www. chameleoncloud.org

A LARGE SCALE RECONFIGURABLE INSTRUMENT FOR COMPUTER SCIENCE EXPERIMENTATION

Kate Keahey
University of Chicago
Argonne National Laboratory
{keahey, priteau}@uchicago.edu

JUNE 19, 2017

THE UNIVERSITY OF CHICAGO TACC NORTHWESTERN UNIVERSITY THE OHIO STATE UNIVERSITY UTSA NSF

SEARCHING FOR A TESTBED

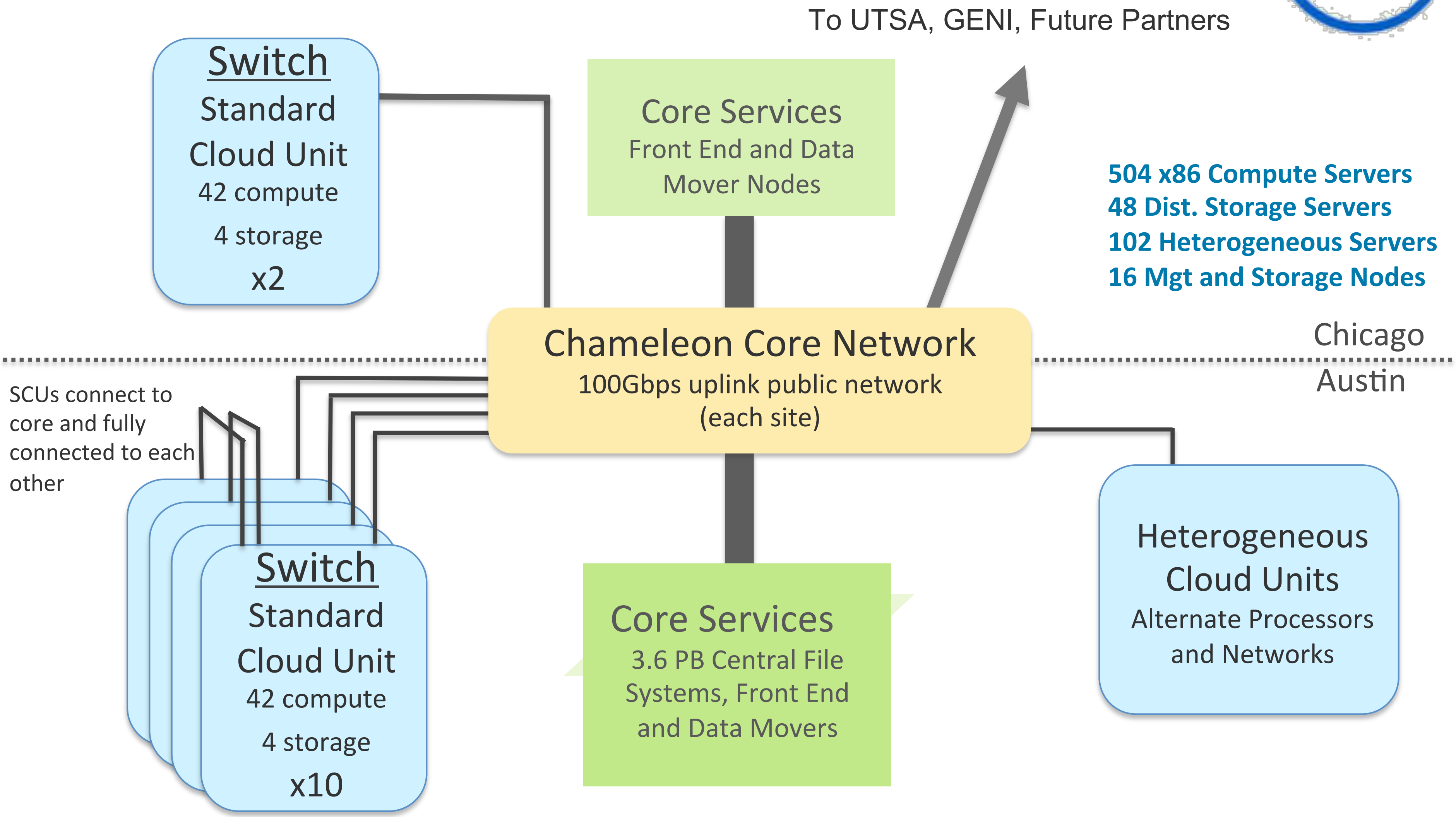
- A personal quest
  - The case of no testbed at all
  - The case of inadequate: “no hardware virtualization”
  - The case of too small: “we think this will scale”
  - The case of shared: “it may have impacted our result”

While the types of experiments we can design are only limited by our creativity, in practice we can carry out only those that are supported by an instrument that allows us to deploy, capture, and measure relevant phenomena.

DESIGN STRATEGY FOR A SCIENTIFIC INSTRUMENT

- **Large-scale:** “Big Data, Big Compute research”
  - ~650 nodes (~14,500 cores), 5 PB of storage distributed over 2 sites connected with 100G network
  - Operated as a single instrument
- **Reconfigurable:** “As close as possible to having it in your lab”
  - Deep reconfigurability (bare metal) and isolation
  - Fundamental to support Computer Science experiments
- Connected: “One stop shopping for experimental needs”
  - Workload and Trace Archive
  - Appliance Catalog
  - Instrumentation and repeatability tools
- Sustainable: “cost-effective to deploy, operate, and enhance”
- Open: “US researchers and collaborators”

CHAMELEON HARDWARE



To UTSA, GENI, Future Partners

Switch
Standard Cloud Unit
42 compute
4 storage
x2

Core Services
Front End and Data Mover Nodes

504 x86 Compute Servers
48 Dist. Storage Servers
102 Heterogeneous Servers
16 Mgt and Storage Nodes

Chameleon Core Network
100Gbps uplink public network (each site)

Chicago

Austin

SCUs connect to core and fully connected to each other

Switch
Standard Cloud Unit
42 compute
4 storage
x10

Core Services
3.6 PB Central File Systems, Front End and Data Movers

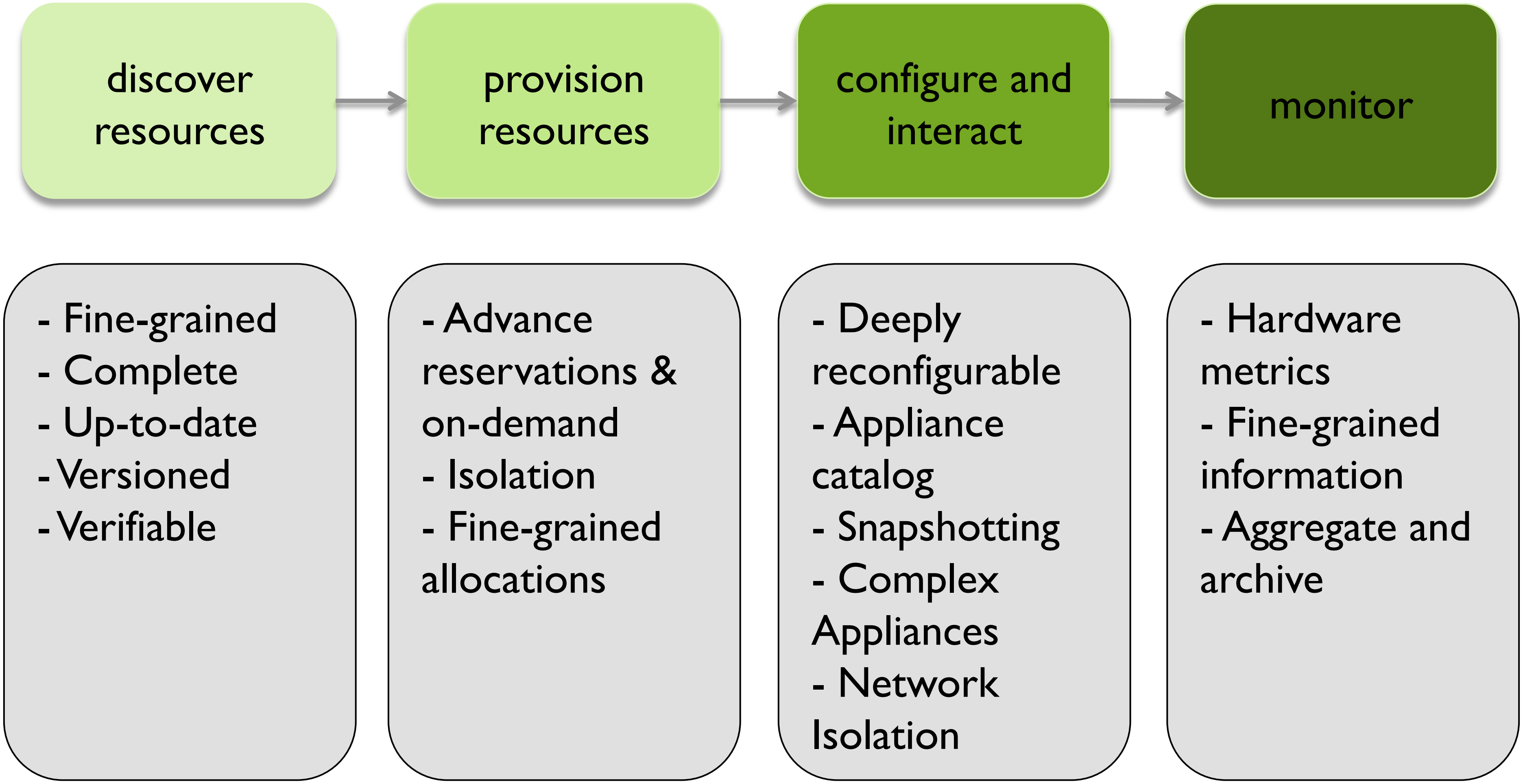
Heterogeneous Cloud Units
Alternate Processors and Networks

Chameleon www.chameleoncloud.org

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 network in one rack with SR-IOV support
  - High-memory, NVMe, SSDs, GPUs, FPGAs
  - ARM microservers (24) and Atom microservers (8), low-power Xeons (8)

EXPERIMENTAL WORKFLOW REQUIREMENTS



discover resources → provision resources → configure and interact → monitor

discover resources

- Fine-grained
- Complete
- Up-to-date
- Versioned
- Verifiable

provision resources

- Advance reservations & on-demand
- Isolation
- Fine-grained allocations

configure and interact

- Deeply reconfigurable
- Appliance catalog
- Snapshotting
- Complex Appliances
- Network Isolation

monitor

- Hardware metrics
- Fine-grained information
- Aggregate and archive

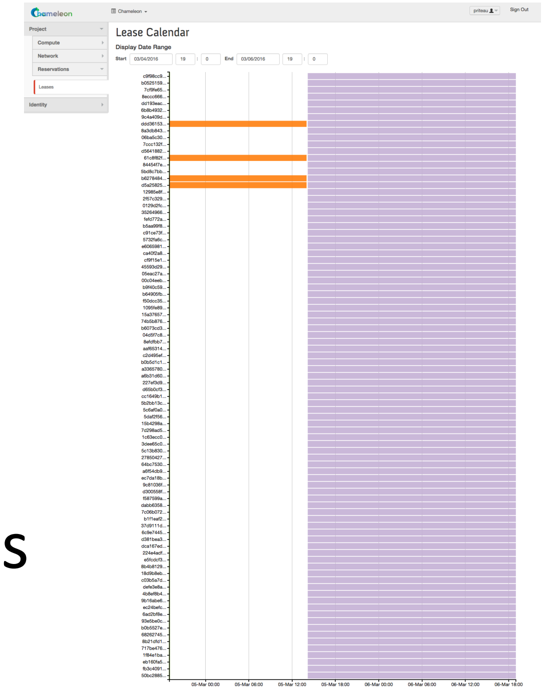
CHI: DISCOVERING AND VERIFYING RESOURCES

- 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.



Lease Calendar

Display Date Range

Start 03/04/2016 19 : 0 End 03/06/2016 19 : 0

- 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

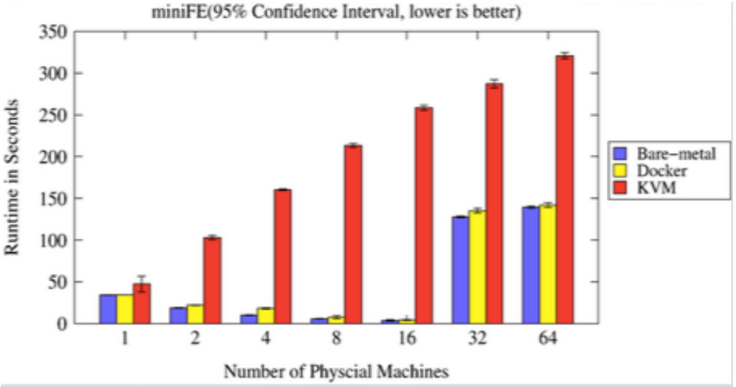
- OpenStack Ceilometer + agents, standard metrics (CPU, memory, network, disk usage, etc.)
- RAPL interface to provide power and energy usage

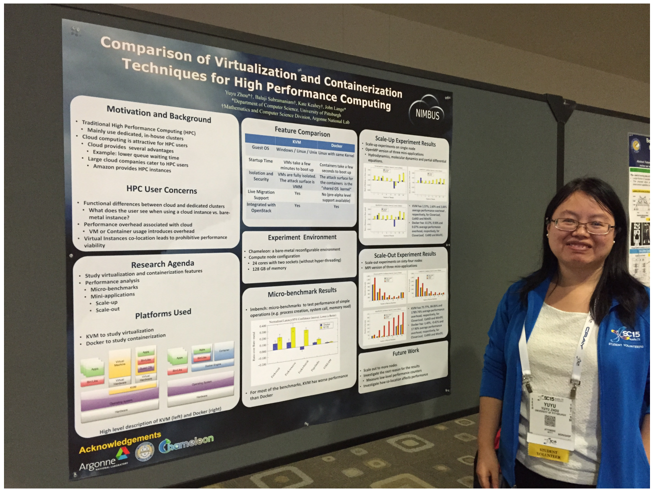
CHAMELEON: TIMELINE AND STATUS

- **10/14: Project starts**
- 04/15: Chameleon Core Technology Preview
- 06/15: Chameleon Early User on new hardware
- **07/15: Chameleon public availability**
- Throughout 2016: New capabilities and new hardware releases
- **Today: 1,400+ users/200+ projects**
- Planning for the next 3-4 years

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

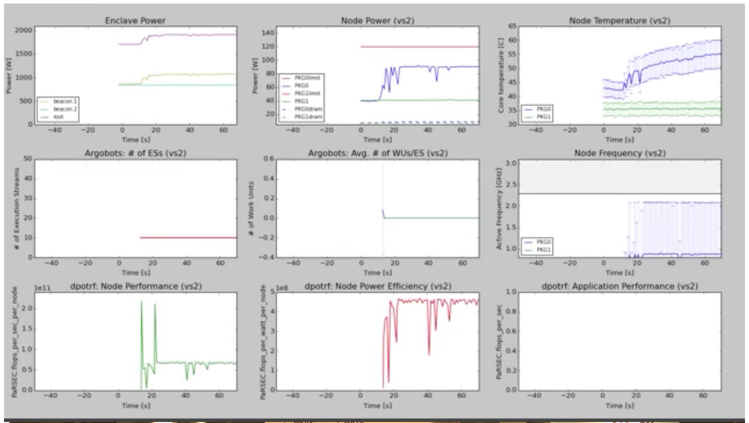



Comparison of Virtualization and Containerization Techniques for High Performance Computing
NIMBUS
Acknowledgements
Argonne
Chameleon

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

EXASCALE OPERATING SYSTEMS

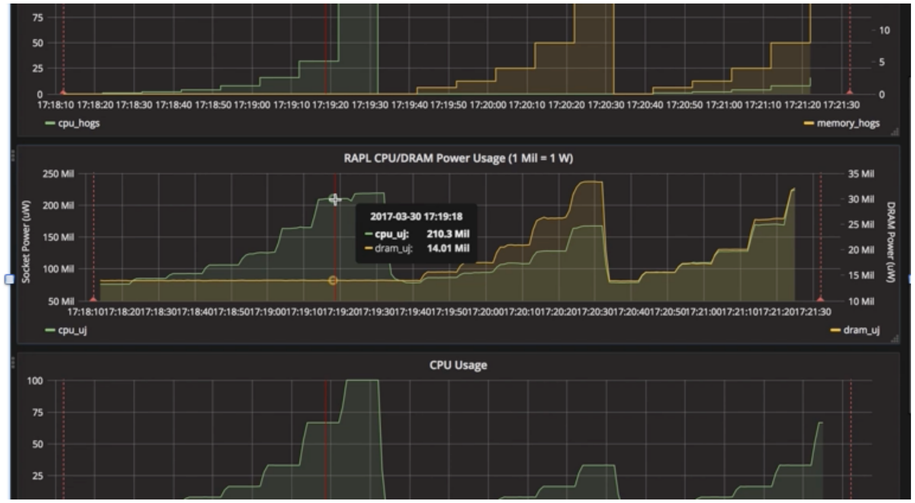
- Swann Perarnau, ANL
- Research: exascale operating systems
- Testbed requirements:
  - Bare metal reconfiguration
  - Boot kernel with varying kernel parameters
  - Fast reconfiguration, many different images, kernels, params
  - Hardware: performance counters, many cores



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

TOWARDS A SCIENTIFIC INSTRUMENT

- Existing elements
  - Testbed versioning
  - Appliances
- Experiment precis: closing the gap between resources, appliances and data
- Experiment logbook: keep better notes
  - Many existing tools (Jupyter, Grafana, etc.)
  - Creative integration with existing technologies
- From summaries to replays



WHO CAN USE CHAMELEON?

- Any US researcher or collaborator
- Chameleon Projects
    - Created by faculty or staff
    - Who joins the project is at their discretion
    - Allocation of 20K service units(SUs)
    - Easy to extend or recharge
- Key policies
    - Lease limit of 1 week (with exceptions)
    - Advance reservations

SUMMARY

- **Open** experimental testbed for **Computer Science research**: 1,400+ users/200+ projects
- Designed from the ground up for a **large-scale** testbed supporting **deep reconfigurability**
- Blueprint for a **sustainable operations model**: a CS testbed powered by OpenStack
- Working towards a **connected** instrument: from insight to repeatability

“We shape our buildings;
thereafter they shape us”

Winston Churchill

We want to make us all dream big

www.chameleoncloud.org

keahey@anl.gov