www. chameleoncloud.org

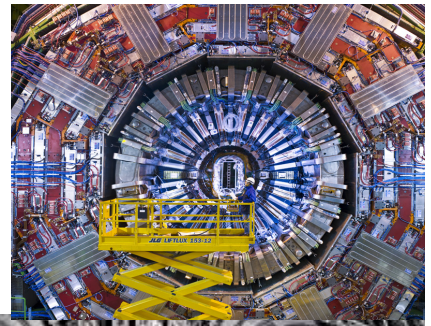
CHAMELEON: A LARGE SCALE, RECONFIGURABLE EXPERIMENTAL INSTRUMENT FOR COMPUTER SCIENCE

Kate Keahey

Joe Mambretti, DK Panda, Paul Rad, Pierre Riteau, Dan Stanzione

A PERSONAL QUEST

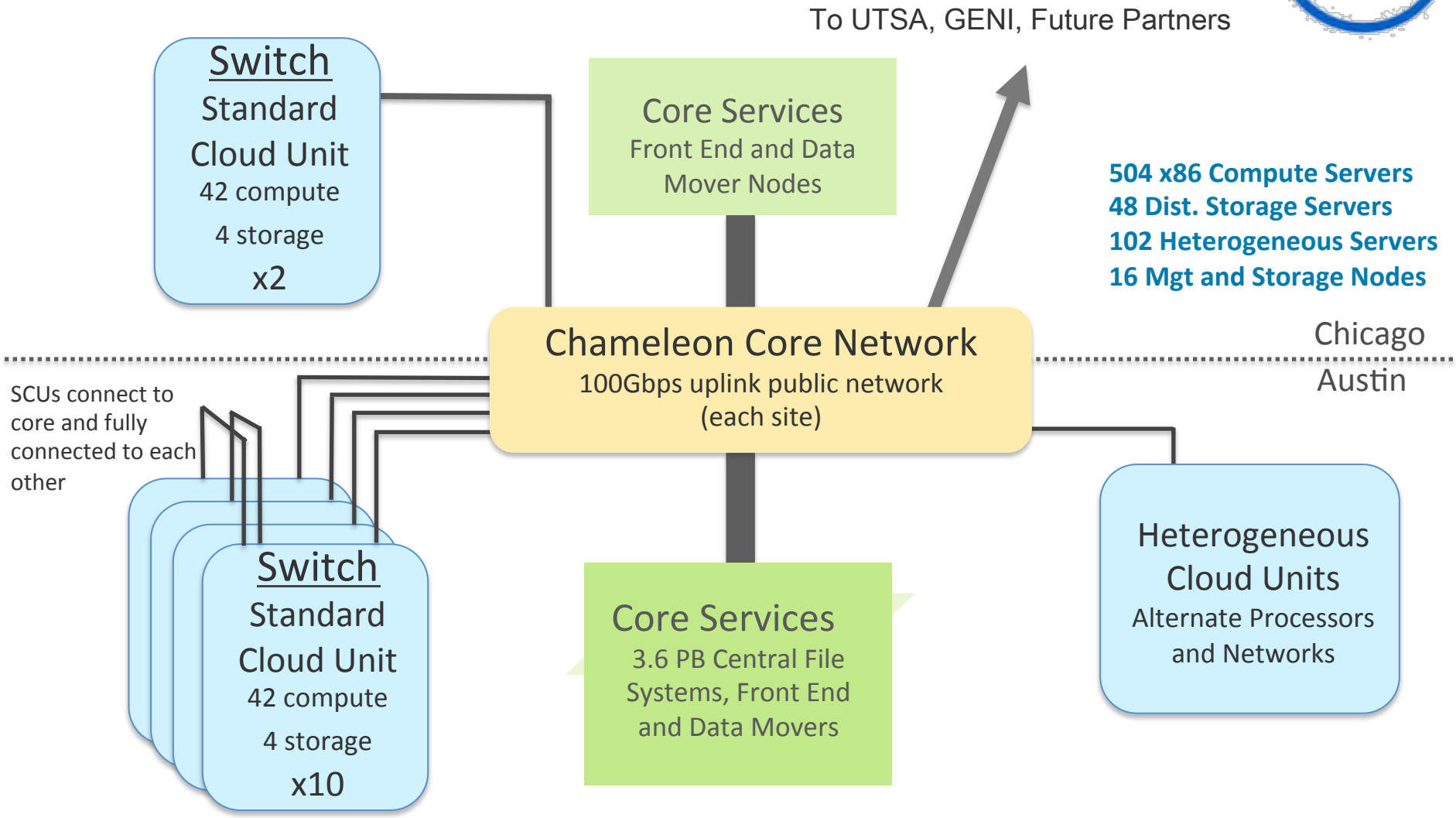
- Searching for an experimental instrument for Computer Science
  - No instrument at all
  - Inadequate: “no hardware virtualization”
  - Too small: “we think this will scale”
  - Shared: “it may have impacted our result”
- Compare with other sciences



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: partnerships with production clouds
  - Appliance Catalog: partnerships with users
- Complementary: "Can't do everything ourselves"
  - Complementing GENI, Grid'5000, and other experimental testbeds
- Sustainable: "Easy to operate, easy to share"

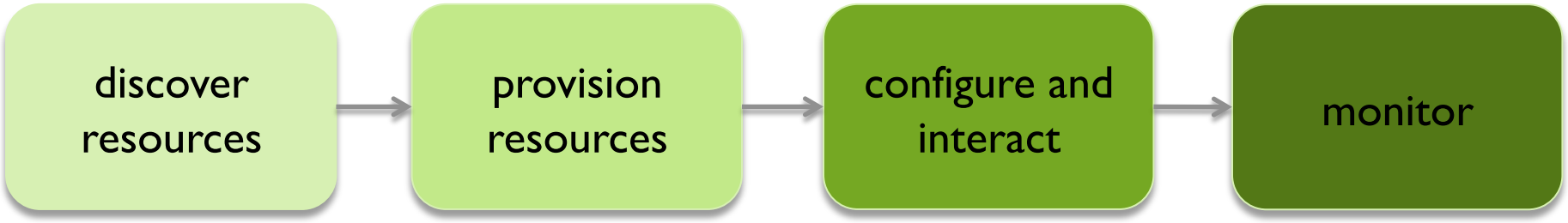
CHAMELEON HARDWARE



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



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

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

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

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

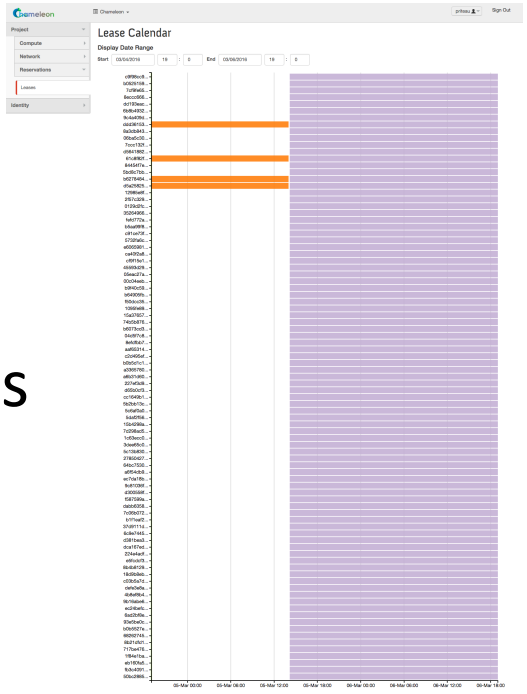
CHI: DISCOVERING AND VERIFYING RESOURCES

- 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.
- Future extensions: match making, testbed allocation management



- OpenStack Nova/Blazar; extensions to Blazar
- Extensions to support Gantt chart displays and several smaller features

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

APPLIANCES AND THE APPLIANCE CATALOG

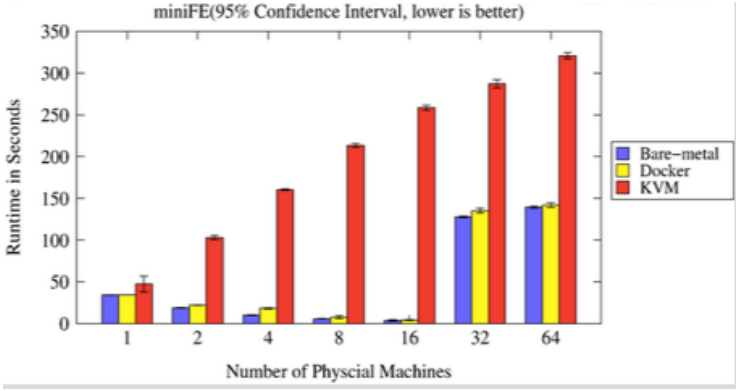
- Chameleon appliance
  - Chameleon bare metal image, same format for UC and TACC
  - Common tools: cc-checks, cc-shapshot, power measurement utility, Ceilometer agent, Heat agent
- System appliances:
  - Base images: CentOS 7, ubuntu (3 versions)
  - Heterogeneous hardware support: CUDA (2 versions),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

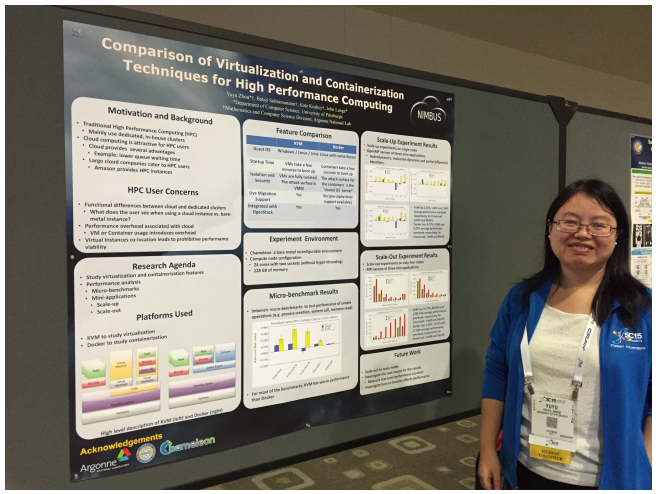
CHAMELEON CORE: TIMELINE AND STATUS

- **10/14: Project starts**
- 12/14: FutureGrid@Chameleon (OpenStack KVM cloud)
- 04/15: Chameleon Technology Preview on FG hardware
- 06/15: Chameleon Early User on new hardware
- **07/15: Chameleon public availability (bare metal)**
- 09/15: Chameleon KVM OpenStack cloud available
- 2016: Heterogeneous hardware releases + new capabilities
- **Today: 1,300+ 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

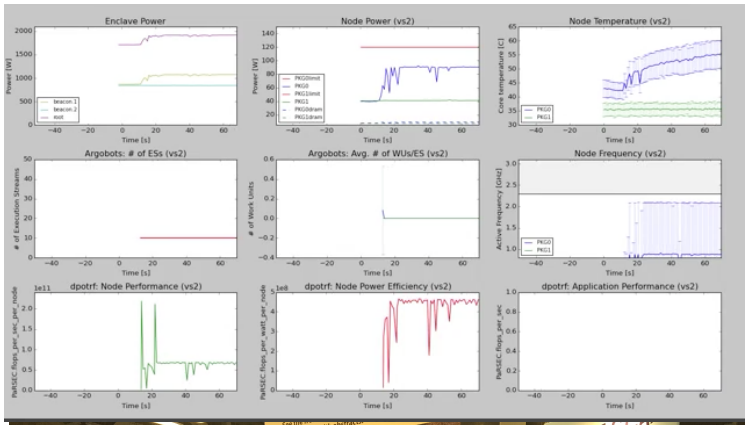




SC15 Poster: "Comparison of Virtualization and Containerization Techniques for HPC"

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"

WHO CAN USE CHAMELEON?

- Any US researcher – or collaborator
- Projects have to be created by faculty or staff
  - Who joins the project is at their discretion
- Key policies
  - Allocation of 20K SUs (extensible, rechargable)
  - Lease limit of 1 week (with exceptions)
  - Advance reservations

PARTING THOUGHTS

- Scientific instrument for **Computer Science research**: 1,300+ users/200+ projects
- Designed from the ground up for a **large-scale** testbed supporting reconfigurable experimentation
- Blueprint for a **sustainable operations model**: building a CS testbed out of commodity components
  - Return on investment, ability to contribute, and sustainable operation
- Towards a scientific instrument: support for repeatability and insight

www.chameleoncloud.org

keahey@anl.gov