www. chameleoncloud.org

WE NEED CLOUDS TO BUILD CLOUDS:
DEVELOPING AN OPEN CLOUD TESTBED USING OPENSTACK

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

SEARCHING FOR A TESTBED

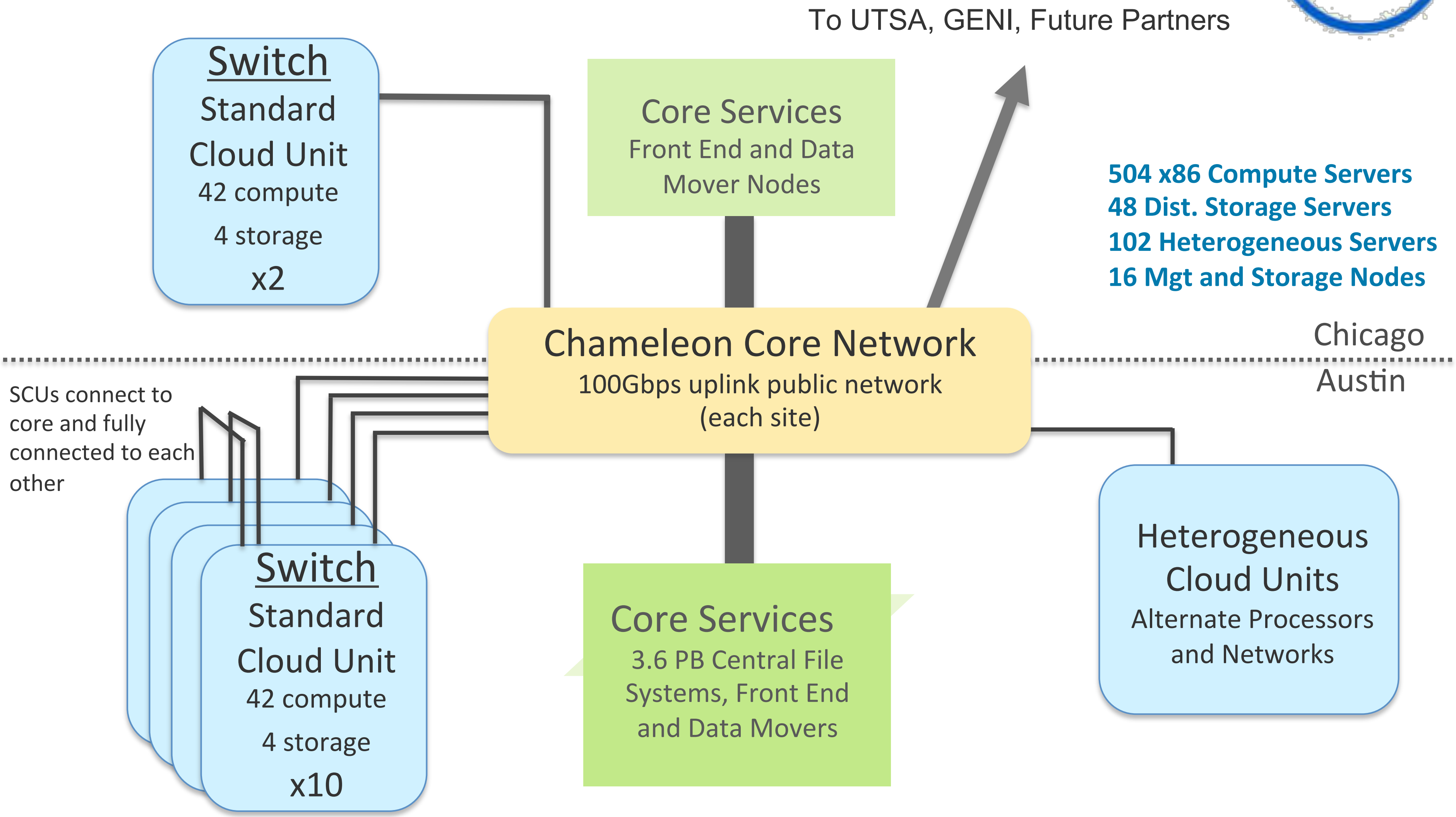
- A personal quest
  - The case of no testbed at all
  - The case of inadequate: “no hardware virtualization”
  - The case of too small: “we think this will scale”
  - The case of shared: “it may have impacted our result”

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

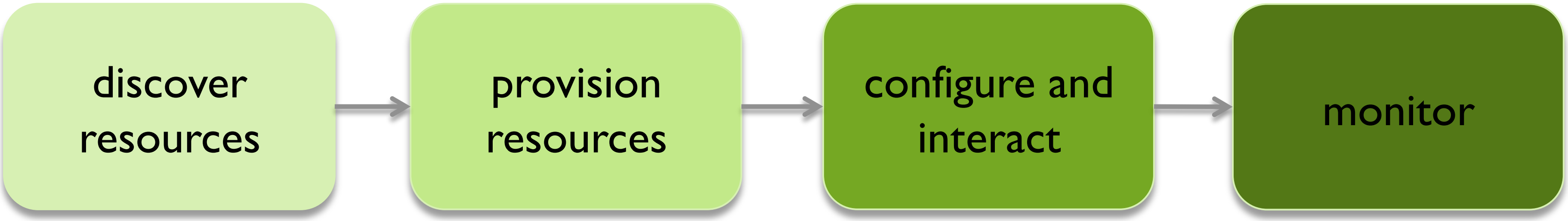
Core Services
3.6 PB Central File Systems, Front End and Data Movers

Heterogeneous Cloud Units
Alternate Processors and Networks

CHAMELEON 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

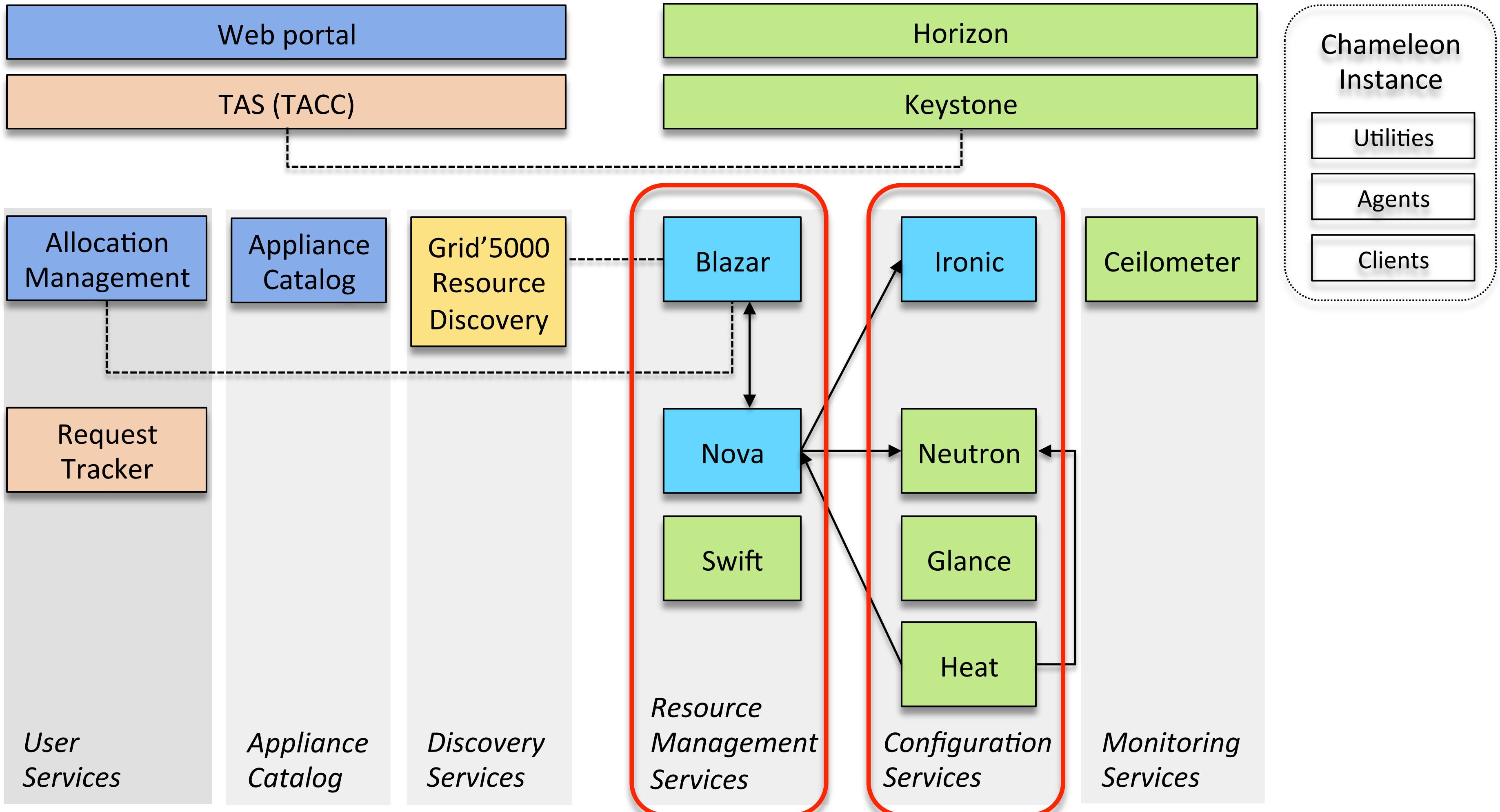
configure and interact

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

monitor

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

CHAMELEON IMPLEMENTATION



Web portal

TAS (TACC)

Horizon

Keystone

Chameleon Instance

Utilities

Agents

Clients

Allocation Management

Request Tracker

User Services

Appliance Catalog

Appliance Catalog

Grid'5000 Resource Discovery

Discovery Services

Blazar

Nova

Swift

Resource Management Services

Ironic

Neutron

Glance

Heat

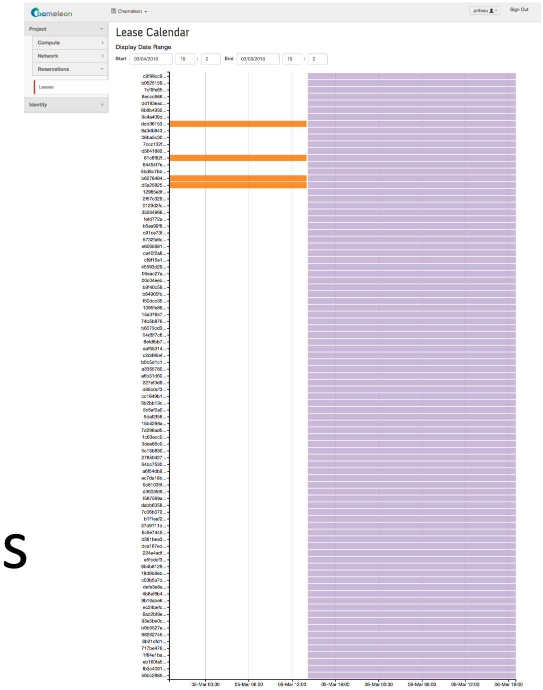
Configuration Services

Ceilometer

Monitoring Services

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

CHAMELEON AND IRONIC

- Snapshotting
  - Our solution: snapshot script inside images
  - Tarball of the root file system into a QCOW2 image
  - Leverages libguestfs tools
  - Supports both whole disk and partition images
- Network isolation with dynamic VLANs
  - Neutron configured with ODL plugin
  - ODL extended to change VLAN port assignment on Dell switch
  - Ironic modified to trigger Neutron port update
  - May be superseded by multi-tenant bare metal networking (Ocata)
- Ironic feature requests
  - Support for multiple networks
  - Gracefully support failures (retry IPMI contact after some time)
  - Faster deployment with kexec

APPLIANCES AND THE APPLIANCE CATALOG

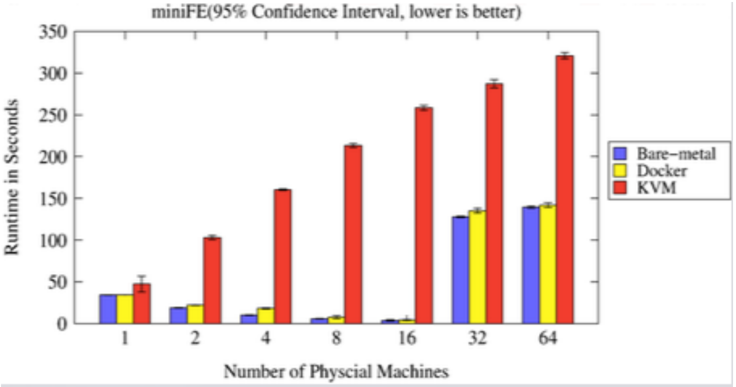
- Chameleon appliance
  - Chameleon bare metal image, same format for UC and TACC
  - Common tools: cc-checks, cc-shapshot, power measurement utility, Ceilometer agent, Heat agent
- System appliances:
  - Base images: CentOS 7, Ubuntu (3 versions)
  - Heterogeneous hardware support: CUDA (2 versions), FPGA
  - SR-IOV support: KVM, MPI-SRIOV on KVM cluster, RDMA Hadoop, MVAPICH
  - Popular applications: DevStack OpenStack (3 versions), TensorFlow, MPI, NFS
- User contributed

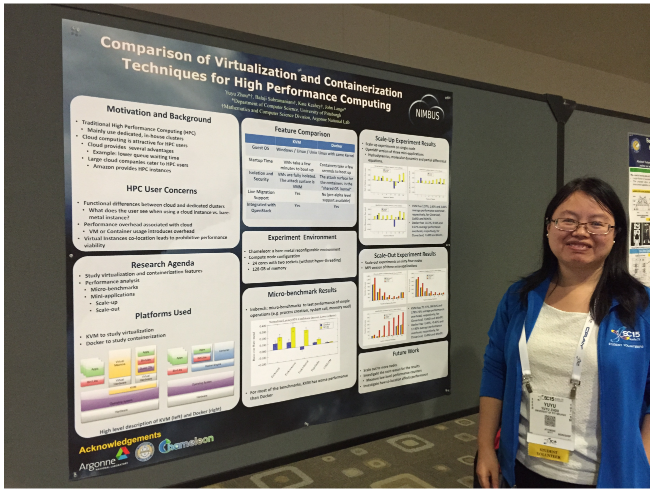
CHAMELEON: TIMELINE AND STATUS

- **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**

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

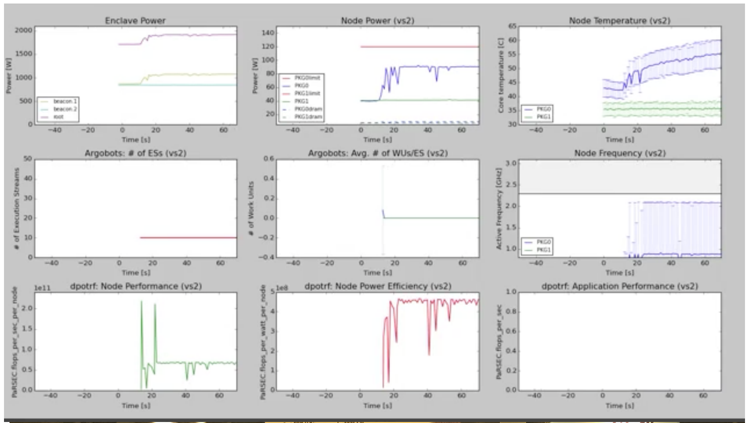



Comparison of Virtualization and Containerization Techniques for High Performance Computing
NIMBUS
Motivation and Background
HPC User Concerns
Research Agenda
Platforms Used
Feature Comparison
Experiment Environment
Micro-benchmark Results
Scale-Up Experiment Results
Scale-Out Experiment Results
Future Work
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



See the ... for schedule with abstracts
EXASC
STARTS HERE
atipa technologies
DDN
MISSISSIPPI STATE UNIVERSITY

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

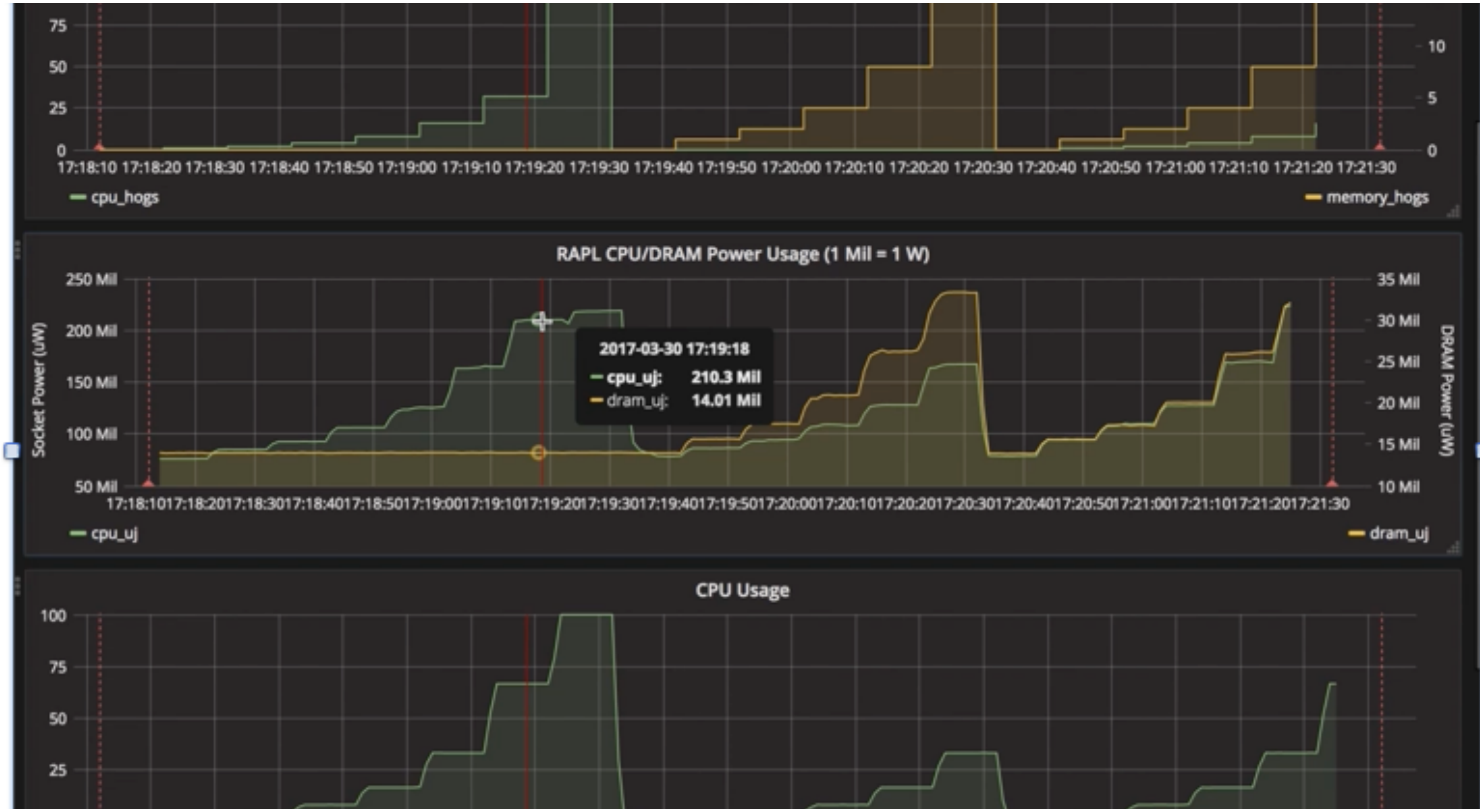
CLOUD WORKLOAD TRACES

- Scientific Working Group activity
- Reviewed existing traces for HPC, Grid computing, etc. to determine requirements
- Defining the trace format
- Populating the traces
  - Export Nova DB info as list of events
  - Optionally combine with telemetry data?
- Evaluating tools to replay workload
  - Can we leverage Rally?
  - Or osic/ops-workload-framework
- Seeking volunteers to provide traces

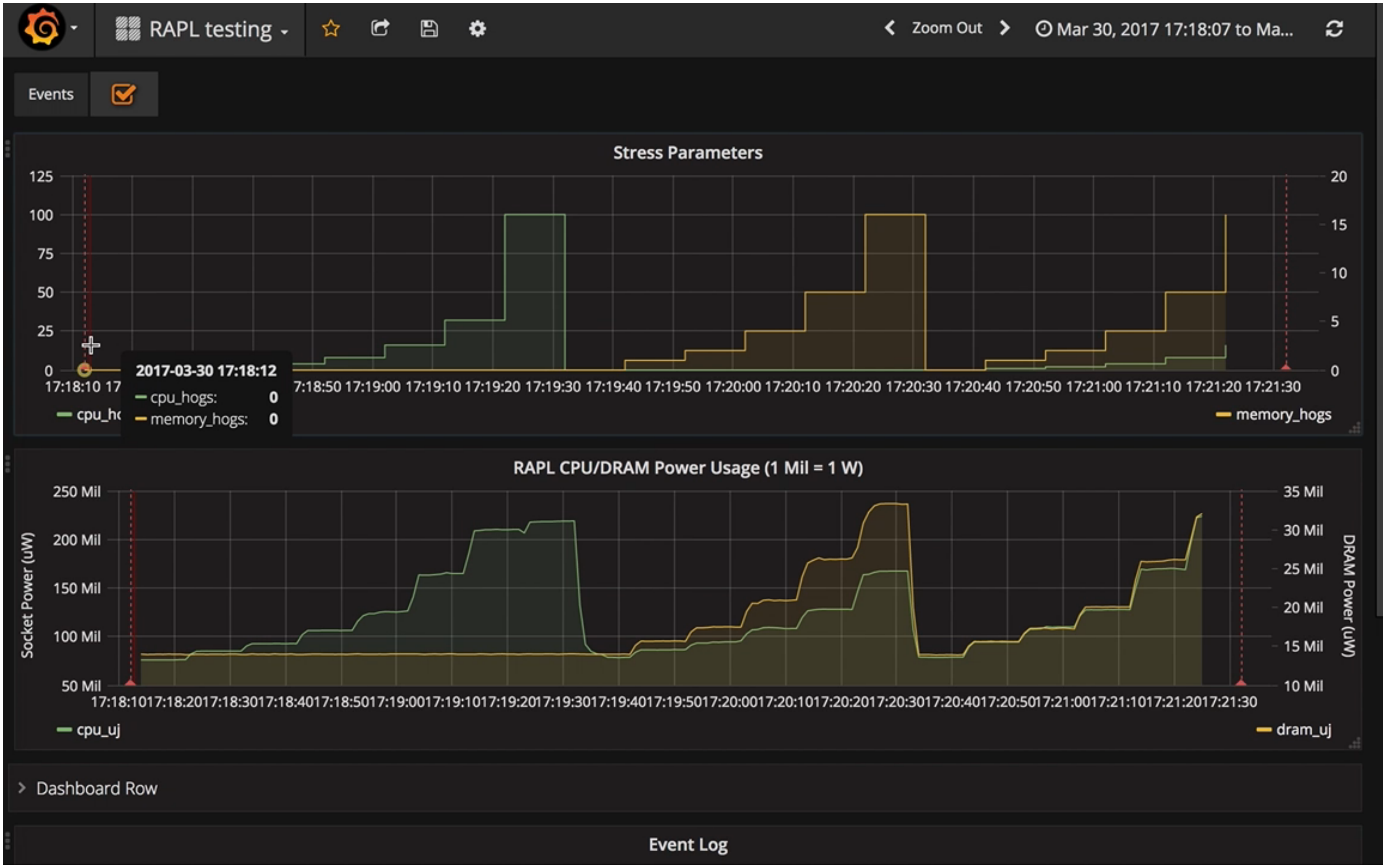
TOWARDS A SCIENTIFIC INSTRUMENT

- Scientific instrument: built for the purpose of observing, measuring, and recording scientific phenomena
- 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
- Experiment logbook: keep better notes
  - Many existing tools (Jupyter, Grafana, etc.)
  - Creative integration with existing technologies

FROM DATA TO INSIGHT



FROM DATA TO INSIGHT



FROM INSIGHT TO REPEATABILITY

- Existing elements
  - Testbed versioning (53 versions so far)
  - Appliance publication, versioning, and management
- Experiment summaries: closing the gap between resource versions, appliances, and data
- From experiment summaries to experiment replays
- Publishing experiment summaries
- Looking for summer students!

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

DEBUNKING CHAMELEON FAKE NEWS

- "I need to have NSF funding to use Chameleon"
  - **Not true:** Chameleon is an **open** testbed: all PIs with **research** projects will be considered
- "I can't do bare metal on Chameleon, all they do is VMs"
  - **Not true:** almost all of Chameleon support **bare metal** reconfiguration, only a very small partition is configured with KVM
- "I can't provision hundreds of nodes on Chameleon"
  - **Not true:** while at any given time hundreds of nodes may not be available, you can make an advance reservation to get hundreds of nodes in near future

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