Shaoqi Wang

I am a fifth-year Ph.D. student in University of Colorado, Colorado Springs. My research interests include Big Data Analytics and System, Distributed ML and DL Systems. Currently, I am working on building dynamic resource scheduler for ML/DL clusters. I plan to graduate in Spring 2020 and am actively seeking a full-time position.

Emails: swang@uccs.edu bingowsq@gmail.com Phone: (+1)7193325477
https://ibingoogle.github.io/swang/index.html
https://www.linkedin.com/in/shaoqi-wang-8b392b105/

Education:

- Pursuing Ph.D. in Computer Science
  University of Colorado, Colorado Springs, USA, Fall 2015 - present
- Master Degree in Engineering
  University of Science and Technology of China, China, Fall 2012 - Spring 2015
- Bachelor Degree in Computer Science
  Anhui Normal University, China, Fall 2008 - Spring 2012

Project Experiences:

- **Intern: Nokia Bell Labs, USA, Jun 2019 – Aug 2019**
  Project: Dynamic GPUs Scheduling for Distributed DL Clusters
  For a DL training system built on top Kubernetes (K8s), the job preemption is inefficient since it requires K8s to stop the running job and then start the incoming job. To this end, we propose and develop a solution that attaches a life-cycle controller service to DL jobs in order to offer efficient preemption in K8s. In addition, we take advantage of the efficient preemption to develop a preemption-based dynamic scheduler in K8s.
- **Research Assistant: DISCO Lab, University of Colorado, Colorado Springs, USA, Fall 2015 – present**
  1. **Dynamic Resource Scheduling for Distributed ML Clusters:** Dec 2018 – Present
     The hyper-parameter search for ML model leads to multiple ML jobs. Existing cluster schedulers with static resource allocation in Parameter Server are largely not tailored to multiple ML jobs. To this end, I am working on building dynamic resource scheduler for multiple ML jobs on clusters. I am implementing the scheduler in Apache Yarn.
  2. **iBatch:** Jun 2018 – April 2019
     Scalability of distributed DL training with parameter server architecture is often communication constrained in large clusters. To this end, I proposed iBatch, a novel communication approach that batches parameter communication and forward computation to overlap them with each other. I implemented iBatch in the distributed DL system Intel BigDL. This work is published in AAAI 2019.
  3. **A-BSP:** Aug 2017 – Aug 2018
     Executing distributed ML jobs on Apache Spark follows BSP model, in which the synchronization is significantly delayed by stragglers. To this end, I proposed a novel BSP-based Aggressive synchronization (A-BSP) model based on the convergent property of iterative ML algorithms, by allowing the algorithm to use the updates generated based on partial input data for synchronization. I implemented A-BSP as a light-weight BSP-compatible mechanism in Spark and also extended it onto Petuum system. This work is published in ACM/IFIP Middleware 2018.
  4. **FlexPara:** Feb 2017 – Dec 2017
     Computational skewness is a significant challenge in multi-tenant data-parallel clusters that introduce dynamic heterogeneity of machine capacity in distributed data processing. To this end, I proposed FlexPara, a parameter partition approach that leverages the non-linear relationship and provisions adaptive tasks to match the distinct machine capacity so as to address the skewness in iterative ML jobs on data-parallel clusters. I implemented FlexPara in Apache Spark. This work is

published in IEEE INFOCOM 2019.

5. **Dawn:** May 2016 – May 2017
   The performance of parallel jobs is often constrained by the cluster's hard-to-scale network bisection bandwidth. To this end, I proposed Dawn, a dependency-aware network-adaptive scheduler that aggregates and co-locates the data and tasks of dependent jobs to improve data locality. I implemented Dawn on Apache Hadoop and Apache Tez. This work is published in IEEE ICAC 2017 and IEEE TPDS.

Technical Skills:

- Programming Language: Java, Python, Scala, Shell script
- Big Data Systems: Apache Hadoop, Apache Spark
- Distributed ML/DL Systems: Spark MLlib, Intel BigDL, Tencent Angel, TensorFlow, Kubernetes
- Machine Learning and Deep Learning Algorithms

Publications and Technical Reports:

1. Pufferfish: Container-driven Elastic Memory Management for Data-intensive Applications
   Wei Chen, Aidi Pi, **Shaoqi Wang**, Xiaobo Zhou
   in ACM Symposium on Cloud Computing (SoCC 2019), California, USA.
2. OS-Augmented Oversubscription of Opportunistic Memory with a User-Assisted OOM Killer
   Wei Chen, Aidi Pi, **Shaoqi Wang**, Xiaobo Zhou
   in ACM/IFIP International Middleware Conference (Middleware 2019), California, USA.
3. Semantic-aware Workflow Construction and Analysis for Distributed Data Analytics Systems
   Aidi Pi, Wei Chen, **Shaoqi Wang**, Xiaobo Zhou
   in ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC 2019), USA.
4. Addressing Skewness in Iterative ML Jobs with Parameter Partition
   Shaoqi Wang, Wei Chen, Xiaobo Zhou, Sang-Yoon Chang, Mike Ji
   in IEEE International Conference on Computer Communications (INFOCOM 2019), Paris, France.
5. Scalable Distributed DL Training: Batching Communication and Computation
   Shaoqi Wang, Aidi Pi, Xiaobo Zhou
   in AAAI Conference on Artificial Intelligence (AAAI 2019), Honolulu, Hawaii, USA.
6. Aggressive Synchronization with Partial Processing for Iterative ML Jobs on Clusters
   Shaoqi Wang, Wei Chen, Aidi Pi, Xiaobo Zhou
   in ACM/IFIP International Middleware Conference (Middleware 2018), Rennes, France.
7. Dependency-aware Network Adaptive Scheduling of Data-Intensive Parallel Jobs
   Shaoqi Wang, Wei Chen, Xiaobo Zhou, Liqiang Zhang, Yin Wang
   in IEEE Transactions on Parallel and Distributed Systems (TPDS), August 2018.
8. Scalable Distributed Machine Learning on Data-Parallel Clusters
   Shaoqi Wang
   Thesis Proposal, Jan 2018.
9. Performance Isolation of Data-Intensive Scale-out Applications in a Multi-tenant Cloud
   Palden Lama, **Shaoqi Wang**, Xiaobo Zhou, Dazhao Cheng
   in IEEE International Parallel and Distributed Processing Symposium (IPDPS 2018), Canada.
10. Characterizing Scheduling Delay for Low-latency Data Analytics Workloads
    Wei Chen, Aidi Pi, **Shaoqi Wang**, Xiaobo Zhou
    in IEEE International Parallel and Distributed Processing Symposium (IPDPS 2018), Canada.
11. Improving Utilization and Parallelism of Hadoop Cluster by Elastic Containers
    Yinggen Xu, Wei Chen, **Shaoqi Wang**, Xiaobo Zhou, Changjun Jiang
    in IEEE International Conference on Computer Communications (INFOCOM 2018), Hawaii, USA.
12. Network-Adaptive Scheduling of Data-Intensive Parallel Jobs with Dependencies in Clusters
    Shaoqi Wang, Xiaobo Zhou, Liqiang Zhang, Changjun Jiang
    in IEEE International Conference on Autonomic Computing (ICAC 2017), Columbus, OH, USA