Ganesh Ajjanagadde

15122 125th Pl NE
Woodinville, WA 98072
(510) 358–5239
gajjanag@alum.mit.edu
US citizen

Education

Doctor of Philosophy (Ph.D.) in Computer Science at Massachusetts Institute of Technology (MIT) (2016-2020)
GPA: 5.0/5.0
Masters of Engineering (M.Eng.) in Electrical Engineering and Computer Science (EECS) at MIT (2015-2016)
GPA: 5.0/5.0
Undergraduate Bachelor of Science (SB) in EECS at MIT (2012-2015)
GPA: 5.0/5.0
Fellowship: Presidential Fellowship, NSF Graduate Fellowship, Claude E. Shannon Research Assistantship Award 2019
Research Advisors: Professor Gregory Wornell, Professor Henry Cohn
Research Areas of Interest: Signal Processing, Applied Mathematics, Coding Theory, Inference/Estimation/ML, Information Theory

Work And Research Experience

Senior Research Scientist/Software Engineer Dec 2023 -
Meta, Inc.

Senior research scientist in AI SW/HW codesign. Tech lead for numerical algorithms codesign. I lead performance optimization work for Meta's proprietary AI accelerator (MTIA), and also co-design novel ML models/algorithms to make efficient use of HW (e.g., 2nd order optimizers for deep learning models that won AlgoPerf contest recently). I have also made substantial contributions to future HW designs, saving x% of chip area, regarding efficient numerical formats (e.g., cascading GEMM recipes to emulate high precision that were independently discovered in academia as well: https://arxiv.org/abs/2303.04353), and through industry influence via vendor engagement regarding numerics. Lately, I have conceptualized and led implementation of proprietary math library for MTIA for high accuracy and great performance, using our MTIA HW features effectively. I am involved in both hands on work (training experiments, performance modeling scripts, ML kernel authoring, performant code prototypes, architecture analyses for HW, etc) as well as cross functional leadership and influence with involvement in road-mapping, goal setting, design documents, etc.

Recognized as an internal expert on low-precision numerics, quantization, and algorithms for high performance.

Software Performance Architect July 2022 - Dec 2023
Apple, Inc.

AI/ML power and performance engineer. Broadly responsible for making Siri workloads run fast while consuming minimal power. I brought a multi-faceted approach to the problem: all the way from the application layer by making ML model inference fast algorithmically to the system layer via good use of the underlying silicon and low level API's. Led critical power/perf efforts for Siri on Vision Pro, Watch, and HomePod. Recognized as AI/ML org expert on performance topics, and serve as

performance expert in Siri architecture working group.

Directly responsible for $>$ 100% power improvements for Siri on Vision Pro enabling Siri to run on device, $>$ 500 ms of Siri response time improvement on Watch via improvements ranging from kernel tunables to Siri natural language algorithmic improvements. On Watch this enabled Siri natural language on highly constrained (1.5 GB RAM) device with no latency regression.

Subsequently, worked on efficient inference of LLM on device, suitable for high powered Apple devices. Also worked on Siri on device for an even more constrained platform than Watch above (1 GB RAM HomePod Mini). Among other things, directly contributed to technical roadmap decision of on device Apple Intelligence (8 GB RAM iPhone, i.e iPhone 15 Pro+ only) via detailed technical analysis of the xnu kernel virtual memory system.

Security/Privacy Software Engineer May 2020 - Jun 2022
Snap, Inc: remote

Tech lead for application privacy at Snap, leading the efforts of 5 engineers. Deep expertise in both applied research and software engineering, with a passion and eye for high performance computation and optimization. Successful track record in combining these skills, learning new domains, and working cross-functionally to deliver great product experiences.

- Developed a novel, high performance, perceptual image hashing algorithm (similar to Apple NeuralHash) suitable for both clients and servers (e.g., hashes an image in less than 5 ms on common iPhones). Research underway to extend the algorithm to videos.
- Expanded the scope and delivered end to end encryption for Snap's AR glasses launched at Snap Partner Summit 2020. Worked on all aspects here: client (iOS, Android, C++) as well as server (Java backend).
- Co-owner of end to end encryption for messaging at Snap. Our system encrypts/decrypts $\approx$ billion 1:1 snaps per day.
- Working on privacy preserving ad measurement technologies, especially relevant in view of iOS 14 changes. Implementation in modern C++ (C++17), with bindings for other languages (Java, Go). In use today by some of Snap's partners.
- Performance geek - delivered a novel app binary size reduction, performance optimizations for our privacy preserving ad measurement technologies, deep optimizations for perceptual hashing etc. Extensive expertise with native (C/C++) code, both here as well as with my open source FFmpeg hacking in the past.
- Promoted at Snap for my work above.

Graduate Student Jun 2015 - May 2020
Research Laboratory of Electronics, MIT, Cambridge, MA

Worked with Professor Henry Cohn on point configuration questions, primarily in Hamming space. These include packing and coding questions. Also worked under Professor Gregory Wornell on some computational imaging problems related to coded apertures, and with Professor Yury Polyanskiy on some classical questions of information theory, such as MAC (multiple access channel). Worked on statistical learning with a focus on classification problems for my Master's.

Research Intern Jun-Aug 2019
Microsoft Research New England, Cambridge, MA

Worked with Prof. Henry Cohn on new lower bounds for the mean squared error of vector quantizers in high dimensions. We obtained the first rigorous improvements since the classical work of Zador, 1968.

Research Intern Jun-Aug 2017
Analog Garage, Analog Devices

Worked on coming up with novel methods for anomaly detection in limited resource environments. Explored both classical learning as well as deep learning approaches.

Draper Laboratory Undergraduate Research and Innovation Scholar Sep 2014-May 2015
Laboratory for Information And Decision Systems, MIT, Cambridge, MA

Worked under Professor Yury Polyanskiy on problems of information theory.

Software Engineer Intern Jun-Aug 2014
Kumu Networks, Santa Clara, CA

Wrote a hardware abstraction layer (HAL) for an RF circuit board.

Links and Skills

- GitHub: `https://github.com/gajjanag`
- Website: `https://gajjanag.github.io`
- Publications: `https://gajjanag.github.io/pubs.html`
- Languages: C++20, C99, Python (+numpy/scipy/matplotlib), CUDA, Julia, Java/Kotlin, Objective C, shell (bash/zsh), x86 asm, LaTeX, MATLAB, Verilog.
- Open source contributions: FFmpeg, Julia.
- Platforms/Tools: FFmpeg hacker, Linux, Android dev, git, Bazel, make, vim, GCP, PyTorch, Tensorflow/Keras, GLPK.