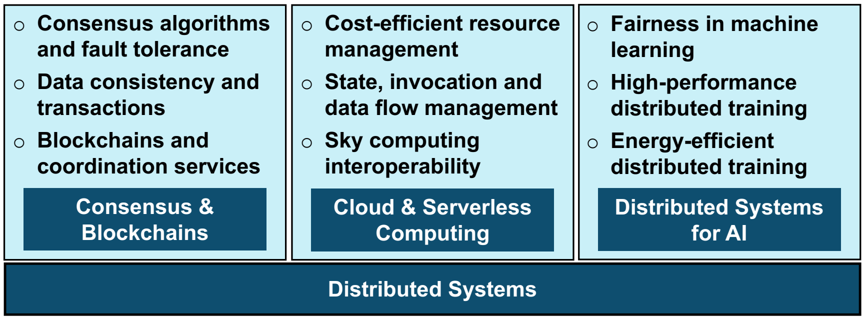
Research Statement

Gengrui (Edward) Zhang
https://gengruizhang.github.io

My current research is focused on the core of distributed systems. Propelled by the exponential increase in computation and storage resources demanded by today's applications, distributed systems have become prevalent, expanding their scales from localized to global-scale deployments. As such, challenges arise in pivotal determinants of system performance, such as coordination, transaction, and replication. For example, how to efficiently manage data consistency in distributed databases of escalating sizes? how to coherently transact on blockchains? And how to synchronize models among thousands of nodes in distributed training systems? Following Lampson's system design principle – "The normal case must be fast; the worst case must make some progress," the design of modern distributed systems begs the following fundamental questions:

1. *How to increase the efficiency of normal operation for higher throughput and lower latency?*
2. *How to scale these systems such that their performance degradation is minimized?*
3. *How to bolster system robustness such that normal operation is least interrupted under a wide range of failures?*

My current research answers these questions in the context of consensus systems, contributing to both practical system design and implementation and the theoretical foundations of distributed computing. My current work centers on three key areas: 1) more scalable and robust consensus algorithms under Byzantine fault tolerance and crash fault tolerance, 2) efficient replication and transactions for distributed databases and blockchains, and 3) versatile blockchain architectures for diverse applications.


○ Consensus algorithms and fault tolerance
○ Data consistency and transactions
○ Blockchains and coordination services
Consensus & Blockchains
○ Cost-efficient resource management
○ State, invocation and data flow management
○ Sky computing interoperability
Cloud & Serverless Computing
○ Fairness in machine learning
○ High-performance distributed training
○ Energy-efficient distributed training
Distributed Systems for AI
Distributed Systems

My long-term research goal is to develop 3H distributed systems – characterized by high performance, high scalability, and high availability. It is focused on designing general and tailored consensus algorithms, fault tolerance, and consistency models to effectively support various distributed applications, including blockchains, databases, cloud computing, and distributed training systems.

Current Research

CR1: Distributed consensus and Byzantine fault tolerance.

Leader-based consensus algorithms utilize designated servers as leaders. SOTA BFT algorithms passively rotate leadership among all servers based on predefined leader schedules. However, blindly rotating leadership makes the system vulnerable, potentially suffering from $\frac{f}{3f+1} \approx 33\%$ faulty leaders in view changes. This vulnerability results in severe performance degradation, especially under frequent leadership rotations.

To address this vulnerability, my research has proposed reputation-based consensus algorithms: `Prosecutor` [1] and `PrestigeBFT` [6]. They translate a server's behavior history into a reputation value, electing reputed servers as leaders. In particular, `Prosecutor` imposes Proof-of-Work computation on suspected faulty servers, suppressing them from becoming new leaders. Moreover, `PrestigeBFT` proposes a reputation mechanism that dynamically discredits misbehaving servers and rewards protocol-obedient servers. The reputation mechanism incentivizes servers to behave correctly, increasing the probability of the election of correct leaders over time.

Impact: My reputation-based BFT algorithms contribute to both the theoretical foundation and implementation of the consensus problem. They extend the traditional state machine replication property to a reputation state, which opens up new directions in the discussion of BFT consensus. Additionally, their implementations obtain high performance and high availability. For example, `PrestigeBFT` achieves 5.4× higher throughput HotStuff and improves system availability under a wide range of failures. While HotStuff struggles to operate at an availability of 37%, `PrestigeBFT` progressively improves its system availability to over 90%.

CR2: Fast distributed replication.

The efficiency of replication has become a crucial performance factor for distributed applications, such as distributed databases and blockchains. Although Paxos and Raft are widely used in these applications, their simple-majority quorum replication is inefficient when applied to large-scale systems, particularly heterogeneous ones. Simple-majority consensus tolerates $f = \lfloor \frac{n-1}{2} \rfloor$ failures in a system of n nodes. At a large scale, however, even though it is improbable that half of the nodes fail simultaneously, the system is bound to await responses from a majority. This issue is exacerbated in heterogeneous systems with nodes of varying configurations, where strong nodes are compelled to wait for slower nodes, resulting in a significant degradation in both throughput and latency.

My research introduces dynamically weighted consensus to more efficiently coordinate agreement among servers. Departing from traditional consensus models relying on majority quorums, weighted consensus dynamically assigns weights to servers based on their responsiveness, preferring fast nodes in the decision-making process. Specifically, `Escape` [2] establishes a weighted voting mechanism that strategically avoids split votes in leader election, ensuring the election of a new leader in a single round of voting. Furthermore, `Cabinet` [8] precisely defines the properties of weighted quorums in replication. It offers a customizable failure threshold t, where $1 \leq t \leq \lfloor \frac{n-1}{2} \rfloor$, along with a tailored weight scheme. It assigns each node a distinct weight and attains system-wide agreement as soon as the $t+1$ nodes with the highest weights have reached an agreement.

Impact: My research on weighted consensus redefines the consensus problem from a practical angle. It creates a logical layer dynamically mapping physical nodes to their weight representations in the consensus process, which presents a more customizable and high-performance solution to consensus applications. For example, `Cabinet` outperforms Raft by 3× to 6× in throughput and latency under YCSB and TPC-C workloads. The performance advantage is sustained under increasing system scales, complex networks, and failures in both homogeneous and heterogeneous clusters, offering a promising high-performance consensus solution.

CR3: Dynamic blockchain systems.

Recently, the widespread deployment of advanced driver-assistance systems (ADAS) in vehicles has prompted concerns from customers, regulators, and lawmakers regarding the allocation of legal responsibility in the event of accidents. To address these concerns, blockchain solutions are being explored as immutable ledgers for recording evidence in Vehicle-to-Everything (V2X) networks. Traditional permissioned blockchains operate in stable networks with a static set of servers. However, V2X networks are highly dynamic and susceptible to frequent disruptions, as vehicles may enter or exit the network at will, which poses a significant challenge for developing robust and scalable blockchain solutions.

My research proposed a novel permissioned blockchain with a new consensus mechanism for V2X networks, namely `V-Guard` [4], which addresses the issue of intermittently connected vehicles. V-Guard establishes a membership management unit that allows transactions to be ordered and committed under different memberships (sets of vehicles). It achieves consensus seamlessly under changing members (e.g., with joining or leaving vehicles) and produces an immutable ledger recording traceable data entries with their corresponding membership profiles. This project is fully open source at https://github.com/vguardbc/vguardbft.

Impact: `V-Guard` is the first blockchain system that achieves consensus with dynamic memberships at high performance. Its peak throughput is 22× higher than HotStuff, 9.5× higher than ResilientDB, and 1.8× higher than Narwhal. This project has filed an international patent [3] and is being used by an industry collaborator. Additionally, `V-Guard`'s architecture can be easily adopted by blockchains operating under unstable networks, such as IoT and supply chain applications.

Future Research

FR1: Software-defined consistency services for distributed systems.

Currently, distributed applications are hard-coded with consistency services based on pre-defined failure assumptions [5]. However, future distributed applications should be equipped with more versatile consistency services that can be defined at an application level to better address various consistency requirements. Building on my previous research in developing consensus systems [1, 2, 6, 4, 7], my future research will explore one-size-fits-all consistency solutions that automate the construction of consistency services with invariants of linearizable, sequential, causal, and FIFO orderings. This work will enable software-defined consistency services, allowing flexible consistency models based on the request at hand. My research will focus on the following aspects:

- Coordination as a utility. My research will construct fine-grained containerized service components, including communication, quorum construction, storage, and cryptography, and allow for multiplexing among different consistency services.
- Software-defined consistency. My research will investigate the possibility of providing customizable consistency guarantees. Upon an invoked consistency request, our system can assemble the services of consistency components. It provides "Legos" of consistency, enabling software-defined consistency services tailored to a wide range of applications.

FR2: Cost-efficient resource management in cloud and sky computing.

Cloud computing is one of the key applications of distributed systems, with resource sharing and management emerging as pivotal factors influencing performance. In the field of resource management, my research has introduced Drone [9], which adaptively configures resource parameters to improve application performance and reduce operational costs facing cloud uncertainties. In the future, my research will continue the exploration of resource management in cloud computing and sky computing.

- Cost-efficient resource management for cloud computing. My future research will continue to explore hybrid scaling strategies of containerized clouds, considering scenarios with diverse workloads, global-scale scalability, and failure recovery.
- Cost-optimization for sky computing. My future research will facilitate the interconnection and interoperability of multicloud in a heterogeneous architecture, where computing and storage services come from different vendors. It will establish cost-aware mechanisms between clouds, quickly and securely exchanging services and enabling seamless data transfer based on user-defined criteria.

FR3: Distributed systems for AI

My future research will develop middleware support for machine learning. The exponential growth in the volume of training data has led to an escalating scale of machine learning systems. This surge is rapidly propelling the deployment of distributed machine learning systems. My future research will explore the fairness and efficiency challenges of distributed training.

- Machine learning fairness. My research will explore the use of blockchains to coordinate the detection of unfair training processes. It will collectively inform the identification and categorization of patterns of bias, including disparities in different groups. When data bias is detected, it will utilize smart contracts to adjust data distribution by incorporating more samples from underrepresented groups. When model bias is detected, the smart contract will adjust the model's weights or architecture to reduce the bias.
- High-performance distributed training. My research will improve the efficiency in distributed training, especially synchronization and communication to minimize latency overhead. It will also integrate traditional fault-tolerant algorithms into distributed training, bolstering robustness against various types of failures and empowering uninterrupted high-performance training.

References

[1] **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*, pages 52–63, 2021.

[2] **Gengrui Zhang** and Hans-Arno Jacobsen. ESCAPE to Precaution Against Leader Failures. In *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*, pages 625–635. IEEE, 2022.

[3] **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).

[4] **Gengrui Zhang**, Yunhao Mao, Shiquan Zhang, Shashank Motepalli, Fei Pan, and Hans-Arno Jacobsen. V-guard: An efficient permissioned blockchain for achieving consensus under dynamic memberships in v2x networks. *arXiv preprint arXiv:2301.06210*, 2023.

[5] **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. *ACM Computing Surveys (CSUR)*, 2023.

[6] **Gengrui Zhang**, Fei Pan, Sofia Tijanic, and Hans-Arno Jacobsen. Prestige BFT: Revolutionizing View Changes in BFT Consensus Algorithms with Reputation Mechanisms. In *In 2024 IEEE 40th International Conference on Data Engineering (ICDE).*, 2024.

[7] **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.

[8] **Gengrui Zhang**, Shiquan Zhang, Michail Bachras, and Hans-Arno Jacobsen. Cabinet: Weighted Consensus Made Fast. In *Under Review*, 2024.

[9] Yuqiu Zhang, Tongkun Zhang, **Gengrui Zhang**, and Hans-Arno Jacobsen. Lifting the Fog of Uncertainties: Dynamic Resource Orchestration for the Containerized Cloud. In *Proceedings of the 2023 ACM Symposium on Cloud Computing (SoCC)*, pages 48–64, 2023.