# Songyuan Li

PH.D. CANDIDATE IN COMPUTER SCIENCE · UNIVERSITY OF EXETER, U.K.

*Department of Computer Science, Innovation Centre, The Queen's Drive, Exeter EX4 4RN*

S.Y.Li@exeter.ac.uk / lisy@ieee.org | songyuanli.github.io | My Google Scholar

## Research Interests

Songyuan has nearly a decade of research experience in distributed systems and networks. His research spans the fields of artificial intelligence (AI) systems, cloud computing, edge computing, and the Internet of Things (IoT). Currently, he focuses on advancing the quality and efficiency of distributed intelligence, including a board of key topics:

- **Distributed machine learning** (e.g., federated learning and distributed data analytics)
- **Edge Intelligence** (e.g., AIoT and resource-efficient ML model inference/training)
- **Generative AI** (e.g., large language models, multimodal models, and mixture-of-experts models)
- **Edge/cloud computing** (e.g., Quality-of-Service optimization and resource management)

## Education

**University of Exeter, U.K.** *Exeter, U.K.*

**Ph.D.** in Computer Science Sept. 2021 – Aug. 2025 (expected)

- Research focus: Edge intelligence, distributed machine learning, and generative AI
- Supervisors: Prof. Jia Hu and Prof. Geyong Min

**Beijing University of Posts and Telecommunications, China** *Beijing, China*

Master's degree in Computer Science and Technology Sept. 2018 – Jun. 2021

- Thesis: QoS-Aware Service Resource Scheduling and Optimization
- Received Outstanding Master's Thesis Award (top 1%)
- Supervisors: Prof. Jiwei Huang and Prof. Bo Cheng

Bachelor's degree in Computer Science and Technology Sept. 2014 – Jun. 2018

- Thesis: QoS Evaluation and Optimization for IoT Services in Edge Computing Architecture
- Received Outstanding Bachelor's Thesis Award (top 3%)

## Research Experience

**High Performance Computing and Networking (HPCN) Research Group** *Exeter, U.K.*

University of Exeter Sept. 2021 – present

- Multi-dimensional resource optimization for multi-exit DNN inference at the network edge
- Incentive mechanism for multi-tenant split federated edge learning of GenAI models
- Personalized federated learning of MoE models for resource-constrained edge environments

**State Key Laboratory of Networking and Switching Technology** *Beijing, China*

Beijing University of Posts and Telecommunications Mar. 2016 – Aug. 2021

- QoS/QoE-aware decentralized resource management and task scheduling in IoT-edge-cloud systems
- Market-oriented resource pricing and demand allocation in cloud environments
- QoS-aware service selection/composition in services computing ecosystem

## Research Areas (Selected)

**Efficient Federated Edge Learning for Large-Scale Generative AI Models** *Jan. 2024 – present*

*supported by EU Horizon, UK EPSRC, etc.*

- Propose a federated foundation model (FM) fine-tuning paradigm at the network edge, named Split Federated Learning (SFL), which alleviates the FM computation burden on local devices through device-edge synergistic fine-tuning
- Tackle the practical challenges of multi-tenant SFL systems, where multiple SFL tenants coexist and independently manage fine-tuning workloads of diversified downstream tasks, each with distinct requirements including FM types, performance targets, and fine-tuning deadlines
- Develop a tailored incentive mechanism that guides multiple SFL tenants to offer strategic price incentives, which solicits high-quality device participation for device-edge synergistic fine-tuning, thereby satisfying their heterogenous FM fine-tuning requirements

**End-to-end DNN Inference Acceleration Solutions for Edge-AI Platforms** *Sept. 2021 – present*

*supported by EU Horizon, UK EPSRC, etc.*

- Perform comprehensive system modelling of the Edge-AI platform, which characterizes the interactions amongst the AI service providers, AI users, and edge infrastructure provider
- Develop a multi-exit device-edge synergistic inference framework that enables AI service providers to offer personalized edge inference services, addressing diversified inference requirements of AI users, in terms of DNN inference accuracy, latency, and task complexity
- Propose a novel DNN inference acceleration solution based on multi-dimensional optimization for edge intelligence

**Performance Evaluation and Optimization of IoT Service System based on Edge Computing Architecture**

*supported by National Natural Science Foundation of China (NSFC) Program* *Jan. 2020 – Dec. 2023*

- Adopt the potential game theory to solve the edge resource allocation problem with QoE maximization in a decentralized manner
- Solve the dynamic QoS-aware task scheduling and resource management problem in mobile edge computing, through designing an efficient optimization algorithm with LP relaxation techniques
- Propose a joint resource allocation and task slicing for mobile multimedia computing based on deep reinforcement learning in edge computing

**Self-adaptive Software Ecosystem with Collaborative Learning among Humans, Machines and Services**

*supported by National Key Research and Development Program of China* *Oct. 2018 – Dec. 2021*

- Propose a price-incentive resource auction mechanism, with the objective of stimulating maximum users willing to purchase cloud resources
- Design a market-oriented cloud pricing strategy which solves the resource pricing and demand allocation for revenue maximization
- Develop a QoS-aware concurrent service selection approach, with the max-min fairness achieved across multiple service requests

**QoS-aware Service Composition in Large-Scale Dynamic IoT Environment** *Mar. 2016 – Dec. 2018*

*supported by National Natural Science Foundation of China (NSFC) Program*

- Design queueing network models for QoS evaluation of IoT services in edge-cloud systems

- Conduct reliability-aware QoS evaluation for recoverable IoT edge services using the modelling techniques of generalized stochastic Petri nets
- Manipulate the Markov Decision Process-based resource allocation and task scheduling in edge computing paradigm weighing energy costs against QoS requirements

## Grants and Funding

- **PhD RA – UK EPSRC Impact Acceleration Account (IAA)**, "Efficient Federated Edge Learning for Large Language Models" (Grant No. EPSRC/014), awarded **63K GBP**, Oct. 2024 – Sept. 2025.
- **PhD RA – UK EPSRC**, "Real-time Federated Learning at the Wireless Edge via Algorithm Hardware Co-design" (Grant No. EP/X019160/1), awarded **192K GBP**, Mar. 2023 – Nov. 2024.
- **Co-I – National Key Research and Development Program of China**, "Self-Adaptive Scheme of Software Ecosystem with Collaborative learning among Humans, Machines and Services" (Grant No. 2018YFB10-03804), awarded **3.3M RMB**, Oct. 2018 – Sept. 2021.
- **Co-I – National Natural Science Foundation of China (NSFC)**, "QoS Evaluation Research for Large-Scale Dynamic Service Environment" (Grant No. 61502043), awarded **210K RMB**, Jan. 2016 – Dec. 2018.

## Publications

Google Scholar: https://scholar.google.com/citations?user=cESBqqkAAAAJ

* *Click the link in the publication title for further details.*

**Journal Publications**

J8. [ToN'25] **S. Li,** J. Hu, G. Min, and H. Huang. **Incentivizing Multi-Tenant Split Federated Learning for Foundation Models at the Network Edge,** IEEE/ACM Transactions on Networking (Under Review), 2025.

J7. [TAAS'24] J. Huang, Y. Leng, J. Bao, **S. Li,** and Y. Chen. **Joint Resource Allocation and Task Slicing for Mobile Multimedia Computing in Edge-based Autonomous Systems,** ACM Transactions on Autonomous and Adaptive Systems (Under Review), 2024.

J6. [TMC'23] **S. Li,** J. Hu, G. Min, H. Huang, and J. Huang. **Dynamic Pricing for On-Demand DNN Inference in the Edge-AI Market,** IEEE Transactions on Mobile Computing (Under Review), 2023.

J5. [TCCN'22] **S. Li,** J. Huang, J. Hu, and B. Cheng. **QoE-DEER: A QoE-Aware Decentralized Resource Allocation Scheme for Edge Computing,** IEEE Transactions on Cognitive Communications and Networking, vol. 8, no. 2, pp. 1059-1073, 2022.

J4. [TNSM'21] **S. Li,** J. Huang, and B. Cheng. **Resource Pricing and Demand Allocation for Revenue Maximization in IaaS Clouds: A Market-Oriented Approach,** IEEE Transactions on Network and Service Management, vol. 18, no. 3, pp. 3460-3475, 2021.

J3. [TNSM'21] **S. Li,** J. Huang, and B. Cheng. **A Price-Incentive Resource Auction Mechanism Balancing the Interests Between Users and Cloud Service Provider,** IEEE Transactions on Network and Service Management, vol. 18, no. 2, pp. 2030-2045, 2021.

J2. [PPNA'20] J. Huang, **S. Li,** and Y. Chen. **Revenue-Optimal Task Scheduling and Resource Management for IoT Batch Jobs in Mobile Edge Computing,** Peer-to-Peer Networking and Applications, vol. 13, no. 5, pp. 1776-1787, 2020.

J1. [IJWGS'18] J. Huang, **S. Li,** Y. Chen, and J. Chen. **Performance Modelling and Analysis for IoT Services,** International Journal of Web and Grid Services, vol. 14, no. 2, pp. 146-169, 2018.

**Conference Publications**

C3. [ICWS'19] **S. Li,** J. Huang, B. Cheng, L. Cui and Y. Shi. **FASS: A Fairness-Aware Approach for Concurrent Service Selection with Constraints,** IEEE International Conference on Web Services, Jul. 8-13, 2019, Milan, Italy.

C2. [ISPA'17] **S. Li** and J. Huang. **Energy Efficient Resource Management and Task Scheduling for IoT Services in Edge Computing Paradigm,** IEEE International Symposium on Parallel and Distributed Processing with Applications, Dec. 12-15, 2017, Guangzhou, China.

C1. [SCC'17] **S. Li** and J. Huang. **GSPN-Based Reliability-Aware Performance Evaluation of IoT Services,** IEEE International Conference on Service Computing, June 25-30, 2017, Honolulu, Hawaii, USA.

**Ongoing Works**

O2. **S. Li,** J. Hu, G. Min, *et al*. **Resource-Efficient Personalized Federated Learning for Mixture-of-Expert Models at the Resource-Constrained Edge.**

O1. H. Huang, **S. Li,** *et al*. **Federated Foundation Models for Accurate Disease Detection.**

## Invited Talks (Selected)

- **Incentivizing Multi-Tenant Split Learning for Federated Foundation Models** *Mar. 2025*
  Flower AI Summit, London, U.K.
- **Accelerating Split Federated Learning for Large-Scale Generative AI Models** *Feb. 2025*
  Lightning Talks in Superintelligence and AI, Institute for Data Science and Artificial Intelligence Exeter, U.K.
- **On-demand DNN Inference Services in the Edge-AI Market** *Jul. 2024*
  Computer Science PGR Research & Career Workshop, University of Exeter, U.K.
- **Models and Solutions of QoS Optimization for Services Computing Ecosystem** *Jul. 2021*
  Seminar at the Services Computing Laboratory, China University of Petroleum - Beijing, China
- **FASS: A Fairness-Aware Approach for Concurrent Service Selection with Constraints** *May 2019*
  Outstanding Young Scholar Symposium co-located with International Conference on Service Science (ICSS) Tianjin, China

## Professional Services

**Conference Organization:**

- **Publicity Chair** of International Conference on Software Engineering and Development, ICSED 2024 and 2023
- **Local Arrangement Chair** of IEEE International Conference on Data, Information, Knowledge, and Wisdom, DIKW 2025
- **Local Arrangement Chair** of International Conference on Ubiquitous Security, UbiSec 2023
- **Session Chair** of IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2023
- **Session Chair** of International Symposium on Intelligent and Trustworthy Computing, Communications, and Networking, ITCCN 2023

**Technical Program Committee (TPC) Member/External Reviewer** (selected)**:**

- IEEE International Conference on Software Engineering and Artificial Intelligence, SEAI 2022-2025

- IEEE International Conference on Ubiquitous Computing and Communications, IUCC 2024 and 2021
- IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2023
- EAI International Conference on Collaborative Computing, CollaborateCom 2023, 2022, and 2020
- IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing, CCGrid 2022
- IEEE International Conference on Ubiquitous Intelligence and Computing, UIC 2022
- IEEE International Conference on Smart Computing, Networking and Services, SmartCNS 2021
- IEEE International Conference on Data Science and Computational Intelligence, DSCI 2021
- IEEE International Conference on Web Services, ICWS 2021
- IEEE Vehicular Technology Conference, VTC 2020-Fall, etc.

**Journal Reviewer** (selected)**:**

- IEEE Transactions on Mobile Computing
- IEEE Transactions on Emerging Topics in Computational Intelligence
- IEEE Transactions on Cognitive Communications and Networking
- IEEE Transactions on Network and Service Management
- IEEE Internet of Things Journal
- Chinese Journal of Electronics
- Journal of Cloud Computing
- Computer Communications
- Scientific Reports, etc.

## Honours & Awards

- **Outstanding Master's Thesis Award** (award rate: approx. 1%) *2021*
  Beijing University of Posts and Telecommunications, China
- **Outstanding Postgraduate Student Award** (award rate: approx. 5%) *2021*
  Beijing Municipal Education Commission, China
- **China National Scholarship** (award rate: approx. 2%) *2019*
  Ministry of Education of the P.R. China
- **Outstanding Bachelor Dissertation Award** (award rate: approx. 3%) *2018*
  Beijing University of Posts and Telecommunications, China

## Work Experience

**Institute of Linguistics, Chinese Academy of Social Sciences** *Beijing, China*

*Android Software Developer* Oct. 2018 – Jan. 2020

- Develop an Android application (Pronunciation Test for Young Children) in Java to collect children's voice recordings and textual feedback for phonological development research.
- Design and implement a Java-based application server to manage test content delivery and data collection for the Android app.

## Skills and Competences

- **Machine Learning:** language models, multimodal models, computer vision models, federated learning,

transfer learning, efficient ML model inference

- **Mathematics:** convex optimization, operational research, probability theory, game theory
- **Programming:** Python, Java, C/C++, MATLAB
- **ML Libraries:** Pytorch, Hugging face Transformers
- **Network Protocols & Standards:** TCP/IP, HTTP/HTTPS, RESTful APIs
- **Soft Skills:** presentation and communication, critical and out-of-box thinking, teamwork, flexibility

## Teaching

**Professional Certification:**

- Associate Fellow of the Higher Education Academy (AFHEA)
- Advanced Certificate in Learning and Teaching in Higher Education (30 credits at NQF Level 7)

**Teaching Experience:**

*Roles:* Co-design the materials for practical sessions/tutorials with module directors, lead the workshops/tutorials, and serve as markers for coursework & exam assessment

- **Department of Computer Science** *University of Exeter, U.K.*

  ECM1414 – Data Structures and Algorithms (1st-year UG Level)
  ECM1407 – Social and Professional Issues of the Information Age (1st-year UG Level)
  ECM2427 – Outside the Box: Computer Science Research and Applications (2nd-year UG Level)
  ECM3420 – Learning from Data (3rd-year UG Level)
  ECMM445 – Learning from Data (PG Level)
  COMM423DA – Work-Based Research Project (NQF Level 7 Apprenticeship)
- **Department of Mathematics and Statistics** *University of Exeter, U.K.*

  MTHM506 – Statistical Data Modelling (PG Level)
- **Doctoral College** *University of Exeter, U.K.*

  Researcher Development Programme (Ph.D. Student Level) including PGR workshop/tutorial delivery

## References

- Prof. Jia Hu, Associate Professor in Computer Science, University of Exeter, U.K. (email: J.Hu@exeter.ac.uk)
- Dr. Shiqiang Wang, Staff Research Scientist, IBM T. J. Watson Research Center, New York, USA (email: shiqiang.wang@ieee.org)