Duplicated from the abstract in HotCRP

Mobile networks enable ubiquitous and on the move connectivity. At the heart of any mobile network is the mobile core. This essential component handles all control functionality (authenticating devices, managing user sessions, mobility, roaming, billing, etc.) and provides a bridge to the Internet. Compared to the small geographical footprint of a base station, the mobile core serves a much wider area; its responsive and reliable operation is essential.

The mobile core handles both control traffic as well as data plane traffic, the latter originating from user applications (web browsing, video calling, etc..), while the former is used by the mobile core to manage the network and the devices connected to it. Control plane performance is critical to overall user experience: a poorly performing control plane will slow down data plane operations. While the mobile core simply routes data plane traffic, control plane traffic is processed inside the core. There is a real risk that an unexpected burst of control traffic can overwhelm the core. Studies have shown that control plane signalling traffic is not only increasing rapidly but is also highly bursty, due to increased numbers of connected devices, higher density of cell towers, and new classes of IoT devices.

Network operators approach this problem by over-provisioning the core to ensure it can handle traffic bursts. However, this approach is wasteful. Autoscaling is a solution that can lead to more efficient resource usage and lower operational expenditure. Autoscaling varies the amount of resources allocated to the core, depending upon the amount of control plane traffic received.

Despite its advantages, autoscaling is not a straightforward solution. Existing mobile core designs are unable to be autoscaled. We have identified two key challenges that prevent this. Firstly, existing designs closely couple the mobile core to the network of base stations that serve users, preventing separate instances of the core from being spawned and scaling up. Secondly, traditional implementations of the mobile core mix the processing of traffic with the state of connected users, preventing scaling as the state cannot be scaled across multiple instances without losing consistency.

We propose a new design, CoreKube, for a scalable, efficient and resilient mobile core. CoreKube addresses the identified challenges through the use of a novel message-focused design, which features truly stateless instances that interface with a common database and with the base stations through a separate frontend. Our implementation is cloud-native, being orchestrated on Kubernetes and therefore is suitable for deployment on both public and private clouds. We show that compared to state-of-the-art core designs, CoreKube efficiently processes control plane messages, scales dynamically while using minimal compute resources and recovers seamlessly from failures.

CoreKube was both presented and demoed at MobiCom 2023 (Best Artifact Award winner, Best Demo shortlisted). Our implementations are open source. This talk would present the CoreKube design, the challenges and requirements that led to its creation, an evaluation compared to the state-of-the-art alternatives, and a video demo of the working implementation.