Memory Concerns for Big Data in HPC

L.R.F. Odiam and K.A. Hawick

The School of Engineering and Computer Science, University of Hull, Cottingham Road, Hull, Yorkshire HU6 7RX

Introduction

Big data is an increasing key aspect in a variety of fields including the global economy, weather prediction, and scientific research of all kinds. With this increase in the amount of data that is needed to be handled becoming an exponential one it is important that it can be handled in a efficient and effective manner which it what lends it to the application of a HPC. This is where many problems can arise. Until very recently the speed of CPUs has increase in accordance to Moores Law with CPUs capable of reaching single core speeds of 4.3 GHz in the case of Intel's i9-7900X and able to have a large number of core per die for example AMD's Ryzen Threadripper 1950X has a total of 16 cores and is able to handle 32 threads. This is not the case for the speed of memory access. In this poster we discuss this issue a give a suggestion of a interesting piece of future technology that could be of aid to this problem.

Big Data

Big data is a relatively new term, It refers to the capture, curation, and analysis of massive data sets[1, 2, 3], these data sets are so large and have such a high level of complexity that it is highly computationally expensive to work with. Big data was first characterized by 3v's this has in recent years been expanded on to form the current 4v's Volume, Velocity, Veracity, and Variety this is show in Figure 1. Until recently with advancements to technology it has not been possible to work with these types off data sets in a time frame that can allow them to be very valuable. Big Data is used in a great variety of fields.

Volume This is the main aspect of any data the makes it into "Big" data and also one of the main concerns when it comes to the handling and storage of the data in concern. Social media is an easy example of the Volume of data that can be involved [4]

Velocity How fast is the data produced. Just think about just twitter for instance where every second 6,000 tweets are tweeted this all adds up to a total of 200 billion tweets year. All of this data needs to be able to reach the server and be stored and indexed so that it is quickly accessible by as many people as possible.

Variety There is a large variety in the type of data the is used in the field of big data anywhere from simple text to video footage depending on the situation in which it is being used. This makes it very difficult to designed an optimal data structure for the storage and analysis of the data.

Veracity Veracity is a new addition to the 4v's. Veracity refers to the the number of abnormalities or noise with in the data that is being archived. This is a very important aspect of Big data as the storage of useless data is a waste an can be very costly. This mean that the "usefulness" of the data needs to be accessed in some manner, which when the volume and velocity of big data is taken into account become a gargantuan task.

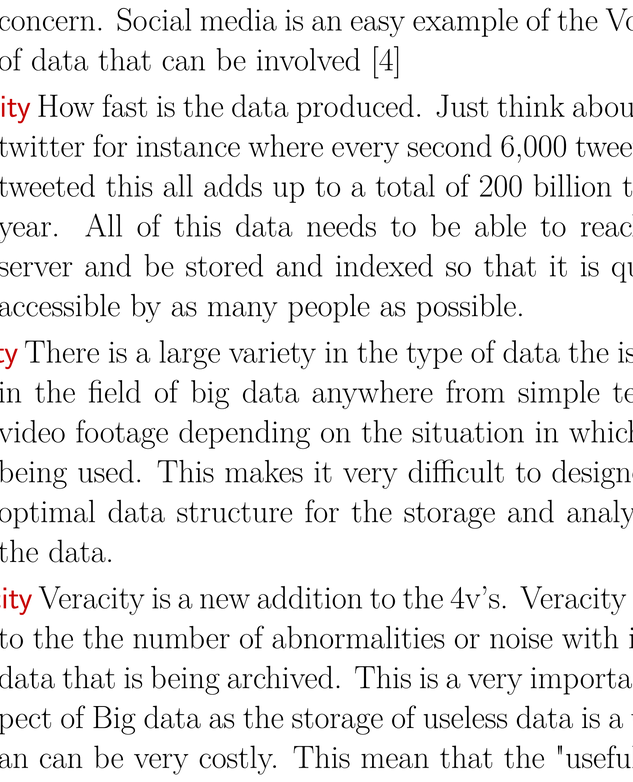

Figure 1: The Four V's of big data

The Memory Wall

This is the term that is used to describe the growth disparity between the speed of CPUs and the speed for off chip memory such as RAM, Hard Disk Drives (HDDs), and Solid State Drives (SSDs)[5]. Figure 2 there is a graph to show this disparity. The issue occurs due to the limitation of the memory bandwidth between the CPUs and other memory sources.

Where as it is older data [5] mention that on average programs reference memory with 20-40% of the instructions meaning in this means that in the best case the program will hit the memory wall when ever $t_{avg} = 5$ the equation for this is $t_{avg} = p \times t_c + (1 - p) \times t_m$ where t_c and t_m are the cache and DRAM access times and p the the chance of a cache miss.

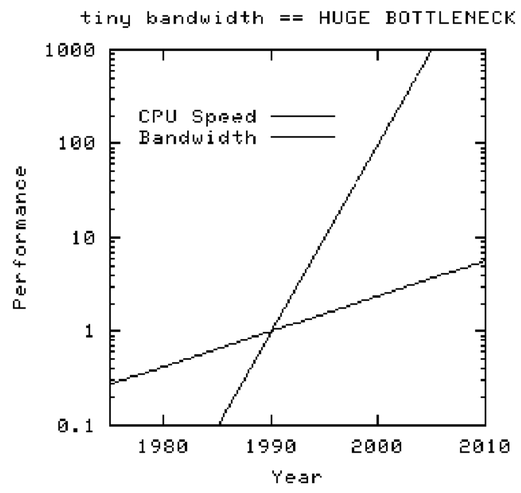

Figure 2: Memory wall showing the discrepancy between CPU speed and bandwidth [6]

Bandwidth is a very important thing in the case of memory. But just increasing the bandwidth does not necessarily improve performance it can instead impact it negatively if the RAM's access latency is fast enough. If this isnt the case it will end up taking up more cycles to get the data even if the cycles are shorter. This means that the CPU can end up doing nothing for a decent portion of its cycles, this latency is measure by the number of cycles it will take the memory to respond.

Memristors

Emerging Technologies for the advancement of memory are all memristors [7, 8, 9]. A memristor the name coming from the phrase "resistor with memory" is a device that regulate the flow of electrical current flow through a circuit and it remembers the amount of charge that flowed through it, this makes it so that a memristor can retain memory without power. Another of these benefits is their ability to be configured as either memory or CPU. The memristor concept was generalized by equations 1 and 2.

$$v = R(w)i \tag{1}$$

$$\frac{dw}{dt} = f(w) \tag{2}$$

In 2008 a Paper in Nature was published about the discovery of the "missing" memristor the found that TiO_2 could be used to form a memristor. By sandwiched the TiO_2 between two platinum electrodes when a positive voltage it applied the Doped side shown in Figure 3 which it doped with oxygen forming a TiO_{2-x} the spreads reducing the amount of TiO_2 in the memristor, which in turn will reduced the resistance[10]. This is then reversed by flipping the voltage.

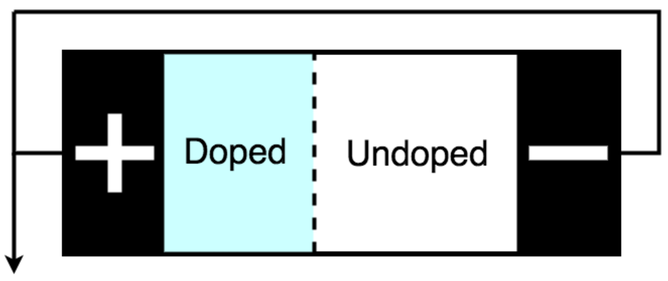

Figure 3: An example of how a memristor is made with two platinum electrode on the end and oxygen doped and undoped TiO_2 in the middle

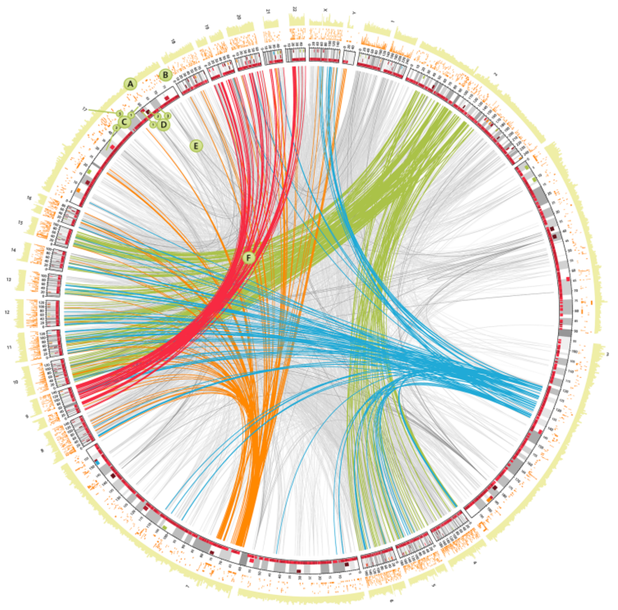

Figure 4: A visualization of big data

Benefits of Memristors

Non-volatile they are able to maintain memory without the use of constant power this can allow for the combination of working memory and long time storage this could even remove boot times form computers.

Low Power this means a cheaper running cost but also that it will produce less heat which in terms means an even lower power consumption though the use of cooling, depending on the situation that the memristors are being use in it could lead to a increase in the number of passively cooled devices.

Non-binary the ability to have a more float point value will allow it to more closely imitate an actual brain making then perfect for learning systems. They have already been proven capable of solving problems such as the finding the shortest path in a maze [11].

Fast a memristor has the ability to be faster then flash memory. this speed comes in part from the memristor density. HP labs have measure a drift velocity of $1ms^1$ meaning it can switch in less the a nanosecond.

Dense this could allow for 1 petabyte to be addressed in 1 cm of space massively reducing latency

Conclusion

In this poster we have discussed the memory wall as it relates to Big Data. This issue can lead to a greater impact on the performance of such systems. We have given a break down of the 4 v's involved with big data and what the entail. Finally we gave a possible future piece of technology that could en up solving these types of issues. With the memristors ability tp contain logic on them the memory bandwidth and latency would be almost completely done away with, the density of memristors could also make the storage of such massive data structures much more feasible especially with the rate that the volume of this data is increasing.

References

[1] C. L. Philip Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information Sciences*, vol. 275, pp. 314–347, 2014.

[2] P. Xuan, F. Luo, R. Ge, and P. K. Srimani, "Dynamic Management of In-memory Storage for Efficiently Integrating Compute- and Data-intensive Computing on HPC Systems," 2017.

[3] D. Laney, "META Delta," *META Group*, vol. 949, no. February 2001, p. 4, 2001.

[4] J. N. Cappella, "Vectors into the Future of Mass and Interpersonal Communication Research: Big Data, Social Media, and Computational Social Science," *Human Communication Research*, pp. 1–14, 2017.

[5] W. a. Wulf and S. a. McKee, "Hitting the memory wall: Implications of the Obvious," *ACM SIGARCH Computer Architecture News*, vol. 23, pp. 20–24, 1994.

[6] S. A. McKee, "Reflections on the Memory Wall," *Proceedings of the 1st Conference on Computing Frontiers*, pp. 162–167, 2004.

[7] S. Kvatinsky, A. E. Friedman, A. Kolodny, and U. Weiser, "Memory Intensive Computing," no. January, 2014.

[8] S. Hamdioui, L. Xie, H. A. D. Nguyen, M. Taouil, K. Bertels, H. Corporaal, H. Jiao, F. Catthoor, D. Wouters, L. Eike, and J. van Lunteren, "Memristor Based Computation-in-memory Architecture for Data-intensive Applications," *Proc. 2015 Des. Autom. Test Eur. Conf. Exhib.*, pp. 1718–1725, 2015.

[9] S. Hamdioui, S. Joshi, S. Kvatinsky, G. Cauwenberghs, L. Xie, N. Wald, K. Bertels, H. M. Elsayed, and H. Corporaal, "Memristor For Computing : Myth or Reality ?," pp. 722–731, 2017.

[10] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found.," *Nature*, vol. 453, no. 7191, pp. 80–3, 2008.

[11] Z. Ye, S. H. M. Wu, and T. Prodromakis, "Computing shortest paths in 2D and 3D memristive networks," pp. 2–12, 2013.