Experience with Large-scale Science on GENI

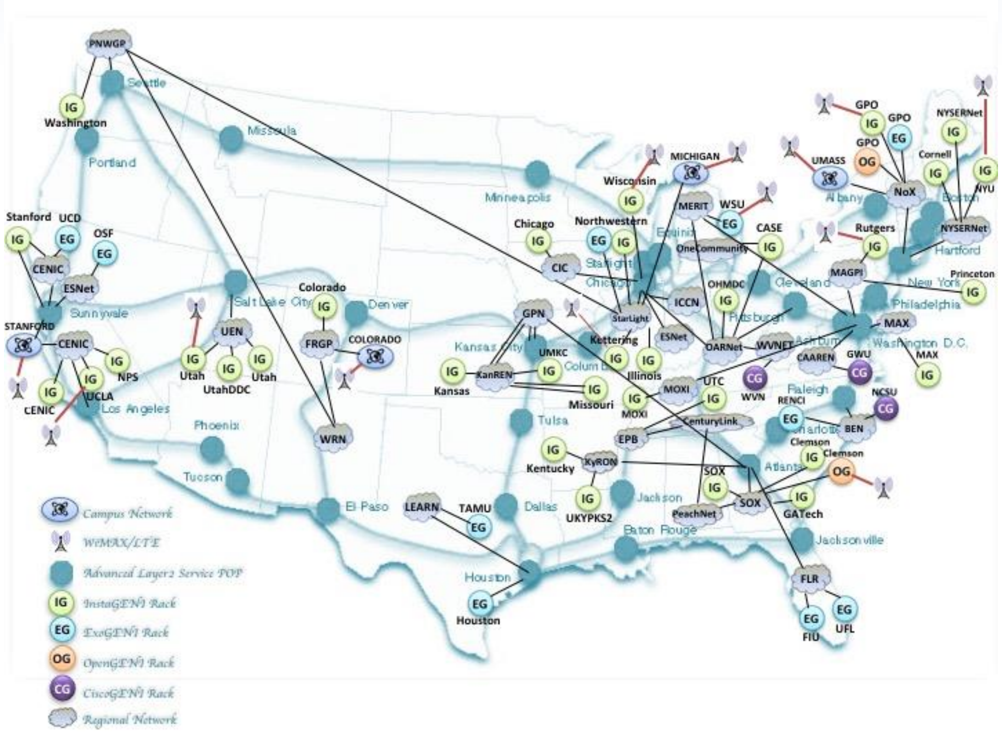
Paul Ruth, pruth@renci.org
RENCI / UNC Chapel Hill

renci

RESEARCH \ ENGAGEMENT \ INNOVATION

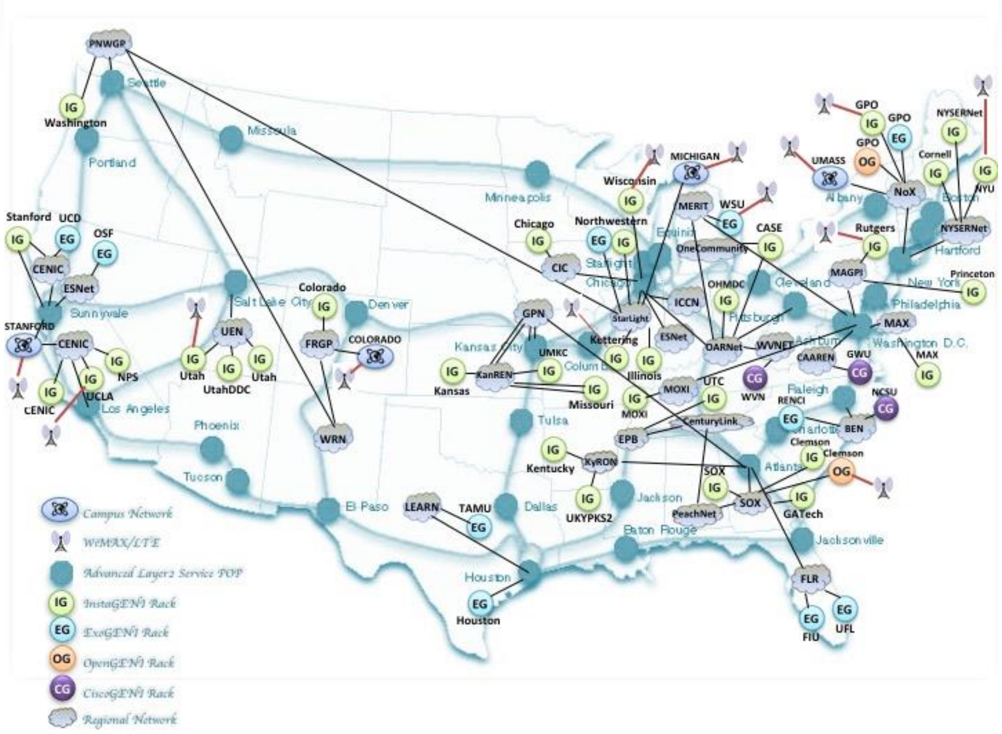
What is GENI?

- Federated Cyberinfrastructure
  - Many domains, people, users sharing resources
  - Common API for diverse resources
  - Deeply programmable
  - Originally targeted networking experiments



What is GENI?

- Federated Cyberinfrastructure
  - Many domains, people, users sharing resources
  - Common API for diverse resources
  - Deeply programmable
  - Originally targeted networking experiments



This talk is about **MY** experience using GENI for domain science

RENCI’s Work Toward Science on GENI

- Solar Fuels (2011, unfunded)
  - Chemistry
- ADAMANT (2013, NSF CC-NIE)
  - Adaptive Data-Aware Multi-Domain Application Network
  - Earthquakes (cybershake), Astronomy (montage)
- GENI Science Shakedown (2013 , GENI Spiral 4)
  - Storm surge (ADCIRC), Genomics (MotifNetwork)
- RADII (2014, NSF CC-IIE)
  - Resource Aware DatacentrIc Collaboration Infrastructure
  - Genomics
- Panorama, Panorama360 (2015 and 2017, DOE)
  - Workflows
  - Physics
- SAFE Superfacilities (2016, NSF CICI)
  - Creating Dynamic Superfacilities the SAFE Way
  - Superfacilities, High-energy physics
- SciDAS (2017, NSF CC*Data)
  - National Cyberinfrastructure for Scientific Data Analysis at Scale
  - Genomics

What sciences am I talking about?

- Computation Science Applications
  - Modeling
  - Simulation
  - Big Data Analytics
  - Scientific Workflows
- Typical Resources
  - Large computing centers
  - Campus compute clusters
  - Public and private clouds
  - Shared data repositories
  - Shared instruments (e.g. telescopes, particle colliders)

How can science benefit from dynamic programmable infrastructure?

Toward Science on GENI

- Initial Effort (2011)
  - Solar Fuels
- Adaptive Applications/Workflows (2013-present)
  - ADAMANT, RADII, Panorama
- Scaling Science Experiments (2013-2015)
  - GENI Science Shakedown
- Multidomain Science (2016-present)
  - SAFE, SciDAS

Toward Science on GENI

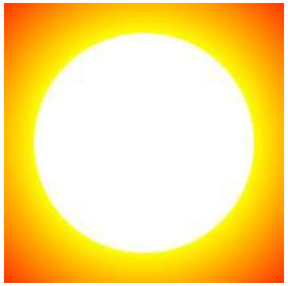
- Initial Effort (2011)
  - Solar Fuels
- Adaptive Applications/Workflows (2013-present)
  - ADAMANT, RADII, Panorama
- Scaling Science Experiments (2013-2015)
  - GENI Science Shakedown
- Multidomain Science (2016-present)
  - SAFE, SciDAS

Initial Effort: Solar Fuels

Creation of storable fuels using solar energy and catalysis

The Science

- Research in Solar Fuels and Photovoltaics will integrate light absorption and electron transfer driven catalysis
- Molecular assemblies to create efficient devices for solar energy conversion through artificial photosynthesis
  - T Meyer, J Papanikolas, C Heyer, Catalysis Letters 141 (2011) 1-7.



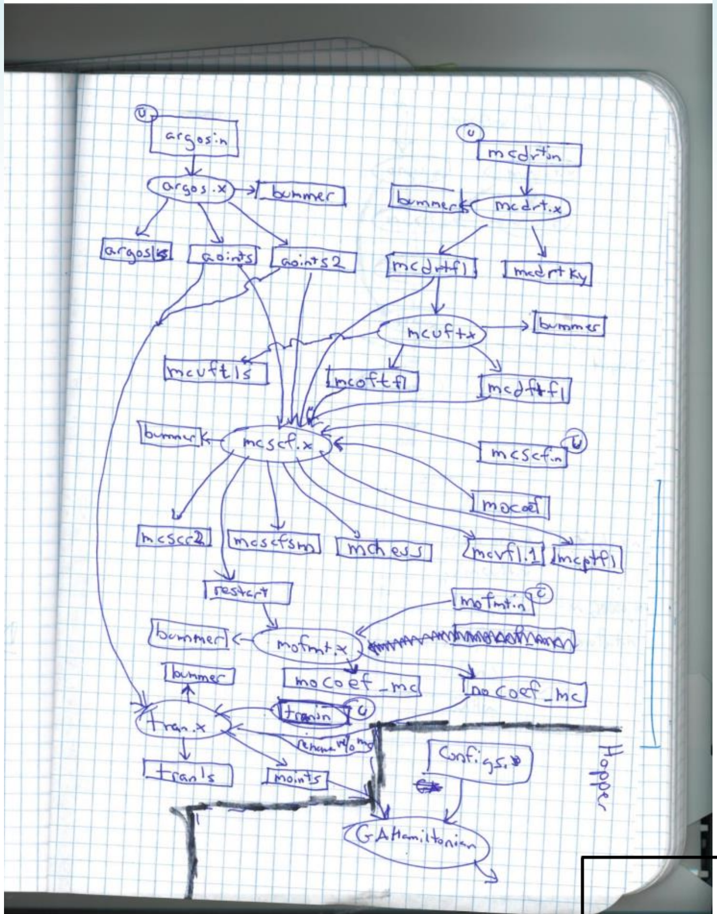
SOLAR FUELS
& NEXT GENERATION PHOTOVOLTAICS

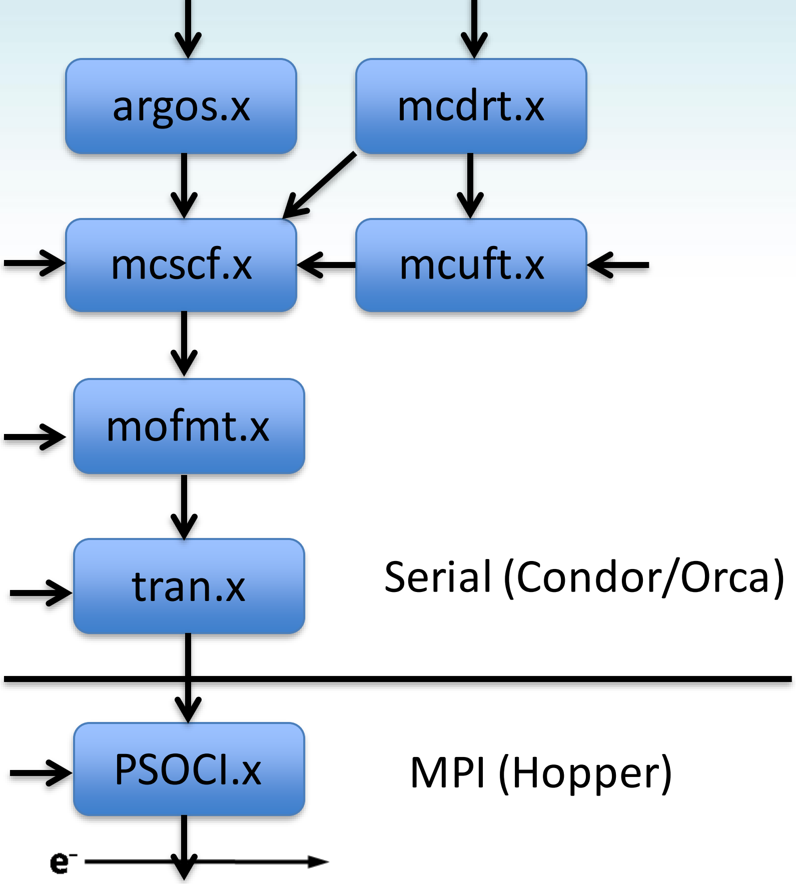
A Theoretical Framework

- Co-design strategy for creation of new scalable codes
- Incorporation of workflow technologies to coordinate, launch, and enhance resilience of the design pipeline
- Apply the developed codes to solve complex problems in electronic structure, kinetics, and synthesis

Collaborations - Working directly with

- Experimentalists (UNC-CH)
- Model and methods developers (Duke, UNC-CH)





Serial (Condor/Orca)

MPI (Hopper)

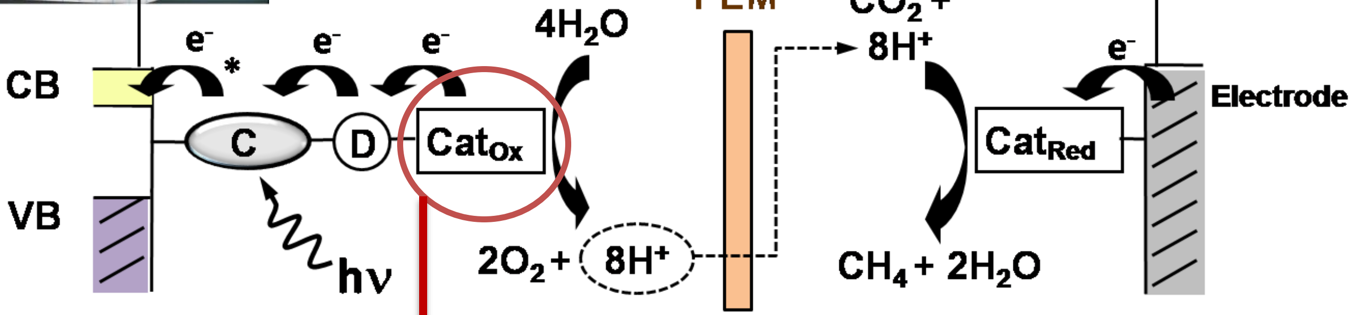


Image provided by UNC-CH EFRC: http://www.efrc.unc.edu/

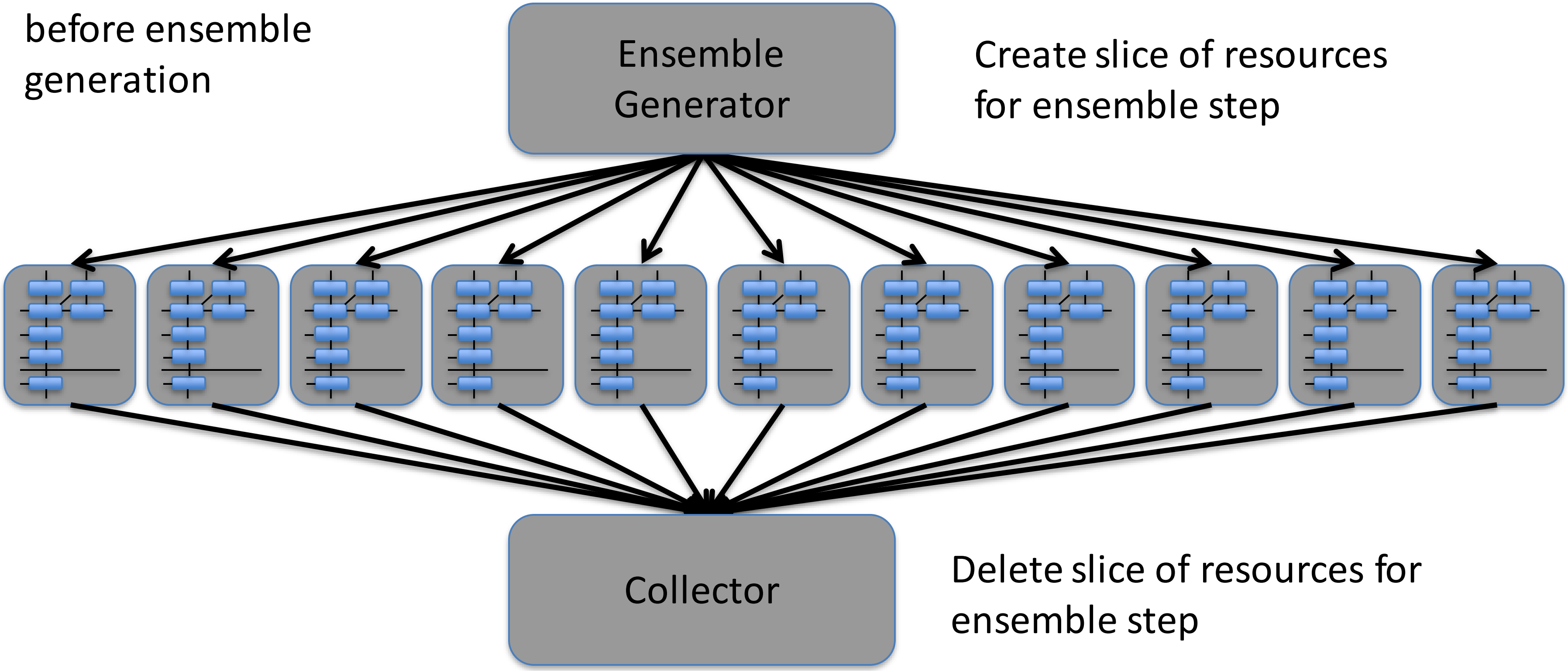
Oxidation catalysts

Renci
Renaissance Computing Institute
Catalyst for Innovation

renci

Solar Fuels: Workflow Ensemble

- Wide step of Ensemble is temporal
- Width may not be known before ensemble generation



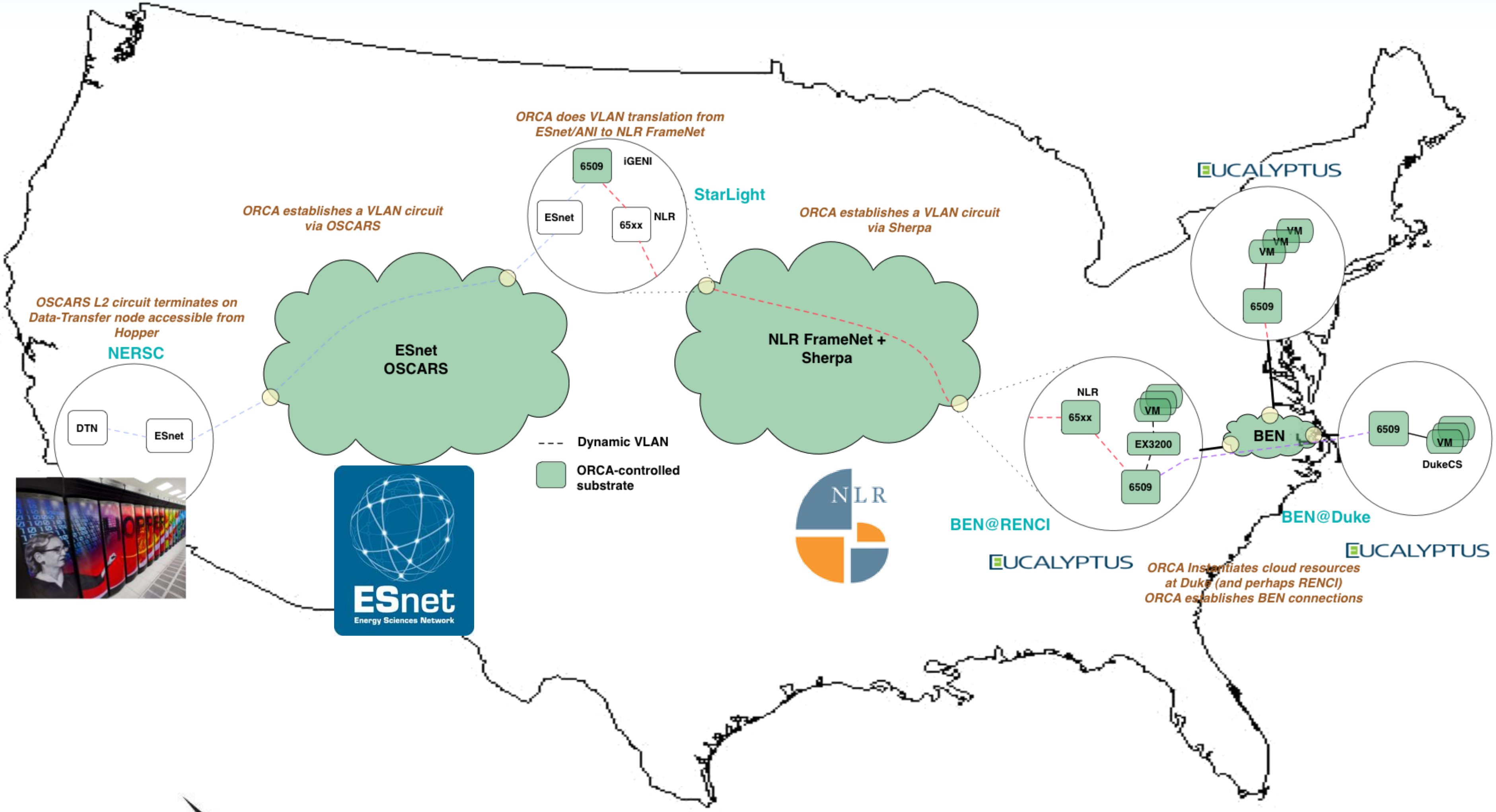
Ensemble Generator

Create slice of resources for ensemble step

Collector

Delete slice of resources for ensemble step

SC11 Demo: Solar Fuels Workflow



ORCA does VLAN translation from ESnet/ANI to NLR FrameNet

ORCA establishes a VLAN circuit via OSCARS

ORCA establishes a VLAN circuit via Sherpa

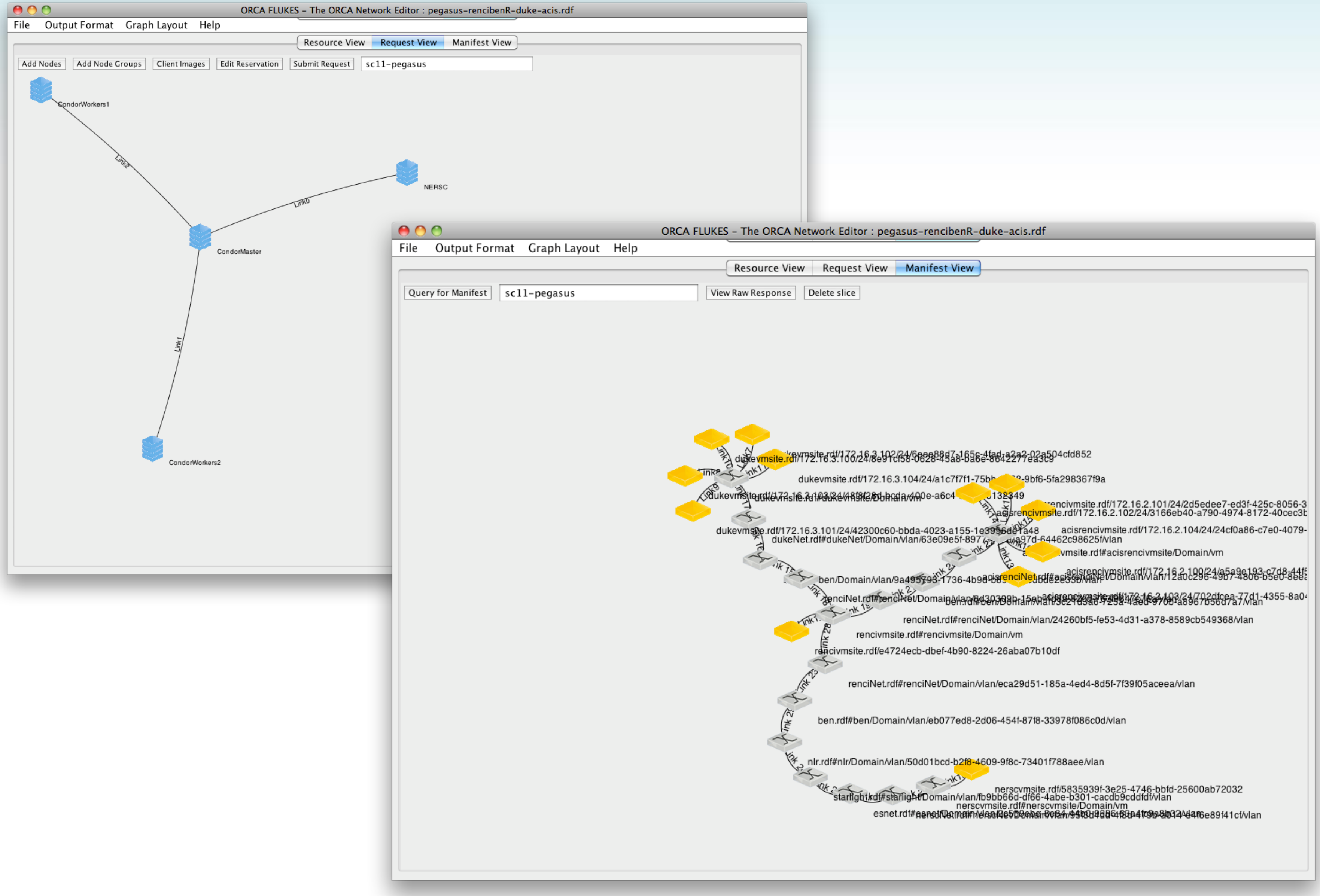
OSCARS L2 circuit terminates on Data-Transfer node accessible from Hopper

ORCA Instantiates cloud resources at Duke (and perhaps RENCI)
ORCA establishes BEN connections

- - - Dynamic VLAN

🟩 ORCA-controlled substrate

renci


ORCA FLUKES – The ORCA Network Editor : pegasus-rencibenR-duke-acis.rdf
File
Output Format
Graph Layout
Help
Resource View
Request View
Manifest View
Add Nodes
Add Node Groups
Client Images
Edit Reservation
Submit Request
sc11-pegasus
CondorWorkers1
Link2
NERSC
Link0
CondorMaster
Link1
CondorWorkers2
ORCA FLUKES – The ORCA Network Editor : pegasus-rencibenR-duke-acis.rdf
File
Output Format
Graph Layout
Help
Resource View
Request View
Manifest View
Query for Manifest
sc11-pegasus
View Raw Response
Delete slice
dukevmsite.rdf/172.16.3.104/24/a1c7f7f1-75b…-9bf6-5fa298367f9a
dukevmsite.rdf/172.16.3.101/24/42300c60-bbda-4023-a155-1e3995de1a48
dukeNet.rdf#dukeNet/Domain/vlan/63e09e5f-897…
acisrencivmsite.rdf/172.16.2.102/24/3166eb40-a790-4974-8172-40cec3b
acisrencivmsite.rdf/172.16.2.104/24/24cf0a86-c7e0-4079-
…vmsite.rdf#acisrencivmsite/Domain/vm
ben/Domain/vlan/9a495793-1736-4b98-…
renciNet.rdf#renciNet/Domain/vlan/24260bf5-fe53-4d31-a378-8589cb549368/vlan
rencivmsite.rdf#rencivmsite/Domain/vm
rencivmsite.rdf/e4724ecb-dbef-4b90-8224-26aba07b10df
renciNet.rdf#renciNet/Domain/vlan/eca29d51-185a-4ed4-8d5f-7f39f05aceea/vlan
ben.rdf#ben/Domain/vlan/eb077ed8-2d06-454f-87f8-33978f086c0d/vlan
nlr.rdf#nlr/Domain/vlan/50d01bcd-b2f8-4609-9f8c-73401f788aee/vlan
nerscvmsite.rdf/5835939f-3e25-4746-bbfd-25600ab72032
starlight.rdf#starlight/Domain/vlan/fb9bb66d-df66-4abe-b301-cacdb9cddfdf/vlan
nerscvmsite.rdf#nerscvmsite/Domain/vm

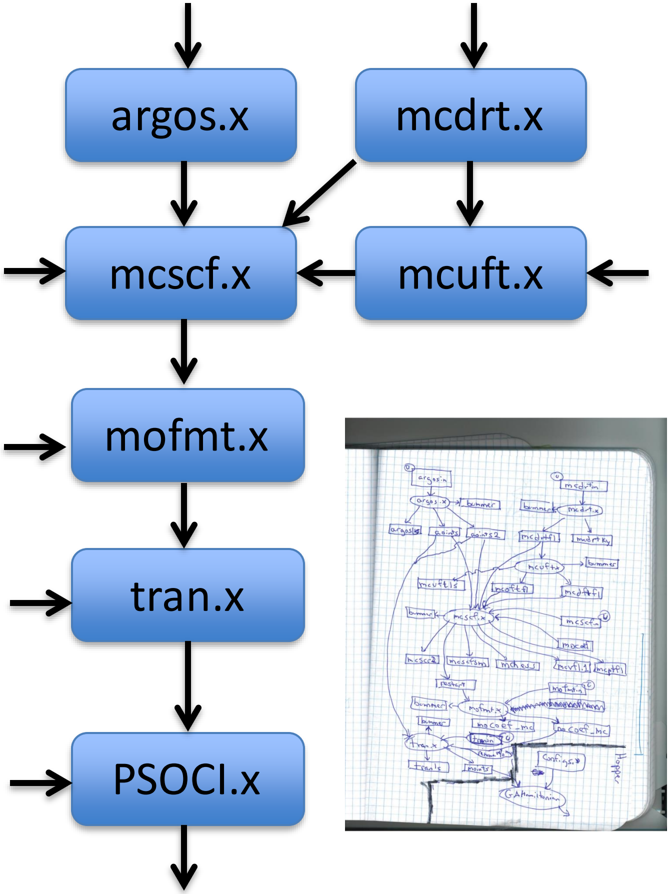
Toward Science on GENI

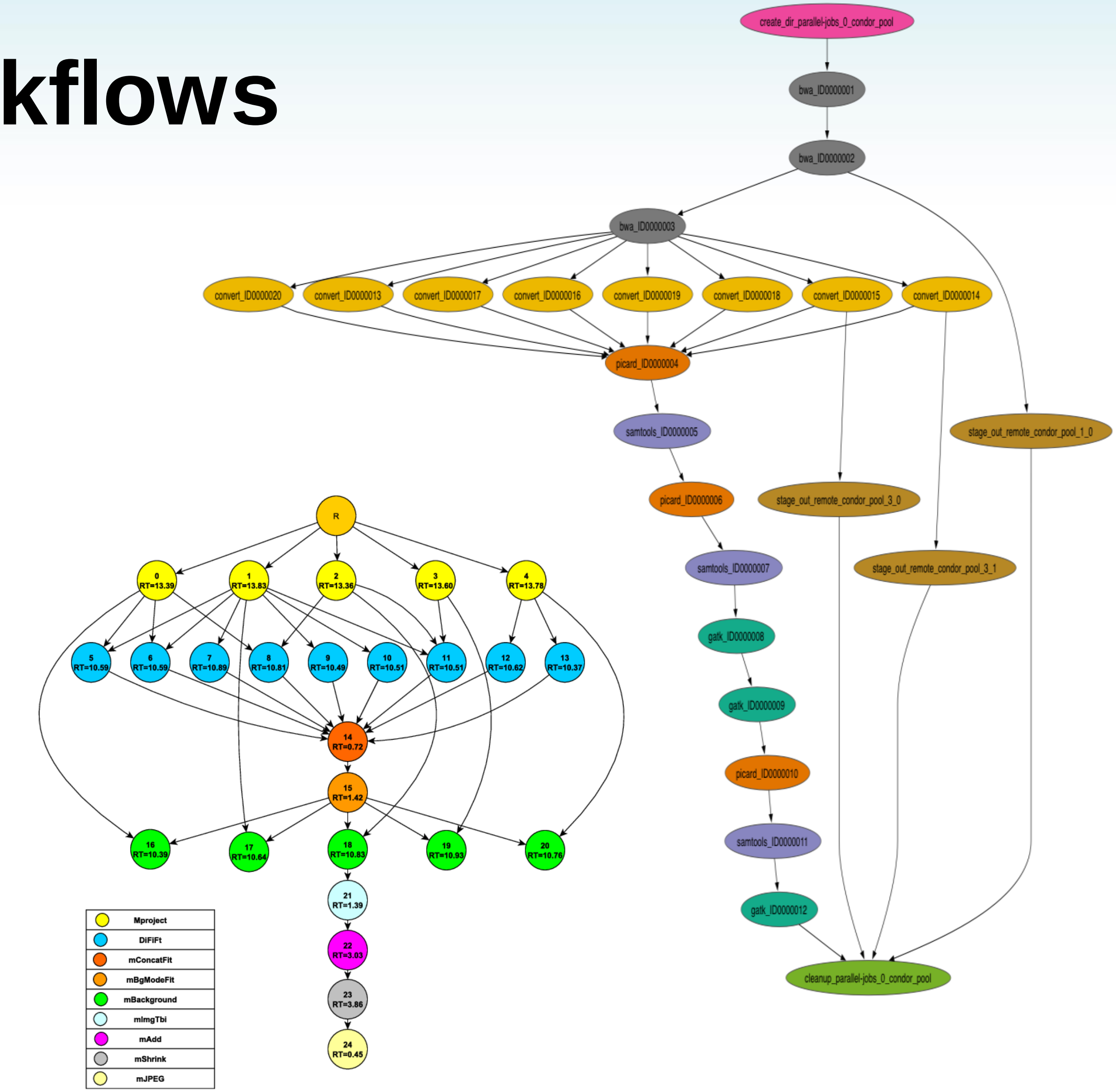
- Initial Effort (2011)
  - Solar Fuels
- Adaptive Applications/Workflows (2013-present)
  - ADAMANT, RADII, Panorama
- Scaling Science Experiments (2013-2015)
  - GENI Science Shakedown
- Multidomain Science (2016-present)
  - SAFE, SciDAS

Adaptive Applications/Workflows

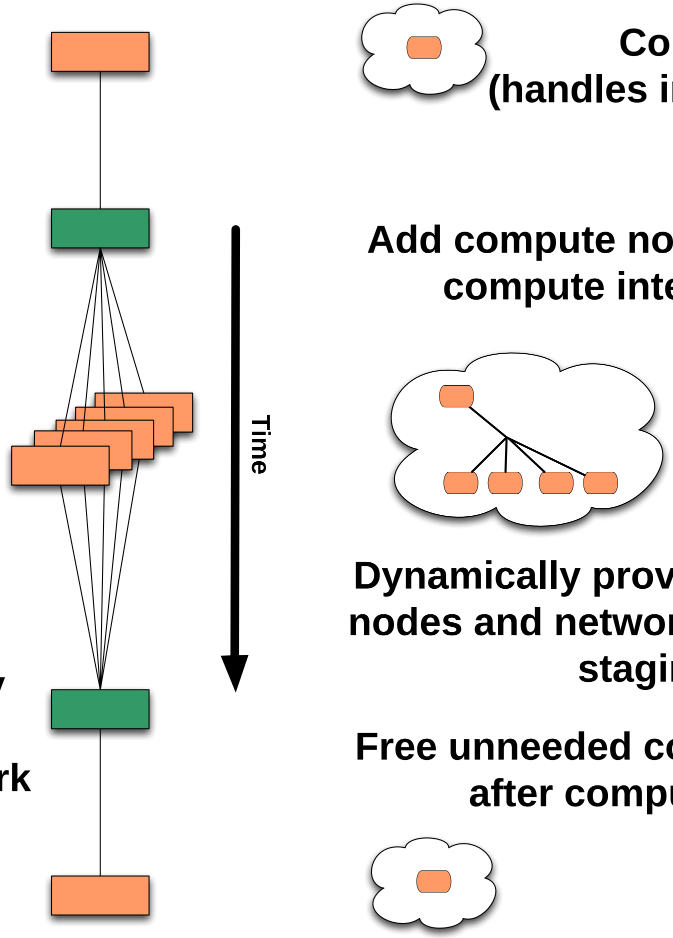
- Workflow Management Systems
  - Pegasus, custom scripts, etc.
- Lack of tools to integrate with dynamic infrastructures
  - Orchestrate the infrastructure in response to application
  - Integrate data movement with workflows for optimized performance
  - Manage application in response to infrastructure
- Scenarios
  - Computational with varying demands
  - Data-driven with large static data-set(s)
  - Data-driven with large amount of input/output data

Scientific Workflows





Scientific Workflows



Workflow

1. Start workflow
2. Dynamically create compute nodes
3. Network intensive workflow staging
4. Dynamically destroy compute nodes and provisioned netowork
5. End workflow

Dynamic Slice

Condor Head Node
(handles initial workflow staging)

Add compute nodes for parallel compute intensive step

Dynamically provision compute nodes and network for workflow staging

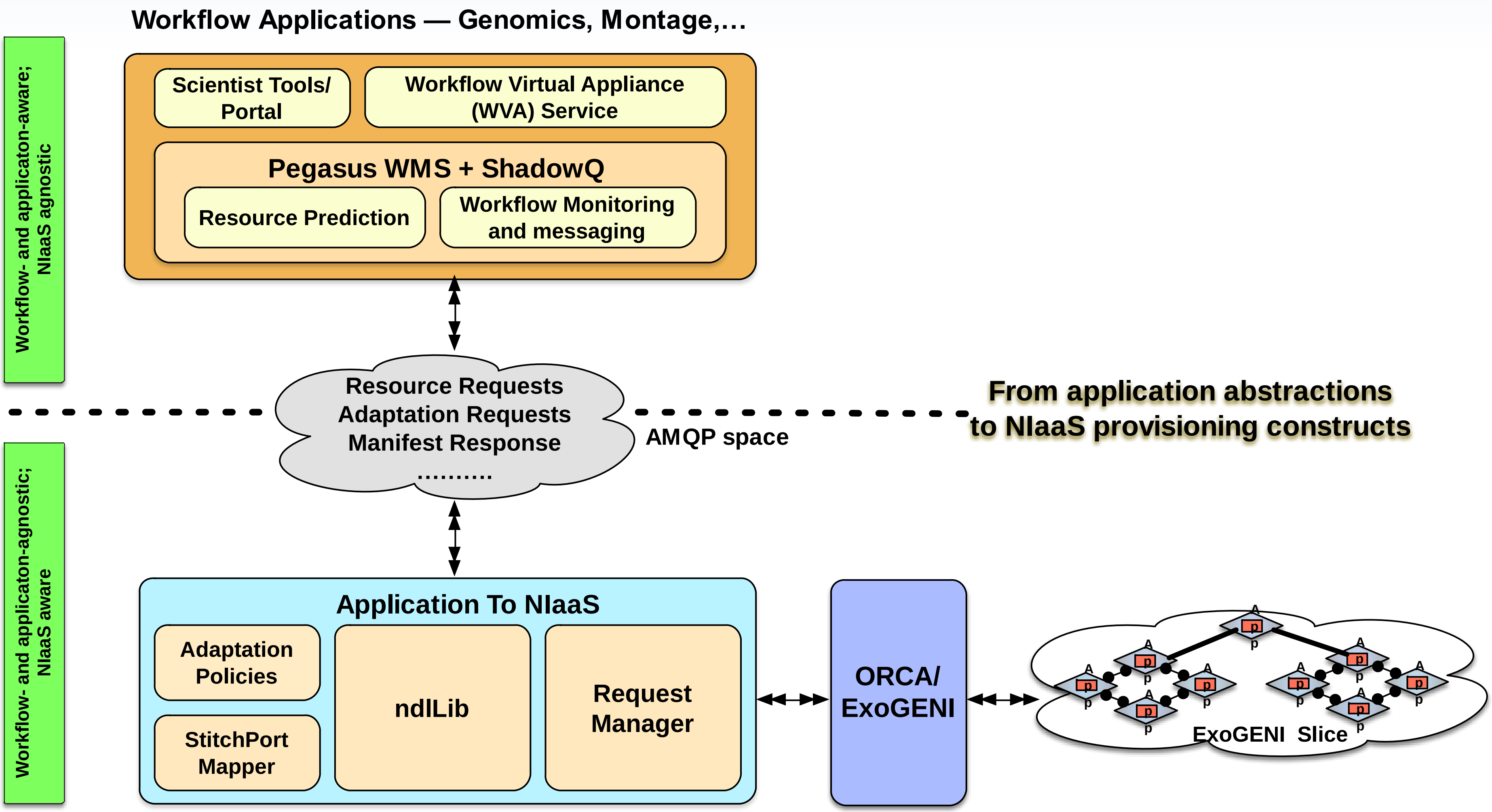
Free unneeded compute nodes after compute step

Pegasus/ExoGENI Collaboration

- Network Infrastructure-as-a-Service (NIaaS) for workflow-driven compute applications.
  - Tools for workflows integration with adaptive infrastructure (ExoGENI).
- Workflows triggering adaptive infrastructure
  - Pegasus workflows using ExoGENI
  - Adapt to application demands
    - Compute
    - Storage
    - Network
  - Integrate data movement into NIaaS
    - On-ramps
  - Target applications
    - Montage Galactic plane ensemble: Astronomy mosaics
    - CyberShake: Probabilistic Seismic Hazard Analysis
    - MapSeq: High-Throughput Sequencing

Multiple funded projects since 2013

Pegasus/ExoGENI



Workflow Applications — Genomics, Montage,…

Workflow- and applicaton-aware; NIaaS agnostic

Scientist Tools/ Portal

Workflow Virtual Appliance (WVA) Service

Pegasus WMS + ShadowQ

Resource Prediction

Workflow Monitoring and messaging

Resource Requests
Adaptation Requests
Manifest Response
……….

AMQP space

From application abstractions to NIaaS provisioning constructs

Workflow- and applicaton-agnostic; NIaaS aware

Application To NIaaS

Adaptation Policies

StitchPort Mapper

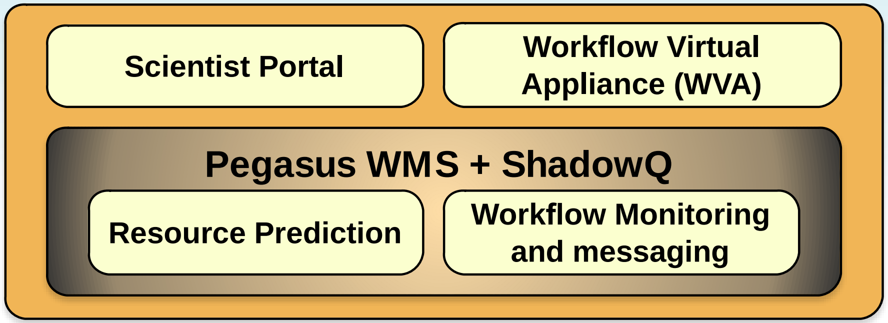
ndlLib

Request Manager

ORCA/ ExoGENI

ExoGENI Slice

Pegasus + ShadowQ



Scientist Portal

Workflow Virtual Appliance (WVA)

Pegasus WMS + ShadowQ

Resource Prediction

Workflow Monitoring and messaging

- ShadowQ
  - Monitors running workflow
  - Makes prediction about future events in the workflow
  - Communicates resource requirements to Request Manager
- Reconstructs current state of running workflow from logs
- Performs set of discrete event simulations to predict future events
- Periodically re-evaluates predictions by running additional simulations
- Estimates finish time of workflow based on current resources and determines the resource level to complete the workflow within a deadline

renci

Toward Science on GENI

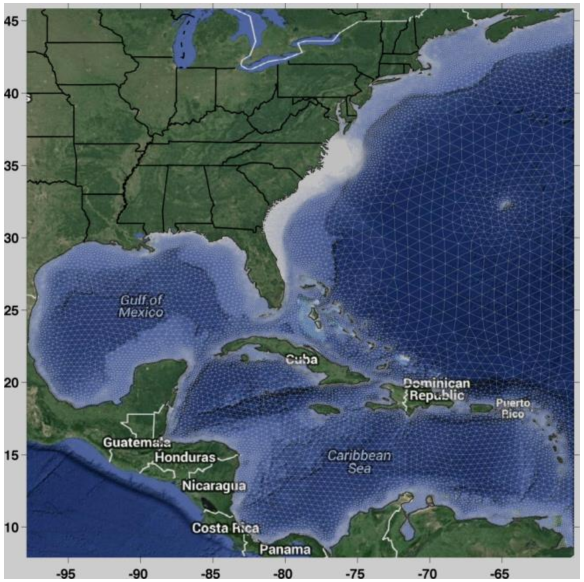
- Initial Effort (2011)
  - Solar Fuels
- Adaptive Applications/Workflows (2013-present)
  - ADAMANT, RADII, Panorama
- Scaling Science Experiments (2013-2015)
  - GENI Science Shakedown
- Multidomain Science (2016-present)
  - SAFE, SciDAS

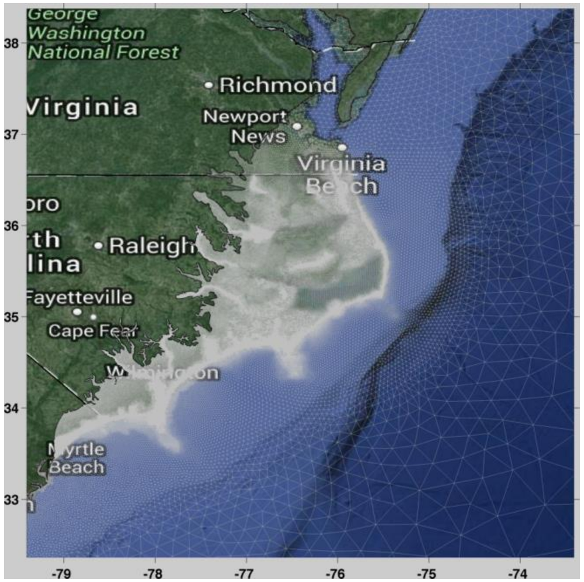
Scaling Science Experiments

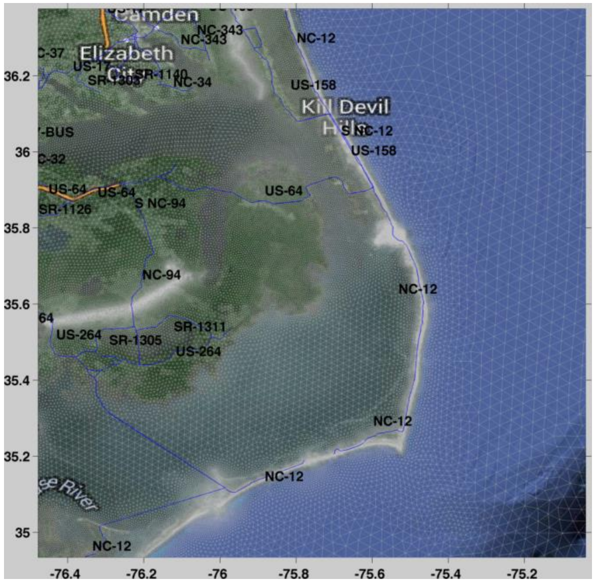
- GENI Science Shakedown
  - –GENI Spiral 4 project
- Project Goals
  - –Apply **the GENI experiment workflow** to domain science applications and evaluate GENI's ability to **run domain science experiments** (performance and ease-of-use).
  - –Build **tools** for domain scientists to create slices from high-level descriptions of high-throughput and high-performance applications.
  - –Provide **feedback to GENI rack developers** on the current capabilities with respect to science applications as well as target areas for improvement.

ADCIRC Storm Surge Model

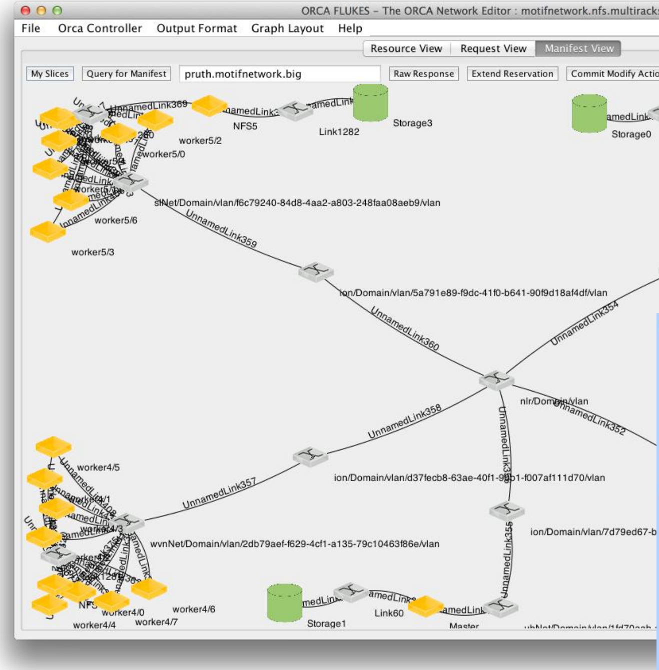
- Finite Element
- Very high spatial resolution (~1.2M triangles)
- Efficient MPI implementation, scales to thousands of cores
- Typically use 256-1024 cores for forecasting applications
- Used for coastal flooding simulations
  - FEMA flood insurance studies, Forecasting systems, Research applications
- Urgent computing
  - 3 hour turnaround requirement



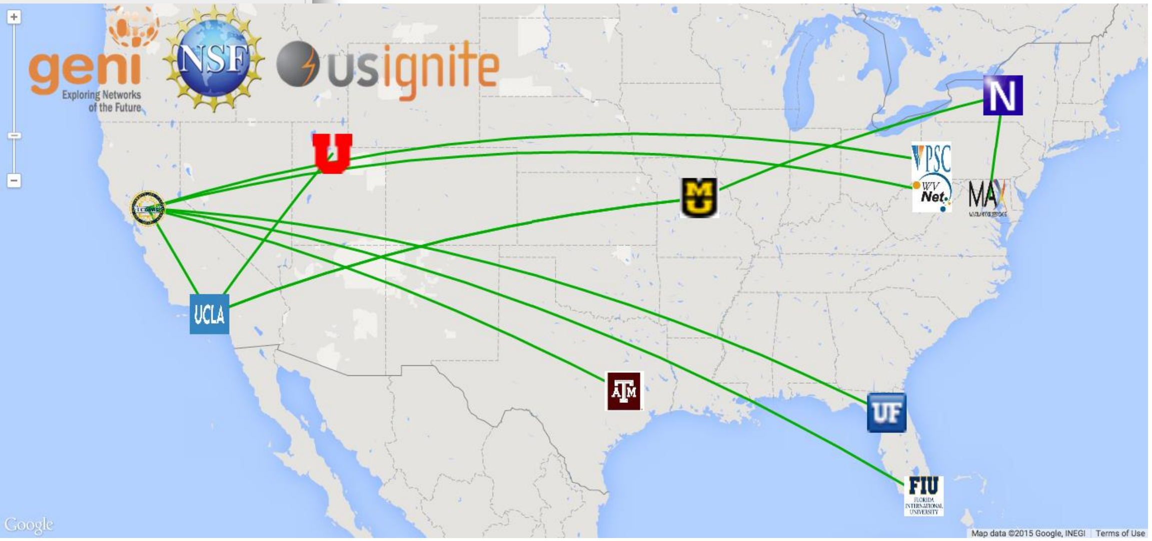




GEC22 Plenary Demo



Both GENI Testbeds (ExoGENI & InstaGENI)

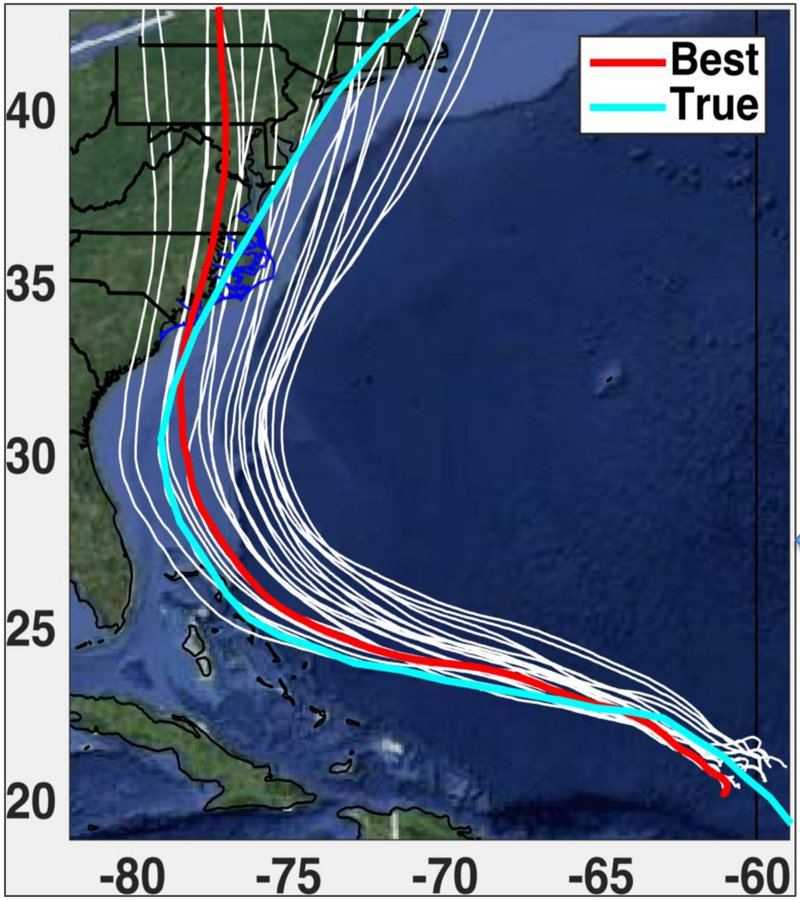

geni
Exploring Networks of the Future
NSF
usignite
U
UCLA
N
PSC
WV Net
MAX
FIU
UF
Google
Map data ©2015 Google, INEGI
Terms of Use

GENI Regional Workshop
University of Kentucky, Lexington, KY
May 14, 2018

Tackling Uncertainty

One simulation is NOT enough!

Probabilistic Assessment of Hurricanes



Research Ensemble

NSF Hazards SEES project

22 members, H. Floyd (1999)

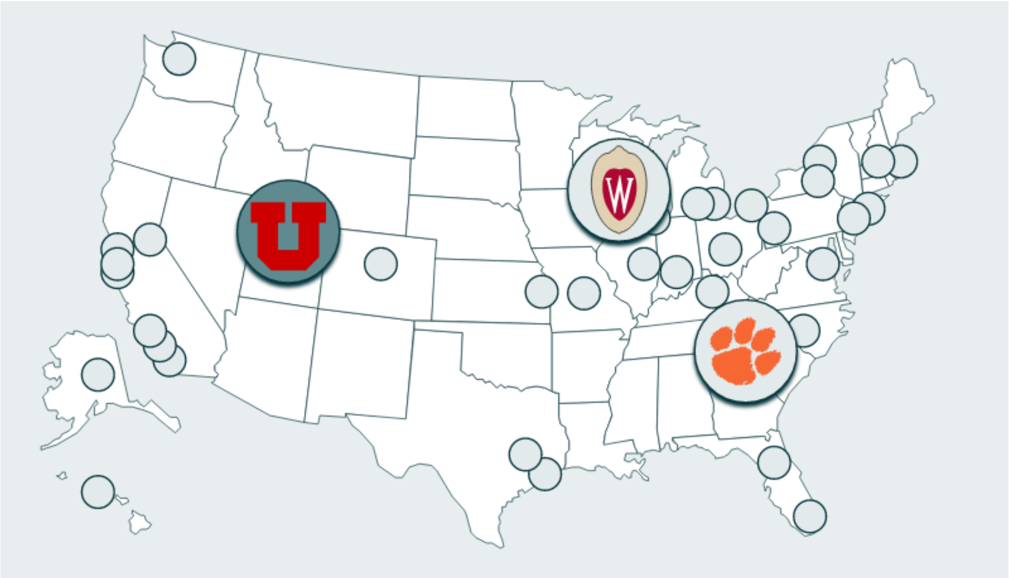
A “few” likely hurricanes

Fully dynamic atmosphere (WRF)

Mid-Scale Infrastructure

CloudLab

- Purpose
  - "Flexible, scientific infrastructure for research on the future of cloud computing. Researchers use CloudLab to build their own clouds, experimenting with new architectures that will form the basis for the next generation of computing platforms"
- Features
  - Scale: ~2000 cores/site
  - Baremetal nodes
  - Heterogeneous architectures
  - Low-latency networks
    - Infiniband
  - GENI API



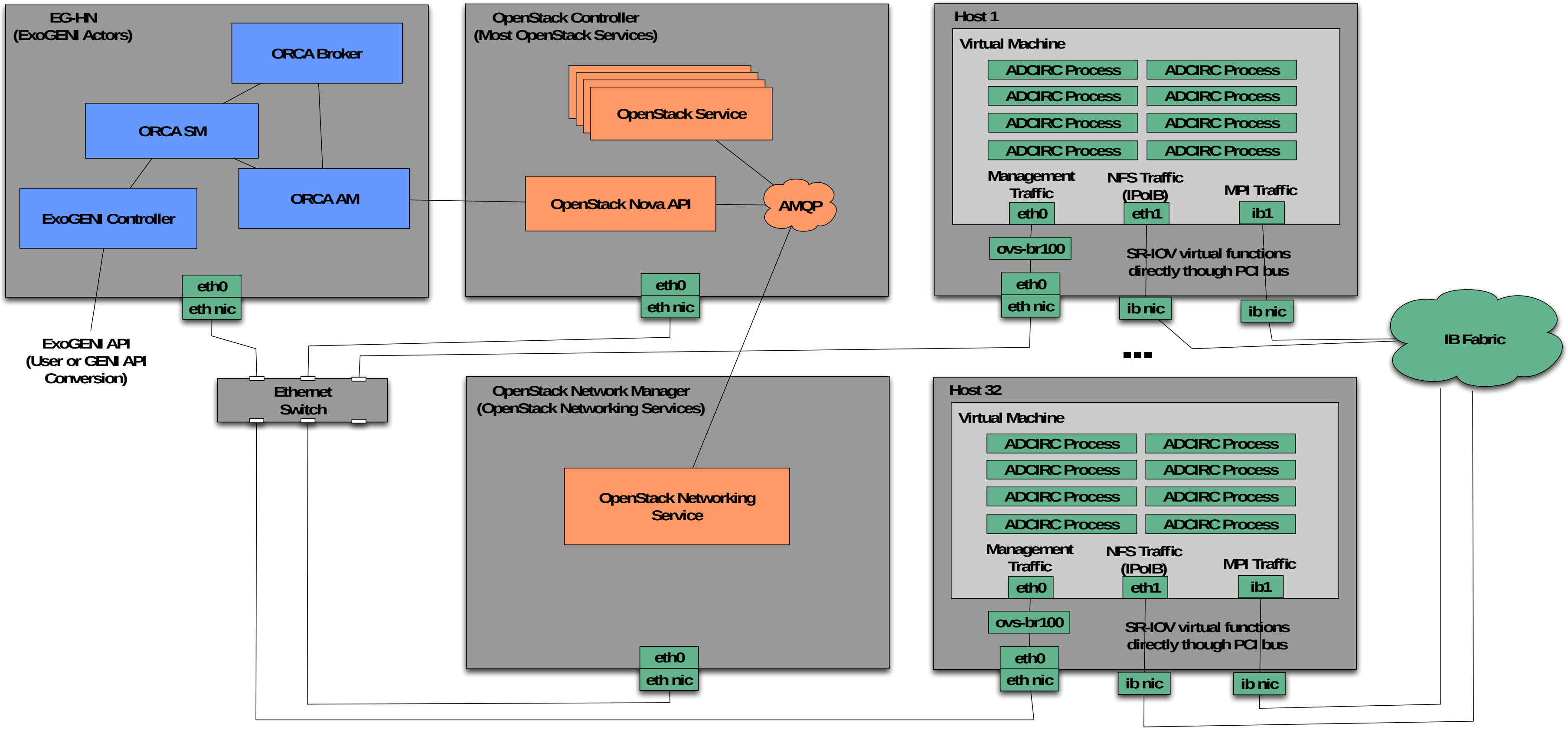
Chameleon

Today we could use Chameleon as well!

ExoGENI on CloudLab

- Goals
  - Experiment with ExoGENI at scale using CloudLab resources
  - Use CloudLab as a testbed for building federated clouds
  - Entice domain science communities to build private cloud federations using ORCA/ExoGENI.
- Completed
  - Stand-alone ExoGENI sites deployed on APT, Clemson, and Wisconsin CloudLab sites.
  - Have tried up to 64 nodes (512 cores)
  - SR-IOV Infiniband access for VMs (APT site only)

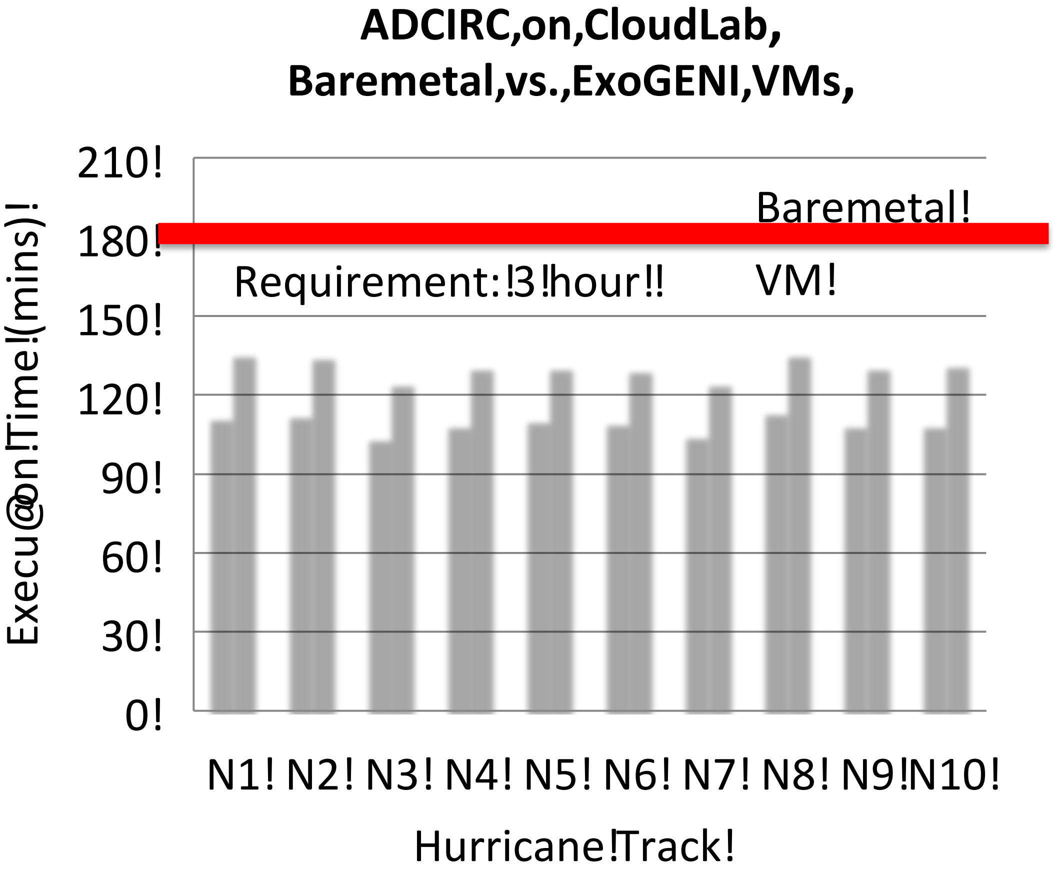
Architecture



GENI Regional Workshop
University of Kentucky, Lexington, KY
May 14, 2018

Performance Results

- Test
  - Baremetal CloudLab nodes vs. ExoGENI VMs on same hardware.
- Runs
  - Baremetal CloudLab nodes vs. ExoGENI VMs on same hardware.
  - 32 nodes (256 cores)
  - 10 input tracks
- GENI
  - ExoGENI ~12 hrs
  - InstaGENI ~18 hrs



Scaling Science Experiments

- Challenges
  - – Reliability. As you scale failure will happen.
  - – Performance capabilities of original sites.
- Lessons Learned
  - – We can use GENI to develop GENI
  - – GENI can support HPC
  - – Scripting, scripting, scripting!

Toward Science on GENI

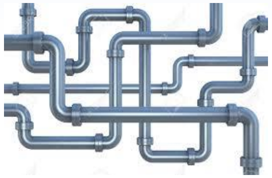
- Initial Effort (2011)
  - Solar Fuels
- Adaptive Applications/Workflows (2013-present)
  - ADAMANT, RADII, Panorama
- Scaling Science Experiments (2013-2015)
  - GENI Science Shakedown
- Multidomain Science (2016-present)
  - SAFE, SciDAS

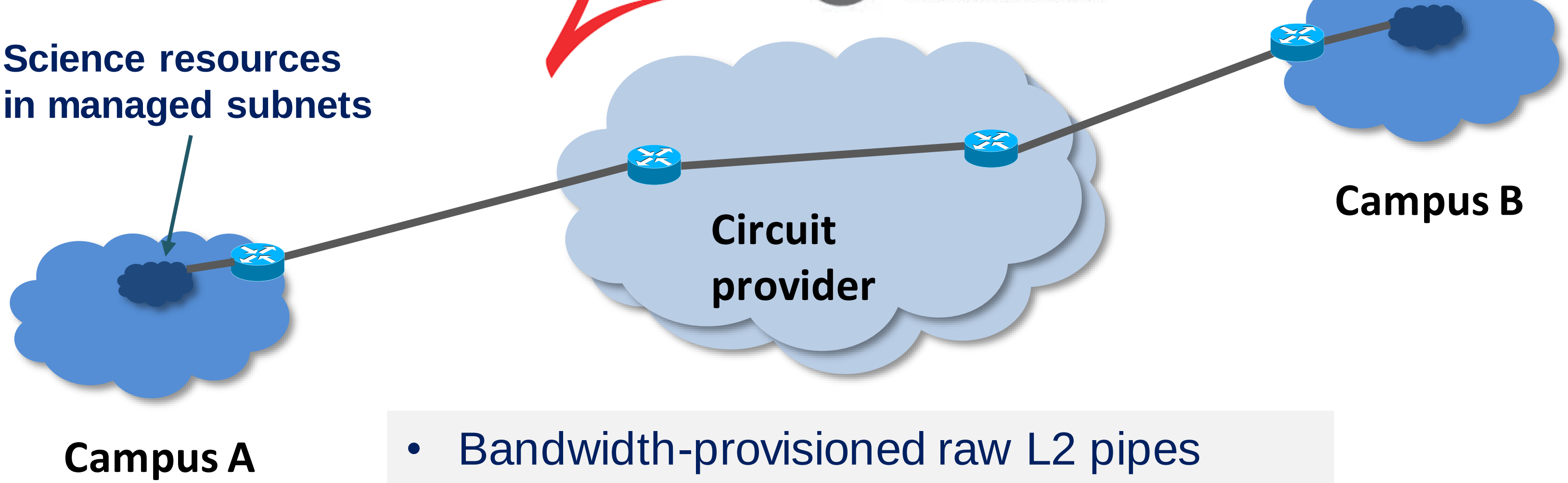
Multidomain Science

DOE coined the term "Superfacility"

- Definition
  - Two or more existing facilities (e.g. instruments, compute resources, data repositories) using high-performance networks and data management software in order to increase scientific output.
- Currently manually created
  - Superfacilities are purpose-built manually for a specific scientific application or community.
  - Trust: "handshake model"
- Ideally automated
  - Advanced Science DMZs and federated Infrastructure-as-a-Service provide the technical building blocks to construct dynamic superfacilities on demand.

Foundation: network circuit fabrics





Science resources in managed subnets

Circuit provider

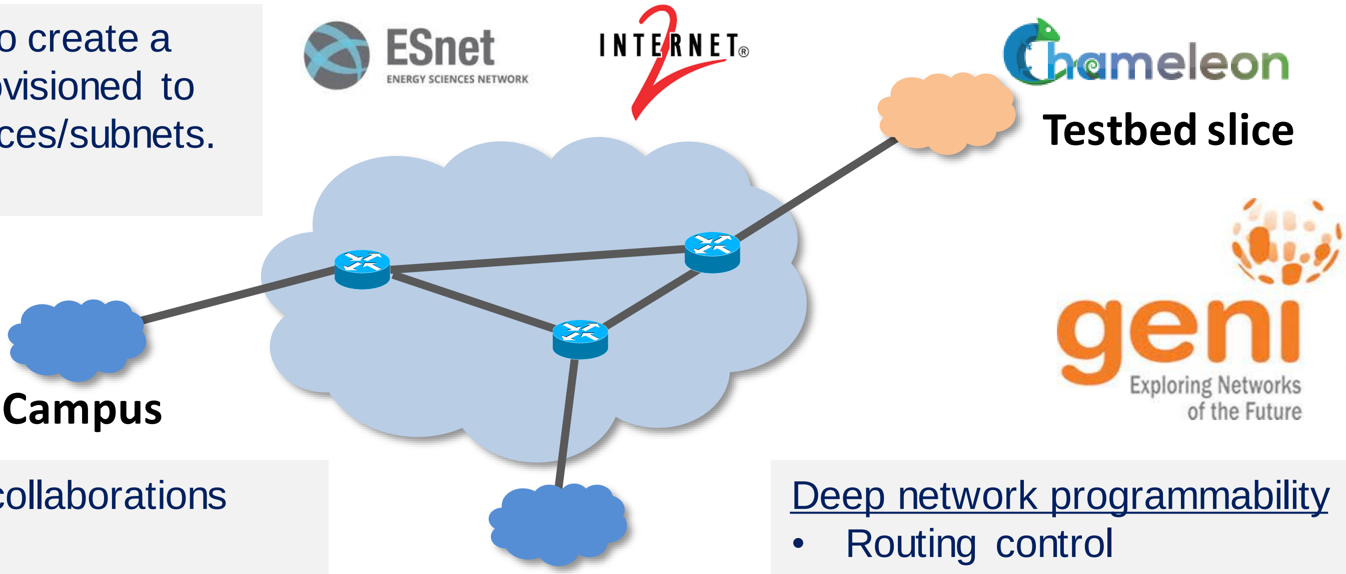
Campus A

Campus B

- Bandwidth-provisioned raw L2 pipes
- Dynamic on demand: edge to edge
- Programmatic hands-free IaaS APIs OSCARS→OESS→NSI

Virtual Science Networks

Idea: Use circuits to create a private network provisioned to link multiple resources/subnets. **Superfacility**



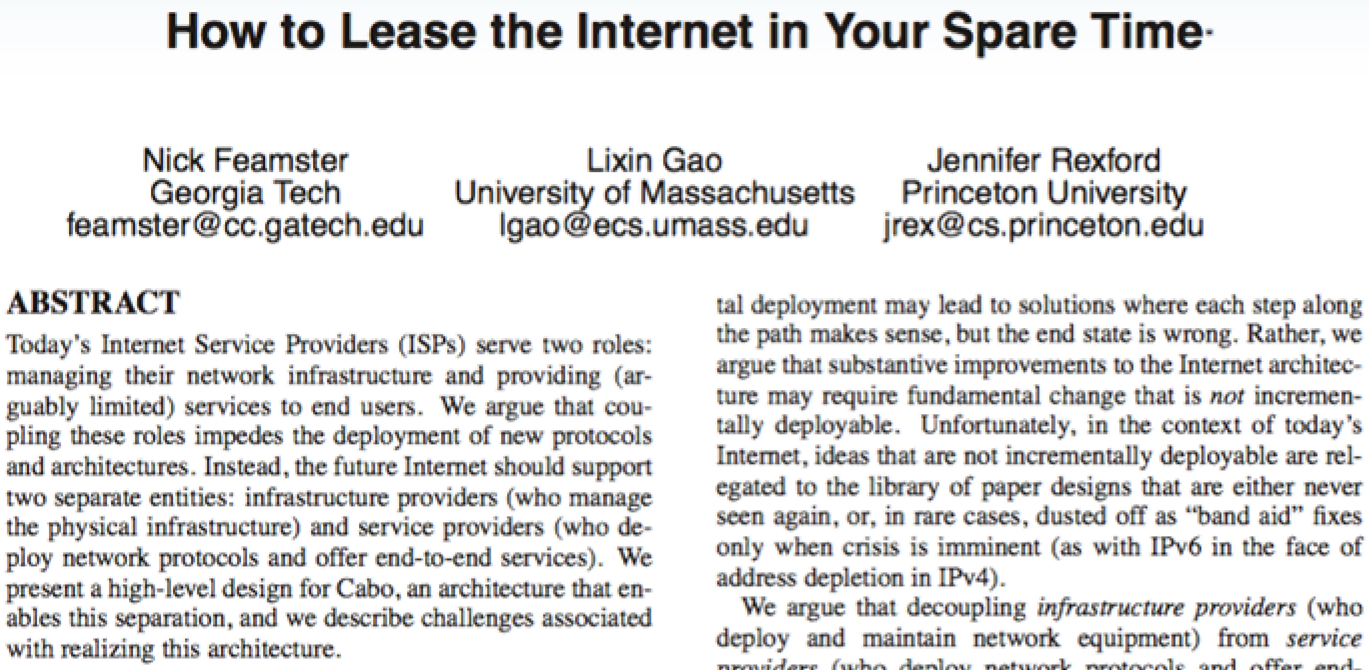
Large Computational Facility or Instrument

- Cross-campus collaborations
- Facility access
- **Examples**: cluster, LCF, testbed
- Resource sharing
- Virtual data enclaves
- Live network services

Deep network programmability

- Routing control
- Security monitoring / NFV
- Dynamic peering
- Topology adaptation
- Elastic edge clouds

Concurrent Architectures



How to Lease the Internet in Your Spare Time

Nick Feamster
Georgia Tech
feamster@cc.gatech.edu

Lixin Gao
University of Massachusetts
lgao@ecs.umass.edu

Jennifer Rexford
Princeton University
jrex@cs.princeton.edu

ABSTRACT

Today's Internet Service Providers (ISPs) serve two roles: managing their network infrastructure and providing (arguably limited) services to end users. We argue that coupling these roles impedes the deployment of new protocols and architectures. Instead, the future Internet should support two separate entities: infrastructure providers (who manage the physical infrastructure) and service providers (who deploy network protocols and offer end-to-end services). We present a high-level design for Cabo, an architecture that enables this separation, and we describe challenges associated with realizing this architecture.

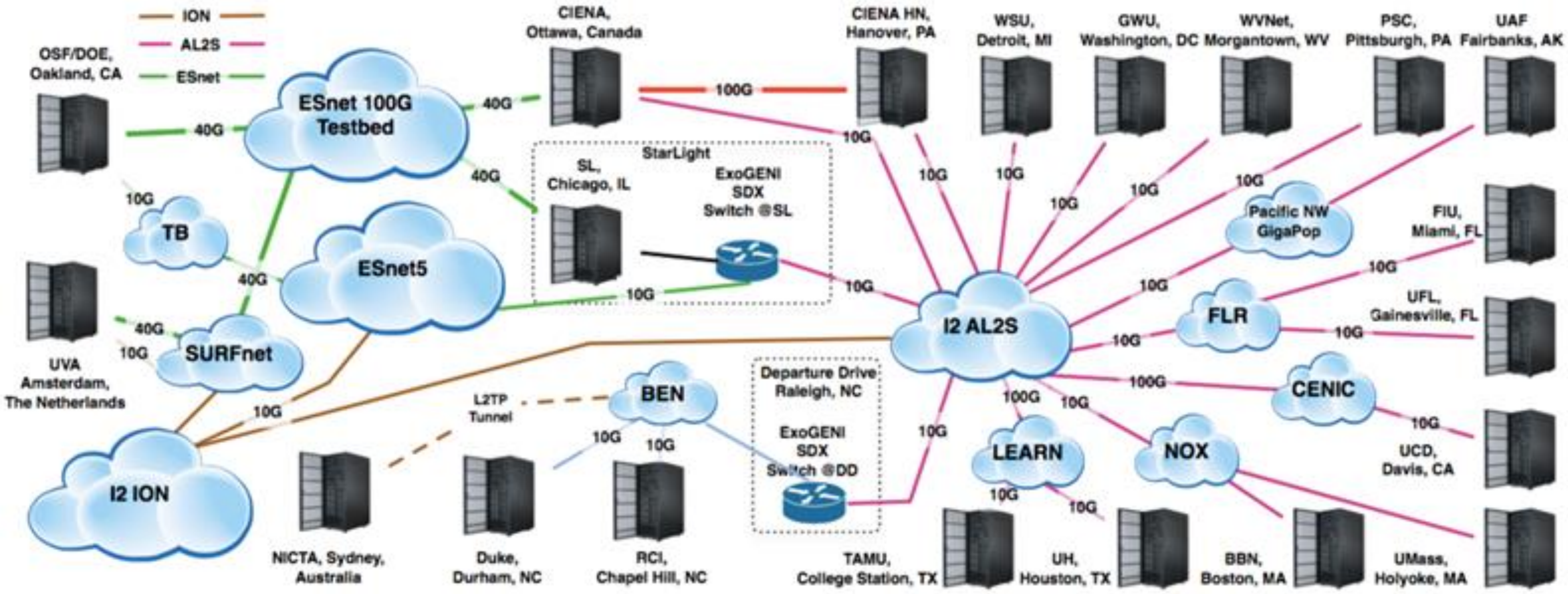
tal deployment may lead to solutions where each step along the path makes sense, but the end state is wrong. Rather, we argue that substantive improvements to the Internet architecture may require fundamental change that is *not* incrementally deployable. Unfortunately, in the context of today's Internet, ideas that are not incrementally deployable are relegated to the library of paper designs that are either never seen again, or, in rare cases, dusted off as "band aid" fixes only when crisis is imminent (as with IPv6 in the face of address depletion in IPv4).

We argue that decoupling *infrastructure providers* (who deploy and maintain network equipment) from *service providers* (who deploy network protocols and offer end-to-end services)[1] is the key to breaking this stalemate.

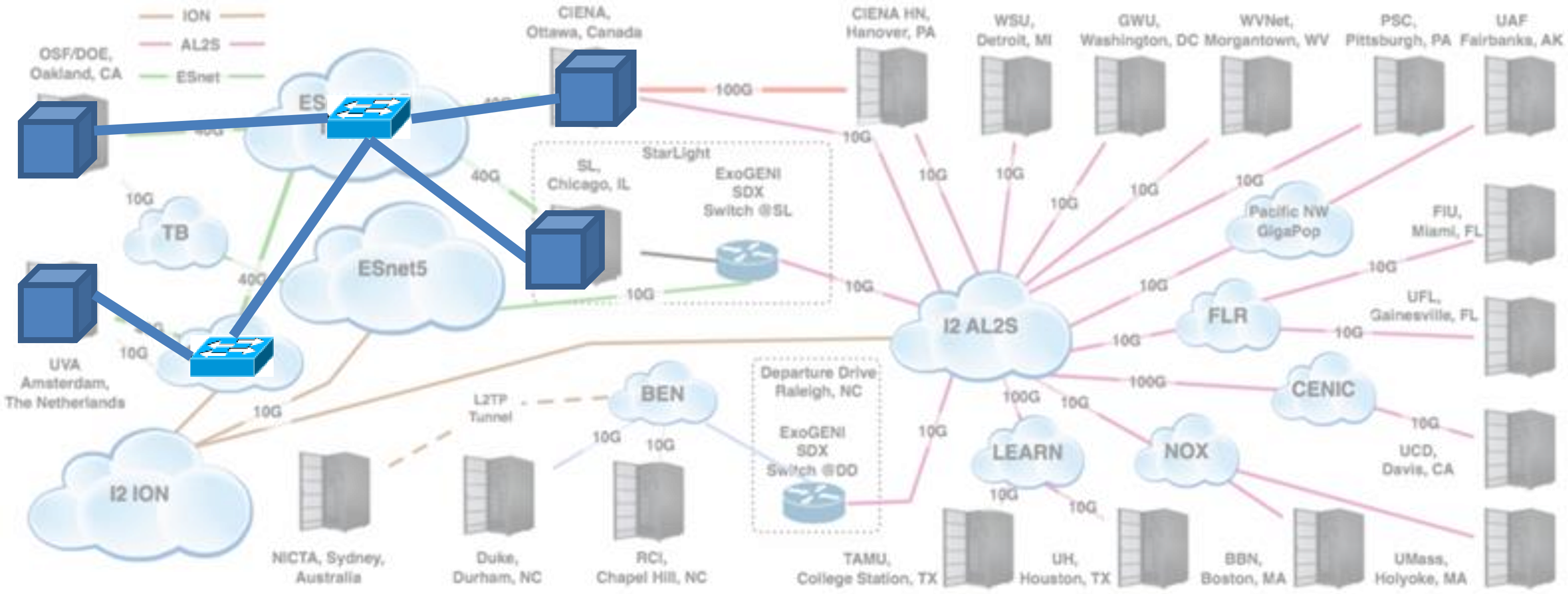
Cabo economic refactoring: "Decouple infrastructure providers from service providers, who offer end-to-end services."

Nick Feamster

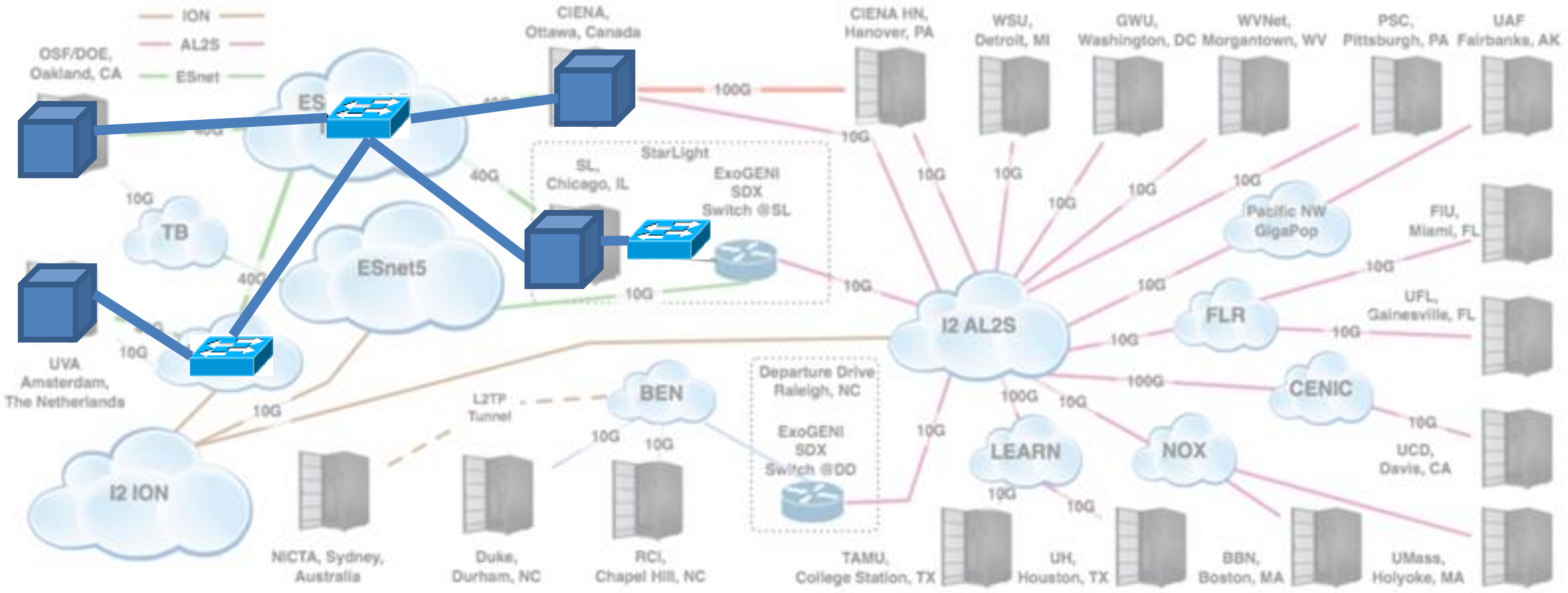
ExoGENI: Stitching



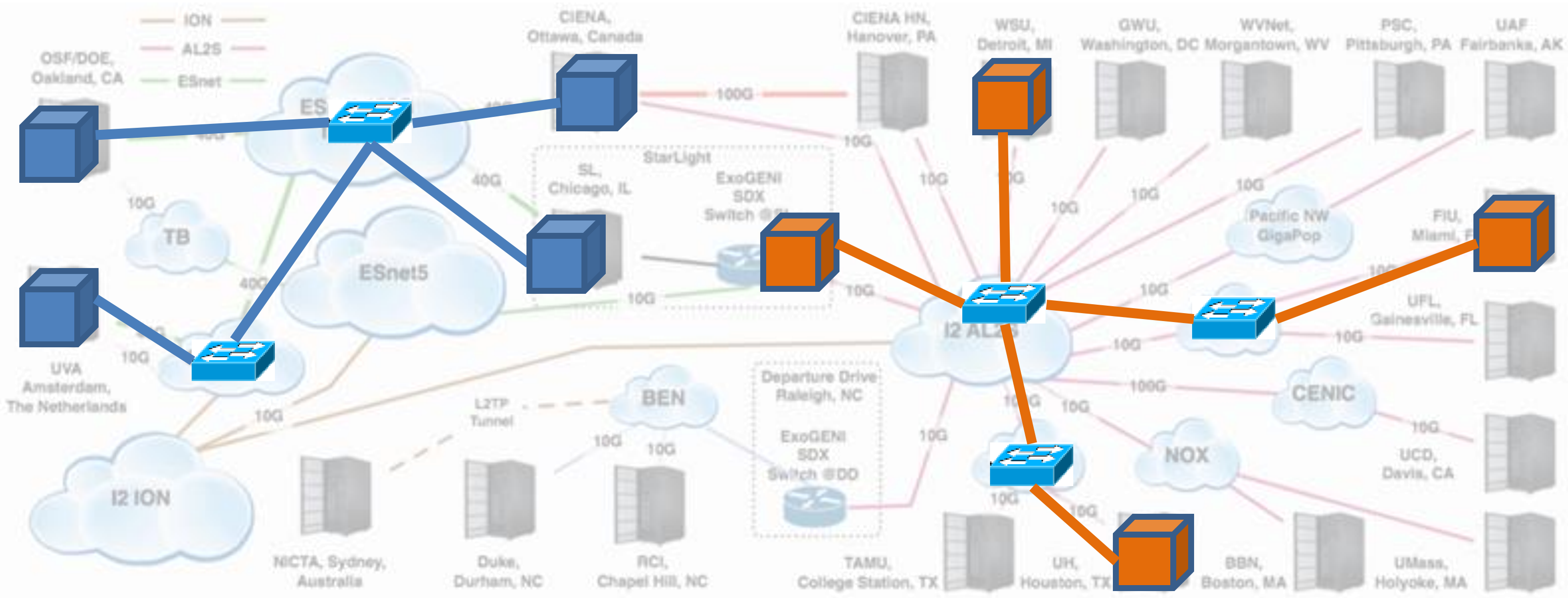
ExoGENI: Stitching



ExoGENI: Inter-Slice Stitching

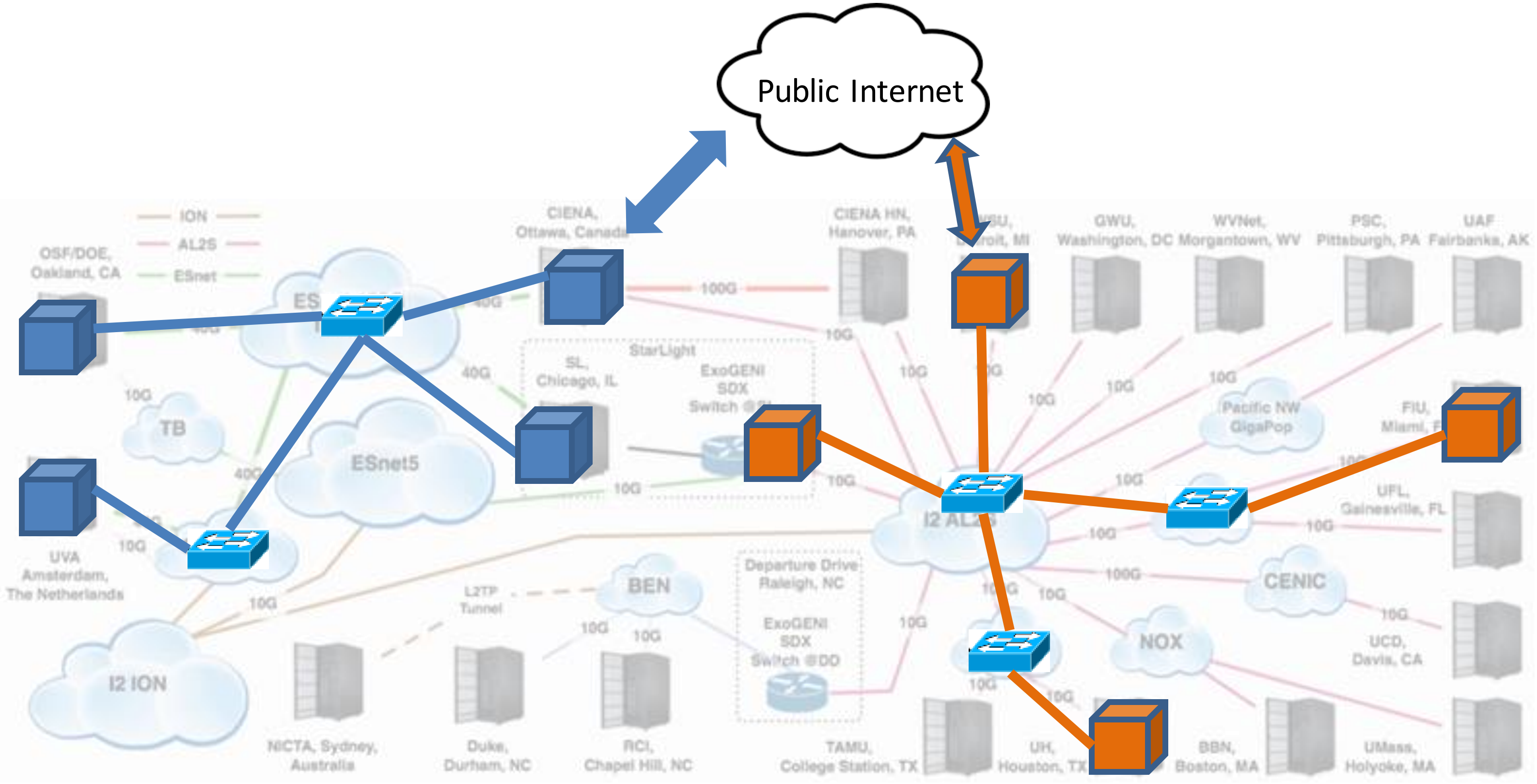


ExoGENI: Inter-Slice Stitching



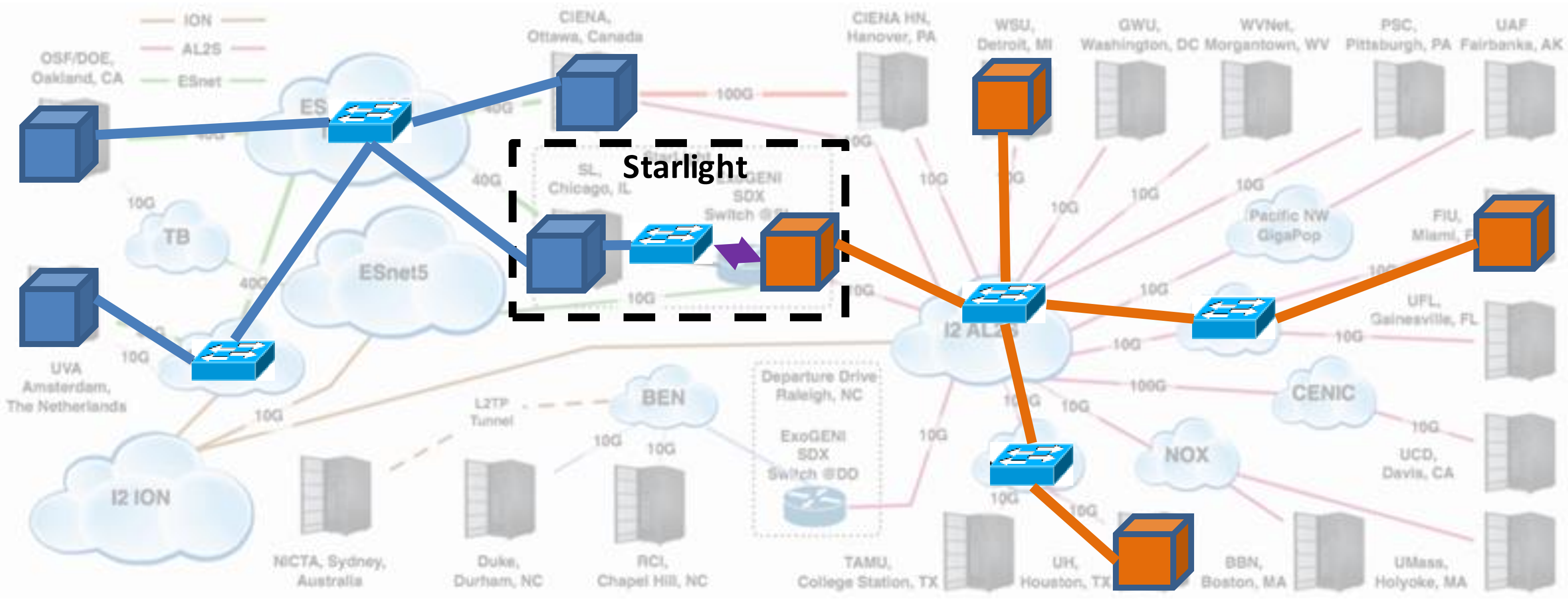
GENI Regional Workshop
University of Kentucky, Lexington, KY
May 14, 2018

ExoGENI: Inter-Slice Stitching



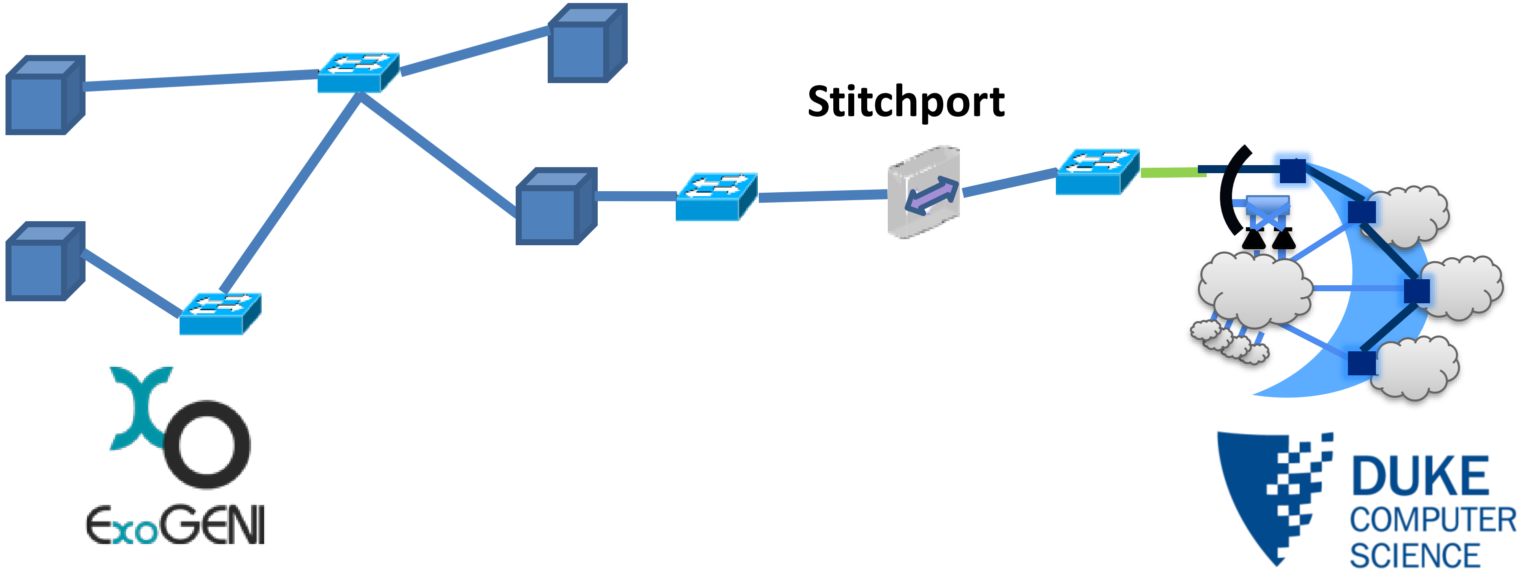
GENI Regional Workshop
University of Kentucky, Lexington, KY
May 14, 2018

ExoGENI: Inter-Slice Stitching

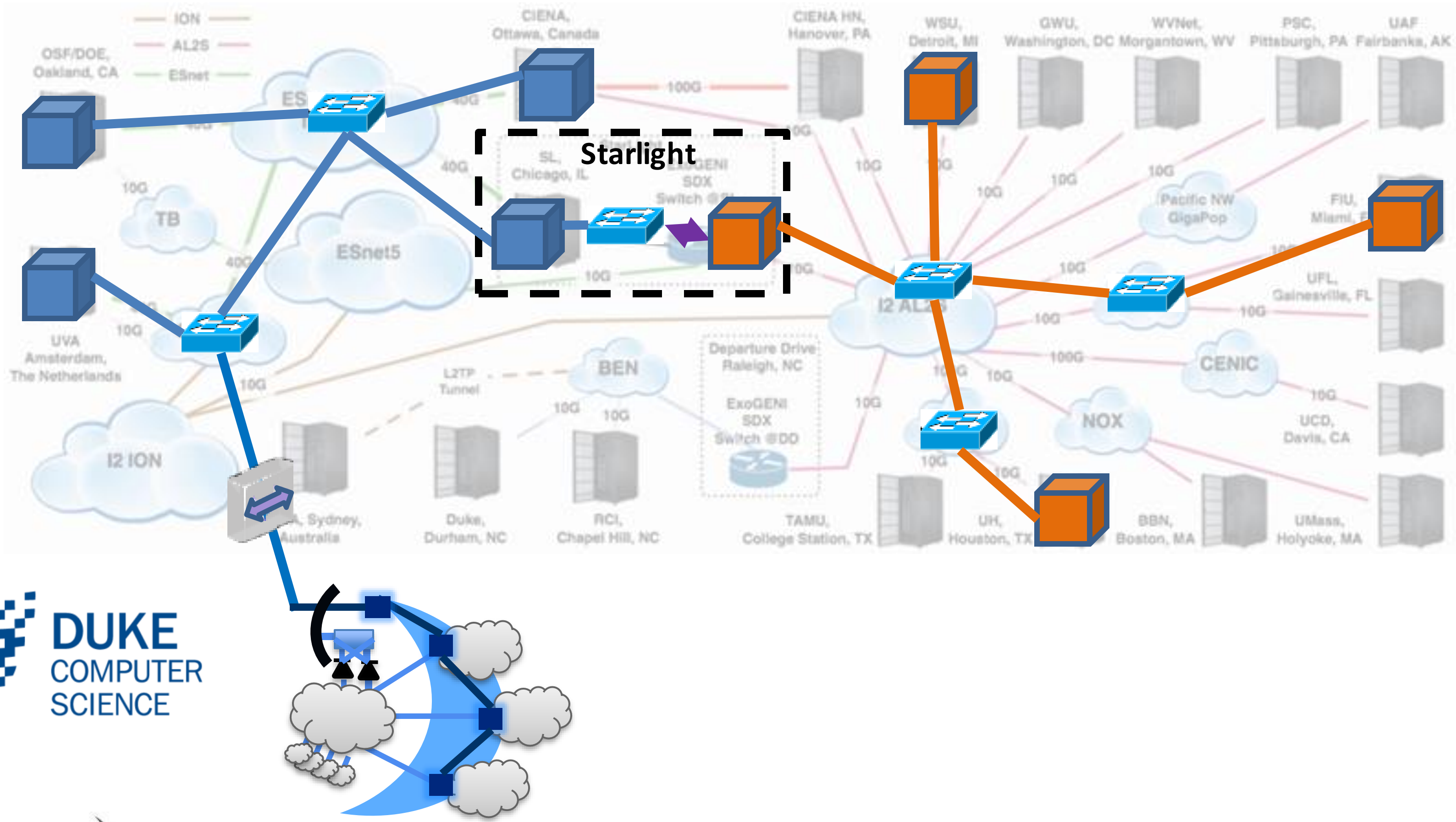


ExoGENI Campus Stitching

Stitchport: Named meeting point linking a layer 2 circuit between ExoGENI and external resources.



ExoGENI Slice-as-a-Service



GENI Regional Workshop
University of Kentucky, Lexington, KY
May 14, 2018

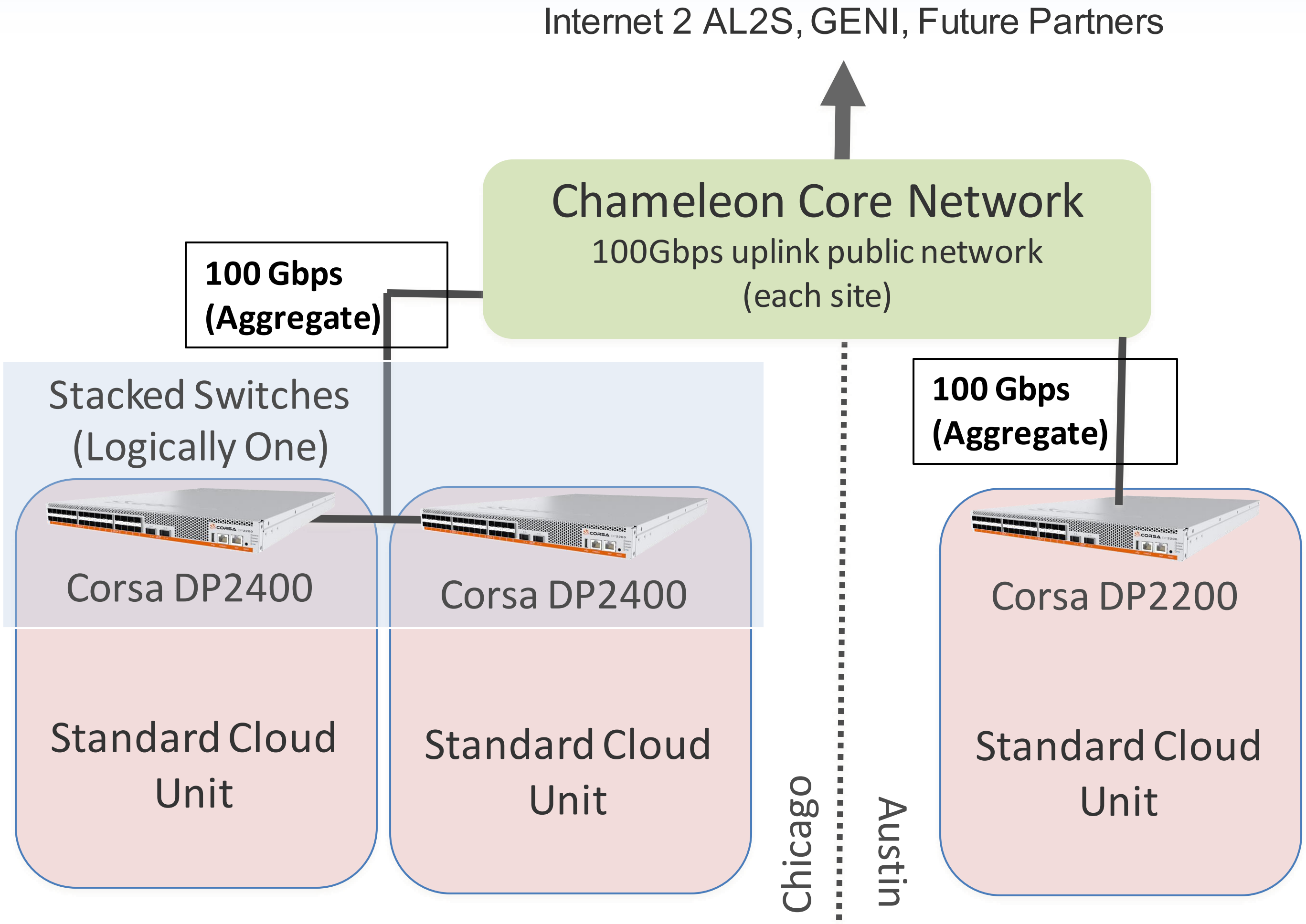
NEW Chameleon Hardware

- 4 new Standard Cloud Units (32 node racks in 2U chassis)
  - 3x Intel Xeon “Sky Lake” racks (2x @UC, 1x @TACC)
  - 1x future Intel Xeon rack (@TACC) in Y2
- Corsa DP2000 series switches
  - 2x DP2400 with 100Gbps uplinks (@UC)
  - 1x DP2200 with 100Gbps uplink (@TACC)
  - Each switch will have a 10 Gbps connection to nodes in the SCU
  - Optional Ethernet connection in both racks
- More storage configurations
  - Global store @UC: 5 servers with 12x10TB disks each
  - Additional storage @TACC: 150 TB of NVMes
- Accelerators: 16 nodes with 2 Volta GPUs (8@UC, 8@TACC)
- Maintenance, support and reserve

Corsa DP2000 Series Switches

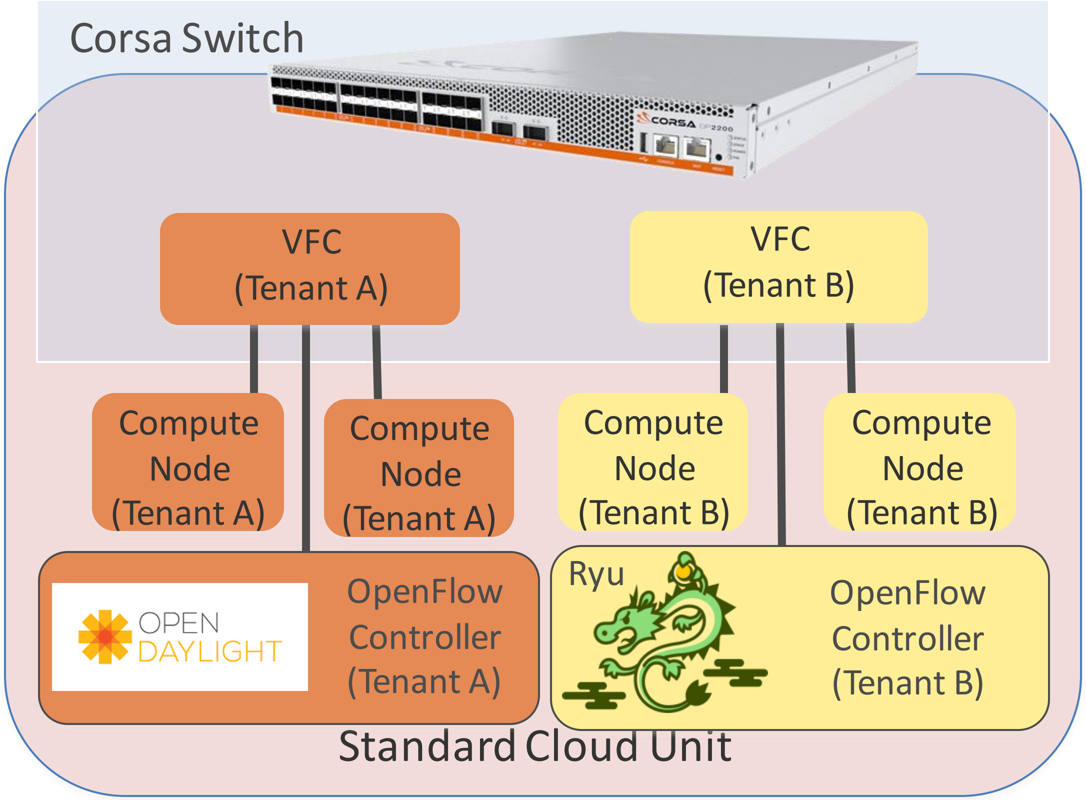
- Hardware Network Isolation
  - Sliceable Network Hardware
  - Tenant controlled Virtual Forwarding Contexts (VFC)
- Software Defined Networking (SDN)
  - OpenFlow v1.3
  - User defined controllers
- Performance
  - 10 Gbps within a site
  - 100 Gbps between UC/TACC (Aggregated)

Network Hardware



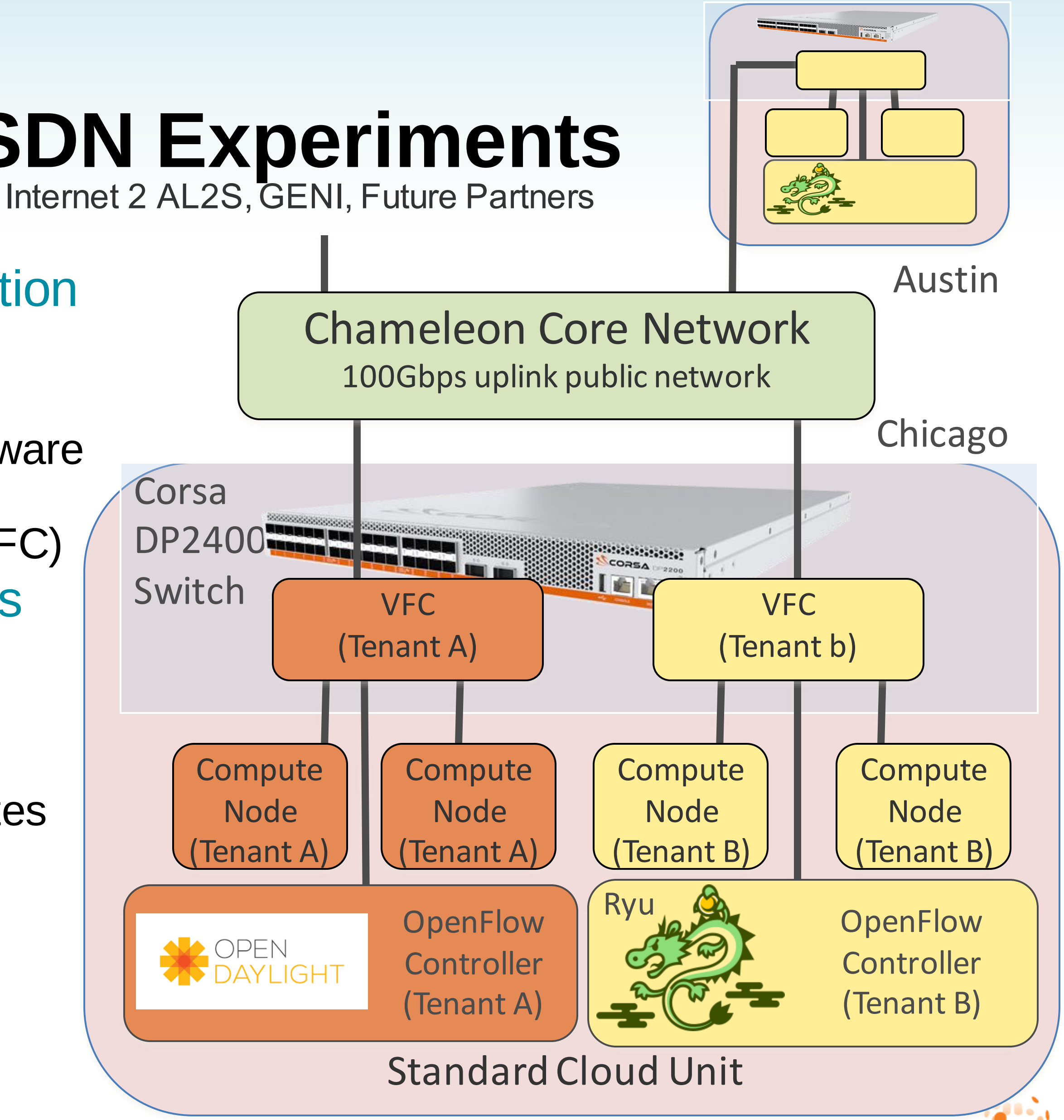
Isolated Virtual SDN Switch

- Isolated Tenant Networks
- BYOC– Bring your own controller: isolated user controlled virtual OpenFlow switches (coming soon)



Chameleon: SDN Experiments

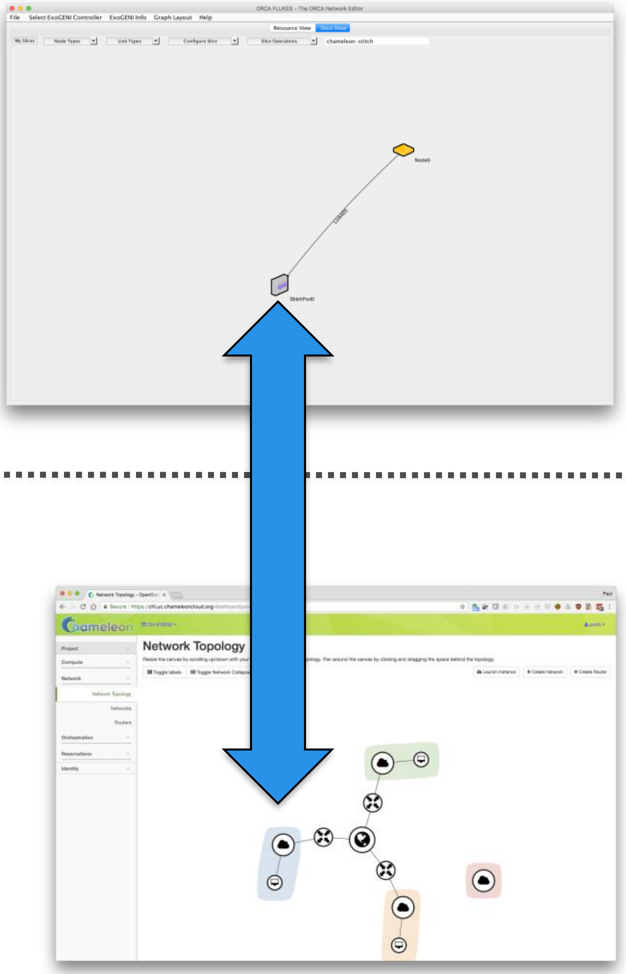
- Hardware Network Isolation
  - Corsa DP2000 series
  - OpenFlow v1.3
  - Sliceable Network Hardware
  - Tenant controlled Virtual Forwarding Contexts (VFC)
- Isolated Tenant Networks
  - BYOC – Bring your own controller
- Wide-area Stitching
  - Between Chameleon Sites (100 Gbps)
  - ExoGENI
  - Campus networks (ScienceDMZs)



GENI Regional Workshop
University of Kentucky, Lexington, KY
May 14, 2018

Chameleon to ExoGENI Stitching

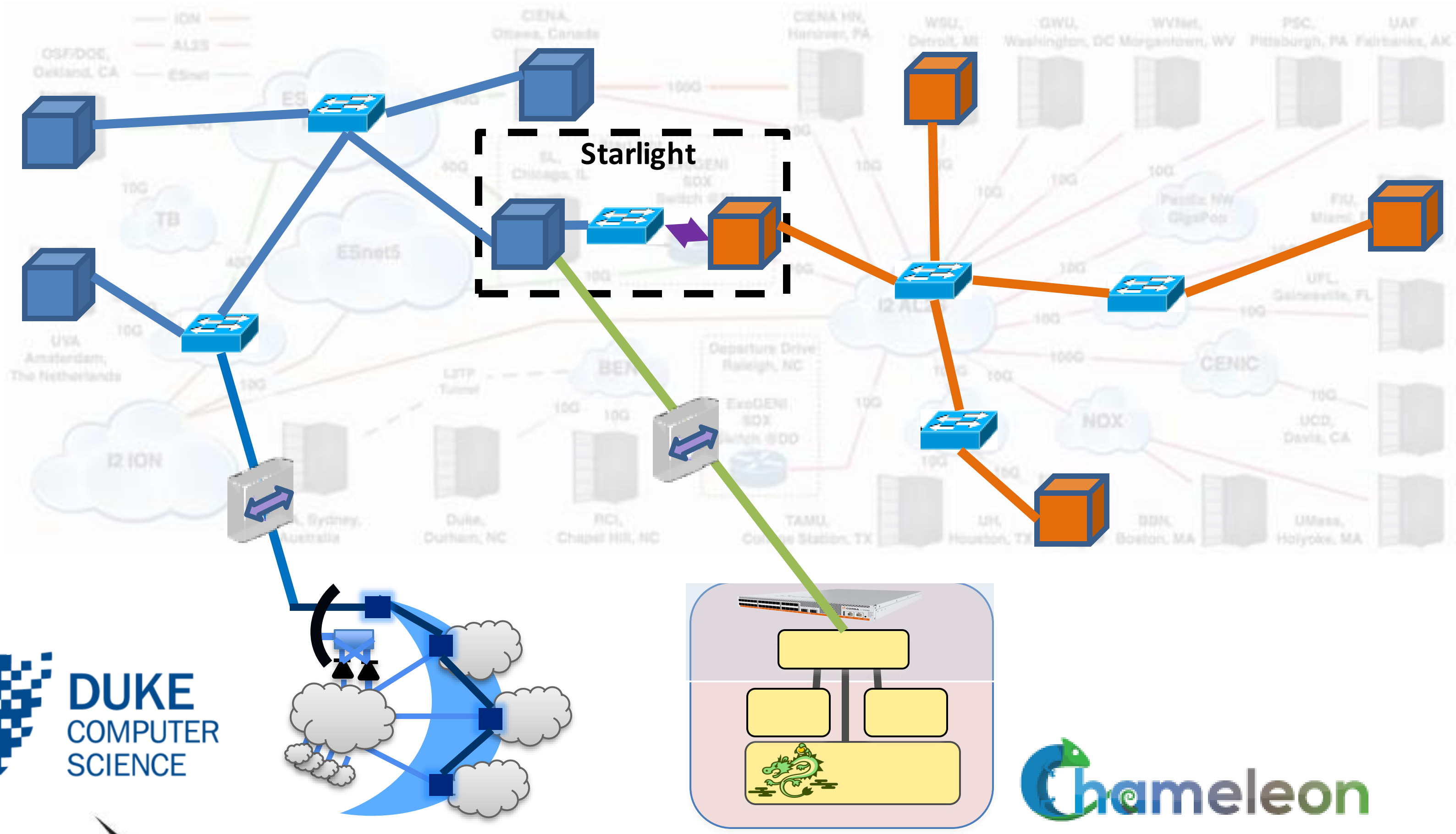
- ExoGENI slice
- Dynamic Chameleon Stitchport


ORCA FLUKES - The ORCA Network Editor
File Select ExoGENI Controller ExoGENI Info Graph Layout Help
Resource View
Slice View
chameleon-stitch
Node0
StitchPort0
Network Topology

- Dynamic VLANs
- Connectivity to ExoGENI Stitchport

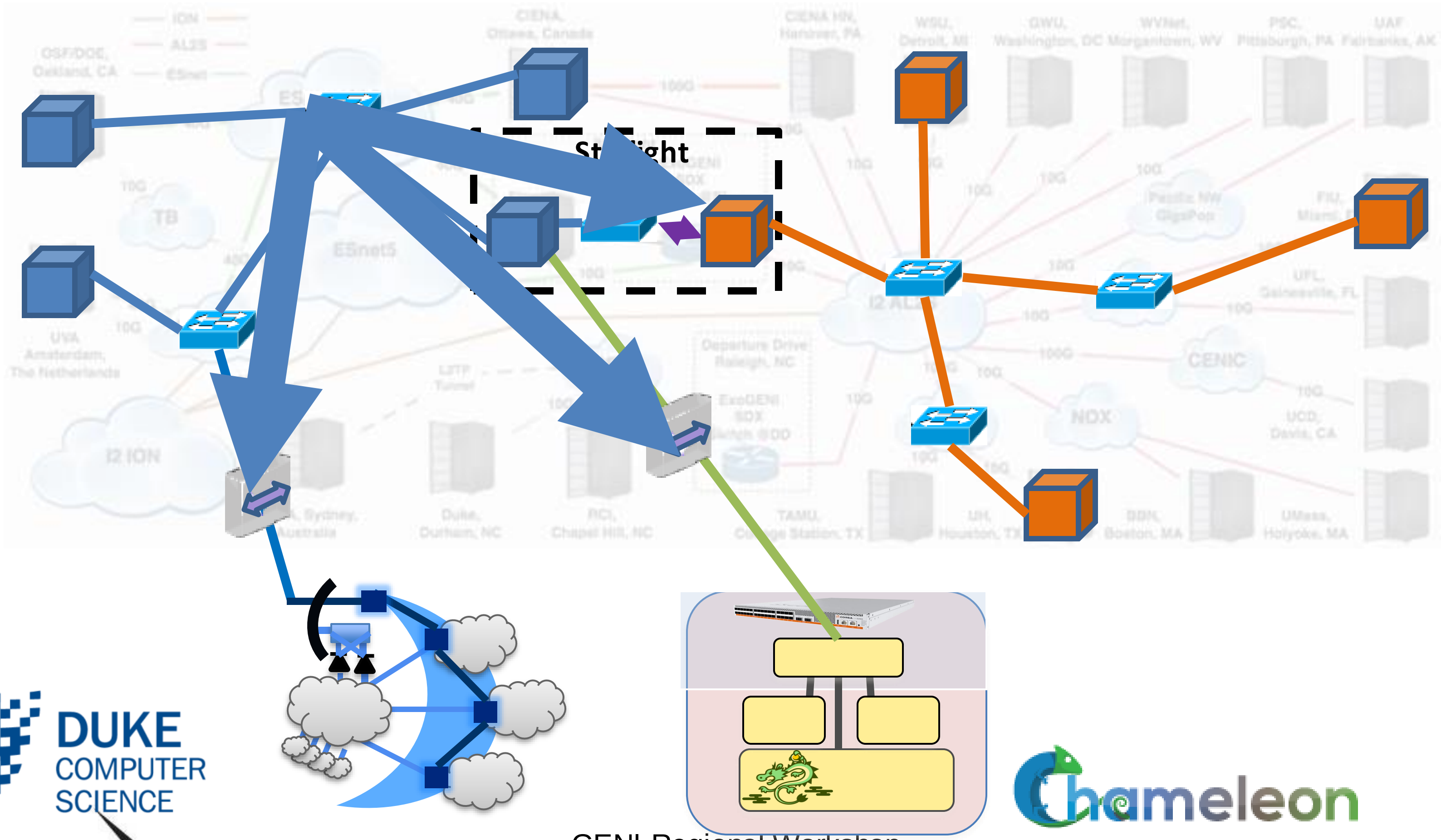
GENI Regional Workshop
University of Kentucky, Lexington, KY
May 14, 2018

Multi-Testbed Experiments



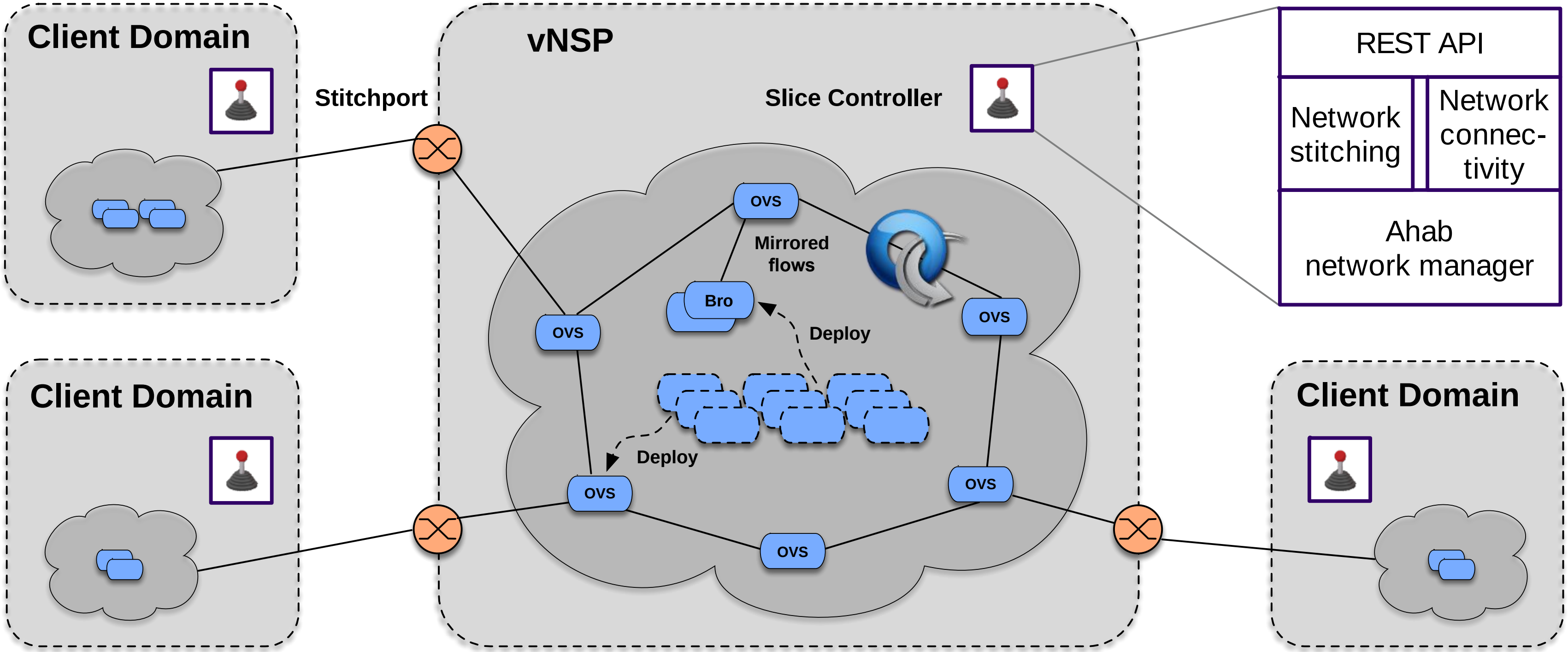
GENI Regional Workshop
University of Kentucky, Lexington, KY
May 14, 2018

Multi-Testbed Experiments



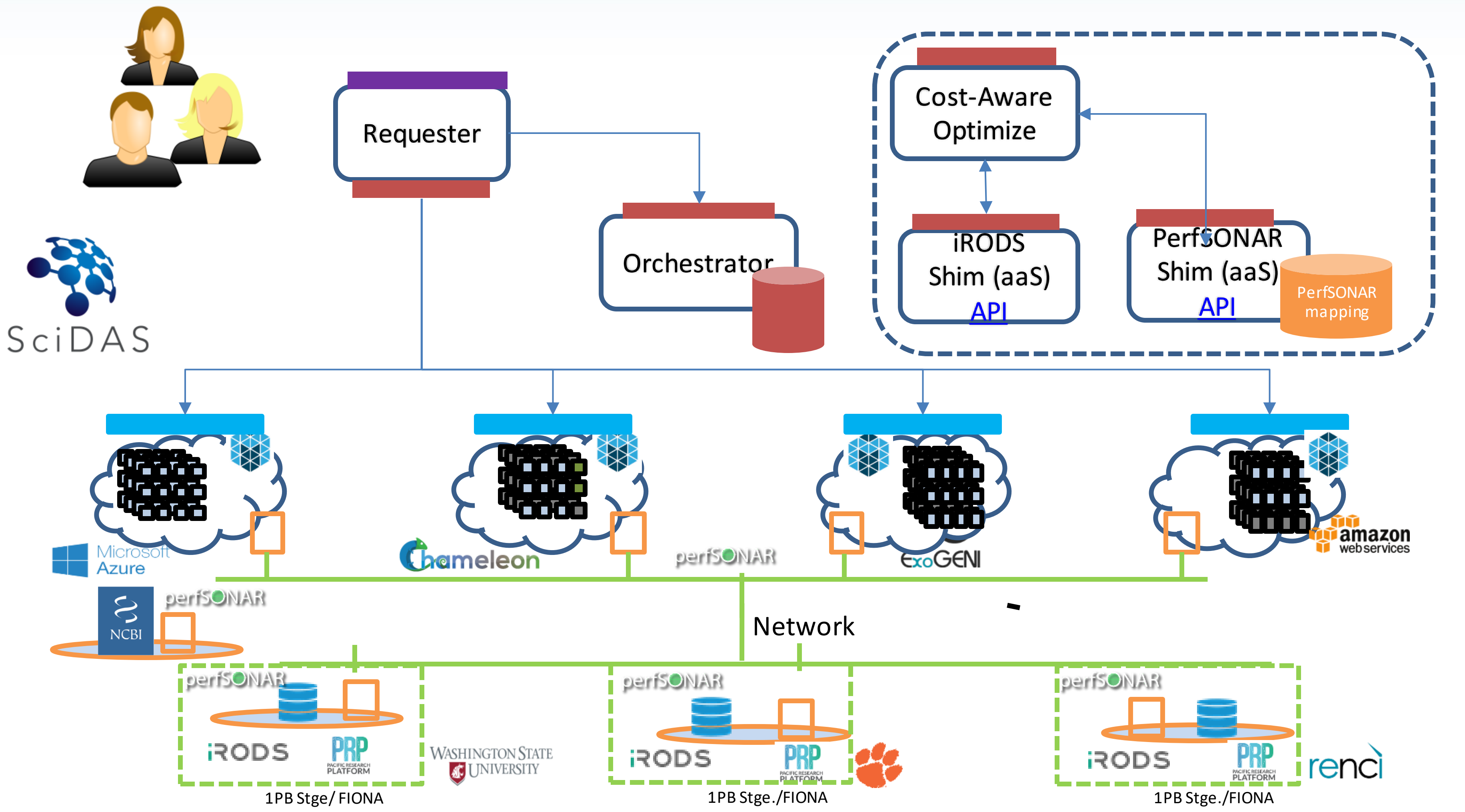
GENI Regional Workshop
University of Kentucky, Lexington, KY
May 14, 2018

SAFE Software Defined eXchange (SDX)



GENI Regional Workshop
University of Kentucky, Lexington, KY
May 14, 2018

Scientific Data Analysis at Scale (SciDAS)



renci

GENI Regional Workshop
University of Kentucky, Lexington, KY
May 14, 2018

geni Exploring Networks of the Future

Thank You

pruth@renci.org

Please join the Chameleon-Federation project before tomorrow's Chameleon tutorial

http://bit.ly/GENI2Chameleon