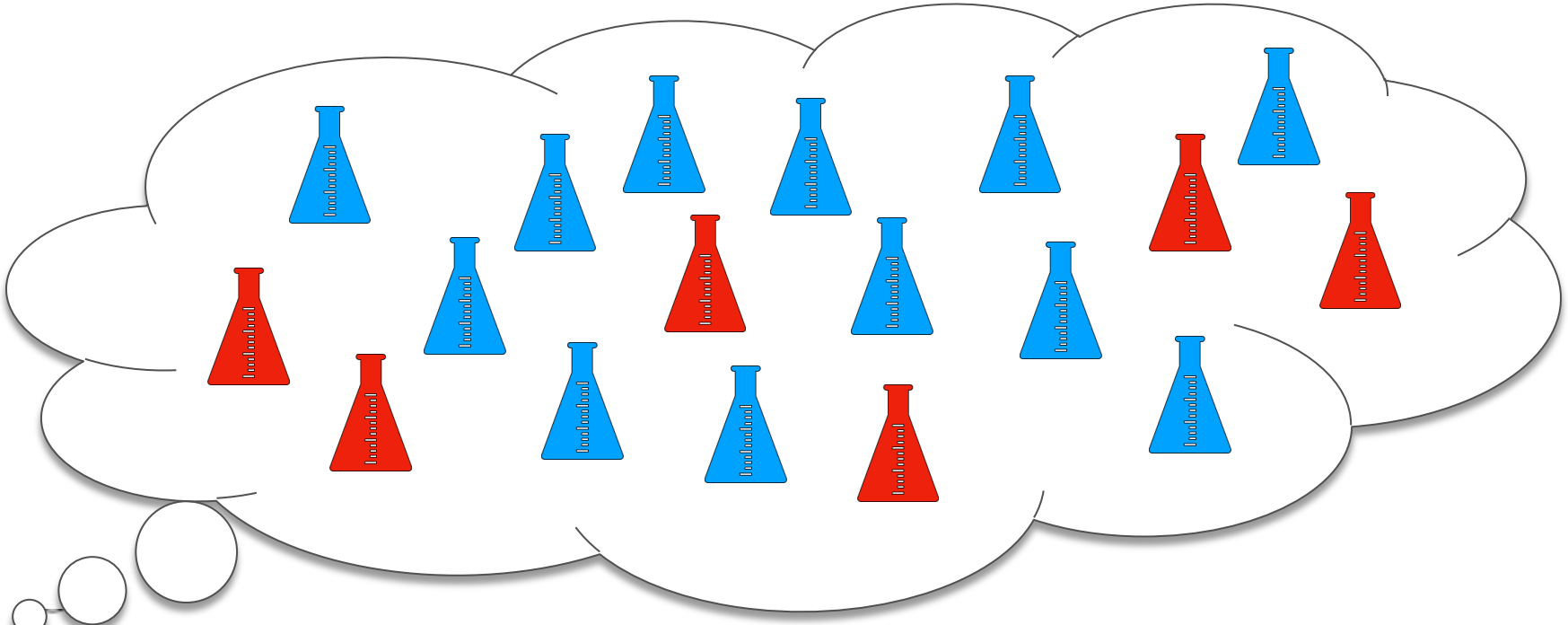
www. chameleoncloud.org

BUILDING A PRODUCTION TESTBED FOR COMPUTER SCIENCE SYSTEMS RESEARCH FROM COMMODITY SOFTWARE

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

April 26, 2018
The Salishan Conference for High Speed Computing

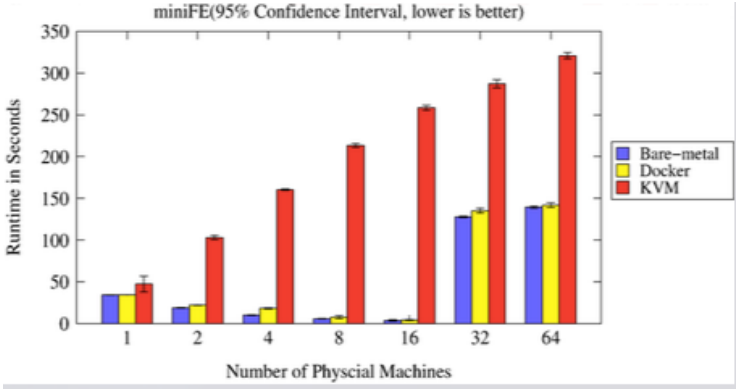
WHY DO WE NEED A TESTBED?

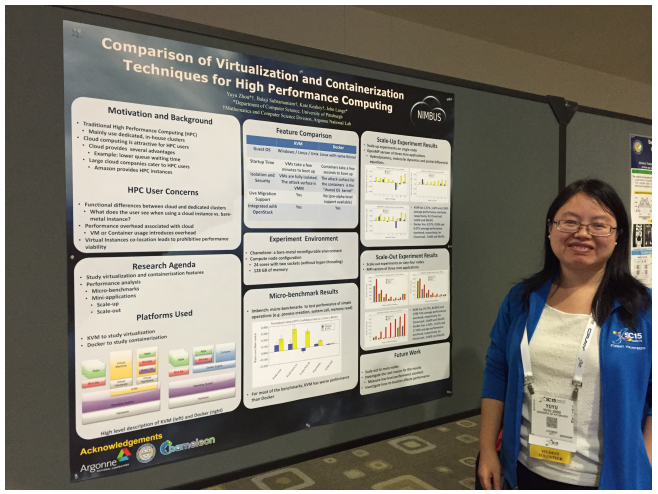


While the types of experiments we can design are only limited by our creativity, in practice we can carry out only those that are supported by an instrument that allows us to deploy, capture (observe and measure), and record relevant scientific phenomena

VIRTUALIZATION OR CONTAINERIZATION?

- Yuyu Zhou, University of Pittsburgh
- Research: lightweight virtualization
- Testbed requirements:
  - Bare metal reconfiguration, isolation, and serial console access
  - The ability to "save your work"
  - Support for large scale experiments
  - Up-to-date hardware

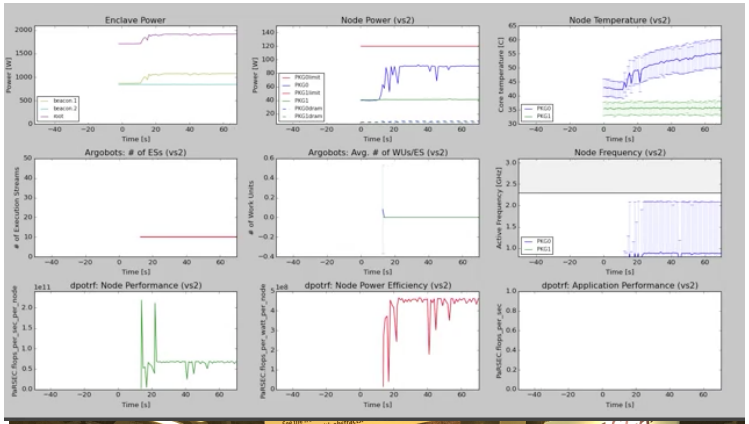




SC15 Poster: "Comparison of Virtualization and Containerization Techniques for HPC"

EXASCALE OPERATING SYSTEMS

- Swann Perarnau, ANL
- Research: exascale operating systems
- Testbed requirements:
  - Bare metal reconfiguration
  - Boot from custom kernel with different kernel parameters
  - Fast reconfiguration, many different images, kernels, params
  - Hardware: accurate information and control over changes, performance counters, many cores
  - Access to same infrastructure for multiple collaborators



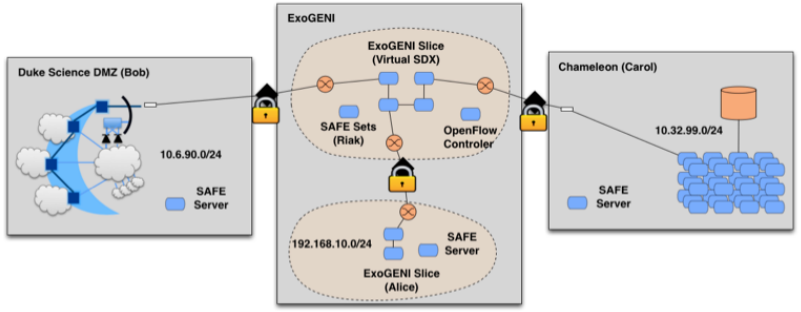
HPPAC'16 paper: "Systemwide Power Management with Argo"

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
  - A selection of pre-configured images
  - Access to the same infrastructure for multiple collaborators

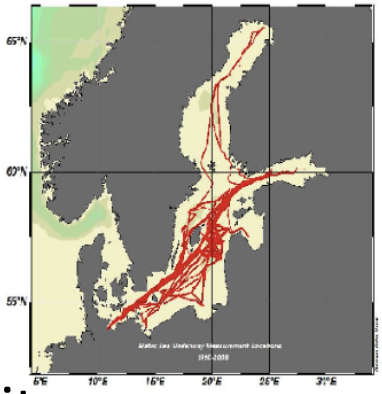
CREATING DYNAMIC SUPERFACILITIES

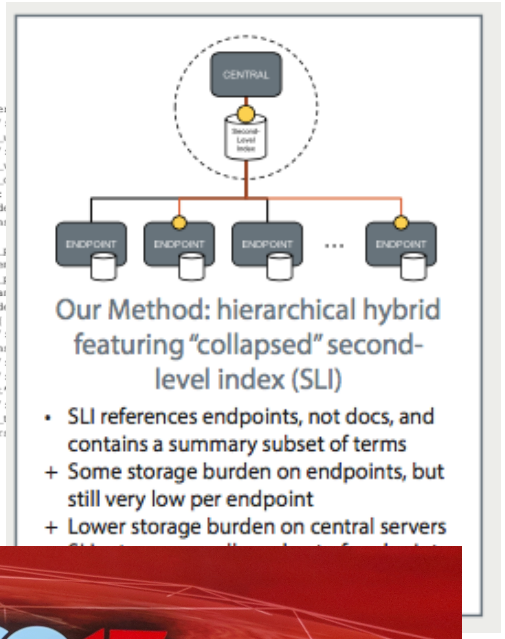
- NSF CICI SAFE, Paul Ruth, RENCI-UNC Chapel Hill
- Creating trusted facilities
  - Automating trusted facility creation
  - Virtual Software Defined Exchange (SDX)
  - Secure Authorization for Federated Environments (SAFE)
- Testbed requirements
  - Creation of dynamic VLANs and wide-area circuits
  - Support for slices and network stitching
  - Managing complex deployments



DATA SCIENCE RESEARCH

- ACM Student Research Competition semi-finalists:
  - Blue Keleher, University of Maryland
  - Emily Herron, Mercer University
- Searching and image extraction in research repositories
- Testbed requirements:
  - Access to distributed storage in various configurations
  - State of the art GPUs
  - Easy to use appliances and complex deployments

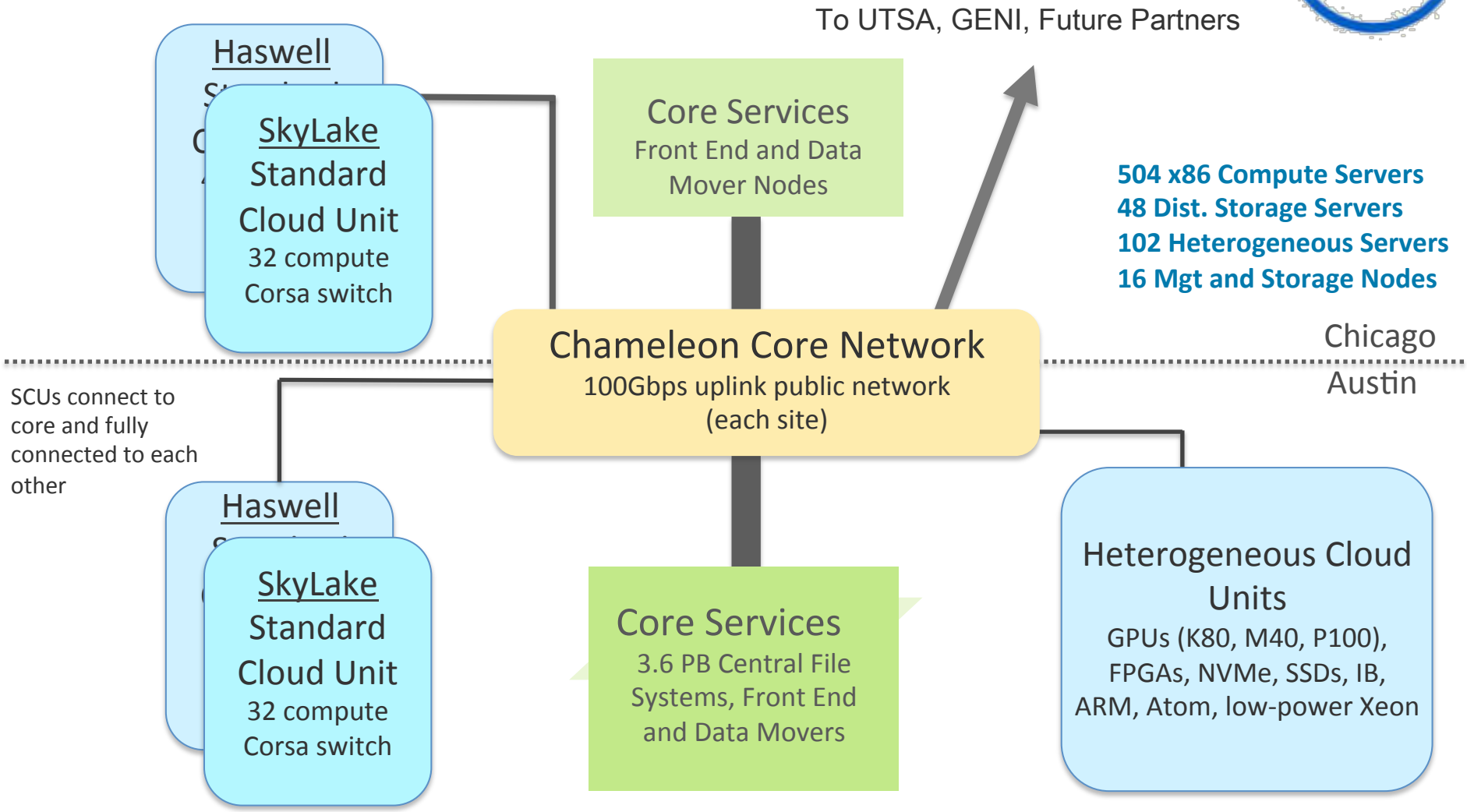




CHAMELEON DESIGN STRATEGY

- **Deeply reconfigurable:** "As close as possible to having it in your lab"
  - Deep reconfigurability (bare metal) and isolation
  - Power on/off, reboot from custom kernel, serial console access, etc.
  - But also – modest KVM cloud for ease of use
- **Combining large-scale and diversity:** "Big Data, Big Compute research"
  - **Large-scale**: ~660 nodes (~15,000 cores), 5 PB of storage distributed over 2 sites connected with 100G network…
  - …and **diverse:** ARMs, Atoms, FPGAs, GPUs, Corsa switches, etc.
  - **Coming soon**: more storage, more accelerators
- Blueprint for a **sustainable** production testbed: "cost-effective to deploy, operate, and enhance"
  - Powered by OpenStack with bare metal reconfiguration (Ironic)
  - Chameleon team contribution recognized as official OpenStack component
- **Open, collaborative** production testbed for **Computer Science Research**
  - Started in 10/2014, testbed available since 07/2015, renewed in 10/2017
  - Currently 2,000+ users, 300+ projects, 100+ institutions, 100+ publications

CHAMELEON HARDWARE



CHAMELEON HARDWARE (DETAILS)

- "Start with large-scale homogenous partition"
  - 12 Haswell Standard Cloud Units (48 node racks), each with 42 Dell R630 compute servers with dual-socket Intel Haswell processors (24 cores) and 128GB RAM and 4 Dell FX2 storage servers with 16 2TB drives each; Force10 s6000 OpenFlow-enabled switches 10Gb to hosts, 40Gb uplinks to Chameleon core network
  - 2 SkyLake Standard Cloud Units (32 node racks); Corsa (DP2400 & DP2200) switches, 100Gb ulpinks to Chameleon core network
  - 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)
- Shared infrastructure
  - 3.6 PB global storage, 100Gb Internet connection between sites
- "Graft on heterogeneous features"
  - Infiniband with SR-IOV support, High-mem, NVMe, SSDs, GPUs (22 nodes), FPGAs (4 nodes)
  - ARM microservers (24) and Atom microservers (8), low-power Xeons (8)
- Coming soon: more nodes, more accelerators, and more storage

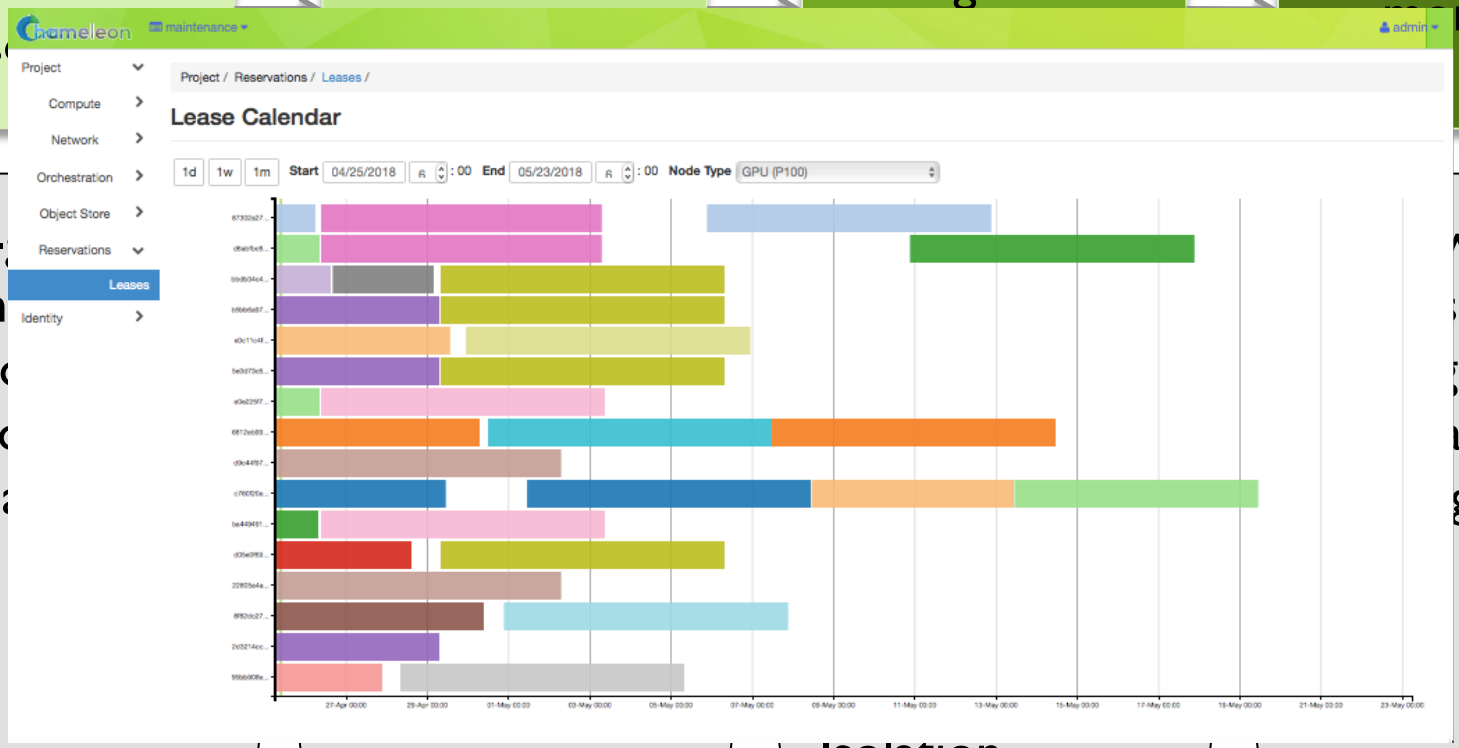
REQUIREMENTS FOR EXPERIMENTAL WORKFLOW

discover

allocate

configure and

nitor



- Fine-
- Com
- Up-t
- Versi
- Verifi

ware

rained

tion

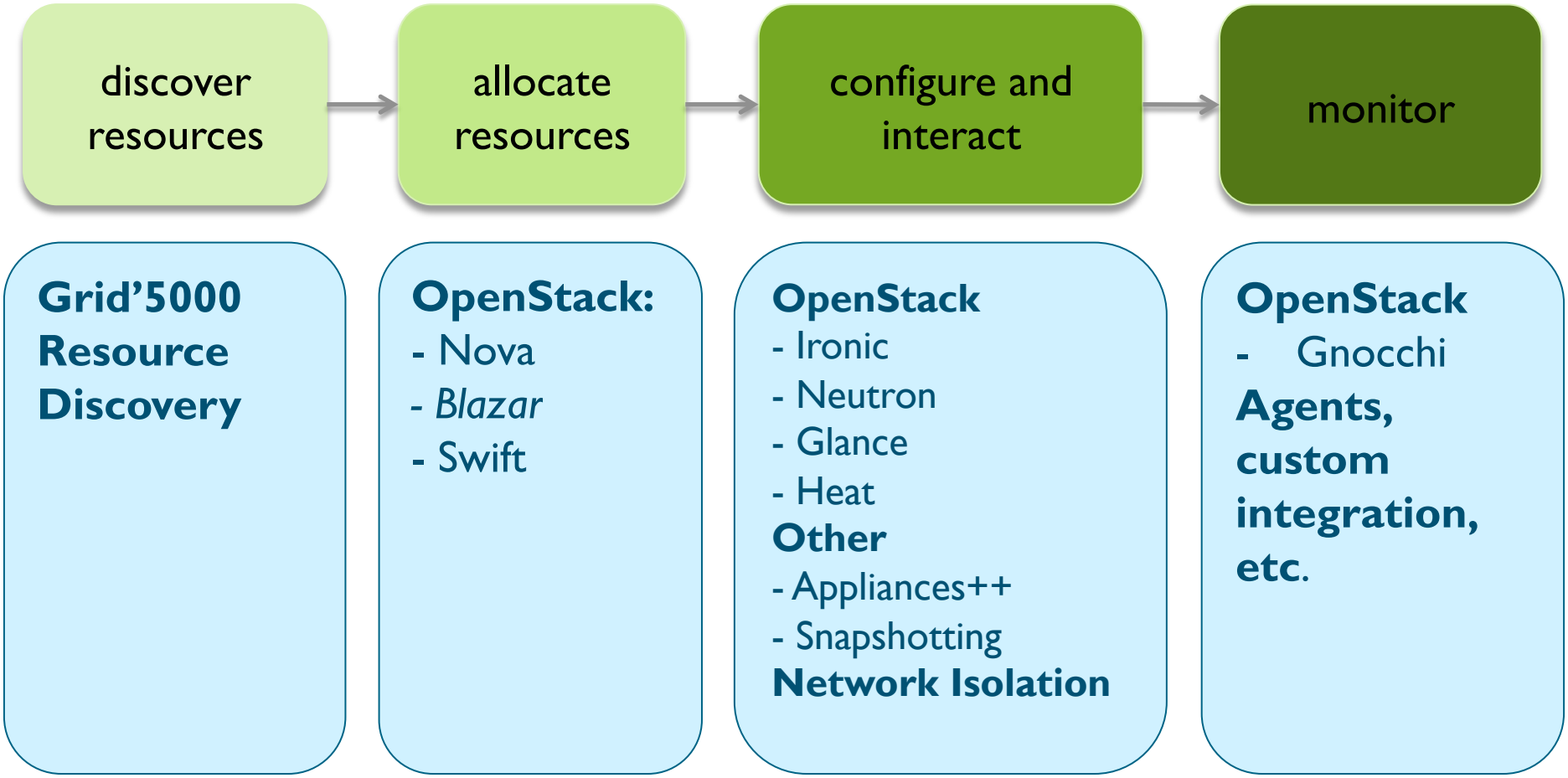
gate and

Isolation

BUILDING CHI (CHAMELEON INFRASTRUCTURE)

- Requirements for core functionality (proposal stage)
  - Interviews with ~20 research groups
- Architecture: **discover**, **provision**, **configure**, and **monitor**
- Technology Evaluation and Risk Analysis
  - Many options: Grid'5000, Nimbus, LosF, OpenStack
  - Final round: Grid'5000 and OpenStack
- Criteria: sustainability as design criterion
  - ***Does it fit our purpose?*** Feature coverage, incl. ease of use
  - ***Can we customize it?*** Open-source, configurable, extendable
  - ***Can we rely on it?*** Stable, scalable, supported
  - Can a CS testbed be built from commodity components?
- A mix of technologies with lots of tweaks (aka "special sauce")
  - Grid'5000 for resource discovery and hardware verification
  - OpenStack for the rest (using Blazar, Ironic, and core OpenStack services)
- Core functionality built in just 3 months after evaluation

SUPPORT FOR EXPERIMENTAL WORKFLOW



CHI = 65%*OpenStack + 10%*G5K + 25%*"special sauce"

WHAT IS OPENSTACK?

- Leading open-source IaaS implementation… and more

Traditional software

OpenStack

- Community: ~ 1,500-2,000 developers contributing to each release including many big companies contributing, e.g. Huawei, Red Hat
- Deployment base:
  - 2017 user surveys logged 1,000 unique deployments (~millions of end users)
  - 60 public cloud data centers, from e.g. Rackspace, OVH
  - Large-scale deployments, e.g. 300K cores at CERN

THE MISSING COMPONENT: OPENSTACK BLAZAR

- **Advanced reservation service** for OpenStack
- Originally developed 2013-2014 in the context of power management research
- From early 2015: adaptation for Chameleon
  - Improve stability, integration with Ironic
  - Dashboard improvements (Gantt chart)
  - Incremental operational improvements
- Fall 2016: revival
  - Joined forces with NTT and others working on capacity reservation for NFV
- **Official OpenStack project** in Sep 2017

BLAZAR
an OpenStack Community Project

OPENSTACK: LESSONS LEARNED

- The good
  - Large community rapidly developing new features
  - Common requirements → shared effort
  - Commodity for sustained use
  - Many users already familiar with OpenStack
- The bad
  - Complexity: need to understand core components
  - Upgrades: rapid development leads to major changes
  - Some users assume Chameleon is like any OpenStack

CHAMELEON PHASE 2: FUTURE DIRECTIONS

- Broaden the set of supported experiments
  - New hardware, new capabilities
- CHI-in-a-box – packaging a CS testbed
- Repeatability and reproducibility
  - **The challenge:** do I invest in making my research reproducible or do I focus on new research?
  - **Experiment précis:** all the information about your experiment in one place
  - Analysis tools: descriptions, visualization, notebooks
  - Active record: Re-examine, share/publish, review, re-play

CHAMELEON 2: NEW HARDWARE

- 4 new Standard Cloud Units (32 node racks in 2U chassis)
  - 3x Intel Xeon "Sky Lake" racks (2x @UC, 1x @TACC) – mostly there!
  - 1x future Intel Xeon rack (@TACC) in Y2
- Corsa DP2000 series switches in Y1
  - 2x DP2400 with 100Gbps uplinks (@UC)
  - 1x DP2200 with 100Gbps uplink (@TACC)
  - Each switch has a 10 Gbps connections to nodes in the SCU
  - Alternative Ethernet connection in both racks
- More storage configurations
  - Global store @UC: 5 servers with 12x10TB disks each
  - Additional storage @TACC: 150 TB of NVMes
- Accelerators: 16 nodes with 2 Volta GPUs (8@UC, 8@TACC)
- Maintenance, support and reserve

CHAMELEON 2 NEW FEATURE HIGHLIGHT: SUPPORT FOR NETWORKING EXPERIMENTS

- Research topics: exploring network programmability, building superfacilities, utilizing high bandwidth
- Building blocks:
  - **Multi-tenant networking** allows users to provision isolated L2 VLANs and manage their own IP address space (since fall 2017)
  - **Stitching** dynamic VLANs from Chameleon to external partners (ExoGENI, ScienceDMZs) (since fall 2017)
  - VLANs + AL2S connection between UC and TACC for **100G experiments** (early user, since Spring 2018)
  - **BYOC– Bring Your Own Controller**: isolated user controlled virtual OpenFlow switches (~Summer 2018)

CHAMELEON 2: CHI-IN-A-BOX

- CHI-in-a-box: packaging a commodity-based testbed
- CHI-in-a-box scenarios
  - **Testbed extension:** join the Chameleon testbed: generalize and package + define operations models
  - **Part-time extension:** define and implement contribution models
  - **New testbed:** generalize policies
- Available Summer 2018

REPEATABILITY: THE FOUNDATION

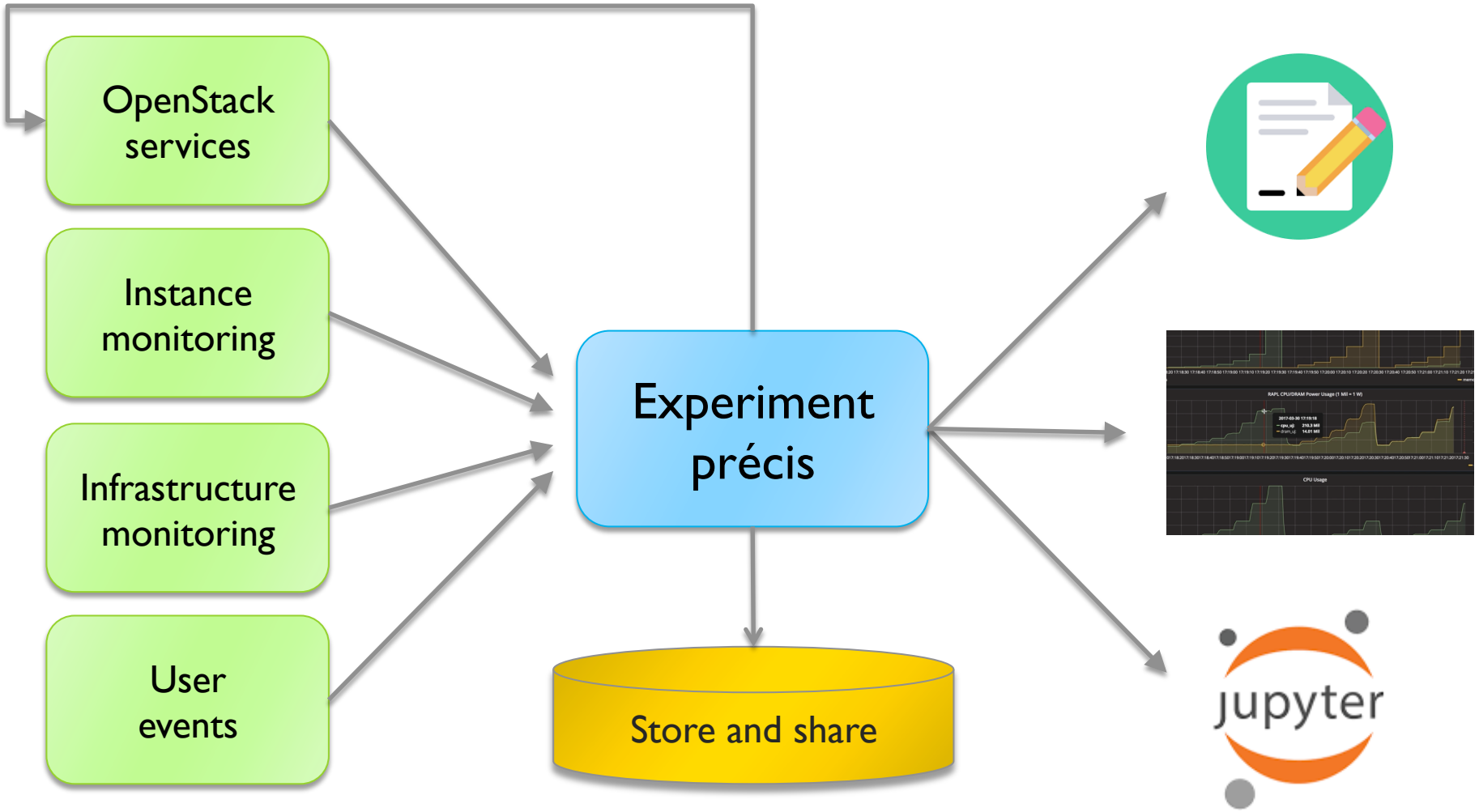
- Testbed versioning
  - Based on representations and tools developed by G5K
  - >50 versions since public availability – and counting
  - Still working on: better firmware version management
- Appliance management
  - Configuration, versioning, publication
  - Still working on: connection between the catalog and glance
- Monitoring and logging
- However… the user still has to keep track of this information

REPEATABILITY: KEEPING TRACK OF EXPERIMENTS

- 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., "start power_exp_23" and "stop power_exp_23"

- **Experiment précis:** information about your experiment made available in a "consumable" form

REPEATABILITY: EXPERIMENT PRÉCIS



PARTING THOUGHTS

- A testbed for Computer Science research
  - **Open, collaborative** production testbed for **Computer Science research**: 2,000+ users/300+ projects
  - Designed from the ground up for a **large-scale** testbed supporting **deep reconfigurability**
  - Moving up the stack: making reproducibility cost effective
  - Blueprint for a **sustainable production testbed**
- Are we there yet?
  - The research frontier does not stay put and will drive the design of scientific instruments that support it …
  - …and they, in turn, will define areas of feasible exploration
- Join us @www.chameleoncloud.org

Questions?

www.chameleoncloud.org

keahey@anl.gov