

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. Strong engineering driven culture, involved hands-on in the design and implementation of the compiler and runtime. Team's responsibilities: Quantization Aware Training for int8 (post-PyTorch), Graph Export (Fx), ML Compiler (MLIR-based), Runtime (MLIR-based) for efficient inference on the car and robot alike.

Technical Highlights of the Team:

- Implemented new ML compiler and runtime: rewrite in MLIR framework
- Developed critical tooling infrastructure for layer-by-layer debugging
- Implemented int8 QAT for model optimization
- Inference of end-to-end neural network based planner model
- Co-designed AI/ML models with engineers and informed hardware design decisions
- For model scaling:
 1. Tensor Parallelism for Transformers
 2. Pipelined Instruction scheduling for Multi-Attention Head
 3. SAT based solver for node level schedule optimizer for GPT/Diffusion based Models
 4. Heterogenous model partitioning across CPU, GPU (OpenCL), and Custom ASIC



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 **Torch-Serve**, 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 Man-

agement for CV-based Deep Learning workloads. Also involved in iPrivacy, Prototyping PID motor controllers, sensor fusion, board bringups.



Amazon SmartHome, Sunnyvale, USA

Senior Software Engineer, Amazon SmartHome Products

2014 - 2017

Low-Level Software Development including 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 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

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 team for Carrier-Grade HA.

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 Architecture, OS
- *Platforms:* Linux, Android, Linux on ARM, EFI, Intel's BIOS, ACPI
- *Dev Tools:* Emacs, Vim, Git, Trace32, GDB, kdb
- *Protocols:* HDMI-CEC, HDCP, I²C, 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