Next Generation Data Transfer Nodes (DTNs) For Global Science: Architecture, Technology, Enabling Capabilities

Joe Mambretti, Director, (j-mambretti@northwestern.edu)
International Center for Advanced Internet Research (www.icair.org)
Northwestern University
Director, Metropolitan Research and Education Network (www.mren.org)
Director, StarLight, PI StarLight IRNC SDX, Co-PI Chameleon, PI-iGENI, PI-OMNINet (www.startap.net/starlight)

TNC 2018
Trondheim, Norway
June 10-14, 2018
Introduction to iCAIR:

Accelerating Leading Edge Innovation and Enhanced Global Communications through Advanced Internet Technologies, in Partnership with the Global Community

- Creation and Early Implementation of Advanced Networking Technologies - The Next Generation Internet All Optical Networks, Terascale Networks, Networks for Petascale and Exascale Science
- Advanced Applications, Middleware, Large-Scale Infrastructure, NG Optical Networks and Testbeds, Public Policy Studies and Forums Related to Optical Fiber and Next Generation Networks
- Three Major Areas of Activity: a) Basic Research b) Design and Implementation of Prototypes and Research Testbeds, c) Operations of Specialized Communication Facilities (e.g., StarLight, Specialized Science Networks)
Topics

- Motivation For Data Transfer Nodes (DTNs)
- Science DMZ 101
- Regional/National/Global DMZs (i.e., Regional, National and Global Research Platforms)
- Programmable Network Techniques and Devices (DTNs In Context)
- DTN 101
- Current DTNs
- Emerging Next Generation DTNs
- Results of Recent DTN Research Experiments
- Future DTNs
- Conclusions
Motivations

• Data Intensive Science - With Today’s Networks, Even R&E Networks, It Is Difficult To Transport Extremely Large Files and Collections of Many Files Over WANs, Especially Over Multi-Domains
• Also, However,
• AI Training/DL
• Distributed Storage Systems
• NFV
• Low Latency Services and Applications (Increasing Robotics)
• Low Latency And High Capacity/Big Data Applications
Science DMZ 101–
Segmentation Of Data Intensive Flows
(Ref:fasterdata.es.net)
Global Research Platform: Building On CENIC/Pacific Wave, GLIF and GLIF GOLEs (e.g., StarLight et al)
Programmable Network Techniques and Devices (DTNs In Context)

- Kernel By Pass
- SDN/SDX & Network Programming Languages
- Programmable Switch ASICs
- Programmable Network Processors
- FPGAs
- Smart/Programmable NICs
- Ref: Barefoot Tofino, Intel FlexPipe, Cavium XPliant, Netronome Agilio.
- P4 Based In-Network Telemetry
- AI/ML/DL Integrated With Network Programming
- RDMA (Remote Direct Memory Access)
- DPDK (Data Plane Development Kit)
- New Backplane/Switch Fabrics
- RoCE (RDMA Over Converged Ethernet)
DTN 101

- DTNs Are Purpose Built Servers Optimized For Data Transfer
- No “Standard” DTN Exists: Many Designs/Configs/Architectures Have Been Implemented (e.g. Memory Intensive, Storage Intensive, Memory and Storage Intensive, Neither Memory or Storage Intensive, 1 G, 10 G 40 G, 100 G, Edge Device, Intermediate Device, WAN Device)
- Multiple OS Stacks (Custom Tuned)
- Multiple DTN Middleware Stacks
- Multiple Transport Protocols
- Multiple Environmental Contexts (Ref Previous Slide)
Flash I/O Network Appliance (FIONA)

- FIONAs PCs [ESnet DTNs]:
  - ~$8,000 Big Data PC with:
    - 10/40 Gbps Network Interface Cards
    - 3 TB SSDs
  - Higher Performance at higher cost:
    - +NVMe SSDs & 100Gbps NICs Disk-to-Disk
    - +Up to 8 GPUs [4M GPU Core Hours/Week]
    - +Up to 196 TB of Disks used as Data Capacitors
    - +Up to 38 Intel CPU cores or AMD Epyc cores
  - US$1,100 10Gbps FIONA (if 10G is fast enough)
- FIONettes are US$300 EL-30-based FIONAs
  - 1Gbps NIC With USB-3 for Flash Storage or SSD
  - Perfect for Training and smaller campuses

Phil Papadopoulos, SDSC &
Tom DeFanti, Joe Keefe & John Graham, Calit2
Emerging Next Gen DTNs

- New Experimental Research Using DTNs
- iCAIR Innovations In DTNs For Data Intensive Science (Current: Developing 5th Gen)
- Focus Is On 100 Gbps And Multiple 100 Gbps
- Agency Innovations: NRL, NASA GSFC, ESnet
- Advanced Industry Components (e.g., NUMA, NVMe, Path Fabrics)
- Design Innovations (Thread Management, Optimal Affinity Bindings, NUMA Optimization)
Non-Volatile Memory Express (NVMe)

- Non-Volatile Memory Host Controller Interface (NVMHCI) Work Group
- Standard Architecture Specification For PCIe SSDs (Designed Specifically For faster Devices vs Traditional)
  - Register Interface
  - Streamlined Commands
  - Attributes
- Optimizes Host ↔ Storage
- NVMe Sends I/O Commands/Results To Shared Memory In Server via PCIe Interface
- Parallel I/O Using Multicore Processors
- One Message Queue Supports: 64,000 Commands
- Supports 65,535 I/O Queues
SUPERMICRO 24X NVMe SUPER SERVER

NVMe Type A: 8 X Intel P3700 800G
NVMe Type B: 8 X SamSung 950 512G
+ M.2 to U.2 Adopter
PowerEdge R740XD Server
2 X Intel® Xeon® Gold 6136
3.0G,12C/24T,10.4GT/s 2UPI,24.75M Cache,Turbo,HT (150W)
192G DDR4-2666
PCI-e Configuration Investigation:
2 X Mellanox ConnectX-5 100GE VPI
4 X Kingston/Liqid AIC NVMe PCI-e
X8 SSD Drives
Optional SAS/SATA Drives

3G / 3G + Clock Is Critically Important
Recent DTN WAN Experimental Research

• Motivated By Large Scale Science
• LHC
• LSST
• ESA
• Geophysical Sciences
• Genomics
• BioInformatics
• Precision Medicine
• Etc
Global Research Platform: Global Lambda Integrated Facility
Available Advanced Network Resources

Visualization courtesy of Bob Patterson, NCSA; data compilation by Maxine Brown, UIC.

www.glif.is
Global LambdaGrid Workshop 2017
Demonstrations, Sydney Australia

International Multi-Domain Provisioning Using AutoGOLE Based Network Service Interface (NSI 2.0)
Using RNP MEICAN Tools for NSI Provisioning
Large Scale Airline Data Transport Over SD-WANs Using NSI and DTNs
Large Scale Science Data Transport Over SD-WANs Using NSI and DTNs
SDX Interdomain Interoperability At L3
Transferring Large Files E2E Across WANs Enabled By SD-WANs and SDXs
Implementing a SCinet DTN (DTN-as-a-Service, DaaS)
A Disaggregated SCinet Optical Layer

Reconfiguration options:
A. Booth to booth connections
B. Booth to WAN connections
C. Booth to switch or router connections
D. WAN to switch or router connections

Examples:
A. B-B
   a. Booth 1001-1 to 1002-1 via optical layer
   b. Booth 1001-1 to 1004-3 via optical layer (assumes OLS2 to OLS4 path)
B. Booth to WAN
   a. Booth 1001-2 to PoP1-1 via OLS2-2 and OLS1-1
   b. Booth 1001-2 to PoP2-B1 via OLS2-2, OLS4, OLS6 and OLS5-1
C. Booth to switch/router
   a. Booth 1001-3 to SW1-A1
   b. Booth 1003-1 to RTR1-A5 (assumes OLS4 to OLS6 path)
D. WAN to switch/router
   a. PoP2-A1 (WAN2) to SW1-3 via OLS3-1 and OLS4-3
   b. PoP2-A2 (WAN2) to RTR1-3 via OLS3-2 and OLS4-7

Notes:
1) Transponders could be from multiple vendors but for near term the links would need to be built with matching transponders.
2) Controllers and orchestration systems are not shown but all Tpndr/OLS systems must be connected
SC17 @ Denver, CO

SCinet NOC Booth #1081

NASA demo @ StarLight/OCC/iCAIR Exhibit Booth #1653

SCinet

HECN X400NVME 24-NVME DISKS

BROCADE SLX-9240

Legend

- 100-GE
- NASA/GSFC-owned

SCinet DCI

x4

xN

ESnet

WASH-CR5

x4

C16-C17

C14-C16

x3

CenturyLink

NASA/GSFC
Greenbelt, MD

HECN X200NVME

DELL Z9100

MAX

x2

Level3/McLean

MAX JUNIPER MX960

Eckington

Dell/3M

NASA/GSFC High End Computer Networking (HECN) Team
Diagram by Bill Fink / Paul Lang - 8/14/2017

R&D Partners

NASA/AMDA

BROCADE

Esnet

STARLIGHT

ARISTA

CenturyLink

Energy Sciences Network

Dell

Intel

Supermicro

Mellanox

Micro SATA Cables

SCinet

Samsung

HECN

High End Computer Networking
Dynamic Distributed Data Processing

NRL will demonstrate:
- Dynamic distributed processing of large volumes of data
- Across geographically dispersed HPC and network resources
- Able to rapidly change resources to meet application demands.
- Software Defined Networking (SDN)
- Pipelined 12 Gbps real-time video processing workflows coexisting with bulk data transfers between processing nodes.
- 3 – 100 Gbps connections DC-to-SC17 floor plus 100 Gbps connections to Chicago and Oakland

“Interconnected and interlocking problems” call for a high performance dynamic distributed data centric infrastructure.
StarLight SDX Geoscience Research Workflow

Input parameter:
1. Date
2. Time (00/06/12/18)

1. Download to SDX DTN
2. grib2
3. Data convert1
4. netCDF
5. Data convert2
6. Convert to video
7. PNG
8. mp4
9. Display
10. Select Data
11. Data Transfer to Other DTN
100 Gbps DTN Optical Testbed

Ciena’s OPⁿ research network testbed

StarLight

Chicago

Ciena

Ottawa

Montreal

Hanover

New York

Baltimore

Wash DC

In Service

Recently Deployed
Additional WAN DTN Testing

- Preparation – Tests Conducted By Se-Young Yu (iCAIR)
- DTNs:
  - @iCAIR : Intel(R) Xeon(R) Gold 6136 CPU @ 3.00GHz, Mellanox ConnectX-5 100G NIC
  - @PACWAVE - LA : Intel(R) Xeon(R) CPU E5-2667 v4 @ 3.20GHz, Mellanox ConnectX-5 100G NIC
  - @UvA : Intel(R) Xeon(R) CPU E5-2630 0 @ 2.30GHz, Mellanox ConnectX-3 40G NIC
  - @CERN : Intel(R) Xeon(R) CPU E31220 @ 3.10GHz, Intel 82599ES 10G NIC
- Tuning Parameters :
  - BIOS, CPU, NIC, TCP Stack, O/S and MTU Tuning Applied
Memory-to-Memory Test Results

Server (X 4 instances): `iperf3 -s ipaddr -p port`
Client (X 4 instances): `iperf3 -c ipaddr -p port -Z -fq-rate=B -t60s`
Where \( B \) is \((\text{link-speed}/4)\).

<table>
<thead>
<tr>
<th></th>
<th>PACWAVE - LA (100G)</th>
<th>UvA (40G)</th>
<th>CERN (10G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sending</td>
<td>89.8% (89.8 Gbps)</td>
<td>89.8% (35.92 Gbps)</td>
<td>95.8% (9.58 Gbps)</td>
</tr>
<tr>
<td>Receiving</td>
<td>87.9% (87.9 Gbps)</td>
<td>92.5% (37 Gbps)</td>
<td>96.4% (9.64 Gbps)</td>
</tr>
</tbody>
</table>
Disk To Memory Test Results

Server (X 8 instances): `nuttcp -S -1 -sdz -P port < file`
Client (X 4 instances): `nuttcp -r -sdz -P port -i1 ipaddr > file`

<table>
<thead>
<tr>
<th>Starlight DTN</th>
<th>PACWAVE - LA (100G)</th>
<th>UvA (40G)</th>
<th>CERN (10G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading from disk</td>
<td>91.68% (91.68 Gbps)</td>
<td>73.06% (29.22 Gbps)</td>
<td>74.77% (7.47 Gbps)</td>
</tr>
<tr>
<td>Writing to disk</td>
<td>53.67% (53.67 Gbps)</td>
<td>99.50% (39.80 Gbps)</td>
<td>77.9% (7.79 Gbps)</td>
</tr>
</tbody>
</table>
The BigData Express Project (BDE)

BDE Research Team
November 2017
BigData Express – Toward Schedulable, Predictable, and High-performance Data Transfer
A Cross-Pacific SDN Testbed
**BigData Express SC’17 DEMO**

- BigData Express: a schedulable, predictable, and high-performance data transfer service
  - QoS-guaranteed data transfer
  - DTN as a service
  - Network as a service
  - Distributed resource brokering/matching

---

A DOE/SC/ASCR-sponsored research project
Software is available at: [http://bigdataexpress.fnal.gov](http://bigdataexpress.fnal.gov)
Test methodology

- 3rd party data transfer
  - Disk-to-Disk
- mdtmFTP vs. GridFTP
- Test scenarios
  - Run1 – 1 data transfer job, transferring a 300GB file
  - Run2 – 2 parallel data transfer job, with each transferring a 300GB file
  - Run4 – 4 parallel data transfer job, with each transferring a 300GB file
  - Run8 – 8 parallel data transfer job, with each transferring a 300GB file
- Metric
  - Aggregate throughput

Source: Wenji Wu
Server 1 Hardware Configuration

- Two NUMA Nodes
  - 28 cores, Intel(R) Xeon(R) CPU E5-2683 v3 @ 2.00GHz
- 64GB MEM
- One 100GE Mellanox ConnectX-4
- 8 NVME Drives

Source: Wenji Wu
Server 2 Hardware Configuration

- Two NUMA Nodes
  - 24 cores, Intel(R) Xeon(R) CPU E5-2687W v4 @ 3.00GHz
- 24GB MEM
- One 100GE Mellanox ConnectX-4
- 8 NVME Drives

```
Source: Wenji Wu
```
Testing environment – 3rd party data transfer

Source: Wenji Wu
### Performance – Aggregate throughput

<table>
<thead>
<tr>
<th>Gb/s</th>
<th>Run1</th>
<th>Run2</th>
<th>Run4</th>
<th>Run8</th>
</tr>
</thead>
<tbody>
<tr>
<td>GridFTP</td>
<td>6.2Gbps</td>
<td>12.24Gbps</td>
<td>20.35Gbps</td>
<td>28.32 Gbps</td>
</tr>
<tr>
<td>mdtmFTP</td>
<td>13.27Gbps</td>
<td>23.80Gbps</td>
<td>28.354Gbps</td>
<td>43.94 Gbps</td>
</tr>
</tbody>
</table>

Source: Wenji Wu

![Bar chart comparing mdtmFTP vs. GridFTP](chart.png)
Future DTNs

- Enhanced Core Components
- Much Faster Channels
- Optimal Use of Multi Core
- Higher Clock Speeds
- Enhanced NUMA Management
- Enhanced Thread Bindings/Affinity Management
- Enhanced In Network Computing
- Integration Into Network Orchestration
- Enhanced Programmability (e.g., DTN + P4 Switch)
- 400 Gbps/Tbps
APIs Based On Messaging and Signaling Protocols
Network Programming Languages
Process Based Virtualization – Multi-Domain Federation –
Policies Cascading Through Architectural Components

Orchestrator(s)

Policy Processes

State Machines

Northbound Interface

Network OSs
SDN Control Systems

Network Hypervisors

Southbound Interface

Security Processes

Policy Processes

State Data Bases

Mon, Measurements
Real Time Analytics

Eastbound Interfaces
Orchestration And AI & Selected ML Frameworks (Of Many):

- Apache Singa
- Caffe
- H2O
- MLlib (Apache Spark)
- Scikit-Learn (Python)
- Shogun (C++)
- TensorFlow
- Theano (Python)
- Torch (~ Scientific Computing)
- Veles (C++, w/ Some Python)
IRNC: RXP: StarLight SDX A Software Defined Networking Exchange for Global Science Research and Education

Joe Mambretti, Director, (j-mambretti@northwestern.edu)
International Center for Advanced Internet Research (www.icair.org)
Northwestern University
Director, Metropolitan Research and Education Network (www.mren.org)
Co-Director, StarLight (www.startap.net/starlight)

PI IRNC: RXP: StarLight SDX
Co-PI Tom DeFanti, Research Scientist, (tdefanti@soe.ucsd.edu)
California Institute for Telecommunications and Information Technology (Calit2), University of California, San Diego
Co-Director, StarLight

Co-PI Maxine Brown, Director, (maxine@uic.edu)
Electronic Visualization Laboratory, University of Illinois at Chicago
Co-Director, StarLight

Jim Chen, Associate Director, International Center for Advanced Internet Research, Northwestern University

National Science Foundation
International Research Network Connections Program Workshop
Chicago, Illinois
May 15, 2015
Emerging US SDX Interoperable Fabric
Source; John Graham UCSD
Potential For Close Integration of DTNs & Large Scale Storage

Building the Open Storage Network

Alex Szalay
The Johns Hopkins University

Institute for Data Intensive Engineering and Science
Summary

- Data Intensive Science As Well As Other Services and Applications Can Benefit From DTNs, Which Enable Enhanced Capabilities For High Performance LAN and WAN Data Transport, Including Customized Flow Management
- Key Enabling Capability: Using DTNs Integrated With Specialized WAN Paths, Including L2 Paths To Optimize E2E Data Flows, Including Disk To Disk
- Core Components Can Be Supplemented By Enhancing Software Stacks, e.g., Jupyter, NSI, MEICAN, P4 Programming, BDE, AI/ML/DL, etc
- Today, Many Components Exist To Create An E2E Services For Data Intensive Science
- Major Opportunity=> Creating DaaS Capabilities and Placing Them Into Production, e.g. Using the Global Research Platform (World-Wide Science DMZ)
www.startap.net/starlight

Thanks to the NSF, DOE, DARPA, NIH, USGS, NASA, Universities, National Labs, International Partners, and Other Supporters