www. chameleoncloud.org

CHAMELEON: BUILDING AN EXPERIMENTAL INSTRUMENT FOR COMPUTER SCIENCE AS APPLICATION OF CLOUD COMPUTING

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

September 28th, 2016
Miami, FL

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

WHY EXPERIMENT?

"Beware of bugs in the above code;
I have only proved it correct, not tried it"

(Donald Knuth)

"In theory there is no difference between theory and practice. In practice there is."

(Yogi Berra)

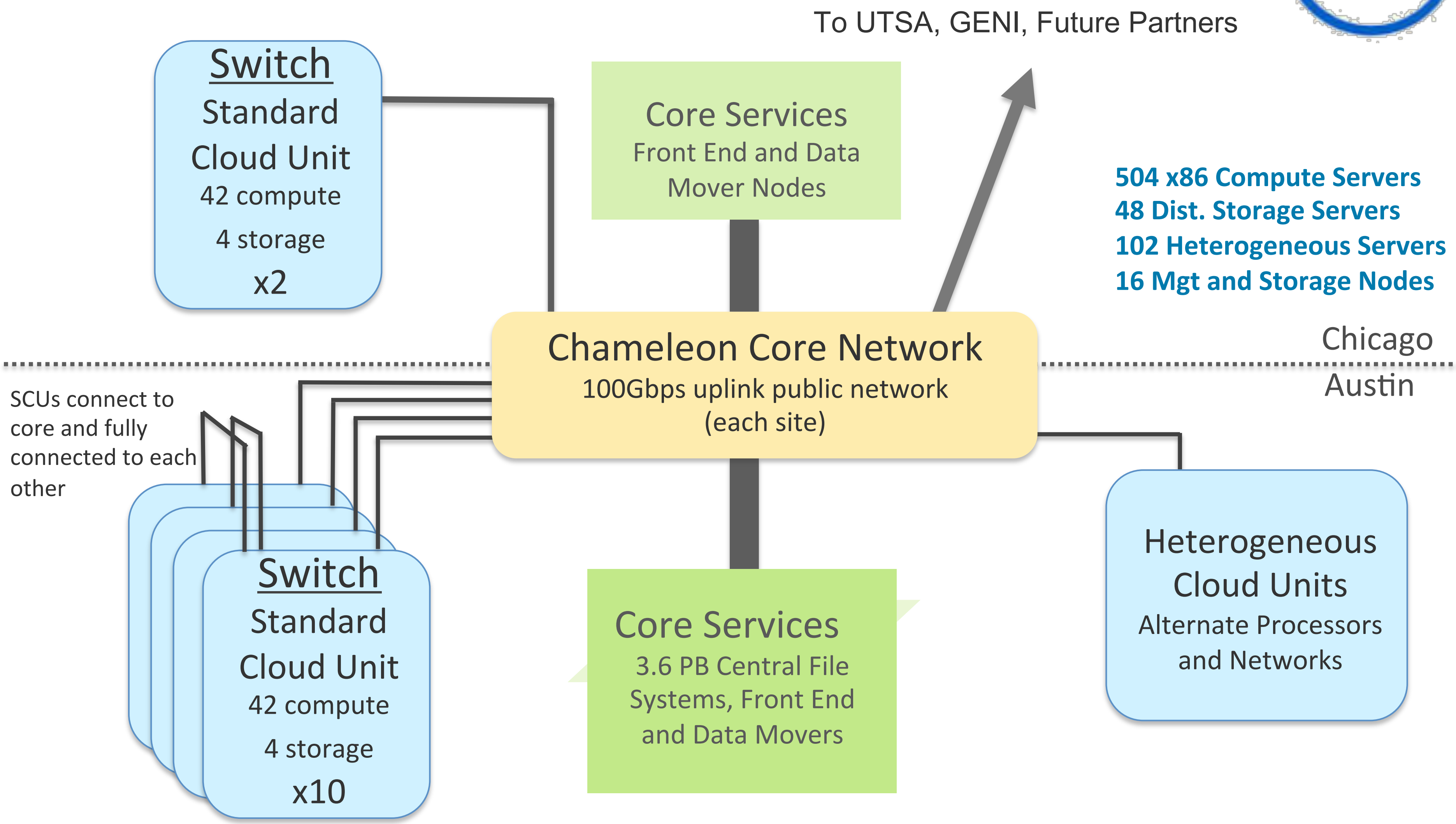
EXPERIMENTS AND MODELS

- Models
  - Essential to understand the problem
  - Are they: Correct?, Too complex? Not complex enough?
  - Need to be discovered by gaining experience about a problem, environment, or solutions
- Experimentation
  - Isolation: why a cloud is not sufficient for cloud research
  - Repeatability: repeat the same experiment multiple times in the same context while varying different factors
  - Reproducibility: the ability to repeat an experiment by a different agency
- Requirements for deep reconfigurability and control

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
- Reconfigurable: "As close as possible to having it in your lab"
  - Deep reconfigurability (bare metal) and isolation
  - Fundamental to support reproducible experiments
- Connected: "One stop shopping for experimental needs"
  - Workload and Trace Archive: partnerships with production clouds: CERN, OSDC, Rackspace, Google, and others
  - Sharing appliances: partnerships with users
- Complementary: "Can't do everything ourselves"
  - Complementing GENI, Grid'5000, and other experimental testbeds
- Sustainable: "Easy to maintain, easy to share"

CHAMELEON HARDWARE



CHAMELEON HARDWARE (MORE DETAIL)

- "Start with large-scale homogenous partition" (deployed)
  - 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 (deployed)
  - 3.6 PB global storage, 100Gb Internet connection between sites
- "Graft on heterogeneous features" (still evolving)
  - Infiniband network in one rack with SR-IOV support (deployed)
  - High-memory, NVMA, SSDs, and GPUs on selected nodes (deployed)
  - FPGAs, ARM microservers and Atom microservers (coming soon)

CAPABILITIES AND SUPPORTED RESEARCH

Development of new models, algorithms, platforms, auto-scaling HA, etc., innovative application and educational uses

Persistent, reliable, shared clouds: modest OpenStack KVM cloud

Repeatable experiments in new models, algorithms, platforms, auto-scaling, high-availability, cloud federation, etc.

Isolated partition, Chameleon Appliances: CHI + Chameleon appliances

Virtualization technology (e.g., SR-IOV, accelerators), systems, networking, infrastructure-level resource management, etc.

Isolated partition, full bare metal reconfiguration: CHI

EXPERIMENTAL WORKFLOW REQUIREMENTS

discover resources	provision resources	configure and interact	monitor
- Fine-grained - Complete - Up-to-date - Versioned - Verifiable	- Advance reservations & on-demand - Isolation - Fine-grained allocations	- Deeply reconfigurable - Appliance catalog - Snapshotting - Super Appliances: “one-click virtual clusters”	- Hardware metrics - Fine-grained information - Aggregate and archive

BUILDING A TESTBED FROM SCRATCH

- Requirements (proposal stage)
- Architecture (project start)
- Technology Evaluation and Risk Analysis
  - Many options: G5K, Nimbus, LosF, OpenStack
  - Sustainability as design criterion: can a CS testbed be built from commodity components?
  - Technology evaluation: Grid'5000 and OpenStack
  - Architecture-based analysis and implementation proposals
- Implementation (~3 months)
- Result: Chameleon Infrastructure (CHI) =
  - 65%*OpenStack + 10%*G5K + 25%*"special sauce"
- Integration environments versus production

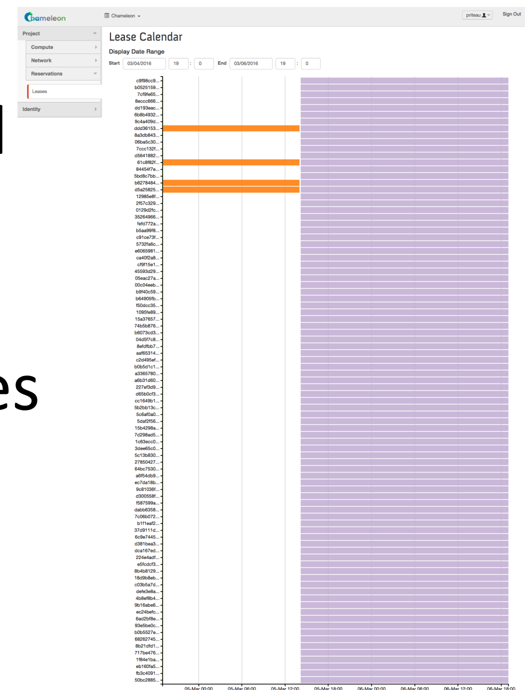
CHI: DISCOVERING AND VERIFYING RESOURCES

- Fine-grained, up-to-date, and complete representation
- Both machine parsable and user friendly representations
- 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 description, automated 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, AR: 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 image sharing
- Map multiple appliances to a lease
- Appliance Catalog
- Handle complex appliances
  - Virtual clusters, cloud installations, etc.
- Interact: shape experimental conditions

- OpenStack Ironic, Glance, and meta-data servers
- Added snapshotting and appliance management

CHI: INSTRUMENTATION AND MONITORING

- Enables users to understand what happens during the experiment
- Instrumentation: high-resolution metrics
- Types of monitoring:
  - Infrastructure monitoring (e.g., PDUs)
  - User resource monitoring
  - Custom user metrics
- Aggregation and Archival
- Easily export data for specific experiments

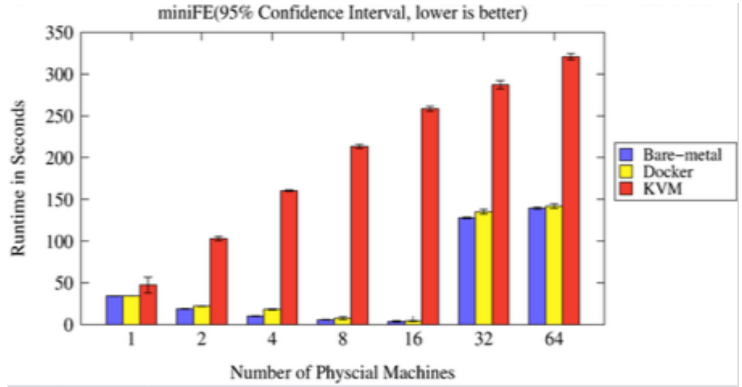
- OpenStack Ceilometer + custom metrics

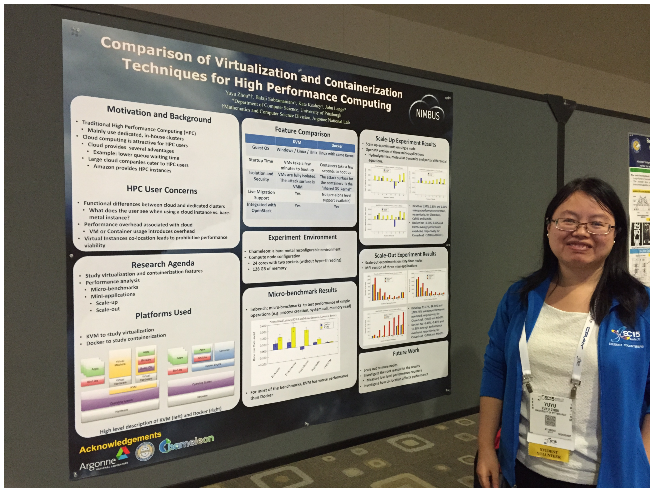
CHAMELEON CORE: TIMELINE AND STATUS

- **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
- 10/15: Identity federation with GENI
- **Today: 1,000+ users/200+ projects**
- 2016: Heterogeneous hardware releases

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

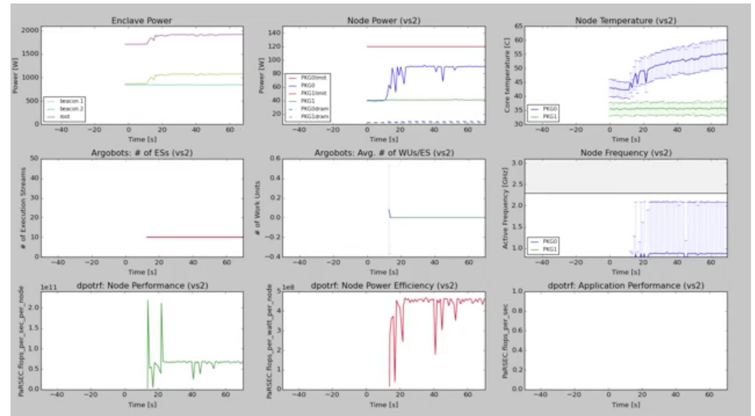




SC15 Poster: "Comparison of Virtualization and Containerization Techniques for HPC"

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

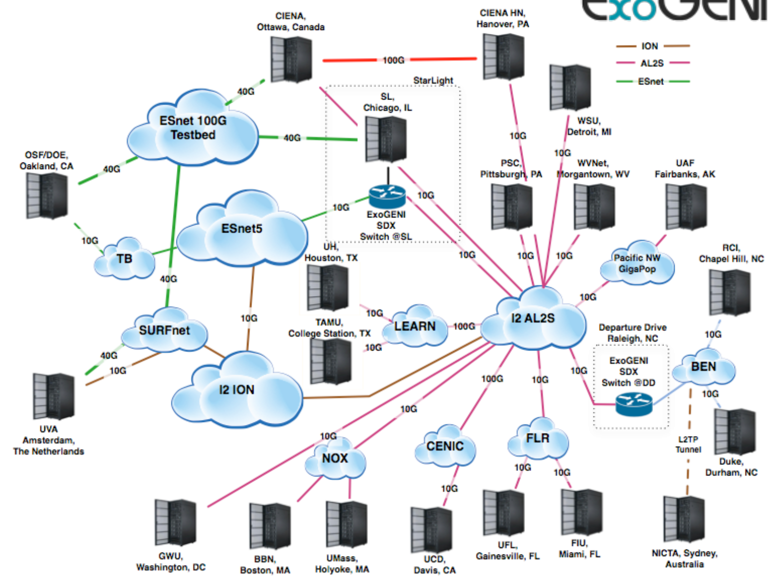
FEDERATING NETWORKS

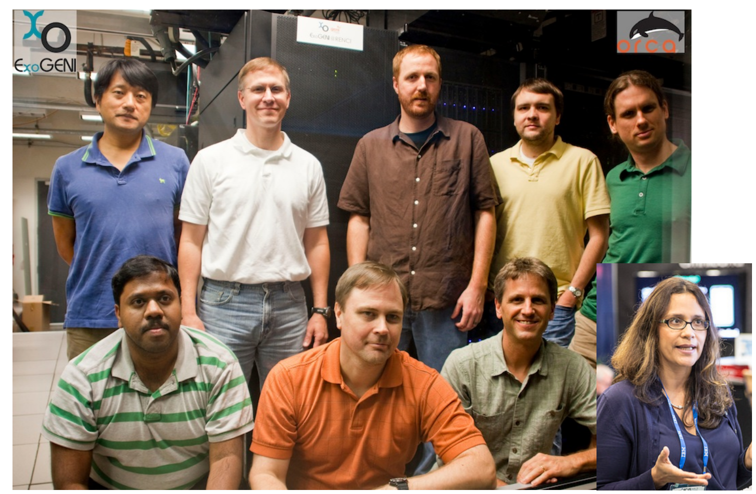
- Paul Ruth, RENCI-UNC Chapel Hill
- Research: Federated Networked Clouds for Domain Science
- Testbed requirements:
  - Deploy ExoGENI on Chamelelon
  - “Stitch” Layer-2 networks between Chameleon and external systems
  - HPC (e.g. Infiniband, SR-IOV, MPI, many cores, performance isolation)

http://www.exogeni.net

geni Exploring Networks of the Future

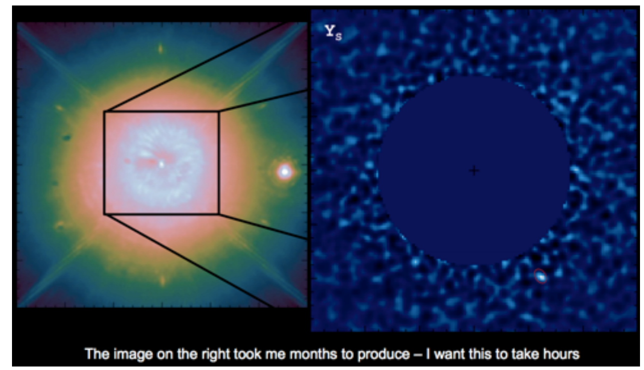
ExoGENI





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



FROM POSSIBLE TO EASY…

Y1: Make things possible

- Develop CHI
- Deploy new hardware

Y2: From possible to easy

Valentine's Day goodies:

- Custom kernel/console
- Liberty upgrade
- Appliance marketplace and appliances

Y2: From possible to easy

Independence Day goodies:

- Heterogeneous hardware
- Object store
- Appliance tools and appliances

… AND ON

PARTING THOUGHTS

- Scientific instrument for CS experimental research
- Open testbed: work on your next research project @

www.chameleoncloud.org

The most important element of any experimental testbed is users and the research they work on

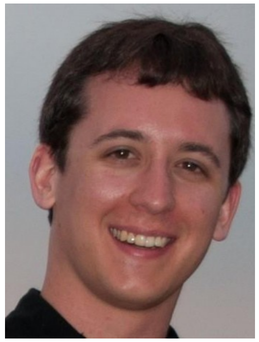
- From vision to reality with Express Delivery
  - Built from scratch in less than a year on a shoestring
  - Operational testbed: 1,000+ users/200+ projects
- Blueprint for a new, sustainable operations model: building a CS testbed as an application of cloud computing: benefits for us, for the broader community, and for other testbeds

CHAMELEON TEAM

Kate Keahey
Chameleon PI
Science Director
Architect
University of Chicago

Paul Rad
Industry Liason
Education and training
UTSA

Joe Mambretti
Programmable networks
Federation activities
Northwestern University



Pierre Riteau
Devops Lead
University of Chicago

DK Panda
High-perf networking
Ohio State University

Dan Stanzione
Facilities Director
TACC