



[www.chameleoncloud.org](http://www.chameleoncloud.org)

## CHAMELEON PHASE 2: TOWARDS A SCIENTIFIC INSTRUMENT FOR COMPUTER SCIENCE RESEARCH

**Kate Keahey**

Mathematics and CS Division, Argonne National Laboratory

Computation Institute, University of Chicago

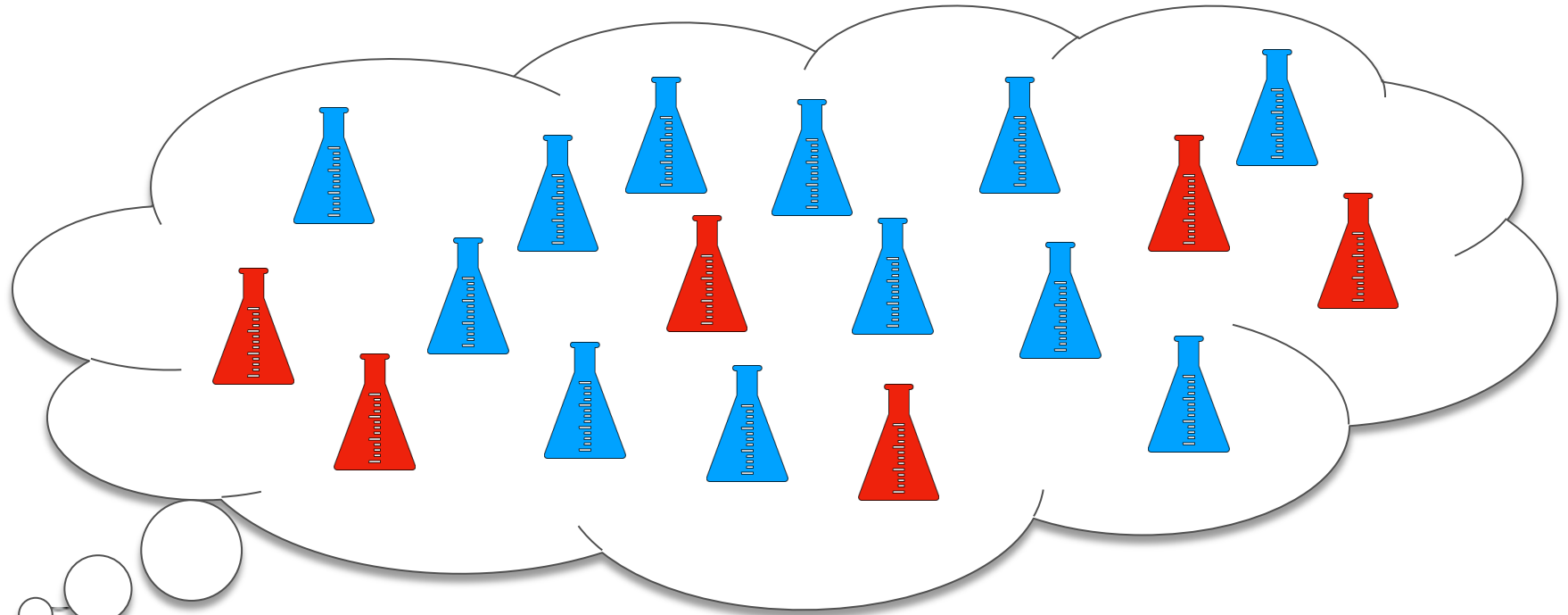
*keahey@anl.gov*

SEPTEMBER 18, 2017

I



# WHY DO WE NEED AN INSTRUMENT?

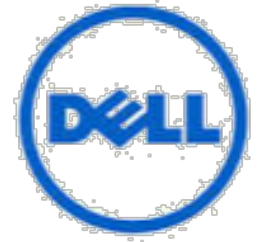


*In practice we can carry out only those experiments that are supported by an instrument that allows us to deploy, capture, and record relevant scientific phenomena*

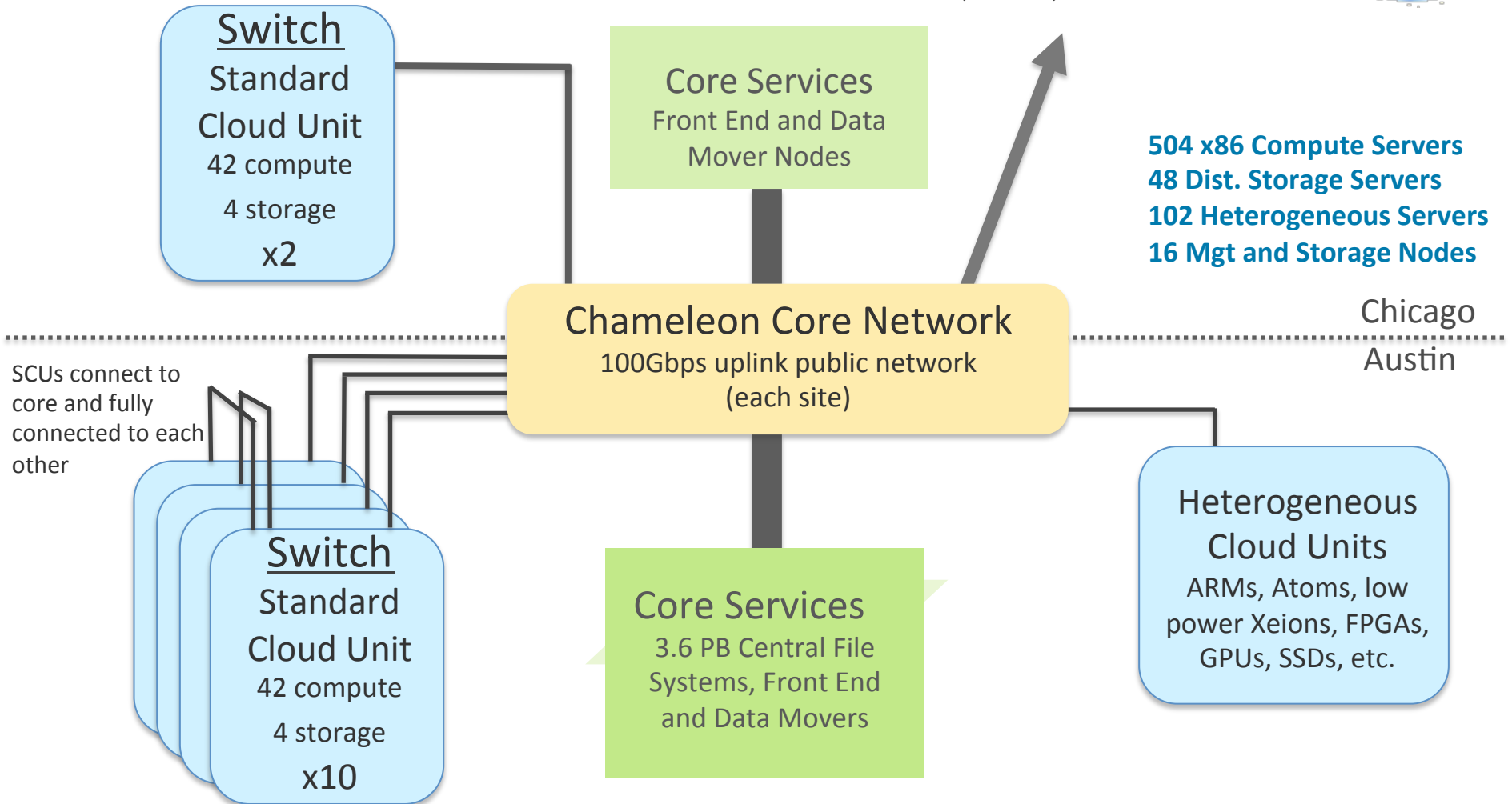
# CHAMELEON PHASE 1 IN A NUTSHELL

- ▶ **Deeply reconfigurable:** “As close as possible to having it in your lab”
  - ▶ Deep reconfigurability (bare metal) and isolation
  - ▶ Power on/off, reboot from custom kernel, serial console access, etc.
  - ▶ But also – modest KVM cloud for ease of use
- ▶ **Large-scale:** “Big Data, Big Compute research”
  - ▶ ~650 nodes (~15,000 cores), 5 PB of storage distributed over 2 sites connected with 100G network...
  - ▶ ...and diverse: ARMs, Atoms, FPGAs, GPUs, etc.
- ▶ Blueprint for a **sustainable** production testbed: “cost-effective to deploy, operate, and enhance”
  - ▶ Powered by OpenStack with bare metal reconfiguration (Ironic)
- ▶ **Open** production testbed for **Computer Science Research**
  - ▶ Project started in 10/2014, testbed available since 07/2015
  - ▶ Currently 1,700+ users, 300+ projects

# CHAMELEON PHASE 1 HARDWARE



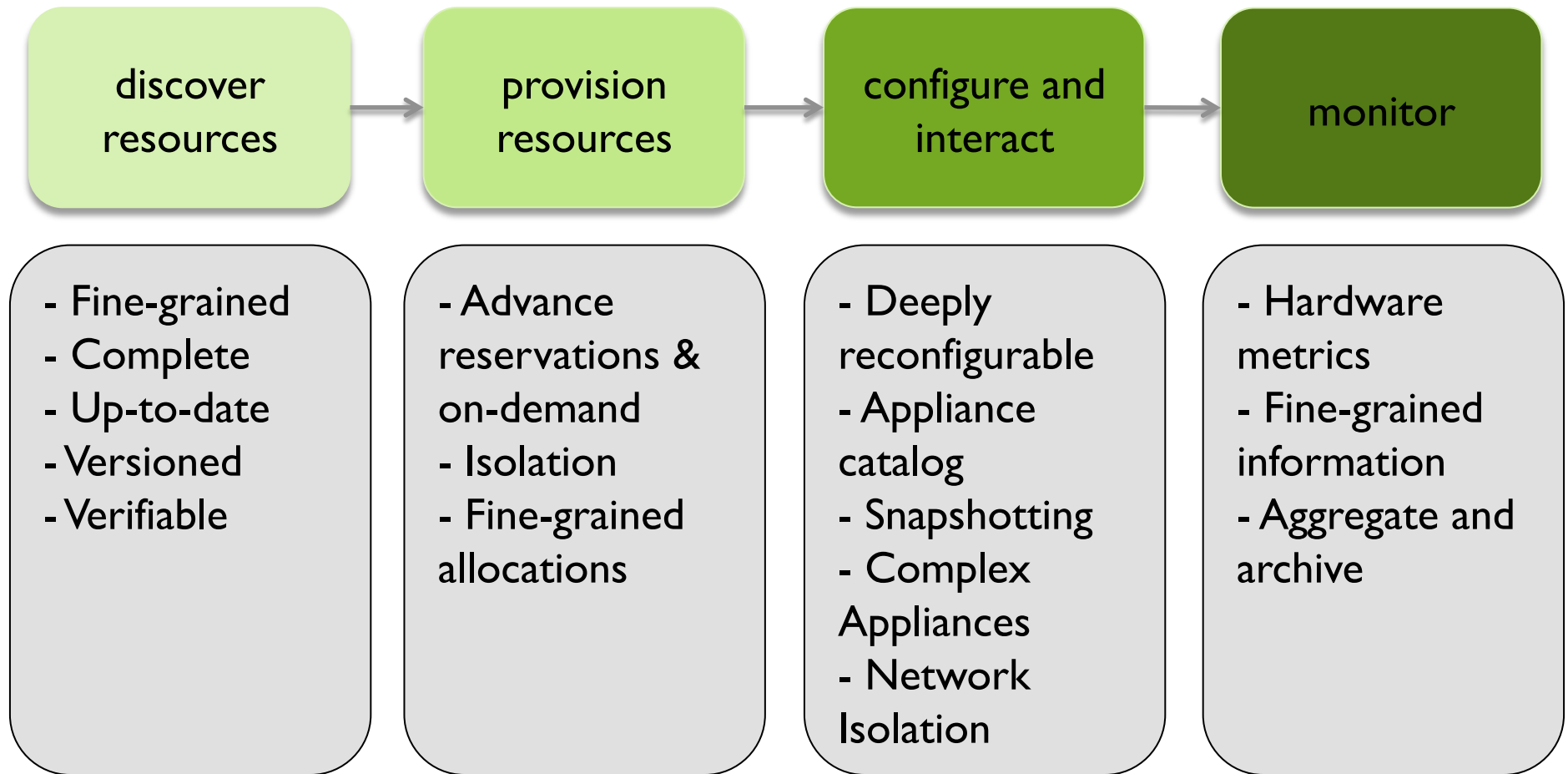
To UTSA, GENI, Future Partners



# CHAMELEON PHASE 1 HARDWARE (DETAIL)

- ▶ “Start with large-scale homogenous partition”
  - ▶ 12 Standard Cloud Units (48 node racks)
  - ▶ Each rack has 42 Dell R630 compute servers, each with dual-socket Intel Haswell processors (24 cores) and 128GB of RAM
  - ▶ Each rack also has 4 Dell FX2 storage server (also Intel Haswells), each with a connected JBOD of 16 2TB drives (total of 128 TB per SCU)
  - ▶ Allocations can be an entire rack, multiple racks, nodes within a single rack or across racks (e.g., storage servers across racks forming a Hadoop cluster)
  - ▶ 48 port Force10 s6000 OpenFlow-enabled switches 10Gb to hosts, 40Gb uplinks to Chameleon core network
- ▶ Shared infrastructure
  - ▶ 3.6 PB global storage, 100Gb Internet connection between sites
- ▶ “Graft on heterogeneous features”
  - ▶ Infiniband with SR-IOV support netw in one rack
  - ▶ High-memory, NVMe, SSDs, GPUs (18 nodes), FPGAs (4 nodes)
  - ▶ ARM microservers (24) and Atom microservers (8), low-power Xeons (8)

# CORE SUPPORT FOR EXPERIMENTAL WORKFLOW



Powered by CHI = 65%\*OpenStack + 10%\*G5K + 25%\*"special sauce"

# LESSONS LEARNED

- ▶ Dates: project start in 10/2014, public release 07/2015
- ▶ In numbers:
  - ▶ 300+ projects, 1,700+ users, 100+ institutions, 30+ states, 100+ reported publications – and counting
- ▶ Strengths
  - ▶ Large-scale homogenous partition (420 nodes/10,080 cores)
  - ▶ Hardware diversity
  - ▶ Support for deep reconfigurability built on top of a commodity technology
- ▶ Opportunities
  - ▶ Platform for networking research
  - ▶ Experiment management tools
- ▶ Welcoming RENCI as a new partner!

## TOWARDS PHASE 2

- ▶ Broaden the set of experiments we can deploy
  - ▶ New hardware, new networking capabilities, a range of usability features, and anything you ask for!
- ▶ Make Computer Science experiments sustainable
  - ▶ Package our testbed and operations model
- ▶ Transform it into a scientific instrument
  - ▶ Add better ways of observing, recording and repeating experiments
- ▶ Working with you!



# NEW HARDWARE IN PHASE 2

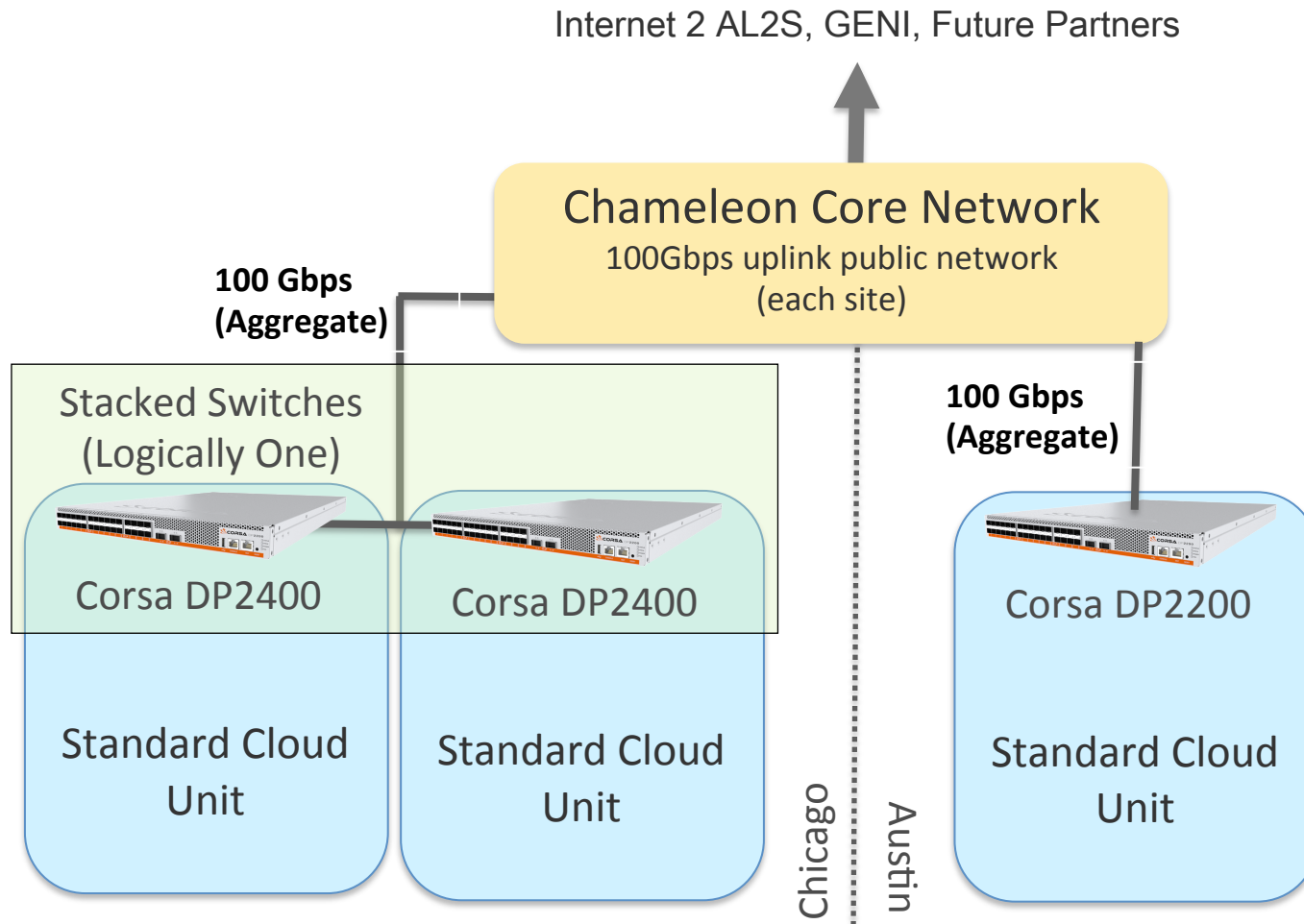
- ▶ 4 new Standard Cloud Units (32 node racks in 2U chassis)
  - ▶ 3x Intel Xeon “Sky Lake” racks (2x @UC, 1x @TACC) in Y1
  - ▶ 1x future Intel Xeon rack (@TACC) in Y2
- ▶ Corsa DP2000 series switches in Y1
  - ▶ 2x DP2400 with 100Gbps uplinks (@UC)
  - ▶ 1x DP2200 with 100Gbps uplink (@TACC)
  - ▶ Each switch will have a 10 Gbps connections 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 NVMe
- ▶ 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 v1.5
  - ▶ User defined controllers
- ▶ **Performance**
  - ▶ 10 Gbps within a site
  - ▶ 100 Gbps between UC/TACC (Aggregated)



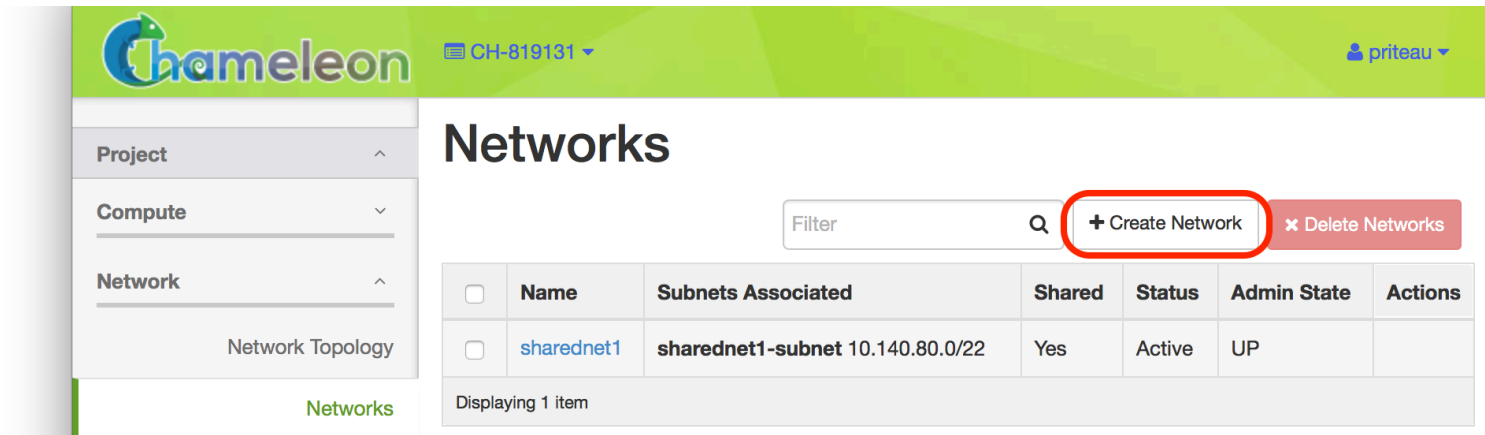
# NETWORK HARDWARE



# NETWORKING RESEARCH

- ▶ Network programmability
  - ▶ How should individual flows/paths through the network (local/wide-area) be routed? What information is needed to control the flow/path of packets? What happens when the network can filter and route using headers other than IP? How do we secure the control of flows/paths?
- ▶ Wide-area isolated networks
  - ▶ How can we build federated superfacilities composed of resources from mid-scale compute infrastructure, centralized data repositories, and institutional resources?
- ▶ High bandwidth
  - ▶ How can we use 100 Gbps wide-area networks to support big-data applications?

# SUPPORT FOR ISOLATED NETWORKS



The screenshot displays the Chameleon Networks management interface. The sidebar on the left includes navigation options for Project, Compute, and Network. The main content area is titled 'Networks' and features a search filter, a '+ Create Network' button (highlighted with a red circle), and a 'Delete Networks' button. Below these is a table with one row of network data.

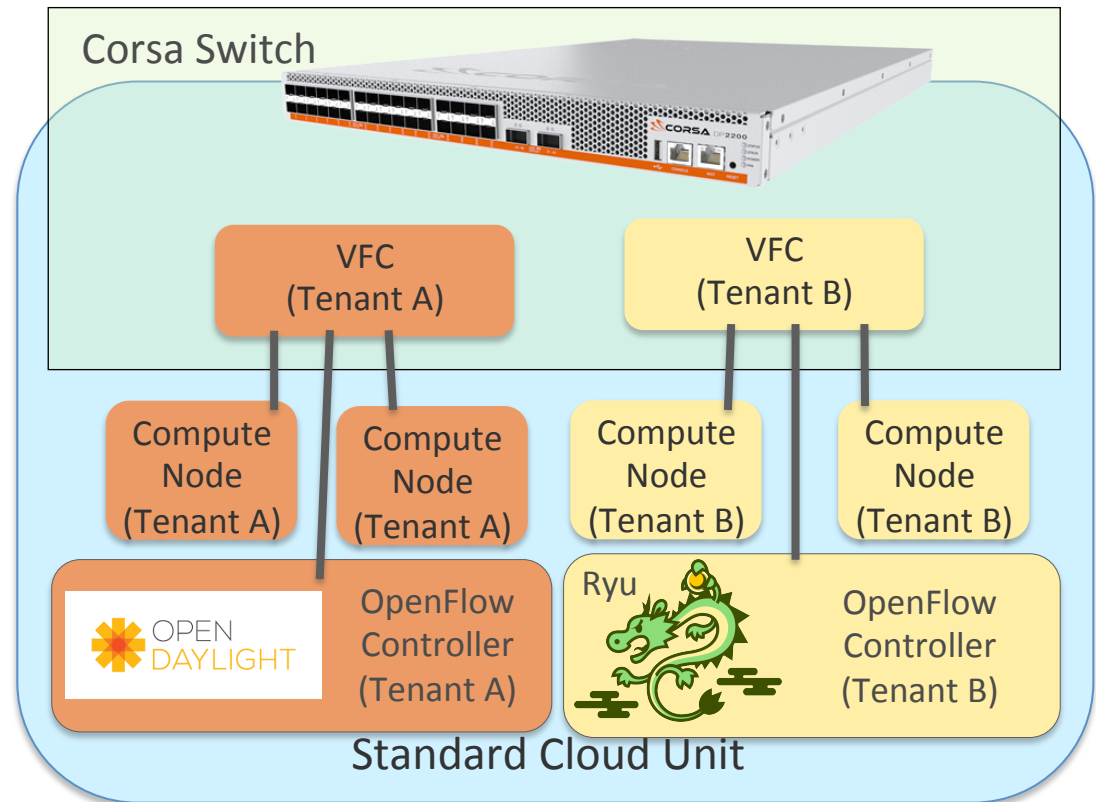
| <input type="checkbox"/> | Name       | Subnets Associated               | Shared | Status | Admin State | Actions |
|--------------------------|------------|----------------------------------|--------|--------|-------------|---------|
| <input type="checkbox"/> | sharednet1 | sharednet1-subnet 10.140.80.0/22 | Yes    | Active | UP          |         |

Displaying 1 item

- ▶ Multi-tenant networking allows users to provision isolated L2 VLANs and manage their own IP address space
- ▶ Currently only available on compute nodes on CHI@UC
- ▶ Migrating to a more extensible implementation
- ▶ Will replace default shared network when all hardware is supported

# ISOLATED VIRTUAL SDN SWITCH

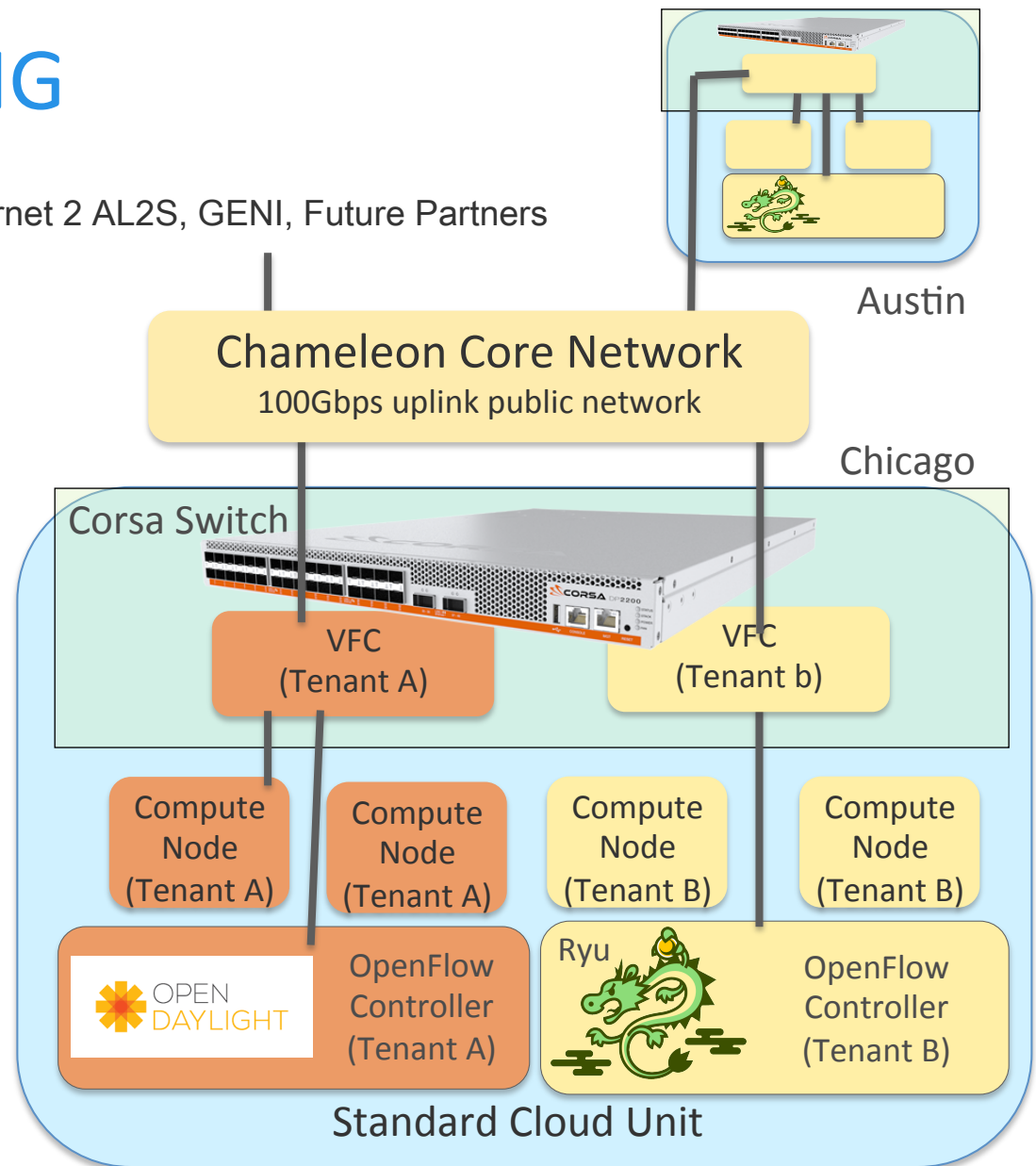
- ▶ Provide Isolated Networks (end of 2017)
- ▶ BYOC– Bring your own controller: isolated user controlled virtual OpenFlow switches (~Summer 2018)



# EXTERNAL STITCHING

- ▶ Stitch dynamic VLANs from Chameleon to external partners (ExoGENI, ScienceDMZs) – end of 2017
- ▶ Support for large flows: Neutron Bypass
- ▶ Support stitching over VFCs (Summer 2018)

Internet 2 AL2S, GENI, Future Partners



# OTHER FEATURES

- ▶ Improved ease of use
  - ▶ Easier console access (Y1), better support for non-x86 architectures (Y1), storage volumes (Y2), multi-region configuration (Y1), single sign-on (Y1), combined reservation and appliance deployment, dynamically add/release nodes from a reservation, model-based descriptions – and many others!
- ▶ Other features
  - ▶ BIOS management, improved monitoring (IPMI data)
- ▶ Your requests here!



# MAKING COMPUTER SCIENCE EXPERIMENTS CHEAP

- ▶ Cost factors
  - ▶ Streamline operations
  - ▶ Package the testbed (Chameleon-in-a box)
- ▶ Lowering operational cost
  - ▶ Understand and resolve typical failures
  - ▶ Preventative management
  - ▶ Processes and documentation

# CHAMELEON IN A BOX

- ▶ Testbed extension: join the Chameleon testbed
  - ▶ Generalize and package
  - ▶ Define operations models
  - ▶ First package expected in Y1
- ▶ Part-time extension
  - ▶ Define and implement contribution models
- ▶ New testbed
  - ▶ Generalize policies



# TOWARDS A SCIENTIFIC INSTRUMENT

Scientific instrument == a device that allows us to **deploy, capture, and record** relevant phenomena



How can we better capture (i.e., observe and measure) or record?

How can we improve our ability to share and repeat experiments?

# DEVELOPING BETTER INSIGHT

- ▶ Everything in a testbed is a recorded event
  - ▶ The resources you used
  - ▶ The appliance/image you deployed
  - ▶ The monitoring information your experiment generated
  - ▶ Plus any information you choose to share with us: e.g., experiment start and stop
- ▶ Experiment summary: information about your experiment made available in a “consumable” form
  - ▶ It could be integrated with many existing tools (Jupyter, Grafana, etc.)...
  - ▶ ... or creatively integrated with existing technologies

# FROM INSIGHT TO REPEATABILITY

- ▶ Existing repeatability elements
  - ▶ Testbed versioning (53 versions and counting)
  - ▶ Appliance publication, versioning, and management
  - ▶ Monitoring and logging data
- ▶ Experiment précis: closing the gap between resource versions, appliances, and data
- ▶ The reproducibility trade-off
  - ▶ Representing work with complex phenomena requires a huge amount of information
  - ▶ Reproducing those complex phenomena is costly
- ▶ From experiment précis to experiment replays
- ▶ Publishing experiment précis

# WORKING WITH THE COMMUNITY

- ▶ Inspiration: research highlights
- ▶ Communication:
  - ▶ Mailing lists, social media, tickets, surveys, etc.
- ▶ Training: training videos, webinars, tutorials
- ▶ Education: creating an educational community
- ▶ Chameleon User Meeting
- ▶ **Personalized Outreach**

# SUMMARY

- ▶ Making a better testbed
  - ▶ Significant new capabilities that will allow you to deploy more experiments
- ▶ Making a good testbed cheaper
  - ▶ Streamlining operations and packaging Chameleon
- ▶ Towards a scientific instrument: exploring a new dimension
  - ▶ Shifting attention to insight, methodology, and repeatability
- ▶ Working with you!

*“We shape our buildings;  
thereafter they shape us”*

*Winston Churchill*







[www.chameleoncloud.org](http://www.chameleoncloud.org)

*Shape the Science:*

[www.chameleoncloud.org](http://www.chameleoncloud.org)

keahey@anl.gov

SEPTEMBER 18, 2017 25

