Hao Wang | Research Statement

My primary research interests are in the fields of large-scale data analytics, distributed computing systems, and machine learning. Big data has dramatically improved public welfare by guiding people to make better decisions with the insights examined from data. Driven by a massive amount of data, navigation applications help drivers avoid congestion, online markets provide personalized recommendations for individual users, and doctors perform accurate diagnosis and treatment of diseases. The prosperity of data analytics and machine learning owes to the advances in distributed computing systems, which make it possible to process large volumes of data in parallel efficiently. As the distributed architecture enables scalable and parallel computation with thousands of servers, new challenges also arise — managing resources, scheduling computing tasks, and optimizing queries across servers are non-trivial. Most of the existing solutions are based on either simplified formulations of distributed systems or heuristics derived from prior experience, such as the network between servers is usually a bottleneck. However, these solutions might only work well in a few cases, because the status of distributed computing systems is in change, just as the fluctuating network bandwidth.

In my research, I have been actively applying innovative machine learning techniques to improve the performance of distributed computing systems. Recently, machine learning algorithms have been widely used to model complex problems and perform tasks of classification and prediction, such as face recognition and sales forecasting. By incorporating machine learning models into distributed computing systems, the systems can learn to optimize their strategies and policies in resource provisioning, task scheduling, and query optimization. With this research philosophy in mind, I build intelligent distributed computing systems with both machine learning algorithms and practical implementations to improve the performance of data analytics and machine learning applications. Therefore, my research is to develop "machine learning for systems" and build "systems for machine learning."

Past and Current Research

My research started from network traffic prediction and bottleneck detection to query optimization and task scheduling. The following projects studied the internal of distributed systems and practiced data-driven methodologies to improve the efficiency and robustness of large-scale data analytics.

Predicting network traffic for distributed computing systems. Distributed computing systems such as Hadoop and Spark are widely deployed for large-scale data analytics and generate massive network traffic. These frameworks provide simplified programming interfaces that ease the development of data analytic applications. These interfaces hide most of the underlying details from developers, such as network conditions. However, most existing scheduling algorithms for data analytics assume that network traffic information is known. It is challenging to accurately predict the network traffic due to the complexity of distributed computing systems and the diversity of data analytic applications.

By observing the execution of popular data analytic frameworks, I found that the frameworks represent computations and data operations with directed acyclic graphs. Such graphs contain the required information to estimate the traffic volume, the data source and destination. I implemented a prototype system based on Spark that measures and predicts network traffic by analyzing the directed acyclic graphs. This work [C6] was accepted by IEEE ICDCS 2015.

Mitigating bottlenecks in wide-area data analytics. The varying bandwidth demands observed in [C6] inspired me to study the performance of wide-area data analytics at runtime. The growing data volume makes it increasingly thorny to speed up the execution of data analytic queries across global datacenters. Existing solutions have been largely motivated by pre-established mantras (*e.g.*, bandwidth scarcity). However, those pre-established mantras are not always true. By examining the execution of data analytic queries, I found resources such as disk I/O and memory capacity can also become performance bottlenecks. Intuitively, mitigating performance bottlenecks can speed up query execution. Unfortunately, such a high-level intuition has not yet been well explored in the literature.

I designed a new system that performs data-driven performance analysis to mitigate bottlenecks and minimize query response times. I also developed time series models to analyze key performance metrics, which enable the system to detect bottlenecks at runtime and avoid assigning tasks to worker nodes in bottlenecks. This work [C4] was published on USENIX HotCloud 2017. I further evaluated both the bottleneck detection algorithms and the scheduling policy. The extended work [J2] was published in IEEE TNSE 2018.

Dynamic query optimization in global data analytics. The bandwidth fluctuations between datacenters excessively inflate the response times of data analytic queries, which are fatal for mission-critical tasks such as topic statistics, content recommendation, and online advertising. Resources in distributed systems, especially the inter-datacenter bandwidth, naturally vary over time during query execution. An "optimal" static query execution plan predetermined by prior heuristic is unlikely to remain optimal, particularly when joining large tables across datacenters. Thus, the execution of queries must be orchestrated dynamically to realize the promised benefits and full potential.

In this project, I proposed a new query optimizer that alters query execution plans on-the-fly. Based on the query cost estimated by a carefully designed machine learning algorithm, the query optimizer selects the optimal query execution plan in response to resource variations. The query optimizer is designed to be compatible with existing query engines such as Spark SQL, Hive, and Pig. This work [C3] was published on ACM SoCC 2018. I further explored deep learning techniques to estimate the query cost and submitted the extended version [J1] to IEEE TPDS.

Distributed machine learning with a serverless architecture. Serverless architectures, represented by AWS Lambda, have emerged as a burgeoning cloud computing model with low maintenance and autoscaling capability. Machine learning systems such as TensorFlow and PyTorch are typically deployed on a dedicated cluster of physical or virtual machines, posing non-trivial cluster management overhead to machine learning practitioners and data scientists. Also, the predetermined resource plan of cluster-based systems can hardly satisfy the varying demands during the life-cycle of machine learning development, leading to either resource shortage or resource waste. Instead, serverless architectures dynamically allocate resources and hide administration details such as capacity planning and maintenance operations from developers. Researchers at Berkeley have successfully implemented PyWren, a scientific computing system prototype on AWS Lambda.

I designed a new framework for distributed machine learning based solely on serverless architectures. To fully utilize the high concurrency and elasticity of serverless architectures, I also proposed a new hybrid synchronous parallelism that seeks to preserve the advantages of both synchronous parallelism and asynchronous parallelism for distributed machine learning. Another highlight is an experience-driven scheduler for serverless architectures, designed to minimize the training time and cost of machine learning jobs. The scheduler learns the best way to trade-off between model quality and resource allocation by deep reinforcement learning. This work [C2] was published on IEEE INFOCOM 2019.

Optimizing federated learning on non-IID data. Collecting private data to a central cluster is hardly possible due to privacy concerns and data protection policies such as GDPR. The emerge of federated learning has addressed this fundamental issue. Federated learning is a decentralized machine learning technique that performs training on devices and exchanges only model updates, preserving data privacy and security. This unique setting of federated learning has addressed critical issues on data privacy and security but also introduced a major statistical challenge — data on each device follow a distinct distribution. Since existing machine learning algorithms expect the training data are independent and identically distributed (IID), the non-IID data on devices can significantly degrade the performance of federated learning.

I proposed an experience-driven framework for federated learning, which actively selects client devices for training to deal with non-IID data. Through both empirical studies and mathematical analysis, I found the correlation between the weights of a model and the distribution of its training data. Based on this correlation, the framework can indirectly profile the data distribution on each device without access to the raw data. I further designed a deep reinforcement learning agent that constructs a specific subset of client devices by analyzing the profiled data distribution. A junior undergraduate student implemented the framework under my supervision, which can speed up the convergence of federated learning up to 42% on non-IID data, compared to existing solutions. This work [C1] has been accepted by IEEE INFOCOM 2020.

Future Research Plans

I am extremely passionate and excited about the potential of my research area. Deep learning models are becoming increasingly complicated and computation-intensive. Meanwhile, mobile phones and IoT devices are taking over more computation. I plan to extend the scope of distributed computing systems and machine learning engineering in various directions.

In the near term, I will continue my study on **distributed machine learning**, which enables deep learning models to solve complicated problems such as understanding human languages. The scale-up of distributed machine learning also makes it imperative to further optimize the scheduling strategy and communication mechanism of distributed systems. Existing solutions mostly rely on assumptions derived from prior experience, due to a lack of effective methods that profile the resource demands and computation patterns of general machine learning workloads. Besides, the availability of heterogeneous hardware requires new load balancing algorithms to coordinate CPUs, GPUs, TPUs, and other accelerators. The ultimate goal is to boost the innovation of machine learning and artificial intelligence.

Another promising topic is **serverless computing**, known as the next-generation of cloud computing. Serverless computing greatly simplifies cloud computing services that developers can directly run their code as cloud functions. A lot of interesting problems arise here: how to efficiently exchange messages among a massive number of cloud functions? How to coordinate cloud functions with the awareness of data locality and hardware affinity? How to handle fine-grained data operations performed by high concurrent cloud functions? How to enable the collaboration between serverless computing and server-based clusters? Answering these questions will further promote the application of cloud computing to all sectors. I believe my research experience on distributed systems and machine learning engineering can give me an edge on solving these problems.

I am also interested in **federated computation**, which follows the principle of "bringing the code to the data, instead of the data to the code" to comply with increasingly restrictive data protection policies. However, it is challenging to execute code on devices such as mobile phones, which have limited computing capacity, unstable network, and varying availability. I plan to build a federated computing system that can dynamically accommodate learning jobs and adapt to the federated environment. I will study a wide spectrum of practical problems including but not limited to: how to scale the system to large numbers of devices, how to maintain robustness against loosely-connected networks and how to select appropriate devices for federated learning. This research will make a substantial amount of personal data and computing devices available for data analytics and machine learning, which build the cornerstone for big pictures such as precision medicine and smart city.

I plan to seek funding supports from NSF, DoD, DARPA, and industrial companies such as Microsoft Research and Google. I will establish a lab committed to research on distributed computing systems and machine learning. I will collaborate with both system and machine learning researchers and continue pushing forward my vision of "systems for machine learning and machine learning for systems" in a constructive way.