Dingming Wu

Senior SWE
Bytedance Inc.

+1-281-928-8685
dmwu0506@gmail.com
dmwu.github.io

Experience

•Bytedance Inc. *Bellevue, WA*

Sr. Software Engineer @ Applied Machine Learning 05/2023 - present

I work on building scalable and efficient ML systems and platforms that support DNN model training and serving for various Bytedance businesses including video recommendation, ads, search and e-commerce. I lead several projects in the team:

- **Parameter Gateway**: a scalable and resource-efficient system that support DNN model parameters updates across geographically distributed regions in realtime, which is a key enabler of Bytedance's global video recommendation.
- **Elastic Model Serving**: a Kubernetes based resource and deployment scheduler for model serving that supports auto-scaling and failure recovery.
- **Collective MatMul with Latency Hiding**: a high-performance NCCL based library that supports overlapping matrix-multiplication with various collective communications (e.g., AllGather, ReduceScatter and AllReduce)

Research Scientist @ Network Infra *01/2020-05/2023*

- I started and led a project that automates the WAN and DC bandwidth resource management for Bytedance's internal workloads. The control plane manages 100s of thousands of agents deployed in Bytedance's end-hosts and perform various network functions such as traffic monitoring, traffic-shaping, rate-limiting and QoS. In the agent dataplane, we use both kernel-based technology (Linux TC, Ebpf) and NIC hardware offloading.

•Alibaba Group *Sunnyvale,CA*

Research Intern 08/2018 - 06/2019

- Worked on **programmable switches**. Modeled and designed a network service function chaining framework on programmable switches for edge clouds. Used P4 programming on Barefoot's Tofino Switch.

•Microsoft *Redmond,WA*

SWE Intern 05/2018 - 08/2018

- Worked on **Azure I/O performance tracing**. Built an I/O performance tracing tool for multi-threaded cache driver of Azure. Gained experiences in Windows Device Driver development.

Data Scientist Intern *05/2017 - 08/2017*

- Worked on **cloud-scale data analytics**. Designed and implemented a data-driven model to detect and predict memory leak of Azure system software. Used Bayesian Networks and correlation analysis..

Education

•Rice University, Houston, TX *GPA: 4.03/4.0*

Ph.D. in Computer Science, Advisor: T. S. Eugene Ng 08/2015 - 01/2020

•Nanjing University, Nanjing, China *GPA: 3.67/4.0*

Master in Computer Science 09/2012-07/2015

•Wuhan University, Wuhan, China *GPA: 3.46/4.0*

Bachelor in Computer Science 09/2008-07/2012

Awards and Honors

•Outstanding Graduate Student, Rank 1st of the CS department of NJU *06/2015*

•Graudate National Scholarship, top 3% of NJU *10/2014*

Academic Services

Program Committee Member:

The 15th International Workshop on Cyberspace Security and Artificial Intelligence (CAI-2023)

Reviewer:

IEEE/ACM Transactions on Networking (TON), 2022-2023
IEEE Transactions on Network and Service Management (TNSM), 2022
The Journal of Supercomputing (SUPD), 2022-2023
MultiMedia Tools and Applications, 2023

Talks

Accelerated Service Chaining on a Single Switch ASIC
Hotnets'19, Princeton, NJ, 11/2019
Say No to Rack Boundaries: Towards a Reconfigurable Pod-Centric DCN Architecture
SOSR'19, San Jose, CA, 04/2019
Masking Failures from Application Performance in Data Center Networks with Shareable Backup
SIGCOMM'18, Budapest, Hungary, 08/2018
HyperOptics: A High Throughput and Low Latency Multicast Architecture for Datacenters
HotCloud'16, Denver, CO, 06/2016
Fast and fine-grained counting and identification via constructive interference in WSNs
IPSN'14, Berlin, Germany, 04/2014

Selected Projects and Publications

- **Rackless Pod-Centric Network Architecture:** we designed a rackless DCN architecture that logically removes the rack boundary of traditional data centers and the inefficiencies that come with it. This is achieved by inserting circuit switches at the network edge between the ToR switches and the servers, and by reconfiguring the circuits to regroup servers across racks based on the traffic patterns
  Publication: Weitao Wang, Dingming Wu, Sushovan Das, Afsaneh Rahbar, Ang Chen, TS Eugene Ng, ***USENIX NSDI, 2022***

- **Accelerated Service Function Chaining on Programmable Switches:** we designed a system that can offload a service chain to a programmable switch to achieve high performance and resource efficiency. Our system can compose multiple network functions into a single program that preserves the original chaining requirements, and exploit features of the switch ASIC to efficiently deploy the composed program on a single switch.
  Publication: Dingming Wu, Ang Chen, T. S. Eugene Ng, Guohui Wang, Haiyong Wang, Accelerated Service Chaining on A Single Programmable Switch ASIC ***ACM HotNets 2019.***

- **Towards a Rackless Network Architecture for Data Centers:** we developed a rackless architecture that removes the rack boundary in DCNs and allows servers to talk to each other with uniform high bandwidth. This is achieved by optimizing the network topology for the changing workloads using circuit switches.
  Publication: Dingming Wu, Weitao Wang, Ang Chen, T. S. Eugene Ng, Say No to Rack Boundaries: Towards a Reconfigurable Pod-Centric DCN Architecture, ***ACM SOSR 2019.***

- **Ultra-Fast and Full-Capacity Failure Recovery in Data Center Networks:** we developed a novel network failure recovery approach that can mask failures from application performance. We use a small number of backup switches that are shared network-wide for repairing failures on demand so that the network quickly recovers to its full capacity without applications noticing the failures. This approach avoids the complications and ineffectiveness of rerouting.
  *Publication: Dingming Wu, Yiting Xia, Xiaoye Sun, Simbarashe Dzinamarira, Xin Huang, T. S. Eugene Ng, Masking Failures from Application Performance in Data Center Networks with Shareable Backup,****ACM SIGCOMM 2018***

- **Convertible Data Center Network Architectures:** We propose a convertible datacenter network architectures, called Flat-tree, which can dynamically change the network topology to combine the benefits of multiple architectures. Flat-tree can be implemented as a Clos network and later be converted to approximate random graphs of different sizes, thus achieving both Clos-like implementation simplicity and random-graph-like transmission performance. Testbed evaluation shows the network core bandwidth is increased by 27.6% just by converting the topology from Clos to approximate random graph.
  Publication: Yiting Xia, Xiaoye Steven Sun, Simbarashe Dzinamarira, Dingming Wu, Xin Sunny Huang, T. S. Eugene Ng, A Tale of Two Topologies: Exploring Convertible Data Center Network Architectures with Flat-tree, ***ACM SIGCOMM 2017***