ECOSYSTEM for RESEARCH NETWORKING

ERN All Hands Meeting

ERN Cryo-EM
Federated Instrument Pilot Project

March 23, 2023

ERN Cryo-EM Federated Instrument Pilot Project

Research community feedback received by the ERN Structural Biology Working Group, through workshops, presentations and discussions.

Benefits

- Adjust experiment parameters live
- Optimize instrument utilization
- Identify target achieved/fruitless runs
- Broaden collaborative efforts, science discovery

Barriers

- Access limitations
- Significant latency
- Insufficient data transfer rates
- HPC queue wait times (public and private)
- Training and expertise

Objectives

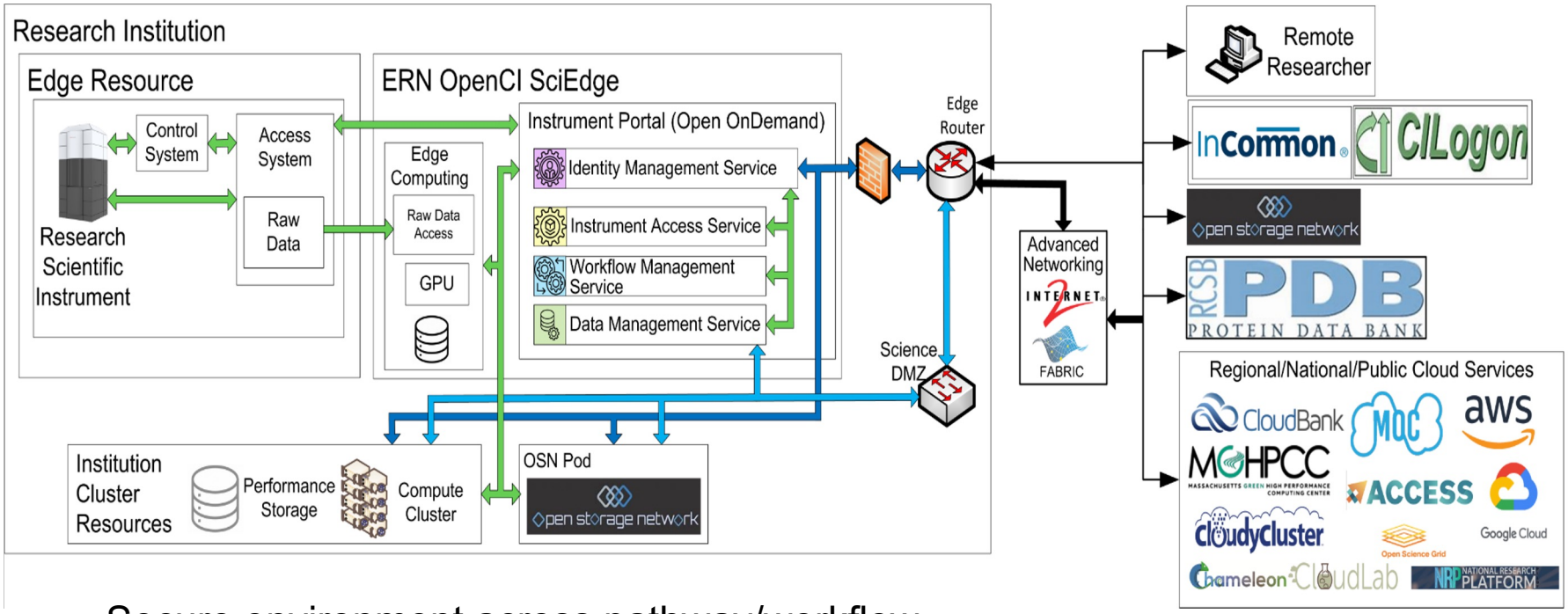
Facilitate and simplify multi-institutional collaborative research by removing many of the barriers encountered when attempting to access remote scientific instruments, and enable real-time parameter adjustment through edge computing.

- Easy to use, secure, web-based resource portal
- Simplified, federated authentication, authorization and access
- Real-time workflows and adjustments
- Edge computing
- Access to additional analysis resources – private and/or public
- Portable, flexible, easily duplicated, managed and maintained
- Secure data management system
- Data pipeline starts at the instrument
- Do not reinvent the wheel
- Share efforts with the research community

The ERN Federated CryoEM Instrument Pilot Project Team

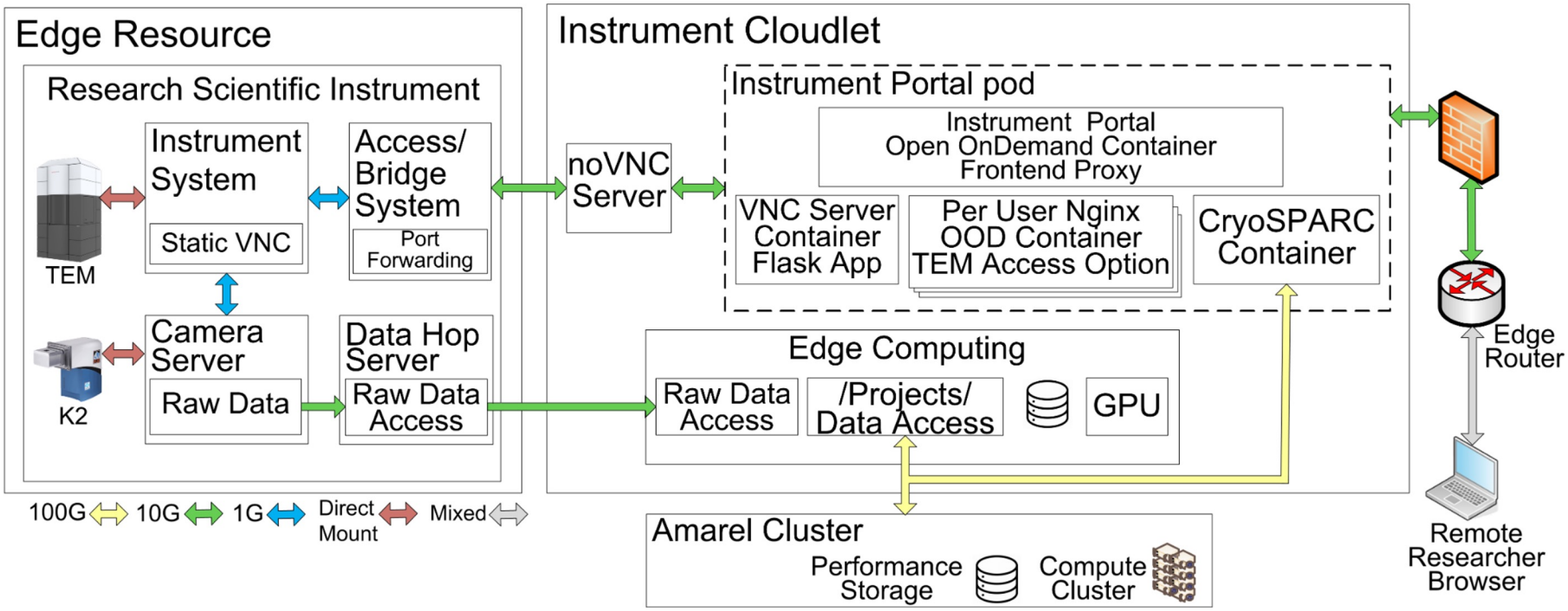
- Ken Dalenberg, Bala Desinghu, Jason Kaelber, Jeremy Schafer, and James Barr von Oehsen - Rutgers University
- Wolf Hey - Penn State University
- John Goodhue - MGHPCC
- Morgan Ludwig - TechSquare
- Boyd Wilson and Cole McKnight - Omnibond
- Michael Zink - University of Massachusetts, Amherst
- Ewa Deelman and Mats Rynge - University of Southern California
- Maureen Dougherty - Ecosystem for Research Networking
- The Open OnDemand team
- The FABRIC team - Matt Zekauskas, Tom Lehman, Paul Ruth, Ilya Baldin

Pilot Project Design

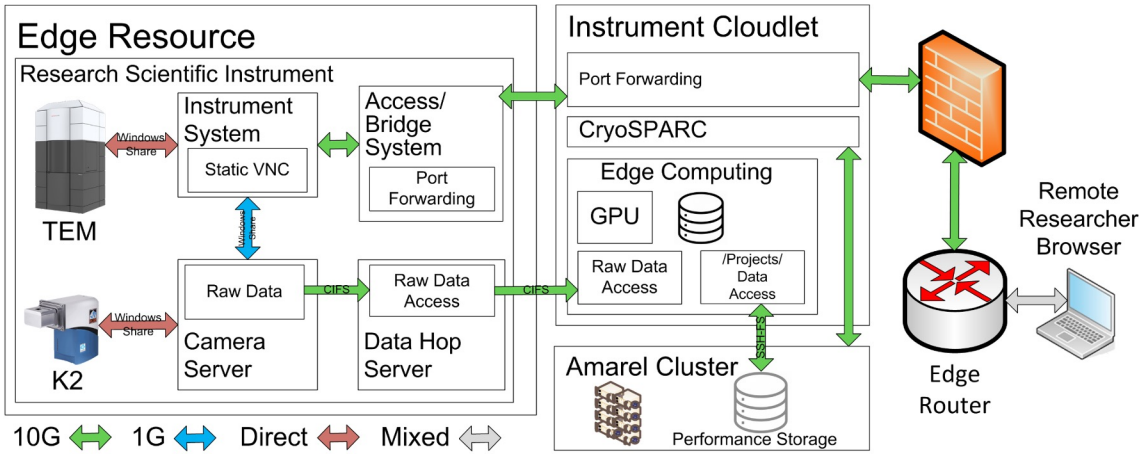


- Secure environment across pathway/workflow
- Common framework for federated authorization, authentication, and access
- Reproducible, reliable, portable, flexible, simplified support
- Edge computing
- Open Source project

Phase 1


Edge Resource
Research Scientific Instrument
Instrument System
Static VNC
Access/ Bridge System
Port Forwarding
TEM
Camera Server
Raw Data
Data Hop Server
Raw Data Access
K2
100G
10G
1G
Direct Mount
Mixed
Instrument Cloudlet
noVNC Server
Instrument Portal pod
Instrument Portal
Open OnDemand Container
Frontend Proxy
VNC Server Container
Flask App
Per User Nginx
OOD Container
TEM Access Option
CryoSPARC Container
Edge Computing
Raw Data Access
/Projects/ Data Access
GPU
Amarel Cluster
Performance Storage
Compute Cluster
Edge Router
Remote Researcher Browser

Step 1 – Basic Remote Accessibility



Remote access of TEM with real-time workflow using edge computing

Workflow

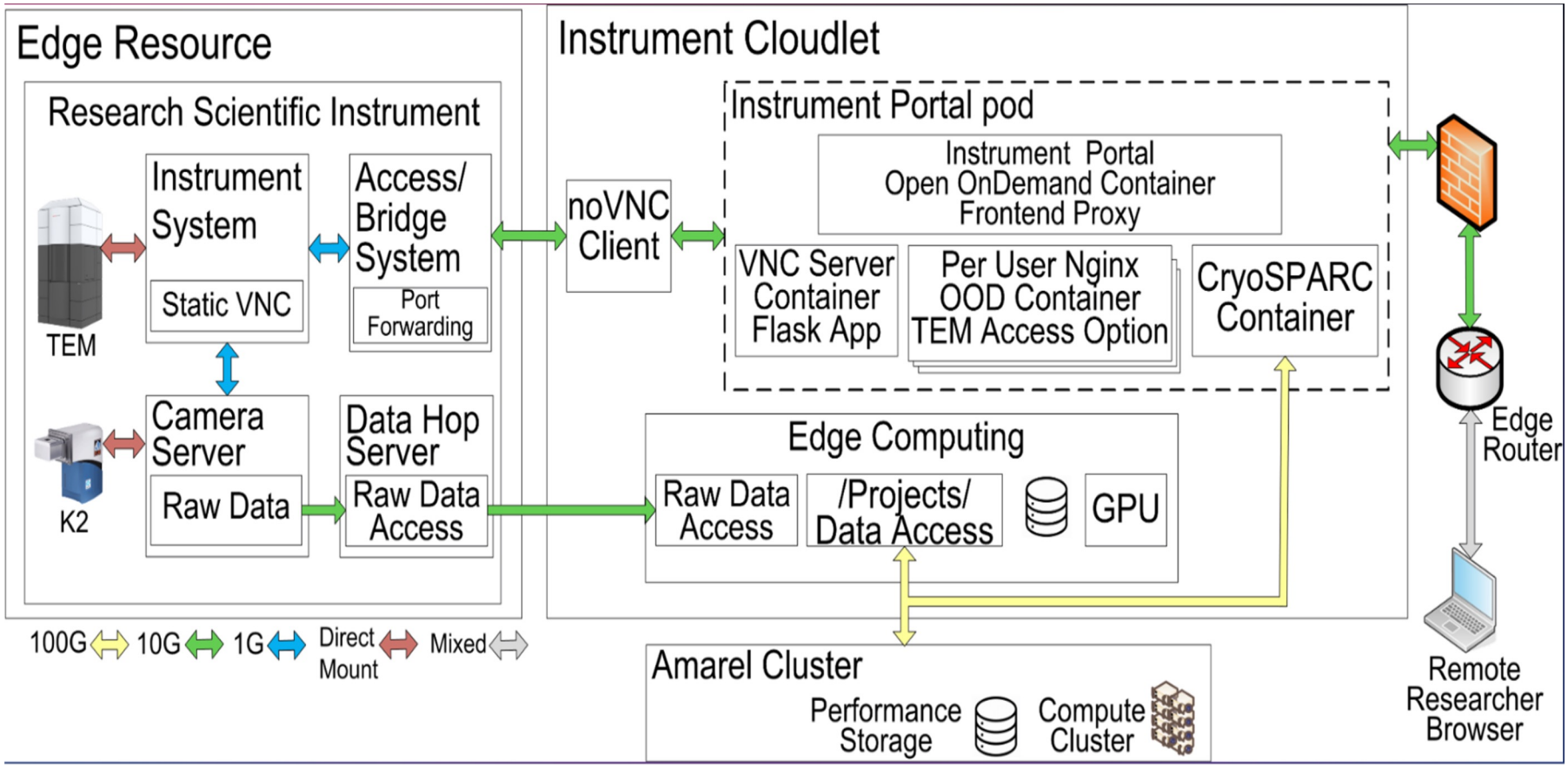
- Off-institution remote user establish Rutgers VPN session
- VNC access through remote web browser to Instrument System's static VNC
- Workflow launched
- Data processing application "cryoSPARC Live"
  - Pre-process raw images with edge computing GPUs
  - Amarel cluster job submission 2D alignment and 3D structure refinement of pre-processed image files
  - Real-time adjustment decisions made: results 3 minutes behind experiment

Step 1 – Results

Workflow Run Results:

- 320 images/hour novel complex of the transmembrane protein TolC
- Processed 2.5TB over 2 days
- Computational output < 3 minutes behind actual data acquisition
- Real-time experiment adjustments made based on quality of incoming TEM data
- Bandwidth measurements confirmed data transfer rate from instrument > 1GB
- Data transfer rate from cloudlet to Amarel cluster < 1GB
- Network I/O reduced by 1-2 orders of magnitude due to Cloudlet edge computing

Step 2 Design



Step 2 Implementation

Cloudlet Updates

- Accessible from the public internet
- SSL Certificates
- OIDC ClientID and Secret from upstream OpenID Connect IDP
- OOD and noVNC Flask application, ansiblized container templates
- Usermap file

Challenges

- SELinux restrictions
- Rootless podman
  - UID:GID configuration limitations
  - Restricts running most SELinux operations
  - Separate user namespace form local host
  - User namespace
- Instrument's SSL httpd port disabled

CryoEM Remote Instrument Demonstration

Phase 1 Conclusion

Remote access to edge scientific instruments for real time analytical workflows using edge computing is both feasible and beneficial

- Federated remote access to scientific instrument in secure environment
- Foster team science and democratization of scientific instruments with emphasis on under-represented and under-resourced colleges and institutions
- Edge computing
  - Real-time decision making and adjustment
  - Decreased network I/O for pre-processed image data
- Reliability, reproducibility, reusability, portability, ease of use/management/support

Lessons Learned

- security: traffic isolation, rootless container, per-user permissions
- expertise: subject matter experts, researcher and technical expertise important

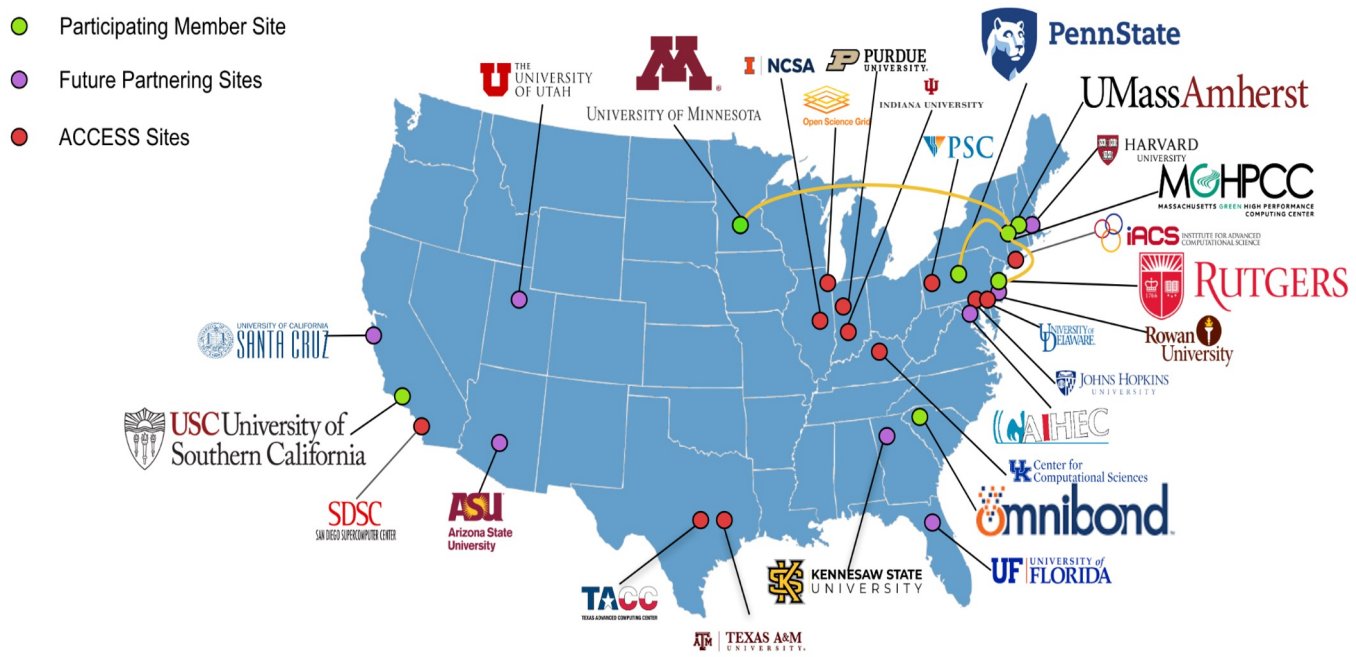
Next Steps

- CryoSPARC containerization
  - Integration with Open OnDemand
- Pegasus Workflow and Data Management System integration
- FABRIC integration
- OSN integration
- Enable access to external researchers
  - engage with interested institutions

Interested in learning more or participating, please contact info@ernrp.org

GITHUB: https://github.com/mghpcc/ERN-Remote-Scientific-Instrument

The ERN Federated CryoEM Instrument Pilot Project Site Map



Participating Member Sites (alphabetical order)

- Massachusetts Green High Performance Computing Center
- Omnibond
- Pennsylvania State
- Rutgers University
- University of Massachusetts, Amherst
- University of Minnesota
- University of Southern California

Future Partnering Sites (alphabetical order)

- American Indian Higher Education Consortium
- Arizona State University
- Harvard University
- Kennesaw State University
- Rowan University
- University of California, Santa Cruz
- University of Florida, Gainesville
- University of Utah

ECOSYSTEM for RESEARCH NETWORKING

Thank You!

Interested in learning more or participating, please contact info@ernrp.org

Website: https://ernrp.org

NSF OAC-2018927 and OAC-1925482