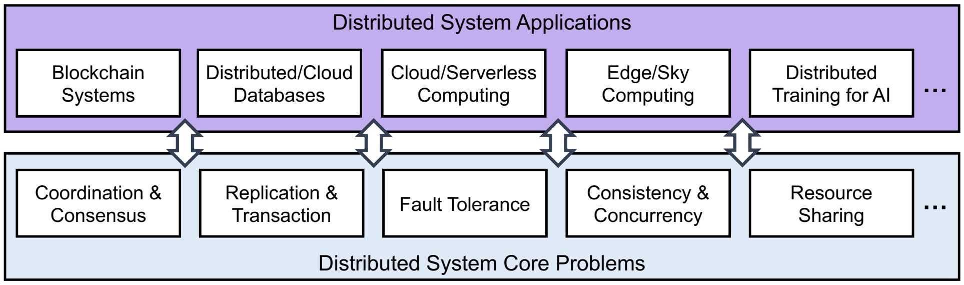
Gengrui Zhang
https://gengruizhang.github.io

Research Statement

My research is at the core of distributed systems. Recent years have seen many of the most significant technological developments of distributed systems with a wide range of applications in use today from relatively localized systems to global-scale systems involving millions of nodes, such as serverless computing, cloud databases, and blockchains. The demand for more scalable, available, and versatile distributed systems has brought forth new challenges in terms of efficiency, robustness, and heterogeneity in system design and architecture.

My research aims to develop algorithms and architectures that build up high-performance, highly scalable, and highly available distributed systems. My Ph.D. research has been bridging distributed system applications and their underlying theoretical foundations, innovating new algorithms and system architectures inspired from real-world scenarios. I have conducted research on the following topics: ① efficient and robust consensus algorithms, ② fault-tolerant protocols for distributed system applications (especially blockchain systems), and ③ fast replication protocols for cloud and distributed databases. My work has been published at prestigious conferences in distributed systems, such as IEEE ICDCS and ACM Middleware [2, 3] (recent work has been submitted to SIGMOD and NSDI [7, 5]); some work has been used by industry collaborators and patented [1, 4].


Distributed System Applications
Blockchain Systems
Distributed/Cloud Databases
Cloud/Serverless Computing
Edge/Sky Computing
Distributed Training for AI
...
Coordination & Consensus
Replication & Transaction
Fault Tolerance
Consistency & Concurrency
Resource Sharing
...
Distributed System Core Problems

My long term research goal is to develop efficient, reliable, and robust algorithms and architectures for large-scale distributed systems, providing computation, coordination, replication, and transaction services as a utility. I envision future research to progress along the following lines of exploration (see above diagram): ① more secure and efficient consensus services (especially blockchain systems); ② ubiquitous coordination among computational resources in cloud, edge, and sky computing; and ③ large-scale distributed training for AIs.

Current Research

CR1: Reputation-based consensus algorithms.

Consensus algorithms conduct state machine replication (SMR) among servers despite failures. Under Byzantine failures, the traditional specification of SMR describes the process of replication but has no language to describe the correctness of participating servers as they may behave arbitrarily and maliciously. This vulnerability leaves the door open for repeated malicious attacks that target leader servers.

My research developed reputation-based consensus algorithms, `Prosecutor` [2] and `PrestigeBFT` [7], that rank server correctness according to their behavior history. `Prosecutor` imposes Proof-of-Work computation on suspected faulty servers during view changes, suppressing Byzantine servers from becoming new leaders. Moreover, `PrestigeBFT` establishes a reputation engine that discredits misbehaved servers using worsening reputations and rewards protocol-obedient servers with improving reputations. Additionally, `PrestigeBFT` enables active view changes where servers proactively campaign for leadership and thus avoids the impact of crash faults during view changes, a major problem faced by state-of-the-art BFT algorithms and blockchain platforms [6].

Impact: My reputation-based consensus algorithms have made impacts on distributed computing theory and system architectures. ① They extend traditional state machine replication properties to a reputation state, opening a new discussion on Byzantine fault tolerance. ② They are the first consensus algorithms that not only achieve high performance but also suppress intentional faults, binding efficiency and robustness. In particular, `PrestigeBFT` achieves 5.4× and 4.2× higher throughput than HotStuff in peak performance under normal operation and sustained Byzantine attacks, respectively.

CR2: Blockchains for vehicle-to-everything (V2X) networks.
Blockchain automotive solutions have been demanded by V2X networks, as automobile manufacturers are becoming a centralized data monopoly that imposes difficulties on regulators and legal systems when delineating legal responsibility between vehicle and driver in the event of an accident. Commonly used BFT algorithms and permissioned blockchains operate in a stable environment, where the system setup affixes the same set of participating servers. However, V2X networks cannot guarantee a stable environment, as vehicles may become online and offline arbitrarily.

My research proposed V-Guard [5], a new permissioned blockchain with a new consensus algorithm for V2X networks, targeting the problem that vehicles are often intermittently connected on the roads. V-Guard separates transaction ordering from consensus and achieves paired consensus that forms two intertwined chains of transactions and configurations, which allows for ordering and consensus to take place in different membership configurations. The source code is publicly available at: https://github.com/vguardbc/vguardbft

Impact: ① V-Guard is the first blockchain architecture that allows consensus to be achieved in a dynamic environment with a high performance. ② This project has filed a US patent [4] and is being used by an industry collaborator. ③ V-Guard's general-purpose architecture can be adopted by applications operating in unstable networks with intermittent connectivity, such as Internet-of-Things, supply chain, and retail applications.

CR3: Fast leader election protocols.
Leader election protocols, electing new leaders through voting-based election mechanisms, are vital to large-scale systems that have a single cluster leader (e.g., GFS and HDFS). However, voting-based mechanisms intrinsically cause competition in leadership candidacy when votes are split among candidates, which prolongs the undesired leader election process.

My research proposes the Escape leader election protocol [3] that investigates Raft's leader election mechanism as a case study and provides solutions that fundamentally resolve split votes. Escape prioritizes servers based on their log responsiveness. It keeps track of servers' logs and assigns configurations that are inclined to win an election to more up-to-date servers, thereby preparing a pool of prioritized candidates. When leader election takes place, Escape can terminate a leader election in one messaging round without suffering from split votes.

Impact: Escape fundamentally resolves the split-vote problem in leader election algorithms, providing a generalized framework that can be adopted by other leader-based systems, such as Zookeeper, Redis, and Azure election protocols.

Future Research

FR1: Software-defined consistency service for distributed systems.
Distributed applications are often demanded to provide various consistency services. However, they often provide only hard-coded consistency services under pre-defined failure assumptions. Future distributed applications should be versatile and support consistency services to be defined at an application level.

Building on my past research on developing consensus algorithms and architectures under various failure models [2, 3, 7, 5, 8], my future research will explore one-size-fits-all architectures that integrate a collection of consistency services with invariants of linearizable, sequential, causal, and FIFO orderings. Software-defined consistency services can innovate the design of blockchain applications and distributed databases, where consistency service can vary by request. My research will focus on the following aspects:

① Coordination as a utility. My research will build up fine-grained consistency service components, including communication, quorum construction, storage, and cryptography, and allow for multiplexing among different consistency services.

② Software-defined consistency. Consistency guarantees can be defined differently across applications by varying requests. As such, the corresponding consistency service can be built by assembling the "Lego" of consistency components while ensuring the invariant of the defined consistency model.

FR2: Cloud, edge, and sky computing.
Computing has become ubiquitous with various computation models, including cloud computing (computation-

centric) and edge computing (data-centric). The development of cloud computing has entered the era of serverless computing, which offers a pay-as-you-go cost model and encompasses automatic, rapid, and unlimited scaling resources up and down per demand. My research will tackle the problem in autuoscaling in cloud computing.

① (Cloud computing.) My research will investigate solutions in serverless computing for achieving fast consensus in horizontal scaling among added nodes (containers) that outperforms vertical scaling under varying workloads.

In contrast to cloud computing's centralized model, edge computing deploys computation near the physical location of either the user or the source of the data. However, since edge devices often have limited computation capability, the coordination between cloud and edge devices becomes necessary when computing on large data sets.

② (Edge computing.) My search will explore efficient coordination algorithms and replication protocols between cloud and edge devices based on varying computation topologies.

Sky computing arises in the context of using multicloud in a heterogeneous architecture (e.g., computing and storage services from different vendors). Sky computing enables interoperability between clouds. It provides compatibility that can seamlessly transfer data among clouds (e.g., from AWS to Google Cloud) based on user-defined criteria.

③ (Sky computing.) My research will focus on developing a peering layer that allows clouds to have agreements with each other about how to fast and securely exchange services.

FR3: Distributed systems for AI.
Recent years have seen explosive growth in the scale and complexity of machine learning applications. The growing gap between the demands of computational resources and the capabilities of even specialized processors leaves system architects no choice but to distribute these computations. For example, Google's TPU v3 Pods consist of up to 1,000 TPUs connected by a high-speed mesh network. While distributed training is becoming the norm, my future research will explore the following topics:

① Coordination in distributed training. My research will focus on data/model parallelism, developing efficient coordination algorithms and protocols for data synchronization, intermediate results aggregation, and supporting scalable training frameworks.

② Fault tolerance. My research will apply and develop fault-tolerant algorithms for distributed training, minimizing the cost of losing intermediate results under failures with proactive recovery from meta states.

References

[1] **Gengrui Zhang**, Tongxin Bai, and Chengzhong Xu. A kind of Second-hand Vehicle Transaction method, apparatus and system based on block chain technology, 2017. CN Patent 106897887 A[P].

[2] **Gengrui Zhang** and Hans-Arno Jacobsen. Prosecutor: An Efficient BFT Consensus Algorithm with Behavior-aware Penalization against Byzantine Attacks. In *Proceedings of the 22nd International Middleware Conference*, 2021.

[3] **Gengrui Zhang** and Hans-Arno Jacobsen. ESCAPE to Precaution against Leader Failures. In *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*, 2022.

[4] **Gengrui Zhang**, Hans-Arno Jacobsen, and Sheng Sun. Method and System for Creating a Distributed Ledger of Verified Vehicle Transactions, 2022. US Patent (Invention Disclosure ID: 10004394).

[5] **Gengrui Zhang**, Yunhao Mao, Shiquan Zhang, Shashank Motepalli, and Hans-Arno Jacobsen. V-Guard: A Fast, Dynamic, and Versatile Permissioned Blockchain Framework for V2X Networks. In *Under view*, 2022.

[6] **Gengrui Zhang**, Fei Pan, Michael Dang'ana, Yunhao Mao, Shashank Motepalli, Shiquan Zhang, and Hans-Arno Jacobsen. Reaching Consensus in the Byzantine Empire: A Comprehensive Review of BFT Consensus Algorithms. *arXiv preprint arXiv:2204.03181*, 2022.

[7] **Gengrui Zhang**, Fei Pan, Sofia Tijanic, and Hans-Arno Jacobsen. Prestige BFT: Making Decentralization Efficient in Distributed Ledgers using Reputation-based Byzantine Fault-Tolerant Consensus Algorithms. In *Under view*, 2022.

[8] **Gengrui Zhang** and Chengzhong Xu. An Efficient Consensus Protocol for Real-time Permissioned Blockchains under Non-Byzantine Conditions. In *International Conference on Green, Pervasive, and Cloud Computing*, pages 298–311. Springer, 2018.