Manoj Raja Rao

Contact

manoj@manojrao.com
805-915-9501

Education

UCLA, Los Angeles, USA

Master of Science, Dept. of Computer Science **2008 – 2010**

- Advisor: Dr. Yuval Tamir
- Research at UCLA's Concurrent Systems Laboratory

RVCE, Bangalore, India

Bachelor of Engg., Dept. of Computer Science&Engg. **2001 – 2005**

Professional Experience

Tesla AI, Palo Alto, USA

Engineering Manager, AI Inference **2022 - Present**

Managing over 23 team members working on AI Inference for Tesla FSD and Robot.
Compiler and Runtime Development for Inference of SOTA FSD Models.
Strong engineering driven, involved hands-on in the design and early implementation phases of the compiler and runtime.
Team is responsible for everything from QAT, Graph Export, ML Compiler, Runtime and deployment for efficient inference on the car and robot. Highlights Of Team's Achievements:

- Designed and Implemented ML Compiler and Runtime from legacy to MLIR framework
- Critical Tooling Infrastructure: Layer-By-Layer Debugging
- Quantization Aware Training
- Export and Deployment Of Tesla's End-To-End Neural Net Based Planner Model

AWS AI, E Palo Alto, USA

Lead (TLM), AWS ML & Deep Learning Inference **2019 - 2022**

Tech Lead / Maintainer *TorchServe - PyTorch's Inference Server*
Link - *https://pytorch.org/docs/master/community/persons_of_interest.html#torchserve* of **TorchServe**, the official Deep Learning Model Server for Facebook's popular Deep Learning Framework - PyTorch Link: my whitepaper on AWS Blog

AWS SageMaker Edge Manager: 2020-2021 Tech Lead of AWS SageMaker Edge, launched in re:Invent 2020. Designed and Implemented the Deep Learning Inference Runtime for Apache-TVM models and model management for Edge Devices via AWS SageMaker. Manage, Deploy, and Serve Deep Learning Models via multiple interfaces efficiently implemented in Modern C++
Link: More info here

AWS AI Personalize Services: 2019 Contributed to AWS ML services to enable Predictive Maintenance solution on large scale industrial data. Perform feature extraction on training data. Hands-on with auto-scaling clusters with AWS EMR for massive data processing. Involved in the
Link:AWS Personalize core personalization algorithm and inference at AWS scale.

Amazon Robotics, Sunnyvale, USA

Senior Software Engineer, Platform Software for Robot **2017 - 2019**

Full stack software development for Robotics project. *Amazon's Robot:* Involved in building Deep Learning platform for the robotics product at Amazon. On-device Power and Performance Management for CV-based Deep Learning workloads. Also involved in iPrivacy, Protoyping PID motor controllers, sensor fusion, board bringups.

Amazon SmartHome, Sunnyvale, USA

Senior Software Engineer, Amazon SmartHome Products **2014 - 2017**

Low-Level Software Development includign Linux Kernel Development for Amazon's Alexa / SmartHome Products based on Android/FireOS.

Amazon's Alexa Devices: Involved in product lifecycle from research, prototyping and productizing next generation devices like Alexa, FireTV, Fire Tablets, and more. BSP, board bring-ups, Power and Thermal management, OS-level through UI Performance Engg.

Qualcomm Innovation Center Inc., San Diego, USA

Senior Software Engineer, Linux Kernel Development for Snapdragon Chipsets **2010 - 2014**

Snapdragon Linux Kernel: Involved in the development of MSM chipsets for Snapdragon's 64-bit CPU architecture. Involved in silicon bring-ups, Linux Kernel Security, Linux Kernel Code Review and Device Tree Code Review.

UCLA, Los Angeles, USA

Graduate Student Researcher **2008 - 2010**

Aylus Networks Pvt Ltd., Bangalore, India

Software Engineer **2006 - 2008**

Developed software modules for the 3G-Telecom Application Server for media share services for 3G users. Developed Service Provisioning System for provisioning users to 3G networks. Developed HA functionality of critical modules.

Huawei Technologies India Pvt Ltd., Bangalore, India

Software Engineer **2005 - 2006**

Worked as Junior Researcher in the R&D for an internal Linux Cluster Middleware for providing Carrier-Grade HA. I was part of the team which applied for patents in Group communication among cluster nodes.

External Links

- TorchServe
- My commits to 3.10 msm Linux Kernel
- My commits to <= 3.4 msm Linux Kernel
- My Technical Blog
- My fledgling podcast series

Technical Skills

- *Languages*: C, C++, Python, Java, Emacs Lisp, Ruby, Perl, Erlang
- *ML/DL FW*: PyTorch, MXNet, scikit-learn
- *OS Dev*: Linux Kernel Development
- *Tracing*: eBPF, ftrace, systrace, perf
- *Research*: Distributed Systems, Fault Tolerance, Reliability, Message Passing in Clusters
- *Concepts*: Deep Learning, Distributed Systems, Cloud Computing, Computer Architecure, OS
- *Platforms*: Linux, Android, Linux on ARM, EFI, Intel's BIOS, ACPI
- *Dev Tools*: Emacs, Vim, Git, Trace32, GDB, kdb
- *Protocols*: HDMI-CEC, HDCP, I^{2}C, MHL-CBUS, USB detection, SPI

RELEVANT GRADUATE COURSES

- Distributed Algorithms, Cloud Computing, Operating Systems, Advanced Scalable Systems, Advanced Parallel Systems, Online Algorithms, Advanced System Design, Advanced Computer Architecture, Wireless and Mobile Computing, Cyber Physical Systems

Miscellaneous Recognition

- Patent idea at Huawei Technologies., All India Ranking of 382 among 100000 participants in Entrance Tests., Huawei Certified .C. Programming Specialist.

Membership

- MENSA, FOSS, Computer Society of India