Dibyendu *"Dev"* Nath

email: dev.nath.cs@gmail.com | *cell:* +1.650.279.5722
web: https://devnath.net | *github:* https://github.com/dnath

Software engineer specializing in large-scale generative AI and ads serving infrastructure, with 10+ years experience.

- Leads Gemini serving infrastructure at Vertex AI, Google Cloud; previously led Google Shopping Ads serving.
- Known for rapidly onboarding new product bets, building strong teams, and mentoring engineers into senior roles.

Experience

Google LLC – Mountain View, CA *Aug 2015 – Present*
Staff Software Engineer – Tech Lead / Manager (TLM)

Vertex AI Gemini Serving – Google Cloud *Mar 2025 – Present*

– Tech Lead / manager for Vertex AI Gemini serving infrastructure.

  – Delivering low-latency, multi-region serving of Gemini models for Google Cloud enterprise customers.
  – Responsible for new Gemini model launches on Vertex AI platform, simultaneously shipping with rest of Google, meeting enterprise compliance needs.
  – Building customer-facing APIs and features as well as safety controls, quota, isolation and scaling policy for these models on Google Cloud.
  – Accountable for service reliability and latency– recently achieved 99.99% availability SLO from 99.9% through reliability focussed efforts.

– Rearchitecting serving path for multimodal inputs (e.g. image, video, audio, pdf) for better scaling with efficient load balancing, and custom sandboxed media processing to keep latency low while isolating untrusted content.

– Grew team from 8 to 20+ in less than a year. Currently leads a 20+ person org across two subteams with engineering managers, focusing on execution rigor, incident readiness, and talent growth.

– Instituted review and launch practices that cut model launch velocity from 1 month to 1 week; drove operational excellence, incident readiness and postmortem culture.

Google Ads Shopping Serving *Aug 2015 – Mar 2025*

– Led the Google Ads Shopping Serving engineering team.

  – Owned low-latency backend serving and indexing infrastructure for Product Listing Ads on google.com, that powers $XXB+ of annual ad revenue.

– Drove redesign of serving stack in order to scale existing systems to meet rapid 40% annual growth in shopping offer inventory (XX billions of offers).

  – Met key goal of drastic improvement of end-to-end data freshness, with a reduction by 54% on average and 91% for high impression offers, while serving XXX kQPS with low serving latency (under 250ms).
  – Recognized with a company-wide Google Tech Impact award in 2023.

– Joined as a new-grad engineer and advanced to Tech Lead / Manager (TLM); Built a high-performing team through hiring, mentoring, and growth planning, apart from deep technical contributions.

University of California, Santa Barbara – Santa Barbara, CA *Sep 2013 – Sep 2015*
Research Assistant, RACE Lab – StochSS : Cloud-based Stochastic Simulation as a Service

– Built and maintained StochSS, a cloud service for running stochastic simulations across heterogeneous compute resources; published in PLOS Computational Biology.

AppFolio Inc. – Goleta, CA *Jun 2014 – Sep 2014*
Software Engineering Intern – RentMatch : Appfolio's Pricing Analytics (Data Science) team

– Built machine learning models to score rental unit similarity using text, amenities and census-derived features.
– Developed a distributed data collection and processing framework for running complex data joins across data centers.

McAfee Inc. – Bangalore, India *Feb 2012 – Aug 2013*
Software Development Engineer – Endpoint Encryption for Files and Folders (EEFF) team

– Delivered features across client endpoints and policy management server for enterprise endpoint encryption product.
– Built 'Kill Pill' prototype for remote deactivation and secure wiping of encrypted removable USB devices.

Education

University of California, Santa Barbara – Santa Barbara, CA *Sep 2013 - Jun 2015*
Master of Science, Computer Science

Advisors: Prof. Chandra Krintz, Prof. Rich Wolski

West Bengal University of Technology – Kolkata, India *Aug 2007 - Jul 2011*
Bachelor of Technology, Computer Science & Engineering

Skills

– Team and org leadership: vision & strategy planning, building roadmaps, day-to-day project management, execution rigor and incident management with a focus on org health.
– Building and growing teams; coaching engineers to senior/staff; fostering a strong engineering culture.
– Distributed and parallel computing, high-throughput serving, low-latency optimization, streaming data pipelines, LLM serving/inference platforms, large-scale indexing and caching systems.
– *Programming:* Extensive experience in C++, C, Python. Proficient in Java, SQL, Go, shell scripts, JavaScript.
– *Protocols & Frameworks:* Protocol Buffers, Stubby/gRPC, HTTP, MapReduce/Flume/Hadoop, memcached/redis.

Publications

Drawert, B., Hellander, A., Bales, B., Banerjee, D., Bellesia, G., Daigle Jr, B.J., Douglas, G., Gu, M., Gupta, A., Hellander, S., Horuk, C., Nath, D., et al 2016. **"Stochastic Simulation Service: Bridging the Gap between the Computational Expert and the Biologist."** PLOS Computational Biology, 12(12), p.e1005220.

D. Nath, S. Ray, S.K. Ghosh, **"Fingerprint Recognition System: Design & Implementation,"** Proceedings of International Conference on Scientific Paradigm Shift In Information Technology & Management, SPSITM'11, January, 2011.

Awards & Honors

- Recipient of company-wide **Google Tech Impact Award** in 2023 and **Google Commerce Tech Impact Award** in 2022 for technical contributions & leadership.
- Ranked **585th** (**99.57 percentile**) out of about 130,000 students in *Computer Science,* in *Graduate Aptitude Test in Engineering, 2011* (Indian Graduate School Admission Exam for IITs/NITs etc).
- Recipient of **National Merit Scholarship** for securing **rank 49** in *State Secondary Examination, 2005* among about 700,000 students.