Sandeep Kumar **Behera**

4088239098 | sandeep.skb@apple.com

"Get comfortable with being uncomfortable."

Summary

Senior Software Engineer with expertise in High Performance Computing, Deep Learning, GPU Architecture and ASIC Design. Amalgamating this expertise to HW-SW co-design, analyze and optimize Inference and Training performance of deep learning networks and ML applications on GPUs and accelerators.

Work Experience

Apple SPG

Cupertino, CA

Sr. Software Engineer — *Aug. 2021 - Present*

- Key player in GPU cloud deployment effort. Used TritonServer, TensorRT, custom ops and other optimizations to improve performance and bring down the cost from $20M to $80K.
- Implemented and guided Quantization Aware Training for Deep Learning models.
- Plan roadmap and scope out projects.
- Mentoring team members.

Aurora Innovations

Mountain View, Ca

Sr. Software Engineer - Tech Lead Manager — *Jan. 2021 - Aug. 2021*

- Improved the Inference latency of Aurora perception model by 1.7x by converting the model to TensorRT, writing custom cuda kernels and other optimizations.
- Evaluated Multi-Instance GPU in A100 for Aurora perception models.

Uber ATG

San Francisco, CA

Sr. Software Engineer — *April 2020 - Jan. 2021*

- Improved the training and inference performance of perception models by an average of 1.5x and 2x respectively by analyzing and optimizing multiple bottlenecks.
- Integrated TensorRT with models using Torch2TRT for Inference.
- Implemented "GroupNorm" operation as a plugin in Torch2TRT to provide a fallback path from TensorRT to PyTorch.
- Designed and Implemented Multi-streaming to launch work from multiple sensors on the autonomous vehicle on different CUDA Streams to parallelize computation and reduce inference and training time.

Nvidia Corp.

Santa Clara, CA

Sr Deep Learning Architect — *Sept. 2017 - April 2020*

- Developed tool to automate the Deep Neural Network performance analysis process.
- Worked on analyzing the difference in Executed performance analysis for Automatic Mixed Precision in TensorFlow which was released at GPU Technical Conference 2019.
- Lead post-silicon performance architect for TitanRTX and T4 data center chips. Executed end-to-end training studies and improved training performance by 24%.
- Implemented multiple features in Python based simulator to explore functionality and performance for the next gen-architecture.
- Coordinated with Framework, cuDNN/cuBLAS, Driver and Unit architecture teams to identify and resolve performance issues.

Nvidia Corp.

Santa Clara, CA

Sr. ASIC Design Engineer — *Feb. 2013 - Sept. 2017*

- Lead designer of the Compute work distributor of GPU compute pipeline.
- Architected, designed and implemented simultaneous compute and graphics in multiple generations of NVIDIA GPU (Maxwell, Pascal, Volta & Turing).
- Designed and implemented compute instruction level preemption in PASCAL architecture of NVIDIA GPU.
- Designed and implemented subcontext dynamic partitioning feature in Volta architecture.
- Worked on micro-architecture, functional and formal verification, synthesis and timing of scheduler and compute work distributor unit in NVIDIA GPUs.

Education

North Carolina State University *USA*

M.S. in Computer Architecture *Aug. 2011 - Dec. 2012*

SASTRA University *India*

B.Tech. in Electrical and Communication Engineering *May. 2007 - May. 2011*

Skills

Programming	C/C++, Python, JAVA, Shell, MySQL, LaTeX
DevOps	AWS, Docker, Kubernetes, Jenkins, CircleCI
Frameworks	PyTorch, TensorFlow, MxNet, TensorRT, Triton
Tools	Nsight, NVprof, Verdi