**Naveen Swamy**

**Work Summary**

* 18+ years of experience in building highly scalable distributed Software using C, C++, Java, and Scala on UNIX/Linux.
* 7+ years of experience working on ML framework internals, building Deep learning tools, scaling DL model training and deploying DL models as a distributed web service.
* Experienced in leading design, implementing, and cross-team coordination to launch all layers of software.

**Master of Science in Computer Engineering December, 2008**

**University of Texas at Dallas, Dallas, TX GPA 3.7/4.0**

**ML Frameworks: Internals of Tensor flow, Apache MXNet, and usage of Pytorch.**

**Programming Languages: C/C++, Python, SQL, PL/SQL, Perl, Scala, Java/J2EE**

**AWS Technologies: Amazon S3, DynamoDB, SWF, EMR, Cloud Formation, Lambda, EC2, Sage Maker**

**Open Source: Apache Spark, Apache MXNet, Tensorlfow**

**Protocols: TCP/IP, SIP, DNS, SS7, SNMP, HTTP, SSL, GRPC.**

**Multi-threaded, interprocess communication, asynchronous and event driven programming on Linux environments.**

**Work Experience**

**Dgraph Labs., Palo Alto, California Sept, 2022 - Till date**

* Design and Build Golang based micro services to support Dgraph Cloud Infrastructure and Dgraph database.
* Design and Build Infrastructure for model hosting and vector databases.
* Design and build Dgraph Cloud Infrastructure from the ground up.

**Cerebras Systems, Inc., Sunnyvale, California June, 2020 - Sept, 2022**

* **ML Frameworks - Member of Technical Staff (ML Engineer)**Worked end-to-end on Tensor Flow including XLA Compile, Tensor Flow core, and other abstraction layers from C++ to the python binding.
* Designed and Implemented Incremental Compilation for translating ML models to Cerebras Wafer Scale Engine (WSE) hardware reducing compile times by 70% for large models like BERT, GPT.
* Implemented Stream Executor subsystem to integrate Tensor Flow backend with Cerebras WSE service which included:
* Lowering of Compile IR to proprietary IR and Graph rewrites.
* Designed and Implemented a GRPC service-client software between Tensor Flow and WSE service.
* Design and implementation of Dense-Sparse Tensor Mapping and Tensor Shading.
* Identified key bottlenecks and refactored compile, runtime components and WSE service achieving 3x reduction in memory footprint and 2x speed-up of compile and weight initializations on very large NLP models.
* Design and Implemented ML framework support for multi-host, multi-box distributed training

**Amazon Web Services, Palo Alto, California May, 2013 - June, 2020**

**Amazon Personalize, AWS - Senior Machine Learning Engineer August, 2019 - June, 2020**

* Scaling Machine learning Training and Infrastructure for the Amazon Personalize Web Service.

**AWS ML Frameworks - Senior Software Engineer September, 2016 - August, 2019**

* Designed and developed Scala Deep Learning APIs for Apache MXNet Deep Learning framework.
* Build Framework and tools for Deep Learning. (Mxnet-spark-inference, deep-cluster)
* Built an automated memory management mechanism for Apache MXNet JVM languages.
* Present and educate Apache MXNet in various conferences. (Spark-summit, odsc-mxnet-workshop)

**Amazon Workdocs, AWS - Software Engineer May, 2013 - September, 2016**

* Designed, Built and Launched Amazon WorkDocs from the ground up, built a scalable asynchronous backend fleet.
* Developed scalable backend systems for processing documents, comments and annotations used for collaboration.

**Yahoo! Inc, Sunnyvale, California - Software Engineer Dec, 2010 - May, 2013**

* Developed distributed rate limiter (Throttling Engine) for Yahoo! Cloud Services.
* Developed User Fiber framework used to build scalable servers.
* Development of HTTP Proxy and Caching Servers, which can handle 150K requests/sec on a Quad-Core box as a part of Yahoo! Cloud Technology.

**O1 Communications, Sacramento, California - Software Engineer Feb, 2009 - Dec, 2010**

* Implemented ENUM (RFC 3761, RFC 3764), Least Cost routing Protocols and integrated with SIP signaling and DNS protocols for O1 VOIP Network.
* Development of Network Monitoring and Management Tools.
* Integration of VOIP & PSTN network.

**ANZ Banking Group, India & Australia - Senior Software Engineer Jun, 2005 - Aug, 2007**

* Develop/Maintain Global Payments Hub to process payment message transactions over the SWIFT network.

**Syntel India Ltd, India - Software Engineer Sept, 2003 - Dec, 2004**

* Software development and Maintenance of Accounts Payable system for American Express Credit card.