James K. Reed

✉ 6jamesr6@gmail.com
jameskreed.com
github.com/jamesr66a/
linkedin.com/in/jamesr66a/

Work Experience

Oct 2022 - Present **Founding Engineer**, *fireworks.ai*, Redwood City, CA.

Working on a best-in-class Generative AI platform for customized Large Language Models (LLMs). Key responsibilities include:

- Deep performance optimization of modern Deep Learning workloads, particularly for open LLMs on NVIDIA GPUs.
- Model scalability optimization, including distributed execution of workloads.
- End-to-end cloud service development for scalable serving of deep learning workloads.

Jan 2021 - Aug 2022 **Staff Software Engineer (E6)**, *Meta (formerly Facebook)*, Menlo Park, CA.

PyTorch team. Experienced engineering lead with a focus on programming language and compiler design and implementation, numeric computing, high-performance software (CPU/GPU), HW/SW co-design, and distributed systems.

- Technical lead of torch.fx program transformation toolkit and PiPPy automatic parallelization toolkit for PyTorch.
- Responsible for design and bring-up of systems and tools for deploying and optimizing deep learning models across devices and at scale (ONNX, TorchScript, Intel CPU quantization, torch.fx, and upcoming Pipeline Parallelism toolkit).
- Adept at moving up and down the computing stack from silicon to Python.

Aug 2018 - Jan 2021 **Senior Software Engineer (E5)**, *Facebook*, Menlo Park, CA.

Feb 2017 - Aug 2018 **Software Engineer (E3-E4)**, *Facebook*, Menlo Park, CA.

May 2016 - Aug 2016 **Software Engineer Intern**, *Facebook*, Menlo Park, CA.

May 2015 - Aug 2015 **Software Engineering Intern**, *Google Machine Intelligence*, Mountain View, CA.

Sept 2014 - Dec 2014 **Software Engineering Intern**, *Google Cloud Platform*, Seattle, WA.

May 2014 - Aug 2014 **Interim Engineering Intern**, *Qualcomm Technologies, Inc.*, San Diego, CA.

May 2013 - Aug 2013 **Interim Engineering Intern**, *Qualcomm Technologies, Inc.*, San Diego, CA.

Publications and Patents

December 2021 **James K. Reed**, Zachary DeVito, Horace He, Ansley Ussery, and Jason Ansel. 2021. *torch.fx: Practical Program Capture and Transformation for Deep Learning in Python.* Accepted to MLSys 2022. arXiv:2112.08429 [cs.LG]. `https://arxiv.org/abs/2112.08429`

Feb 2020 Nadav Rotem, Abdulkadir Utku Diril, Mikhail Smelyanskiy, Jong Soo Park, and **James Kenneth Reed**. 2020. *Systems and methods for employing predication in computational models*. US Patent #10553207

Education

Education

- Bachelor of Science in Computer Engineering
  - Minor in Computer Science
- Virginia Tech, Blacksburg, VA
- Graduated December 2016, Summa Cum Laude

Public Speaking

September 2022 **MLSys 2022**, *torch.fx: Practical Program Capture and Transformation for Deep Learning in Python*, `https://mlsys.org/Conferences/2022/Schedule?showEvent=2141`.

March 2022 **NVIDIA GTC 2022**, *Optimizing & Deploying PyTorch Models for High-Performance Inference*, `https://bit.ly/3keHK5I`.

December 2021 **PyTorch Developer Day**, *Easy Python Code Transformations with torch.fx*, `https://www.youtube.com/watch?v=fbtVDqp3lv8`.

April 2021 **PyTorch Ecosystem Day**, *torch.fx: A New Python-to-Python Code Translation Framework for PyTorch Code (talk and poster)*, `https://assets.pytorch.org/pted2021/posters/C5.png`.

June 2020 **PyTorch YouTube**, *TorchScript and PyTorch JIT | Deep Dive*, `https://www.youtube.com/watch?v=2awmrMRf0dA`.

October 2020 **NVIDIA GTC 2020**, *Named Tensors, Model Quantization, and the Latest PyTorch Features*, `https://developer.nvidia.com/gtc/2020/video/s22145-vid`.

December 2019 **NeurIPS 2019 Expo**, *Production Scale PyTorch*.

May 2019 **PyCon 2019 Workshop**, *Production-scale PyTorch: TorchScript and the PyTorch JIT*, `https://us.pycon.org/2019/schedule/presentation/381/`.

Skills

Programming Python, PyTorch, C++, High-performance x86 AVX, CUDA C++, LaTeX

Technical Design High-Performance Numeric Computing, Programming Languages and Compilers, Computer Architecture and Design, Performance Analysis and Engineering, Operating Systems and Embedded, Deep Learning, Distributed Systems

People Requirements Gathering/API/UX design, Technical Leadership, Technical Interviewing and Recruiting, Public Speaking