Abstract

The research described in this dissertation explores three distinct technological domains: programmable networks, network-wide distributed blockchain technologies, and cloud gaming. Nevertheless, these domains may appear distinct, the first two often coexist and interact closely in modern distributed applications, and the last one is an example of a modern distributed application. In the context of enhancing overall performance, it's crucial to recognize that each component can potentially serve as a bottleneck that needs improvement. The primary research question underlying this work is centered on optimizing network-wide distributed applications to minimize latency, reduce resource consumption, and alleviate network overhead. Through tailored solutions for each domain, such as novel mechanisms for state replication in 5G networks, optimization of blockchain internal processes for latency-sensitive applications, and development of efficient algorithms for game engine module placement in distributed gaming environments, this dissertation aims to address these challenges comprehensively. By acknowledging the interconnectedness of these technological landscapes and their impact on performance, this research contributes to a holistic approach to improving network-wide distributed applications.

As the first research domain, we consider modern 5G networks capable of providing ultra-low-latency and highly scalable network services by employing modern networking paradigms, such as Software-Defined Networking (SDN) and Network Function Virtualization (NFV). The latter enables performance-critical network applications to be run in a distributed fashion directly inside the infrastructure. Being distributed, those applications rely on sophisticated state replication algorithms to synchronize states among each other. In such a scenario, we propose STARE, a novel state replication system tailored for 5G networks. At its core, STARE exploits stateful SDN to offload replication-related processes to the data plane, ultimately reducing communication delays and processing overhead for virtual network functions. We provide a detailed description of the STARE architecture

alongside a publicly available P4-based implementation. Furthermore, our evaluation shows that STARE is capable of scaling to large networks while introducing low communication overhead.

Regarding the second research domain, blockchains offer trust and immutability in untrusted environments, but they are typically not fast enough for latency-sensitive applications. Hyperledger Fabric (HF) is a common enterprise-level platform offered as a fast Blockchain-as-a-Service (BaaS) by cloud providers. In HF, every new transaction requires a preliminary endorsement by multiple mutually untrusted organizations, contributing to the delay in storing the transaction in the blockchain. The endorsement policy is specific for each application and defines the required approvals by the Endorser Peers (EPs) of the involved organizations. Given an input endorsement policy, we study the optimal choice to distribute the endorsement requests to the proper EPs. We propose the OPtimal ENdorsement (OPEN) algorithm, devised to minimize the latency due to network delays and the processing times at the EPs. Through extensive simulations, we show that OPEN can reduce the endorsement latency by up to 70% compared to the state-of-the-art solutions and can approximate well the optimal policies while offering a negligible implementation overhead compared to them.

In the third research domain, online gaming has seen a significant surge in popularity, becoming a dominant form of entertainment worldwide. This growth has necessitated the evolution of game servers from centralized to distributed models, leading to the emergence of distributed game engines. These engines allow for the distribution of Game Engine Modules (GEMs) across multiple servers, improving scalability and performance. However, this distribution presents a new challenge: the game engine module placement problem. This problem involves strategically placing GEMs to maximize the number of accepted placement requests while minimizing the delay experienced by players, a critical factor in enhancing the gaming experience. We show that the problem can be formulated as an Integer Linear Programming (ILP) model. It provides an optimal solution but suffers from high computational complexity, making it impractical for real-world applications. To address this challenge, our research introduces two novel heuristic algorithms, MAP-MIND and MAP-MIND*. The MAP-MIND algorithm demonstrates superior performance, achieving near-optimal delay and more than 92% GEM request acceptance in the worst heterogeneous scenarios. The MAP-MIND* algorithm, while slightly under-performing MAP-MIND in terms of delay, proves to be significantly faster, making it a viable alternative for real-world applications with equal GEM request acceptance. The trade-off between the two algorithms offers a flexible approach to GEM placement, balancing performance and computational efficiency.