TECHNOLOGY ADVANCES FUELLING A NEW WAVE OF HPC PERFORMANCE

Technology advances fuelling a new wave of HPC performance

Today we live in a world of data — so much so that there's often too much information available to make clear decisions in a timely fashion.

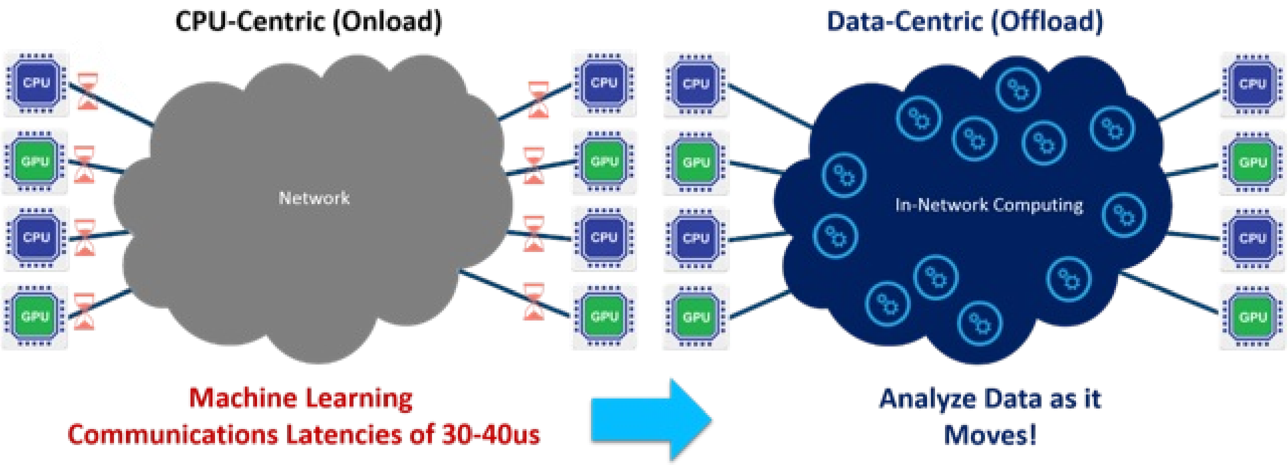
By Scot Schultz, SNIA Ethernet Storage Forum.

Since data analytics has become an essential function within today's high-performance enterprise datacenters, clouds and hyperscale platforms, solutions are necessary to better enable and more accurately gauge decisions based on analyzing tremendous amounts of data. With the right data analysis technologies, what was once an overwhelming volume of disparate information can be turned into a simplified, clear decision point. This encompasses a wide range of applications, ranging from security, financial, image and voice recognition, to autonomous cars and smart cities. The technology building blocks include better compute elements, graphics processing units (GPUs) and quite possibly the most important element of technology today - very fast intelligent network pipes to transfer all the data.

Improving the Network with Smart Offload Engines

In legacy architectures of days past, data could only be processed by a compute element such as a CPU or GPU and is commonly referred to as a CPU-Centric architecture. Having to wait for the data to reach the CPU before it can be analyzed created a delay or a potential performance bottleneck. Because of the increasing amount of data today, a CPU-Centric architecture can limit overall data processing, where as a data-centric architecture *can be used to o*vercome unnecessary latency. The fundamental concept behind this new architecture addresses the need to move computing to the data, instead of moving data to the computing, and results in a more modern approach for the datacenter to execute data analytics-*everywhere data exists*.

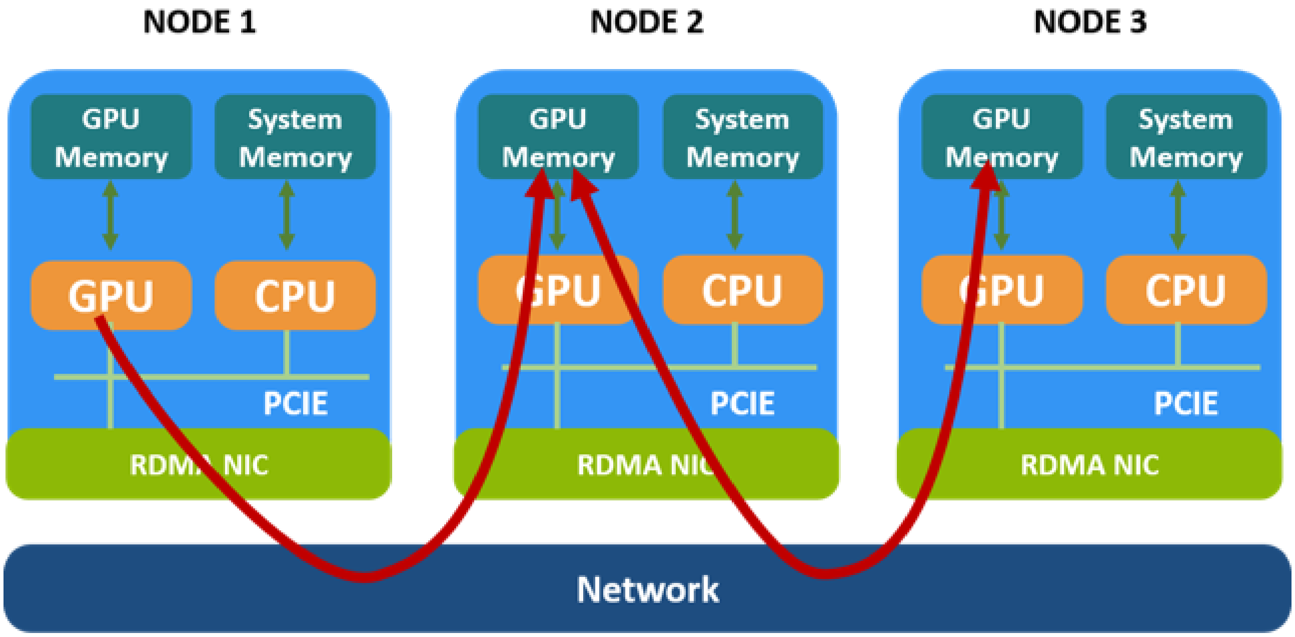
We are witnessing the introduction of several advanced capabilities and acceleration engines to address the challenges of the modern-day datacenter by delivering an intelligent network to act as a "co-processor", which shares the responsibility for computation. By placing the computation for data-related algorithms on an intelligent network, it is possible to dramatically improve both application performance and scalability. The first In-Network Compute capabilities include aggregation and reduction functions that perform integer and floating-point operations on data flows at wire speed, enabling the most efficient data-parallel reduction operations, which is extremely important for both high performance computing (HPC) and artificial intelligence (AI) workloads. In-Networking Computing technologies enable in-network aggregations to minimize the overall latency of reduction operations by reducing the bandwidth they require and their calculation time – thanks to the network devices performing the calculation on the fly.



GPUDirect Technology is one example of In-Network Computing to enable a Data-Centric Architecture

The rapid increase in the performance of graphics hardware, coupled with recent improvements in its programmability, have made graphic accelerators a compelling platform for computationally-demanding tasks in a wide variety of application domains. GPU-based clusters are used to perform compute-intensive tasks. Since GPUs provide high core count and floating point operations capabilities, high-speed networking is required to connect between the platforms in order to provide high throughput and the lowest latency for GPU-network-GPU communications.

The main performance issue with deploying platforms consisting of multiple GPU nodes has involved the interaction between the GPUs, or the GPU-network-GPU communication model. Prior to GPUDirect technology, any communication between GPUs had to involve the host processor and required buffer copies of data via the system memory.



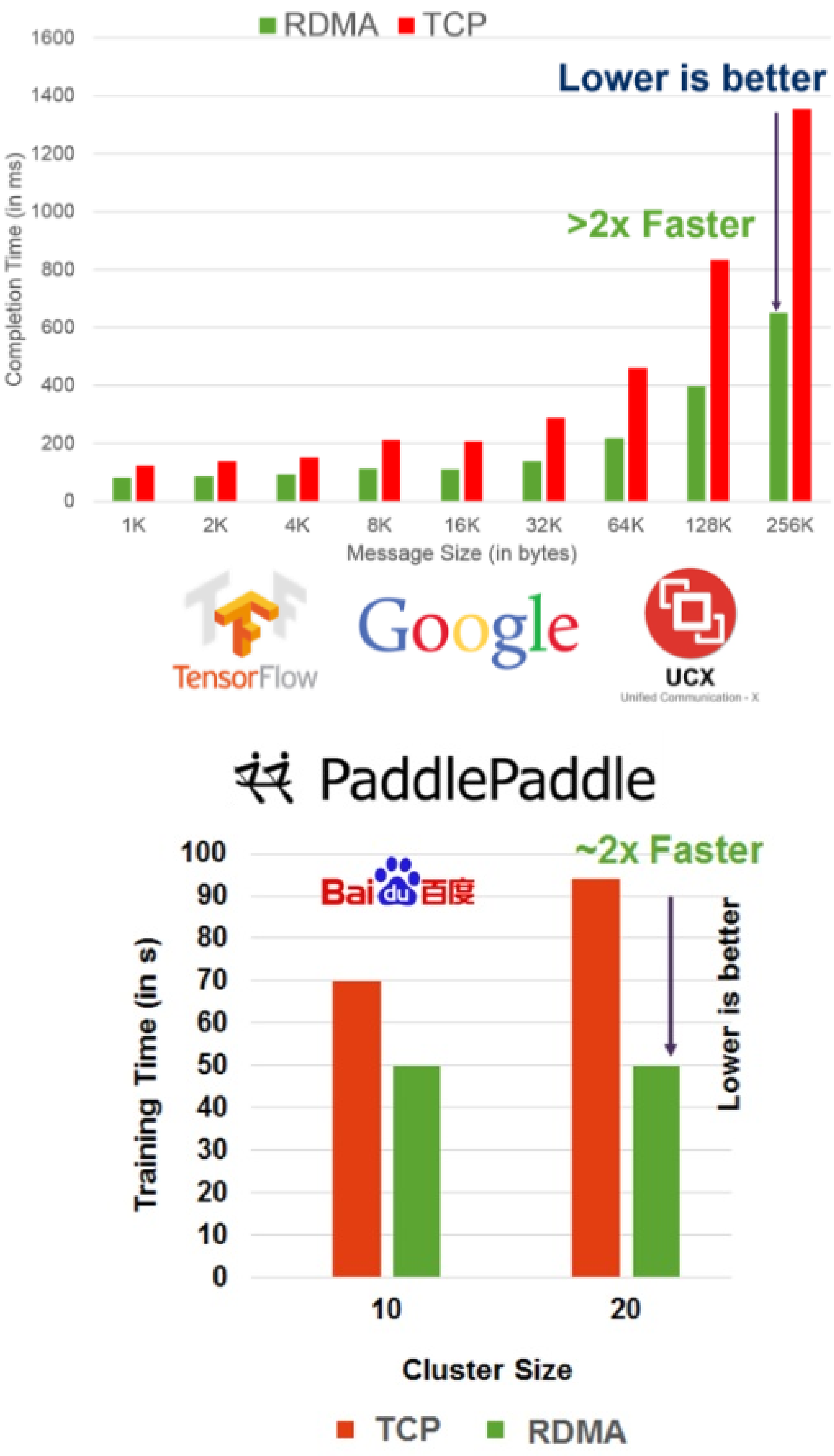
GPUDirect enables direct communications between GPUs over the network

GPUDirect is a technology implemented within both remote direct memory access (RDMA) adapters and GPUs that enable a direct path for data exchange between the GPU and the high-speed interconnect using standard features of PCI Express. GPUDirect provides significant improvement of an order of magnitude, for both communication bandwidth and communication latency between GPU devices of different cluster nodes, and completely offloads the CPU from involvement, making network communication very efficient between GPUs.

GPUDirect technology has moved through several enhancements since introduced, and the recent GPUDirect version 3.0, is also called GPUDirect RDMA. GPUDirect 4.0 or GPUDirect ASYNC is planned to be introduced in the near future, and will further enhance the connectivity between the GPU and the network. Beyond the data path offloads of GPUDirect RDMA, GPUDirect ASYNC will also offload the control path between the GPU and the network, further reducing latency operations at an average of 25%.

Remote Direct Memory Access (RDMA) Doubles AI Performance

RDMA usually refers to three features: Remote direct memory access (Remote DMA), asynchronous work queues, and kernel bypass. Remote DMA is the ability of the network adapter to place data directly to the application memory. RDMA is also known as a “one-sided” operation in the sense that the incoming data messages are processed by the adapter without involving the host CPU. Kernel bypass allows user space processes to do fast-path operations directly with the network hardware without involving the kernel. Saving system call overhead is a big advantage, especially for high-performance, latency-sensitive applications, such as machine learning workloads.



RDMA enables 2X higher performance for Tensorflow, PaddlePaddle and others.

While at the early stages, most of the AI software frameworks were designed to use the TCP communication protocol; now most, if not all (TensorFlow, Caffe-2, Paddle PaddlePaddle and others), include native RDMA communications due to the performance and scalability advantages of the latter.

The capabilities of In-Network Computing are set to continue and further increase performance with the future ‘Smart’ interconnect generations. Today In-Network Computing includes network operations, data reduction and aggregation algorithms, and storage operations. Future capabilities may include tighter integration of the middleware functionality, communications libraries and various elements of the machine learning frameworks themselves.

Looking Further Out

The amount of data being parsed for data analytics, and the amount of data leveraged to make real-time decisions will only continue to increase. Therefore we will see greater demand on the interconnect itself to provide faster data movement and more importantly, to execute data algorithms on the data while it is being transferred. It will not be practical to move the data to the compute elements; as previous, it will be mandatory to perform computational operations on the data where it resides.

As for the interconnect speeds and feeds, 2018 will usher in 200 gigabit per second speeds, and by 2019/2020 the capability to move data at 400 gigabit per second will be available. By 2022 we’ll approach moving data at nearly one terabit per second. We should expect the world to undergo an amazing transformation in how we interact with computers in just four short years due to In-Network Computing and higher transfer speeds. Autonomous self-driving vehicles, humanitarian research, personalized medicine, homeland security and even seamlessly interacting as a global society regardless of language or location are just a few of the exciting elements we will experience within our lifetime; and this is just the tip of the proverbial iceberg that will advance our knowledge and understanding of our place in the universe for generation to come.

About the SNIA Ethernet Storage Forum

The SNIA Ethernet Storage Forum (ESF) is committed to providing vendor-neutral education on the advantages and adoption of Ethernet storage networking technologies. Learn more at http://www.snia.org/esf.