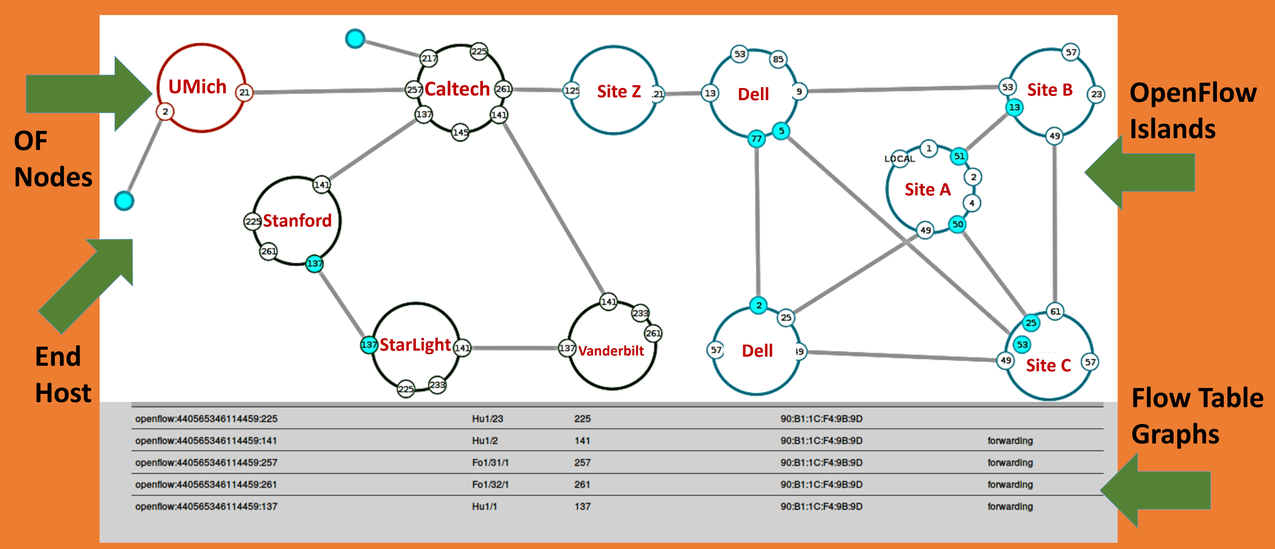
Data Transfer Nodes and Demonstration of 100-400 Gbps Wide Area Throughput Using the Caltech SDN Testbed

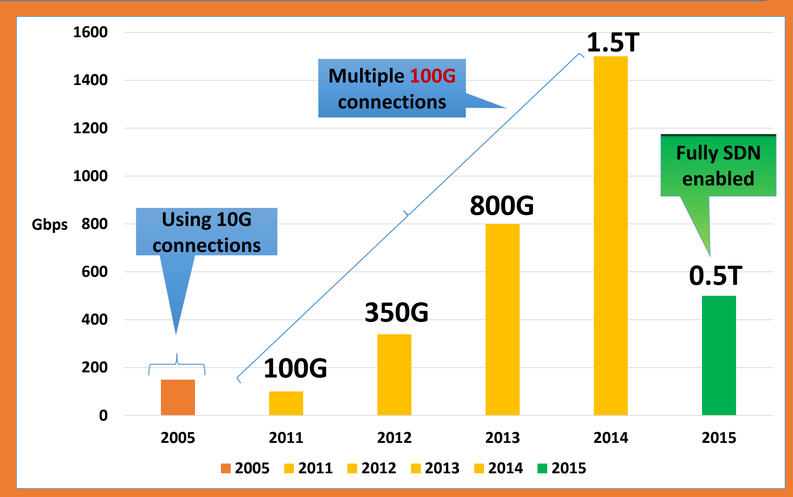
Azher Mughal, California Institute of Technology

Software Defined Network (SDN) testbed, connecting OpenFlow (OF) islands across the Wide Area Network



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

SC05 (Seattle): 155Gbps
SC11 (Seattle): 100Gbps
SC12 (Salt Lake): 350Gbps
SC13 (Denver): 800Gbps
SC14 (Louisiana): 1.5Tbps
SC15 (Austin): ~ 500Gbps
SC16 (Salt Lake): ~ 2.5Tbps

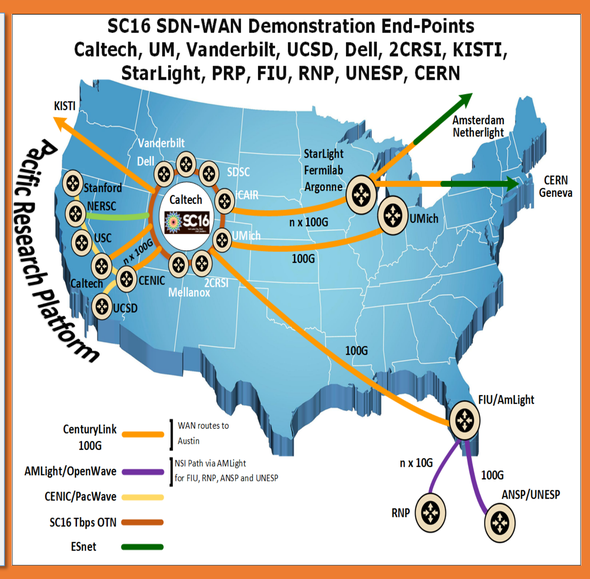


SC16 WAN Network connectivity: Pacific Research Platform, KISTI, Brazil, CERN

At SC16, multiple 100GE WAN links extended to regional PoPs including CENIC: Los Angeles, Sunnyvale; PacWave: Seattle; FIU/FLR: Miami and StarLight: Chicago

Dynamic circuit setup demonstration planned within the Pacific research, RNP in Brazil and at the show floor using NSI protocol

Setup through SDN controller that can do application layer traffic engineering to redirect and control flows within the core and at the edges



SC16 Multi Terabit Network & SDN network and traffic engineering with Application Layer Traffic Optimization

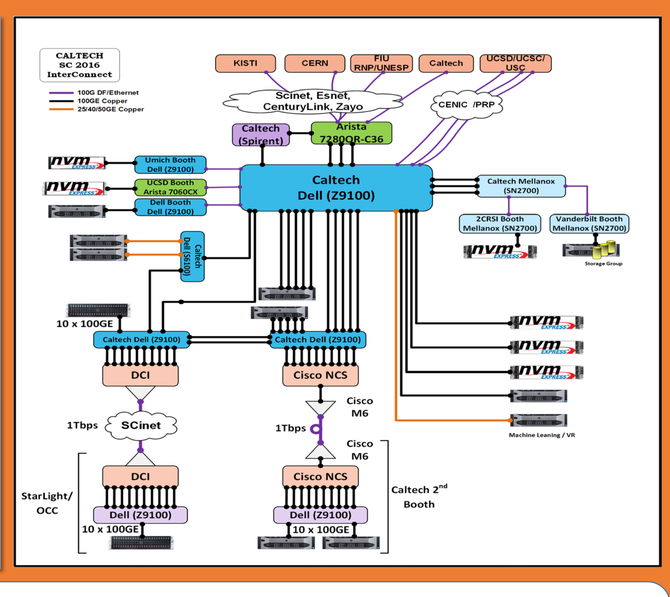
SDN Infrastructure includes:

OpenFlow switches from Arista, Dell, Inventec and Mellanox

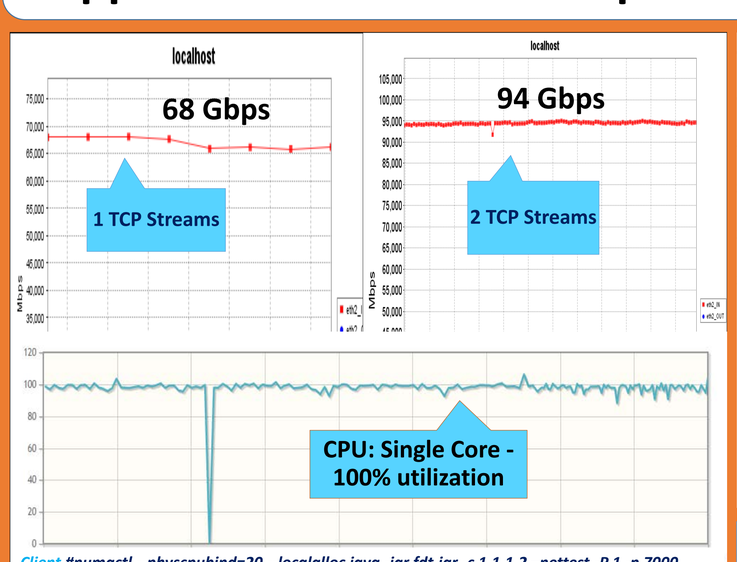
NVME over Fabric solution from HGST, Intel and LIQID storage cards

Network cards from Chelsio, Mellanox and QLogic

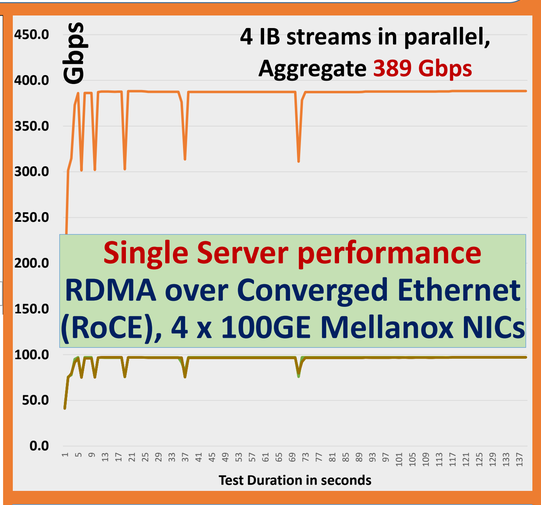
Tbit/s solutions from Ciena, Cisco and with the help of SCinet team



FDT 100 Gbps data transfers across two systems Application readiness and performance out of the box



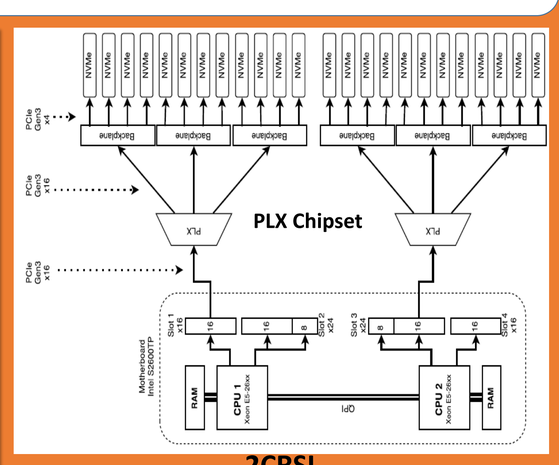
Client #numactl --physcpubind=20 --localalloc java -jar fdt.jar -c 1.1.1.2 –nettest -P 1 -p 7000
Server #numactl --physcpubind=20 --localalloc java -jar fdt.jar -p 7000



Transmission across 4 Mellanox VPI NICs.
Only 4 CPU cores are used out of 24 cores.

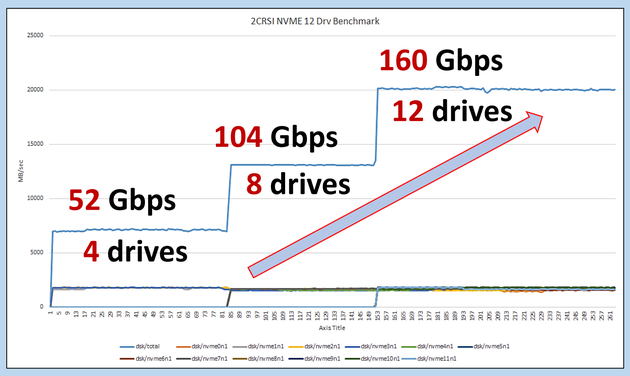
PCIe switch design to distribute large number of NVME drives across multiple processors, maximum bus utilization

- Current CPU generation offers 40 PCIe lanes
- At minimum, each NVME drive is PCIe x4 width, but not efficient to fully utilize a complete x4 slot's capacity
- Due to this limitation, a PCIe switching framework is needed to efficiently distribute the bandwidth to vide create a larger super fast DTN server
- Each CPU provides one PCIe x16 Gen3 to connect a group of 12 NVME drives
- While each CPU can provide additional one x16 bus for a 100GE NIC, to design a DTN system with 200Gbps



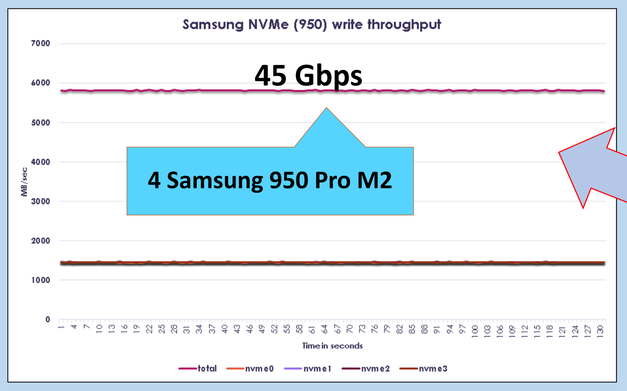
2CRSI

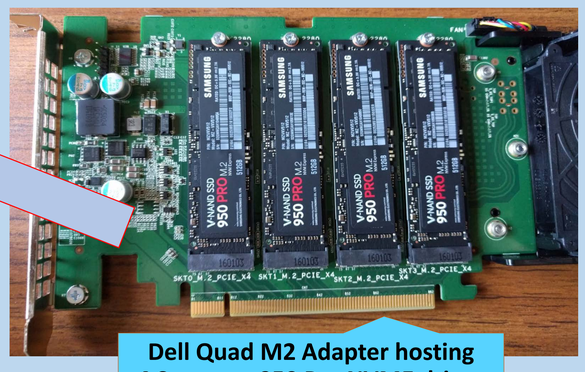
200Gbps disk throughput, Design Requirements



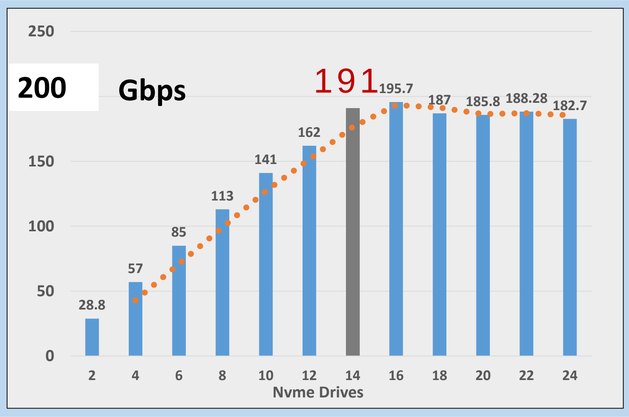
Design scalability with respect to number of drives

Disk throughput results across 12 Intel NVME drives with 4 drives added in each step





Dell Quad M2 Adapter hosting 4 Samsung 950 Pro NVME drives

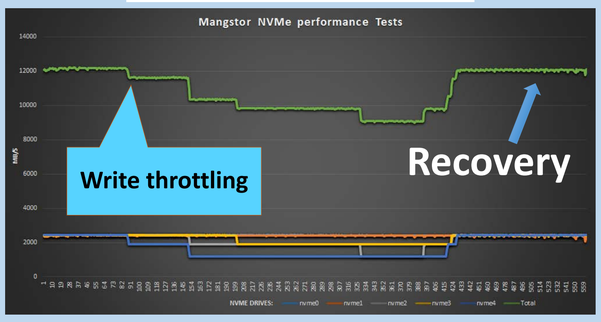


Disk throughput across 24 Intel NVME DCP 3700 2.5" drives in 2CRSI server. 191 Gbps is the saturation point observed using only 14 drives. At this point throughput almost remained constant

This means 1) To design a system with 200Gbps only 14 drives are enough 2) With 2 PCIe x16 slots each connects to one CPU, a saturation point is reached

Disk FIO disk benchmarks across pool of drives

12GB/s using 5 drives



Considerable performance degradation to about 30% of overall write throughput due to excessive heat blocked within the server, drives recover after installing additional fans

Airflow is very important in consistent write performance of the overall system design

http://supercomputing.caltech.edu/