www. chameleoncloud.org

CHAMELEON: BUILDING A RECONFIGURABLE EXPERIMENTAL TESTBED FOR CLOUD RESEARCH

Kate Keahey

keahey@anl.gov

AMGCC 2015
September 21
Boston, MA

WHY EXPERIMENT?

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

(Donald Knuth)

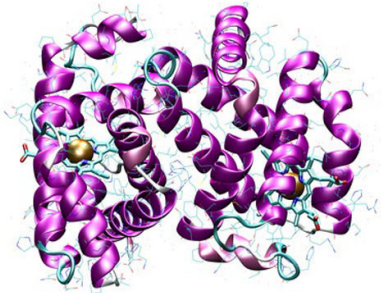
"In theory there is no difference between theory and practice. In practice there is."

(Yogi Berra)

EXPERIMENTS AND MODELS

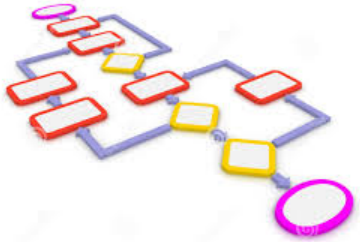
- Models
  - Essential to understand the problem
  - Correctness, tractability, complexity
- Experimentation
  - Isolation: why a cloud is not sufficient for cloud research
  - Repeatability: repeat the same experiment multiple times in the same context while varying different factors
  - Reproducibility: the ability to repeat an experiment by a different agency
  - Fine-grained information everywhere
- Requirements for deep reconfigurability and control

CLOUD RESEARCH CHALLENGES



BigData Management and Analytics

Highly Distributed Cloud Frameworks



BigData Algorithms

Research at Scale:
Big Data, Big Compute, Big Instrument

Collaboration at Scale

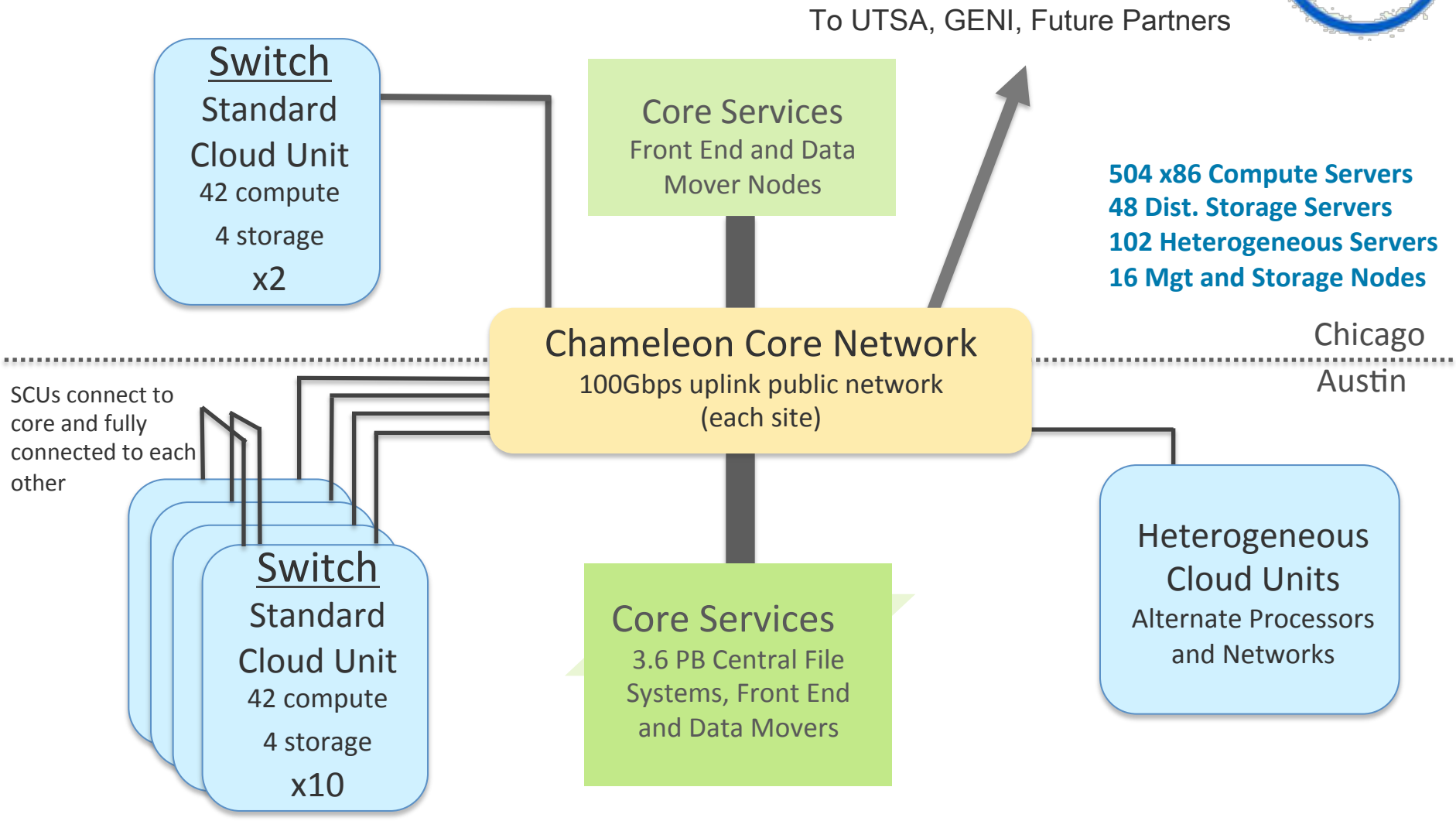
Short Response at Large Scale

Cloud Data Security

CHAMELEON DESIGN STRATEGY

- Large-scale: "Big Data, Big Compute, Big Instrument research"
  - ~650 nodes (~14,500 cores), 5 PB disk over two sites, 2 sites connected with 100G network
- Reconfigurable: "As close as possible to having it in your lab"
  - From bare metal reconfiguration to clouds
  - Support for repeatable and reproducible experiments
- Connected: "One stop shopping for experimental needs"
  - Workload and Trace Archive
  - Partnerships with production clouds: CERN, OSDC, Rackspace, Google, and others
  - Partnerships with users
- Complementary: "Can't do everything ourselves"
  - Complementing GENI, Grid'5000, and other experimental testbeds

CHAMELEON HARDWARE



STANDARD CLOUD UNIT

- Each of the 12 SCUs is comprised of a single 48U rack
  - Allocations can be an entire SCU, multiple SCUs, or within a single one.
- A single 48 port Force10 s6000 OpenFlow-enabled switch connects all nodes in the rack (with an additional network for management/control plane).
  - 10Gb to hosts, 40Gb uplinks to Chameleon core network
- An SCU has 42 Dell R630 compute servers, each with dual-socket Intel Xeon (Haswell) processors and 128GB of RAM
- In addition, each SCU has 4 DellFX2 storage servers, each with a connected JBOD of 16 2TB drives.
  - Can be used as local storage within the SCU, or allocated separately (48 total available for Hadoop configurations)

HETEROGENEOUS CLOUD UNITS

- One of the SCUs will also contain Connectx3 Infiniband network
- Additional HCUs will contain:
  - Atom microservers
  - ARM microservers
  - A mix of servers with:
    - High RAM
    - FPGAs (Xilinx/Convey Wolverine)
    - NVidia K40 GPUs
    - Intel Xeon Phis

CHAMELEON CORE HARDWARE

- Shared Infrastructure:
  - In addition to distributed storage nodes, Chameleon will have 3.6PB of central storage, for a *persistent* object store and shared filesystem.
  - An additional dozen management nodes will provide data movers, user portal, provisioning services, and other core functions within Chameleon.
- Core Network
  - Force10 OpenFlow-enabled switches will aggregate the 40Gb uplinks from each unit and provide multiple links to the 100Gb Internet2 layer 2 service.

CAPABILITIES AND SUPPORTED RESEARCH

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

Persistent, reliable, shared clouds

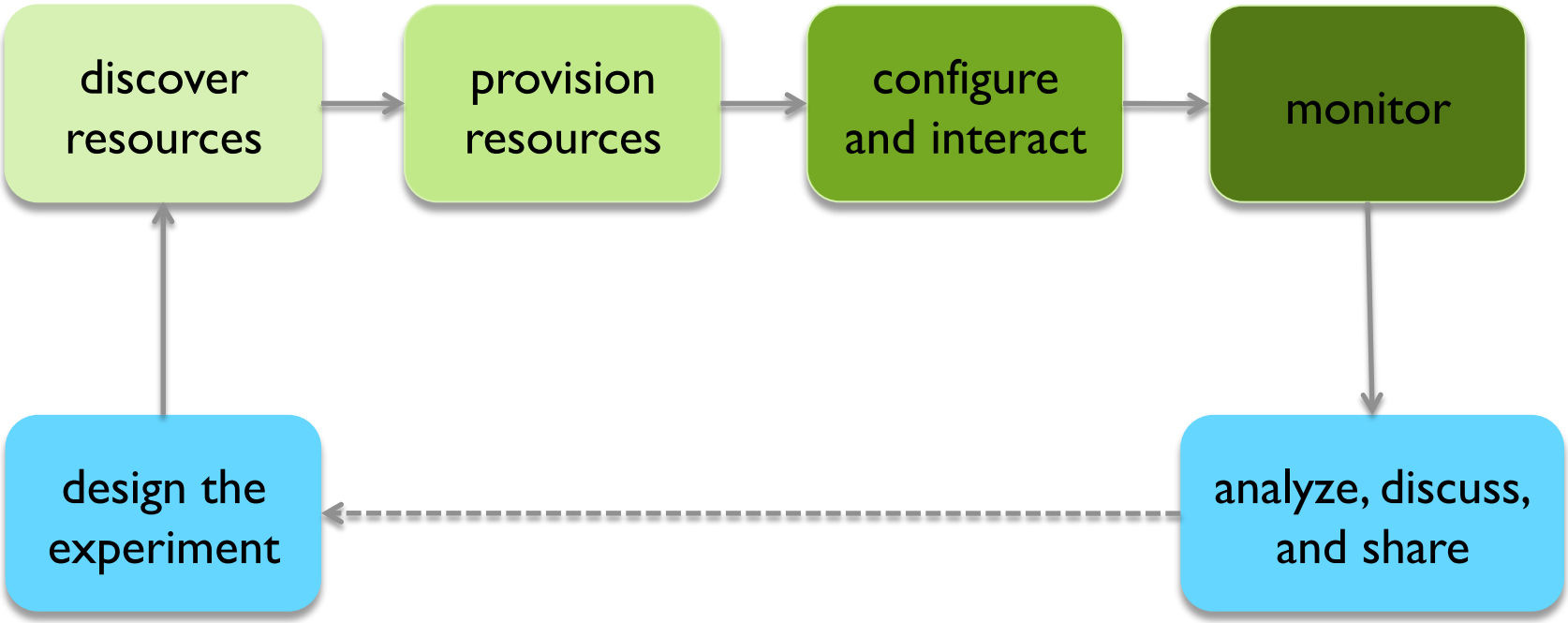
Repeatable experiments in new models, algorithms, platforms, auto-scaling, high-availability, cloud federation, etc.

Isolated partition, Chameleon Appliances

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

Isolated partition, full bare metal reconfiguration

USING CHAMELEON: THE EXPERIMENTAL WORKFLOW



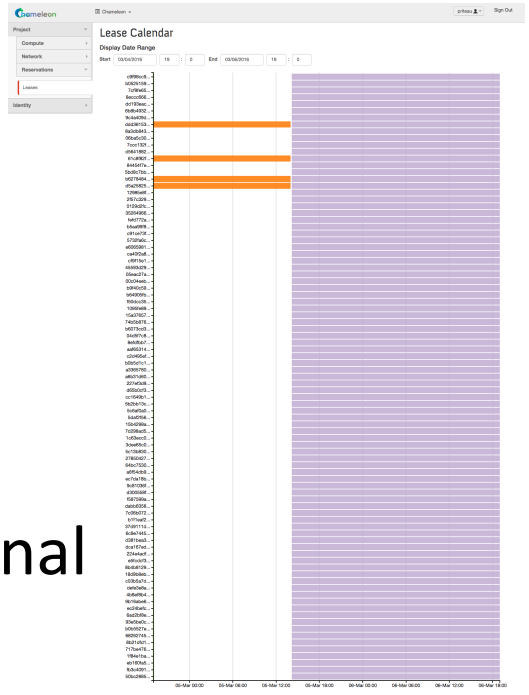
CHI: SELECTING AND VERIFYING RESOURCES

- Complete, fine-grained and up-to-date representation
- Machine parsable, enables match making
- Versioned
  - "What was the drive on the nodes I used 6 months ago?"
- Dynamically Verifiable
  - Does reality correspond to description? (e.g., failures)

- Grid'5000 registry toolkit + Chameleon portal
  - Automated resource description, automated export to RM
- G5K-checks
  - Can be run after boot, acquires information and compares it with resource catalog description

CHI: PROVISIONING RESOURCES

- Resource leases
- Allocating a range of resources
  - Different node types, switches, etc.
- Multiple environments in one lease
- Advance reservations (AR)
  - Sharing resources across time
- Upcoming extensions: match making, internal management



- OpenStack Nova/Blazar
- Extensions to support Gantt chart displays and other features

CHI: CONFIGURE AND INTERACT

- Map multiple appliances to a lease
- Allow deep reconfiguration (including BIOS)
- Snapshotting for image sharing
- Efficient appliance deployment
- Handle complex appliances
  - Virtual clusters, cloud installations, etc.
- Interact: reboot, power on/off, access to console
- Shape experimental conditions

- OpenStack Ironic, Glance, and meta-data servers

CHI: MONITORING

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

- OpenStack Ceilometer

CHAMELEON ALLOCATIONS AND POLICIES

- Projects, PIs, and users
- Service Unit (SU) == one hour wall clock on a single server
- Soft allocation model
- Startup allocation: 20,000 SUs for 6 months
  - non-trivial set of experiments
  - 1% of 6 months' tesbed capacity
- Allocations can be extended or recharged

CHAMELEON BEST PRACTICES

- Start small
  - Prototype experiment with a handful of nodes
  - Use short reservation (e.g., for a full work day)
- Snapshot images for fast re-deployment
  - Use scripting, environment customization, and experiment
- Test on incrementally larger scales
- Issues
  - Do not delete instances and reservations
  - Communicate UUIDs & other info to help@

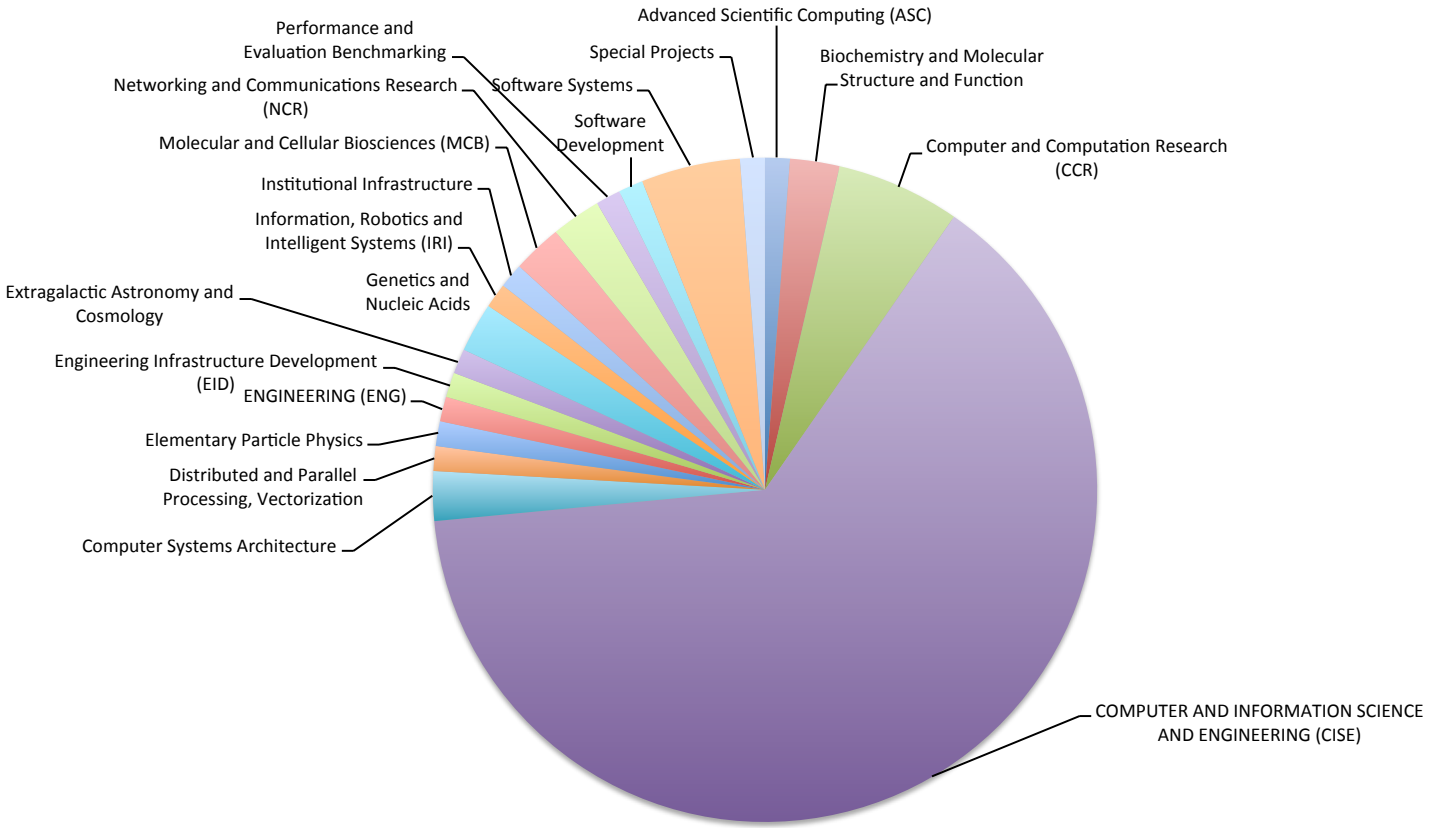
BUILDING CHI: CHAMELEON BARE METAL

- Defining requirements (proposal stage)
- Developing architecture
- Technology Evaluation and Risk Analysis
  - Rough requirements based analysis
  - Technology evaluation: Grid'5000 and OpenStack
  - Implementation proposals
- Implementing CHI
- Technology Preview deployment
- Early User and public availability

CHAMELEON AVAILABILITY TIMELINE

- 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 homogenous hardware
- 07/15: Chameleon Public availability
- 09/15: Chameleon KVM OpenStack cloud available
- 10/15: Global storage available
- 2016: Heterogenous hardware available

CHAMELEON PROJECTS



Overall: 93 projects, 174 users, 59 institutions

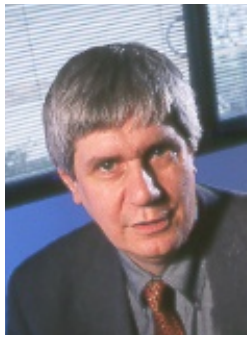
PLANNED CAPABILITIES

- Outreach
- Incremental capabilities
  - Better snapshotting, sharing of appliances, appliance libraries
  - Better isolation and networking capabilities
  - Better infrastructure monitoring (PDUs, etc.)
  - Deeper reconfiguration
- Resource management
  - Rebalancing between KVM & CHI partitions
  - Matchmaking
- Federation activities

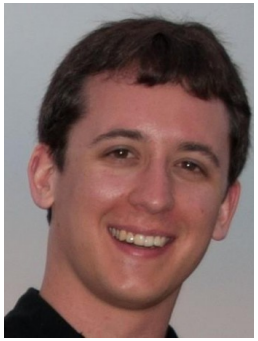
CHAMELEON TEAM

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

PARTING THOUGHTS

- 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

- Building operations for long-term sustainability
- Creating a forum for collaboration between research community and practitioners
  - Workshops, traces, funding opportunities and other forms of engagement