Pump Up the Volume: Processing Large Data on GPUs with Fast Interconnects

Link : https://dl.acm.org/doi/pdf/10.1145/3318464.3389705

Technical Question: This paper investigates how the fast interconnect such as NVLink 2.0 can solve the data transfer bottleneck between CPUs and GPUs. What are the disadvantages of current interconnect and the benefits of NVLink 2.0 besides speed?

Traditional interconnects, such as the Peripheral Component Interconnect Express (PCIe), have become impediments in GPU-accelerated data processing systems due to their restricted bandwidth capacity and elevated latency. As GPUs have undergone rapid advancements in computational capabilities, the necessity for interconnects with greater bandwidth and reduced latency has increased. NVIDIA's NVLink 2.0 is specifically designed to overcome these constraints and offers several distinct advantages compared to PCIe-based interconnects:

• *Augmented Bandwidth*: NVLink 2.0 provides a substantially increased bandwidth in comparison to PCIe, facilitating rapid data transfers between Central Processing Units (CPUs) and GPUs, as well as among multiple GPUs in parallel. This becomes particularly vital for processing large-scale data sets, where data transfer times can substantially impact overall performance and efficiency.

• *Diminished Latency*: NVLink 2.0 exhibits lower latency than conventional PCIe

Background: In the research paper "Pump Up the Volume: Processing Large Data on GPUs with Fast Interconnects," the primary focus is on the GPU Volume Processing (GVolP) algorithm, which is designed to efficiently handle large-scale data sets on Graphics Processing Units (GPUs). Although the paper does not directly address the drawbacks of existing interconnects or the advantages of NVLink 2.0 beyond speed, it emphasizes the significance of high-performance interconnects for achieving optimal results in GPU-based data processing systems.

GVolP is designed to take full advantage of the parallel processing capabilities and high memory bandwidth of GPUs, enabling efficient processing of large-scale data in main-memory database systems. The paper also investigates the impact of fast interconnects, such as NVIDIA's NVLink, on the performance of the GVolP algorithm.

The key contributions of the paper include the introduction of GVolP, an evaluation of the impact of high-speed interconnects on GVolP's performance, and a detailed experimental evaluation that compares GVolP's performance with CPU-based query processing

algorithms and other GPU-based approaches. The results demonstrate that GVolP significantly outperforms traditional CPU-based methods and other GPU-based techniques, particularly when handling large-scale data sets.