DTN-as-a-Service and AIDTN Support For 400G WAN

Jim Chen

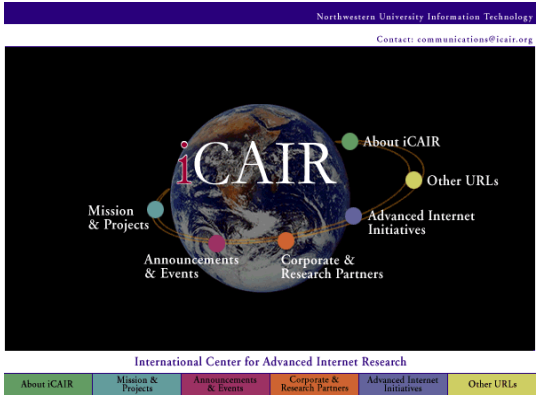
5GRP, September 16 2024, Osaka Japan

StarLight International National Communications Exchange Facility
International Center for Advanced Internet Research
Northwestern University

iCAIR

STARLIGHT SDX

International Center for Advanced Internet Research (iCAIR) and StarLight



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

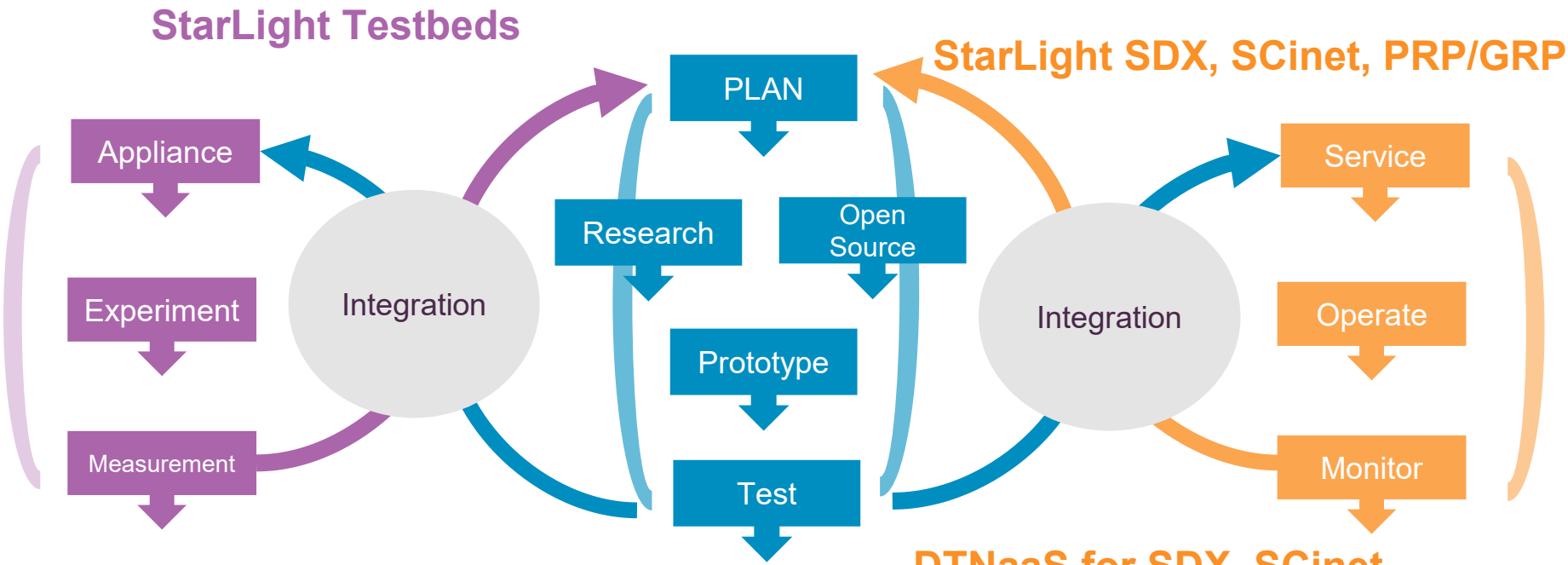
StarLight –
"By Researchers
For Researchers"

- Creation and Early Implementation of Advanced Networking Technologies - The Next Generation Internet All Optical Networks, Terascale Networks, Networks for Petascale Science
- *Advanced Applications, Middleware, Large-Scale Infrastructure, NG Optical Networks and Testbeds, Public Policy Studies and Forums Related to NG Networks*
- *Three* Major Areas of Activity: a) Basic Research b) Design and Implementation of Testbeds and Prototypes c) Operations of Specialized Communication Facilities (e.g., StarLight)
- Supporting Over 112 100 Gbps Paths, 30 400 Gbps Paths

iCAIR

STARLIGHT SDX

StarLight Software Defined Exchange (SDX) CD/CI/CD Innovation Workflow



iCAIR

STARLIGHT SDX

iCAIR/StarLight Research Collaborations

-DTN and Data Movement

"AIDTN: Towards a Real-Time AI Optimized DTN System With NVMeoF ". IEEE Transactions on Parallel and Distributed Systems. PP. 1-12. 10.1109/TPDS.2023.3260806. Yu, Se-Young & Zeng, Qingyang & Chen, Jim & Chen, Yan & Mambretti, Joe. (June 2023).

"Analysis of NVMe over Fabrics with SCinet DTN-as-a-Service", Cluster Computing, Aug 2022

S. Yu, J. Chen, F. Yeh, J. Mambretti, X. Wang, A. Giannakou, E. Pouyoul, M. Lyonnais

-P4 & Programmable Infra: Testbed, In-Band Network Telemetry (INT) and DDoS Detection

"Sketch-based entropy estimation: a tabular interpolation approach using" The 6th European P4 Workshop (EuroP4'23) Dec 2022, Y. Lai, et al.

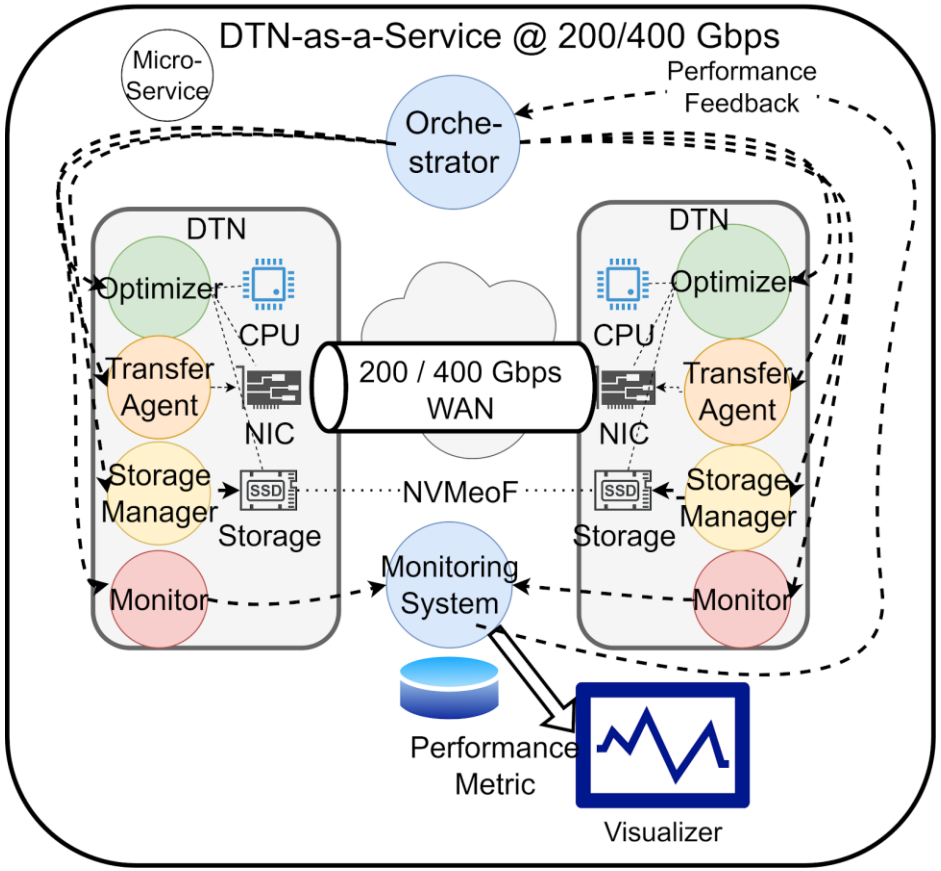
"Tabular Interpolation Approach Based on Stable Random Projection for Estimating Empirical Entropy of High-Speed Network Traffic" IEEE Access, 27 September 2022, Y. Lai, et al.

"P4MT: Designing and Evaluating Multi-Tenant Services for P4 Switches", 2021 The Asia-Pacific Network Operations and Management Symposium (APNOMS), B. Chung, et al.

-Micro Services and Kubernetes

"Automatic Policy Generation for Inter-Service Access Control of Microservices" 30th USENIX Security Symposium/USENIX Security '21, X. Li, et al. Aug 11, 2021.

200/400 Gbps DTN-as-a-Service in High-Performance Research Platform

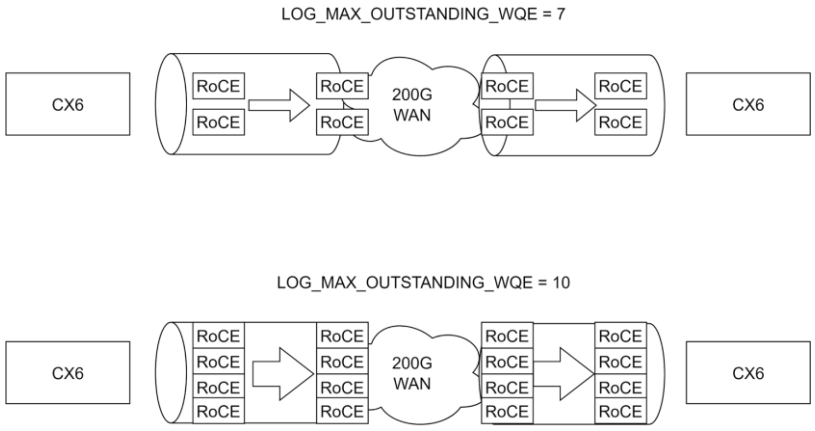


- 200/400 Gbps end-to-end high-performance data transfer over WAN
- DTN-as-a-Service with microservice architecture, optimizing and transferring using containers
- NVMeoF with streaming support
- Performance monitoring and visualization using opensource platforms (Prometheus, Grafana, and sFlow)

iCAIR

StarLight SDX

Nvidia ConnectX-6 Performance Challenge in WAN



- Longer RTT requires a larger buffer for in-flight packets
- RoCE performance suffers from a smaller buffer in long distances, especially for applications with smaller transmit queue (TX)
- ConnectX-5 and 6 define LOG_MAX_OUTSTANDING_WQE for outstanding packets on the wire using host memory
- Increasing it will load more packets into the pipe at the same time. Observed ~1.9x performance increase in 200 Gb with 88ms RTT pipe.
- Yet, this potentially increases pressure on the switch buffer size when used with flow control. Requires deep buffer switch to minimize packet loss.

LOG_MAX_OUTSTANDING_WQE=7

TX_size/ Msg_size	256	512
4M	42.05 Gb	48.15 Gb
8M	69.5 Gb	96.04 Gb

LOG_MAX_OUTSTANDING_WQE=10

TX_size/ Msg_size	256	512
4M	56.34 Gb	149.22 Gb
8M	125.89 Gb	188.5 Gb

iCAIR

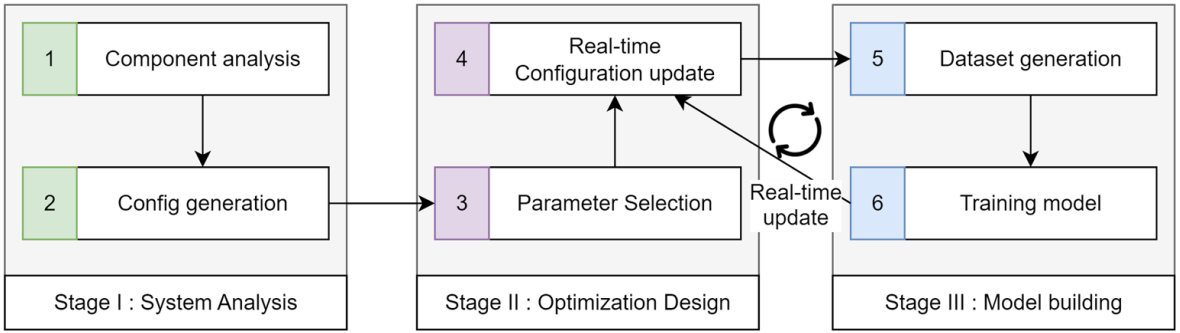
STARLIGHT SDX

AIDTN: Towards a Real-Time AI Optimized DTN System with Classic storage and with NVMeoF

Introduction

1. How can we optimize the end-to-end performance of big data movement in **real-time**?
2. How can we extend DTN performance optimization for **remote data streaming** in addition to traditional data transfer over WAN?
3. How can we deploy an optimization system to **existing research platforms** with managed training data to enable transition-to-practice?

What AIDTN will do?



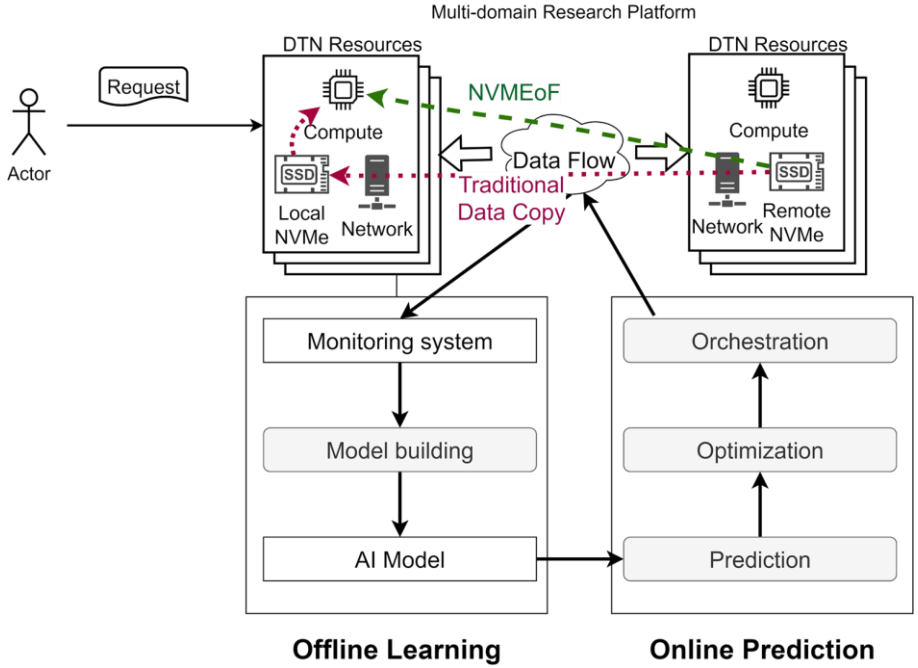
- **Analysing system and generate configs**
- **Real-time parameter adjustment**
- Monitor and generate performance data
- **Exporting data to machine learning system**

- Invoke and generate performance data for data movement (TCP and NVMeoF)
- **Managing historical performance data**
- Visualizing performance
- etc…

iCAIR

STARLIGHT SDX

Real-time optimization with both traditional and NVMeoF

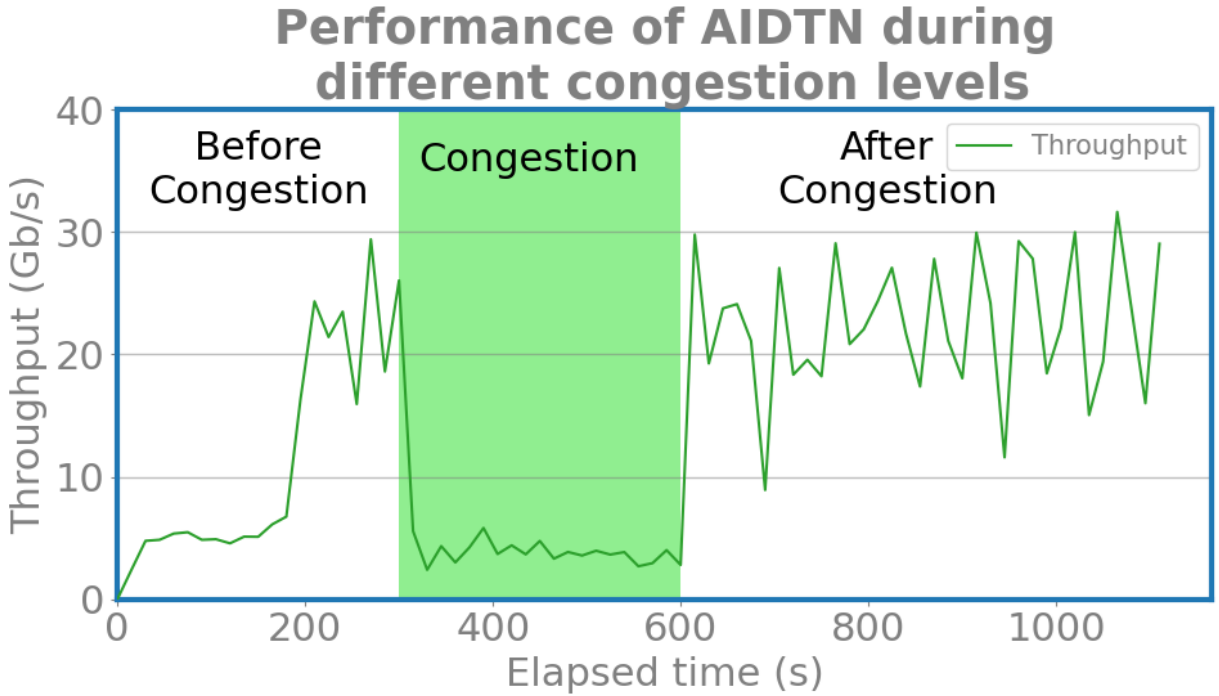


- Network, system, and storage features to improve the performance prediction
- Incorporating NVMeoF specific configurations
- Evaluate AIDTN prototype in MRP and NRP/PRP
- Manage the sparsity of training data

iCAIR

STARLIGHT SDX

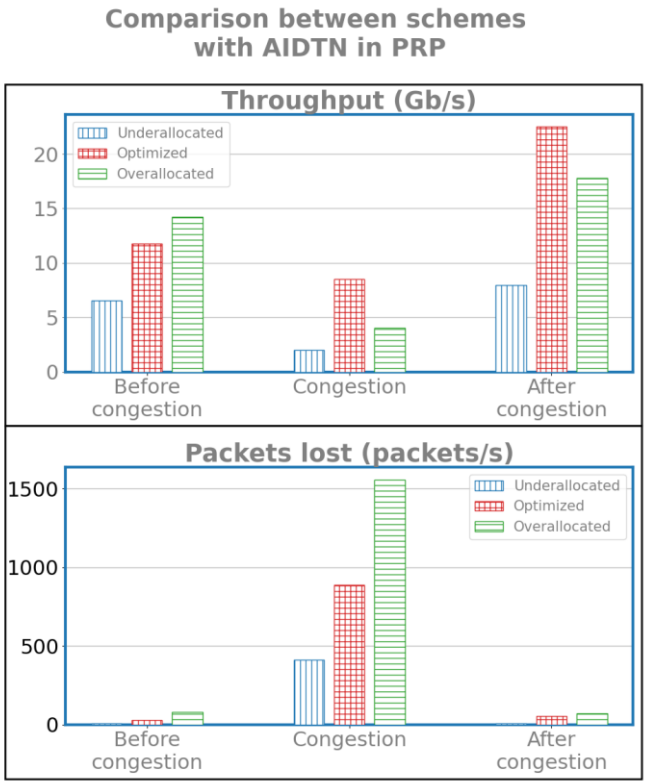
Is real-time adjustment effective?

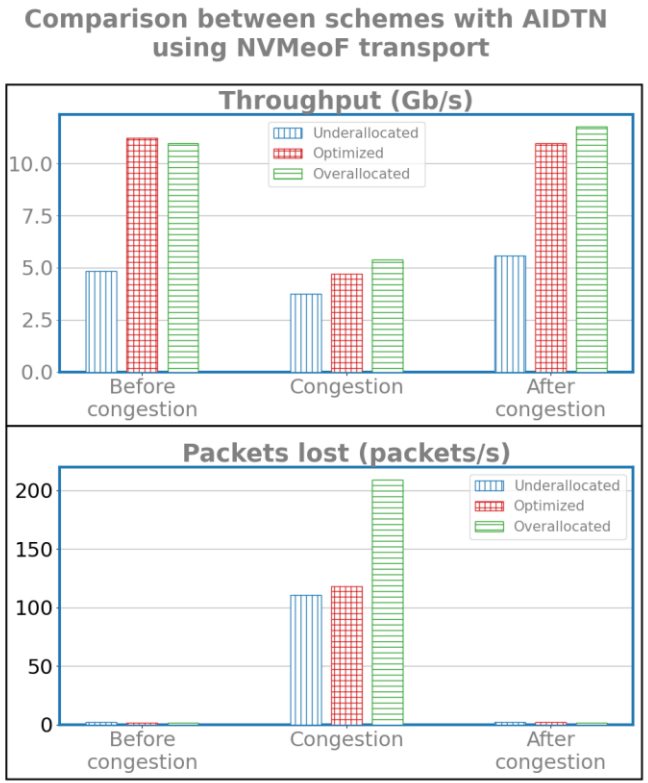


iCAIR

StarLight SDX

Is real-time adjustment effective?



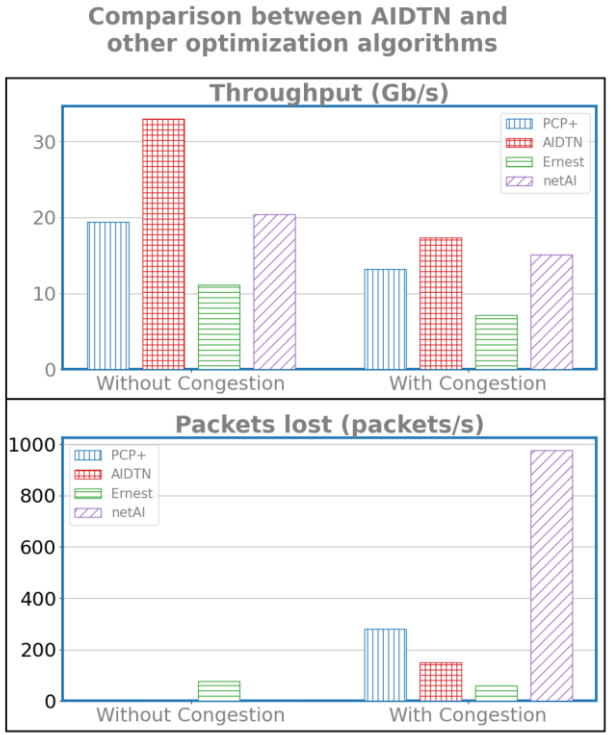


- 60% faster completion time with real-time adjustment
- Significantly less packet losses
- Prediction Error small as RMSE = 0.16

iCAIR

STARLIGHT SDX

Do we have better AI algorithms?



- We use two-step (Historical data learning + real-time data learning) prediction
- BI-LSTM for temporal(real-time) data and XGBoost for historical data
- Provide best performance among other prediction models
- Reduce packet losses drastically when congestion

- Prototype in LLM/SLM(new)

iCAIR

STARLIGHT SDX

400G Data Center Services Over WAN Prototype

-A recent cyber infrastructure innovation is enhanced high efficiency/high performance data transport within data centers.

-iCAIR/StarLight is developing and prototyping extended data center services over WANs – nationally and internationally.

-The recent Nvidia GTC announcement indicated RDMA and RoCE are the primary data network services for AI data centers.

-We are extending data center 400G RoCE data movement services to WAN, including by prototyping RoCE services for data transport.

iCAIR

STARLIGHT SDX

400G Data Center Services Over WAN Prototype

-The prototype service transits 5 different WAN testbed providers:

a. Internet2
b. NA-REX
c. ESnet
d. FABRIC
e. CENI & Partners

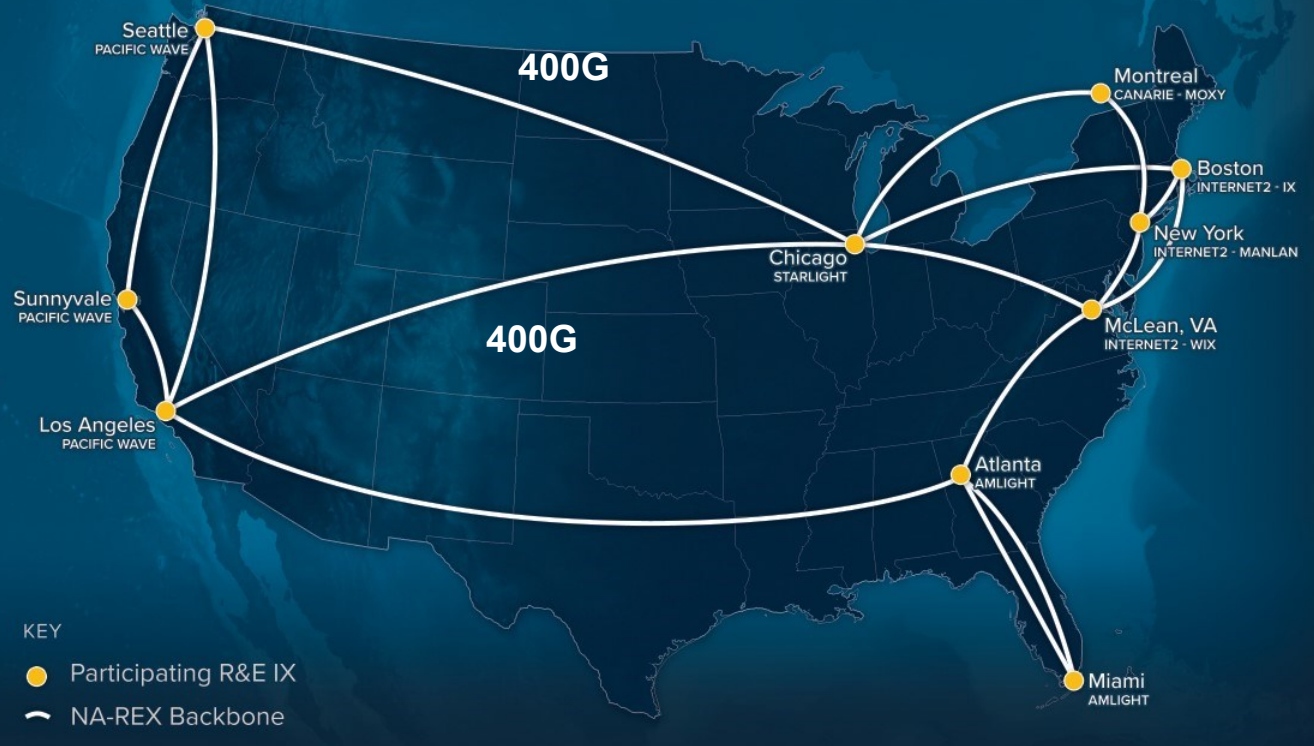
-The DTNs in StarLight for this prototype are highly customized Dell PCI-e Gen5 R760s, each with CX7 single 400G port interfaces.

-The same 2 systems are used for all tests.

-Tuning applied depends on the distance/latency.

NA-REX

North America Research & Education Exchange Collaboration



October 2023

AMPATH PATHWAY TO THE AMERICAS

canarie

ESnet ENERGY SCIENCES NETWORK

INTERNET2

International Networks at Indiana University

STARLIGHT The Optical STAR TAP

CENIC

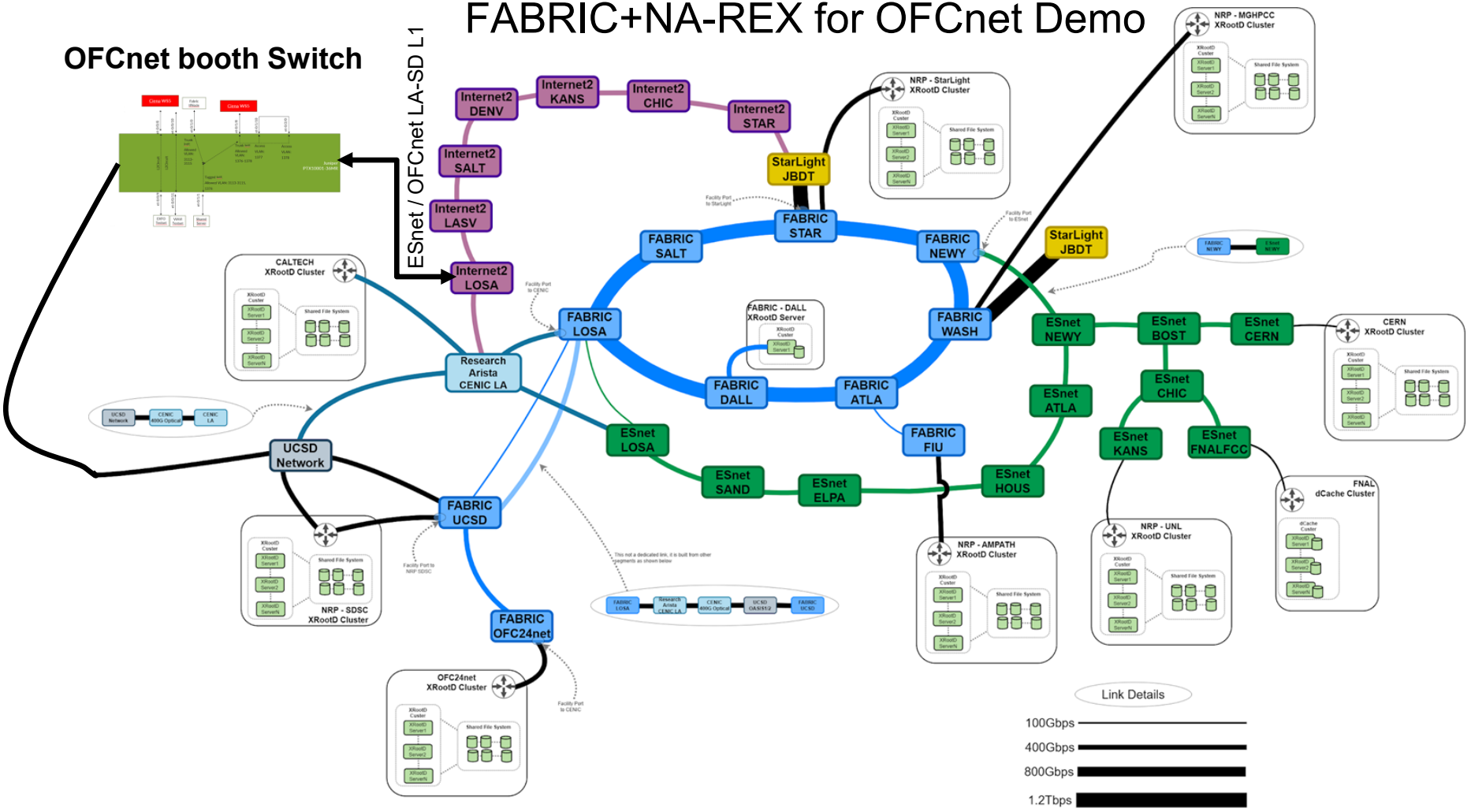
PACIFIC NORTHWEST GIGAPOP

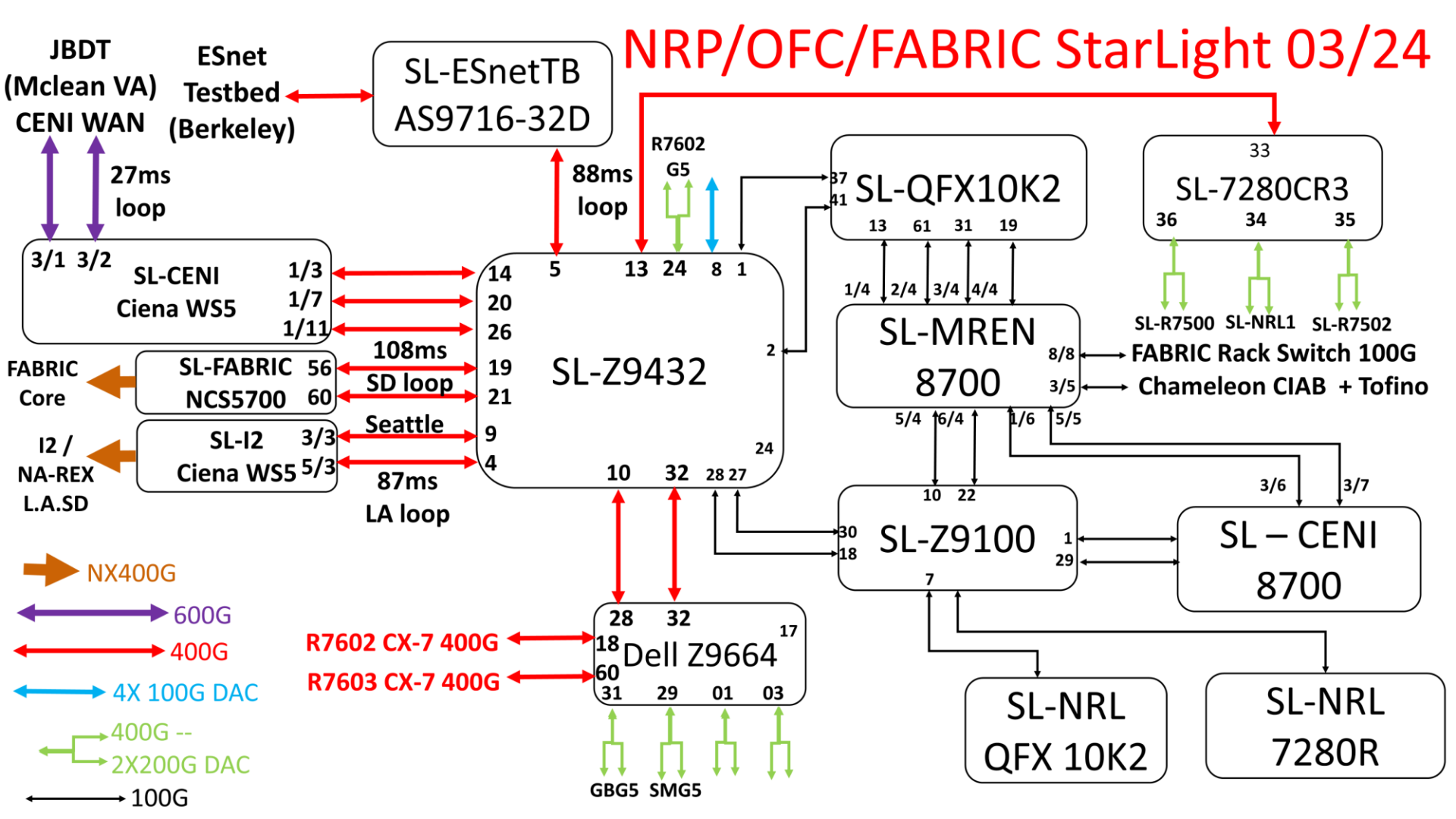
PACIFIC WAVE

CENI TESTBED

Source: Ciena / Scott Kohlert

FABRIC+NA-REX for OFCnet Demo




NRP/OFC/FABRIC StarLight 03/24
JBDT
(Mclean VA)
CENI WAN
ESnet
Testbed
(Berkeley)
SL-ESnetTB
AS9716-32D
27ms
loop
88ms
loop
R7602
G5
SL-QFX10K2
SL-7280CR3
SL-CENI
Ciena WS5
SL-Z9432
SL-MREN
8700
SL-R7500
SL-NRL1
SL-R7502
FABRIC Rack Switch 100G
Chameleon CIAB + Tofino
FABRIC
Core
SL-FABRIC
NCS5700
108ms
SD loop
I2 /
NA-REX
L.A.SD
SL-I2
Ciena WS5
Seattle
87ms
LA loop
SL-Z9100
SL – CENI
8700
R7602 CX-7 400G
R7603 CX-7 400G
Dell Z9664
SL-NRL
QFX 10K2
SL-NRL
7280R
GBG5
SMG5
NX400G
600G
400G
4X 100G DAC
400G --
2X200G DAC
100G

Extend Data Center Services Over 400G WAN

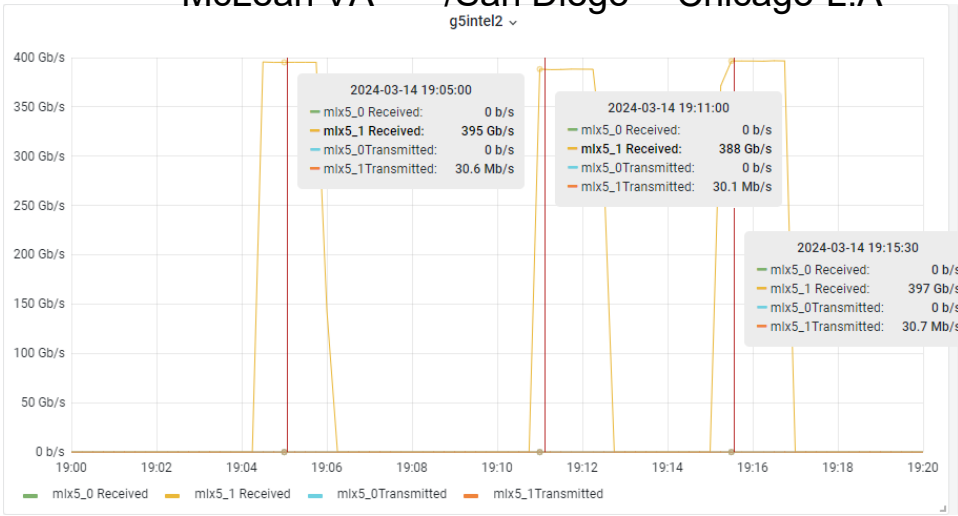
Prototype Solution Initial Results:

Single stream RDMA/RoCE over 400G network at different distance

(1) CENI Chicago - McLean VA

(2) NA-REX Chicago -L.A /San Diego

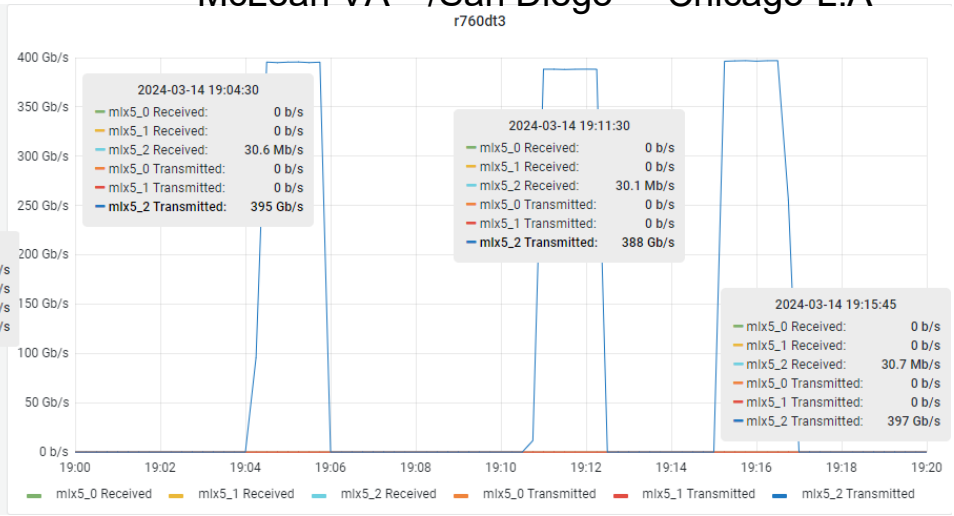
(3) FABRIC + NA-REX Chicago-L.A



(1) CENI Chicago - McLean VA

(2) NA-REX Chicago -L.A /San Diego

(3) FABRIC + NA-REX Chicago-L.A



SL loopbacks: (1) Rtt 27 ms @ **395G** (2) Rtt87 ms @ **388G** (3) Rtt 108 ms @ **397G**

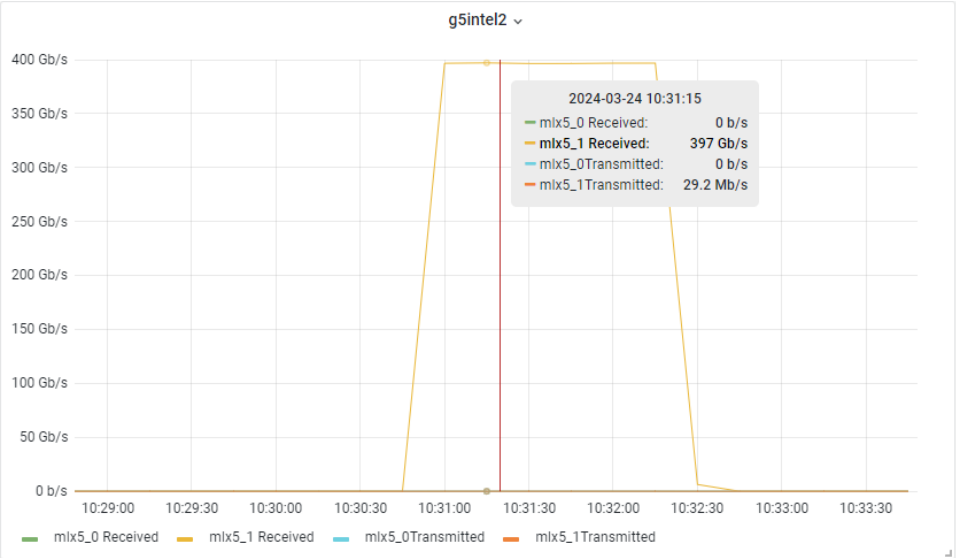
iCAIR

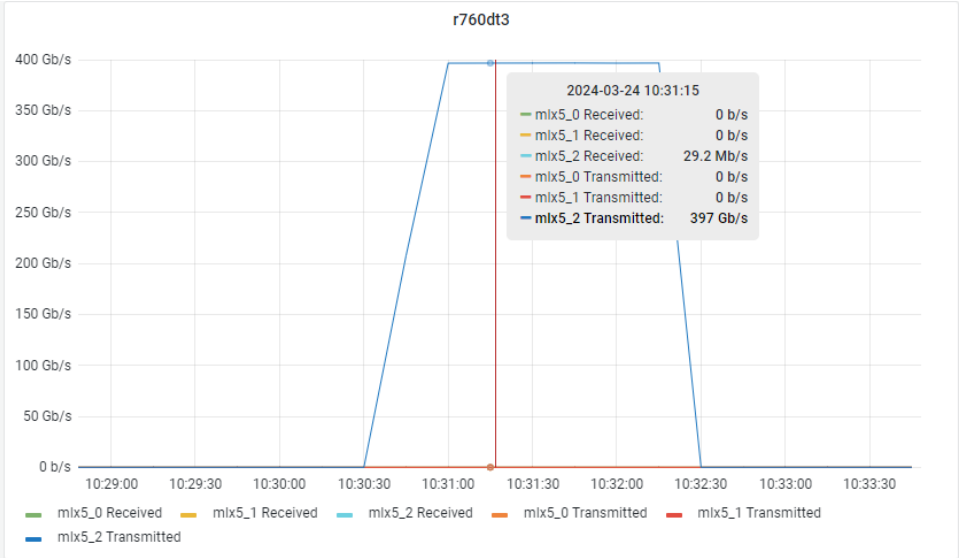
STARLIGHT SDX

Extend Data Center Services Over 400G WAN

Prototype Solution Initial Results:

Single Stream RDMA/RoCE Over 400G Network





Chicago-San Diego OFCnet loopback: Rtt 96.4 ms, Peak @ 397G X 2

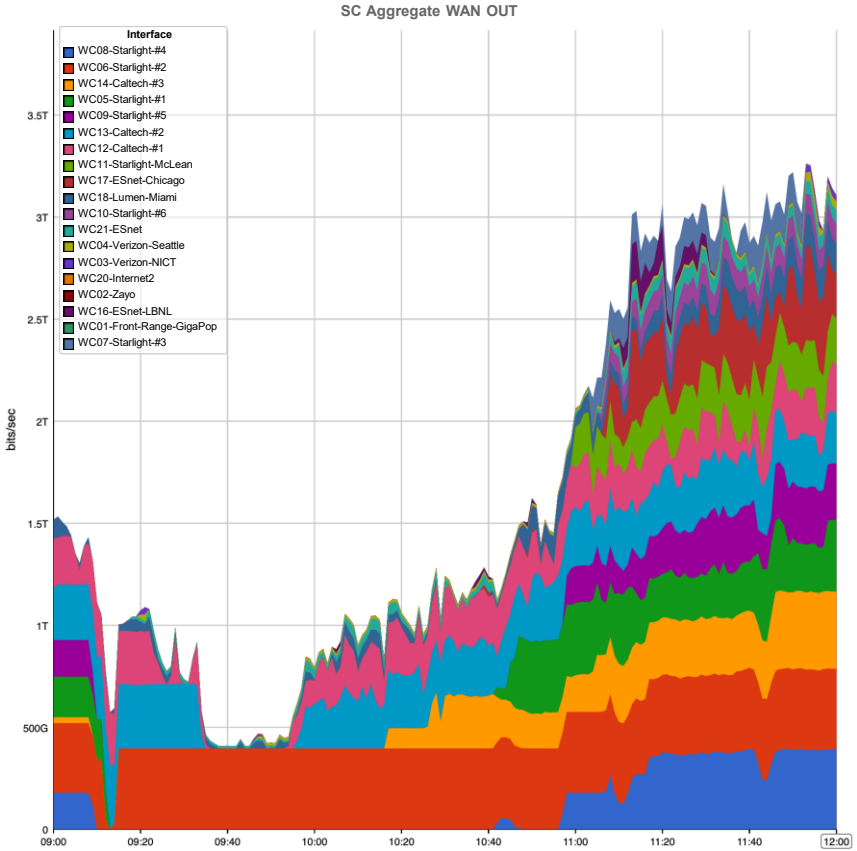
iCAIR

March 24, 2024

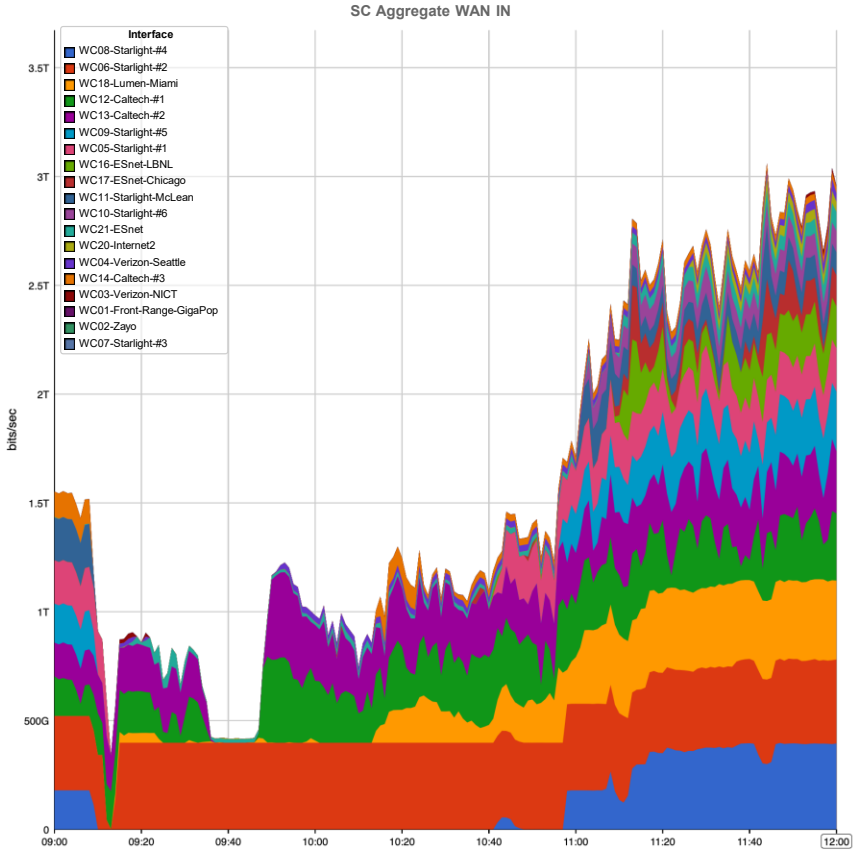
STARLIGHT SDX

SC23 Bandwidth Challenge

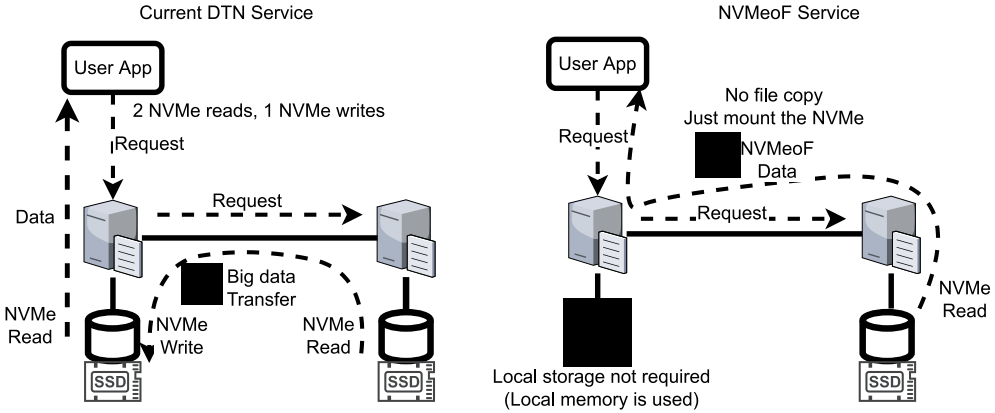
StarLight contributes 4 of Top 5

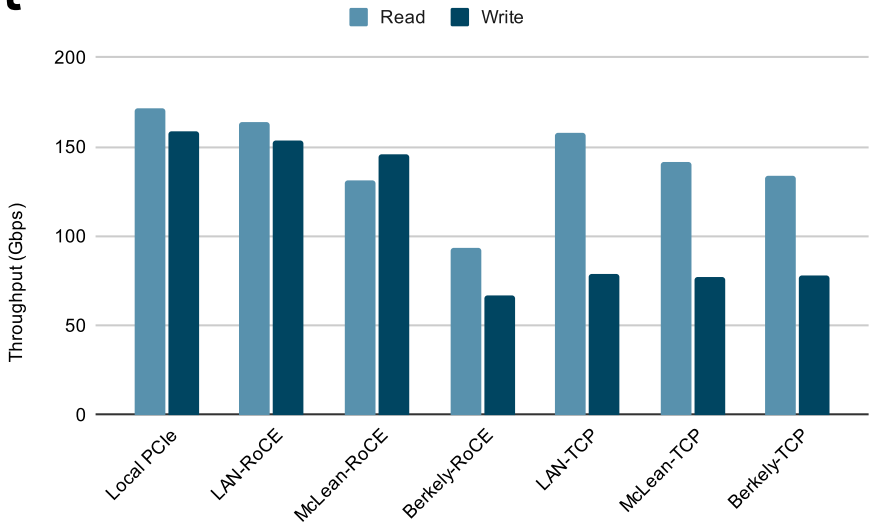


StarLight contributes 2 of Top 5



High-performance NVMe-over-Fabrics with SmartNICs for data transport



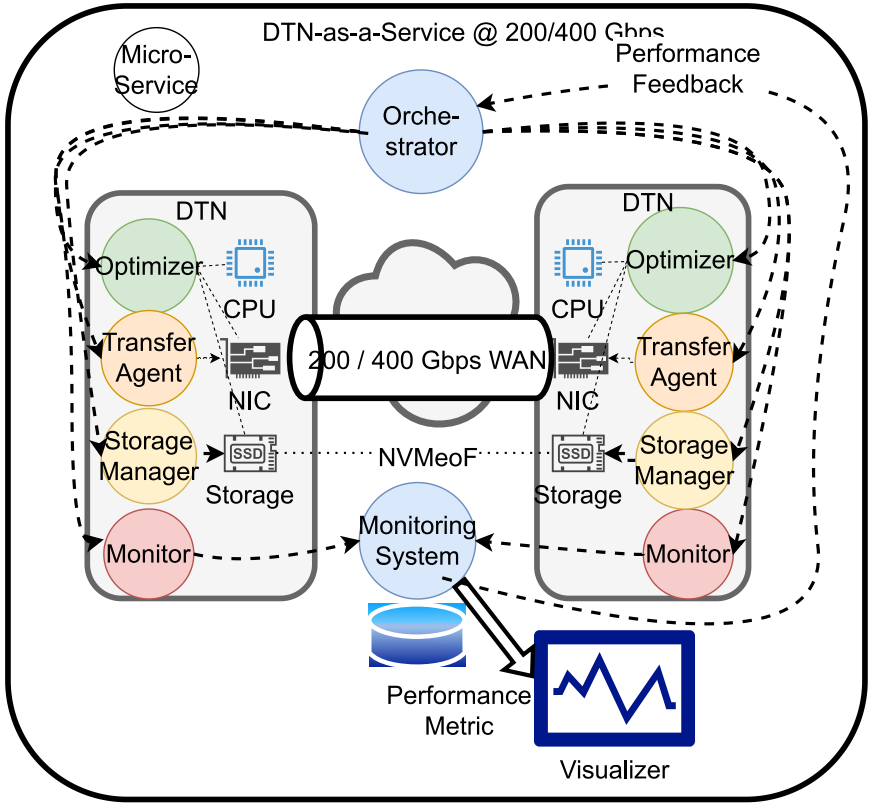


- Low overhead NVMeoF data streaming over RoCE and TCP
- High-performance data streaming over 130 Gbps using TCP (~ 88ms)
- High-performance remote data writing over 140 Gbps using RoCEv2 (~22ms)

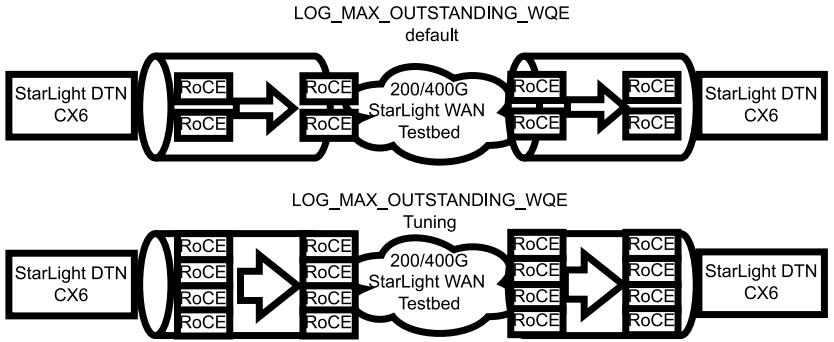
iCAIR

STARLIGHT SDX

DTN-as-a-Service with K8s integration for high-performance data transport



RoCE optimization for 200/400G StarLight WAN



- Implementing Cloud-native service for data transport
- Service optimization for Cloud environment with 200/400G WAN
- Jupyter controller for Easy-of-use

SC24 ATLANTA NOV 17-22

scitags.org

Flow and Packet Marking for Global Scientific Computing

DTN-KIT-100g

KIT Karlsruher Institut für Technologie

DFN DEUTSCHES FORSCHUNGSNETZ

DTN-UVIC-100g

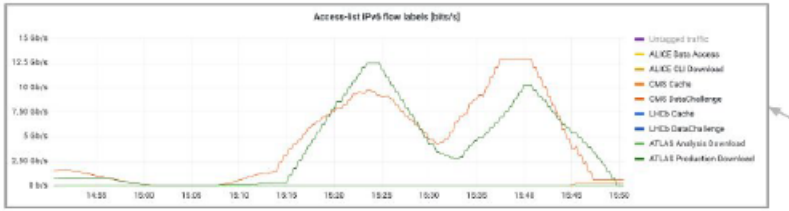
University of Victoria

1.Clients requesting data transfers from/to DTN-SC22-400g while passing science domain and activity fields via transfer protocols.

ESnet

canarie

4.High performance tests using eBPF-TC filters to test encoding of the science domains and activity fields in the IPv6 flow label at scale.



3. P4 programmable switch at CERN collecting the science domains and activity bits encoded in the packets.

CERN-LHCONE

P4 EdgeCore Wedge

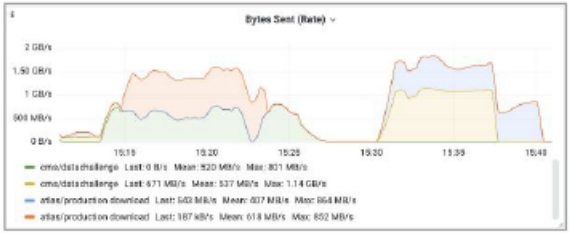
CERN

DTN-CERN-100g
2.6GHz/32 cores
SSD, 100Gbps

CERN

DTN-SC22-400g
R7503 2.6 GHz
NVMe 2.0
2x200 Gbps

XRootD



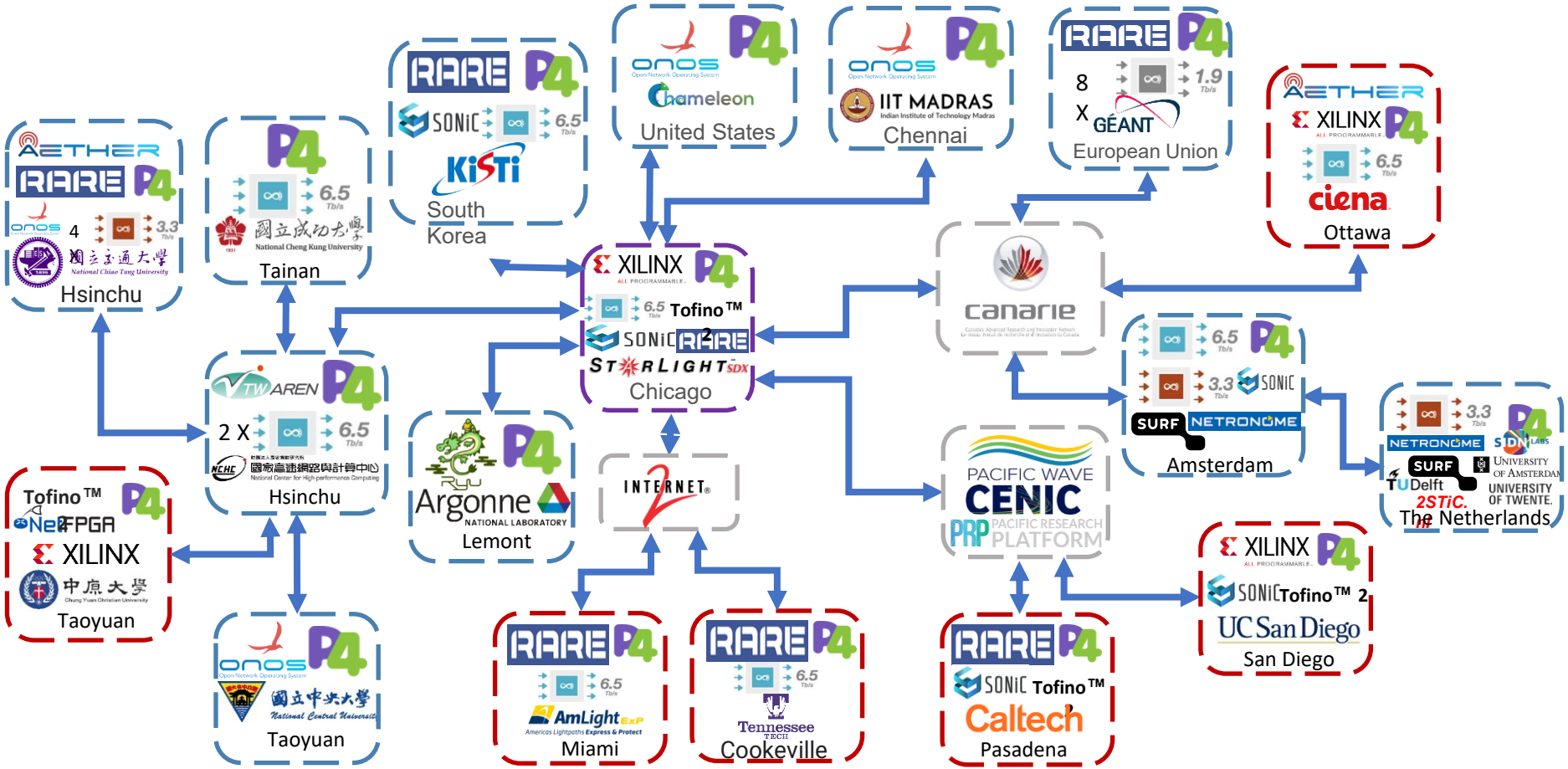
2. XRootD storage responds to the client requests and marks the data transfer packets with the corresponding science domain and activity.

5.Sampling of the low level TCP/IP metrics, which can be used by sites and R&Es to better understand the scientific flows.

200 Gbps
100 Gbps
50 Gbps

SC24 STARLIGHT

International P4 Experimental Networks (iP4EN)



Global Research Platform (GRP) was established to create a worldwide Science DMZ, a distributed environment for data-intensive research, particularly advanced networking capabilities for large-scale data transfers that builds on a variety of testbed activities being developed by global science partners. Innovative concepts for advanced architectures, services, and technologies, when proven through experiments, trials, and demonstrations on research testbeds, at scale, are migrated to production environments.

iCAIR

STARLIGHT SDX

SC24 Atlanta, GA hpc creates.

NRE 111

SC24 Network Research Exhibition: Demonstration Preliminary Abstract

StarLight DTN-as-a-Service and SciTag Prototype for High-Performance Data Transport with Research Platforms

Jim Chen, Ting-Yu Lin, Fei Yeh, Joe Mambretti
International Center for Advanced Internet Research - Northwestern University, jim-chen, tingyu.lin, fyeh, j-mambretti@northwestern.edu

Se-Young Yu, Kiran Vasu, ESnet, youf3, kvasu@es.net

AND PARTNERS

iCAIR

SC24 Atlanta, GA hpc creates.

NRE 118

SC24 Network Research Exhibition: Demonstration Preliminary Abstract

International P4 Experimental Networks for The Global Research Platform and Other Research Platforms

Jim Chen, Ting-Yu Lin, Fei Yeh, Joe Mambretti
International Center for Advanced Internet Research - Northwestern University, jim-chen, tingyu.lin, fyeh, j-mambretti@northwestern.edu

AND PARTNERS

STARLIGHT SDX

Dr. Yu-kuen Lai

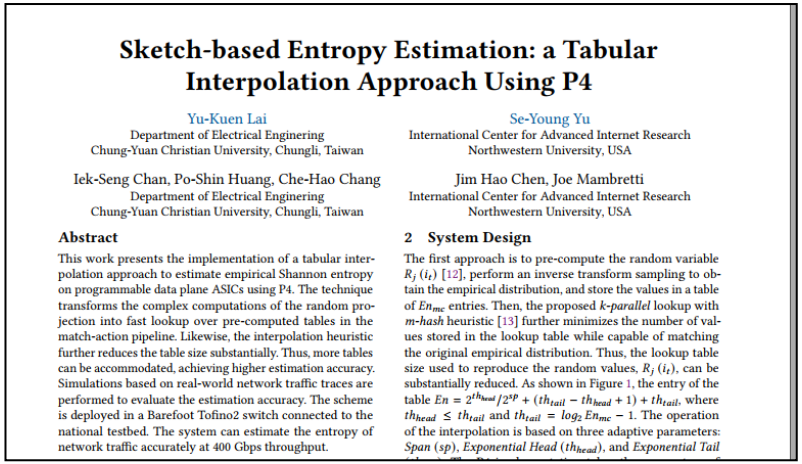
- **Professor**
  - CYCU Electrical Engineering Department
  - IEEE Senior Member
  - CYCU Distinguished Teaching Award
  - CYCU Excellence Award of Academia and Industry Collaboration
- **Director**
  - ✓ Computer Networks and Systems Research Laboratory
  - ✓ Interdisciplinary Program of Electrical and Computer Engineering
- **Education**

  M.S. & Ph. D in Electrical and Computer Engineering,
  North Carolina State University, USA
- **Selected Publications**

•Yu-Kuen Lai, Mnh-Hung Nguyen, "A Real-Time DDoS Attack Detection and Classification System Using Hierarchical Temporal Memory", APSIPA Transactions on Signal and Information Processing:Vol. 12: No. 2, e8. , Special Issue on Learning, Security, AIoT for Emerging Communication/Networking Systems., Apr. 3, 2023. http://dx.doi.org/10.1561/116.00000147
•Yu-Kuen Lai, C. -L. Tsai, C. -H. Chuang, X. -W. Ku and J. H. Chen, "Tabular Interpolation Approach Based on Stable Random Projection for Estimating Empirical Entropy of High-Speed Network Traffic," in IEEE Access, vol. 10, pp. 104934-104953, 2022, doi: 10.1109/ACCESS.2022.3210336.
•Yu-Kuen Lai, Se-Young Yu, Iek-Seng Chan, Po-Shin Huang, Che-Hao Chang, Jim Hao Chen, Joe Mambretti, "Sketch-based Entropy Estimation: a Tabular Interpolation Approach Using P4", 5th European P4 Workshop (EuroP4), December 6-9, 2022, Rome, Italy.

Shannon Entropy Estimation: A Tabular Interpolation Approach

Y. -K. Lai, C. -L. Tsai, C. -H. Chuang, X. -W. Ku and J. H. Chen, "Tabular Interpolation Approach Based on Stable Random Projection for Estimating Empirical Entropy of High-Speed Network Traffic," in IEEE Access, vol. 10, pp. 104934-104953, 2022, doi: 10.1109/ACCESS.2022.3210336.



Sketch-based Entropy Estimation: a Tabular Interpolation Approach Using P4

Yu-Kuen Lai
Department of Electrical Enginering
Chung-Yuan Christian University, Chungli, Taiwan

Se-Young Yu
International Center for Advanced Internet Research
Northwestern University, USA

Iek-Seng Chan, Po-Shin Huang, Che-Hao Chang
Department of Electrical Enginering
Chung-Yuan Christian University, Chungli, Taiwan

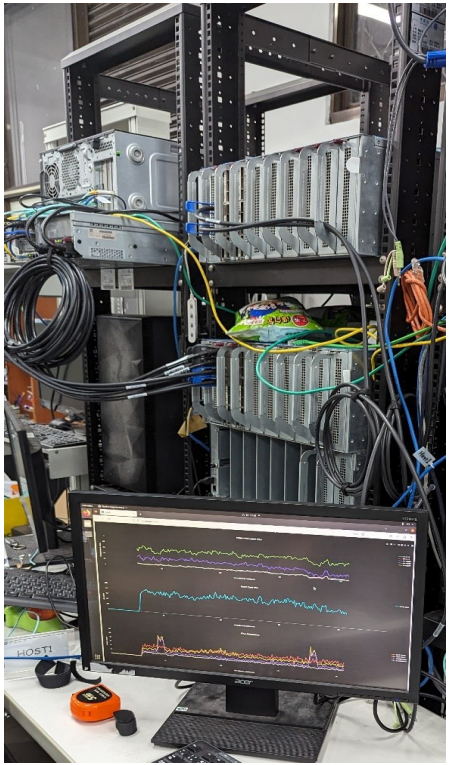
Jim Hao Chen, Joe Mambretti
International Center for Advanced Internet Research
Northwestern University, USA

Abstract

This work presents the implementation of a tabular interpolation approach to estimate empirical Shannon entropy on programmable data plane ASICs using P4. The technique transforms the complex computations of the random projection into fast lookup over pre-computed tables in the match-action pipeline. Likewise, the interpolation heuristic further reduces the table size substantially. Thus, more tables can be accommodated, achieving higher estimation accuracy. Simulations based on real-world network traffic traces are performed to evaluate the estimation accuracy. The scheme is deployed in a Barefoot Tofino2 switch connected to the national testbed. The system can estimate the entropy of network traffic accurately at 400 Gbps throughput.

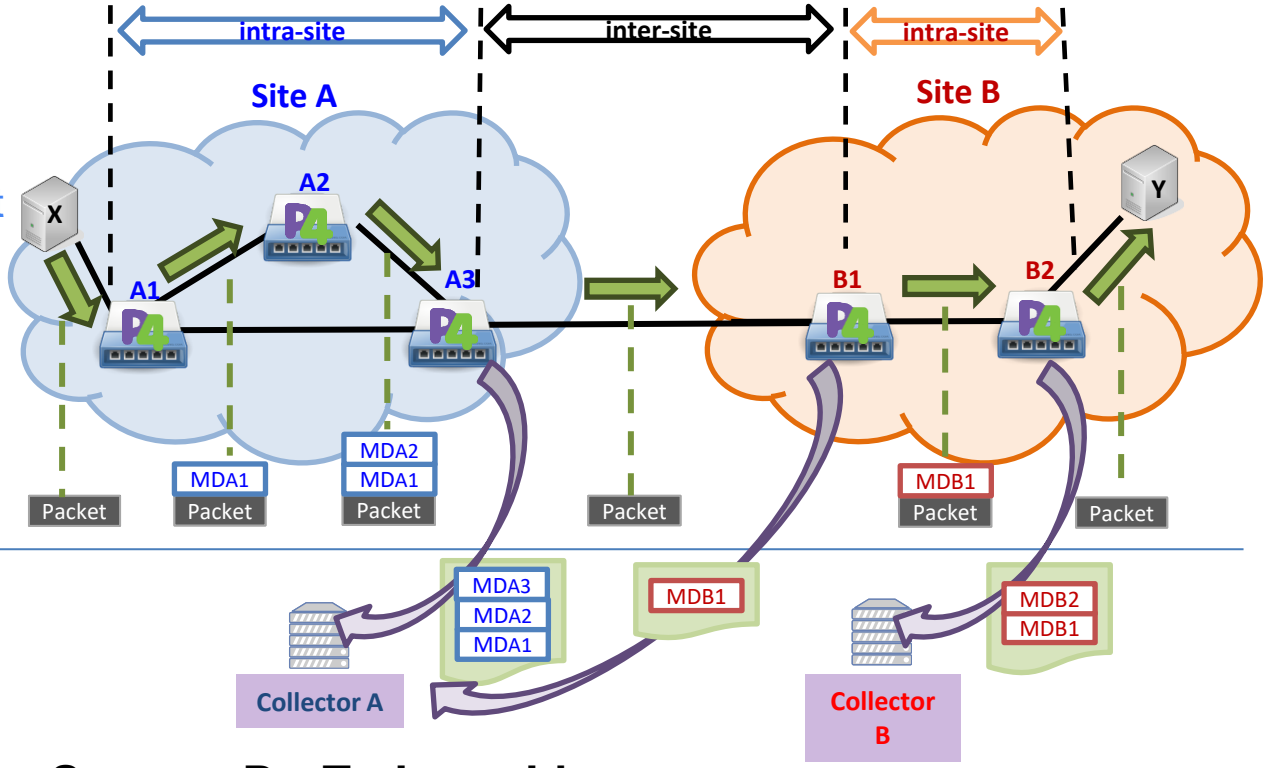
2 System Design

The first approach is to pre-compute the random variable $R_j(i_t)$ [12], perform an inverse transform sampling to obtain the empirical distribution, and store the values in a table of En_{mc} entries. Then, the proposed *k-parallel* lookup with *m-hash* heuristic [13] further minimizes the number of values stored in the lookup table while capable of matching the original empirical distribution. Thus, the lookup table size used to reproduce the random values, $R_j(i_t)$, can be substantially reduced. As shown in Figure 1, the entry of the table $En = 2^{th_{head}}/2^{sp} + (th_{tail} - th_{head} + 1) + th_{tail}$, where $th_{head} \leq th_{tail}$ and $th_{tail} = log_2 En_{mc} - 1$. The operation of the interpolation is based on three adaptive parameters: *Span* (sp), *Exponential Head* (th_{head}), and *Exponential Tail*



Cross-Site Network Telemetry based on Programmable Network Technology

- For inter-site network monitoring, we could analysis data from INT collectors of the sits that the packet traversed.
- However, inter-site links are not monitored by any INT collector
- We propose that site collector collects its intra-site and outing inter-site links INT status
- By integrating the data from all INT collectors, we can depict the whole network status in realtime.
- We will utilize our testbed with international partners to develop and evaluate the proposed solution.



Source: Dr. Te-Lung Liu

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

Questions?

jim-chen@northwestern.edu

iCAIR

STARLIGHT SDX