Chameleon

www. chameleoncloud.org

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

Kate Keahey

keahey@anl.gov

ON*VECTOR

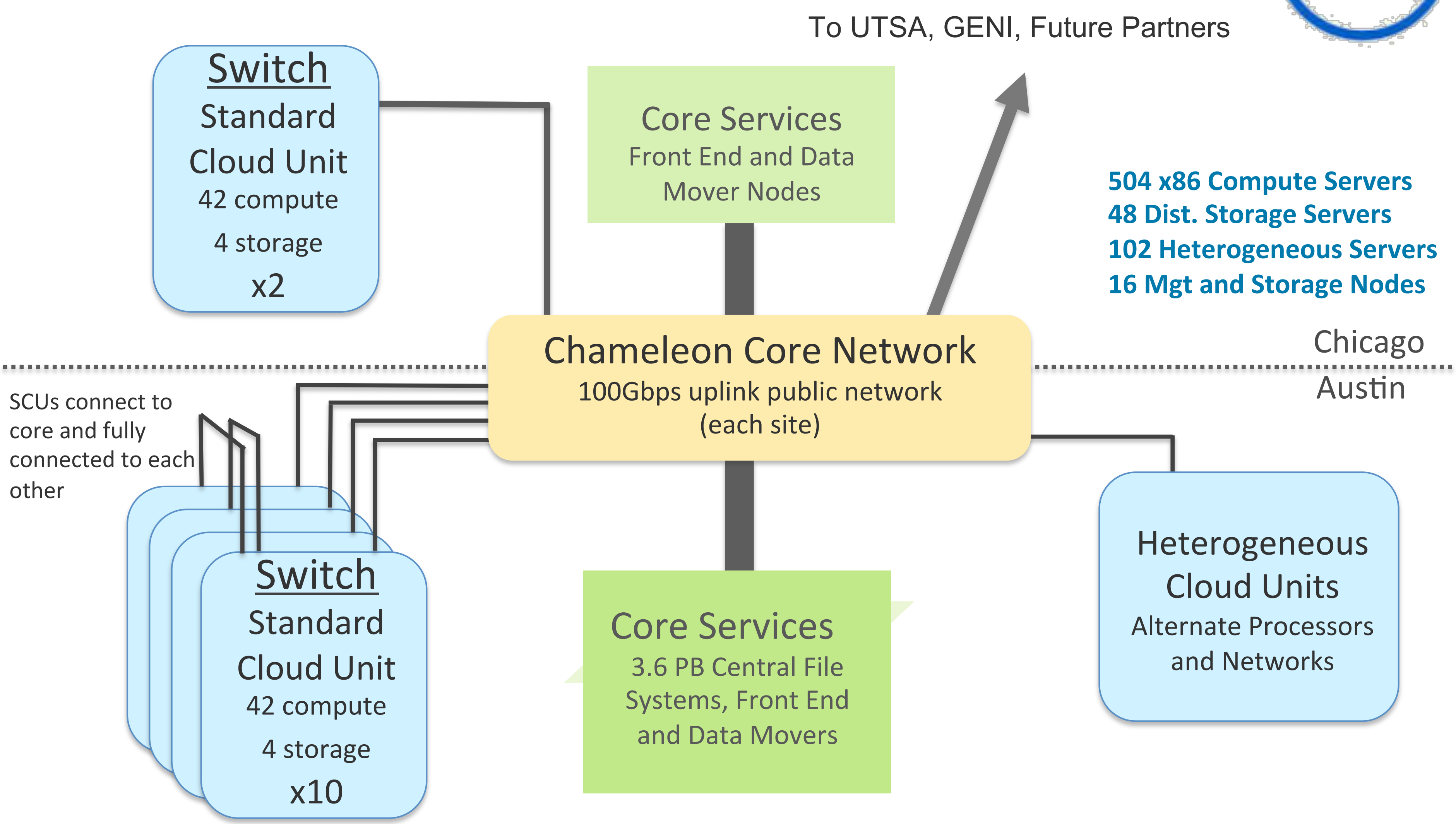
February 29, 2016,
San Diego, CA

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

DESIGN STRATEGY FOR A SCIENTIFIC INSTRUMENT

- 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”
  - Bare metal reconfiguration, operated as a single instrument
  - 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
- Sustainable: “Easy to maintain, easy to share”

CHAMELEON HARDWARE



To UTSA, GENI, Future Partners

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

Switch
Standard Cloud Unit
42 compute
4 storage
x2

Core Services
Front End and Data Mover 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

- Standard Cloud Units (SCU) (deployed)
  - Each of the 12 Standard Cloud Units is a single 48U rack
  - 42 Dell R630 compute servers, each with dual-socket Intel Xeon (Haswell) processors and 128GB of RAM
  - 4 DellFX2 storage servers, each with a connected JBOD of 16 2TB drives (total of 128 TB per SCU)
  - Allocations can be an entire SCU, multiple SCUs, or within a single SCU, or across SCUs (e.g., storage servers for Hadoop configurations)
  - 48 port Force10 s6000 OpenFlow-enabled switches 10Gb to hosts, 40Gb uplinks to Chameleon core network
  - Connectx3 IB network in one rack
- Shared infrastructure (deployed)
  - 3.6 PB global storage, 100Gb Internet connection between sites
- Heterogeneous Cloud Units (to be procured in Y2)
  - ARM microservers, Atom microservers, SSDs, GPUs, FPGAs

CAPABILITIES AND SUPPORTED RESEARCH

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

Persistent, reliable, shared clouds

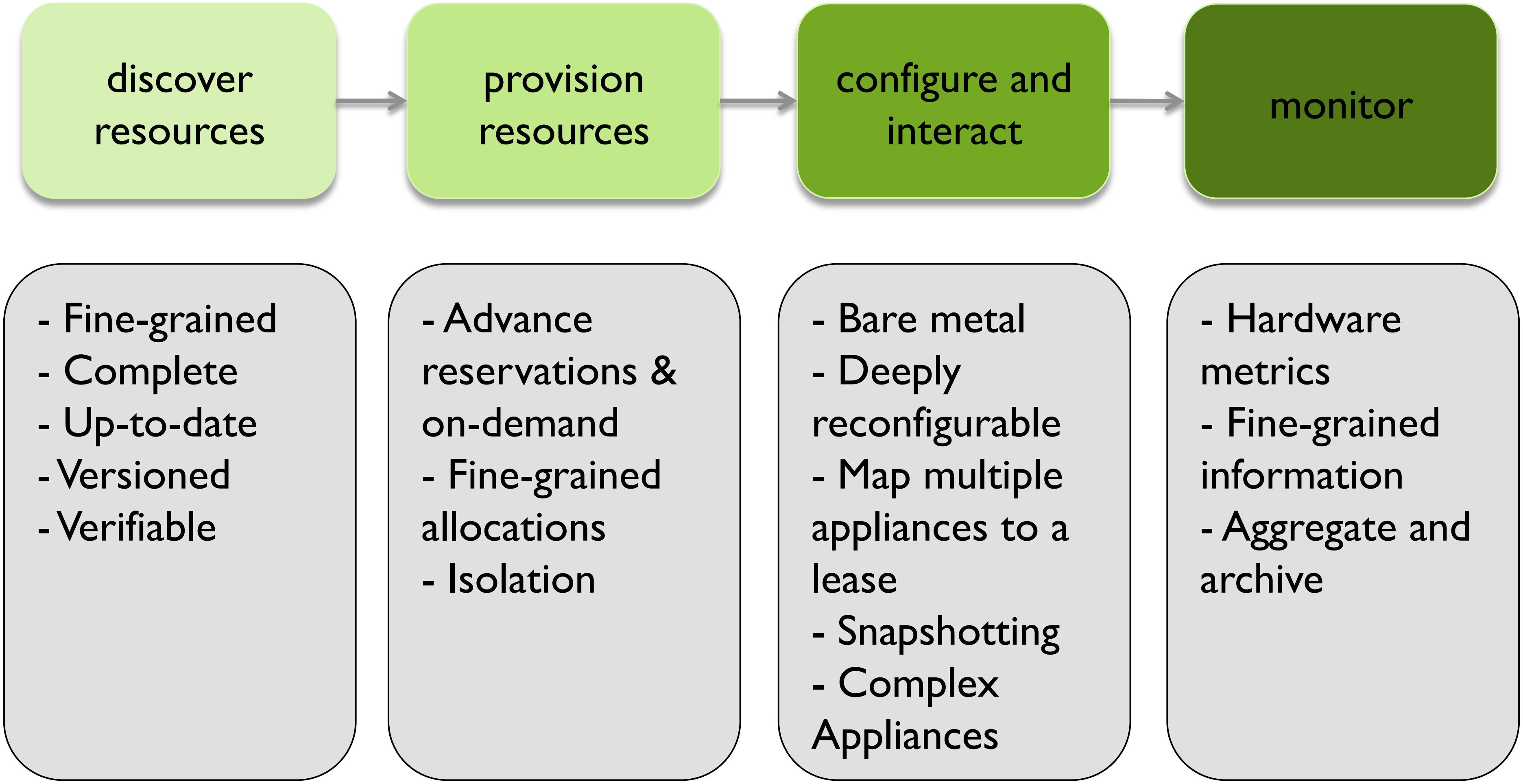
Repeatable experiments in new models, algorithms, platforms, auto-scaling, high-availability, cloud federation, etc.

Isolated partition, Chameleon Appliances

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

Isolated partition, full bare metal reconfiguration

IMPLEMENTING THE EXPERIMENTAL WORKFLOW



discover resources → provision resources → configure and interact → monitor

discover resources

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

provision resources

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

configure and interact

- Bare metal
- Deeply reconfigurable
- Map multiple appliances to a lease
- Snapshotting
- Complex Appliances

monitor

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

BUILDING A TESTBED FROM SCRATH

- 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)
- CHI = OpenStack + G5K + special sauce

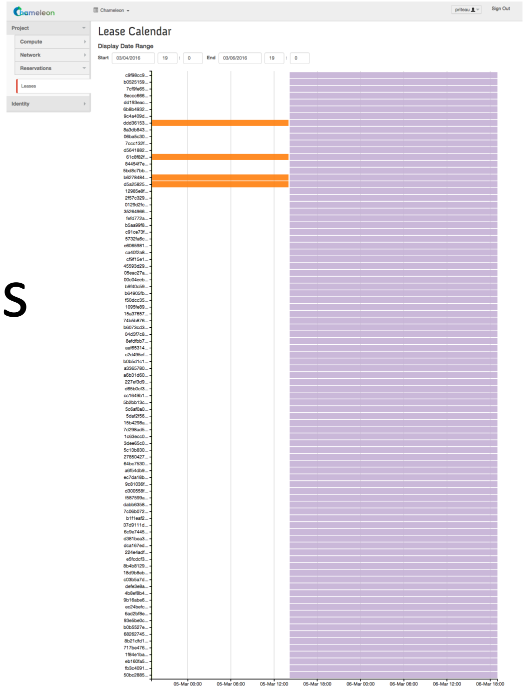
CHI: DISCOVERING AND VERIFYING RESOURCES

- 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
- Fine-grain allocation of a range of resources
  - Different node types, switches, etc.
- Isolation between experiments
- Future extensions: match making, testbed allocation management



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

CHI: CONFIGURE AND INTERACT

- Bare Metal
- Allow deep reconfigurability (access to console)
- Map multiple appliances to a lease
- Snapshotting for image sharing
- Efficient appliance deployment
- Handle complex appliances
  - Virtual clusters, cloud installations, etc.
- Interact: shape experimental conditions

- OpenStack Ironic, Glance, and meta-data servers
- Plus 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

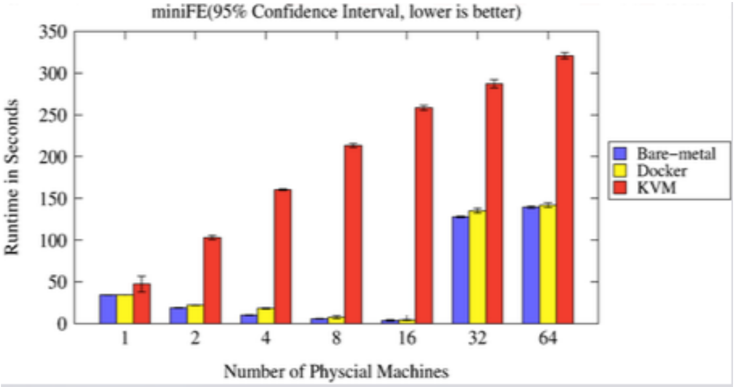
- OpenStack Ceilometer + custom metrics

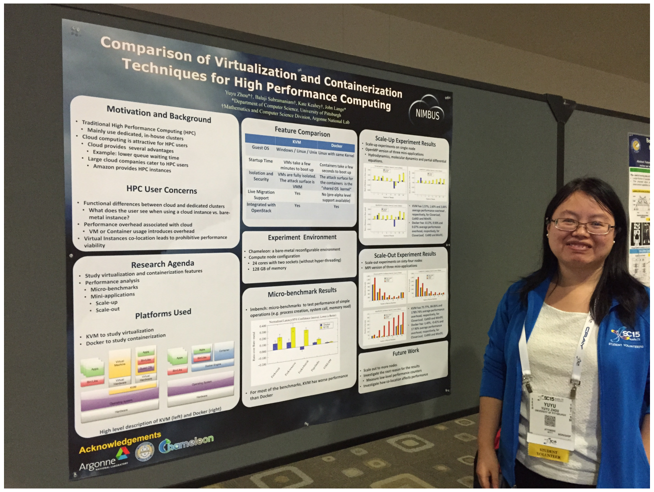
CHAMELEON TIMELINE AND STATUS

- 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: Interoperability with GENI
- Today: 650+ users/160+ projects
- 2016: Heterogeneous hardware available

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

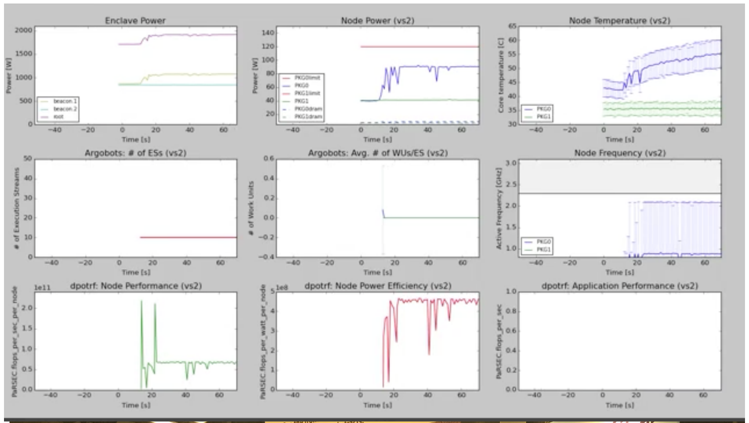



Comparison of Virtualization and Containerization Techniques for High Performance Computing
NIMBUS
Chameleon
Argonne

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.
atípa technologies
DDN
MISSISSIPPI STATE UNIVERSITY
EXASCA...
STARTS HERE

CLASSIFYING CYBERSECURITY ATTACKS

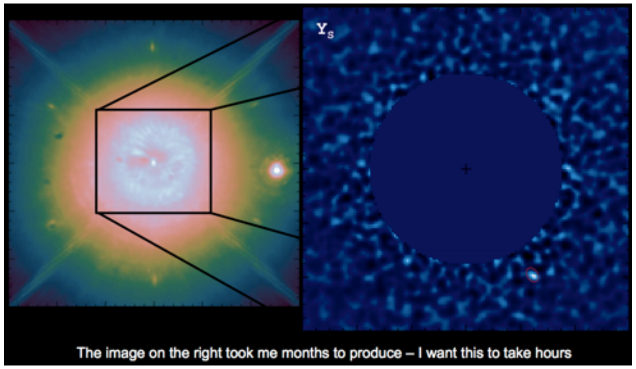
- Jessie Walker & team, University of Arkansas at Pine Bluff (UAPB)
- Research: modeling and visualizing multi-stage intrusion attacks (MAS)
- Testbed requirements:
  - Easy to use OpenStack installation
  - Access to the same infrastructure for multiple collaborators

TEACHING CLOUD COMPUTING

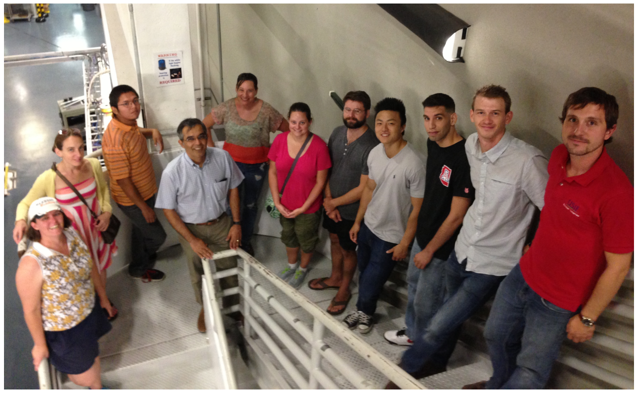
- Nirav Merchant and Eric Lyons, University of Arizona
- ACIC2015: project-based learning course
  - Data mining to find exoplanets
  - Scaled analysis pipeline by Jared Males
  - Develop a VM/workflow management appliance and best practice that can be shared with broader community
- Testbed requirements:
  - Easy to use IaaS/KVM installation
  - Minimal startup time
  - Support distributed workers
  - Block store: make copies of many 100GB datasets

Introduction to Imaging Extrasolar Planets

Jared Males
UA Steward Observatory



The image on the right took me months to produce – I want this to take hours



IN THE PIPELINE…

- Y1 theme was “making things possible”: focus on infrastructure
- Y2 theme is “from possible to easy”: focus on users
- Outreach
- Experiment management
  - Appliances: snapshotting, sharing, appliance marketplace, community
  - Experiment Blueprint: automation and preservation
- Functionality: from possible to easy
  - Better reconfiguration capabilities
  - Better networking capabilities
  - Better infrastructure monitoring (PDUs, etc.)
  - Allocation management
  - And others

PARTING THOUGHTS

- Scientific instrument for CS experimental research
- 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

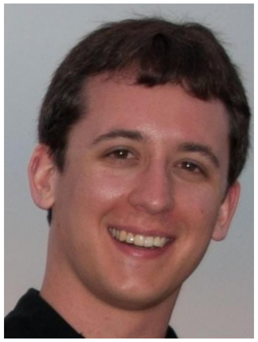
- From vision to reality with Express Delivery
  - Built from scratch within a year on a shoestring
  - Operational testbed: 650+ users/160+ projects
  - Exciting research projects on a range of topics
- Sustainability as a design criterion: building a CS testbed as an application of cloud computing: 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