Chen Wang

IBM T. J. Watson Research Center
1101 Kitchawan Rd
Yorktown Heights, NY 10598

Office: 914-945-1773, Cell: 914-792-8658
Email: Chen.Wang1@ibm.com, wangchen615@ieee.org
LinkedIn: https://www.linkedin.com/in/chenw615/
Website: https://researcher.watson.ibm.com/researcher/view.php?person=ibm-Chen.Wang1
Google Scholar: https://scholar.google.com/citations?user=JL6iWLgAAAAJ

RESEARCH INTERESTS

- **Data-driven resource management in the Cloud:** Control flow graph mining for Kubernetes & OpenShift; Workload profiling; Automated intelligent resource provisioning; Resource auto-scaling; Resilient scheduling
- **Deep learning in the Cloud:** The run-time estimation for deep learning jobs; Deep learning platform in Container Cloud
- **Cloud measurement:** Public Cloud CDN comparison; Networking measurements for Cloud Infrastructures and Services
- **Quality of Experience (QoE) driven Management in Video Streaming System**: QoE based adaptive server selection; Distributed QoE monitoring system; QoE based networking anomaly detection, localization, and diagnosis
- **Video processing and retrieval system:** Video shot classification; Content-based video copy detection system**;** Object extraction from videos, including embedded text and logo

EDUCATION

Carnegie Mellon University (CMU) — Pittsburgh, PA
Ph.D. in Electrical and Computer Engineering — August 2017
Dissertation: QoE based management and control for large-scale VoD in the Cloud
Ph.D. Advisors: Hyong Kim (ECE, CMU), Ricardo Morla (FEUP, Porto, Portugal)

Carnegie Mellon University — Pittsburgh, PA
Master of Science in Electrical and Computer Engineering — May 2014
GPA: 3.92/4.0

Xi'an Jiaotong University (XJTU) — Xi'an, Shaanxi, China
Master of Science in Information Engineering — July 2010
GPA: 90.2/100.0

Xi'an Jiaotong University — Xi'an, Shaanxi, China
Bachelor of Science in Information Engineering — July 2007
GPA: 87.1/100.0

WORKING EXPERIENCE

IBM Thomas J. Watson Research Center — 1101 Kitchawan Rd, Yorktown Heights, NY, 10598
Research Staff Member — 09/05/2017 – Now
Manager**:** Dr. Alaa Youssef

- **Data-driven Resource Management for Container Cloud** — **Role: Leading**

- kCFG: Detecting Anomalies in Kubernetes via mining Control Flow Graphs from logs.

Apply template mining for logs from the control plane of Kubernetes to extract control flow graphs for Kubernetes resources. Discover common and abnormal patterns from the extracted control flow graphs. Propose to use Graph Neural Network (GNN) to classify the control flow graphs for anomaly detection in Kubernetes. The work has 2 patent applications filed.

- kubeRL: A Reward-based Resource Management Framework in Kubernetes

Perform log analysis on large-scale Kubernetes clusters running heterogeneous workload, discovering the fact that recurring failures in Kubernetes clusters are caused by either admitting workloads with similar features or scheduling workloads to nodes with specific features. Propose and develop a reward-based admission controller to prevent admitting recurring workload prone to fail and a reward-biased scheduler extender in Kubernetes to prevent scheduling a particular workload to nodes where it is prone to fail. Contribute the scheduler and the admission controller software to the open-source community at https://github.com/IBM/kube-safe-scheduler/tree/master/kubeRL . File 2 patent applications about this work.

- kAdvisor: A Container Resource Profile Advisor for Heterogeneous Workload in Kubernetes Cluster.

Apply statistical analysis to profile workload using container resource usage metric data in the container cloud. Propose an intelligent resource profile advisor to ensure efficient resource utilization in large scale Kubernetes clusters running heterogeneous workload. The resource profile advisor can help reduce resource utilization for more than 50% for enterprise clusters running various machine learning services while maintaining similar service availabilities. Present one poster paper in ACM SOCC 2018.

- AI4DL: Mining Behaviors of Deep Learning Workloads for Resource Management in Container Cloud.

Discover that containers running different deep learning (DL) workloads exhibit repeating patterns and similitude of resource usage. Present a methodology to discover resource usage behaviors for the training workloads of Deep Learning (DL) models. Specifically, the approach combines Conditional Restricted Boltzmann Machines (CRBMs) and clustering techniques to discover common sequences of behaviors (phases) of containers running the model training workloads in clusters providing IBM Deep Learning Services. Results show that typical patterns of resource usage behaviors discovered by AI4DL can help reduce CPU and memory overprovisioning by 30%, while maintaining the under-provisioning within 5% through time. The work has 1 paper presented in USENIX HotCloud 2020 and 1 paper presented in IEEE CLOUD 2020. The AI4DL package is open sourced at https://github.com/HiEST/AI4DL.

- DoppelGANger: Generating High-fidelity, Synthetic Time Series Datasets.

Propose to use DoppelGANger to synthesize container resource usage in enterprise clouds with fidelity, as well as preserving business secrets and privacies. DoppelGANger is a synthetic data generation framework based on generative adversarial networks (GANs). DoppelGANger learns features and characteristics from the resource usage of millions of containers running the heterogeneous workload in production clouds. It then generates synthetic container resource usage data that reflects the characteristics of the real workload but does not reveal any personal or business information. The DoppelGANger and the synthetic dataset are available in the open-source community to help researchers in the Cloud computing field. The work was presented at ACM IMC 2020 and was in the best paper finalist. DoppelGANger package is open sourced at https://github.com/fjxmlzn/DoppelGANger.

- **Data-driven Optimization for Deep Learning Service in the Cloud** **Role: Participant**

- AI Gauge: A Runtime Estimation Service for Deep Learning in the Cloud

Propose AI Gauge, a cloud service to estimate runtime and cost for training deep learning models under different configuration options on the cloud. Mainly work on the design of AI Gauge service in the container cloud platform. AI Gauge can accurately predict the remaining time of running jobs based on its runtime progress (< 10% relative error) and can accurately predict the total runtime for a job before it starts with a 7-8% relative error on average. The work has 1 paper presented in SBAC-PAD 2019.

- FfDL: A Flexible Multi-Tenant Deep Learning Platform

Performance analysis for FfDL, a deep learning platform open-sourced and used at IBM. Help examine FfDL quantitatively through a detailed empirical evaluation. The work includes the overheads introduced by the platform for various DL models, the load and performance observed in a real case study using FfDL within our organization, and the frequency of various faults observed, including faults that we did not anticipate. The work has 1 paper presented in MIDDLEWARE 2019.

Carnegie Mellon University 5000 Forbes Ave, Pittsburgh, PA, 15213
Research Assistant 09/01/2016 – 08/31/2017
Advisor**:** Prof. Hyong Kim and Prof. Ricardo Morla

- **QRank: Inferring ISP Roles based on End-user QoE for Large-scale VoD service**

- Propose and develop QRank, a scalable ranking system, to rank the ISPs based on end-user QoE at run time. QRank uses run-time user QoE to study and discover the roles of ISPs in the video streaming services under the assumption of abolishing network neutrality.
- Evaluate QRank in the production Cloud and obtain exciting insights from extensive experiments. Find that 91.4% of QoE anomalies are associated with 15.32% of users. These users experience QoE anomalies persistently and recurrently. More than 95% of persistent and

recurrent QoE anomalies are identified in less than 10 transit networks. QRank results indicate that the limited capacity in transit networks are the primary cause of QoE anomalies.

- **Evaluation of Production Cloud CDNs for Video Streaming Services**

- Compare Production Cloud providers, including Amazon Web Service (AWS), Microsoft Azure, and Google Cloud, regarding their geographical coverage, scalability, stability, and cost for video streaming services. Users' Quality of Experience (QoE) is the evaluation metric. The findings of these providers are obtained and presented in IEEE CLOUD 2018.

INESC TEC Rua Dr. Roberto Frias, Porto, Portugal, 4200-465
Research Assistant 09/01/2013 – 05/31/2016
Advisor**:** Prof. Hyong Kim and Prof. Ricardo Morla

- **QoE based Management and Control for VoD in the Cloud**

- Model user Quality of Experience (QoE) as rewards in the Management system of VoD. Propose and develop QoE based Management System for VoD in the Cloud (QMan). QMan monitors all users' QoE at run time and applies reinforcement learning techniques to adaptively select CDNs/servers for users to optimize the overall user QoE.
- Evaluate QMan in a VoD deployed in Clouds (servers/CDNs) with 300+ DASH clients worldwide emulated in PlanetLab. Results show that the QMan (with server selection) improve 90th percentile QoE more than 40% compared to the DNS based server selection system; QMan (with CDN selection) improves 90% percentile QoE over 20% compared to VoD using single CDN.

- **QWatch: Detecting and locating QoE anomalies for VoD in the Cloud**

- Design a scalable monitoring system that detects and locates anomalies based on the end-user QoE at run time. Evaluate QWatch in a controlled VoD system and a VoD system deployed in Microsoft Azure Cloud. QWatch correctly detects and locates QoE anomalies in servers, routers and client devices.

PUBLICATIONS

1. J. Berral, D. Buchaca, C. Herron, C. Wang, A. Youssef. *"Theta-Scan: Leveraging Behavior-Driven Forecasting for Vertical Auto-Scaling in Container Cloud", accepted to 2021 IEEE International Conference on Cloud Computing (CLOUD). IEEE, 2021*
2. Buchaca, David, J. Berral, **C.Wang**, A. Youssef. *"Proactive Container Auto-scaling for Cloud Native Machine Learning Services." 2020 IEEE 13th International Conference on Cloud Computing (CLOUD). IEEE, 2020.*
3. Lin, Zinan, A. Jain, **C. Wang**, G. Fanti, V. Sekar. *"Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions." Proceedings of the ACM Internet Measurement Conference, IMC 2020 (Citation: 14)*
4. J. L. Berral, **C. Wang**, and A.Youssef. *"AI4DL: Mining behaviors of deep learning workloads for resource management." in 12th USENIX Workshop on Hot Topics in Cloud Computing, HotCloud 2020 (Citation: 3)*
5. K. R. Jayaram, V. Muthusamy, P. Dube, V. Ishakian, **C. Wang,** B. Herta, S. Boag, D. Arroyo, A. Tantawi, A. Verma, Falk Pollok, Rania Khalaf. *"FfDL: A Flexible Multi-Tenant Deep Learning Platform", ACM/IFIP International Middleware Conference, Middleware 2019 (Citation: 4)*
6. P. Dube, T. Suk, **C. Wang**. "*AI Gauge: A Runtime Estimation Service for Deep Learning in the Cloud"*, *accepted to the International Symposium on Computer Architecture and High Performance Computing, SBAC-PAD 2019 (Citation: 3)*
7. **C. Wang**, H. Kim. *"Touchdown on the Cloud: Impact of the Super Bowl on Internet, Cloud and Web Services", in 2019 IEEE Fifth International Conference on Big Data Service and Applications (BigDataService)*. IEEE, 2019 *(Citation: 1)*
8. M. F. Aktas, **C. Wang**, A. Youssef, M. Steinder. *"Resource Profile Advisor for Containers in Cognitive Platform", in Proceedings of the ACM Symposium on Cloud Computing (SOCC).* ACM, 2018.
9. **C. Wang**, A. Yayaseelan, H. Kim. *"Comparing Content Delivery Networks for Adaptive Video Streaming"*, *in IEEE 11th International Conference on Cloud Computing (CLOUD).* IEEE, 2018 *(Citation: 9)*
10. **C. Wang**. *"QoE based Management and Control for Large-scale VoD System in the Cloud"*. Ph.D. Dissertation, 2017.
11. **C. Wang**, H. Kim and R. Morla. "*Identifying Persistent and Recurrent QoE anomalies for DASH streaming in the Cloud*", *in 9th IEEE International Conference on Cloud Computing Technology and Science (CLOUDCOM)*. IEEE, 2017 *(Citation: 2)*
12. **C. Wang**, H. Kim and R. Morla. "*QWatch*: Detecting and Locating QoE anomaly for VoD in the Cloud", *in 8th IEEE International Conference on Cloud Computing Technology and Science (CLOUDCOM)*. IEEE, 2016 *(Citation: 6)*
13. **C. Wang**, H. Kim and R. Morla. *"Users Know Better: A QoE based Adaptive Control System for VoD in the Cloud",* in *IEEE Global Communication Conference (GLOBECOM).* IEEE, 2015 *(Citation: 5)*
14. **C. Wang**, H. Kim, and R. Morla. "QoE driven Server Selection for VoD in the Cloud", in *IEEE International Conference on Cloud Computing (CLOUD)*. IEEE, 2015. *(Citation: 11)*
15. **C. Wang**, H. Wang. *"Utilization of Temporal Continuity in Video Text Detection"*, in *Second IEEE International Conference on Multimedia and Information Technology (MMIT)*. IEEE, 2010. *(Citation: 3)*

16. N. Nan, G. Liu, **C. Wang**. *"A BoF model based CBCD system using hierarchical indexing and feature similarity constraints"*, in *Second International Conference on Internet Multimedia Computing and Service (ICIMCS).* ACM, 2010. *(Citation: 2)*
17. Z. Li, G. Liu, X. Qian, **C. Wang**. *"Scale and rotation invariant Gabor texture descriptor for texture classification"*, *Visual Communications and Image Processing (VCIP)*. SPIE, 2010. *(Citation: 11)*
18. Z. Li, G. Liu, X. Qian, **C. Wang**, Y. Ma and Y. Yang. *"A video text detection method based on key text points"*, *Advances in Multimedia Information Processing-PCM*. Springer, 2010. *(Citation: 2)*
19. N. Nan, G. Liu, X. Qian, **C. Wang**. *"An SVM-based soccer video shot classification scheme using projection histograms"*, *Advances in Multimedia Information Processing-PCM*. Springer, 2008. *(Citation: 7)*

PATENTS

20. **C. Wang**, A. Kanso, S.V. Costache, A. Youssef, *"Reward-based admission controller for resource requests in the cloud." U.S. Patent No. 10,915,369. 9 Feb. 2021. (Citation: 1)*
21. **C. Wang**, S. V. Costache, A. Youssef, A. Kanso, T. Suk, A. N. Tantawi, *"Reward-biased scheduler extender for Cloud systems"*, *Application Filed*
22. **C. Wang**, L. Wu, *"Method and System for Automatic Online Log Template Mining for Multi-source Aggregated Logs", Application filed.*
23. **C. Wang**, A. Kanso, A. Youssef, *"A Method For Identifying Harmful Containers Through Scheduling Decisions", Application filed.*
24. L. Wu, **C. Wang**, *"Method and System for RDF-to-Text Generation with Graph-augmented Structural Neural Encoders", Application filed.*
25. L. Wu, **C. Wang**, *"Graph Learning to Rank: Querying Knowledge Graph with Deep Sub-Graph Matching Networks", Application filed.*
26. L. Wu, L. Yu, **C. Wang**, D. Wang, *"System and Method for Visual Question Generation with Answer-awareness and Region-reference", Application filed.*
27. **C. Wang**, A. Youssef, J. Berral, *"An Evolving Anomaly Learner: Mining Resource Execution Graph Patterns from Streaming Logs Aggregation", Application to be filed.*

Open-Source Software

28. **C. Wang**, *"KubeRL: a software framework implementing a reward biased scheduler extender in Kubernetes",* https://github.com/IBM/kube-safe-scheduler/tree/master/kubeRL
29. **C. Wang**, A. Qadeer, A. Tantawi, *"Trimaran: Real Load Aware Scheduling Plugins",* https://github.com/kubernetes-sigs/scheduler-plugins/tree/master/pkg/trimaran
30. A. Qadeer, **C. Wang**, *"Load Watcher"*, https://github.com/paypal/load-watcher
31. J. Smith, **C. Wang**, et al., *"An Operator for running the Vertical Pod Autoscaler on OpenShift",* https://github.com/openshift/vertical-pod-autoscaler-operator.git

PROFESSIONAL SERVICES

Conference and Workshop Organization

- Industry Track Co-Chair for 22nd ACM/IFIP International Conference on Middleware (MIDDLEWARE 2021)
- Services Congress Publicity Chair for IEEE World Congress on Services (IEEE SERVICES 2021)
- Publicity Chair for the Seventh International Workshop on Container Technologies and Container Clouds (WoC 2021)
- Session Chair for IEEE 11th International Conference on Cloud Computing (IEEE CLOUD 2018): Cloud and Big Data Analytics Workshop

Program Committee

- 2021 IEEE International Symposium on Women in Services Computing (WISC 2021)
- The 22nd ACM/IFIP International Conference on Middleware (MIDDLEWARE 2021)
- The 3rd IPDPS Workshop on Scalable Deep Learning over Parallel and Distributed Infrastructure (ScaDL 2021)
- The 4th Workshop on Hot Topics in Cloud Computing Performance (HotCloudPerf 2021)
- The 2nd IPDPS Workshop on Scalable Deep Learning over Parallel and Distributed Infrastructure (ScaDL 2020)
- The Third Workshop on Hot Topics in Cloud Computing Performance (HotCloudPerf 2020)
- The 5th International Workshop on Container Technologies & Container Clouds (WoC'19)
- The Second Workshop on Hot Topics in Cloud Computing Performance (HotCloudPerf 2019)
- International Conference on Cloud Computing (SCF CLOUD 2019)
- The IEEE Fourth International Workshop on Container Technologies and Container Clouds (WoC'18)

Reviewers

- ACM Transactions on Multimedia Computing, Communications and Applications (TOMM) (2020-now, 3 review)
- ACM Transactions on Knowledge Discovery from Data (TKDD) (2019-now, 1 review)
- IEEE Transaction on Cloud Computing (TCC) (2017-now, 16 reviews)
- IEEE Transaction on Service Computing (2018-now, 2 reviews)
- The Workshop on Scalable Deep Learning over Parallel and Distributed Infrastructure (2020-now, 5 reviews)
- International Workshop on Container Technologies & Container Clouds (2018-now, 10 reviews)
- The Workshop on Hot Topics in Cloud Computing Performance (HotCloudPerf) (2019-now, 5 review)
- International Conference on Cloud Computing (SCF CLOUD 2019, 1 review)
- IEEE 2018 BIGDATA Congress (1 review)
- The International Workshop on Big Data (2017, 1 review)
- IEEE Symposium on Industrial Electronics and Applications (ISIEA 2012, 1 review)

Conference presentations and invited talks

- Oral presentation entitled "Comparing Content Delivery Networks for Adaptive Video Streaming", in IEEE 11th International Conference on Cloud Computing (CLOUD), San Francisco, CA. July 2018.
- Invited talk entitled "QoE based management and control for large-scale VoD in the Cloud", Research Lab of Artificial Intelligence at Accenture, San Jose, CA. February 2017.
- Invited talk entitled "QoE based management and control for large-scale VoD in the Cloud", IBM T. J. Watson Research Center, Yorktown Heights, NY. December 2016.
- Oral presentation entitled "QoE driven Server Selection for VoD in the Cloud", in IEEE International Conference on Cloud Computing (CLOUD), New York City, US. July 2015.
- Oral presentation entitled "Users Know Better: A QoE based Adaptive Control System for VoD in the Cloud", in IEEE Global Communication Conference (GLOBECOM), San Diego, US. December 2015

Award Judge:

- Reviewer for the IBM Research Accomplishment Nominations within Hybrid Cloud Pillar, 2019.

Supervisors

- February 2021 – present, work with PhD student, Beomyeol Jeon, from University of Illinois at Urbana-Champaign on "Resource Management for Elastic Machine Learning on Multi-tenant Clusters"
- October 2020 – present, work with PhD student, Xiao Sun, from Stony Brook University on "Unified Resource Management for Heterogeneous Machine Learning Workload on Multi-tenant Clusters"
- October 2020 – present, work with PhD student, Haoran Qiu, from University of Illinois at Urbana-Champaign on *"Operationalizing FIRM: An Intelligent Fine-grained Resource Management Framework for SLO-Oriented Microservices"*
- September 2019 – present, work with researcher Josep Berral, PhD student David Buchaca and undergraduate student Claudia Herron from Barcelona Supercomputing Center on *"Reinforcement Learning for Container Vertical Auto-scaling"*, *"Leveraging Periodicity Forecasting for Cloud Vertical Auto-Scaling"*, *"Anomaly Forecasting based on Execution Graphs for Kubernetes Clusters"* and *"Online container resource resizing via resource usage similarity analysis."*
- September 2018 – July 2020, work with PhD student, Zinan Lin, from Carnegie Mellon University on *"Generating High-fidelity, Synthetic Time Series Datasets with DoppelGANger"*.
- September 2018 – May 2019, work with PhD student, Samarth Gupta, "Modeling and Detection of cascading failures via non-intrusive measurements in the Cloud"
- May 2018 – August 2019, work with PhD student, Mehmet Fatih Aktas, from Rutgers University on "*Resource Profile Advisor for Containers in Cognitive Platform*" and "*Identifying predictability for containers in the Cloud: statistical analysis on container resource usage*"
- January 2017 – August 2017, work with graduate student, Andal Yayaseelan, from Carnegie Mellon University on *"Comparing Content Delivery Networks for Adaptive Video Streaming"*.

MEMBERSHIPS

- IEEE Membership
- IEEE Women in Engineering Membership
- IEEE Big Data Community
- IEEE Computer Society Technical Committee on Cloud Computing
- IEEE Computer Society Technical Committee on Services Computing

AWARDS and HONORS

- IBM Second Plateau Invention Achievement Award, 2021
- IBM First Patent Issuance Invention Achievement Award, 2020
- IBM First Plateau Invention Achievement Award, 2020
- IBM Manager's Choice Award, 2019
- IBM First Patent Invention Achievement Award, 2018
- IBM Manager's Choice Award, 2017
- AWS Research Grants, 2017, AWS Cloud Credits for Research
- Microsoft Azure Award for Research, 2016, Award opportunities: Microsoft Azure for Research
- NSF Student Travel Grant Award, 2015, GLOBECOM

SKILLS

Programming Language: *Proficient* in Go, Python, C, C++, Java, and Matlab; *familiar with* Django, AngularJS, D3.js, Node.js, jQuery, HTML, CSS, MySQL, Perl, Shell

Cloud/CDN Platform: *Experience* in Docker, Kubernetes, OpenShift, Operators, IBM Cloud, Google Cloud, Amazon AWS, Microsoft Azure, Hadoop

Knowledge: *Solid foundation* in microservices, cloud-native computing, cloud resource management, machine learning systems, networking, distributed systems, information theory, signal processing, video/image processing; good understanding in large-scale distributed system, load balancing, anomaly detection, fault tolerance, DASH streaming, Video-on-Demand system, CDN system and video search system.

TEACHING EXPERIENCE

Carnegie Mellon University Pittsburgh, PA, US

18757 – Network Management and Control Spring 2013

18345 – Introduction to Telecommunication Networks Spring 2013