Parallel Patterns using Heterogeneous Computing

Mr Andreas Vermeulen
-
University of St Andrews
Saint Andrews, Fife KY16 9AJ
-
University of Dundee
Nethergate,Dundee DD1 4HN
-
a.f.vermeulen@dundee.ac.uk

Dr Vladimir Janjic
-
University of St Andrews
Saint Andrews, Fife KY16 9AJ
-
vj32@st-andrews.ac.uk

Mr Andy Cobley
-
University of Dundee
Nethergate, Dundee DD1 4HN
-
acobley@computing.dundee.ac.uk

ABSTRACT

First report on the joint research between University of St Andrews and University of Dundee to formulate an enhanced version of the ***Research Information Factory*** *using* ***Heterogeneous Computing*** *and* ***Parallel Patterns*** *against a* ***Cassandara*** *and* ***Spark*** *data lake.*

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Theory, Framework, Application, Research,Hardware,3D Torus Grid

Keywords

research, research information factory, RIF, RIFF, RIFC ,heterogeneous computing, parallel patterns, data lake, cassandra, spark, opencl, fastflow, cuda, 3D torus

1. INTRODUCTION

The landscape for processing data from raw research data into actionable research knowledge is exceeding the limits of previously accepted processing methodologies and processing capacity of new computer equipment is creating new opportunities for procesing data quickly and cost efficiently. Larger and faster central processing units are nolonger the answer to the enhancement in processing throughput, the future is heterogeneous systems and design patterns to utilise these systems. The joint research between St Andrews and Dundee investigates the area of parallel patterns using heterogeneous systems that officiallly started March 2015. The fundamental research covers three stages: Fundamental research on heterogeneous systems, fundamental research on parallel patterns for data processing analysis and a practical implementation into an enhanced version of the Research Information Factory.

2. RESEARCH INFORMATION FACTORY

The Research Information Factory (*RIF*) is a processing appliance consisting of a framework and a cluster that supports the conversion of raw research data into knowledge using parallel patterns.

2.1 Research Information Factory Framework

The Research Information Factory Framework (*RIFF*) is a data processing framework that was designed during the period 2006 to 2011 and officially published as part of a MSc in Business Intelligence project (2012) and adapted during a Pg Cert in Data Science (2013) [15] to support unstructured and structured data patterns. The framework contains a set of guidelines to process raw data into knowledge. Framework uses a five layer process Research Layer (spesific research requirements), Utility Layer (common processing utilities), Audit, Balance and Control Layer (schedule jobs, collect audits, control patterns) Operational Management Layer (active processing controls) Functional Layer supports the core processing pattern of **R**etrieve-**A**ssess-**P**rocess-**T**ransform-**O**rganise-**R**eport (*R-A-P-T-O-R*). The planned research will enhance the framework with the required parallel patterns to generate the rules from fundamental principals. A version 2.0 of the framework already exist and the new research will develop version 3.x.

2.2 Research Information Factory Cluster

The Research Information Factory Cluster (*RIFC*) is a cluster appliance using commondity computer equipment to process the RIFF parallel patterns. This new custom parallel cluster appliance design [15] (Synaptic Assimilator) is part of the results of the joint research.

3. HETEROGENEOUS COMPUTING

Heterogeneous computing systems [11] uses more than one type of processors. Systems gain performance enhancements by ability to utilise dissimilar processors to execute common processing requirements. The research we are conducting is investigating using central processing unit (CPU), graphical processing unit (GPU) and field-programmable gate array ($FPGA$) processor to perform the required processing from the RIFF. This research is the fundamental discovery process of what parallel patterns requires what combination of Heterogeneous Computing. Started with a prototype system using CPU and GPU combinations to build the core building blocks in the form of cluster processing cells and related processing patterns. The central processing unit (CPU) is designed with few cores optimised for sequential serial processing. The CPU will perform patterns that require requires large amounts of control changes. The graphical processing unit (GPU) is designed as a massively parallel architecture of thousands of smaller, efficient cores designed for handling multiple tasks simultaneously. The spesilised Application-Specific Integrated Circuit ($ASIC$) solution is capable of performing fixed spesific tasks like network connectivity, mapping information. Field-Programmable Gate Array ($FPGA$) is a set of programmable logic blocks and programmable interconnects allow the same FPGA to be used in many different applications. The Field-Programmable Gate Array ($FPGA$) is designed to use hardware description language (HDL) to dynamically setup the logic flow processing patterns. The research investigates the challenges and opportunities [1] to find the process patterns that achieves the set goals of the research.

4. PARALLEL PATTERNS

Parallel patterns is the fundimental building blocks of any data processing requirement. The research is into how heterogeneous computing changes the design and implimentation of common parallel patterns (task parallelism, pipelines, recursive splitting and geometric decomposition of data processing). Researching common strategies (actors, shared queue, fork/join, loop parallelism and master/worker) processing using following libraries (CUDA [7], OpenCL[5][8][9][13], FastFlow [6] and ZeroMQ [10]).

5. EFFICIENCY AND ENERGY-AWARENESS

The increasing drive for processing power demand increase energy levels to perform tasks. Our research targets efficiency in terms of processing time, programming effort, energy-awareness for each design pattern. Research calculates efficiency of processing in Floating-point Operations Per Second per Watt ($FLOP/S/W$). We investigate systems like nVidia Jetson TK1 [4] and Tilera TILE-Mx100 processor [14].

6. DATA LAKE

A massive, easily accessible data repository using commodity computer hardware to store data with four dimensions: volume, variety, velocity and veracity. This requirement generates data sources of a size larger than what a single system can handle. These systems requires parallel patterns with heterogeneous computing clusters to assess the data sources with efficiency.

6.1 Cassandra Data Processing Engine

Cassandra database [2] is proven capasity for scalability and high availability data processing. Linear scalability and fault-tolerance on commodity hardware supports massively scalability and high processing rate required for the parallel patterns.

6.2 Spark Processing Engine

Spark Engine [12] is a fast and generalised processing of large-scale data processing using an advanced Directed Acyclic Graph (DAG) execution engine that supports cyclic data flow and in-memory computing. This matches the parallel patterns requirements for the research process. The RIFC will currently support three node cluster for handling spark based design patterns applied to a Cassandra cluster [3].

7. RESEARCH METHODOLOGY

The research into the fundamental behavior of a selection of heterogeneous computing components forming a cluster cell i.e 4/8 core CPU with 172/384/768 core GPU. The behavior is evaluated on a basis of a fixed size data set processed via a spesific parallel pattern changing parameters like combinations of CPU, GPU with different clock speeds, memory allocations and measuring time to complete task, energy requirements and effort in amount of code required for each parallel pattern and heterogeneous computing combination.

8. CONCLUSION

The research is in early stages and we need to stabilise the experiment parameters and determine the true scope of the final research question. On achievement of a stable experiment setup we will report back details on final findings and progress to final publication.

9. REFERENCES

[1] Ashfaq A. Khokhar, Viktor K. Prasanna,Muhammad E. Shaaban, Cho-Li Wang, (1993); Heterogeneous Computing: Challenges and Opportunities.

[2] Cassandra, (2014), Apache Cassandra; http://cassandra.apache.org/

[3] Datastax, (2014), Getting Started with Apache Spark and Cassandra; http://planetcassandra.org/getting-started-with-apache-spark-and-cassandra/

[4] NVIDIA Jetson TK1, (2014), NVIDIA Jetson TK1 development kit.

[5] Khronos, (2014), OpenCL; https://www.khronos.org/opencl/

[6] M. Aldinucci, M. Danelutto, P. Kilpatrick, and M. Torquati, (2011), FastFlow: high-level and efficient streaming on multi-core in Programming Multi-core and Many-core Computing System

[7] nVidia, (2014), CUDA Tool Kit.

[8] OpenCL 1.2, (2011), OpenCL 1.2 Reference Card.

[9] OpenCL 2.0, (2013), OpenCL 2.0 Reference Card.

[10] Pieter Hintjens (2014), ÃŸMQ - The Guide.

[11] Qiang Wu,Yajun Ha ; Kumar, A. ; Shaobo Luo ; Ang Li ; Mohamed, S., (2014), A heterogeneous platform with GPU and FPGA for power efficient high performance computing.

[12] Spark, (2014), Spark Lightning-fast cluster computing.

[13] Stone, J.E. ; Gohara, D. ; Guochun Shi,(2010), OpenCL: A Parallel Programming Standard for Heterogeneous Computing Systems, Comput. Sci. Eng. 12, 66 (2010).

[14] Tilera, (2007), Tilera Processor Family.

[15] Yuichiro Ajima, Shinji Sumimoto, Toshiyuki Shimizu, (2009), ”Tofu: A 6D Mesh/Torus Interconnect for Exascale Computers,” IEEE Computer, vol. 42, no. 11, November 2009, pp. 36-40.