Feb 5, 2019
Rohan Hulsure
CSCI 5593, REPORT 1

Report on "An Architecture for Integrated Near-Data Processors"

In recent times, data has been playing a major role in computer systems. There has been a substantial growth in production and storage of digital media. With growing data existing computer systems are getting worse at data handling. Increased memory bandwidth with reduced memory latency does not help to keep up with increased processing capabilities thus having degraded performance by compute nodes at handling data-intensive workloads. Scaled up hardware can be solution but even these suffer from high latency and relatively small storage volume. To bridge this gap the processing should take place as close to the data as possible avoiding interconnecting bottlenecks. This is also referred as near data processing (NDP). The author put forth an extension to an existing CPU architecture that enables near data processing capabilities close to main memory eventually increasing the performance of data intensive applications.

The architecture consist of two level memory controller: at CPU level we have memory-technology agnostic memory controllers and with this technology specific memory controllers tightly coupled to main memory. A lightweight NDP messaging extension added to the high speed serial link of memory channel which carries both original traffic as well as communication between manager and access point of NDP. A NDP manager is introduced for handling virtual memory coherence, remote accesses and communication capabilities for NDP in memory system. An access point (NDP-AP) as a single hardware component attached to the system bus to reduce software overhead and CPU overload. NDP manager access NDP-AP regarding all its needs. Hardware managed extended coherence solution is also introduced to enforce coherence between CPU, NDP and main memory. All the virtual address and address translation mechanism are implemented at NDP-M

The proposed architecture ensures that workloads can continue working seamlessly as expected lowering the impact of CPU limitations. The benchmarks shows that proposed architecture provides excellent performance. This extended architecture can be used as a base for future research in how near data processing can be more tightly coupled with CPUs and other processing devices

Reference:

[1] Erik Vermij, Leandro Fiorin, Rik Jongerius, Christoph Hagleitner, Jan Van Lunteren, Koen Bertels, "An Architecture for Integrated Near-Data Processors", Journal ACM Transactions on Architecture and Code Optimization (TACO), Volume 14 Issue 3, September 2017, Article No. 30

URL: https://dl-acm-org.aurarialibrary.idm.oclc.org/citation.cfm?id=3127069