ISSN **2 3 2 1 - 8 0 7 X**
Volume 12 Number13
Journal of Advances in chemistry

BIG DATA ANALYTICS METHODS USING GPU : A COMPREHENSIVE SURVEY

Dr. R. Kingsy Grace[1], S. Manju[2]
[1]Associate Professor, [2]Assistant Professor
kingsydhasphd@gmail.com
[1,2]Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu, India

ABSTRACT

Big data analytics is eventual discovery of knowledge from large set of data thus leading to business benefits. Its biggest challenge is the ability to provide information within reasonable time. The traditional analytics methods might fail to produce efficient result when data handled is of large size. As part of enhancing the performance, the researchers incorporated Graphical Processing Unit (GPU) on big data. GPU being the soul of computer delivers high performance by using its multi core parallel architecture. This paper investigates some methods of integrating GPU on analytics of big data that solely delivered high performance when compared to conventional schemes.

Keywords: GPU, Big data analytics, HPC.

1. INTRODUCTION

The exponential growth of information in every field had led to the explosion of big data. It is the digital evolution come revolution which has almost laid its part in every business life [1]. As result of this evolution the volume of data is growing enormously large, that at each second numerous amount of data is created and shared on internet. The data in this digital universe had already grown from Giga bytes to zetta bytes. But not all the existing data are meaningful and clear as the way it goes “we starve for knowledge while we drown with data”. Hence analytics is the key technique to be followed to extract meaningful information as form of patterns, knowledge, business trends and preferences. As a result to extract knowledge, more remedial data analyzing methods like data reduction, Principle Component Analysis, Sampling, incremental learning, Support Vector Machine (SVM)etc., were introduced. Principle Component Analysis [2] and sampling [3] methods aimed at reducing the data volume to enhance data analysis.These process could be grouped under major category of supervised or unsupervised learning which were found unsuitable for current data volume [4]. In effect to this inability more parallel processing concepts were also introduced like specialized SVM. This attempt to improve computational speed of big data was not satisfactory in effect to the distributed computational requirements [5]. In view of this GPU computing is used with multi core parallel architecture. GPU was initially used for graphical computations rendering images. At later stage, they are used for general purpose computations which has more number of cores. Each of the core has more arithmetic and logical unit, functional units to speed up the performance. Each functional unit is used to process a thread of execution achieving parallelism thereby accounting to High Performance Computing (HPC).GPU differs from CPU by the way how all transistors are used. In CPU most of them are used for caching data while in GPU all of them are used for processing. Some of the GPU systems that deliver high performance are NVIDIA GPU, AMD. The frameworks for programming GPU are OpenCL, Compute Unified Device Architecture (CUDA). The rest of the paper is organized as follows. Section 2 deals with the big data analytics methods in GPU. Section 3 presents the summary of all the discussed methods ad section 4 concludes the paper.

2. BIG DATA ANALYTICS METHODS IN GPU

This section deals with big data analytics methods using GPU and its supported platforms governing towards HPC. This paper also provides a comparison of traditional analytics method, benchmarks and algorithms with big data analytics methods.

2.1 Big Data Analytics on Cloud using GPU

The solution to handle fastest growing data was provided in form of platforms or infrastructures such as i) Hadoop platform [9] using Nvidia CUDA architecture [10] for processing or computation, ii) Titan using HDFS for distributed and massive storage. This kind of advanced infrastructure alone cannot improve performance while the way of data analysis i.e. big data analytic methodologies should also be focused to accountable performance elevation. The one such data analytic technique was enforced on cloud to scale up performance level [1]. The online storage technologies are well preferred for this technologies to fulfill larger storage and computing requirements. With this handling big data within rational time was not an unachievable task. Cloud services also laid way for parallel and distributed computations where no tasks are loaded on a single machine accounting to a fault tolerant and reliable system.

2.2 Message Passing Interface (MPI) /OpenMP on Beowulf

Jorge et al. have proposed a parallelization technology tested on high performance oriented infrastructure by exploiting multi core architectures. MPI is the defacto language independent communication standard for parallel computing. MPI enables programs to run on distributed memory systems. MPI concentrates more towards HPC while it doesn’t support multiprocessing programming [6]. To make it possible, OpenMP an application programming interface was introduced that supports multiprocessing job on shared memory. This multicore architecture was tested on beowulf cluster

by running two supervised machine learning algorithms on google cloud platform. The data are stored as equal parts in all the machines ensuring distributed storage. MPI resulted in consistent performance and speedy processing.

2.3 Spark on Hadoop

Jorge et al. have also evaluated the previously discussed machine learning algorithms by integrating spark on hadoop platform by using the same google cloud service [6]. On comparison spark was found lagging in computation speed with MIP. But spark on hadoop was found better in data management infrastructure and in fault tolerance. As well it allows dynamic addition of nodes to existing virtual machines.

2.4 Slab Lower Upper Matrix Decomposition (LUD) code on Sandy Bridge, PHI Processor and Tesla

To assess the big data system configuration Rao et al. have proposed an idea of evaluating LUD on Intel many integrated core Phi processors, sandy bridge and on tesla [7]. The LUD software package was the input which was developed by Lockheed Martin Corporation. These packages was tested under the clusters Bluegrit, Bluewave and NASA-GFSC's discover cluster systems. The performance valuation was done on three different systems namely,

i) Linux cluster with IBM iDataplex having many integrated phi coprocessors,
ii) Bluegrit cluster as a distributed memory on Intel Xeon x5670 integrated with Nvidia tesla M2070 GP.
iii) Bluewave cluster on quad core Intel(R) Xeon(R).

Rao et al. have presented a performance comparison of slab LUD code on Nehalem, Westmere, Sandy Bridge, NVidia Tesla M2070 GPU and Phi processors. The basic configuration is shown in Table 1. The results shows that Sandy Bridge outperforms others processors such as Nehalem, Westmere, NVidia Tesla M2070 GPU and Phi processors.

Table 1 : Configuration of Rao et al. method

	Cluster	Processor & components used	Interconnect	Hosted at	Result
Case 1	Linux	an IBM iDataplex with 480Intel Many Integrated Core (Phi) co-processors hosted on 2 Hex-core 2.8 GHz Intel Xeon Westmere Processor	Dual Data Rate Infiniband (DDR)	NASA Center for Computational Sciences (NCCS)	4 times faster than Bluegrit
Case 2	Bluegrit	A single Intel(R) Xeon(R) CPU E5504 (Nehalem) running at 2.00GHz, a single Intel(R) Xeon(R) CPU X5670 (Westmere), NVIDIA Tsla M2070 GPU	Lack of Infiband	University of Maryland, Baltimore County (UMBC)	Faster than discover GPU
Case 3	Bluewave	160 IBM iDataplex compute nodes, with 2 quad core Intel(R) Xeon(R) Nehalem X5560 running at 2.80GHz	DDR	UMBC	Nehalam processor on Bluewave was 30% faster than bluegrit

2.5 GPU Map Reduce (GPMR)

Stuart et al. have proposed a GPU based map reduce library for large scale computing [8]. GPMR was designed to handle data movement, data management and GPU access challenges. GPMR being a standalone model targets on modifying MapReduce by including batching maps and reduces via chunking to obtain better utilization. Additionally adding accumulation to map substage, adding a partial reduction substage was also done. The library was tested on the benchmarks like Matrix Multiplication (MM, to multiply two large square matrices), Sparse Integer Occurrence (SIO, to count the number of times each integer had appeared in a large dataset); Word Occurrence (WO, to count the number of times each word had occurred in a text corpus); Linear Regression (LR, to compute a linear model of a set of data), and K-Means Clustering (KMC, to partition a set of data points into clusters). The result of implementing GPMR on various benchmarks is listed in Table 2.

Table 2 : Results of GPMR on various Benchmarks

Benchmarks	Results of GPMR
MM	Faster,Scalable, Efficient use of GPU was observed in out-of-core implementation of MM that uses the GPU much more efficiently.
SIO	No much difference, Yield was as same as CPU
WO	Achieved reduced reduction time by assigning a key to warp
LR	Speed up in order of magnitude when compared to traditional CPU method
KMC	Poor when compared to CPU thus leading because of intermediate key value pairs and loading of threads to its own point

3. SUMMARY

The comparison on different big data analytics methods in the literature using GPU are shown in the Table 3.

Table 3 : Comparison on Different Big data Analytics Methods on GPU

Big Data Analytics Methods	GPU and Frameworks used	Algorithms used	Datasets used	Performance
Big data analytics on cloud using GPU	DOT, CUDA	BIRCH,DBSCAN,GPU based SVM, RKM,TKM, etc.,	Not Mentioned	Machine Learning algorithms perform better than other data mining algorithms.
MPI/ OpenMP on Beowulf	multi-core clusters architectures such as Beowulf	K-Nearest Neighbors (KNN) and Pegasos SVM	HIGGS Data Set from the UCI Machine Learning Repository.	Consistent performance and Speedy processing.
Spark on Hadoop	multi-core clusters architectures such as Beowulf	Spark KNN and Spark Pegasos SVM	HIGGS Data Set from the UCI Machine Learning Repository.	Better in data management infrastructure and in fault tolerance
Slab Lower Upper Matrix Decomposition (LUD) code	Sandy Bridge, phi processor and Tesla	LUD Code	Not Mentioned	Bluegrit was Faster than discover GPU. Nehalam processor on Blueware was 30% faster than bluegrit.
GPU Map Reduce (GPMR)	GPU based map reduce library	MM,SIO,WO,LR,KMC	Not Mentioned	MM was Faster, Scalable and used GPU much more efficiently.

4. CONCLUSION

In this scientific era, most of the data are digital and transferred through internet. The growth of the data is found to be exponential. To extract knowledge from these data there is a need for efficient data analytics techniques and frameworks. This paper has presented some of the big data analytics techniques using GPU in the literature. The GPU architecture is used for big data analytics to improve performance. A detailed comparison on the existing big data analytics methods in GPU is also elaborated.

REFERENCES

1. ChunWei Tsai, ChinFeng Lai, HanChieh Chao and Athanasios V.Vasilakos, Big Data Analytics : A Survey, 2015. Journal of Big Data, Vol.2, No.21, 1-32.
2. Ding C, He X., 2004. K-means Clustering via Principal Component Analysis, In Proceedings of the Twenty-first International Conference on Machine Learning, 1–9.
3. Kollios G, Gunopulos D, Koudas N, Berchtold S., 2003. Efficient Biased Sampling for Approximate Clustering and Outlier Detection in Large Data Sets, IEEE Trans Knowl Data Eng., Vol.15, No.5, 1170–1187.
4. X. Wu, X. Zhu, G. Q. Wu, and W. Ding., 2014. Data Mining with Big Data, IEEE Transactions on Knowledge and Data Engineering, Vol.26, No.1, 97–107.
5. Y. You, S. L. Song, H. Fu, A. Marquez, M. M. Dehnavi, K. Barker, K. W. Cameron, A. P. Randles, and G. Yang. Mic-svm: 2014. Designing a Highly Efficient Support Vector Machine for Advanced Modern Multi-Core and Many-Core Architectures, In IEEE International Parallel and Distributed Processing Symposium, 809–818.
6. Jorge L.Reyes-Ortiz,Luca Oneto, Davide Anguita, 2015. Big Data Analytics in the Cloud: Spark on Hadoop vs MPI / OpenMP on Beowulf, In Proceedings of INNS Conference on big data, Vol.53, 121-130.
7. RaghavendraShruti Rao, Dr. Milton Halem and Dr. John Dorband, 2015. Big Data Analytics Performance for Large Out-of- Core Matrix Solvers on Advanced Hybrid Architectures, In Proceedings of International Conference on Computational Science, Vol.51, 2774-2778.
8. Jeff A. Stuart, John D. Owens, 2011. Multi-GPU Map Reduce on GPU Clusters, In Proceedings of the IEEE International Parallel & Distributed Processing Symposium, 1068-1079.
9. Apache Hadoop, February 2, 2015. [Online]. Available: http://hadoop.apache.org.
10. CUDA, February 2, 2015. [Online]. Available: http://www.nvidia.com/object/ cuda_home_new.html.

Dr. R. Kingsy Grace is currently working as Associate Professor in the Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, Coimbatore, India. She has completed her Ph.D under Anna University. She is the alumnus of Noorul Islam College of Engineering, India and Karunya Institute of Technology, Coimbatore, India. Her area of interest includes Grid Computing, High Performance Computing and Sentiment Analysis.

S.Manju is currently working as Assistant Professor in Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, Coimbatore, India. She has completed her Master of Engineering in Computer Science and Engineering from Sri Krishna College of Engineering and Technology in 2016 and Bachelor of Engineering in Computer Science and Engineering from Dr.Mahalingam College of Engineering and Technology in 2014. Her area of interests includes Wireless Networking and High Performance Computing.