DATA STORAGE MANAGEMENT

PROJECT A

X18110088 (M.Sc. DATA ANALYTICS)

COHORT A

Context: 'We have entered the era of BIG DATA. To support the increasing demand for information services it is necessary for service providers to utilise networking and storage technology resources in an efficient and effective manner. There are opportunities for organisations to establish and leverage distributed computing environments in order to process large volumes of data.'

Requirement: An investigation into the utilisation of solid-state Flash storage in modern Big Data and predictive analytics processing environments. Present and provide a report on the utilisation Flash storage and protocols such as NVMe for analytics processing in terms of latency, cost, reliability etc. Comparisons between utilising Flash and alternative storage infrastructure options should be made.

Contents

1. Abstract

The term big data was invented by Doug Laney. Big data refer to the three main terms Volume, Velocity, Verity. This represents that the size of the data is huge. It is structures as well as unstructured. Velocity represents the speed through which the data is collected finally verity which is the variation in the dataset. Big data is large Data set and it is difficult to Handel it in traditional way. These datasets are extremely important as it contains hidden information which shows the pattern, behaviour, trends which is useful for the business users to make decision. To handle such high volume of the data and for fast processing requires High performance Storage. SSD Devices based on NAND flash memory are useful depending on big data application. Two types of SSDs enterprise class or personal storage SSDs are used according to the application requirement. These can be used as host cache, Network cache, all-SSDs storage arrays, or hybrid storage arrays with an SSD tier. Before selection of the SSDs it is important to have an understanding of the workload performance and clear understanding of the requirement. (Qiumin Xu1, nd)

2. Introduction :

2.5 quintillion bytes of data is generated in a single day. There are different sources of the data through which such large number of data is generated.there are total 3.7 billion humans use the internet daily which is large number and almost every person spend half of our on mobile web search. In the world there are total 5 billion searches a day with Google. Over the social media Snap chat users share 527,760 photos,120 user joins linkdin,414600 user watch you tube video,456000 tweets on twitter and finally Instagram user post 46740 photos per minute. This how data is generated. (Marr, 2018)

3. utilisation of solid state Flash storage in modern Big Data and predictive analytics processing environments.

This is the speed through which high volume of the data is generated, So it is very important to handle the data and get the useful information through this large size of the data. There are the amount of data that is qualify as big is debatable. It normally in tens terabytes or Peta bytes.To get the insect-full information, we need to apply different statistical traditional technique and it is not possible to get the information from such vast amount of data with the noise or veracity. So to handle the complex calculations researcher developed "**Predictive analytics**" or **User behaviour analytics** to manage the big data. (Qiumin Xu1, nd)This analytical methods contains different statistical methods such as predictive modelling, machine learning, and data mining. These methods are used to analyze the trends in historical and transaction data. There is huge challenge of extreme scale and because of this it in important to Handel and process the data cost effectively and quickly. **Distributed computing, massive parallel processing** all of these requires faster hardware with huge capacity. Efficient memory performance and capacity is necessary as big data requires fast processing throughput. Different memory are used in enterprise level for this purpose such as DRAM,HDD,SSDs cost of SSDs are less as that of DRAM but bandwidth and access time is less than DRAM.SSDs are much better than HDD. (Goodwin, nd)

Very high speed and electrically programmable memory is used in the Solid state Flash Storage. It performs various operation such as read and write as well as I/O operations in flash. flash memory is Non volatile memory and so if the power is not there, data still maintain. flash memory is available from USB to enterprise all flash arrays. Solid state defines no moving parts. There are lots of advantages of flash storage solution In the enterprise than HDD such as improving the performance of application, cost impact on the enterprise, Future-proof infrastructure. (Qiumin Xu1, nd)

Data files and process are distributed among the different nodes. so it is possible to replace all the hard disk drives with the SSDs in the big data system but this will require lot of cost and mainly unnecessary use of the SSDs where HDD are doing good. In the Big data system some process are IO intensive and other process are compute intensive. IO intensive operation that read or write huge amount of data. In this case SSDs are most useful .In the Compute Intensive Operation Requires computation. SSDs are not useful in this case if the data is fit into the memory and for processing huge amount of data in and out of memory to hard disk drive causes CPU to wait for the next block of data so in this case High speed NVMe is useful.

Three Different type of SSDs Deployment. Server side cache, storage side cache, SSD storage array. Use of SSD is again depend whether environment is CPU bound or I/O bound. When the environment is CPU bound, which requires more processing of the data then faster processer is useful and for reading Sequential data Continuously it is I/O bound so Use of SSDs is useful in the big data analytics. If the data is Recursive then it is better to use Server side cache and if the data is Sequential it is good to use all-SSD storage array to get the high performance across the dataset. (Goodwin, nd)

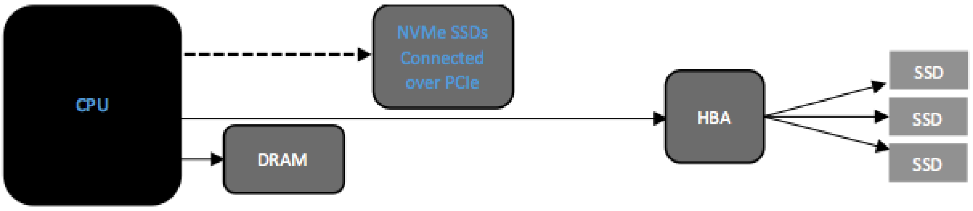
4.Utilization Flash storage and protocols such as NVMe for analytics processing in terms of latency, cost, reliability

What is NVMe?

Non volatile memory Express is high performance and scalable storage protocol. this Protocol which is directly connected to CPU through PCIe

And the speed offer by PCIe gen 3.0 is 2X than SATA interface.

NVMe supports ten thousands of parallel queues and support Concurrent commands

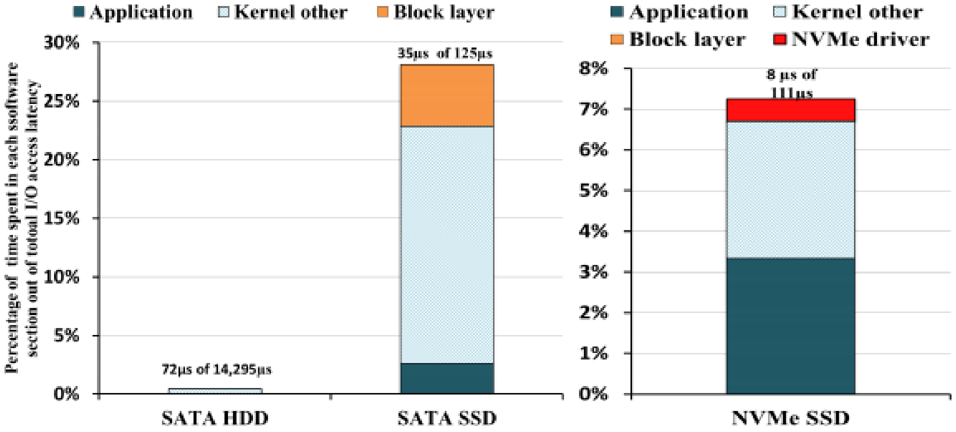


(Westerndigital)

NVMe take advantage on parallel and low latency data path .this performance significant higher performance and lower latency comparing to SAS and SATA protocol. Use of the CPU cycle are more for conventional protocol compare to NVMe and this will cost business real money. Budget of the company is not changing as per the speed in which the data handling is required. Workload handling is very careful by NVMe with smaller infrastructure so the organisation can reduce total cost of ownership. (Westerndigital)

Latency

Protocol such as SAS and SATA developed and used to connect slow hard drive theses protocol either connect through HBA or controller build into chipset. NVMe has lot of benefits over the SATA and SAS. Connectivity is much simpler and usefs PCIe bus. Researchers from the University of Southern California, San Jose State University and Samsung fio(an I/O traffic generator) and *blktrace* (a tool to trace I/O activity) and calculate I/O latency also track request from application to device and back again
Below two graph shows the comparison of the SATA HDD,SATA SSD AND NVMe SSD from the graph we can clearly see SATA HDD is much slower(14ms) compare to SATA SSD(125µs) and NVMe ssd (111µs).the overhead of NVMe is much lower so performance of the storage is not compromised due to use of the low latency storage device. (Qiumin Xu1, nd)



(Qiumin Xu1, nd)

Cost:

The cost of 512Gb SATA HDD is Euro €63, The cost 512Gb SATA SSD and NVMe is €84 and €99.so clearly Cost of the NVMe SSD is Higher and this is There are two different reasons for this SSDs normally use NAND flash technology which requires high cost and complex process is used to assemble the component of the SSDs so this is step increase the cost associated with SSDs. (SSD)

Reliability :

SSD is same as that of HDD but store data in different way. such that data is stored in interconnected flash memory chips. Data is maintain when there is no power present SSD are typically faster and more reliable.

5 Comparisons between utilising Flash and alternative storage infrastructure

Impact on performance of application:

Enterprise Applications like Teradata, Informatica, Oracle, My-Sql, Virtual machine and database like Hadoop and NoSQL require the fast processing and flash memory is much useful than HDD in such scenario also it is important to identify on which node it is useful to implement flash memory rather than HDD. This is very important from the perspective of IT employees to focus on business goals and this will improve the productivity. So with the use of SSD accelerates the application performance. (Patrizio, 2017)

Cost impact on Enterprise:

Due to the use of different flash storage solutions over the traditional ways of storing the data performance of the system improved as well as capacity increased. Due to reduction in the space, power and cooling canter huge amount of the cost is saved in long term for the enterprise (Patrizio, 2017)

Future proof infrastructure

It is important to choose the flash storage that support changing business model NvMe ready is useful to eliminate the cost. it is wise decision to choose flash system that supports cloud integration.
Now a days SSDs uses 3D TLC NAND and also growth in NVMe over NVMe-of and SCM offer huge help to the enterprise or data center. (Patrizio, 2017)

References

Adshead, A. (n.d.). *Big data storage: Defining big data and the type of storage it needs*. Retrieved from https://www.computerweekly.com/podcast/Big-data-storage-Defining-big-data-and-the-type-of-storage-it-needs

Goodwin, P. (nd). *Techtarget*. Retrieved from https://searchstorage.techtarget.com/answer/Are-SSDs-necessary-to-perform-analytics-on-big-data-efficiently

Marr, B. (2018, 5 21). Retrieved from https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#53fe913060ba

Patrizio, A. (2017, 8 14). Retrieved from https://www.hpe.com/us/en/insights/articles/crystal-ball-corner-the-future-of-enterprise-storage-1708.html

Qiumin Xu1. (nd). Retrieved from https://www.cs.utah.edu/~manua/pubs/systor15.pdf

Westerndigital. (n.d.). Retrieved from https://blog.westerndigital.com/wp-content/uploads/2018/02/What-is-NVMe-interface-and-controller-750x168.png