Information Technology

Matrix Processor with Localized Memory to Increase Data Throughput

View U.S. Patent Application Publication No. US-2018-0113840 in PDF format.

WARF: P160414US01

Inventors: Jing Li, Jialiang Zhang

The Wisconsin Alumni Research Foundation (WARF) is seeking commercial partners interested in developing innovative computer architecture for high-speed matrix operations.

The new hardware technique reduces the memory bottleneck between external and local memory for matrix-type calculations and allows increased throughput up to fivefold.

Overview

Matrix calculations such as matrix multiplication are foundational to a wide range of emerging computer applications such as machine learning and image processing. Matrix calculations cannot be fully exploited by a conventional general-purpose processor, so there is interest in developing a specialized matrix accelerator, for example, using field programmable gate arrays (FPGAs) to perform matrix calculations. In such designs, different processing elements of the FPGA could simultaneously process different matrix elements using portions of the matrix loaded into local memory associated with each processing element.

UW–Madison researchers led by Prof. Jing Li have recognized that there is a severe memory 'bottleneck' in the transfer of matrix data between external memory and the local memory of FPGA-type architectures. This bottleneck results from both the limited size of local memory compared to the computing resources of the FPGA-type architecture and from delays inherent in repeated transfer of data from external memory to local memory. Exacerbating this problem, computational resources are growing much faster than local memory resources.

The Invention

To address this challenge, Li's team has developed computer architecture combining local memory elements and processing elements on a single integrated circuit substrate. In this way, data stored in a given local memory resource normally associated with a given processing unit is shared among multiple processing units. The sharing may be in a pattern following the logical interrelationship of a matrix calculation (e.g., along rows and columns in one or more dimensions of the matrix).

Sharing reduces memory replication (the need to store a given value in multiple local memory locations), thus reducing the need for local memory and unnecessary transfers of data between local memory and external memory. This permits high speed processing on local memories (on-chip) and reduces energy consumption associated with a calculation.

Applications

- Hardware support for efficient data sharing in vector processing

Key Benefits

- Allows increase in throughput by at least 2-5 times
- More efficient data sharing between compute and memory resources on chip
- Unlike existing techniques, no performance tradeoff
- Readily implemented on an FPGA-type device

- Scalable

Stage of Development

The researchers have reconfigured a state-of-the-art FPGA prototyping platform to test their concept. The technology outperformed the best known method for performing matrix multiplication (the source of bottlenecks).

The development of this technology was supported by WARF Accelerator. WARF Accelerator selects WARF's most commercially promising technologies and provides expert assistance and funding to enable achievement of commercially significant milestones. WARF believes that these technologies are especially attractive opportunities for licensing.

Additional Information

Related Technologies

- WARF reference number P150232US01 describes innovative hardware that blends compute and storage capabilities for increased efficiency.

Tech Fields

- Information Technology : Hardware

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