www. chameleoncloud.org

CHAMELEON: A DEEPLY RECONFIGURABLE, LARGE SCALE INSTRUMENT FOR COMPUTER SCIENCE EXPERIMENTATION

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

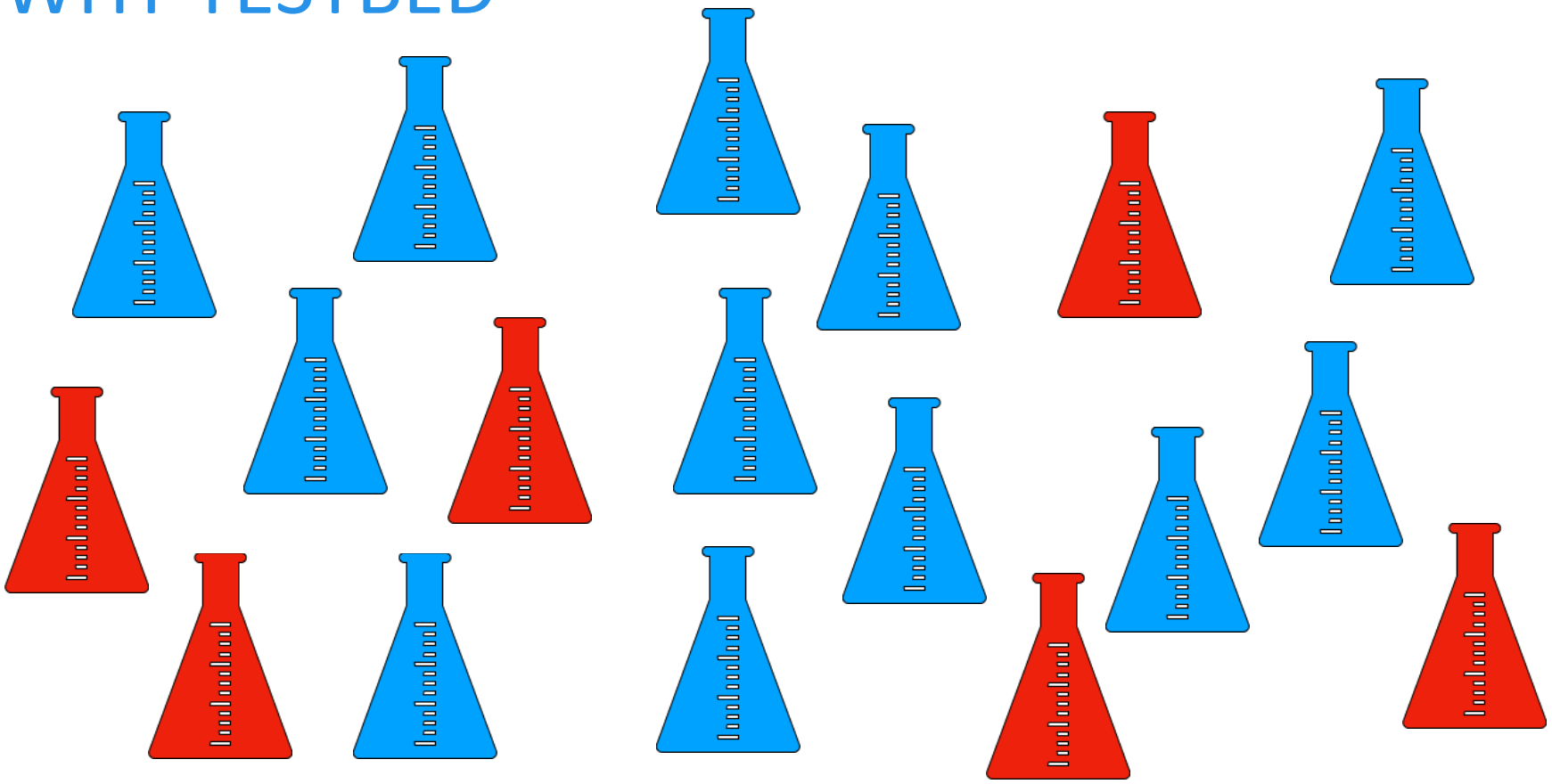





SEARCHING FOR A TESTBED

- The case of no testbed at all
- The case of **outdated**:
  - "no hardware virtualization"
- The case of too **small**:
  - "we think this will scale"
- The case of **shared**:
  - "it may have impacted our result"

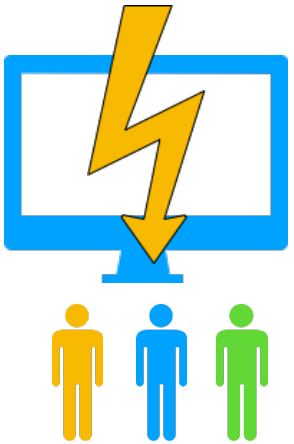
WHY TESTBED



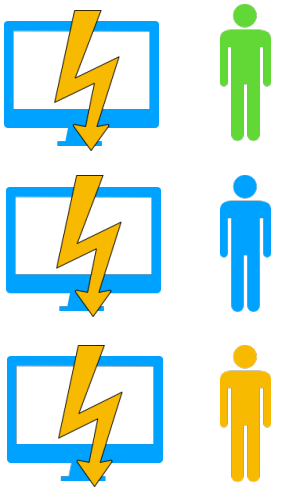
While the types of experiments we can design are only limited by our creativity, in practice we can carry out only those that are supported by an instrument that allows us to deploy, capture (observe and measure), and record relevant scientific phenomena

TOWARDS A PRODUCTION TESTBED

- A testbed is thought of as a place where “things can break”
  - Integration/staging environment used before deployment in production
  - Available to a relatively small group of people and thus not scalable



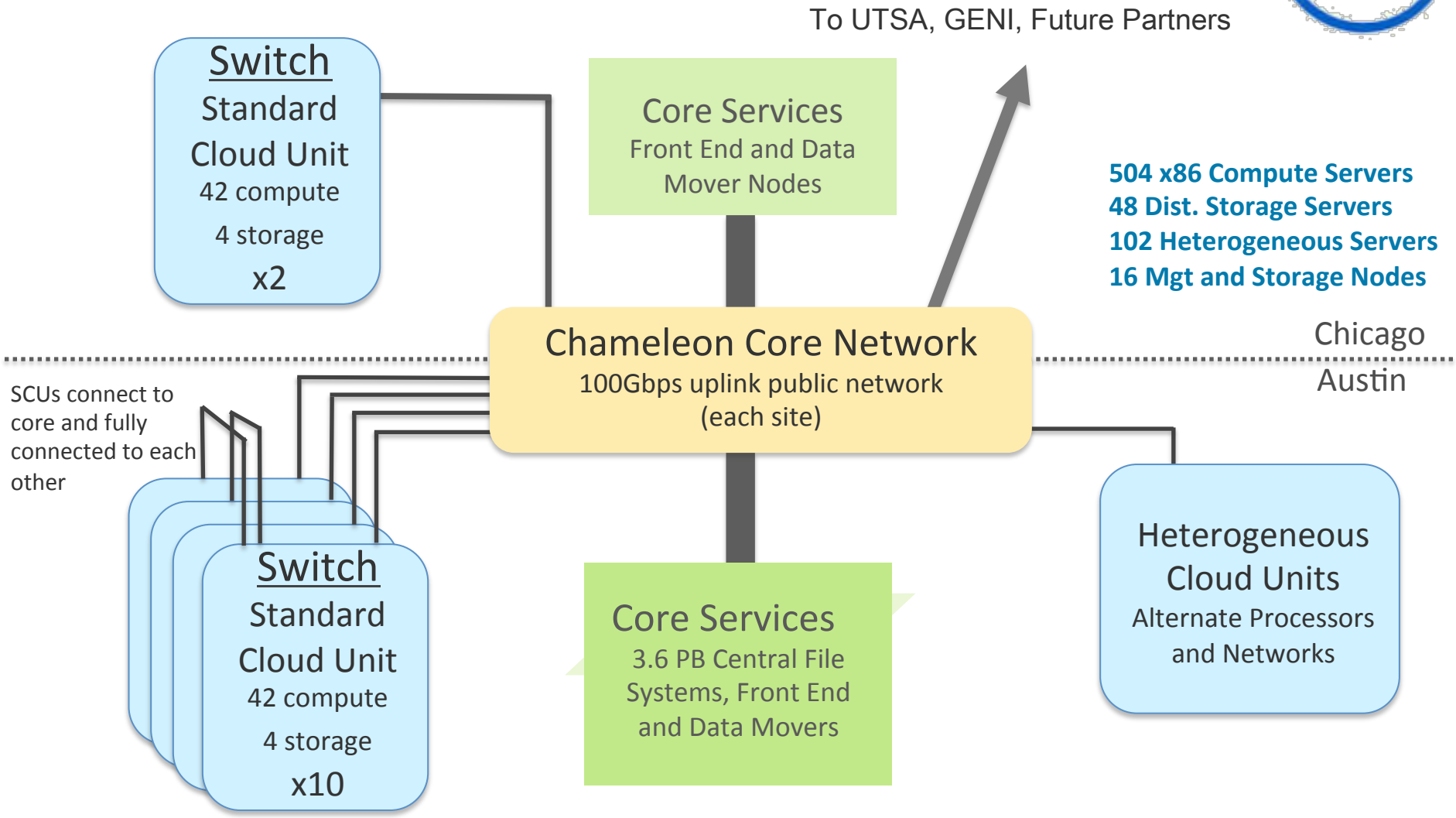
- Production Testbed
  - An oxymoron?
  - Defined by a set of production services that allow you to obtain a personal testbed (“testbed as a service”)



DESIGN STRATEGY FOR A SCIENTIFIC INSTRUMENT

- **Large-scale:** "Big Data, Big Compute research"
  - ~650 nodes (~15,000 cores), 5 PB of storage distributed over 2 sites connected with 100G network – and counting…
  - 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, Appliance Catalog, Instrumentation and repeatability tools
- **Complementary**: "Can't do everything ourselves"
  - Complementing GENI, Grid'5000, and other experimental testbeds
- **Sustainable**: "cost-effective to deploy, operate, and enhance"
- **Open**: "US researchers and collaborators"

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 with SR-IOV support netw in one rack
  - High-memory, NVMe, SSDs, GPUs (18 nodes), FPGAs (4 nodes)
  - ARM microservers (24) and Atom microservers (8), low-power Xeons (8)

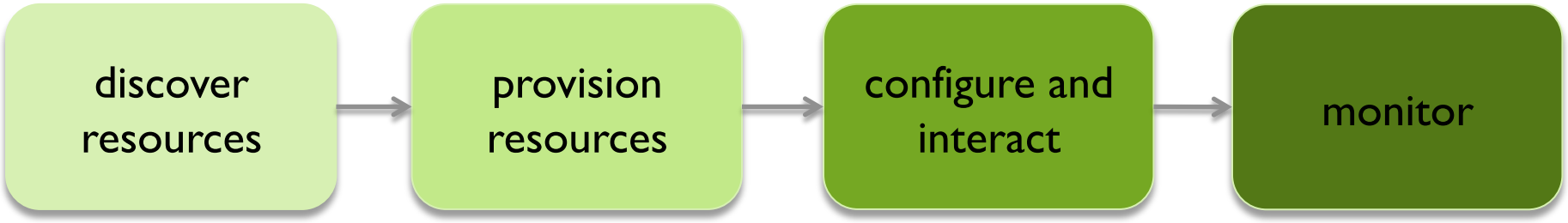
BUILDING A TESTBED FROM SCRATCH

- 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 of core capabilities (~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 (a blueprint for a Production Testbed)

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

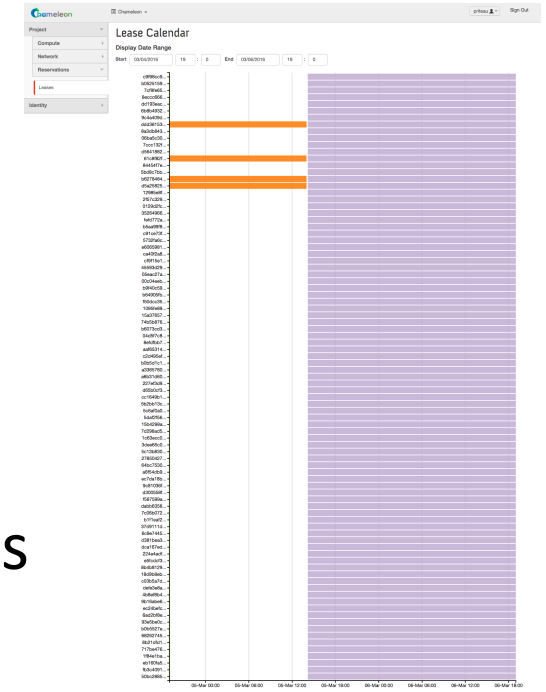
- Hardware metrics
- Fine-grained information
- Aggregate and archive

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.

- Based on OpenStack Nova/Blazar
- Revived Blazar project (ex. Climate), part of core reviewer team
- Extended Horizon panel with calendar displays
- Added Chameleon usage policy enforcement

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

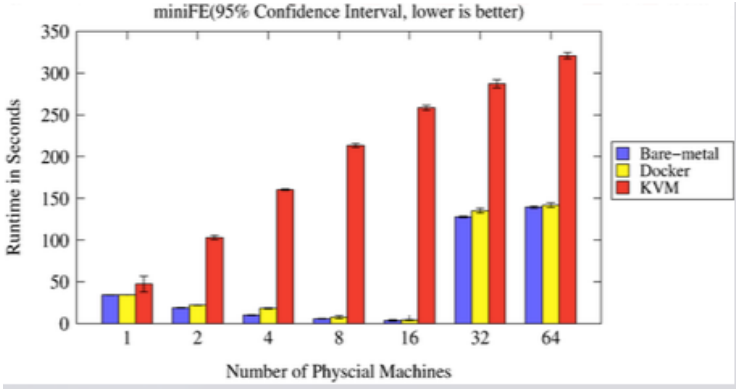
- Chameleon appliance
  - Chameleon bare metal image, same format for UC and TACC
  - Common tools: cc-checks, cc-shapshot, power measurement utility, Ceilometer agent, Heat agent
- System appliances:
  - Base images: CentOS 7, Ubuntu (3 versions)
  - Heterogeneous hardware support: CUDA (2 versions),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

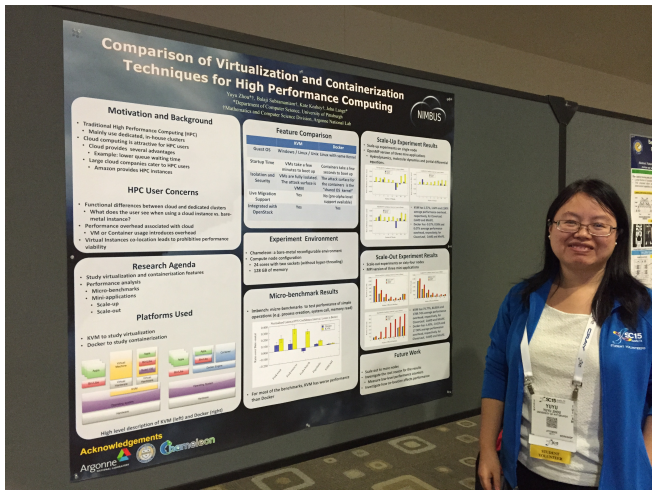
CHAMELEON: TIMELINE AND STATUS

- **10/14: Project starts**
- 04/15: Chameleon Core Technology Preview
- 06/15: Chameleon Early User on new hardware
- **07/15: Chameleon public availability**
- Throughout 2016: New capabilities and new hardware releases
- **Today: 1,400+ users/250+ 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

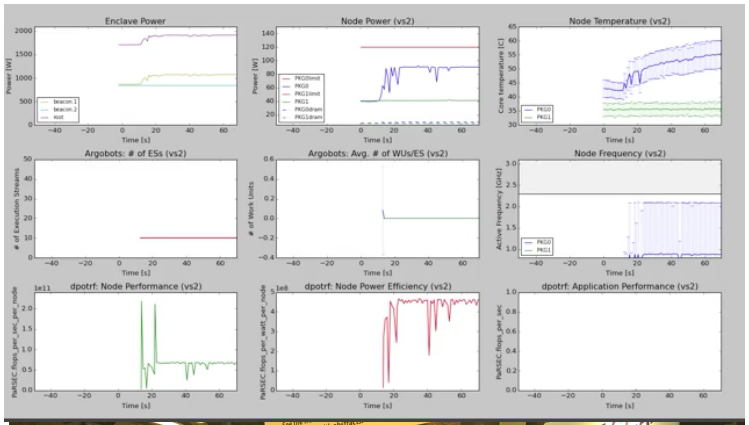




SC15 Poster: "Comparison of Virtualization and Containerization Techniques for HPC"

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"

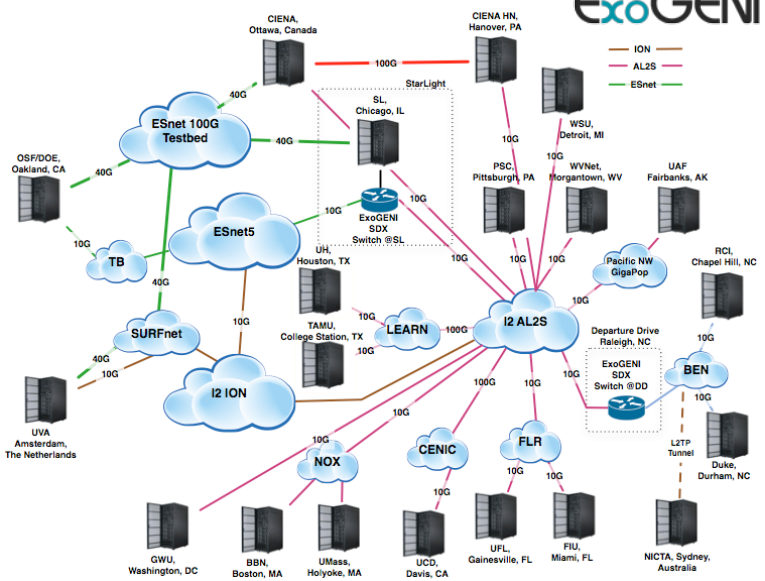
CLASSIFYING CYBERSECURITY ATTACKS

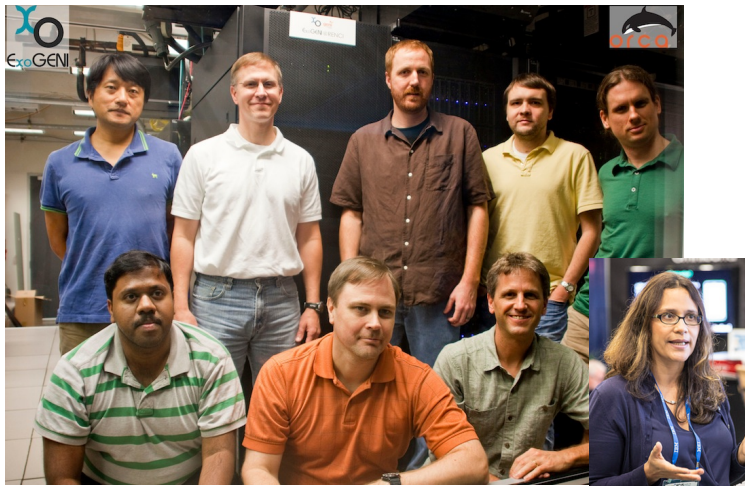
- Jessie Walker & team, University of Arkansas at Pine Bluff (UAPB)
- Research: modeling and visualizing multi-stage intrusion attacks (MAS)
- Testbed requirements:
  - Easy to use OpenStack installation
  - Access to the same infrastructure for multiple collaborators

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)

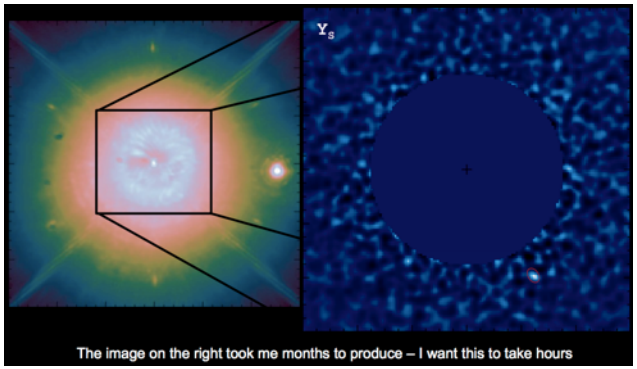
http://www.exogeni.net

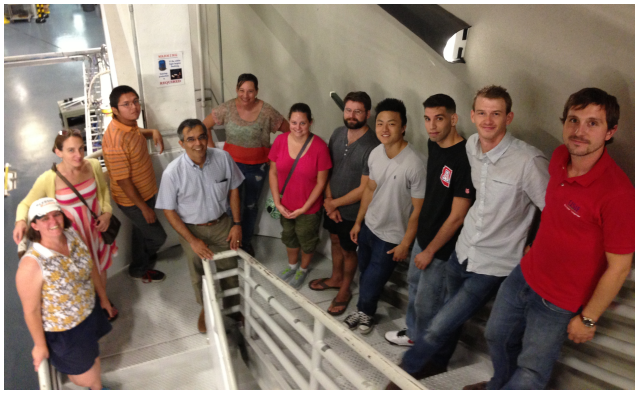




TEACHING CLOUD COMPUTING

- 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

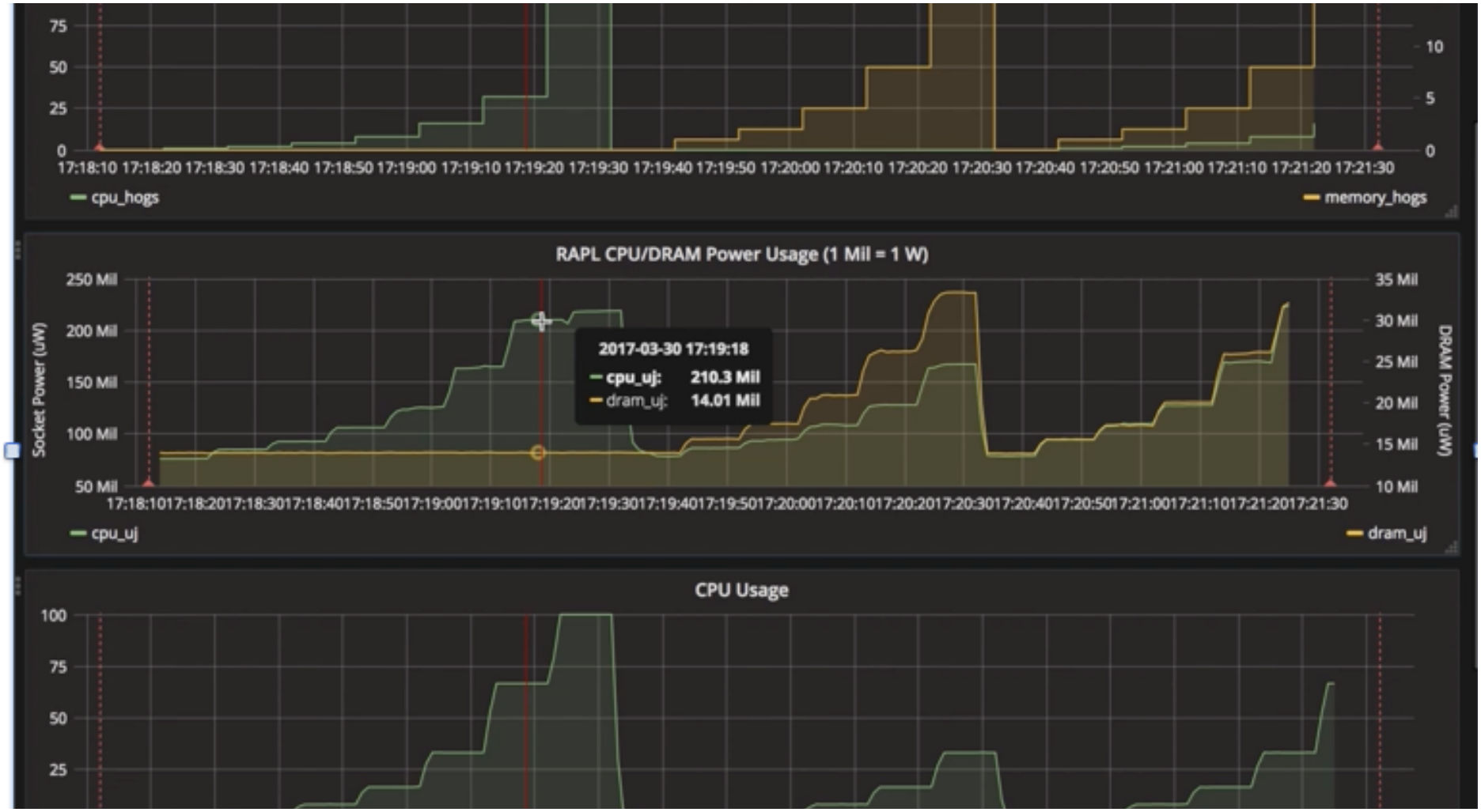




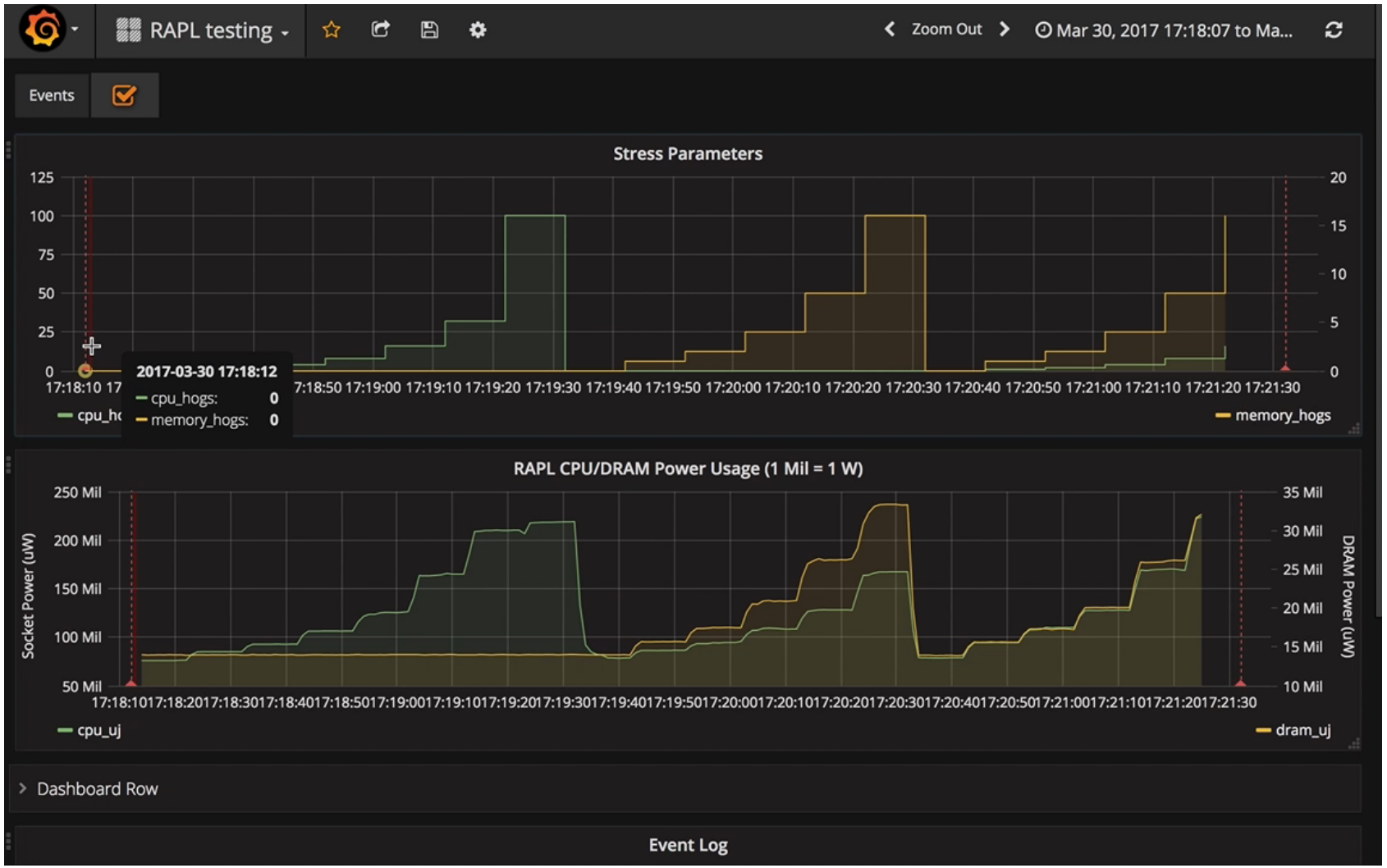
TOWARDS A SCIENTIFIC INSTRUMENT

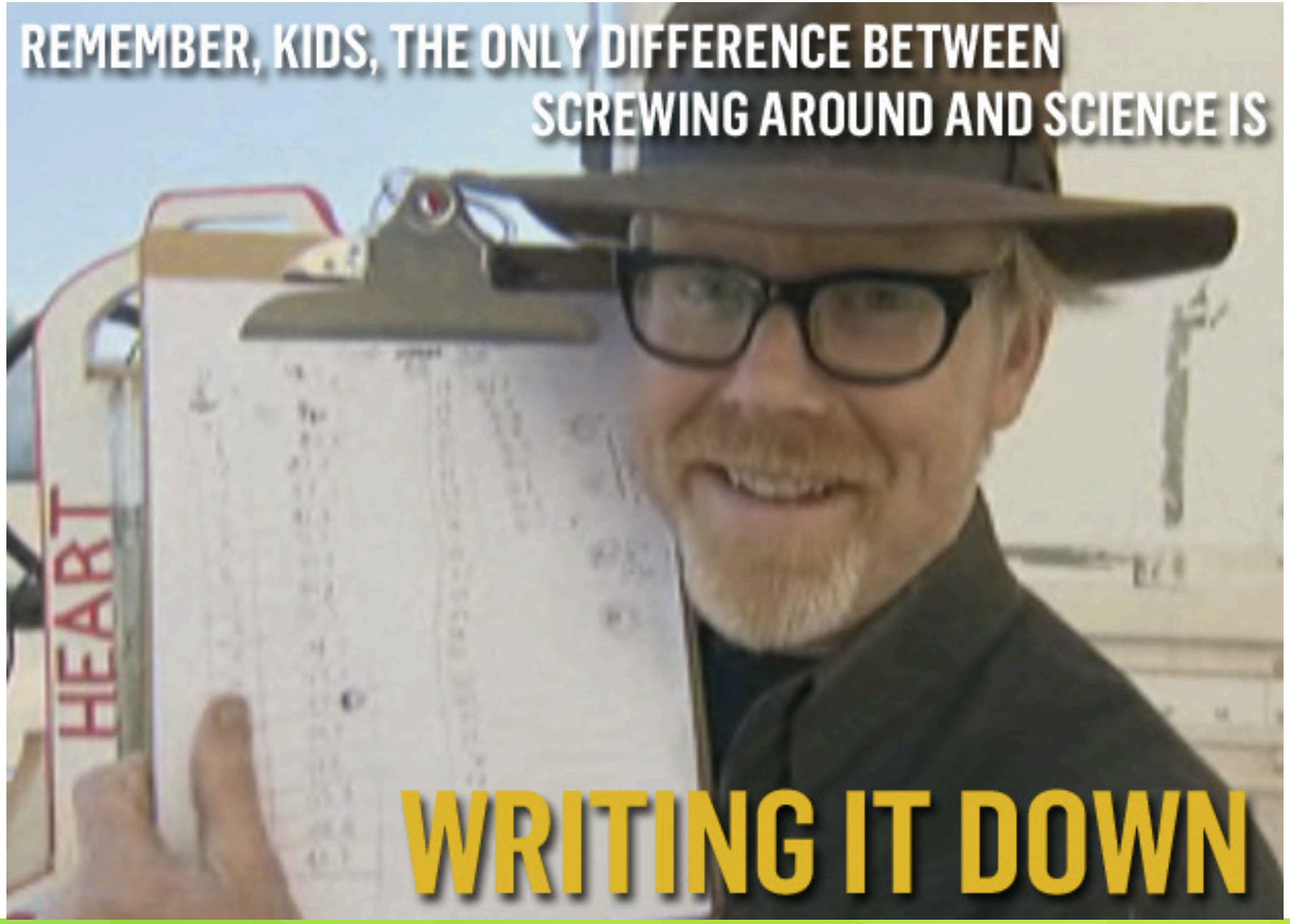
- Scientific instrument: an instrument that allows us to deploy, capture, measure, and record relevant scientific phenomena
- Everything in a testbed is a recorded event
  - The resources you used
  - The appliance/image you deployed
  - The monitoring information your experiment generated
  - Plus any 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

FROM DATA TO INSIGHT



FROM DATA TO INSIGHT




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

FROM INSIGHT TO REPEATABILITY

- Existing elements
  - Testbed versioning (53 versions so far)
  - Appliance publication, versioning, and management
- Experiment summaries: closing the gap between resource versions, appliances, and data
- 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
- Publishing experiment summaries

WHO CAN USE CHAMELEON?

- Any US researcher or collaborator
- Chameleon Projects
  - Created by faculty or staff
  - Who joins the project is at their discretion
  - Allocation of 20K service units(SUs)
  - Easy to extend or recharge
- Key policies
  - Lease limit of 1 week (with exceptions)
  - Advance reservations

DEBUNKING CHAMELEON MYTHS

- "I need to have NSF funding to use Chameleon"
  - **Not true:** Chameleon is an **open** testbed: all PIs with **research** projects will be considered
- "I can't do bare metal on Chameleon, all they do is VMs"
  - **Not true:** almost all of Chameleon support **bare metal** reconfiguration, only a very small partition is configured with KVM
- "I can't provision hundreds of nodes on Chameleon"
  - **Not true:** while at any given time hundreds of nodes may not be available, you can make an advance reservation to get hundreds of nodes in near future

SUMMARY

- **Open** production testbed for **Computer Science research**: 1,400+ users/250+ projects
- Designed from the ground up for a **large-scale** testbed supporting **deep reconfigurability**
- Blueprint for a **sustainable production testbed**: powered by OpenStack
- Working towards a **connected** instrument: from insight to repeatability
- Come to Chameleon User Meeting, Sept 13-14 2017: www.chameleoncloud.org/user-meeting-2017

"We shape our buildings; thereafter they shape us"

Winston Churchill

Help us all dream big:

www.chameleoncloud.org

keahey@anl.gov