Contribution ID: **382** Type: **not specified**

SMC-ERM: A fast remote memory communication method based on SMC socket

Wednesday, 18 September 2024 15:30 (30 minutes)

Shared Memory Communication (SMC) is a high-performance, socket-based stack that operates within kernel space. By leveraging shared memory technology, SMC enhances communication speeds while preserving the TCP socket API for userspace. Consequently, most TCP applications can seamlessly transition from TCP to SMC to achieve better performance without requiring any code modifications.

Recent AI training demands increasingly higher bandwidth, making userspace RDMA widely adopted in AI applications. TCP device memory aims to eliminate memory copying between main memory and GPU memory, and has made considerable progress. However, while SMC-R natively supports RDMA, enabling zero-copy functionality is straightforward for SMC-R, it remains nearly impossible for SMC. This limitation arises due to the need to maintain compatibility with the TCP socket-based API and the constraints of the in-kernel ring buffer used for communication.

To fully unleash the potential of SMC and meet the high bandwidth requirements, we propose a new set of simple APIs built upon the SMC socket API, which we call ERM (Extended Remote Memory). With ERM, users can perform direct read/write operations on remote memories without any memory copying, akin to RDMA, but with much simpler usage. The core benefits of SMC-ERM include:

1. Ease of Use: Socket-based API that reuses the SMC socket for establishing connections, requiring only the extension of datapath APIs.
2. Direct Memory Access: Offers RDMA-like direct memory access with comparable performance.
3. Kernel-Space Management: Device and memory management occur in kernel space, eliminating the need for a large userspace stack to manage RDMA devices.

This talk will introduce the SMC-ERM concept to the community for the first time, covering its design, usage, and performance metrics compared to TCP and RDMA.

Primary authors: LI, Dust (Alibaba Cloud); SHI, Wei (Alibaba Cloud)

Presenter: LI, Dust (Alibaba Cloud)

Session Classification: Networking Track

Track Classification: Networking Track