Popular Science Summary

Casper Vikström, Kesheng Wang

June 2024

The race for computing power that drives the advancements of popular AI and machine learning applications, such as LLMs like ChatGPT, often relies on dedicated hardware. These performance requirements for speed and efficiency emphasize the need for high-speed data transfers between different components. Whether it is hundreds of high-end GPUs performing inference or clusters with multiple dedicated accelerators, high-speed communication is crucial for making everything work efficiently. A popular high-speed serial link choice and the link used and implemented in this paper is a SerDes channel, which is commonly used in Gigabit Ethernet, PCIe, and other data transmission protocols.

High-speed serial data allows for the extension of an existing SoC by incorporating additional SoCs and their resources like larger memory. This enables the storage of weights or data outside the accelerator chip, significantly reducing cost and area, as on-chip memory (such as SRAM) tends to be more expensive and larger in area.

In this thesis an existing SoC structure was modified to include a custom communication protocol over a SerDes channel, allowing read and write operations to be issued on a receiving SoC. This demonstrates the capabilities of a dedicated high-speed SerDes communication channel for resource sharing. The work includes the creation of the physical SerDes channel along with the communication for the Wishbone bus connected to the SoC's CPU.

The proposed implementation proves that resource sharing is possible and promising but some improvements towards the reliability is left for future work.