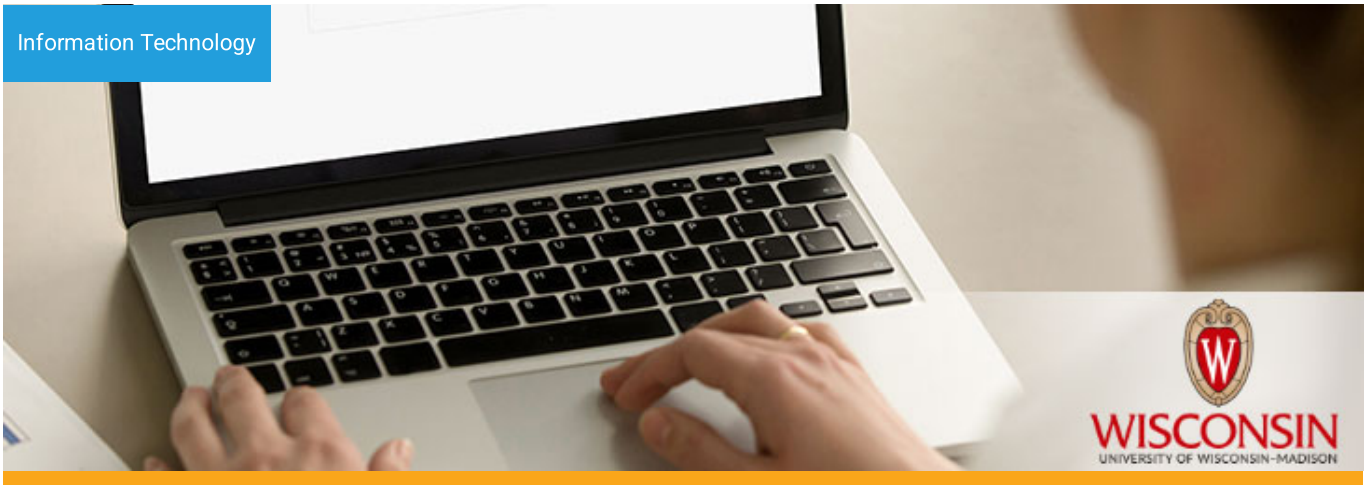
Information Technology



Improved GPU Performance by Memory-Link Compression

View U.S. Patent No. 9,189,394 in PDF format.

WARF: P120224US01

Inventors: Nam Sung Kim

The Wisconsin Alumni Research Foundation (WARF) is seeking commercial partners interested in developing a method for high-speed data transfer between graphic processing units and their off-chip memories.

Overview

Graphic processing units (GPUs) are specialized for graphics but also work with conventional computer processing units (CPUs) to accelerate different applications. In normal operation, the CPU loads data and instructions into GPU memory, which executes the task and returns the data.

Problems of long latency – waiting for the GPU's numerous computational elements to access its off-chip memory – can be accommodated by context switching. This method flexibly switches to different threads when a given thread faces a memory access delay. In many important memory-bound applications, however, context switching may still be too slow. This is especially true when GPUs are used for general-purpose computation where the rate of instructions increases.

The Invention

A UW–Madison researcher has developed a GPU design for faster data transfer by compressing and decompressing data passed between the units and their memories.

The computational elements of the GPU are adapted to receive, execute and output data through connected memory channels. A compressor/decompressor associated with each channel prepares the data for reading and storage.

Applications

- Hardware microarchitecture and compression software

Key Benefits

- Increases effective bandwidth of memory channels
- Faster processing with less power
- Readily implemented on GPU hardware
- Avoids circuitry overhead and delays
- Provides lossy compression and decompression
- Simplifies data handling

Stage of Development

The lossless and lossy compression techniques have been demonstrated to improve performance of memory-bound workloads by 26 percent and 41 percent on average.

Additional Information

Related Technologies

- WARF reference number P110254US01 describes a low voltage operating cache structure that works to maximize energy efficiency in processors.

Publications

- Sathish V., Schulte M. and Kim N.S. 2012. Lossless and Lossy Memory-link Compression Techniques for Improving Performance of Memory-bound GPGPU Workloads. IEEE/ACM Int. Conf. on Parallel Architecture and Compilation Techniques (PACT)

Tech Fields

- Information Technology : Computing methods, software & machine learning

For current licensing status, please contact Jeanine Burmania at jeanine@warf.org or 608-960-9846