Advanced Software Architectures and Technologies in High Performance Computing and Data Centers

J.J. Vegas Olmos, Liran Liss, Tzahi Oved, Zachi Binshtock, and Dror Goldenberg
Mellanox Technologies Ltd., Yokneam Ilit Industrial Zone, Hermon, Yokneam, 20692, Israel
Author e-mail address: juanj@mellanox.com

Abstract: This paper reviews advanced software architectures and technologies that support in-networking computing and improve the overall performance of data centers and high-performance computing clusters; the ability to converge software and hardware allows for new solutions, such as artificial intelligence, to be deployed massively.

1. Introduction

Telecommunications networks of different types served for hundreds of years to convey information; it was however during the 80s, that the telecommunication industry started to work towards the concept of providing any served to anyone, anywhere, mainly thanks to the introduction of wireless technologies. Arguably, it was however the Telecommunications Act of 1996 [1] and the subsequent digitalization revolution what took us where we are. Services and digitalization of the underlying technologies created a paradigm consisting of a cloud and end user split. Cloud (or more generally, centralized infrastructure) consisted mainly of massive data centers. The end user eco-system developed over mobile and fiber-based last mile systems. In parallel, the scientific community developed high-performance computing (HPC) tools, mainly through large HPC clusters serving the computational needs of the scientific community.

Artificial intelligence (AI) is now shacking this ecosystem. AI can be widely described as machines that mimic cognitive functions that are normally associated with the human mind, such as learning and problem solving. AI is not new: the period from the 40-60s saw the theoretical foundations of the science laid down (the first full-scale humanoid robot, WABOT-1, dates back from 1972 and was built by Waseda University [2]), an AI winter period during the 70s due to a lack of public funding, the first boom of the 80s followed by the second AI winter in the 90s, and the current big data, deep learning and AI general intelligence wave. It is this new wave the one that identified the need to employ telecom/datacom networks as enablers of AI systems, thereby generating a push towards a convergence.

This convergence between the telecom/datacom infrastructure and AI is represented in Figure 1 in a simplified manner. Figure 1 shows a split, consisting in large data centers (DCs) that conform the commonly known “cloud”, an “edge” computing layer to interface with the end-user, and a final heterogeneous layer of end-users. The new cloud consists of DC resources, where aisles of HPC capacity are placed to support specific AI needs (i.e. AI training). The cloud is connected to the edge by high-capacity inter-DC interconnects. The edge acts as a buffer by serving the end-users by AI inference work; acceleration (or CPU offloading) is present at both the cloud and the edge, either through FPGAs, ASICs or SoC tailored to the needs of the end-user system being served (i.e. 5G, media broadcasting, automobile, healthcare...).

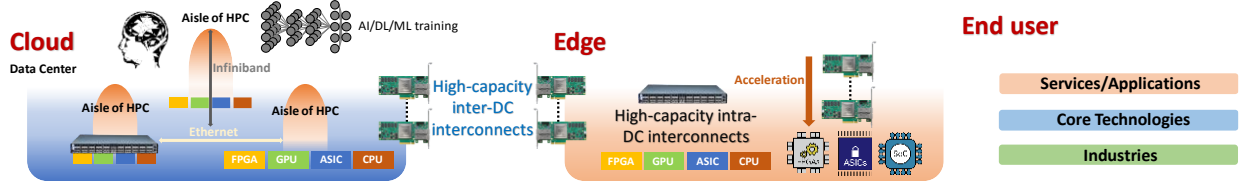

Fig. 1. Cloud and edge computing, the new split.

This paper highlights some of the new developments pushed by new Software Architecture technologies; during the presentation a browsing of new paradigms, solutions, market’s technology pulls and opportunities for R&D will be presented and discussed.

2. Enabling Technologies and Concepts

One of the main concepts supporting the current network developments enabling AI is in-network computing (also known as in-network computation or NetCompute) [3] . In-network computing refers to the ability by existing devices in the network (network interface cards (NIC) and switches) to conduct computations and traffic forwarding. In a traditional model, operations are conducted in either the kernel module or the virtual switch, both through

software. In-network computing relies on hardware offload, where indeed traffic may be initially managed in virtual switches, then transferred to the kernel, to be finally processed fully in hardware. Figure 2 represents these concepts, which in combination with some enabling technologies, boost dramatically the network performance.

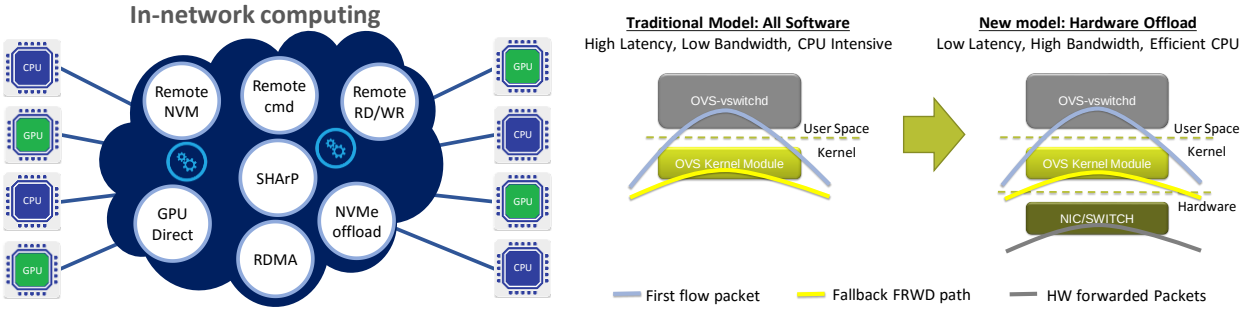

Fig. 2. In-network computing paradigm, where data is processed on the fly while traversing the network, and the hardware offload concept.

1.1. Remote Direct Access Memory (RDMA)

RDMA can be defined as a set of technologies allowing direct memory access from the memory of one processing unit into that of another without involving either one's operating system [4] . This feature is attractive as the transfer is not done by CPUs, caches or context switches, thereby transfers occurring in parallel with other system operations, reduced latency and effectively reducing the CPU load.

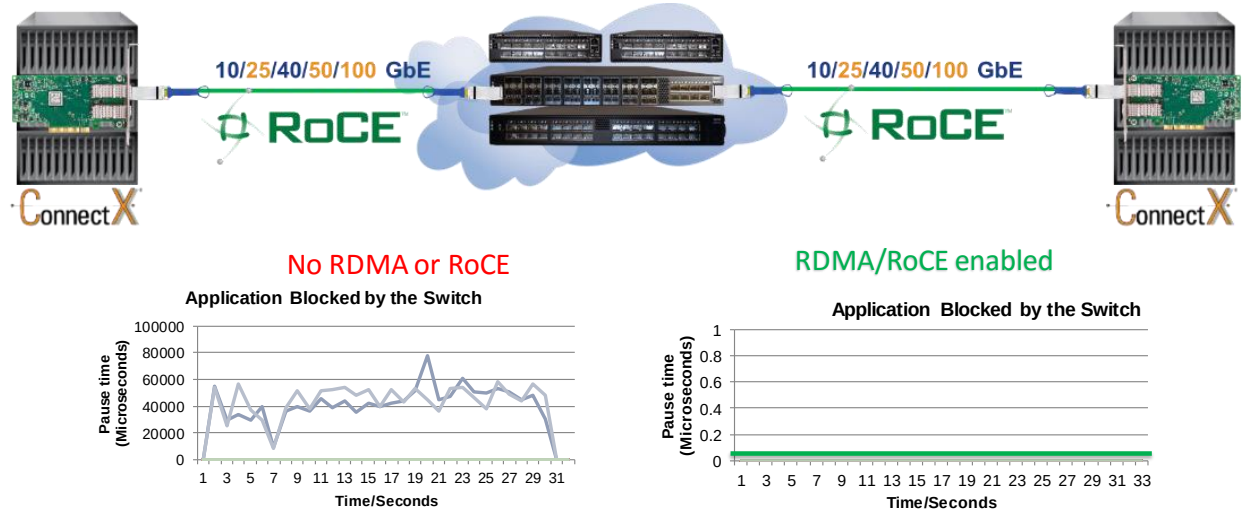

Fig. 3. RDMA or RoCE allow for lower switch blocking as memory management is off-loaded.

RDMA can be extended to operate over Ethernet (RDMA over Converged Ethernet – RoCE); RoCE was initially thought to bring Infiniband applications onto a common Ethernet converged fabric. At practical level, enables the design of large Ethernet-based data centers that behave as HPC in terms of memory management [5] .

1.2. GPUDirect

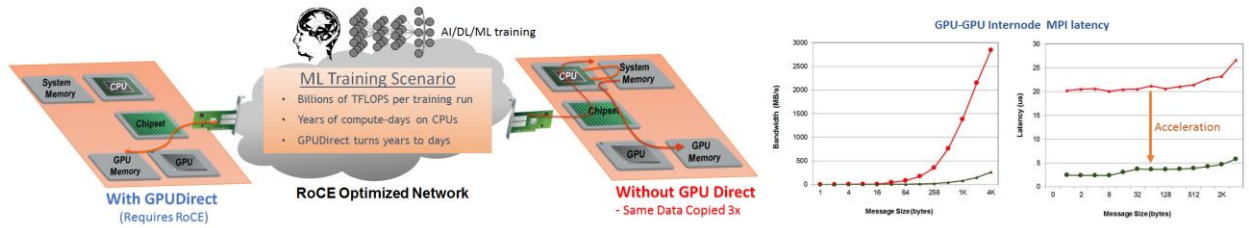

Fig. 4. GPUDirect allows to move data straight into the GPU block without copies within the system.

GPUDirect RDMA is a technology that enables a direct path for data exchange between a GPU and a third-party peer device (a NIC for example, connecting to the network). GPUDirect is very relevant in machine learning training scenarios, where intensive use of GPUs is needed, and therefore, latency reductions are critical [6] .

1.3. Cryptography acceleration

Security is a major concern and technologies such as Internet Protocol Security (IPsec) (which includes protocols for establishing mutual authentication between agents and negotiation of cryptographic keys) and Transport Layer Security (TLS) (which provides communications security over a computer network) are quickly penetrating the cloud and edge ecosystem. The gold standard for cryptography is effectively AES-GCM 128/256; Crypto-acceleration brings to the cloud/edge a full securitization of the communications at the edge, while drastically reducing the added latency and freeing up CPU resources, which can be used for AI processes instead.

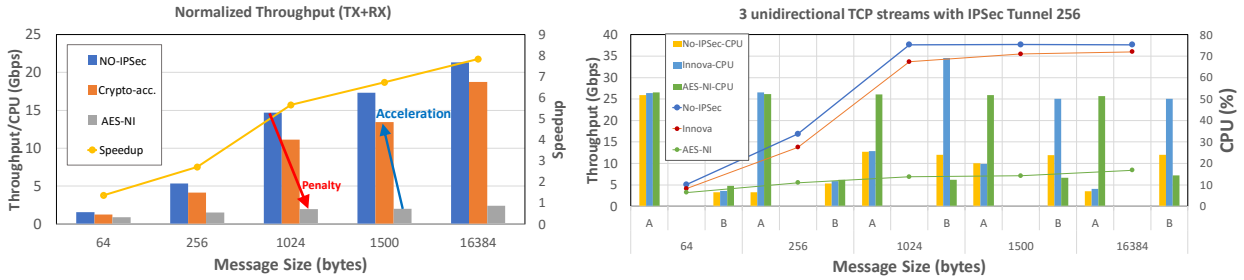

Fig. 5. Performance simulations of crypto-acceleration over different configurations; negligible impact on the overall performance when introducing crypto-acceleration.

1.4. NVM Express

The ability to store and retrieve all the information generated or processed in cloud/edge environments in a transparent and low-latency manner is critical; non-volatile memory (NVM) express is an open logical device specification for accessing non-volatile storage media via PCI Express (PCIe) bus [7] . NVMe SNAP (Software-defined Network Accelerated Processing) enables hardware virtualization of NVMe storage, enabling the integration of storage solutions into server deployments. NVMe SNAP allows migration of bare-metal cloud and virtual machine (VM) migration [8] .

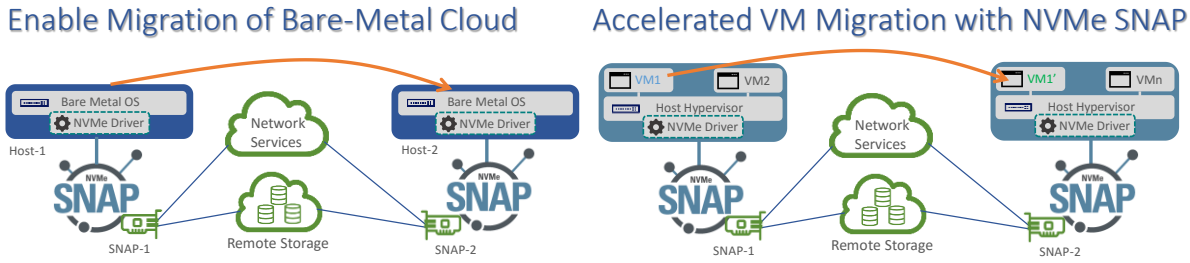

Fig. 6. NVMe and Software-Defined Network Accelerated Processing) for bare-metal and VM migration.

3. Conclusions

This paper presents some technologies that are changing the landscape of data centers and high-performance computing clusters. The introduction of artificial intelligence processes is propelling the adoption of novel technological solutions, which many times involve both software and hardware approaches designed in conjunction.

Acknowledgements

We would like to thank all the staff at the Software Architecture team of Mellanox Technologies for valuables inputs in the different research topics.

References

[1] The Telecommunications Act of 1996, Title 3, sec. 301, fcc.gov.
[2] Humanoid History – WABOT.
[3] Y. Tokusashi, H. Dang, F. Pedone, R. Soule, N. Zilberman, "The case for in-network computing on demand," ACM EuroSys, Art. 21, 2019
[4] S. Gugnani et al., "Switch-X: accelerating OpenStack swift with RDMA for building an efficient HPC cloud," IEEE/ACM CCGRID, 2017.
[5] RoCE Initiative, http://www.roceinitiative.org/
[6] G. Shainer et al., "The development of Mellanox/NVIDIA GPUDirect over Infiniband – a new model for GPU to GPU communications," Computer Science – Research and Development 26(3-4), June 2011.
[7] NVM Express – scalable, efficient and industry standard, https://nvmexpress.org/
[8] Software-defined Network Accelerated Processing, https://www.mellanox.com