Dhruv Matani

E-mail: dhruvbird@gmail.com | LinkedIn: dhruvbird | Medium: @dhruvbird | Meta Engineering: dhruv-matani | PyTorch

Profile Summary:

A SWE Technical Lead and Software Architect with 16+ years of hands-on experience building teams and projects from 0 to 1 as well as successfully scaling teams and projects. I have a unique blend of deep learning and software engineering skills with a track record of bring state of the art AI research to production.

Areas of Expertise:

- On-Device Artificial Intelligence (AI)
- Model Optimization and Efficient Model Architectures
- Machine Learning (ML) infrastructure
- Performance, reliability, efficiency, scalability
- API design and high throughput distributed systems
- Bootstrapping teams
- Cross functional team leadership
- Open Source (PyTorch, gcc)
- Hands on leadership style
- Attention to detail

Work Experience:

Meta Platforms (Facebook Inc) **March 2013 - Present**

PyTorch (AI Frameworks and AI Infrastructure) - Technical Lead

- Leading key initiatives to bring PyTorch to the On-Device ecosystem internally and externally as a unified framework. Developing PyTorch for powering state of the art *research to production* workstreams
- Influenced technical direction and investments
- Enabling Vision, Speech, Text, and Personalization based ML experiences on Edge Devices
- Spearheaded cross-functional (XFN) efforts for the Meta family of apps, enabling Mobile Vision, On-Device Personalization, Spark AR, Federated Learning (on-device training) and Analytics
- Unlocked value for Meta via the following critical initiatives
  - Achieved up to 3x reduction in model latency by developing optimized CNN and Transformer Attention based models for on-device image segmentation, object and key-point detection, inpainting, denoising, super-resolution, bandwidth prediction, advertiser integrity, personalization, and voice transcription
  - Optimized models for on-device execution using quantization, distillation, efficient model architectures, and Efficiency best practices
  - 90% reduction in library size and 60% reduction in model loading latency and runtime startup cost for PyTorch On-Device (patent pending)
  - Led the development of benchmarking infrastructure for heterogeneous devices and built a self-serve platform for ML Model Lifecycle Management, scaling developer productivity 5x through auto-generated tests and continuous integration (CI)
  - Mentored 20+ engineers to adopt a reliability-first culture, resulting in enhanced productivity and established best practices for logging and data collection for internal stakeholders and leadership
  - Drove the creation of productivity-enhancing tooling such as Model Inspector and Model Inventory
  - Established critical outage criteria for the entire team across all functions, ensuring a reliable product for customers
  - Aligned engineering and business goals, resulting in successful product outcomes
- PyTorch Evangelism
  - Published 12 articles on On-Device Machine Learning and PyTorch on KDNuggets, Towards Data Science and PyTorch Blog with 20k+ readers
  - Authoring an expert written chapter on PyTorch for the “Efficient Deep Learning” book
  - Served as a judge for the Global PyTorch Hackathon 2021 (1900+ participants worldwide) and 2021 Meta XR Hackathon ($700k prize money)

Data Platform - Technical Lead Engineering Manager

- Managed the Logging Team at Meta and successfully utilized analytics and machine learning data logging to drive crucial company-wide metrics such as time spent, DAU, and advertiser revenue on both backend and app platforms.
- Achieved phenomenal efficiencies (40% space reduction, and 80% compute reduction, saving $300million/year by bending the cost curve) by re-architecting the on-wire serialization format for loading data into the Facebook Data Warehouse. This affects both batch and real-time workloads (patent pending)
- Optimized 8% fleet-wide compute costs by re-architecting the PHP Logger framework to leverage type-safety and runtime efficiency offered by Hack and HHVM
- Drove a 5% reduction in fleet-wide compute though an efficiency program targeted at data sampling
- Super-charged developer productivity 50x by schematizing logging and data consumption by building the statically-typed C++/Python/Hack Logger and Reader libraries for all backend use-cases across Meta
- Led the semantic warehouse project to support semantically-rich and structured types across Ent, Logger, Presto and Spark

- Unified datasets across batch and real-time workloads by launching the real-time data warehouse
- Kick started the Engineer Embedding Program to unite Data Engineering and Software Engineering teams to enhance collaboration
- Drove up favourability in the team from 72% to 94% by investing in career growth, team wellbeing, and upholding a culture of mutual respect

Ads Targeting and Custom Audiences – Software Engineer

- Boosted advertiser ROI by 18% by launching the custom-audiences backend, enabling targeted ad delivery using advertisers' own audience sets

Data Platform and Infrastructure - Scuba and MySQL - Software Engineer

- 98.7% reduction (2.5 hours to 120 seconds) in Scuba restart time via Fast Database Restarts in Scuba (SIGMOD’14)
- Achieved 6x data compression by implementing a performant in-memory columnar storage engine for Scuba
- Accelerated data ingestion by 4x and boosted Scuba's reliability by adding an extra 9's (equivalent to 99.9% uptime or 3.5 additional days/year)
- Successfully migrated Scuba across 4000+ machines to different regions without any downtime or service disruption by utilizing rolling migrations, demonstrating expertise in managing large-scale projects and ensuring seamless transitions.
- Led Scuba’s disaster recovery; replicating data in real time across machines in multiple geographic regions
- Optimized corrupted table restore time 52x by created an innovative per-table Backup and Restore tool for MySQL databases, significantly enhancing data recovery capabilities by enabling faster and more precise restoration of individual tables in case of corruption, surpassing previous methods that only allowed for restoration of entire databases. This granular approach has greatly improved the reliability and efficiency of MySQL database management.

Directi Pvt. Ltd. **Jan ‘09 – Jun ‘11**

Senior Software Engineer

- Launched the Open Source XMPP BOSH Server for internal and external use
- Developed the critical online update logic and the business logic for .pw desktop chat client and chat server respectively
- Led the development of an online community driven translation tool for internationalization of web applications
- Led problem-setting and interviewer training and developed an online platform to streamline recruitment of software engineers at scale for the IIT graduate hiring process.

Mukesh Patel School of Technology Management & Engineering **Jul ’08 – Jan ‘09**

Lecturer

- Instructed two core computer science courses to more than 60 students each, covering Systems Architecture & Programming and Operating Systems-II.

Calsoft Pvt. Ltd. **Jul ’06 – Apr‘08**

Senior Software Engineer

- Optimized the Ingres Database Server for ScaleMP’s vSMP architecure
- Achieved 8x speedup for log parsing, and 3x improvement in operator productivity by implementing automated installation of Linux (via TFTP) for ScaleMP
- Video ad serving: Developed an HTTP and RTSP Caching Proxy (with re-encoding and down-sampling) for YouTube flash and 3gp video content

Skills:

AI Frameworks and Tools	PyTorch, Kaggle
Programming Languages	C, C++, Python, Hack, Javascript, PHP, Bash, Object Pascal, SQL
Presentation technologies	HTML, LaTeX
Platforms	Linux, MS-DOS, Windows
DBMS	MySQL, SQLite, Hive, Presto, SparkSQL, PostgreSQL, Oracle
Tools	Jupyter Notebooks, Google Colab, node.js, Mercurial, git

Education:

Master of Science (M.S.), Computer Science
Stony Brook University, Stony Brook (NY)
Bachelor of Engineering (B.E.), Computer Engineering
University of Mumbai, Mumbai, India

Publications:

- Efficient PyTorch: Tensor Memory Format Matters 2021
  https://pytorch.org/blog/tensor-memory-format-matters/
- Fast Bitmap Fit: A CPU Cache Line friendly memory allocator for single object allocations 2021
  https://arxiv.org/abs/2110.10357
- A Simple Solution to the Level-Ancestor Problem (1 citation) 2019
  https://arxiv.org/abs/1903.01387
- Fast Database Restarts at Facebook (47 citations) 2014
  https://research.facebook.com/publications/fast-database-restarts-at-facebook/
- Avoiding locks and atomic instructions in shared-memory parallel BFS using optimistic parallelization (13 citations) 2012
- Partial deamortization of the Packed Memory Array *http://dhruvbird.com/pdpma.pdf*
- Compressing the human genome against a reference (6 citations) 2011
  http://dhruvbird.com/genome_compression.pdf
- An O(k log n) algorithm for prefix based ranked autocomplete (15 citations) 2011
  http://dhruvbird.com/autocomplete.pdf
- An O(1) algorithm for LFU (Least Frequently Used) cache replacement (48 citations) 2010
  http://dhruvbird.com/lfu.pdf
- A technique for extracting song lyrics from web pages without knowing their structure 2006
  http://dhruvbird.com/liblyric.pdf
- A distributed approach for solving a system of linear equations (2 citations) 2004

Contributions:

- Improved libstdc++-v3 (g++ C++ Standard Template Library) with numerous bug-fixes and optimized critical applications up to 2x by adding a cache-optimized single-object allocator (bitmap_allocator) to libstdc++v3
- Built node-xmpp-bosh, an Open-Source web proxy to serve XMPP over HTTP
- Architected and built the infrastructure powering the analysis in the book "Who is Bigger" – by Steven Skiena and Charles Ward
- Provided key insights into the complexity analysts of the R1Q algorithm mentioned here