SmartNICs--What's Working and What's Next? ESnet, CERN Successes; AutoGOLE/SENSE Orchestration

Sixth National Research Platform (6NRP) Workshop
La Jolla, California USA
January 30, 2025, 2pm PT

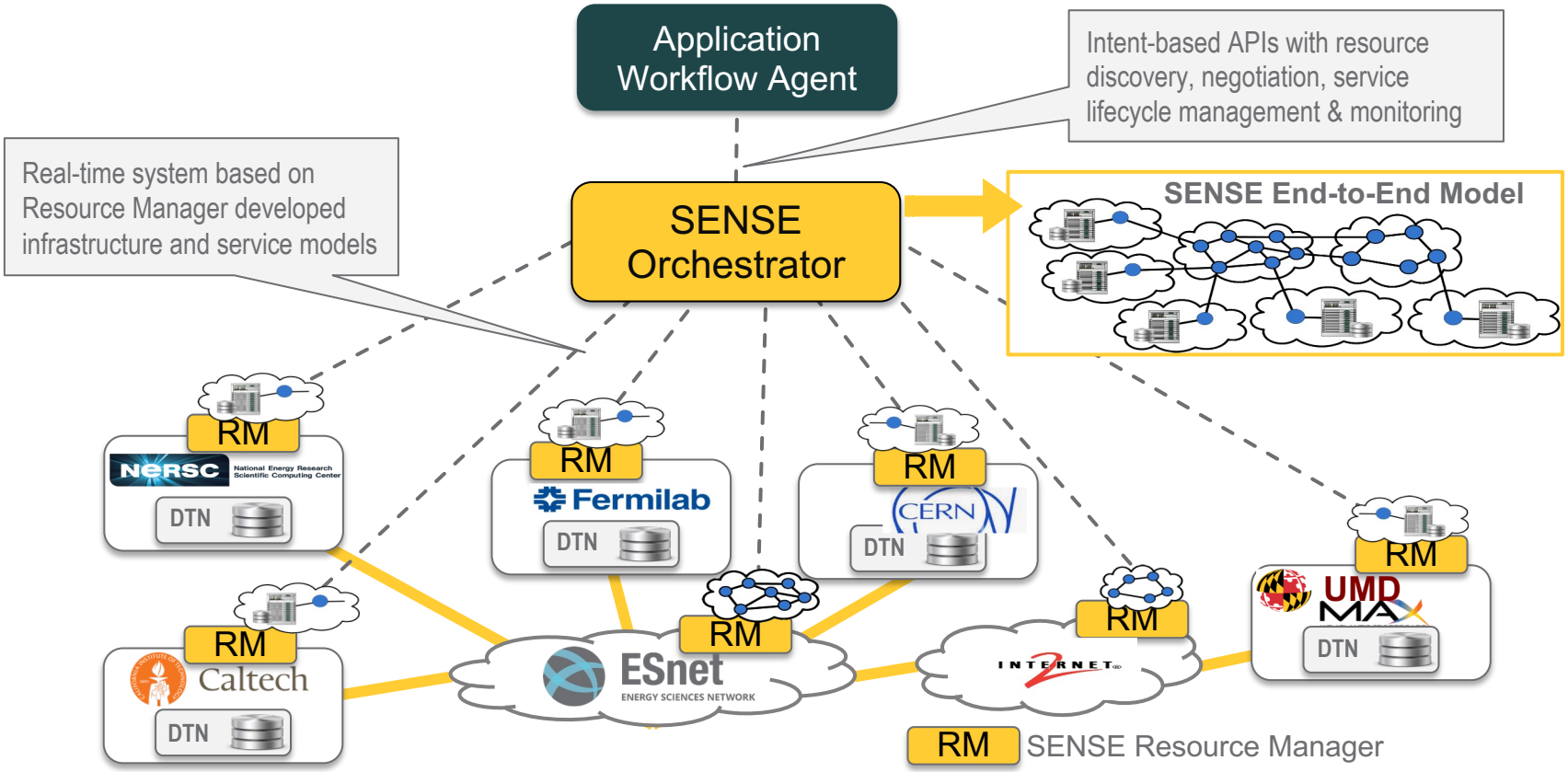
Moderator: Tom Lehman, ESnet

Panelist:

- Justas Balcas, ESnet
- Joe Mambretti, Northwestern University
- Harvey Newman, Caltech
- Mohammad Sada, SDSC, UC San Diego

Justas Balcas Slides

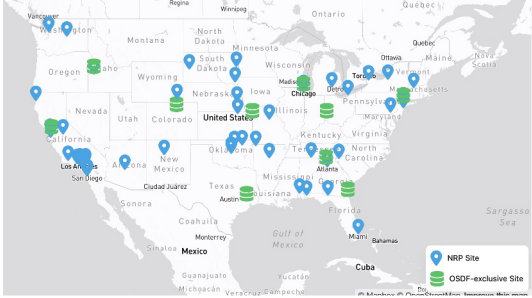
The SENSE Architecture



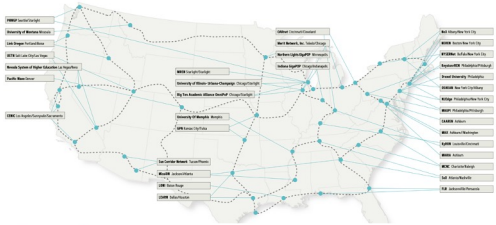
ESnet

Network Control that makes SENSE

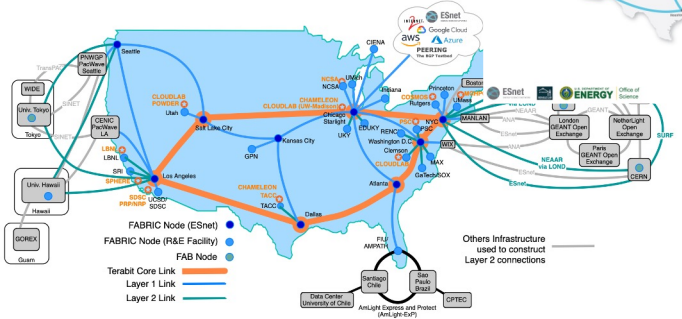
- **82 Servers (15 of them on NRP), 22 Sites, 20 Network Domains (ESnet, Internet2, CENIC, Ampath, Geant, PacificWave, Surf, Kreonet, HEAnet)**
- **Empowering NRP/Sites/Users with simplified network automation over many NRENs:**
  - Resource Managers are adaptable based on NRENs Requirements: links, cli, netconf, restconf, bandwidth guarantees.
  - Standardized configurations using custom Ansible modules for multi-vendor ecosystems (Dell OS9/10, Arista, SONiC, FreeRTR, Cisco Nexus, Juniper, etc.).
  - Supports VLAN translation, BGP control, IPv4/IPv6 assignments, ping, traceroute.
  - Intent-based APIs enable efficient resource management and service lifecycle control.
  - NRP Kubernetes Operator - see Mohammads slides
  - LHC (ATLAS+CMS) use cases for elephant flows.



Internet2 Network Connections
internet2.edu/connectors 7.2024

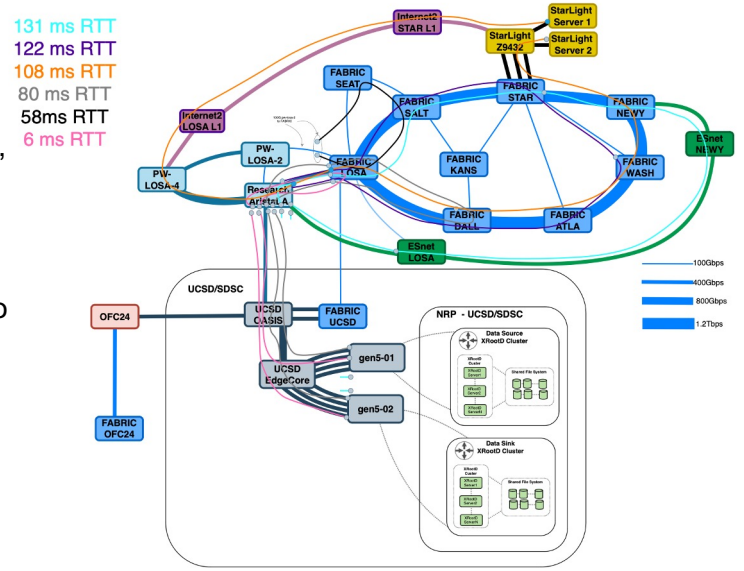


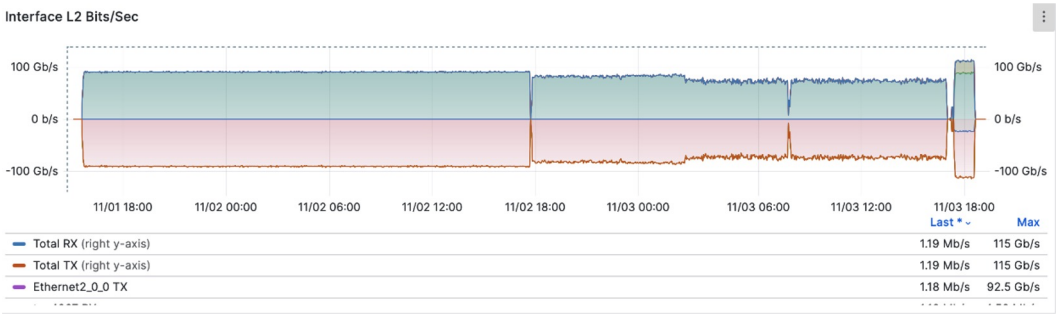
ESnet6



NRP + SENSE + SmartNICs + Fabric

- **NRP Node at CERN connected to VPP/FRR on Fabric (L2/L3 Control)**
- **Use DPDK/VPP on Fabric:**
  - Offloading big flows directly to NICs reduces reliance on traditional switches.
  - Supports advanced offloads like VLAN insert/strip, TCP/UDP checksum, and large receive offload (LRO).
- **Benefits:**
  - Utilizing DPDK and VPP for high-speed, low-latency data processing:
    - Up to **40% lower latency** and enhanced throughput compared to kernel routing.
    - All Runs in containers
  - Happy Network Engineers (no need access to Switch/Router)
- **Future Work**
  - Researching memory/core requirements, advanced offload optimizations, Calico VPP
  - Bluefield-3 and run VPP/DPDK/Router on the NIC

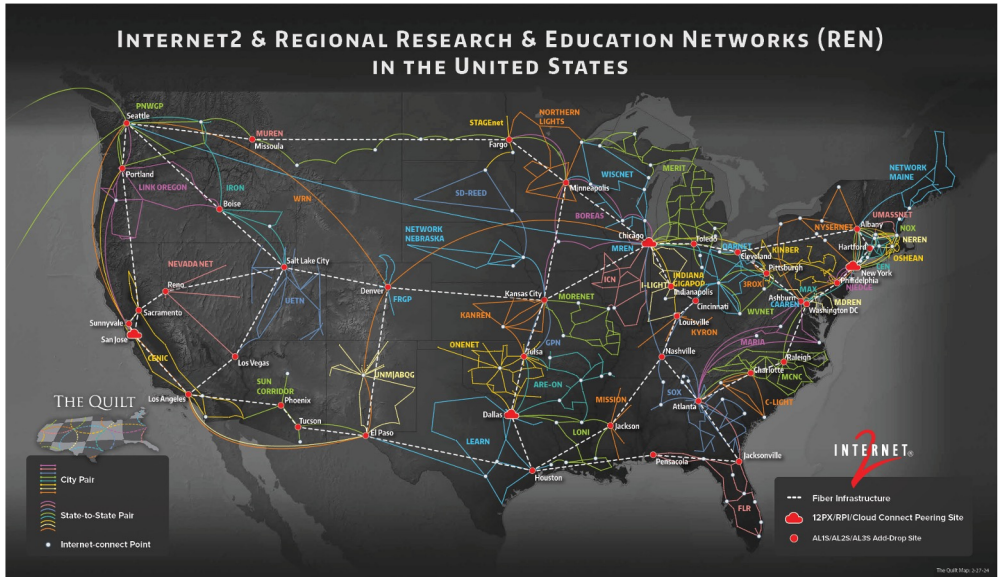




Looking forward

- SENSE and NRP Expansion (Internet2, NYCERnet, MGHPCC planned already)
- OFC25, FABRIC Webinar, FABRIC KNIT10 Demos for **Advanced Networked Services for Domain Science Workflow Innovation** (FABRIC, UCSD/SDSC, ESnet, Ciena, NRP).

My dream is to have coverage over most NRENs and provide seamless request for bandwidth guarantees for users!



Joe Mambretti Slides

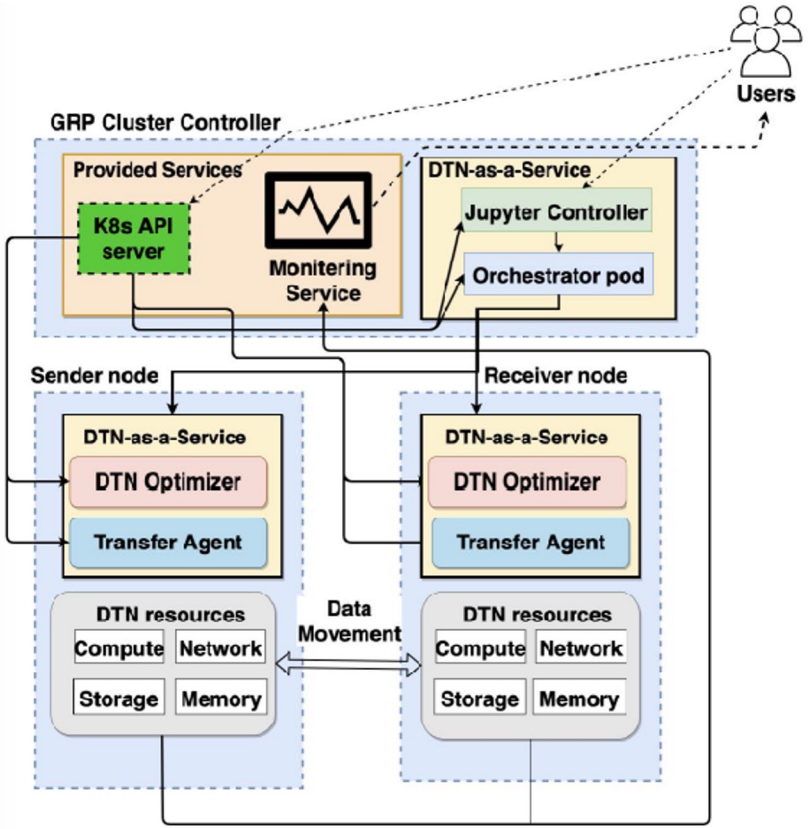
GRP Cluster with DTN-as-a-Service

DTN-as-a-Service(DTNaaS) provides a data movement workflow in GRP k8s cluster:

1. Deploy DTNaaS workloads via k8s API server
2. Use Jupyter to optimize and run transfers
3. Observe performance from monitoring service

GRP DTNaaS Components:

- Orchestrator: controller of DTNaaS to manage agent and optimizer pods via REST API.
- Transfer Agent: run transfer jobs
- DTN Optimizer: optimize the DTN resources for workflow
- Jupyter: web interface to run DTNaaS interactively



Smart NICs + DTNs

Building Blocks For 400G/800G/Tbps WANs

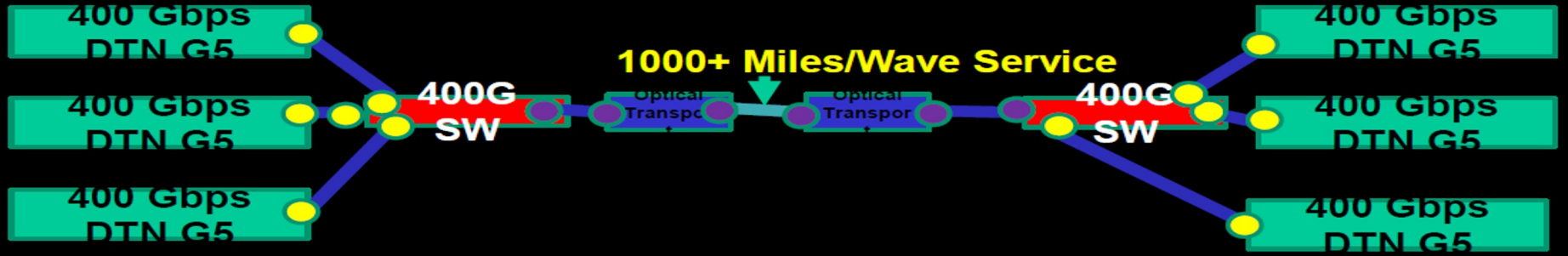
iCAIR

STARLIGHT SDX

1.2 Tbps WAN Service Prototype for Data Intensive Science

StarLight International/National Communications Exchange Facility, Chicago, Il

Joint Big Data Testbed McLean, Va



LR4 Transceiver + Smart NIC

X Transceiver

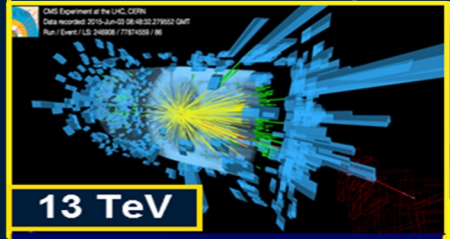
High Capacity WAN Services, Trafic Mangement, In-Band Workflow Pipelining, etc

STARLIGHT℠

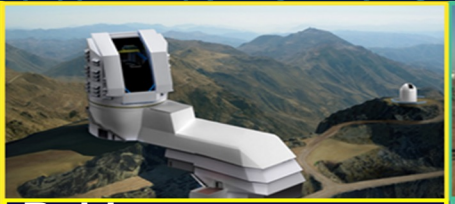
Harvey Newman Slides

Global Network Advancement Group

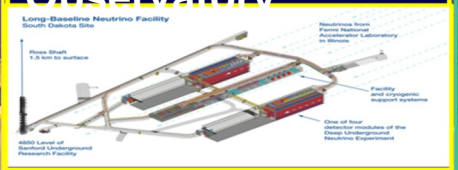
Next Generation Network-Integrated System for Data Intensive Sciences

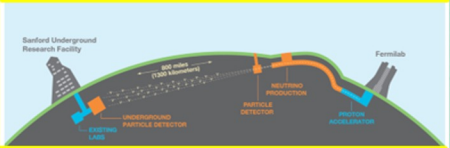


LHC



Rubin Observatory





LBNF/DUNE

SKA

LHC Run 3 and HL-LHC

Rubin Observatory

SKA

BioInformatics

Earth Observation

Gateways to a New Era

GNA-G Global Network Advancement Group

6NRP Workshop SmartNIC Panel

January 30, 2025

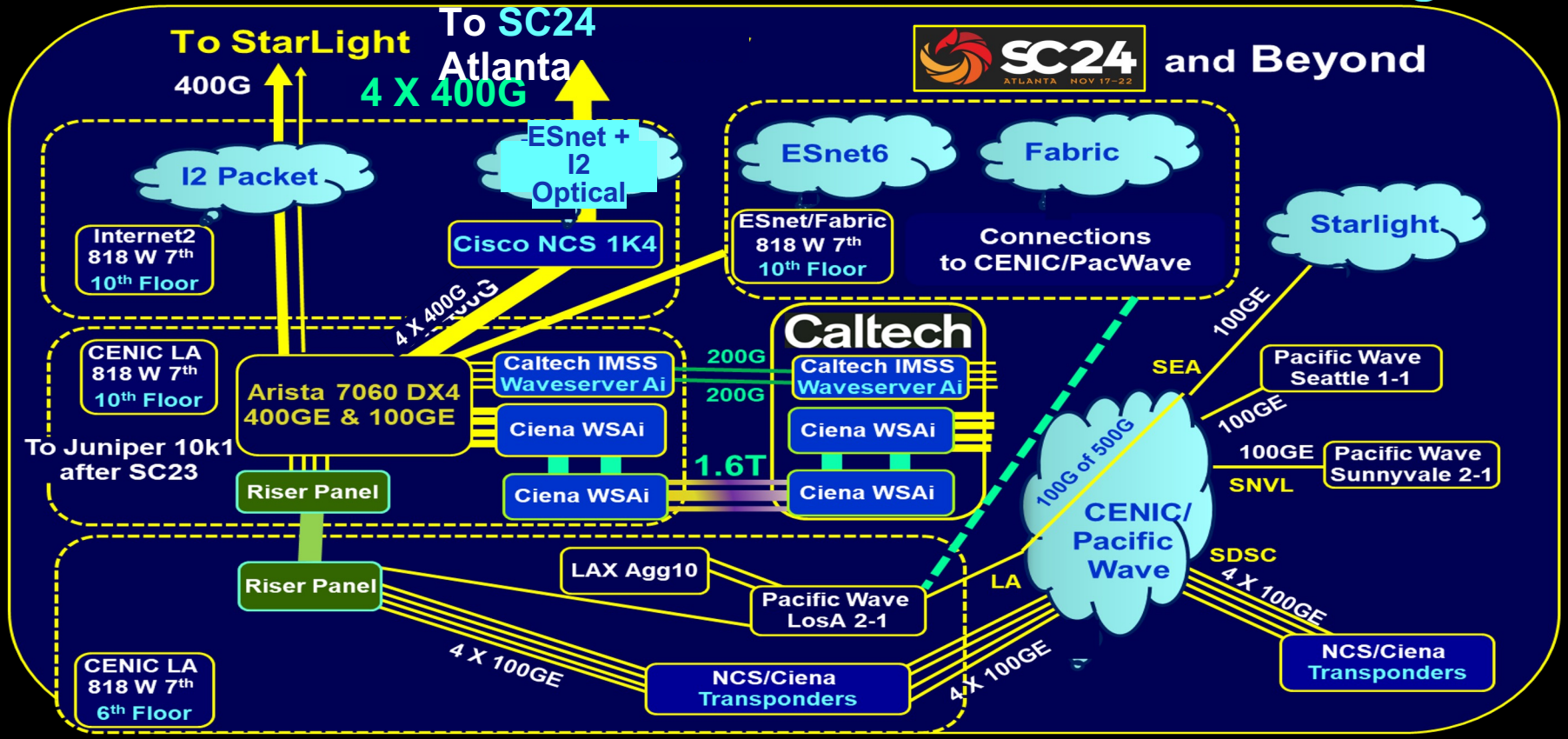
California Institute of Technology 1891

Global Petascale to Exascale Workflows for Data Intensive Sciences

Advances Embedded and Interoperate within a 'composable' architecture of subsystems, components and interfaces, organized into several areas; coupled to rising Automation

- **Visibility: Monitoring and information tracking and management** including IETF ALTO/OpenALTO, BGP-LS, sFlow/NetFlow, Perfsonar, Traceroute, Qualcomm Gradient Graph congestion information, Kubernetes statistics, Prometheus, P4/Inband telemetry, ***InMon***
- **Intelligence: Stateful decisions using composable metrics (policy, priority, network- and site-state, SLA constraints, responses to 'events' at sites and in the networks, ...)**, using NetPredict, Hecate, GradientGraph, Yale Bilevel optimization, Coral, Elastiflow/Elastic Stack
- **Controllability: SENSE/AutoGOLE/SUPA,** P4, segment routing with SRv6, SR/MPLS and/or PolKA, BGP/PCEP
- **Network OSes and Tools: GEANT RARE/freeRtr, SONIC;** Calico VPP, Bstruct-Mininet environment, ...
- **Orchestration: SENSE, Kubernetes (+k8s namespace),** dedicated code and APIs for interoperation and progressive integration

A New Generation Persistent 400G Super-DMZ: Ciena, Arista, CENIC, Pacific Wave, ESnet, Internet2, Caltech, UCSD, StarLight++



SC24: 4 X 400G on ESnet, I2 Atlanta-LA: Ciena, Caltech and CENIC Using WS Ais and a dark fiber pair. Bringing 4 X 400GE via 2 800G Waves direct to the campus

CENIC, ESnet and Internet2 at the LA PoP

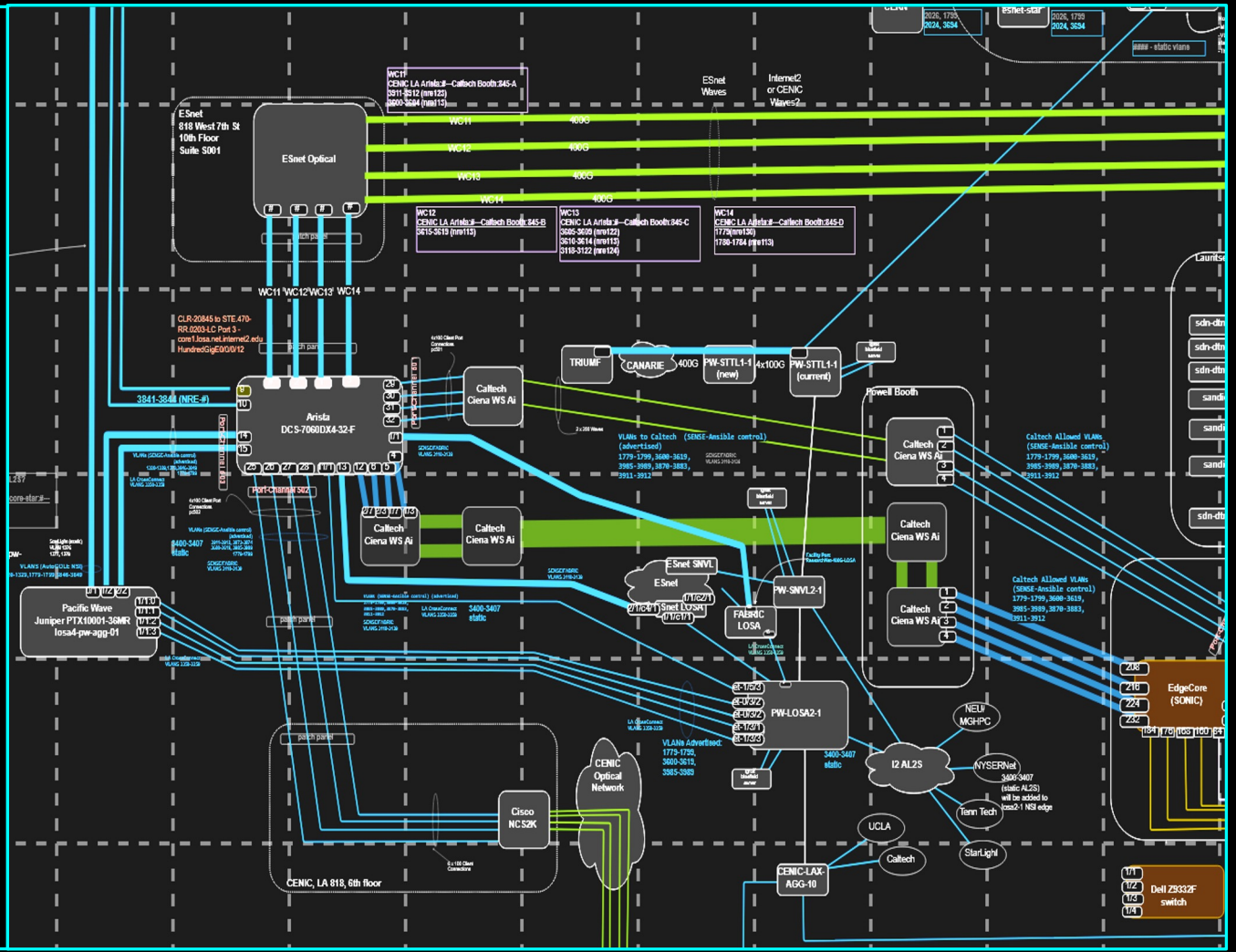
400G + 4 X 100G to Caltech via WS Ais

4 X 400G LA-Atlanta via ESnet, Internet2

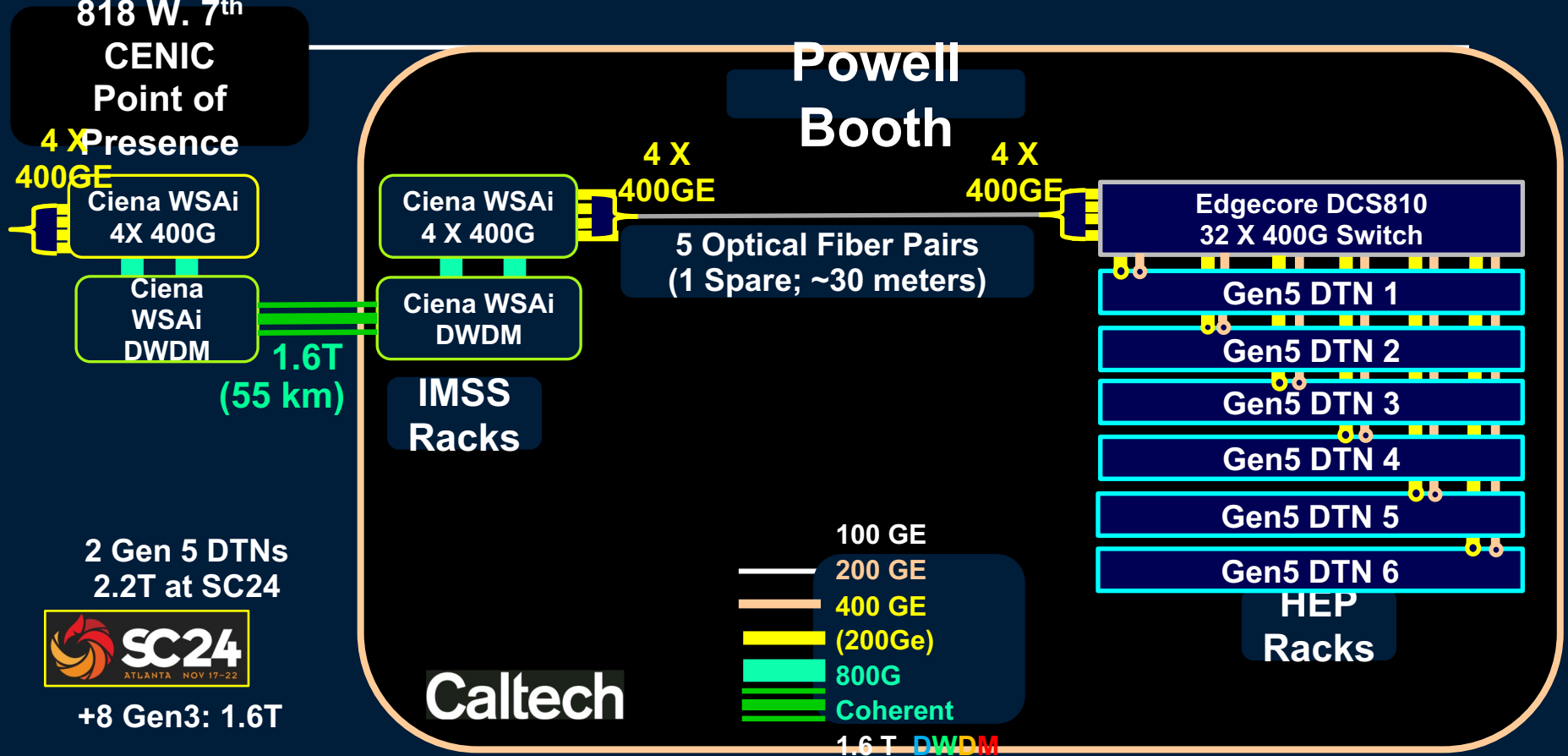
4 x 100G to UCSD/SDSC

2 X 400G to Pacific Wave via CENIC

Permanent:
400G NA-REX Prototype
400G to ESnet Production



Simplified Caltech – LA Layout for SC24



DTN: ASUS RS520A-E12-RS12U

PCIe 5.0 Ports: Two x16, Two x8, 1 OCP 3.0 x16

8 US CMS DTNs: CPU EPYC 9374F 3.85 GHz, to 4.3 GHz 32 Core

42	
41	
40	Tofino1 BUR001
39	
38	Tofino1 BUR002
37	
36	Dell Z9432F 32 X 400G Switch
35	
34	Dell 730XD DTN 2 X 100G UCSD 1 (2U)
33	
32	
31	Dell 730XD DTN 2 X 100G UCSD 2 (2U)
30	
29	
28	Dell Z9100 32 X 100G Switch
27	
26	Dell S60 Switch
25	
24	Console
23	
22	
21	Dell 730XD DTN 2 X 100G UCSD3 (2U)
20	
19	Dell 730XD DTN 2 X 100G UCSD4 (2U)
18	
17	
16	Dell 730XD DTN 2 X 100G UCSD5 (2U)
15	
14	
13	Dell 730XD DTN 2 X 100G UCSD6 (2U)
12	
11	
10	Dell 730XD DTN 2 X 100G NEU 1 (2U)
9	
8	
7	Dell 730XD DTN 2 X 100G SANDIE 9 (2U)
6	
5	
4	PDU, cables etc.
3	
2	
1	

NIC Setup at SC24: 2 X 1.2 Tbps

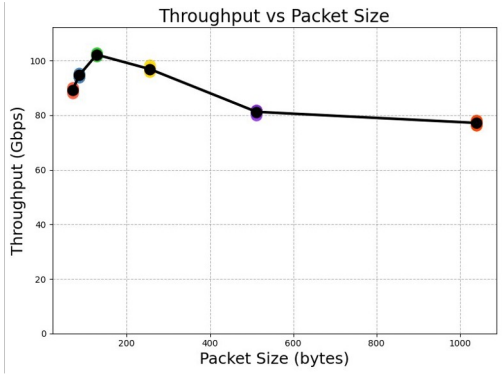
Two ConnectX-7 2 X 400GE;
One ConnectX-6 200GE
One Broadcom OCP3.0 2 X 100G

Now SDN Testbed + SC24 Rack Connected at 4 X 400G to the Production Tier2 Facility

Mohammad Sada Slides

SmartNIC resources on Nautilus:

- 32 Xilinx Alveo FPGAs as 2x100Gbps P4-programmable SmartNICs
- 3 Xilinx Alveo U55C FPGAs
- 24 Intel Stratix 10 NX2100 as 6x100Gbps SmartNICs
- 7 BlueField SmartNICs with DOCA Flow for building packet processing
- P4-programmable Tofino switches enabling dynamic packet processing
- ConnectX-6 2x100Gbps NICs



Network Experiments on Nautilus:

SENSE Path Provisioning

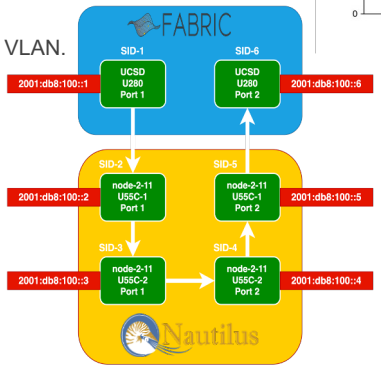
- Uses ESnet SENSE to establish an L2 path between Nautilus nodes.
- Dynamically configures and activates VLAN interfaces across networks.

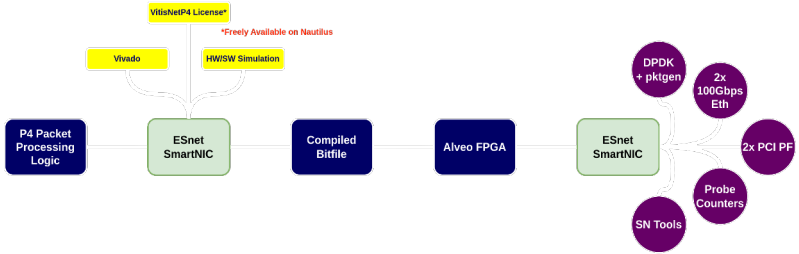
Multus Network Definition

- Multus CNI enables multiple pod interfaces.
- A NetworkAttachmentDefinition (NAD) is created for the SENSE VLAN.
- Connects pods to the configured L2 path.

Experiment Pod Deployment

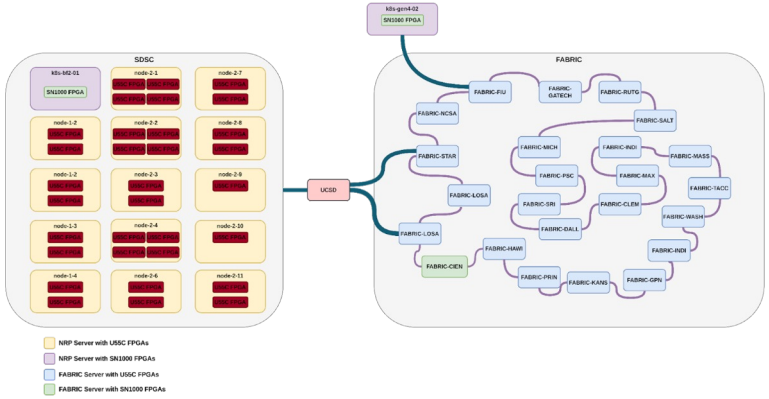
- Pods deploy with a default network (control) and VLAN (data).
- Multus attaches VLAN for data-plane communication.



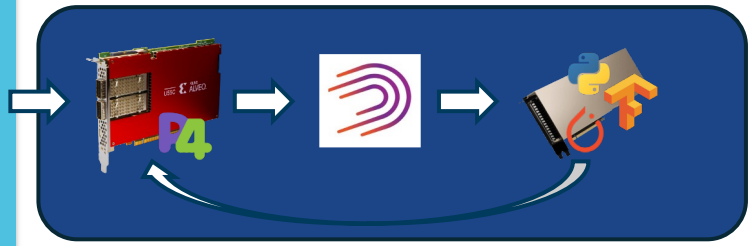


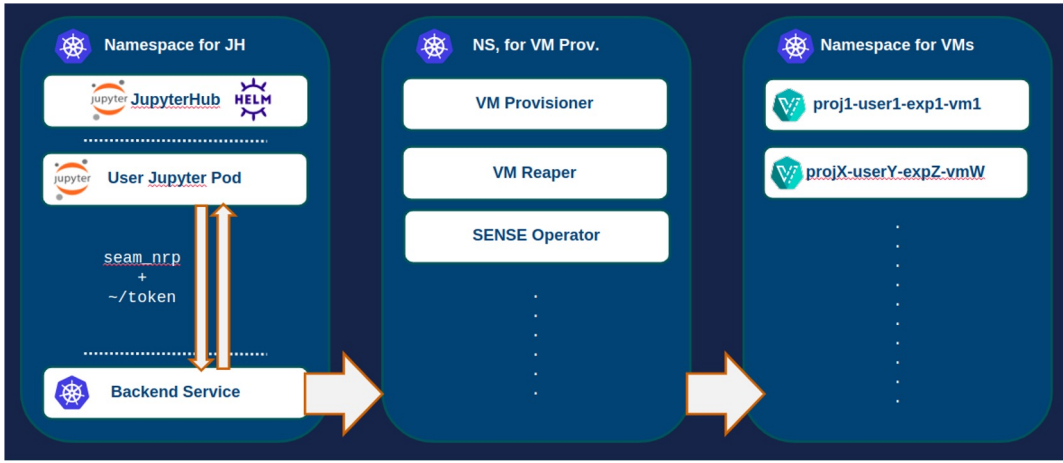
The SENSE Operator allows users to declare a network path using a custom Kubernetes resource: SensePath.

```
apiVersion: 6nrp.example.com/v1
kind: SensePath
metadata:
name: my-sense-path
spec:
  uri1: "urn:ogf:network:nrp-nautilus.io:2020:node-2-6.sdsc.optiputer.net:enp65s0f1np1"
  uri2: "urn:ogf:network:nrp-nautilus.io:2020:node-2-7.sdsc.optiputer.net:enp65s0f1np1"
  bandwidth: 1000
  vlan_tag: 3115
```



- Offload packet processing with **SRv6**
- **In-Network AI Acceleration:**
  Filter, compress, and preprocess datasets **before** reaching GPUs
- Train **ML models on network telemetry**
- **Cross-domain inter-testbed networking** experiments.
- Paving the way for ***autonomous networking and self-optimizing distributed systems***





Questions?