The National Research Platform & SCinet: Live, Multi-Domain AI/HPC Workflows Across a National Cyberinfrastructure

Submitted by Mohammad Firas Sada (mfsada@es.net)
on behalf of the involved parties listed herein

SC26 Network Research Exhibition

1 Abstract

The San Diego Supercomputer Center and the Prototype National Research Platform (NSF Award 2112167, Category II) present the NRP, a community-owned federation of over 80 institutions. At SC26, NRP's Nautilus Kubernetes Cluster directly integrates SCinet-connected compute nodes on the exhibit floor for live, interactive AI/ML and HPC workflows. Beyond the local demonstration, this NRE conducts wide-area network experiments with the ESnet SENSE Orchestrator and NRP's SENSE Operator, provisioning on-demand, high-bandwidth, deterministic QoS paths across SCinet, StarLight, FABRIC, Caltech, and the Chameleon Cloud, transforming the show floor into a national-scale research instrument.

2 Goals

- Live, interactive AI & scientific visualization on SCinet-attached Nautilus nodes, with attendee-driven workflows.
- Reproducible multi-user access via JupyterHub, Kubernetes, SLURM, and CILogon, backed by real-time NRP infrastructure.
- Heterogeneous hardware orchestration (GPUs, DPUs, AMD/Xilinx Alveo, Intel Agilex/Stratix FPGAs) under Kubernetes.
- Federated WAN experiments with the ESnet SENSE Orchestrator: dynamic high-bandwidth, low-latency L2/L3 paths across SENSE-enabled domains.
- Production scientific workloads on the SC26 cluster with guaranteed low preemption, zero reservation, and opportunistic use of underutilized capacity across research domains (OSG, AI-driven digital twins, HTC).
- Reproducible-by-construction: every demo is code in a Jupyter notebook or Kubernetes manifest, so attendees can replay or extend the exact workflow on NRP (or on their own institution's resources) after the conference.

3 Resources

- >1500 GPUs, 50+ FPGAs, and 7 DPUs across 500+ Nautilus nodes.
- Dedicated VLANs and L2 segments SCinet → NRP via programmable switches and SmartNICs.
- User interaction: JupyterHub, CILogon login, Matrix community support.
- ESnet SENSE Orchestrator + NRP SENSE Operator for provisioning.

4 Infrastructure

- Kubernetes cluster with Multus and ESnet SENSE integration for programmable L2 QoS paths from SCinet to NRP, StarLight, FABRIC, and others.
- Live SCinet-connected Nautilus nodes.
- 32 P4-programmable Alveo U55C FPGAs.
- Qualcomm AI100 Ultra cards (GenAI).
- Storage backends: CephFS, S3-compatible stores, OSDF.

The cluster is accessed exactly as production NRP users access it: via JupyterHub for interactive serve or via SLURM/Kubernetes manifests for batch and distributed workloads, with CILogon single sign-on across the federation. Multus binds each workload's secondary interfaces to dedicated VLANs, so a notebook cell that streams from a remote GPU cluster or an FPGA at StarLight is exercising the same plumbing a production remote job would.

5 Impact

The NRP is a living, community-driven cyberinfrastructure that erases traditional silos between AI, HPC, storage, and networking. Bringing SCinet-connected Nautilus nodes into production Kubernetes and SLURM at SC26 lets researchers run real, interactive AI and scientific visualization live from the floor, bridging coast-to-coast resources. Live WAN experiments with the ESnet SENSE Orchestrator demonstrate dynamic, high-bandwidth connectivity across multiple administrative and network domains, enabling distributed workflows spanning the Open Science Grid, FPGAs at StarLight, GPU clusters at UCSD, and edge sites, orchestrated as one reproducible, multi-tenant platform. The experiment is itself a template for research communities: capacity is pooled, contributed, and reclaimed opportunistically on a national scale, and the network that carries it is provisioned the same way: on demand, per workflow, with QoS guarantees rather than permanent private circuits.

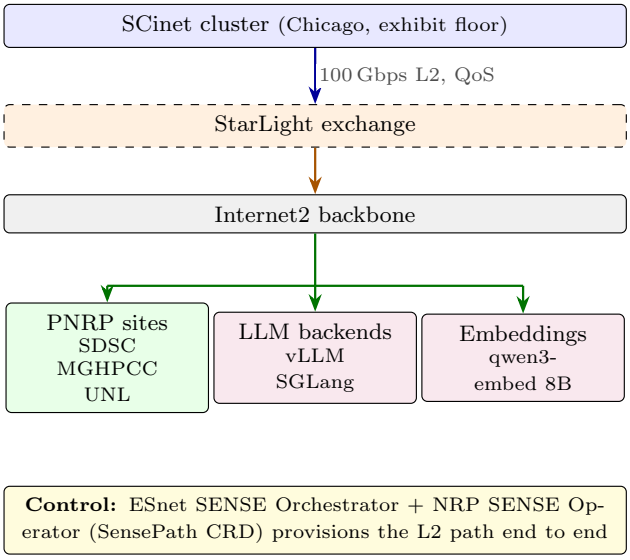


Figure 1: Exhibit-floor cluster + federated WAN paths from SCinet to NRP sites (shared NRE 1/2 topology).

6 Interactive Experience at SC26

Attendees interact with live research workflows on Nautilus nodes at the SCinet NOC. Real-time dashboards (Grafana, Prometheus, Observable, Jupyter, inMon) and immersive visualizations expose every layer, from code and AI models to network QoS metrics and WAN path utilization. The Nautilus cluster guarantees low-priority preemption on SCinet nodes, ensuring dedicated resources for all NREs. Featured workflows include:

- 3D Cell Reconstruction with AI (bioimaging)
- Digital twins for space exploration
- Wildfire monitoring & immersive visualization (atmospheric science)
- LLM as a Service in education; GenAI architectures for scientific research
- Open Science Grid (OSG) experiment hosting across the federation
- Live WAN-aware AI training over SENSE-provisioned bandwidth between SC26 and remote FABRIC/StarLight nodes

7 Involved Parties

ESnet / LBNL: Mohammad Firas Sada (mfsada@es.net); Tom Lehman (tlehman@es.net); Xi Yang (xiyang@es.net); Justas Balcas (jbalcas@es.net); Alberto Jimenez (ajimenez@es.net); Kiran Vasu (kvasu@es.net); Chin Guok (chin@es.net); Inder Monga (inder@es.net); Dmitry Mishin (dimm@lbl.gov); **UCSD / SDSC:** John Graham (jjgraham@ucsd.edu); Frank Würthwein (fkw@ucsd.edu); Tom DeFanti (tdefanti@ucsd.edu); Tom Hutton (hutton@sdsc.edu); Mahidhar Tatineni (mahidhar@sdsc.edu); Daniel Diaz (d4diaz@ucsd.edu); Chris Hughes (crhughes@sdsc.edu); Aashay Arora (aaarora@ucsd.edu); Diego Davila (didavila@ucsd.edu); **RNP:** Marcos Felipe Schwarz (marcos.schwarz@rnp.br); **iCAIR / Northwestern:** Joe Mambretti (j-mambretti@northwestern.edu); Jim Hao Chen (jim-chen@northwestern.edu); Fei Yeh (fyeh@northwestern.edu); **Caltech:** Harvey Newman (newman@hep.caltech.edu); Preeti Bhat (preeti@caltech.edu); Andres Moya (amoya@caltech.edu); **Fermilab:** Phil Demar (demar@fnal.gov); Asif Shah (sshah@fnal.gov); Dmitry Litvintsev (litvinse@fnal.gov); **UMass:** Eduardo Bach (eduardobach@umass.edu); Rafael Coelho Lopes de Sa (rclsa@umass.edu); Jessa Westclark (jessicaw@umass.edu); **UNL:** Derek Weitzel (dweitzel@unl.edu); Huijun Zhu (huijun.zhu@unl.edu); Sam Albin (sam.albin@unl.edu); Ashton Graves (agraves10@unl.edu); **FIU:** Jeronimo Bezerra (jbezerra@fiu.edu); Julio Ibarra (julio@fiu.edu); **Internet2:** Chris Cox (ccox@internet2.edu); James Deaton (jdeaton@internet2.edu); Matt Zekauskas (matt@internet2.edu); Christian Hitchcock (chitchcock@internet2.edu); Karl Newell (knewell@internet2.edu); **Others:** Christopher Bruton (CENIC, cbruton@cenic.org); Neil McKee (inMon, neil.mckee@inmon.com); Heidi Morgan (USC ISI, hlmorgan@isi.edu); Everson Scherrer Borges (IFES, everson@ifes.edu.br); Seungmin Kim (Yonsei, ehf@yonsei.ac.kr).

References

https://nrp.ai • https://nrp.ai/ai-hubs • https://nrp.ai/llms • https://sense.es.net • https://nrp.ai/documentation/admindocs/cluster/sense-multus • https://www.gna-g.net/autogole/