Analysis of CPU Pinning and Storage Configuration in 100 Gbps Network Data Transfer

International Center for Advanced Internet Research
Northwestern University

Se-young Yu

Jim Chen, Joe Mambretti, Fei Yeh

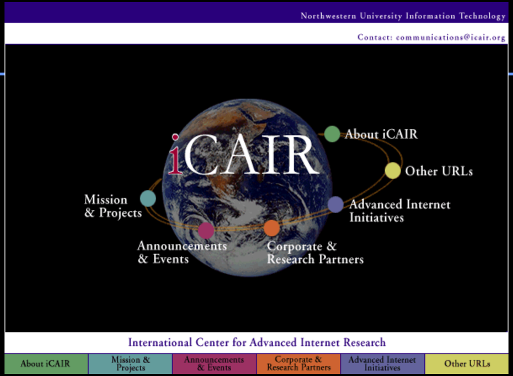
INDIS Workshop, Co-Located With
ACM International Conference for High Performance Computing, Networking, Storage, and Analysis

Dallas, TX

November 11, 2018

iCAIR

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)**

StarLight – "By Researchers For Researchers"

StarLight: Experimental Optical Infrastructure/Proving Ground For Next Gen Network Services
Optimized for High Performance Data Intensive Science
Multiple 100 Gbps (57+ 100 G Paths)
StarWave
100 G Exchange
Innovating First of a Kind Services and Capabilities

View from StarLight

Abbott Hall, Northwestern University's Chicago Campus

iCAIR: Founding Partner of the Global Lambda Integrated Facility
Available Advanced Network Resources

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

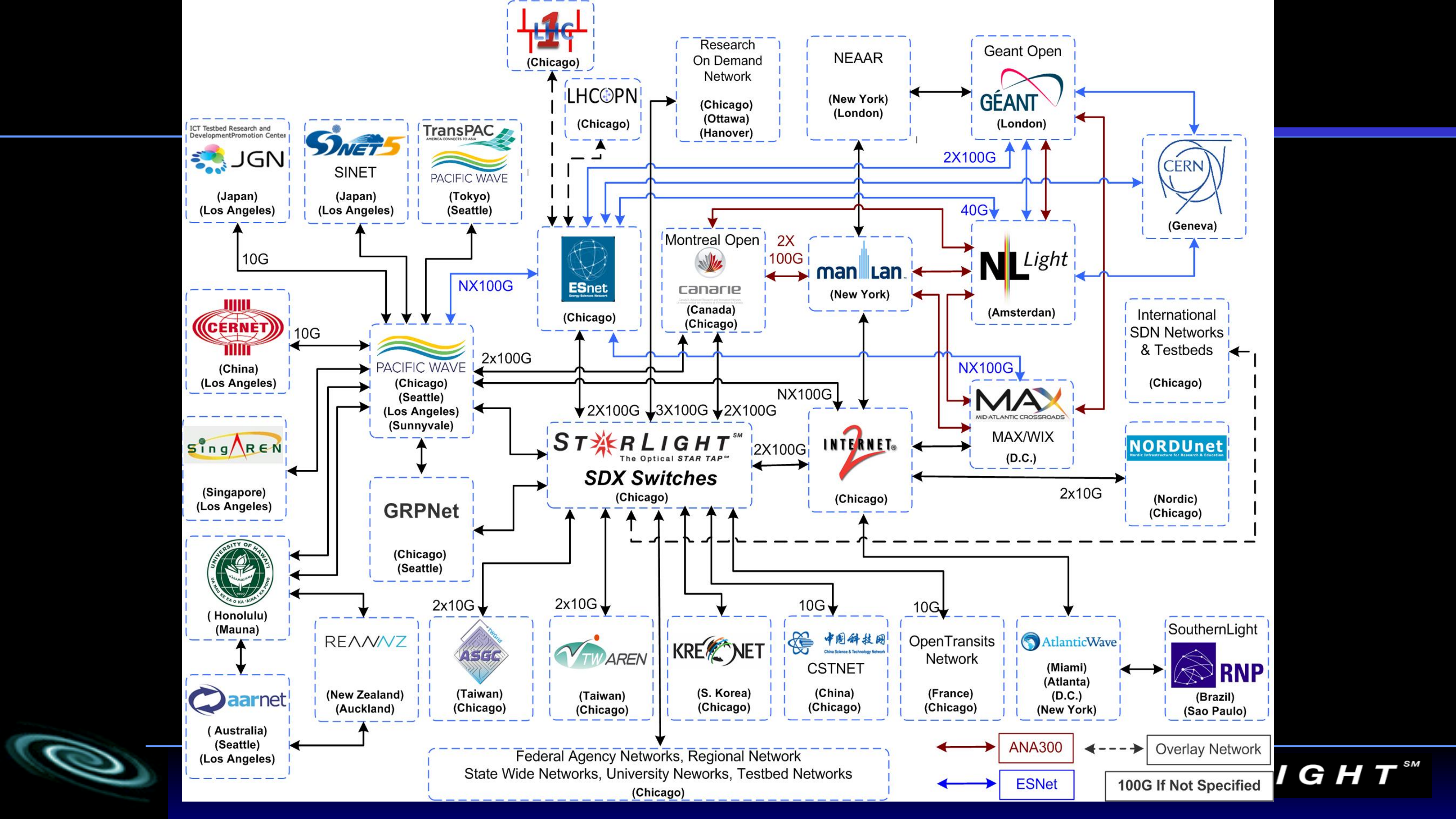
www.glif.is

iCAIR

STARLIGHT℠

StarLight SDX

- Context: Providing DTN Based Services For High Performance Data Flows Across the Globe, Especially For Data Intensive Science Integrated With an International SDX
- National Science Foundation International Research Network Connections RXP: StarLight SDX A Software Defined Networking Exchange for Global Science and Education (NSF ACI-1450871)



LHC (Chicago)
LHCOPN (Chicago)
Research On Demand Network
(Chicago) (Ottawa) (Hanover)
NEAAR
(New York) (London)
Geant Open
GÉANT
(London)
CERN
(Geneva)
ICT Testbed Research and DevelopmentPromotion Center
JGN
(Japan) (Los Angeles)
SINET5
SINET
(Japan) (Los Angeles)
TransPAC
PACIFIC WAVE
(Tokyo) (Seattle)
ESnet
(Chicago)
Montreal Open
canarie
(Canada) (Chicago)
2X 100G
manLan
(New York)
NLLight
(Amsterdan)
2X100G
40G
10G
NX100G
CERNET
(China) (Los Angeles)
PACIFIC WAVE
(Chicago) (Seattle) (Los Angeles) (Sunnyvale)
2x100G
International SDN Networks & Testbeds
(Chicago)
NX100G
MAX
MID-ATLANTIC CROSSROADS
MAX/WIX
(D.C.)
2X100G
3X100G
2X100G
SingAREN
(Singapore) (Los Angeles)
STARLIGHT
The Optical STAR TAP
SDX Switches
(Chicago)
2X100G
INTERNET2
(Chicago)
NORDUnet
(Nordic) (Chicago)
2x10G
GRPNet
(Chicago) (Seattle)
(Honolulu) (Mauna)
2x10G
2x10G
10G
10G
REANNZ
(New Zealand) (Auckland)
ASGC
(Taiwan) (Chicago)
TWAREN
(Taiwan) (Chicago)
KREONET
(S. Korea) (Chicago)
CSTNET
(China) (Chicago)
OpenTransits Network
(France) (Chicago)
AtlanticWave
(Miami) (Atlanta) (D.C.) (New York)
SouthernLight
RNP
(Brazil) (Sao Paulo)
aarnet
(Australia) (Seattle) (Los Angeles)
Federal Agency Networks, Regional Network
State Wide Networks, University Neworks, Testbed Networks
(Chicago)
ANA300
Overlay Network
ESNet
100G If Not Specified
STARLIGHT SM

Data Transfer Nodes (DTNs)

- Purpose-built systems (network appliances)
- High-Performance networks and I/O
- Optimized for 10-100 Gbps transfers
- Providing high-performance transfer tools to applications, processes, users
- Faster disks
- SSD, RAID, SAN

100 Gbps Transfer Challenges

- Modern x86 CPUs cannot transfer data at 100 Gbps with a single flow
  - Requires multiple concurrent flows (4 - 8)
- I/O devices create many more interrupts
  - NIC and storage devices through MSI-X
- HDD RAIDs cannot read/write at 12.5 GB/s

DTN Bottlenecks and Potential Solution

- Bottlenecks:
  - Slow disk speed, even with multiple NVMe
  - The multi-processors in DTNs start to create bottlenecks
- Responses:
  - Need to optimize CPU and storage to improve performance in network applications
  - Need to handle applications with increasing number of cores without increasing processor speed

Non-Uniform Memory Access (NUMA)

- Allows multiple processors to access memory simultaneously
- Forms a cluster-like logical processing unit called a NUMA node
- Memory controller handles memory access between NUMA nodes
- However, there is overhead in accessing memory in different NUMA nodes

How to use NUMA better

- Bind processes and NIC into the same NUMA node
- Processes can be bound
  - In specific core
    Each process is not allowed to be scheduled to another core
  - Within NUMA node
    Processes can be scheduled within local NUMA nodes
  - Anywhere
    Processes can be scheduled anywhere regardless of NUMA
- Binding a single memory-to-memory transfer process to local NUMA node improves throughput
- Q: Does this technique also improve throughput in disk-to-disk transfer?

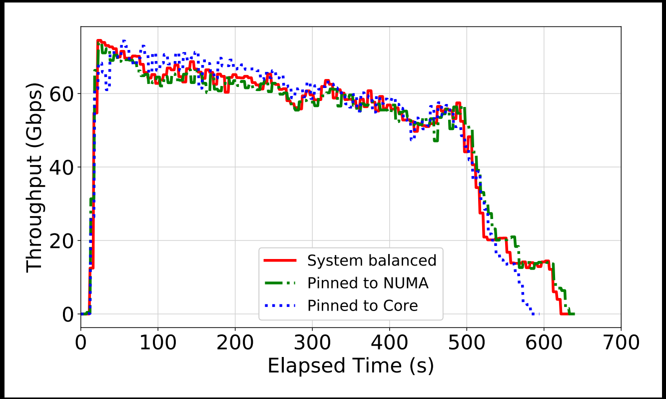
How to use NUMA better

- Hypothesis: It is better to bind disk transfer processes in local NUMA nodes
  - Put a NIC and NVMe devices in the same NUMA
  - Limit transfer processes to that NUMA
  - > Less foreign NUMA memory access

Why Test Initially Within Local Area Network?

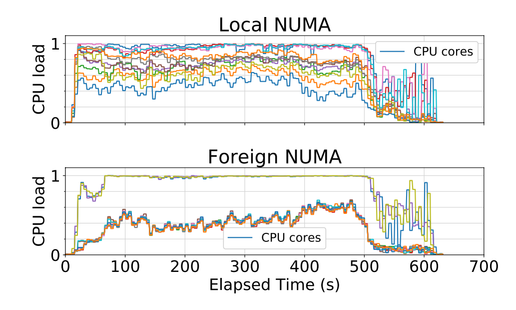
- Easier to establish a clean path
  - Minimizing packet loss effect
  - 100 Gbps DTN with NVMe is not widely available until recently
- NUMA access latency difference between the nodes is in 100s of nanoseconds where TCP is in milliseconds
  - Larger TCP access latency may disguise affinity effect
  - Need to highlight the difference in affinity settings

How To Use NUMA Better (Individual Disk) -Pinned to Core

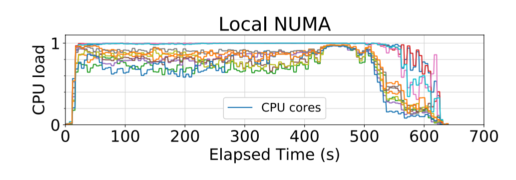


- **Throughput limited by NVMe under test ~ 60 Gbps**
- **Binding processes to specific core reduced overhead**

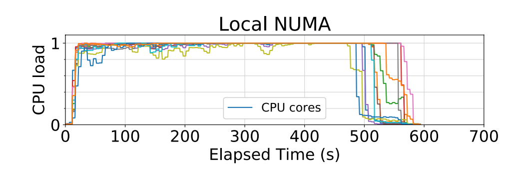
CPU load on System Balanced



CPU load on Pinned to NUMA



CPU load on Pinned to Core



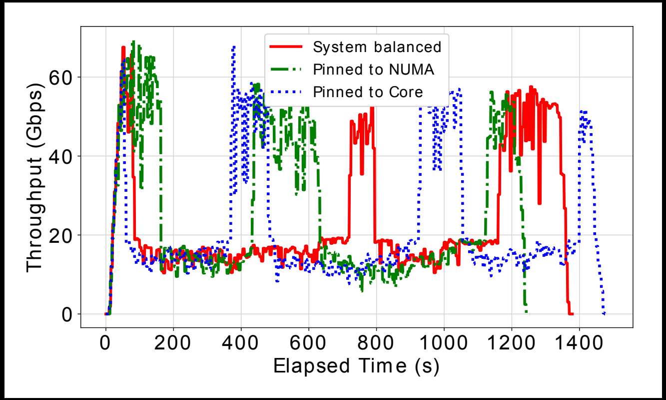
iCAIR

STARLIGHT℠

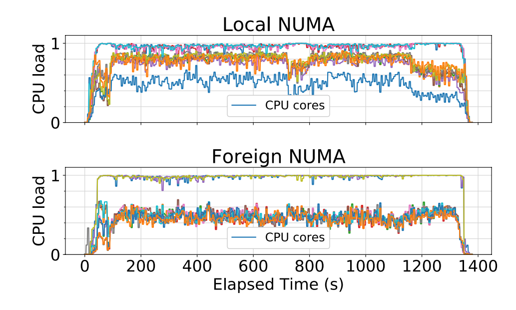
RAID vs Non-RAID

- RAID provides better management for multiple storage systems
- Is RAID overhead significant in 100 Gbps transfer?
- Are there any differences among different process binding schemes?

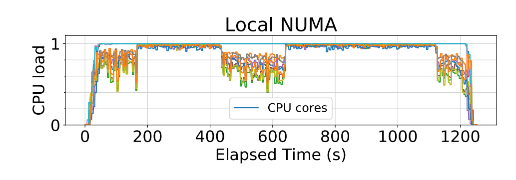
How to use NUMA better (RAID) - Pinned to NUMA



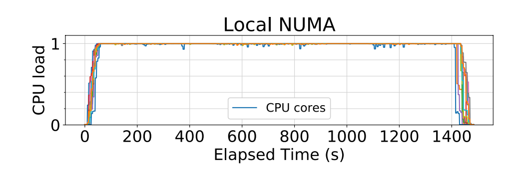
CPU load on System Balanced



CPU load on Pinned to NUMA



CPU load on Pinned to Core

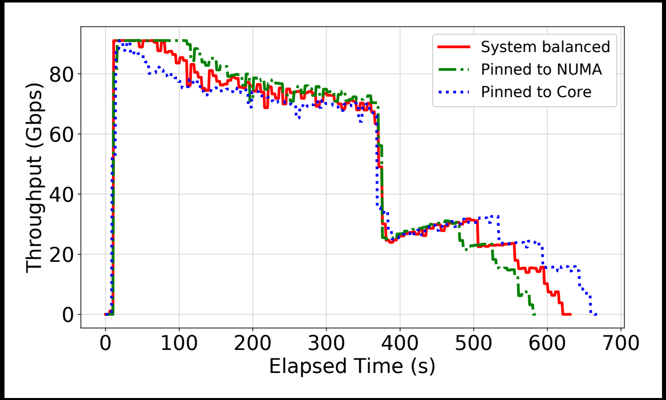


- **Software RAID created more overhead at almost 100%**
- **Under heavy loads, flexibility within NUMA helps**

NVMe over Fabrics

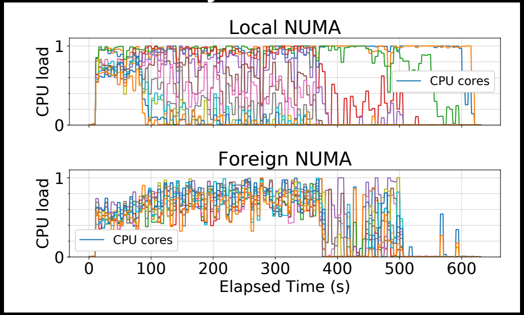
- Connects remote NVMe devices to local machine
- Access to the NVMe block device
- File copy tools can be used
- Good for streaming instead of copy and paste

NVMe over Fabrics - Pinned to NUMA

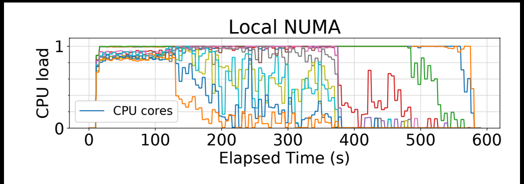


- Faster peak throughput, but similar completion time – some processes have delayed completion time
- Pinning to NUMA allows faster completion time

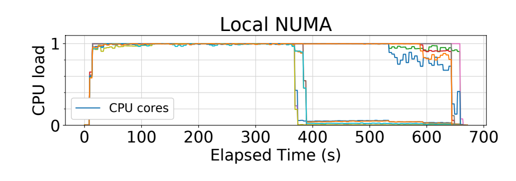
CPU load on System Balanced



CPU load on Pinned to NUMA



CPU load on Pinned to Core



Starlight SDX DTN-as-a-Service Software Stack

- Building a loss-less DTNs in 100 Gbps network
  - Developing a NUMA-aware testing framework to identify packet loss between DTNs
  - Systematic tuning and optimization of the DTNs for 100 Gbps disk-to-disk transfer
  - Passive monitoring system for bottleneck detection in high-speed network data transfer

X-NET: SCinet Data Transfer Node(DTN) Service

TEAM MEMBERS

Jim Chen	NWU/STARLight
Gonzalo Rodrigo	Apple/LBL
Ana Giannakou	LBL
Eric Pouyoul	ESnet
Fei Yeh	NWU/STARLight
Se-Young Yu	NWU/STARLight
Xiao(Shawn)Wang	NWU/STARLight
David Wheeler	NCSA/UIUC

ABLE TO DO:

1) Develop 100G network fiber/link/vlan/route verification procedures with a portable tester to shorten set up time and improve readiness.
2) Prototype user experiment environment isolation & management solutions: Docker/ Kubernetes/Rancher/VM, also plan to evaluate other Docker Integration
3) Design AI-Enabled DTN use case and workflow prototype

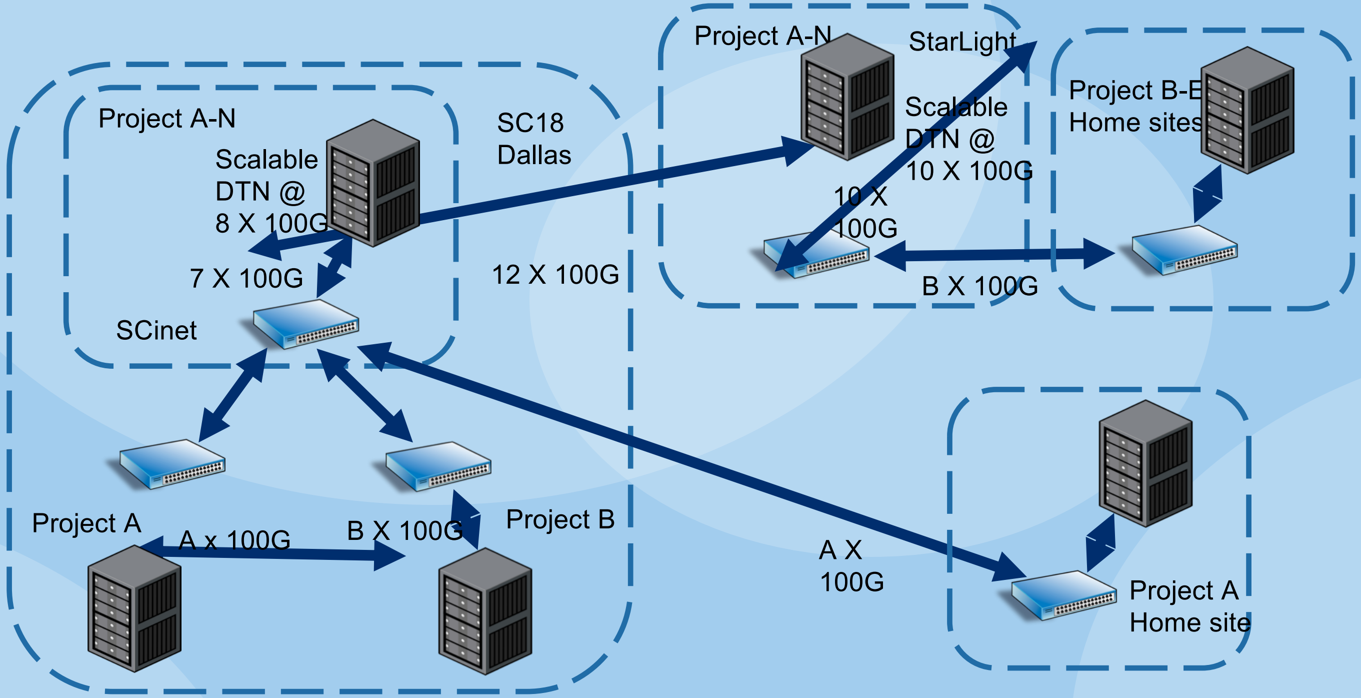
Related & Supported Paper:

1) "Analysis of CPU Pinning and Storage Configuration in 100 Gbps Network Data Transfer" -Se-Young Yu & others.
2) "BigData Express: Toward Schedulable, Predictable, and High-performance Data Transfer" -Wenji Wu & other
3) "Flowzilla: A methodology for Detecting Data Transfer Anomalies in Research Networks." -Anna Giannakou & others
4) And Additional Papers

Issues & Recommendations:

- DTN user cases
- Prepare for 100G network data connectivity end to end tests
- DTN performance tuning over network

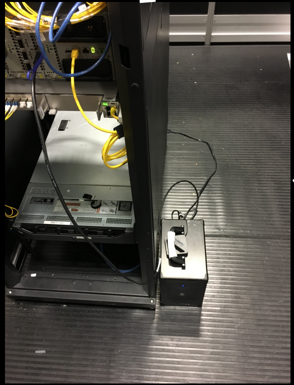
X-NET:
SCinet Data Transfer Node(DTN) Service



SC18 X-NET Faucet and SCinet DTN Team Collaboration: Faucet Demo with 100G DTN Probe in DNOCs

DNOC 2644

Allied Telesis SBx908 GEN2

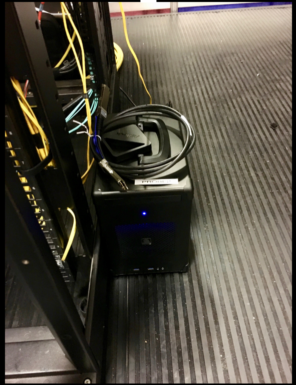


NOC FAUCET

100 Gbps Flows

DNOC 420

Cisco C9500-32C



iCAIR

STARLIGHT℠

SC18 X-NET Faucet and SCinet DTN Team Collaboration: Faucet Demo with 100G DTN Probe in DNOCs

```
dnoc2644-faucet>sho int port1.8.1 | inc rate
    input average rate : 30 seconds 211.46 Mbps, 5 m
    output average rate: 30 seconds 53.28 Gbps, 5 mi
    input peak rate 444.78 Mbps at 2018/11/10 16:41:
    output peak rate 103.24 Gbps at 2018/11/10 16:42
```

Please Visit Booth 2851 For test configuration

Conclusions

- Storage systems are still common bottlenecks in 100 Gbp disk-to-disk transfer
- CPU affinity settings in NUMA can reduce processor overhead
  - NB: Actual affinity setting may differ
- Software RAID for NVMe may not work well in 100 Gbps for now
- NVMe Over Fabrics has less overhead in 100 Gbps network transfers

Future Work

- **Disk-to-disk transfers Across WANs**
  - 100 Gbps DTNs with NVMe are now available
  - RoCE v2 requires very clean path
    - Congestion control based on ECN (like DCTCP)
  - NVMe over Fabrics TCP
  - Hardware NVMe RAID
- **Apply to the following projects**
  - Starlight SDX DTNs
  - SCinet DTN
  - Chameleon Large Flow Appliance
  - Investigation of "intelligence/automation" techniques

www.startap.net/starlight

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

iCAIR

STARLIGHT[SM]

Thank you

Any Question?