineers.

* Tech lead for the Generative AI / LLM capabilities in the AI Platform. In this role, I partnered with the modeling teams to convince leadership to take a strategic bet on fine-tuning open source models and together we were able to satisfy the product needs at 75x cost efficiency compared to the original GPT-4 model.
* responsible for extracting the best unit economics out of a large chunk of company's GPU fleet,

* influencing many strategic technical decisions around LLM and LPM (ranking / personalization) inference, model authoring frameworks, training & fine-tuning capabilities, ML Ops, hardware choices, genAI application stack
* hiring and performance reviews
* cross-company initiatives
* technical bets

Senior Staff Software Engineer
September 2016 - July 2020 (3 years 11 months)
Mountain View, California

1) Notifications Platform:
- Started owning LinkedIn's notifications ecosystem at a time of lack of technical vision, too many disconnected developments, high-touch onboarding, and low client satisfaction.
- Over the last 3 years, my team has built a central notifications platform which supports offline/nearline/realtime use cases, configurability, ML models, operations (delivery time: hours -> seconds/minutes), increasingly self-serve UI-driven integration (onboarding time: months -> 2-4 weeks) & e2e debugging.
- Today, the platform / use cases are the *largest driver of user sessions* and significant contributors to other business metrics.
- Likely one of the most advanced notifications platforms among companies that depend on recommendations or network driven engagement.

2) Marketing Technology: With increasing spend ($$$M/year), the goal was to enable internal marketers to seamlessly reach "the right audience in the right channels at the right time with the right message" to maximize ROI & optimize time to market. The team spanned Marketing Ops, Data Science, Analytics Infra, Enterprise Productivity, Consumer Eng. We put together a design and a beta release to enable self-serve campaign creation, centralize data in a customer data platform (CDP), deliver across multiple channels, recommend / filter using ML, by leveraging in-house & third party technology.

3) Dynamic Configuration: Changing application behaviour using configs can be calendar time consuming due to the number of repositories touched, multiple service restarts, testing and repeating the whole cycle if bugs are discovered. The goal was to minimize this time by eliminating service restarts & providing easy config management. My team worked across orgs to define a phased approach and delivered a dynamic config management experience

natively integrated with LinkedIn's experimentation engine, and onboarded multiple clients.

Staff Software Engineer
June 2015 - September 2016 (1 year 4 months)
Mountain View, California

1) Realtime LinkedIn -> User device Push: Pub/sub and persistent connection infrastructure that powers instant conversation-like experiences by low latency delivery of events to connected apps and also doubles up to avoid capacity overload of polling. I led the design, implementation and rollout of the v1 platform functionalities and instant messaging features. The platform later continued to evolve to support Presence/Chat as a first class feature, and the events traffic increased 10X with newer use cases and organic engagement.

2) Messaging: Data organization, mid-tier, security, technical design reviews, feature development, engineering practices. Co-led the core architecture re-design of LinkedIn's messaging platform with two other tech leads from ground up to replace the legacy stack, scale an order of magnitude.

Working in the Messaging org has been an exhilarating experience as the org scaled from 15 engineers to 55 within one year, much of the software was refreshed, and a good amount of long-lasting leverage was created. The async design review process piloted by our team was later adopted by all product engineering orgs which amounted to ~1.5-2k engineers.

3) Project Voyager: As a short term gig to help ship LinkedIn's new Flagship app, I led the Front-End API server side development of the 'Network Brief/ Recommendations' features. This brief stint provided me a steep learning curve in the FE world and consumer-facing product development.

4) Site Speed: Worked to optimize site speed of various consumer product areas such as Growth, Identity, Feed, Messaging, Contacts/PYMK, Groups, Pulse, etc. Drove large chunks of horizontal optimization efforts such as killing redirects, and contributed to tools for performance bottlenecks detection using real user monitoring data.

Senior Software Engineer
July 2014 - June 2015 (1 year)
Mountain View, California

FollowFeed - Distributed system to ingest and retrieve Feed index with support for filtering, relevance and A/B testing. It replaced a SenseiDB based solution.

From the project's conception, I worked on all aspects of FollowFeed v1.0 - caching and persistence using using embedded KV store (RocksDB), data ingestion and querying, integrating relevance and A/B testing, bootstrap, operationalizing and client onboarding, performance optimizations focused on GC/network/OS/request fanout for mitigating the 99th percentile and increasing Quality of Service.

Major accomplishments: 5X reduction in the 99th pct latency of serving feed index, ability to handle 3X throughput compared to Sensei, 20X longer retention, 50% CapEx reduction.

Software Engineer
August 2012 - July 2014 (2 years)
Mountain View, California

Apache Kafka - Open Sourced distributed publish-subscribe system. Kafka is used by almost every service in production at LinkedIn, and also by many other companies. (kafka.apache.org).

I contributed to the stabilization/operations of Kafka's replication features, wrote a hadoop job to push data from HDFS to Kafka with features such as event count auditing, drove a backward incompatible client rollout with 80 clients, solved some major pain points in tools and auto-deployments.

University of Illinois at Urbana-Champaign
Research Assistant
August 2010 - May 2012 (1 year 10 months)

- Worked with PGI and Cray accelerator compilers that can parallelize and offload compute intensive loops to GPUs with hints from programmer in the form of directives.

- Demonstrated the abilities, or the lack thereof, of aforementioned compilers to automatically apply loop transformations to remove dependences between loop iterations, using a home-made microbenchmark suite.

- Recommended a series of transformations to convert sequential programs to a format compatible for compiling with the PGI and Cray accelerator compilers. Demonstrated their impact on CPU/GPU computation and communication, and on overall performance.

Facebook
Software Engineering Intern
May 2011 - August 2011 (4 months)
Palo Alto, CA

- Designed a NUMA-aware request scheduler for HPHP (HipHop) and performance optimizations.

- A new scheduling and load balancing scheme that binds threads to NUMA nodes and prohibits memory sharing across threads running on different NUMA nodes.

- The performance tests conducted on different types of machines with vanilla linux kernel showed a reduction of 5-10% in cpu time required to satisfy a web request.

Tensilica
Intern
January 2010 - June 2010 (6 months)

- Rendering of image data using a master-slave hierarchy of TX multiprocessors in parallel.
- Implemented a few instructions of JVM on Xtensa Configurable processors.

Indian Statistical Institute, Kolkata
Summer Intern
May 2009 - July 2009 (3 months)

I wrote algorithms for distributed computing by autonomous robots (denoted by points on 2D plane) using essential tools of maths and computational geometry and optimized the time complexity.

Education

University of Illinois at Urbana-Champaign
MS, Computer Science · (2010 - 2012)

Birla Institute of Technology and Science
B.E.(Hons.), Computer Science · (2006 - 2010)