The Case for Non-Volatile RAM in Cloud HPCaaS

Yehonatan Fridman¹,², Re‘em Harel¹,³,⁴ and Gal Oren³,⁵

¹Department of Computer Science, Ben-Gurion University of the Negev, Israel
²Israel Atomic Energy Commission
³Scientific Computing Center, Nuclear Research Center – Negev, Israel
⁴Department of Physics, Nuclear Research Center – Negev, Israel
⁵Department of Computer Science, Technion – Israel Institute of Technology, Israel
fridyeh@post.bgu.ac.il, reemha@bgu.ac.il, galoren@cs.technion.ac.il

Abstract—HPC as a service (HPCaaS) is a new way to expose HPC resources via cloud services. However, continued effort to port large-scale tightly coupled applications with high interprocessor communication to multiple (and many) nodes synchronously, as in on-premise supercomputers, is still far from satisfactory due to network latencies. As a consequence, in said cases, HPCaaS is recommended to be used with one or few instances. In this paper we take the claim that new piece of memory hardware, namely Non-Volatile RAM (NVRAM), can allow such computations to scale up to an order of magnitude with marginalized penalty in comparison to RAM. Moreover, we suggest that the introduction of NVRAM to HPCaaS can be cost-effective to the users and the suppliers in numerous forms.

Index Terms—NVRAM, Intel Optane™, HPC, Cloud, HPCaaS, Infiniband, DDR5

I. SUPERCOMPUTING AND THE CLOUD

As the demand for more extensive calculations and simulations increases, huge amounts of data should be analyzed, manipulated, and stored. These needs ushered in a new Exascale era of stronger and more complex supercomputers [1]–[3]. This era was recently introduced with the first actual Exascale supercomputer – Frontier [4]. Frontier consists of more than 8 million compute cores that provide the peak computing power of 1.102 Exaflop/s.

Traditionally, said class of supercomputers can perform (almost) any kind of computation – from loosely to tightly coupled and from CFD simulations to deep-learning optimizations – with scalable performances compared to a single core or machine [5]. This scalability is mandatory for R&D breakthroughs, as exact and more detailed calculations are the key to better science, and they tend to be latency-sensitive [6]. Thus, the network architecture choice in modern supercomputers is of high speed, wide bandwidth, low latency, RDMA transmission enabled network, namely InfiniBand [7], [8], that is used to construct a complete system without (almost) any interference [9], [10]. With switch latencies close to 100ns, and end-to-end latency of less than 1µs (end-point to end-point), InfiniBand interconnect achieves comparable performances to shared memory communication for high-performance and data-intensive applications. In scaling measurements of Summit supercomputer, not long ago the world’s fastest supercomputer, latency experiments from/to all 4,608 nodes showed a latency of 1.3µs for 0B messages and 3.54µs for 4KB messages with Mellanox EDR InfiniBand [11].

Alongside the evolution of ‘classic’ supercomputing, a new architecture for HPC was raised – The Cloud, specifically HPC as a service (HPCaaS), which exposes HPC resources via cloud services [12]. The recent significant developments in the cloud allows better compute nodes than those in supercomputers, with denser computing power and much-improved access and elasticity [13]–[15]. As such, many small to low-medium scale calculations that usually needed access to supercomputers to be executed now can easily port to cloud services [15]. However, continued efforts to port large-scale tightly coupled applications with high interprocessor communication to multiple (and many) nodes synchronously, as in on-premise supercomputer, is still far from satisfactory and currently measured with a few hundred cores before scalability crashes [17]. This scalability issue results from the fabric (Ethernet, in this way or the other) with latency as high as one to two orders of magnitude than Infiniband on-premise supercomputers, even with dedicated optimizations [17]–[22]. As a result, the cloud is usually recommended to be used in such scenarios with just one powerful instance (or a few), with as much computing power, memory and storage [16], [23], [24]. For example, the new AWS Hpc6a.48xlarge contains 96 cores with 384GB of RAM [13], and the majority of cloud services already make available local SSD storage through SCSI and even NVMe interfaces [25]–[29].

From the business perspective, the economic model of the public cloud, that much prefers elasticity over connectivity speed, can not support Infiniband networking efficiently and cost-effectively, let alone supplying a large amount of resources simultaneously, spontaneously and instantly [30]. A unique exception that does not prove the rule is a supercomputer that was fully ported to the cloud: In Nov 2021, and for the first and only time since then, new supercomputers (mainly No. 10 on the TOP500 list [31], the Voyager-EUS2) were introduced entirely to one of the major cloud suppliers. As such, Voyager-EUS2 enjoys the two technologies’ benefits and disregards their weaknesses, except the most critical one in the cloud, the elasticity, that is bounded to the system resources.

In a seminal paper from 2009, Napper and Bientinesi wondered “Can Cloud Computing Reach The TOP500?” [32], and their concerns remain valid. Specifically, their observation that the flop/s obtained per dollar spent decrease exponentially with increasing computing cores and correspondingly, the cost
for solving a linear system increases exponentially with the problem size – is still very much in contrast to existing scalable supercomputers. "If cloud computing vendors are serious about targeting the HPC market," they stated, "different cost models must be explored. An obvious first step would be to offer better interconnects or nodes provisioned with more physical memory to overcome the slower network” [32].

II. Non-Volatile RAM in HPC Use Cases

Three primary resources of HPC (node-wise, excluding communication) are computing power, main memory and storage. Each component develops differently, which creates significant gaps that impair the scalability and performance of these systems. Large capacity of main memory is required to execute large problems, while its low latency is required to constantly feed the computing units with data. While computing power has increased exponentially over the years (mainly due to integrating heterogeneous hardware like CPUs and GPUs) [31], [33], DRAM scales slowly and presents no significant improvement in latency or bandwidth: The world’s first DDR5 DRAM chip was officially launched in 2020 [34], a decade after the DDR4 was introduced in 2011; and it has about the same latency as DDR4, just four times the density, and is only targeted to double the bandwidth. For example, in the flagship Aurora supercomputer [45], each compute node will hold the top density possible of 16 dies times 64GB per die, meaning ~1TB. The DRAM is also expensive relative to capacity, creating a significant gap with the increasing demand for computational power. Worse, the storage performances of HDDs and SSDs are extremely low in comparison to the DRAM, and exhibit no significant improvements in technology over the years. There are still 2-3 orders of magnitude between the performances of DRAM and storage devices [36], and as such they create a huge bottleneck to scientific applications [37]. However, DRAM is significantly more expensive than standard storage devices [36] and its power consumption is higher, due to the cost of frequent refresh operations [38]. Thus, building cost-effective, resilient, fast and power-efficient systems with large memory and storage spaces is a major challenge faced by the HPC community [39], [40].

A new building block in supercomputers is non-volatile RAM (NVRAM). Next-generation supercomputers are expected to integrate NVRAM devices. For example, Aurora [41] is planned to integrate the emerging Intel Optane™ NVRAM. NVRAM provides better density and energy efficiency than DRAM while providing DRAM-like performance: NVRAM comes in 128GB, 256GB and 512GB capacities (vastly larger than DRAM cards) [42] and its read/write latencies are 400% higher than those of DRAM, while the read/write bandwidths were 37% and 8% of those of DRAM, respectively [43], [44]. When configured in App-Direct mode these devices are byte-addressable and can be used by processes as NVRAM, while when configured in Memory or Flat mode these devices provide an extension to the volatile memory pool of the application, enabling the execution of much larger problems in a small number of compute nodes [45].

Prior work on the use of NVRAM in HPC environments was focused on three main direct use-cases: (1) as fast storage for diagnostics [46]–[48], (2) as fast persistence area for checkpointing [46], [49], and (3) as memory expansion to enable larger memory scientific workloads [45], [46], [50]. [51]. The two former use-cases rely on the non-volatility of NVRAM for fast storage as a substitute for standard storage mediums [46]–[48]. The third use-case investigates the benefits of NVRAM in Memory mode, using its large capacity to enhance HPC capabilities [45], [46], [50]. In particular, a comprehensive investigation [46] studies the main use-cases of Intel Optane™ NVRAM in HPC to improve scientific applications’ performances and fault tolerance.

III. The Case for Non-Volatile RAM in HPCaaS

It was shown that various HPC applications can benefit from scaling-in applications into fewer nodes by integrating NVRAM as memory expansion — in terms of energy efficiency and performance [45]. Patil et al. [45] investigated the trade-off between performance degradation due to the higher latency of NVRAM and the savings in intra-node communication overheads when executing various scientific applications on a single NVRAM node (which contains DRAM and NVRAM) rather than on 4 DRAM-only nodes. This investigation assumed an Infiniband connection between the 4 DRAM-only nodes, and found that the shared memory buffer communication on the NVRAM node has up to half the latency and 3x the bandwidth than the Infiniband-connected DRAM-only nodes. Patil inferred that compute-bound applications, which utilize cache locality well, are able to run larger problem sizes on fewer compute nodes with a DRAM-NVRAM hybrid memory system, experiencing minimal performance degradation while providing significant cost and energy savings.

As a consequence of previous conclusions by Patil et al. [45], and given the Ethernet-based cloud connectivity, we hereby take the claim that NVRAM is an essential, and even crucial component in future cloud HPCaaS instances, when considering memory and I/O bounded applications in general, or small to mid-range tightly-coupled computations in particular. We notice that even in current NVRAM capacity, it is possible – with only a few factors of performance degradation in comparison to RAM [46], [51], [52] – to achieve up to an order of magnitude more byte-addressable memory, with much lower costs [53], [54]. In weak-scaled computations, that rely on RAM extensions, the presence of such memory can save not only the need to extend to other instances/nodes (and by that to reduce dramatic latency overheads), but also to significantly reduce the costs – either for the clients (by requesting less resources and achieving more scalability) or the cloud supplier (with energy and scheduling savings). Furthermore, in applications that heavily rely on I/O, either for diagnostics or even for recoverability, the local NVRAM can serve as a reliable, vast and fast storage and memory combined, reducing the need to extend external distributed file systems over high latency communication.
ACKNOWLEDGMENT

This work was supported by Pazy grant 226/20, ISF grant 1425/22, and the Lynn and William Frankel Center for Computer Science. Computational support was provided by the NegevHPC project [55].

REFERENCES

[1] S. Heldens, P. Hijma, B. V. Werkhoven, J. Maassen, A. S. Belloum, and R. V. Van Nieuwpoort, “The landscape of exascale research: A data-driven literature analysis,” ACM Computing Surveys (CSUR), vol. 53, no. 2, pp. 1–43, 2020.

[2] A. G. Hoekstra, B. Chopard, D. Coster, S. Portegies Zwart, and P. V. Coveney, “Multiscale computing for science and engineering in the era of exascale performance,” Philosophical Transactions of the Royal Society A, vol. 377, no. 2142, p. 20180144, 2019.

[3] D. E. Bernholdt, S. Boehm, G. Bosilca, M. Gorenflo Venkata, R. E. Grant, T. Naughton, H. P. Pritchard, M. Schulz, and G. R. Vallee, “A survey of mpi usage in the us exascale computing project,” Concurrency and Computation: Practice and Experience, vol. 32, no. 3, p. e4851, 2020.

[4] D. Schneider, “The exascale era is upon us: The frontier supercomputer may be the first to reach 1,000,000,000,000,000,000 operations per second,” IEEE Spectrum, vol. 59, no. 1, pp. 34–35, 2022.

[5] F. Alexander, A. Almgren, J. Bell, A. Bhattacharjee, J. Chen, P. Colella, D. Daniel, J. DeSlippe, L. Diachin, E. Draeger et al., “Exascale applications: skin in the game,” Philosophical Transactions of the Royal Society A, vol. 378, no. 2166, p. 20190056, 2020.

[6] R. Murphy, “On the effects of memory latency and bandwidth on supercomputer application performance,” in 2007 IEEE 10th International Symposium on Workload Characterization. IEEE, 2007, pp. 35–43.

[7] G. F. Pfister, “An introduction to the infiniband architecture,” High performance mass storage and parallel I/O, vol. 42, pp. 617–632, p. 102, 2001.

[8] P. Grun, “Introduction to infiniband for end users,” White paper, InfiniBand Trade Association, vol. 55, 2010.

[9] Y. Ajima, “Optical connection of top-level supercomputers: current status and future expectations,” Optical Interconnects XXII, vol. 12007, pp. 11–19, 2012.

[10] T.-J. Lu, M.-C. Lai, and J.-S. Chang, “A survey of high-performance interconnection networks in high-performance computer systems,” Electronics, vol. 11, no. 9, p. 1369, 2022.

[11] V. G. Vergara Larrea, W. Josifov, M. J. Brim, R. D. Budiardja, D. Maxwell, M. Ezzel, C. Zimmer, S. Boehm, W. Eliasif, S. Oral et al., “Scaling the summit: deploying the world’s fastest supercomputer,” in International Conference on High Performance Computing. Springer, 2019, pp. 330–351.

[12] M. Parashar, M. AbdelBaky, I. Rodero, and A. Devarakonda, “Cloud computing study: Technology challenges in achieving exascale systems,” in Proceedings of the 5th Workshop on Exascale Computing: Resilience, Reliability and Energy Efficiency. Springer, 2020.

[13] A. N. Laboratory, “Aurora,” https://www.alcf.anl.gov/aurora/, 2022.

[14] K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, K. Hill, J. Miller et al., “Exascale computing study: Technology challenges in achieving exascale systems,” in Defense Advanced Research Projects Agency Information Processing Techniques Office (DARPA IPTO), Tech. Rep., vol. 15, 2008.

[15] A. N. Laboratory, “Aurora,” https://www.alfn.gov/aurora/2022.
