Weitao Wang

8181 Fannin St, APT 918, Houston, TX 77054
wtwang@rice.edu • (+1) 281-236-7550 • weitaowang.site/about

Education

Rice University, Houston, Texas — Aug. 2018 – May 2024 (Expected)

- Ph.D. in Computer Science (Candidate)
- **Courses:** Algorithms Design, Database Implementation, Artificial Intelligence, Data Mining, Operating systems, Cloud Computing, Computer Architecture, Computer Networks, Computer Security
- **Research Interests:** Application-Infrastructure Co-Design, including distributed systems, heterogeneous systems, data center network, programmable hardware, systems for AI, and AI-powered systems.

Shanghai Jiao Tong University, Shanghai, China — Sept. 2014 – Jun. 2018

- B.S. in Engineering • IEEE Honor Class • Major GPA: 90.45 / 100 (Rank: 4 / 69)

Industry Experiences

High Precision Datacenter-scale Clock Synchronization | Google Cloud — May 2022 – Present

- A new large-scale clock synchronization system designed to achieve nanosecond-level precision
- Use periodical probes to detect and eliminate both random errors and systematic error
- Leverage redundancy design to achieve both agile and reliable clock synchronization services
- Reduce the clock synchronization error from tens of nanoseconds to 0.99 ns average and 2.75 ns maximum

Next-generation Data Center CC Via Deployable INT | Google Cloud — May 2021 – May 2022

- Poseidon: a new congestion control for next-generation network with the in-network telemetry (INT)
- Achieve ultra-low latency (< 50 μs), high utilization (> 96%), and max-min fairness (first ever)
- Reduce the average job completion time by 42.4% and tail completion time by 99.1% for RPC workload
- Validate on both commodity and programmable switches, wide deployment expected in next several years

Minimize the Precision Loss in WCMP Deployment | Google Cloud — May 2020 – Aug. 2020

- Explore ILP approaches to minimize the WCMP precision loss within the data center network
- Build a simulator with Integer Linear Programming solvers based on SCIP + Cpp
- Reduce the precision loss up to 60% compared to the current strategy
- Project repository with 6000+ lines of code (Cpp + Proto3 + CMake)

Selected Research Experiences

Schedule MXDAG with In-depth Resource Sharing Knowledge — Apr. 2021 – Present

- Schedule the tasks in MXDAG (a cluster APP abstraction) precisely by understanding the resource sharing
- Parse the code to obtain MXDAG and communicate with the cluster for precise source allocation
- Reduce the job completion time and minimize the resource usage for all the cloud service users

Monitor and Mitigate Cluster Applications' Stragglers In The Network — Apr. 2021 – Present

- Identify and mitigate the stragglers for the distributed applications insides the data center network
- Use the programmble hardware to monitor and identify stragglers with low overhead
- Reallocate the network and computation resources to mitigate the straggler processes

A Default Off Network Diagnose System with Programmable Switches — Mar. 2020 – May 2021

- Achieve network-wide monitoring with zero overhead and reactive diagnosing with low latency
- Implement the prototypes on both **Barefoot Tofino** Switches and **NS3-Bmv2** simulator
- Reduce the memory overhead by more than **99%** comparing to a record-all monitoring solution

A Reconfigurable Pod-Centric DCN Architecture — Oct. 2018 – May 2020

- Optimize the data center network traffic with the dynamically reconfigurable network topology
- Deploy a **Hadoop**/**MPI**/**Memcached** datacenter prototype with 16 servers and 5 **openflow** switches

- Achieve an average path length **35%** shorter and improve the job completion times by **1.1-2.7x**.

Support High Throughput Low Delay Multicast with Optical Network Apr. 2019 – Mar. 2020

- Support multicast traffic with optimal bandwidth provisioning using a dedicated network
- Implemented a **Hadoop/MPI** network prototype with 16 servers and 5 OpenFlow switches
- Speedup raw broadcast **2.6x** and improve end-to-end application performance by up to **23%**.

Selected Publications

- [NSDI'23] Poseidon: Efficient, Robust, and Practical Datacenter CC via Deployable INT
  Weitao Wang, Masoud Moshref, Yuliang Li, Gautam Kumar, T. S. Eugene Ng, Neal Cardwell, Nandita Dukkipati
- [NSDI'22] RDC: Relieving Data Center Network Congestion with Topological Reconfigurability at the Edge
  Weitao Wang, Dingming Wu, Sushovan Das, Afsaneh Rahbar, Ang Chen, T. S. Eugene Ng
- [NSDI'22] SpiderMon: Harnessing Wait-For Relations for Performance Debugging with Programmable Switches
  Weitao Wang, Xinyu Crystal Wu, Praveen Tammana, Ang Chen, T. S. Eugene Ng
- [ToN'22] Shufflecast: An Optical, Data-rate Agnostic and Low-Power Multicast Architecture for Next-Generation Compute Clusters
  Sushovan Das, Afsaneh Rahbar, Xinyu Crystal Wu, Zhuang Wang, **Weitao Wang**, Ang Chen, T. S. Eugene Ng
- [HotNets'21] MXDAG: A Hybrid Abstraction for Cluster Applications
  Weitao Wang, Sushovan Das, Xinyu Crystal Wu, Zhuang Wang, Ang Chen, T. S. Eugene Ng
- [OptSys'21] Abstractions for Reconfigurable Hybrid Network Update and A Consistent Update Approach
  Weitao Wang, Sushovan Das, T. S. Eugene Ng
- [OptSys'21] Towards All-optical Circuit-switched Network Cores: Mitigating Traffic Skewness at the Edge
  Sushovan Das, **Weitao Wang**, T. S. Eugene Ng
- [SoSR'20] Grasp the Root Causes in the Data Plane: Diagnosing Latency Problems with SpiderMon
  Weitao Wang, Praveen Tammana, Ang Chen, T. S. Eugene Ng
- [SoSR'19] Say No to Rack Boundaries: Towards A Reconfigurable Pod-Centric DCN Architecture
  Dingming Wu, **Weitao Wang**, Ang Chen, T. S. Eugene Ng

Skills

Programming language: C/C++, Python, P4, MATLAB, Java, Labview, Verilog, HTML5/CSS3

Tools & Platforms: Tofino, Openflow, Spark, Hadoop, BMv2, Mininet, NS3, Gurobi, SCIP