Chameleon

www. chameleoncloud.org

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

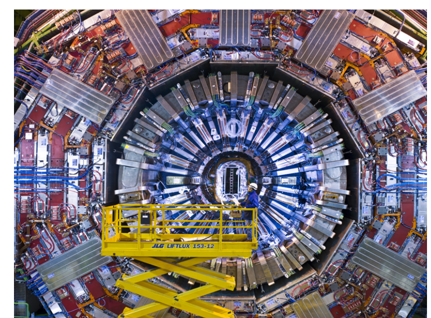
Kate Keahey

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

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

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"

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

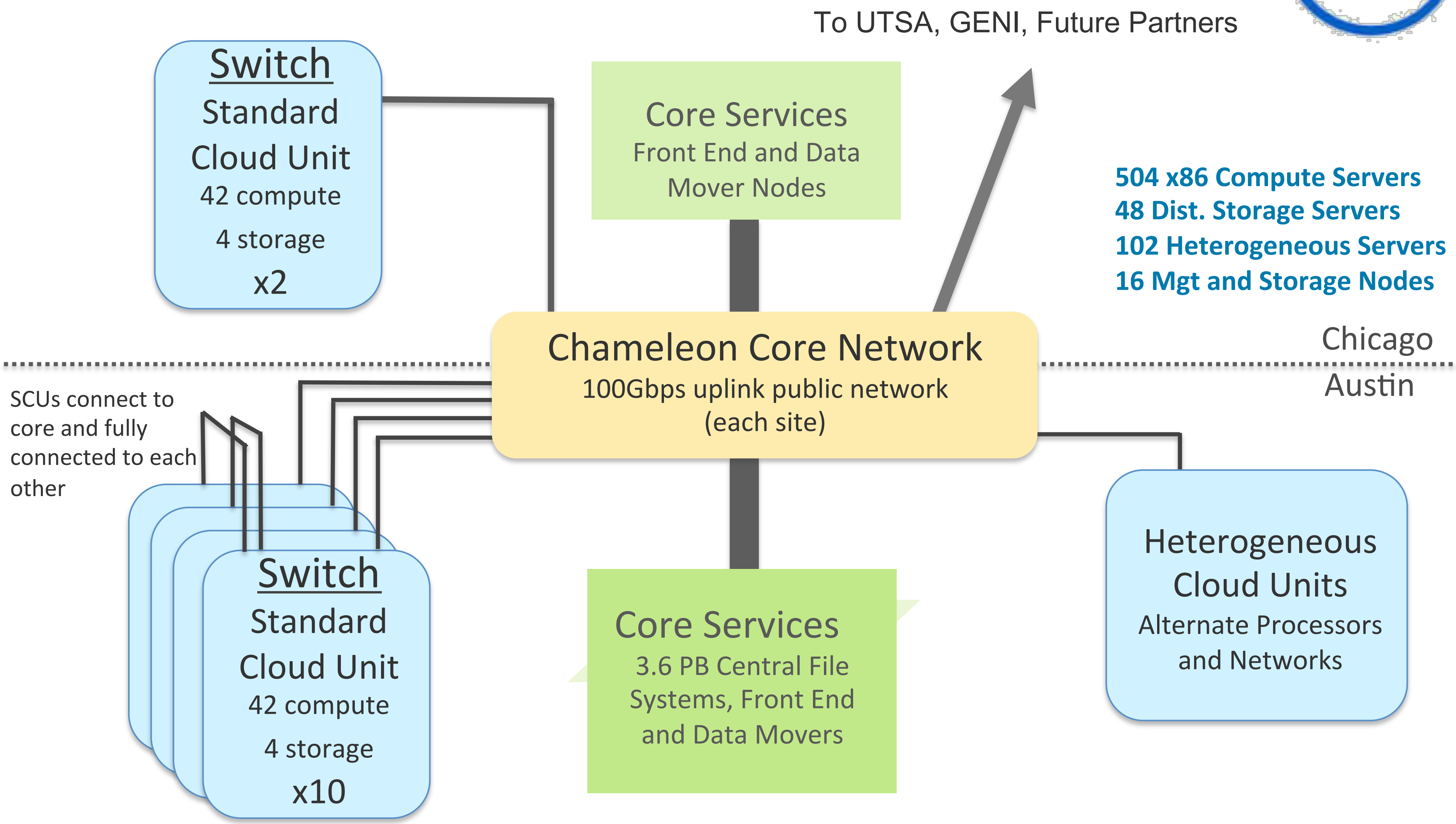
Repeatable experiments in new models, algorithms, platforms, auto-scaling, high-availability, cloud federation, etc.

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

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)

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)
- Today: Chameleon Infrastructure (CHI) =
  - 65%*OpenStack + 10%*G5K + 25%*”special sauce”

WORKING WITH OPENSTACK

- The Good
  - **Leverage community contributions:** whole disk image boot (Liberty), console access, multi-tenant networking, better support for non-x86
  - **Contribute our work:** revival of Blazar project (advance reservations), by collaboration with other organizations (NTT, NEC, HP). Aiming for Blazar to become an official "big tent" OpenStack project.
  - **Basis for operational sustainability**: developing base in scientific institutions (Jetstream, Bridges), having trained staff lowers barriers and costs to adoption
  - **Working with a cloud open source community:** participation in the scientific working group, defining cloud traces, sharing insights, etc.
- The Bad
  - **Complex:** implementing the testbed required high level of skill and persistence – but it can now be packaged for others to use

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

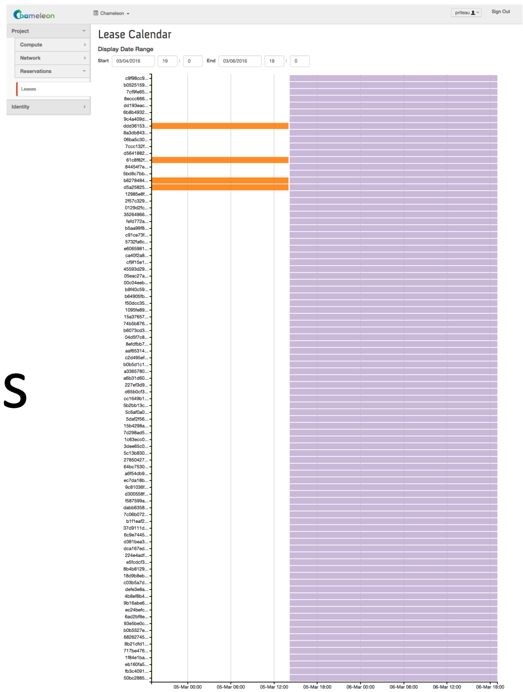
CHI: DISCOVERING AND VERIFYING RESOURCES

- 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

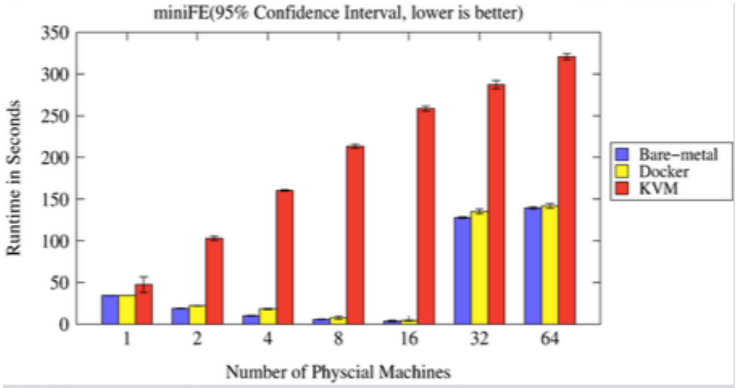
- Chameleon appliance
  - Chameleon bare metal image, same format for UC and TACC
  - Common tools: cc-checks, cc-shapshot, power measurement utility, Ceilometer agent, Heat agent
- System appliances:
  - Base images: CentOS 7, ubuntu (3 versions)
  - Heterogeneous hardware support: CUDA (2 versions),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

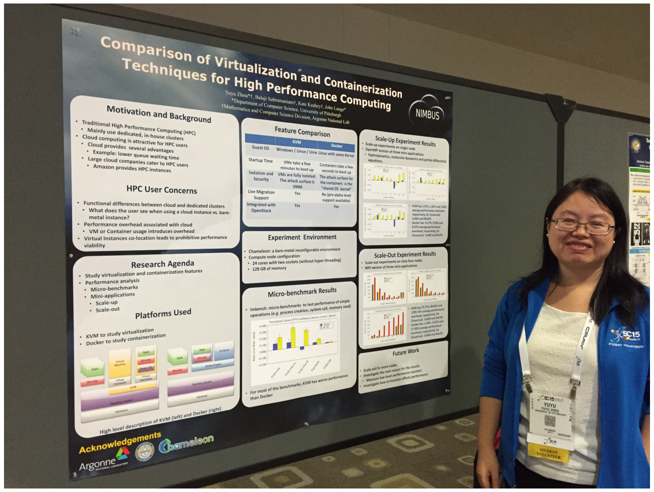
CHAMELEON CORE: TIMELINE AND STATUS

- **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

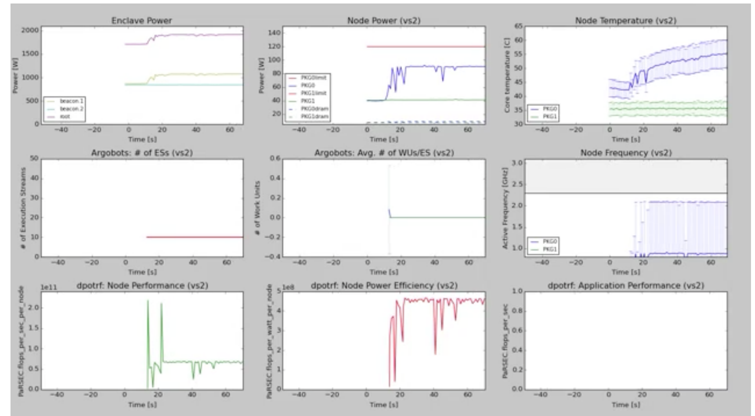




SC15 Poster: "Comparison of Virtualization and Containerization Techniques for HPC"

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"

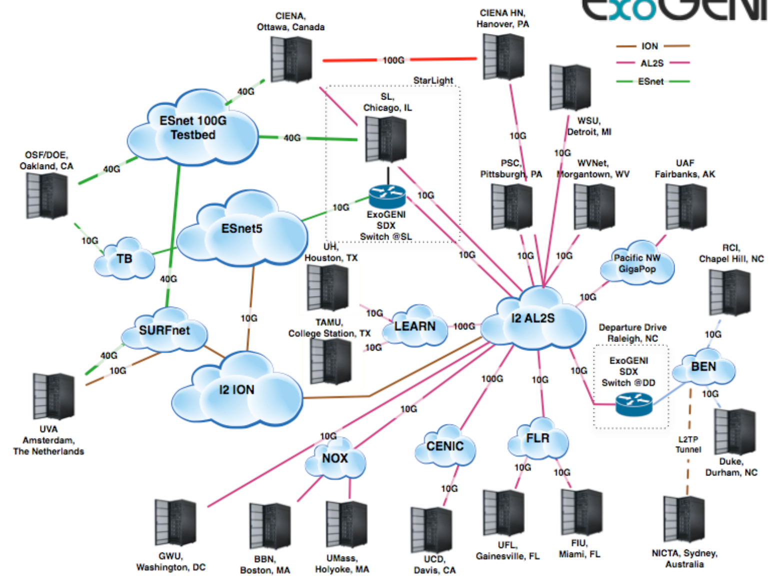
CLASSIFYING CYBERSECURITY ATTACKS

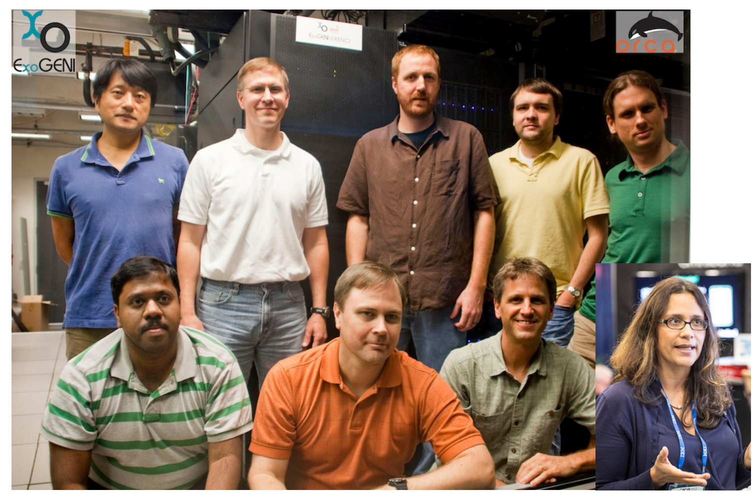
- Jessie Walker & team, University of Arkansas at Pine Bluff (UAPB)
- Research: modeling and visualizing multi-stage intrusion attacks (MAS)
- Testbed requirements:
  - Easy to use OpenStack installation
  - Access to the same infrastructure for multiple collaborators

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)

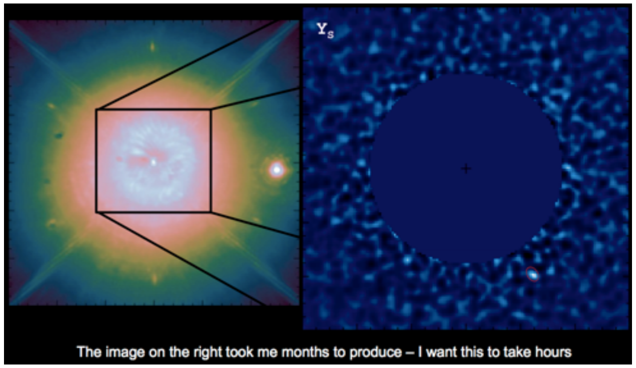
http://www.exogeni.net

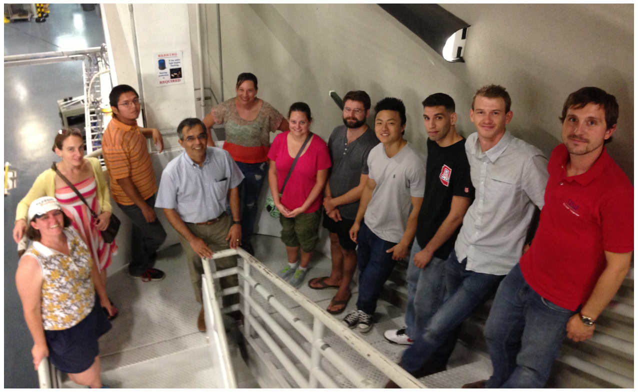




TEACHING CLOUD COMPUTING

- 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

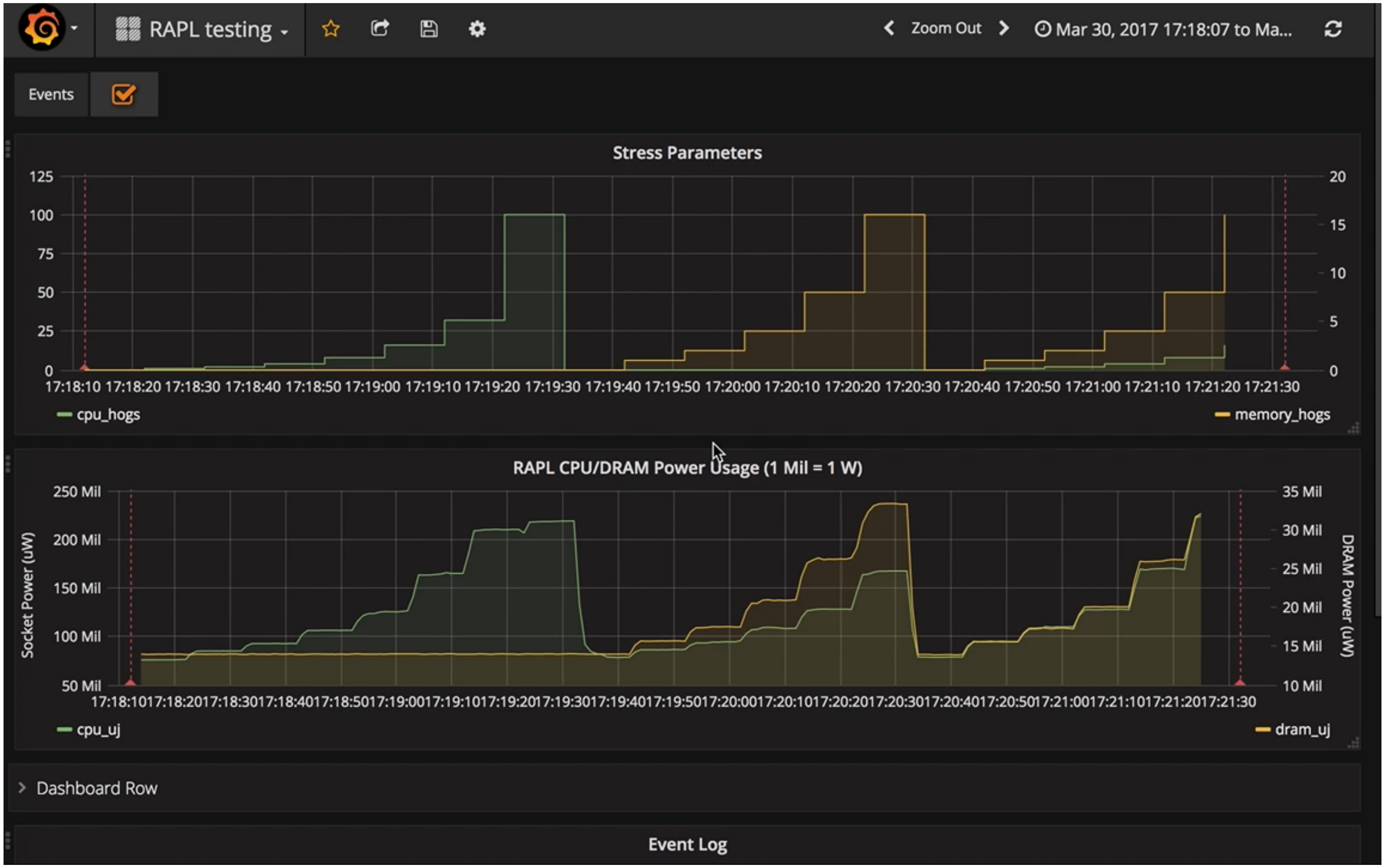


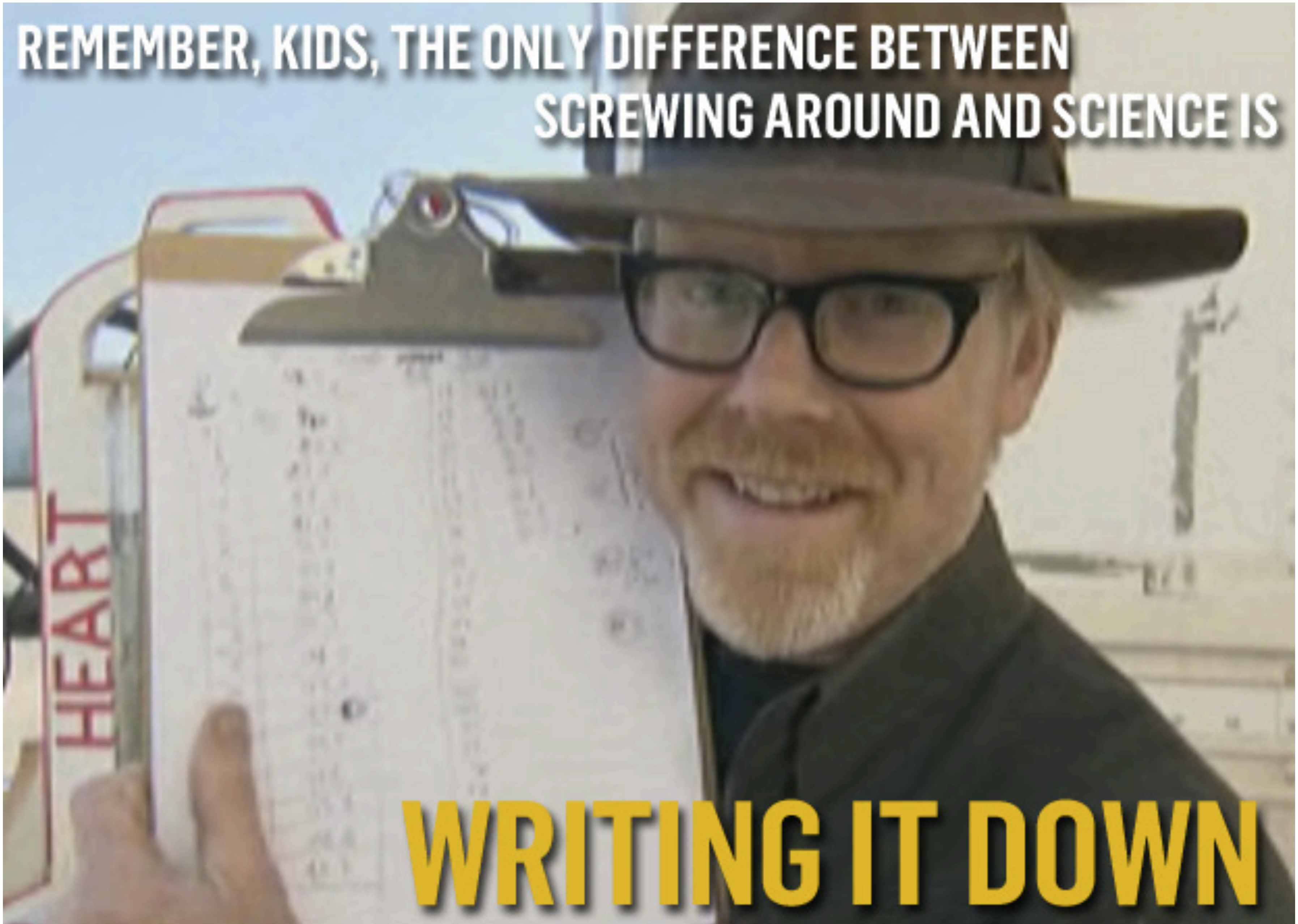


TOWARDS A SCIENTIFIC INSTRUMENT

- 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

FACILITATING UNDERSTANDING




REMEMBER, KIDS, THE ONLY DIFFERENCE BETWEEN
SCREWING AROUND AND SCIENCE IS
WRITING IT DOWN
HEART

COMPLEX INFORMATION MADE SIMPLE

- Testbed description
  - Fine-grained, complete, up-to date, and versioned
  - 53 versions since Chameleon public availability
- Appliance Management
  - Tools for appliance management, versioning, and publication
- Closing the gap: experiment summaries
  - Connections between testbed versions, resources requested, resources allocated, appliances, data, etc.

TOWARDS REPRODUCIBILITY

- The reproducibility trade-off
  - Representing work with complex phenomena requires a huge amount of information
  - Reproducing those complex phenomena is costly
- From experiment summaries to experiment replays
- Publication of experiments, appliances, and data
- Steps towards “publishing the process”
  - Facilities for image generation
- Looking for summer students!

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
- Operational testbedBlueprint for a **sustainable operations model**: building a CS testbed out of commodity components: return on investment, leveraging our investment, and sustainable operation
- Working towards facilitating that help keep track of your work and making it easier to repeat

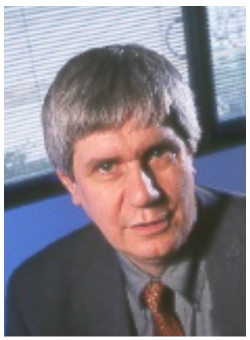
www.chameleoncloud.org

keahey@anl.gov

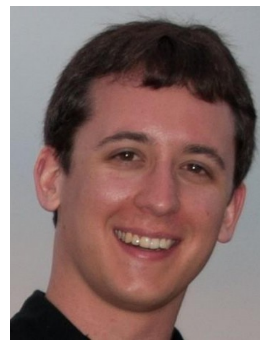
CHAMELEON TEAM

Kate Keahey
Chameleon PI
Science Director
Architect
University of Chicago

Paul Rad
Industry Liaison
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