Bhuvan Prakash Karuturi

Bellevue, WA | bhuvankaruturi123@gmail.com | +1 469-230-0476 LinkedIn: bhuvan-karuturi | GitHub: bhuvankaruturi

SUMMARY

Results-oriented Senior Software Engineer (L5) and Tech Lead with 5+ years of experience specializing in ML accelerator infrastructure, distributed systems, and hardware/software co-design. Proven track record of leading complex, multi-organizational initiatives from concept to launch for large-scale AI compute environments. Highly skilled at bridging low-level system performance (virtualization, networking, kernel) with high-level ML framework execution (JAX, PyTorch) to optimize massive-scale training and inference workloads.

SKILLS

- **Languages:** C++, Python, Go, Java, SQL
- **ML & Infrastructure:** JAX, PyTorch, Cloud TPUs, Kubernetes (GKE), Google Compute Engine (GCE), Distributed Systems
- **Systems & Tools:** TPU Compiler & Runtime (libtpu), Virtualization, NUMA Architecture, Performance Profiling/Benchmarking, CI/CD, Git

EXPERIENCE

Google LLC | Bellevue, WA *Senior Software Engineer (L5) & Tech Lead, Cloud TPU ML Infrastructure* | Jan 2022 – Present

- **Next-Generation Cloud TPU NPIs:** Led end-to-end system design and execution for multiple next-generation Cloud TPU New Product Introductions (NPIs). Delivered the compute engine stack turnup 3 months ahead of hardware availability, navigating capacity delays to successfully pull in the development schedule by 4 weeks.
- **Hardware-less Simulation Platform:** Co-led a year-long cross-organizational initiative (10 teams across 5 departments) to architect and deliver a distributed Cloud TPU Pod Simulator. Accelerated cloud hardware development lifecycles by 6+ months by establishing a foundational platform for testing compiler, runtime, and cluster scheduling systems prior to silicon availability.
- **ML Compiler & Runtime (libtpu) Optimization:** Directed SDK integration for tier-1 AI/LLM research customers. Designed and delivered a runtime SDK within 2 months that enabled accelerator sub-slicing, drastically improving hardware resource utilization for large-scale inference workloads.

- **System-Level Performance Engineering:** Increased ML model training performance by 30% on Cloud TPU instances by resolving virtualization (vIOMMU) degradation, implementing static hugepages, and optimizing multi-PCIe root configurations to eliminate non-local memory access.
- **Cross-Functional Technical Leadership:** Drove alignment across networking, kernel, and ML framework teams to resolve systemic performance bottlenecks and influence host NUMA configurations. Established new disaggregated serving benchmarks to set network performance baselines for future ML inference workloads.
- **Ecosystem Integration & Open Source:** Streamlined the ML framework execution pipeline by standardizing build configurations, transitioning Python package hosting to PyPI, and engineering a release pipeline that ensures seamless compatibility with the JAX stable stack and custom Python interpreters.
- **Operational Excellence:** Serve as a primary incident commander for high-priority compute anomalies, proactively unblocking high-value enterprise clients. Consistently recognized for mentorship, onboarding, and authoring comprehensive system design and debugging documentation.

Amazon.com Services LLC | Seattle, WA *Software Development Engineer Intern (ML Focus)* | May 2021 – Aug 2021

- Developed a machine learning-based classification system using features derived from 4 distinct internal systems to identify fraudulent transactions, demonstrating the ability to reduce manual review labor costs by 73%.

Ohio-At-Home Healthcare Agency | Columbus, OH *Full Stack Web Developer Intern* | Aug 2020 – May 2021

- Built the company's mobile-first provider portal frontend (Angular, Material Design) and 6 microservices for business-logic handling, utilizing asynchronous message queues.

Capgemini India | Hyderabad, India *Software Engineer* | May 2017 – May 2019

- Customized and optimized a large-scale enterprise ERP system using Java, SQL, and PL/SQL, earning the Project STAR Award for prompt and efficient technical delivery.

EDUCATION

The University of Texas at Dallas | Richardson, TX *Master of Science, Computer Science* | GPA: 3.8/4.0 | Aug 2019 – Dec 2021

Acharya Nagarjuna University | Guntur, India *Bachelor of Technology, Electrical and Electronics Engineering* | GPA: 3.7/4.0 | Sep 2013 – Apr 2017