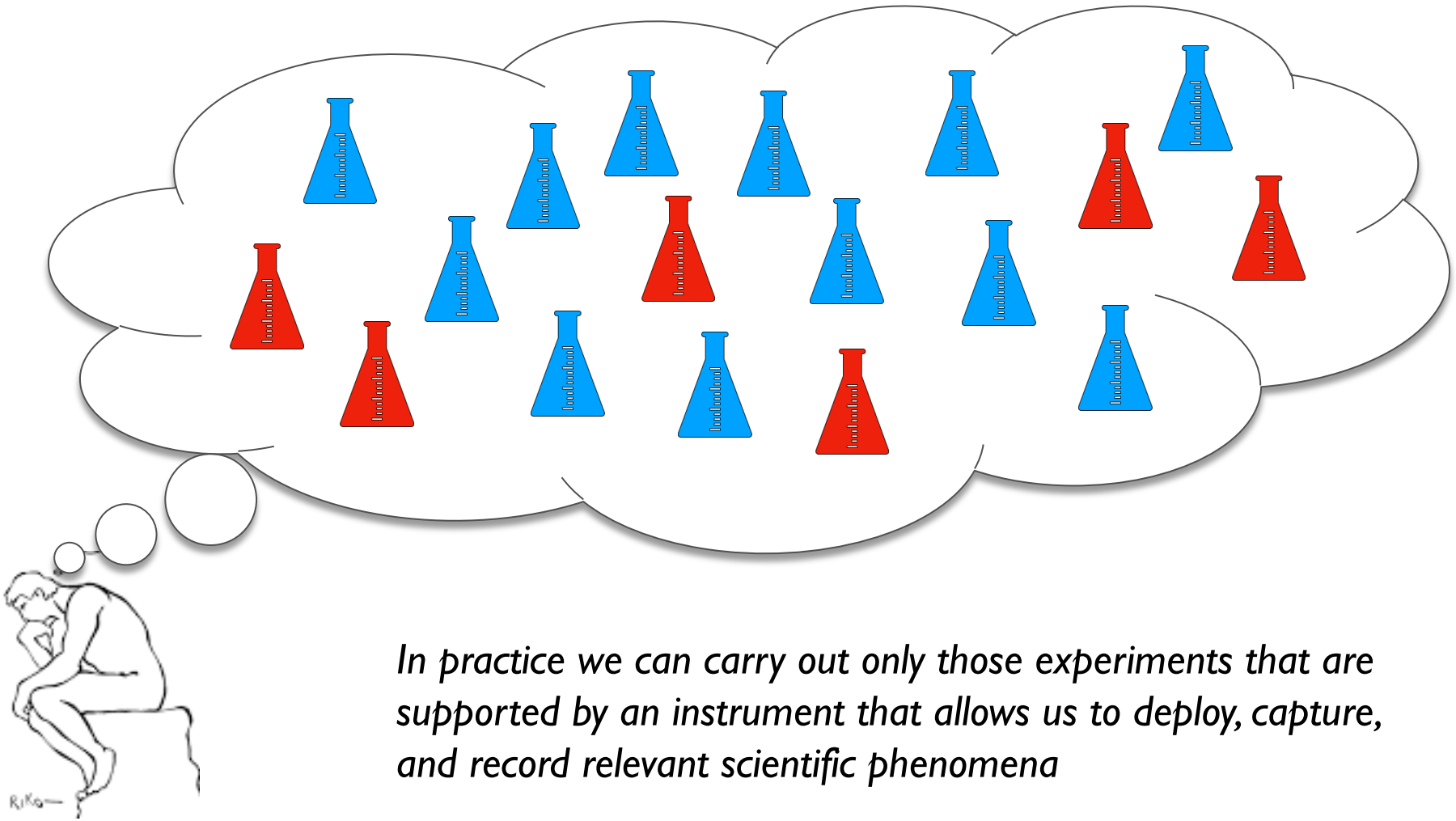
www. chameleoncloud.org

CHAMELEON: TOWARDS AN EXPERIMENTAL INSTRUMENT FOR COMPUTER SCIENCE RESEARCH

Kate Keahey
Mathematics and CS Division, Argonne National Laboratory
Computation Institute, University of Chicago
keahey@anl.gov

THE UNIVERSITY OF CHICAGO TACC NORTHWESTERN UNIVERSITY renci NSF

WHY DO WE NEED AN INSTRUMENT?

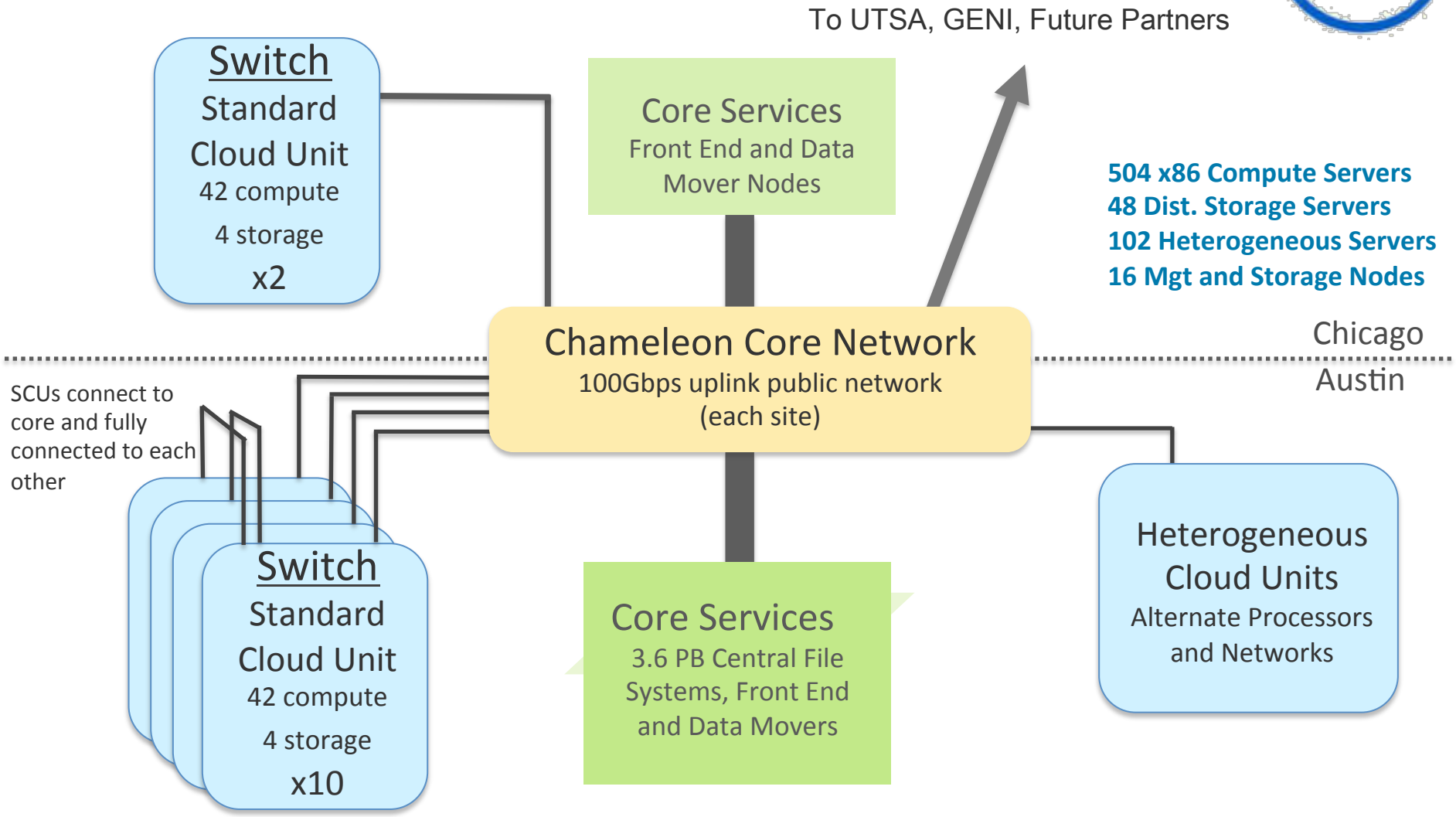


In practice we can carry out only those experiments that are supported by an instrument that allows us to deploy, capture, and record relevant scientific phenomena

CHAMELEON 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"
  - **Large-scale**: ~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
  - Just renewed for phase 2

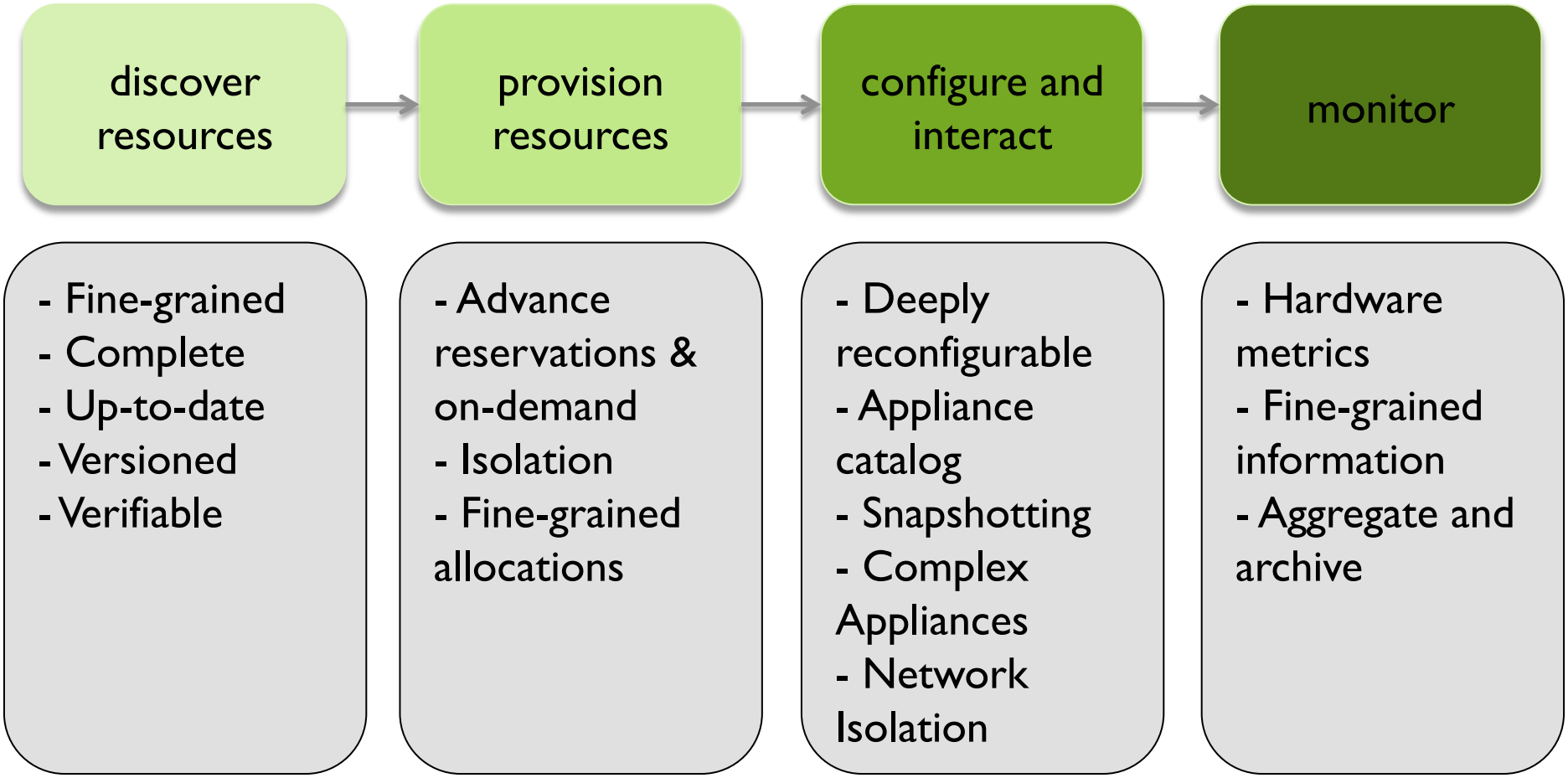
EXISTING CHAMELEON HARDWARE



EXISTING CHAMELEON 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 (22 nodes), FPGAs (4 nodes)
  - ARM microservers (24) and Atom microservers (8), low-power Xeons (8)

CORE SUPPORT FOR EXPERIMENTAL WORKFLOW



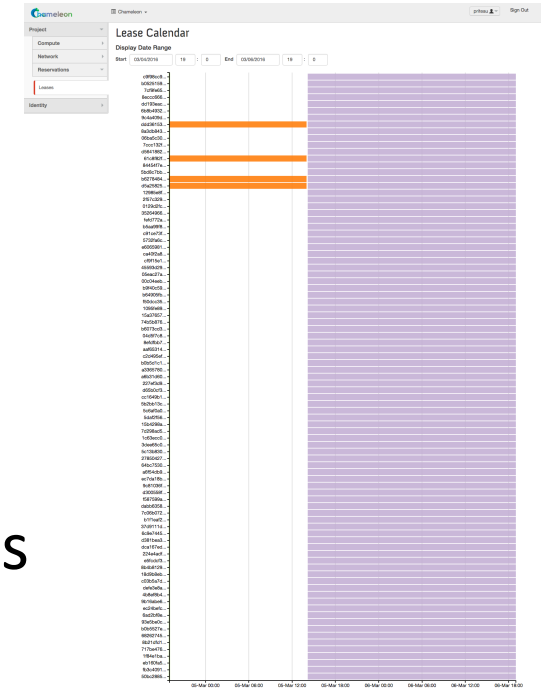
Powered by CHI = 65%*OpenStack + 10%*G5K + 25%*"special sauce"

CHI: DISCOVERING AND VERIFYING RESOURCES

- Fine-grained, up-to-date, and complete representation
- Testbed versioning
  - "What was the drive on the nodes I used 6 months ago?"
- Dynamically verifiable
  - Does reality correspond to description? (e.g., failure handling)

- Grid'5000 registry toolkit + Chameleon portal
  - Automated resource discovery (lshw, hwloc, ethtool, etc.)
  - Scripted export to RM/Blazar
- G5K-checks
  - Can be run after boot, acquires information and compares it with resource catalog description

CHI: PROVISIONING RESOURCES



- Resource leases
- Advance reservations (AR) and on-demand
  - AR facilitates allocating at large scale
- Isolation between experiments
- Fine-grain allocation of a range of resources
  - Different node types, etc.

- Based on OpenStack Nova/Blazar
- Revived Blazar project (ex. Climate), part of core reviewer team
- Extended Horizon panel with calendar displays
- Added Chameleon usage policy enforcement

CHI: CONFIGURE AND INTERACT

- Deep reconfigurability: custom kernels, console access, etc.
- Snapshotting for saving your work
- Map multiple appliances to a lease
- Appliance Catalog and appliance management
- Handle complex appliances
  - Virtual clusters, cloud installations, etc.
- Support for network isolation

- OpenStack Ironic, Neutron, Glance, meta-data servers, and Heat
- Added snapshotting, appliance management and catalog, dynamic VLANs
- Not yet BIOS reconfiguration

CHI: INSTRUMENTATION AND MONITORING

- Enables users to understand what happens during the experiment
- Instrumentation metrics
- Types of monitoring:
  - Infrastructure monitoring (e.g., PDUs)
  - User resource monitoring
  - Custom user metrics
- Aggregation and Archival

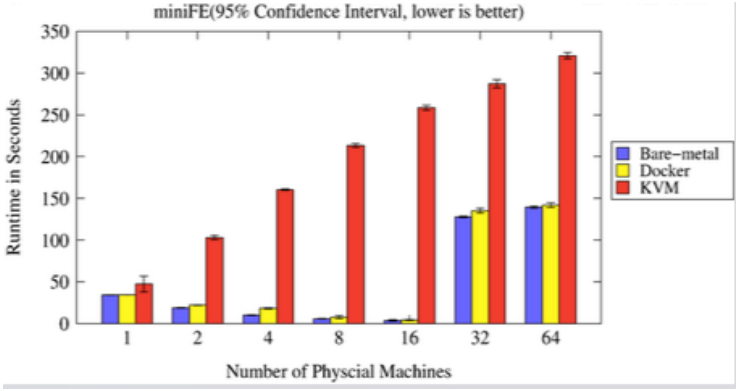
- OpenStack Ceilometer + agents, standard metrics (CPU, memory, network, disk usage, etc.)
- RAPL interface to provide power and energy usage

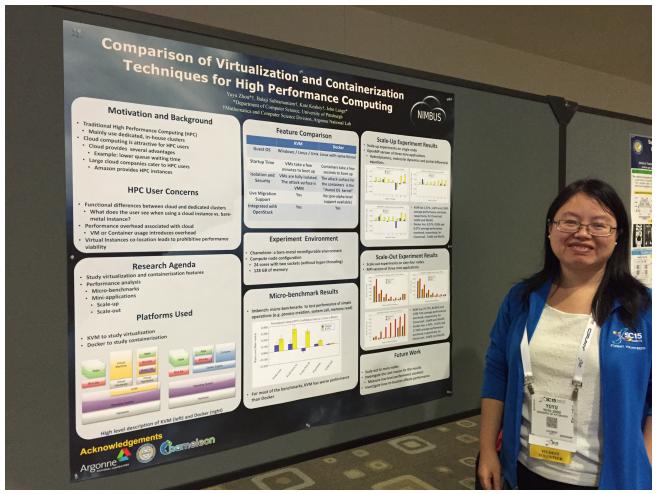
CHAMELEON: TIMELINE AND STATUS

- **10/14: Project starts**
- 04/15: Chameleon Core Technology Preview
- 06/15: Chameleon Early User on new hardware
- **07/15: Chameleon public availability**
- 2016&2017: New capabilities and new hardware releases
- **Today: 1,900+ users/300+ projects**

VIRTUALIZATION OR CONTAINERIZATION?

- Yuyu Zhou, University of Pittsburgh
- Research: lightweight virtualization
- Testbed requirements:
  - Bare metal reconfiguration
  - Boot from custom kernel
  - Console access
  - Up-to-date hardware
  - Large scale experiments

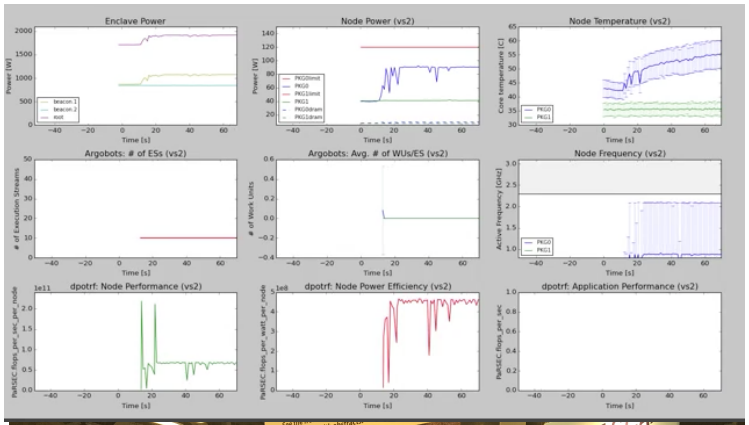




SC15 Poster: "Comparison of Virtualization and Containerization Techniques for HPC"

EXASCALE OPERATING SYSTEMS

- Swann Perarnau, ANL
- Research: exascale operating systems
- Testbed requirements:
  - Bare metal reconfiguration
  - Boot kernel with varying kernel parameters
  - Fast reconfiguration, many different images, kernels, params
  - Hardware: performance counters, many cores



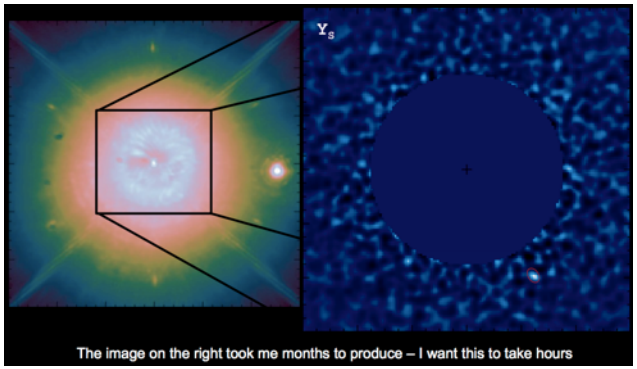
HPPAC'16 paper: "Systemwide Power Management with Argo"

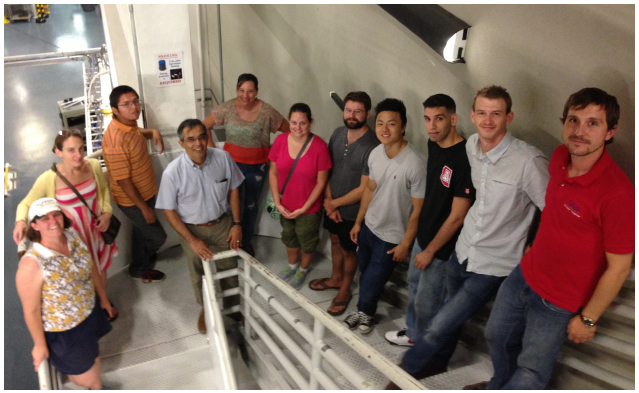
CLASSIFYING CYBERSECURITY ATTACKS

- Jessie Walker & team, University of Arkansas at Pine Bluff (UAPB)
- Research: modeling and visualizing multi-stage intrusion attacks (MAS)
- Testbed requirements:
  - Easy to use OpenStack installation
  - Access to the same infrastructure for multiple collaborators

TEACHING CLOUD COMPUTING

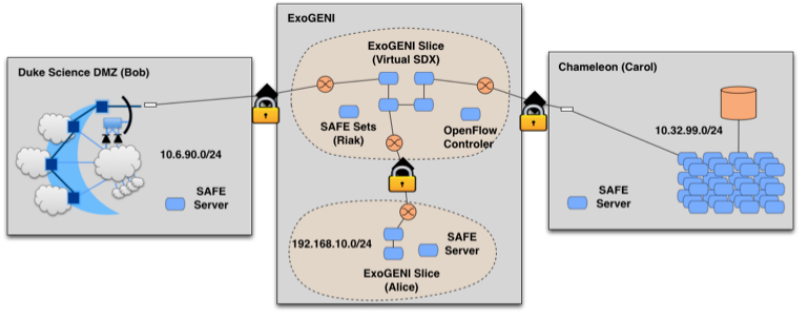
- Nirav Merchant and Eric Lyons, University of Arizona
- ACIC2015: project-based learning course
  - Data mining to find exoplanets
  - Scaled analysis pipeline by Jared Males
  - Develop a VM/workflow management appliance and best practice that can be shared with broader community
- Testbed requirements:
  - Easy to use IaaS/KVM installation
  - Minimal startup time
  - Support distributed workers
  - Block store: make copies of many 100GB datasets



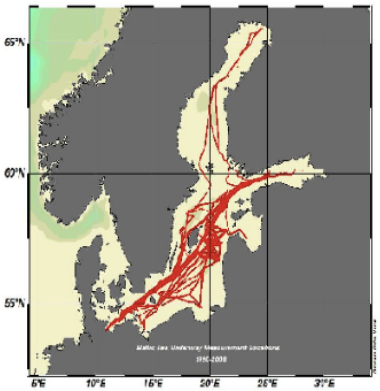


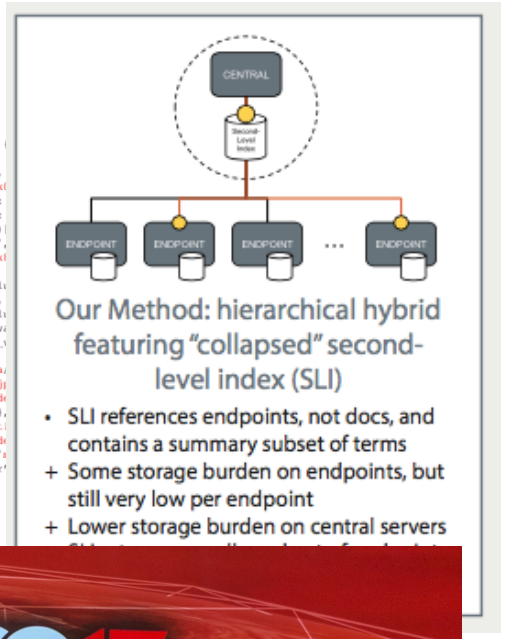
CREATING DYNAMIC SUPERFACILITIES

- NSF CICI SAFE, Paul Ruth, RENCI-UNC Chapel Hill
- Creating trusted facilities
  - Automating trusted facility creation
  - Virtual Software Defined Exchange (SDX)
  - Secure Authorization for Federated Environments (SAFE)
- Testbed requirements
  - Creation of dynamic VLANs
  - Support for slices and network stitching



DATA SCIENCE RESEARCH





- ACM Student Research Competition semi-finalists:
  - Blue Keleher, University of Maryland
  - Emily Herron, Mercer University
- Searching and image extraction in research repositories
- Testbed requirements:
  - Access to distributed storage in various configurations
  - State of the art GPUs
  - Easy to use appliances

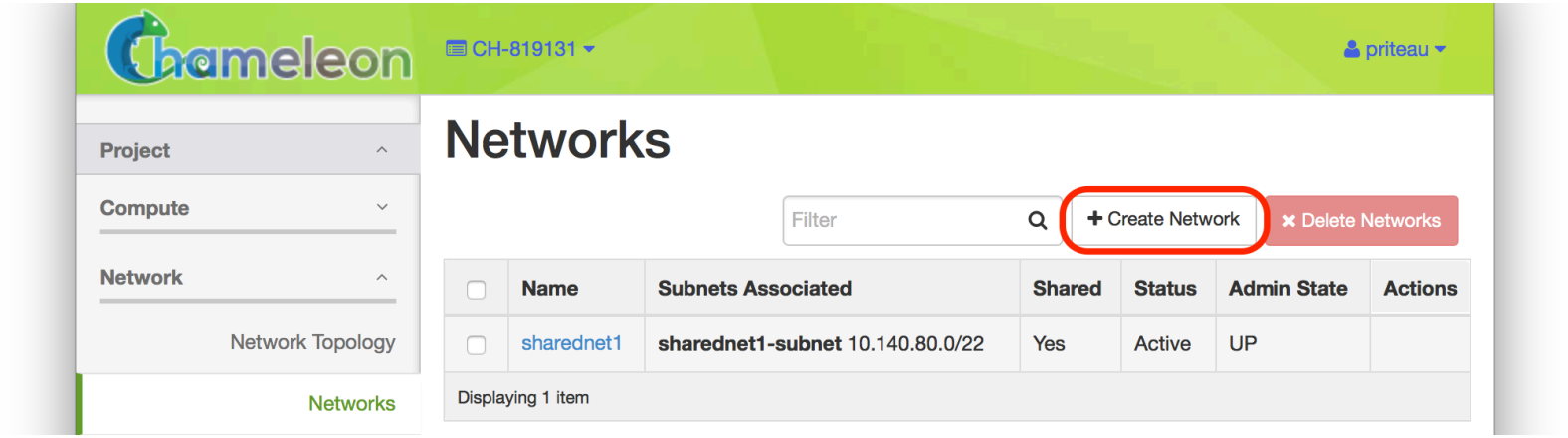
TOWARDS A SCIENTIFIC INSTRUMENT

- **Deploy:** Broaden the set of supported experiments
- **Capture:** observe, monitor, and measure
- **Record:** a comprehensive "active record"
  - Re-examine, share/publish, review, re-play

DEPLOY: NEW HARDWARE

- 4 new Standard Cloud Units (32 node racks in 2U chassis)
  - 3x Intel Xeon "Sky Lake" racks (2x @UC, 1x @TACC) -- almost there!
  - 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
  - Alternative 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

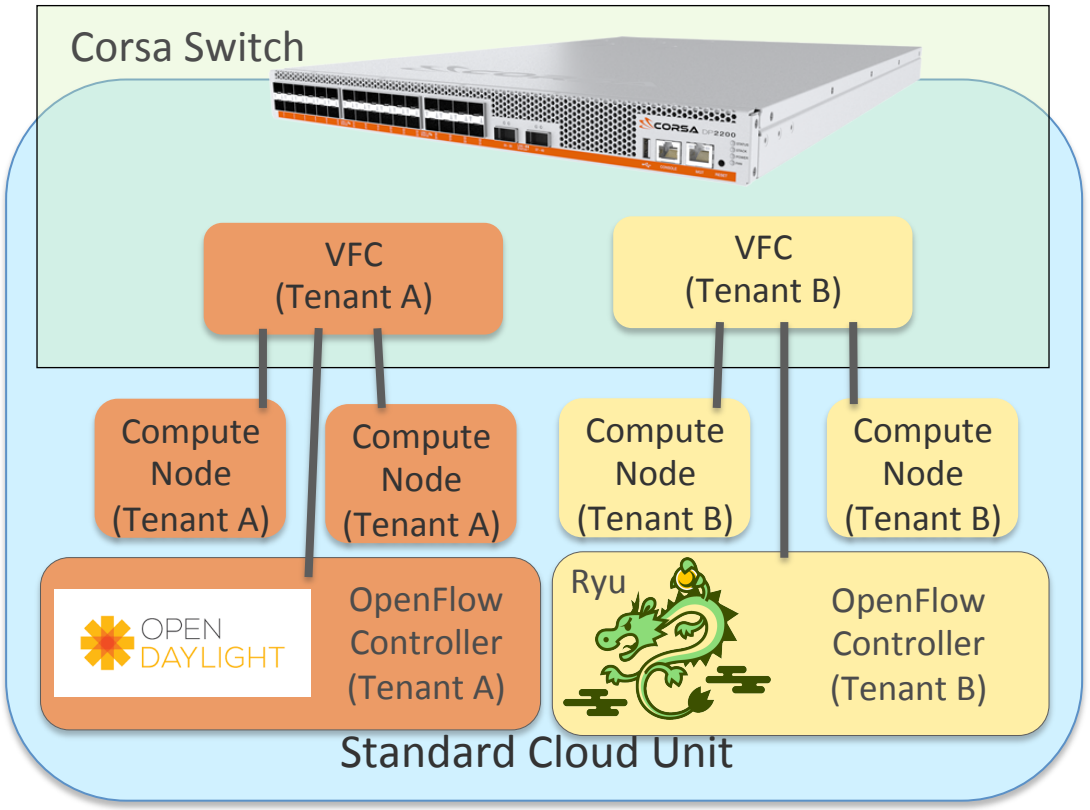
DEPLOY: NETWORKING BUILDING BLOCKS



- Research topics: exploring network programmability, building superfacilities, utilizing high bandwidth
- Building blocks:
  - **Multi-tenant networking** allows users to provision isolated L2 VLANs and manage their own IP address space – available now!
  - **Stitching** dynamic VLANs from Chameleon to external partners (ExoGENI, ScienceDMZs) – currently in Early User preview, available by end of 2017

DEPLOY: ADVANCED NETWORKING

- BYOC– Bring your own controller: isolated user controlled virtual OpenFlow switches (~Summer 2018)
- Support for large flows: Neutron Bypass
- Support stitching over VFCs (Summer 2018)



CAPTURE: THE FOUNDATION

- Testbed versioning
  - Fine-grain representation
  - Automated discovery and updates
  - >50 versions since public availability – and counting
  - Still working on: better firmware version management
- Appliance management
  - Configuration, versioning, publication
  - Still working on: repository vs catalog connection
- Monitoring and logging
  - Making it accessible in easier ways
- However… the user still has to keep track of this information

CAPTURE: KEEPING TRACK OF EXPERIMENTS

- 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., "start power_exp_23" and "stop power_exp_23"

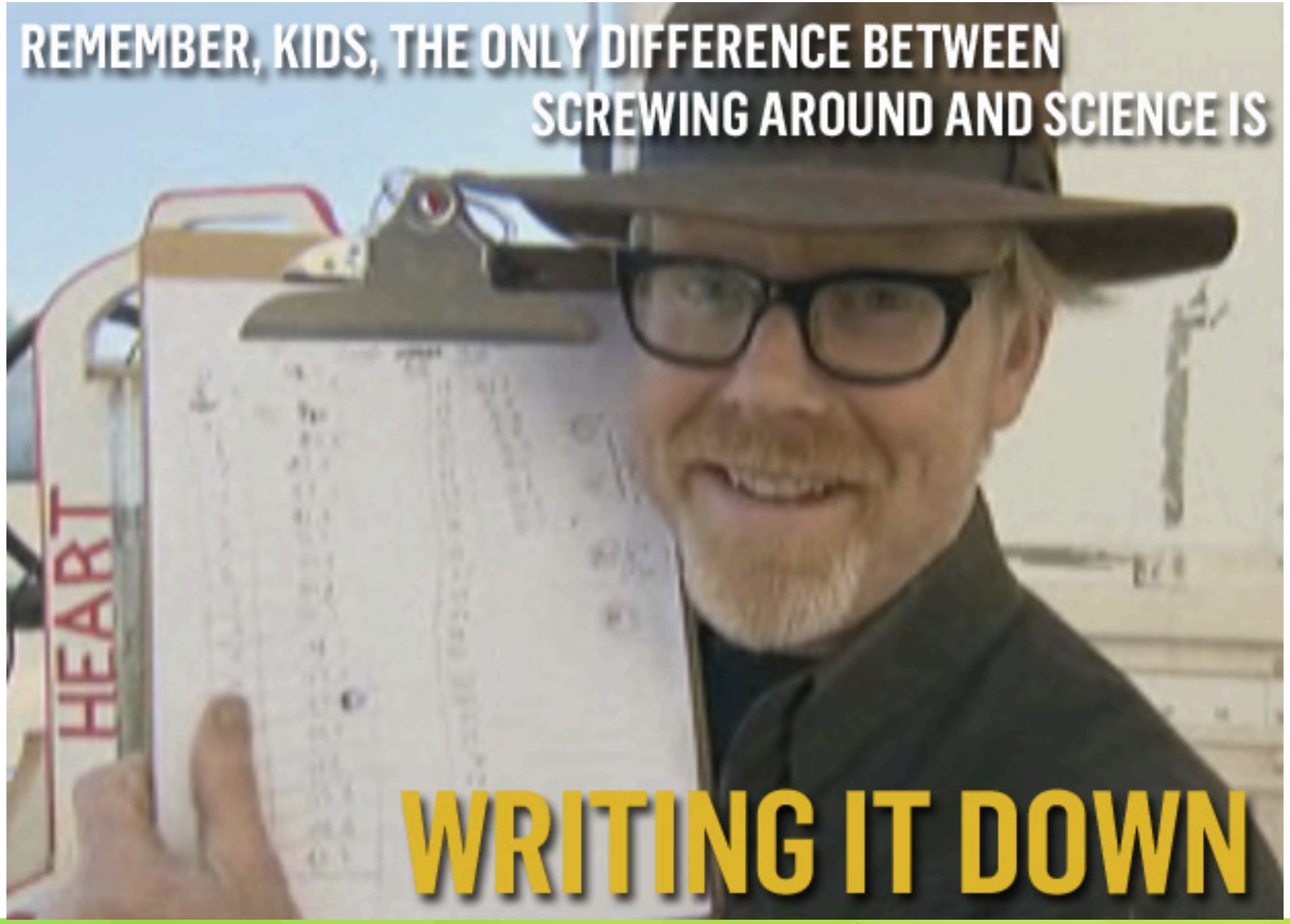
- **Experiment précis:** information about your experiment made available in a "consumable" form
- (Bonus: it can be integrated with many existing tools, e.g., Jupyter or Grafana)…

VISUALIZING DATA FROM EXPERIMENTS

Visualizing data from experiments –
video from Chameleon YouTube channel at:
https://www.youtube.com/watch?v=9EcdF9csFb4&feature=youtu.be

RECORD: MOVING TOWARDS REPEATABILITY

- Experiment précis
  - Recording the experiment for you: closing the gap between resource versions, appliances, and events
  - "Active record" that can be given to a reviewer or shared with others
- Publishing experiment précis
- Integration with popular tools
- From experiment précis to experiment replays
  - Model-based experiment capture
  - Re-play tools


REMEMBER, KIDS, THE ONLY DIFFERENCE BETWEEN
SCREWING AROUND AND SCIENCE IS
HEART
WRITING IT DOWN

PARTING THOUGHTS

- A testbed for Computer Science research
  - **Open** production testbed for **Computer Science research**: 1,900+ users/300+ projects
  - Designed from the ground up for a **large-scale** testbed supporting **deep reconfigurability**
  - Blueprint for a **sustainable production testbed**: powered by OpenStack
- Towards an instrument: capture, record, replay
  - Making repeatability/reproducibility cost-effective
  - Integrating with popular tools
  - Helping you leverage incentives (ACM badges, conference awards, etc.)

Help us all dream big:

www.chameleoncloud.org

keahey@anl.gov