INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

IN-MEMORY BIG DATA MANAGEMENT

VRUSHALI P. BONDE[1], PROF. MAYUR S. BURANGE[2]

1. M.E. (CSE) 2ND Sem, PRPCOET, Amravati, India.
2. PRPCOET, Amravati, India.

Accepted Date: 15/03/2016; Published Date: 01/05/2016

Abstract: In this innovative world, organizations like amazon, Google, Facebook etc. are facing tremendous increase in data. This leads to the problem of storing, analyzing processing and managing terabytes or petabytes of data. The requirements of storing this huge amount of data doesn't fulfilled by On-disks database management. The need of new and speedy data processing system replaced by In-memory data management. In-memory Big Data management is capable to process data faster. There are various technologies to process data in challenging way like Memory hierarchy, Non uniform memory access (NUMA), Non-volatile random access memory (NVRAM) etc. We also provide some key factors like simple scalable streaming system and piccolo that need to be considered in order to achieve efficient In-memory data management and processing.

Keywords: In-memory systems, Nonuniform memory access (NUMA), Nonvolatile random access memory (NVRAM), simple scalable streaming system (S4), Piccolo

Corresponding Author: MS. VRUSHALI P. BONDE

Access Online On:

www.ijpret.com

How to Cite This Article:

Vrushali P. Bonde, IJPRET, 2016; Volume 4 (9): 294-299

PAPER-QR CODE

INTRODUCTION

The explosion of Big Data has prompted much research to develop systems to support ultra-low latency service and real-time data analytics. The memory system is one of the most critical components of modern computers. It has attained a high level of complexity due to the many layers involved in memory hierarchy, application software, operating system, cache, main memory and disk. Existing disk-based systems can no longer response due to the high access latency to hard disks. Jim Gray's insight that "Memory is the new disk, disk is the new tape" is becoming true today we are witnessing a trend where memory will eventually replace disk and the role of disks

HDD	NVM	DRAM
Data management system for Relational Data ➢ Asterix ➢ MySql ➢ Oracle	Clustrix FAWN SICT	Data management system for Relational Data ➢ H-Store ➢ Hyper
Generic Data processing system ➢ Hadoop ➢ Hyrack		Generic Data processing system ➢ Spark ➢ Piccolo
HDD Based Big Data storage system ➢ LogBase ➢ Hadoop HDFS		Memory Based Big Data storage system ➢ MongoDB ➢ RAM Cloud

Table.1. Landscape of Disk And main memory system[1].

Database systems have been evolving over the last few decades, mainly driven by advances in hardware, availability of a large amount of data, collection of data at an unprecedented rate, emerging applications and so on[1]. Table. 1 shows state-of-the-art commercial and academic systems for disk-based and in-memory operations. It is reasonable to assume that the entire database fits in main memory? Yes, for some application. In some cases, the database is of limited size [7].

I. LITERATURE SURVEY

Growing main memory capacity has fueled the development of in-memory big data management and processing. Jim Gray's insight that "Memory is the new disk, disk is the new tape" is becoming true today[1]. We are witnessing a trend where memory will eventually replace disk and the role of disks must inevitably become more archival. In-memory database systems have been studied in the past, as early as the 1980s. However, recent advances in hardware technology have invalidated many of the earlier works and re-generated interests in hosting the whole database in memory in order to provide faster accesses and real-time analytics. Most commercial database vendors have recently introduced in-memory database processing to support large-scale applications completely in memory. With the increasing demand of real time data processing, traditional (on-disk) database management systems are in tremendous pressure to improve the performance. With the increasing amount of data, which is expected to touch 40ZB (1ZB = 1 billion terabytes) by 2020, means 5247 GB of data per person, and with traditional DBMS architecture, it is becoming more and more challenging to process the data and to produce analytical results in almost real time[10]. Non-uniform memory access is architecture of the main memory subsystem where the latency of a memory operation depends on the relative location of the processor that is performing memory operations.

II. TECHNOLOGIES FOR IN MEMORY SYSTEM

This is the techniques and concepts which is important for efficient In-Memory database system including memory hierarchy, Non uniform memory access (NUMA), transactional memory and Non virtual random access memory (NVRAM).

A. Memory Hierarchy

Memory system is a hierarchy of storage devices with different capacities, costs, and access times. CPU registers hold the most frequently used data [8]. The memory hierarchy is defined in terms of access latency and the logical distance to the CPU. It consist of Register, caches, Main

memory. In modern architecture, data can't processed until it is not stored in register. Performance of data program highly depend on utilization of memory hierarchy. Good temporal locality and spatial locality is usually depend on efficiency optimization.

B. Non uniform memory access

Non-uniform memory access (NUMA) is an architecture of the main memory subsystem where the latency of a memory operation depends on the relative location of the processor that is performing memory operations. Non-Uniform Memory Access machine, performance depends heavily on the extent to which data reside close to the processes that use them[2]. The reason for employing NUMA architecture is to improve the main memory bandwidth and total memory size.

Data partitioning is one of the technique in database is used to minimize data transfers across different data domain, within compute node and across compute node. Data shuffling in NUMA systems aims to transfer the data across the NUMA domains as efficiently as possible, by saturating the transfer bandwidth of the NUMA interconnect network [1].

C. Transactional memory

Transactional memory is a concurrency control mechanism for shared memory access, which is analogous to atomic database transactions [1]. There are mainly two types i. e. software transactional memory and Hardware transactional memory. STM has limited practical application. HTM is efficiently useful in atomic operations/ transactions.

D. Nonvolatile random access memory

Nonvolatile random access memory (NVRAM) is a category of Random access memory that retains stored data even if the power is switched off. NVRAM can be architected as the main memory in general-purpose systems with well-designed architecture [1]. The effectively advantages of NVRAM is to provides excellent performance when compared to other nonvolatile memory products and less power is required for NVRAM so the backup guarantee can be ensured for up to 10 years[9].

Advanced NVRAM technologies, such as phase change memory, Spin-Transfer Torque Magnetic RAM, and Memristors, can provide orders of magnitude better performance than either conventional hard disk or flash memory[8]. Phase Change Memory (PCM) devices offer more density relative to DRAM, and can help increase main memory capacity of future systems while remaining within the cost and power constraints [5].

III. IN-MEMORY DATA PROCESSING SYSTEM

In–Memory data processing is more important in Big data to process large data in less amount of time. There are two types of data processing system such as Batch processing system (Example. Piccolo) and Real time processing system (Example. Simple scalable streaming system).

A. In- Memory Batch/ Big Data processing system

Piccolo is the Big Data processing system. Piccolo is a new data centric programming model for writing parallel in-memory applications in data centers [4] It borrows ideas from existing data-centric systems to enable efficient application implementations.

B. In-Memory Real time processing system

Simple scalable streaming system (S4) is the example of real time processing system. S4 (Simple Scalable Streaming System) is a type of streaming processing and distributed stream processing engine. S4 is a general-purpose, distributed, scalable, partially fault-tolerant, pluggable platform that allows programmers to easily develop applications for processing continuous unbounded streams of data [2]. Computation is performed by processing elements (PEs) which are distributed across the cluster, and messages are transmitted among them in the form of data events, which are routed to corresponding PEs based on their identities [1]. Keyed data events are routed with affinity Processing Elements (PEs), which consume the events and do one or both of the following: (1) emit one or more events which may be consumed by other PEs, (2) publish results [2].

The design goals were as follows:

a. Provide a simple Programming Interface for processing data streams.

b. Minimize latency by using local memory in each processing node and avoiding disk I/O bottlenecks.

c. Use a decentralized and symmetric architecture; all nodes share the same functionality and responsibilities.

IV. CONCLUSION

In this demanding world, In-memory data management and processing becomes increasingly interesting. Data storage shifted from disks to main memory can lead various improvements.

In-memory Big Data management provides a very efficient way to support highly available and performance oriented database management system. Ideal future memory technology should be nonvolatile, low cost, highly dense, energy efficient, fast and with high endurance. Such ideal technology would be universal memory. NUMA is to consider only a simple, two level memory hierarchy, even if the actual NUMA memory system is more complex. Simple scalable streaming system (S4) and Piccolo provides the In-memory Big Data storage processing system. We highlighted some strong design concept from which we can learn concrete system design principles.

V. ACKNOWLEDGMENT

We would like to thanks to co-author to guides us properly and gives his valuable time. We would also thanks to all journal and authors for their great work and development in In-memory Big Data management.

REFERENCES

1. Hao Zhang, Gang Chen, Beng Chin Ooi, Kian-Lee Tan, Meihui Zhang, “In-Memory Big Data Management and Processing: A Survey”, IEEE Transactions On Knowledge And Data Engineering, Volume 27, Issue 7, Page No. 1920, July 2015.
2. Leonardo Neumeyer, Bruce Robbins, Anish Nair and Anand Kesari, “S4: Distributed Stream Computing Platform”, August 2014.
3. Mohit Kumar Gupta, Vishal Verma, Megha Singh Verma, “In-Memory Database Systems - A Paradigm Shift”, International Journal of Engineering Trends and Technology, ISSN: 2231-5381, Volume 6, Issue 6, Page No. 333, Dec 2013.
4. R. Power and J. Li, “Piccolo: Building fast, distributed programs with partitioned tables,” in Proc. 9th USENIX Conf. Operating Syst. Des.Implementation, 2010.
5. Moinuddin K. Qureshi, Vijayalakshmi Srinivasan, Jude A. Rivers, “Scalable High Performance Main Memory System Using Phase-Change Memory Technology”, ACM, June 2009.
6. William j. Bolosky, Robert P. Fitzgerald, Michael L. Scott, “Simple but Effective Techniques for NUMA Memory Management”, ACM
7. Hector Garcia-Molina, Kenneth Salem, “Main Memory Database System: An Overview”, IEEE Transactions on Knowledge and Data Engineering, Volume 4, Issue 6, Page No. 509, Dec 1992.
8. www.csapp.cs.cmu.edu
9. www.techopedia.com/defination
10. IDC Digital Universe Study. (2012) The Digital Universe in 2020 : Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. [Online]. Available: http://idcdocserv.com/1414.