Ram Raghunathan

ram@llama.is • https://www.linkedin.com/in/ramraghunathan

Experienced machine learning infrastructure tech lead focused on accelerating iteration and innovation through high performance, scalable, and cost efficient infrastructure paired with user-friendly tooling

Core Competencies

- Implementing distributed infrastructure while balancing performance and cost
- Ensuring operational reliability through system visibility and proactive maintenance
- Managing peers through open communication and active knowledge transfer
- Balancing trade-offs between short-term team needs and long-term business needs

Technical Knowledge

- Python, C++, C, Typescript
- Redshift, Memcached
- AWS, CDK, CI/CD
- Kubernetes, gRPC

Recent Work Experience

Twitch
San Francisco, CA

Software Engineer
May 2021 – Present

High velocity data exploration through user-focused tools and infrastructure

- Drove five large zero-to-one engineering efforts working backwards from a desired objective to requirements, scoped project plan, and execution of the same
- Reduced detection time of production system faults from over 50 days to 2 days by developing a low-friction and accessible monitoring and alerting framework for production pipelines
- Accelerated scientists' ability to discover value from Twitch data and quickly answer customer questions by designing and implementing an easy-to-use data science platform closely integrated with Twitch data sources
- Educated and empowered non-technical team members to adhere to software development and security best practices via self-service processes by planning, communicating, and providing technical aid for migrating existing systems to company-standard code hosting, continuous deployment pipelines, and system design patterns

Quora
Mountain View, CA

Software Engineer
Mar. 2018 – May 2021

Fast and reliable distributed machine learning infrastructure

- Owned Quora Feed backend and increased performance and ongoing reliability by driving regular cross-team review and projects such as more actionable monitoring and a complex migration of content filtering system to handle increased scale demands
- Demonstrated strong management skills by aiding team quarterly planning, prioritizing and scoping of multi-team projects, providing guidance to ML teams on system level planning, design, and impact measurement, and mentoring junior colleagues to excel as effective engineers and make significant contributions
- Achieved double-digit percentage decrease in Tensorflow inference time by driving several multi-team projects including gRPC, Tensorflow, and Kubernetes tuning, improving feature caching schema, and reducing request size by implementing 32-bit float support in Thrift
- Improved understanding and debugging of ranking pipeline by designing and developing a framework for ML engineers to create standard visualizations of their pipelines
- Regularly presented to colleagues on complex system architecture and future work

Prior Work Experience

Carnegie Mellon Univ.
Pittsburgh, PA

Research Assistant
Aug. 2013 – Sep. 2017

Low overhead automatic memory management for parallel programs

- Developed and implemented a theory of hierarchical memory management for nested parallel programs and proved key properties and characteristics about the theory
- Achieved high scalability and performance with an intuitive parallel programming paradigm

Tower Research Capital
New York, NY

Infrastructure Developer
Jul. 2011 – Jul. 2013

High-performance core infrastructure for market data access

- Increased trader velocity by developing a more composable API for market data access.
- Decreased hardware costs through multi-tenancy of strategies by designing and implementing a low latency single producer, multiple consumer inter-process communication system.

Education

Carnegie Mellon University
M.S. in Computer Science September 2017

- Advised by Umut Acar
- Researched theory and practice of automatic memory management for nested parallel programs

Carnegie Mellon University
B.S. in Computer Science May 2011

- Graduated with University and College Honors
- Senior thesis: "Design and Implementation of a Power-Aware Load Balancer" advised by Mor Harchol-Balter
- Winner of School of Computer Science Alumni Award for Undergraduate Excellence for Best Senior Thesis

Publications

Adrien Guatto, Sam Westrick, **Ram Raghunathan**, Umut A. Acar, Matthew Fluet. Hierarchical memory management for mutable state. *In Proceedings of the 23rd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP '18). ACM, New York, NY, USA, 81-93*

Ram Raghunathan, Stefan K. Muller, Umut A. Acar, and Guy Blelloch. Hierarchical memory management for parallel programs. *In Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming (ICFP 2016). ACM, New York, NY, USA, 392-406*

Umut A. Acar, Guy Blelloch, Matthew Fluet, Stefan K. Muller, and **Ram Raghunathan**. Coupling memory and computation for locality management. *1st Summit on Advances in Programming Languages (SNAPL 2015), volume 32 of Leibniz International Proceedings in Informatics (LIPIcs)*

Anshul Gandhi, Mor Harchol-Balter, **Ram Raghunathan**, Michael Kozuch. AutoScale: dynamic, robust capacity management for multi-tier data centers. *ACM Trans. Comput. Syst. 30, 4, Article 14 (November 2012)*

Anshul Gandhi, Mor Harchol-Balter, **Ram Raghunathan**, Michael Kozuch. Distributed, robust auto-scaling policies for power management in compute intensive server farms. *Open Cirrus Summit 2011*