Sudnya Diamos

San Jose, CA. mailsudnya@gmail.com

OBJECTIVE

To develop new technology that advances the state of the art in deep learning and high performance computing.

EXPERIENCE

Attune: Mountain View, CA

Data Science and Algorithms Engineer – March 2014 - December 2016

- A core developer of Attune's personalization as a service platform. This system procures data from clients, applies personalization algorithms, and bills clients using A/B test results.
- Co-designed the data ingestion framework supporting multiple customers.
- Developed personalization algorithms using collaborative filtering and matrix factorization.
- Developed a continuous A/B testing/billing tool to measure Attune's impact.
- When required, I stepped in on a platforms role and developed a Java client to make REST API calls to Attune's service, currently deployed by Attune's clients.

NVIDIA: Santa Clara, CA

Senior Architect - GPU Streaming Multiprocessor – February 2012 - February 2014

- Worked on the architecture design team that created the first general purpose programmable graphics accelerator core built by NVIDIA for mobile devices (Tegra).
- Debugged complex workloads (e.g. Cuda Nested Parallelism uScheduler, computer vision, OpenGL shaders, etc) on a full chip simulator involving multiple interacting units.
- Evaluated new features and ISA changes using simulator modeling and designing directed performance tests. This system was used to evaluate the RTL design and performance bottlenecks.
- Implemented architecture and simulator support for precise exceptions.

NVIDIA: Santa Clara, CA

GPU Architecture Engineer – March 2010 - February 2012

- Ported an architecture simulation framework to enable evaluation of the next generation GPUs.
- Wrote bringup tests for a new feature in the next generation GPU.
- Developed a tool to translate application traces into unit tests that run on RTL.
- Worked on an instrumentation tool that captures GPU processor state for performance studies.

RELEVANT PROJECTS

Audio Captioning System

- Converts from raw audio into an english language description using end-to-end deep learning.
- Network architecture includes recurrent and convolutional layers, trained from raw audio.
- Extended Connectionist Temporal Classification cost function to handle multiple labels.
- Created a new dataset by crawling LibreSpeech, freesound.org, and freemusicarchive.org.

Spatial Transformer Networks

- From scratch reimplementation of Spatial Transformer Networks in Tensorflow.
- Performed performance analysis using roofline model. Training time within 2x of normal CNN with the same number of parameters.
- Worked with Deep Learning researchers to apply these models to generative image modeling.

EDUCATION

Georgia Institute of Technology, Atlanta, GA
- MS - Computer Science – August 2007 - December 2008

Vishwakarma Institute of Technology, Pune University, India
- BE - Computer Engineering – August 2002 - July 2006

SELECTED PUBLICATIONS

Sudnya Padalikar and Gregory Diamos. "GPU-RPC: Exploiting The Latency Tolerance of CUDA Applications." In *NVIDIA Research Summit*, San Jose, California, USA, September 2009.

SELECTED PRIOR PROJECTS

A Massively Parallel Simulator
- Built a massively parallel simulator (CUDA based) to simulate future parallel processor architectures on current GPUs.
- Explicit separation between functional and timing model.
- Achieves high performance by exploiting data structure locality, hierarchical synchronization, and minimal state-per-thread.

Parallel Discrete Event Simulator
- The simulator is partitioned into models which describe the system being simulated and a kernel that manages events and time synchronization.
- Detailed timing models for network links, ethernet devices, and layer 3 and 4 protocols.
- Sequential and parallel implementations of the simulator kernel, each with identical interfaces. The parallel version implements the (Chandy-Misra-Bryant) time synchronization algorithm using MPI.

SKILLS

Languages:
- C++, Java, Python, C, CUDA, NoSQL (Cassandra), Octave, Intel x86 assembly, NVIDIA GPU Assembly, Scripting (Zsh, Csh, Bash).

Libraries:
- STL, Boost, OpenCV, MPI, Pthreads, OpenMP, BLAS, LAPACK, NumPy, TensorFlow.

REFERENCES

- Will be provided on request.