Data transfer nodes and demonstration of 100-400 Gbps wide area throughput using the Caltech SDN testbed

A Mughal¹, H Newman²
California Institute of Technology
E-mail: °¹ azher@hep.caltech.edu, °² newman@hep.caltech.edu

Abstract. We review and demonstrate the design of efficient data transfer nodes (DTNs), from the perspective of the highest throughput over both local and wide area networks, as well as the highest performance per unit cost. A careful system-level design is required for the hardware, firmware, OS and software components. Furthermore, additional tuning of these components, and the identification and elimination of any remaining bottlenecks is needed once the system is assembled and commissioned, in order to obtain optimal performance. For high throughput data transfers, specialized software is used to overcome the traditional limits in performance caused by the OS, file system, file structures used, etc. Concretely, we will discuss and present the latest results using Fast Data Transfer (FDT), developed by Caltech. We present and discuss the design choices for three generations of Caltech DTNs. Their transfer capabilities range from 40 Gbps to 400 Gbps. Disk throughput is still the biggest challenge in the current generation of available hardware. However, new NVME drives combined with RDMA and a new NVME network fabric are expected to improve the overall data-transfer throughput and simultaneously reduce the CPU load on the end nodes.

1. Introduction
Since 2005, the Caltech high energy physics group (HEP) has participated in the Supercomputing Conferences (SC) for demonstrations of advanced research work in the field of data transfers at high throughputs, custom written application, network-system performance analysis using intelligent monitoring systems. For these demonstrations we design data transfer nodes (DTNs) using highly customized versions of commercially available hardware. Software on the DTNs is also customized to achieve low latency and maximum transfer throughput. The software tuning consists of custom kernel build for network and storage, kernel TCP/IP stack tuning, NIC and storage adapters tuning as well as NUMA pinning for each and every piece of hardware or software. An overview of the history of data transfers by Caltech in the past Supercomputing conferences is shown in Figure 1. The network throughput over the years is given in Table 1.

2. Caltech at Supercomputing 2016
At Supercomputing 2016 (SC16), the setup from Caltech and collaborators included multiple 100GE WAN links extended to national and regional PoPs including CENIC (Los Angeles, Sunnyvale); PacWave (Seattle); FIU/FLR (Miami) and StarLight (Chicago). It comprised a total of five 100 GE WAN connections. In addition, a star topology was laid from Caltech booth...
Table 1. Data Transfer throughput over the years at the Supercomputing conference

| Supercomputing          | Throughput |
|-------------------------|------------|
| 2005 Seattle            | 155 Gbps   |
| 2011 Seattle            | 100 Gbps   |
| 2012 Salt Lake          | 350 Gbps   |
| 2013 Denver             | 800 Gbps   |
| 2014 Louisiana          | 1.5 Tbps   |
| 2015 Austin             | 500 Gbps   |
| 2016 Salt Lake          | 2.5 Tbps   |

Figure 1. History of data transfers by Caltech at Supercomputing conferences since 2005, moving from 1/10/100 Gbps to Tbps and beyond.

to other collaborating booths, all at 100 GE links. These links were terminated on different SDN switches.

A customized version of the OpenDaylight\(^1\) SDN controller was deployed to provide application layer traffic engineering over parallel paths to redirect and control flows within the core and at the edges. This customization provides feedback to the northbound applications and offers the underlying infrastructure as a service (IAAS) and moves away from the traditional approach of network equipment configurations.

Wide area dynamic circuit setup demonstration were performed within the Pacific research, UNESP in Brazil and at the show floor using the NSI protocol over the 100G links. The WAN connectivity for the demonstrations from Caltech and partners at the Supercomputing 2016 is shown in Figure 2.

3. Design of 100 Gbps data transfer node using the TCP/IP protocol

A DTN node above 40 Gbps requires careful design in terms of hardware selection, data transfer tools as well as in terms of operating system stack tuning. Symmetrical multiprocessing (SMP) assignment is key to achieve consistent performance during every data transfer operation. We used Intel Haswell processor with 3.2 GHz clock rate processors, DDR4 memory, the SuperMicro X10 series motherboard and Mellanox ConnectX-4 VPI 100GE network card in designing this particular DTN node. We have standartized all the software across the CentOS 7.x releases which provide a very reliable and responsive OS environment.

\(^1\) www.opendaylight.org
Figure 2. WAN connectivity for the demonstrations from Caltech and partners at the Supercomputing 2016.

Figure 3. FDT 100 Gbps data transfers across two systems. Application readiness and performance is achieved out of the box.

We used a specialized but very light weight java application called as fast data transfer (FDT) [1] for the data transfer exercises shown in Figure 3. Single stream FDT transfer shows 68 Gbps, while dual stream FDT results show an aggregate throughput of 94 Gbps. In the case of single stream data transfer individual CPU core shows 100% utilization, which indicates limitations of the CPU core itself. If a higher clocked processor were provided, then single stream performance could be further enhanced.

4. DTN node disk I/O performance
Storage technology and host interface protocols have advanced quite drastically in the last few years. With the introduction of the NVMe stack and NVME SSD drives, disk I/O latency is reduced to new levels. Figure 4 (left) shows effective disk throughput while adding four drives during each test. One can observe that the performance increase is steady. Figure 4 (right) also shows that in a large NVME based system with 24 drives, 14 drives are enough to reach to the full PCIe Gen3 x16 bandwidth with peaks of 191 Gbps.
Figure 4. 200 Gbps disk throughput, design requirements.

Figure 5. RDMA based data transfer across 4 Mellanox VPI NICs. Only 4 out of 24 CPU cores are used.

5. Design of 400 Gbps data transfer node using the RoCE protocol
Performance of TCP/IP stack has shown challenges beyond 40 Gbps and limitations above 100 Gbps in terms of resource utilization. TCP acknowledgement is another limiting factor when data transfer nodes are beyond certain round trip time (RTT). An alternative approach to achieve high throughput within a data center or over the WAN is to use RDMA over converged Ethernet (RoCE) which offers data transfers with very low latency and CPU offload. Using a pair of Supermicro servers with four Mellanox 100 Gbps VPI NICs in each server connected back to back, we successfully achieved network throughput of 389 Gbps. The results are shown in Figure 5.

6. Summary
With the increased deployment of DTN servers in the data centers and in the Science DMZs [2], there is need for detailed hardware and system level optimizations and for an optimal design in order to achieve maximum data transfer throughput. In this paper we have shown results and examples that demonstrate that this optimization is possible. The chosen system designs allow to achieve a consistent, stable network throughput.
References
[1] Z. Maxa, B. Ahmed, D. Kcira, I. Legrand, A. Mughal, M. Thomas and R. Voicu, “Powering physics data transfers with FDT,” J. Phys. Conf. Ser. 331, 052014 (2011). doi:10.1088/1742-6596/331/5/052014
[2] "Science DMZ," https://fasterdata.es.net/science-dmz