Apache airavata

Science Gateway Architectures: Open Source, Distributed Computing

Marlon Pierce, Suresh Marru
Science Gateway Group, Indiana University
marpierc@iu.edu, smarru@iu.edu

SciGaP

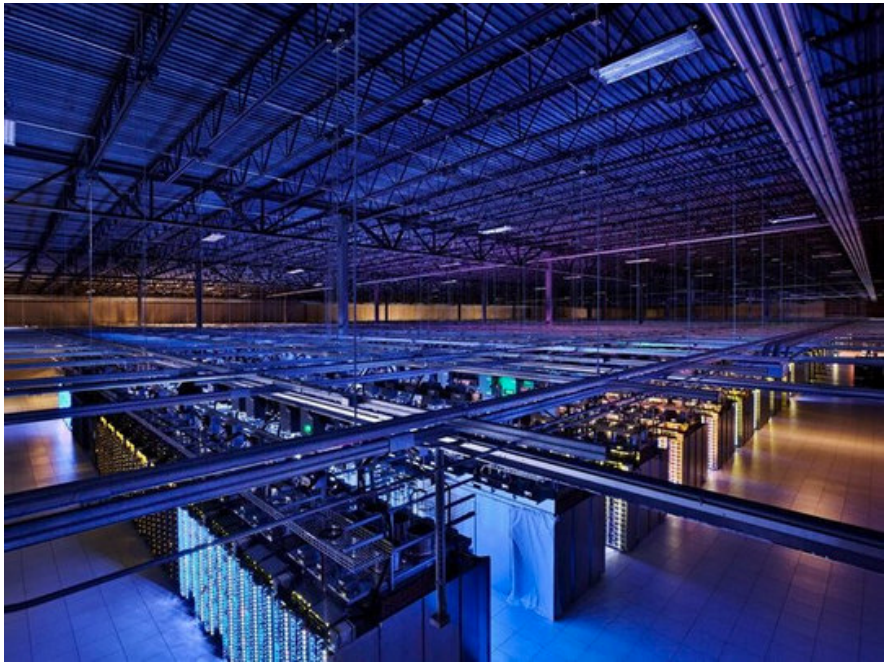
I590 Special Topics Course: Spring 2016

http://homes.soic.indiana.edu/classes/spring2016/info/i590-marpierc

Or

http://s.apache.org/i590-spring2016

SciGaP



Goal 1: teach basic distributed computing concepts by applying them to science gateways.

SciGaP

You Don’t Choose Chaos Monkey…
Chaos Monkey Chooses You

Goal 2: explore new architectures, methodologies, and technologies: Microservices, DevOps

SciGaP

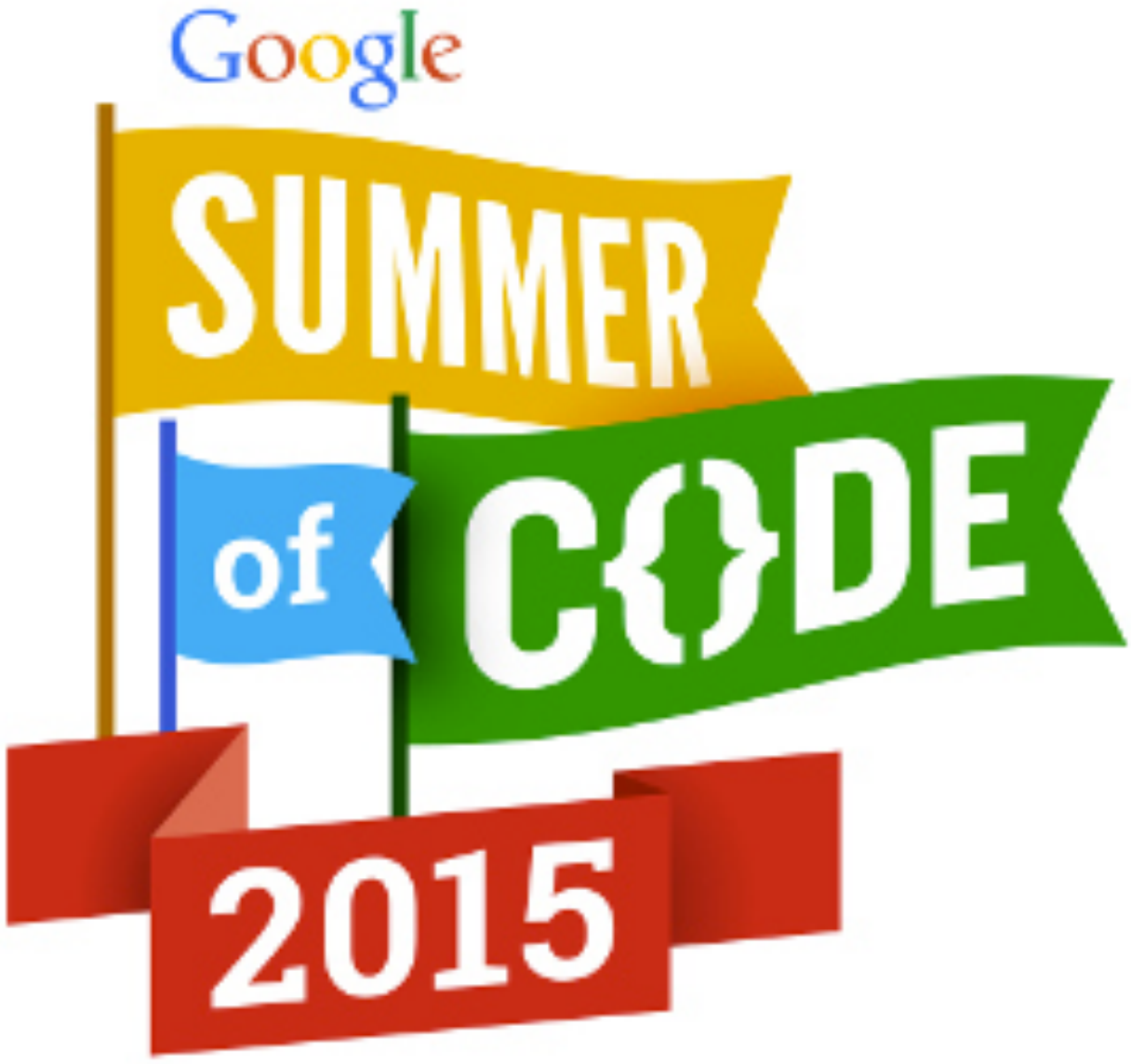
Goal 3: teach open source software practices

SciGaP

Structure of the Course

- Students will be divided into development teams
- Each team will develop an open source science gateway system from scratch.
  - Biweekly assignments
  - Grades entirely from projects
- Each assignment will illustrate one or more distributed computing topics.
- Students will learn how to operate services as well as write good software.
- Students will learn open source methodologies.

SciGaP



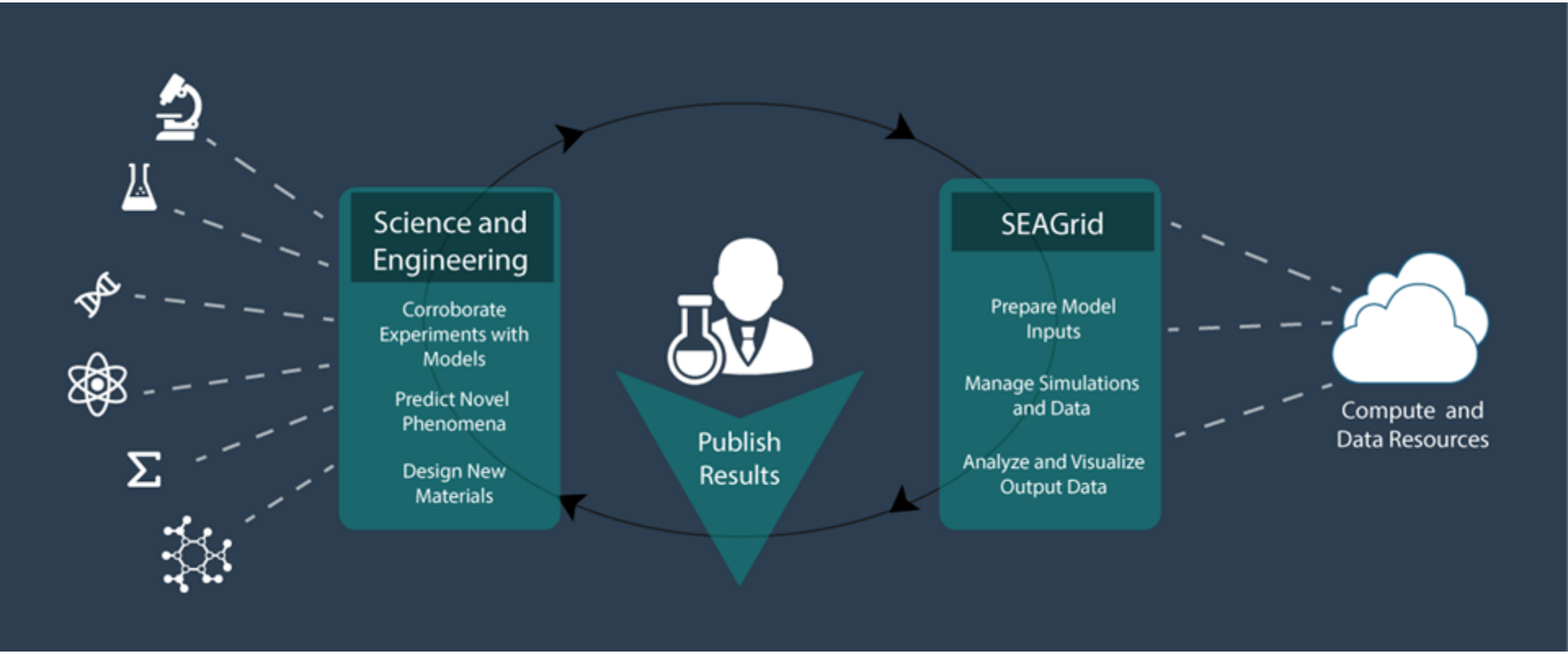
Student Application Deadline: March 25, 2016

SciGaP

What Is a Science Gateway?

SciGaP

SEAGrid.org



SEAGrid.org is an Apache Airavata-powered gateway

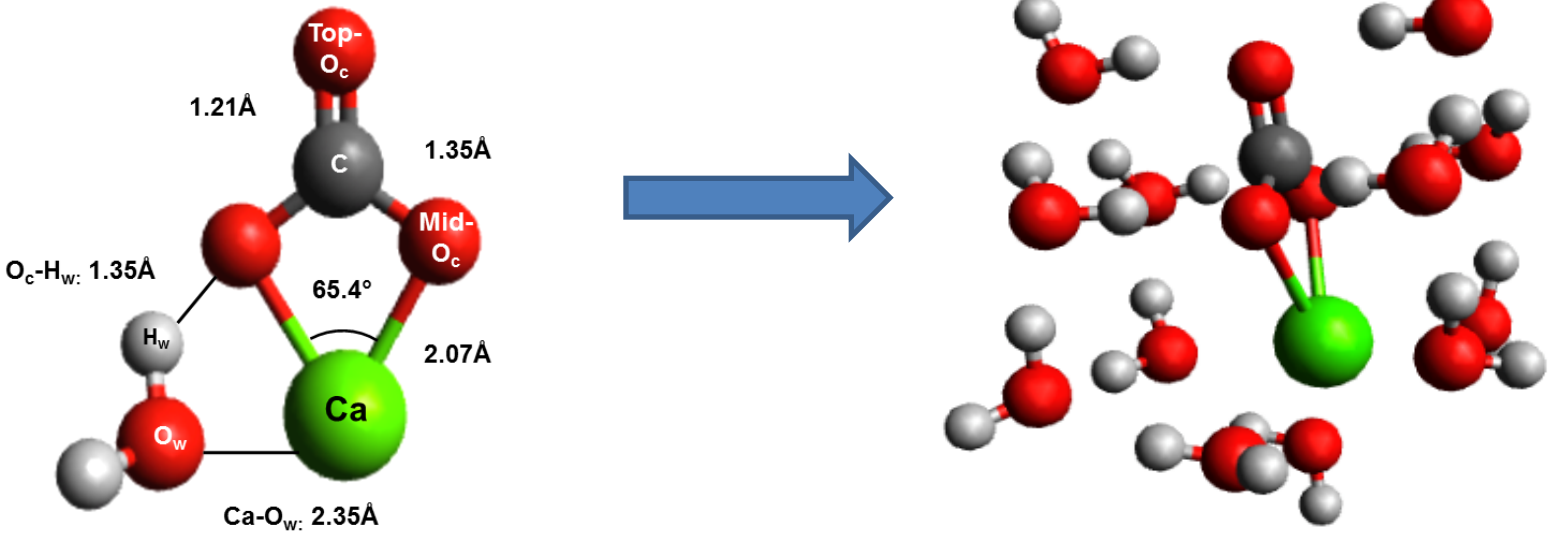
sgg@iu.edu

SciGaP

What is the chemistry of hydrated calcium carbonate?

- Bio-mineralization of skeletons and shells
- Geological C02 sequestration
- Cleanup of contaminated environments

Lopez-Berganza, et al. *J Phys. Chem. A*(2015)

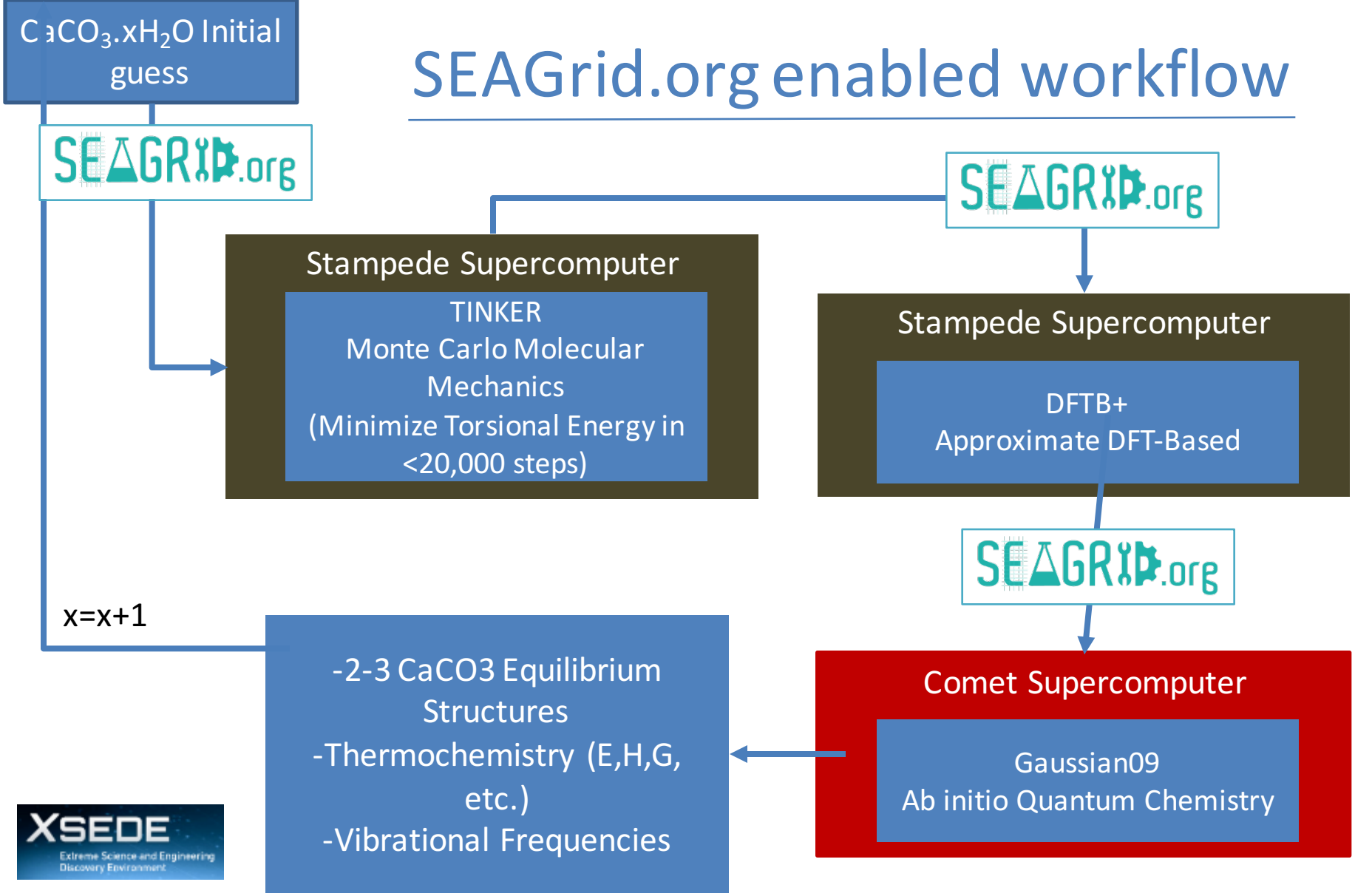


CaCO3.1H2O

CaCO3.12H2O

sgg@iu.edu

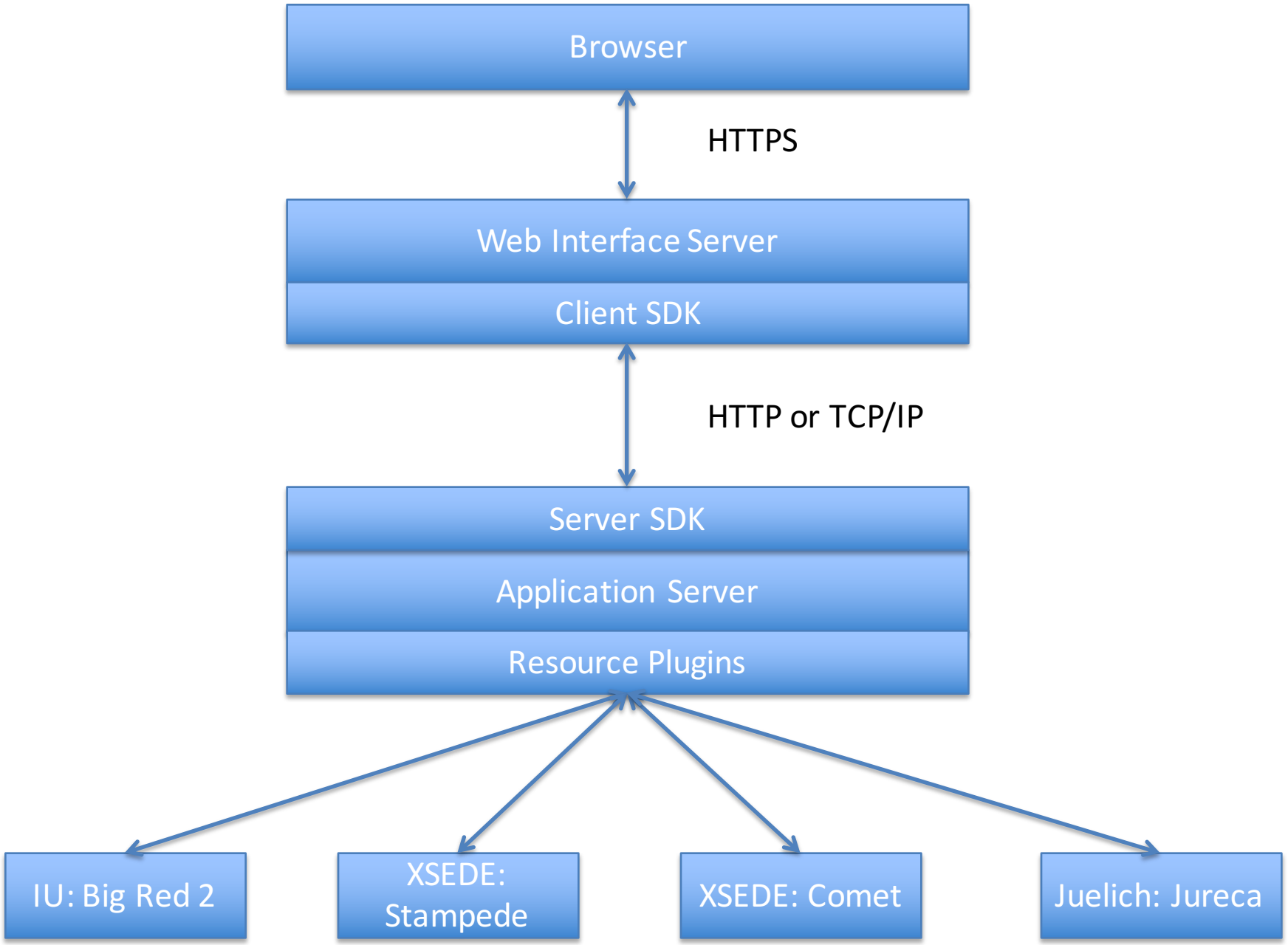
SEAGrid.org enabled workflow



Lopez-Berganza, et al. *J Phys. Chem. A*(2015)

sgg@iu.edu

SciGaP


Browser
HTTPS
Web Interface Server
Client SDK
HTTP or TCP/IP
Server SDK
Application Server
Resource Plugins
IU: Big Red 2
XSEDE: Stampede
XSEDE: Comet
Juelich: Jureca

Challenges for Gateways

- Providing a rich user experience
- Defining an API for the application server
- Defining the right sub-components for the application server.
- Implementing the components, wiring them together correctly.
- Supporting multiple gateway tenants
- Fault tolerance for components
- State management
- Continuous delivery
- Security management
- Supporting full scientific exploratory cycle

SciGaP

How open is open source?

- What's missing?
  - Open source licensing
  - Open Standards
  - Open Code (GitHub)

We also need open governance

SciGaP

Contributing to the Airavata Framework

Component	Research Opportunities
Registry	Better support for Thrift-generated objects; NoSQL and other backend data stores; fault tolerance
Orchestrator	Pluggable scheduling: work stealing internal scheduling implementations, super-scheduling over multiple resources
GFAC	Apache Storm workflow management; elastic packaging.
Messenger	Investigate AMQP, Kafka, and other newer messaging systems
Workflow Interpreter	Alternative workflow processing engines.
Overall	Rewire component micro-services.

We took our software to the Apache Software Foundation to encourage more scientific collaborations. Community-owned software. **GET INVOLVED THROUGH GSOC.**

More Information

- I590 Course URL: http://s.apache.org/i590-spring2016
- Contact:
  - marpierc@iu.edu, smarru@iu.edu
  - Join dev@airavata.apache.org, users@airavata.apache.org, architecture@airavata.apache.org
- Websites:
  - Apache Airavata: http://airavata.apache.org

SciGaP

SciGaP

Developing Computational Science Gateways using Apache Airavata

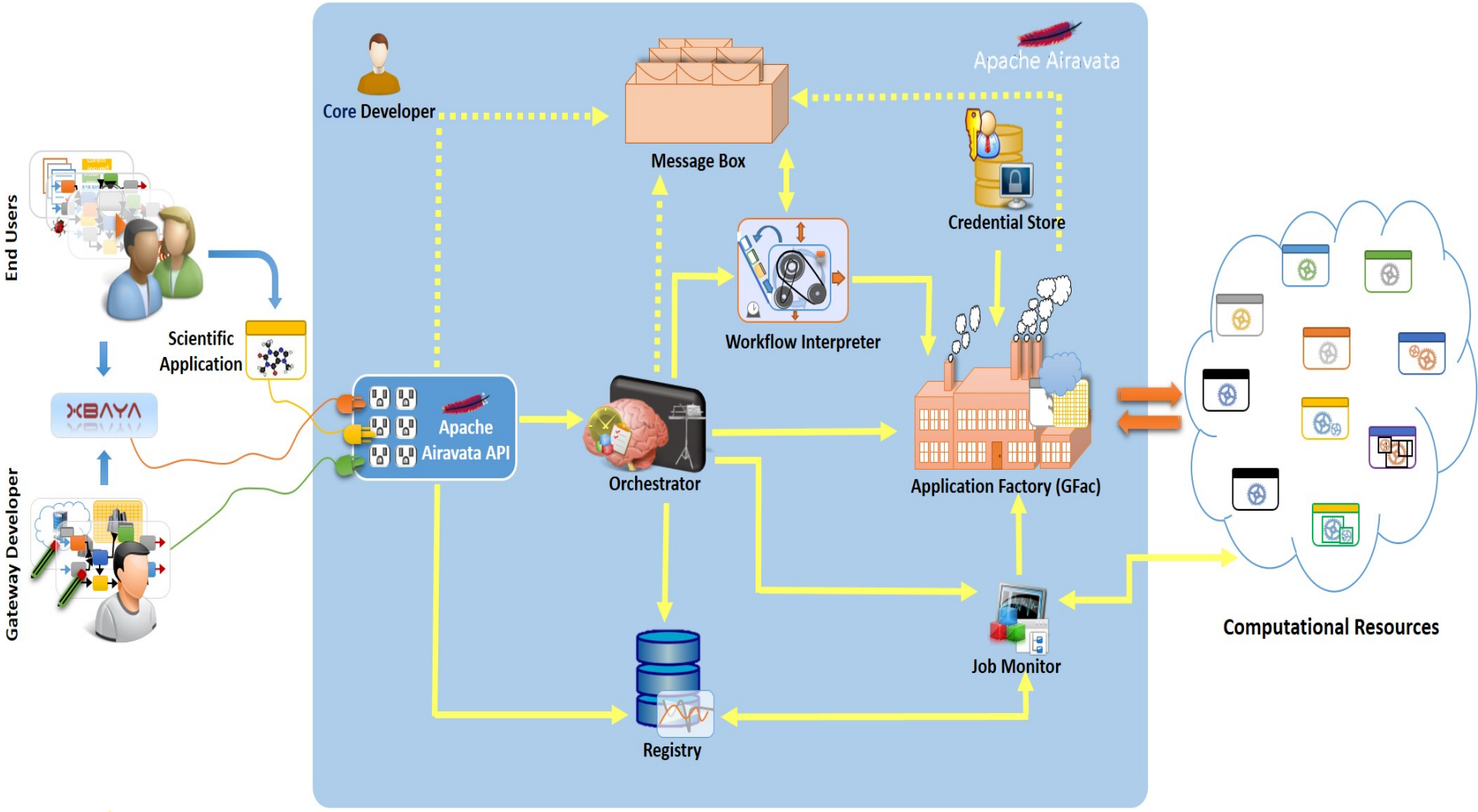
SciGaP

What Is Apache Airavata?

- Science Gateway software system to
  - Compose, manage, execute, and monitor distributed, computational workflows
  - Wrap legacy command line scientific applications with Web services.
  - Run jobs on computational resources ranging from local resources to computational grids and clouds
- Airavata software is derived from NSF-funded academic research.

SciGaP

Airavata Components



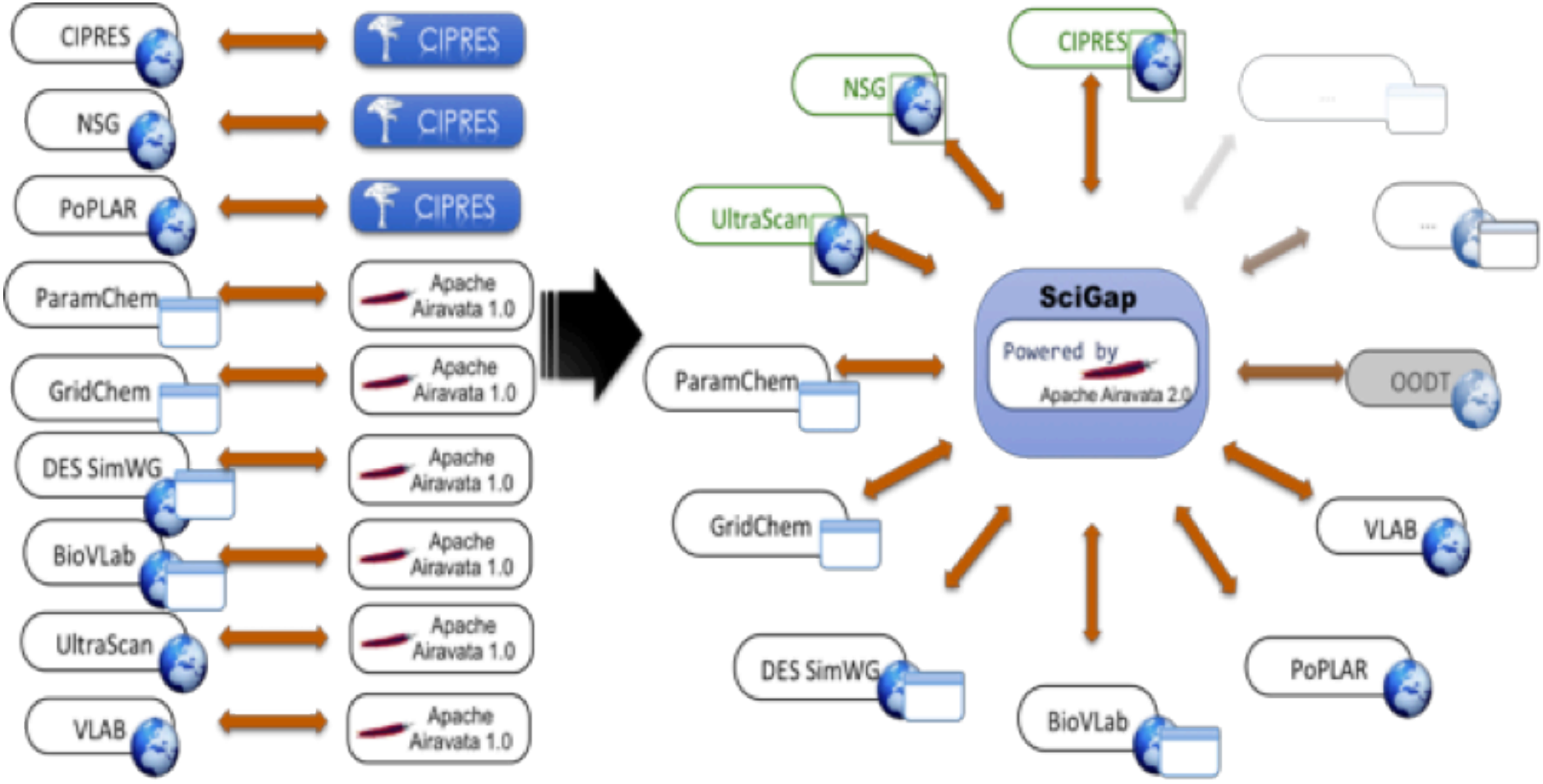
Airavata and SciGaP

- SciGaP: Airavata as a multi-tenanted Gateway Platform as a Service
- Goal: We run Airavata so you don't have to.
  - Scalable support
- Challenges:
  - Centralize system state
  - Make Airavata more cloud friendly, elastic

http://scigap.org

SciGaP

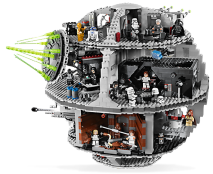
SciGaP Vision



SciGaP

Open Source Software and Governance

- Open source projects need diversity, governance.
  - Sustainability
- Incentives for projects to diversify their developer base.
- Govern
  - Software releases
  - Contributions
  - Credit sharing.
  - Members are added
  - Project direction decisions.
  - IP, legal issues
- Our approach: Apache Software Foundation



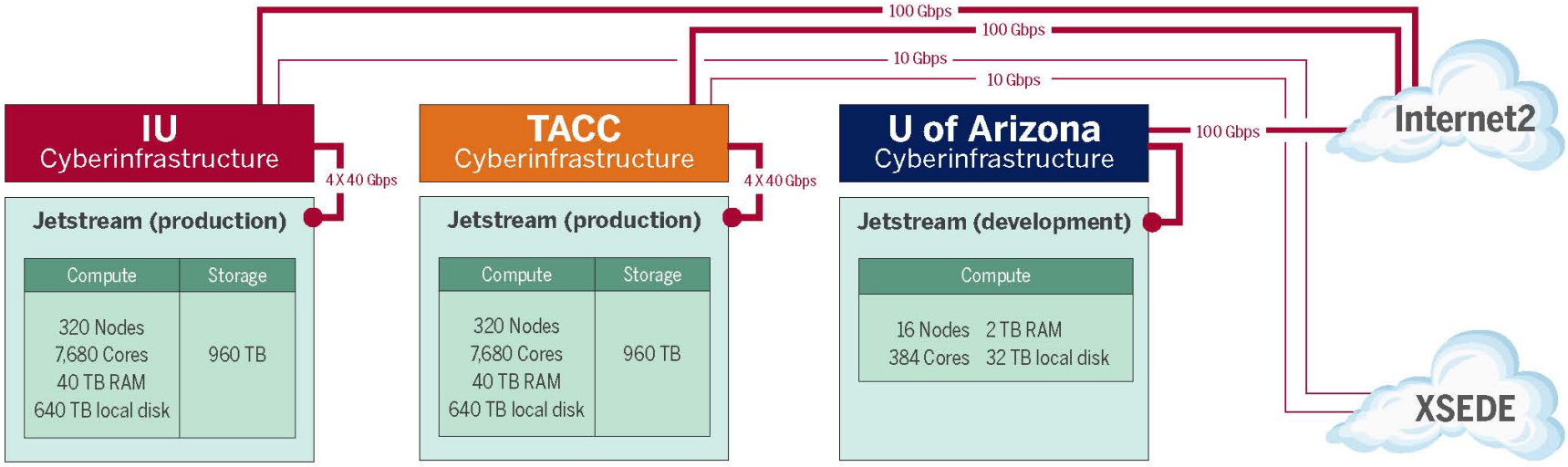
Compete

Collaborate

SciGaP

SciGaP

Jetstream



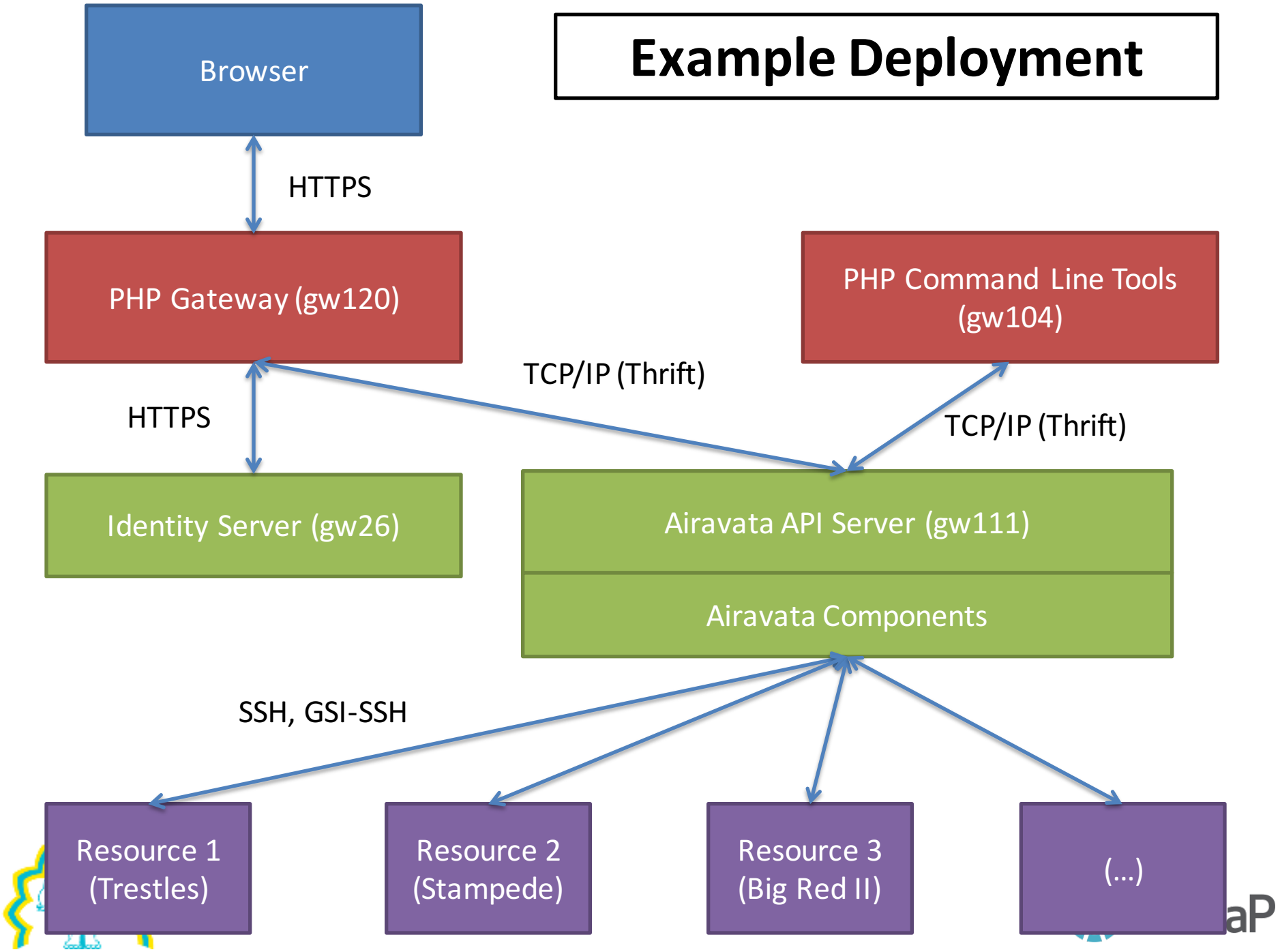
- Geographically Distributed Cloud, 0.5 PetaFLOPS
- High-speed connections to Internet2 and local connections to Wrangler disk storage at IU and TACC
- Globus-based large scale file movement

SciGaP

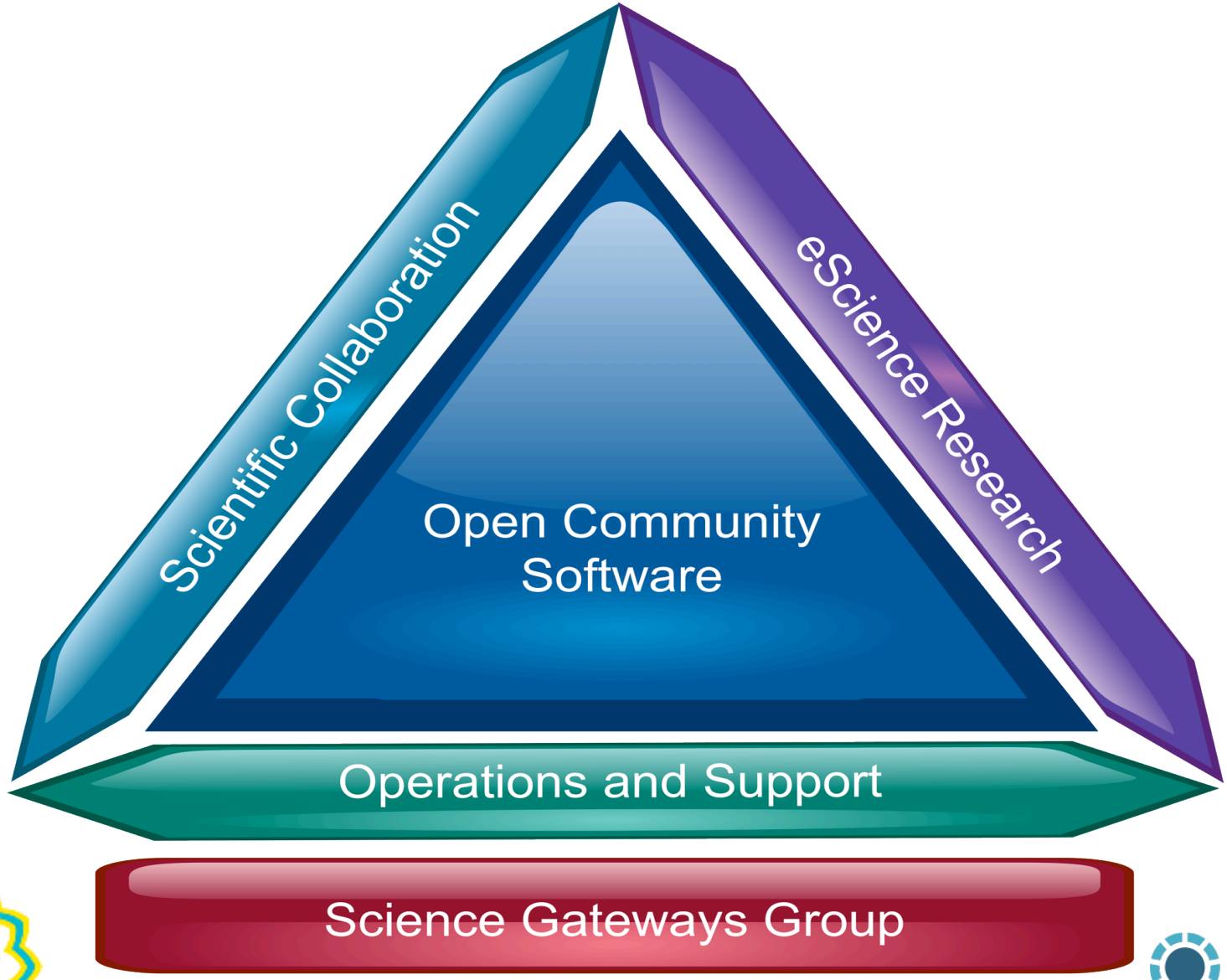
Science Gateway Communities

Community	Gateway Capabilities
Universities and Academic Departments	Provide simplified access to computing resources for students, faculty, and staff. Gateways should support MOOCs
Shared Instrument Facilities	Simplify access to instruments, support research derived from common data products: UltraScan, LIGO, DES, LSST
Multisite Collaborations and Virtual Organizations	Funded or unfunded, including self-organized communities, who need to collaborate and use a common pool of resources: LEAD, ENZO
Businesses and Services	Provide Software as a Service for a particular application suite
Small Research Groups	"Long tail" of science, need to preserve the work that is done.

SciGaP


Example Deployment
Browser
HTTPS
PHP Gateway (gw120)
PHP Command Line Tools (gw104)
TCP/IP (Thrift)
HTTPS
TCP/IP (Thrift)
Identity Server (gw26)
Airavata API Server (gw111)
Airavata Components
SSH, GSI-SSH
Resource 1 (Trestles)
Resource 2 (Stampede)
Resource 3 (Big Red II)
(...)

A Balancing Act



SciGaP

What is Cyberinfrastructure?

Scientific Distributed Computing

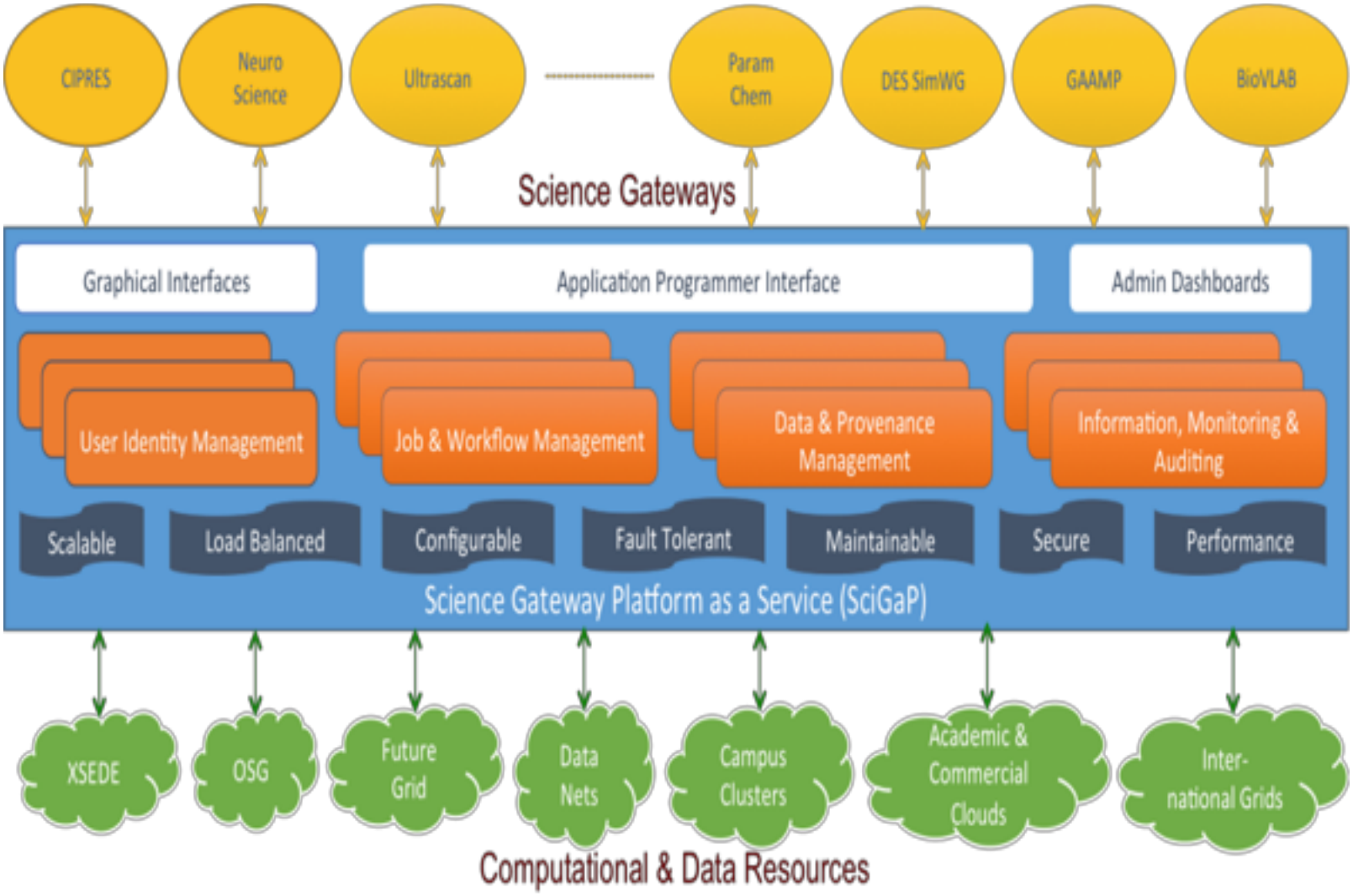
SciGaP

What Is Cyberinfrastructure?

"Cyberinfrastructure consists of computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked together by software and high performance networks to improve research productivity and enable breakthroughs not otherwise possible."

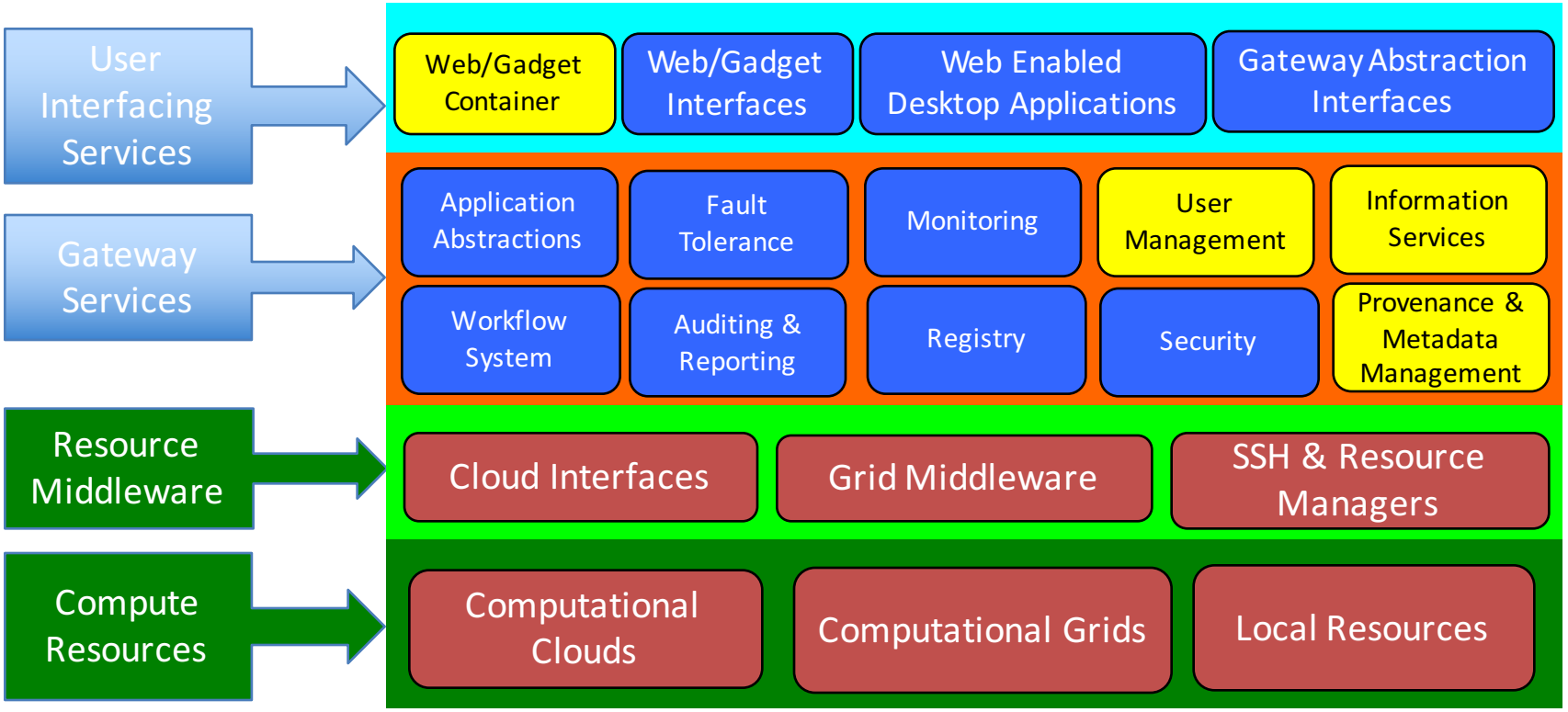
–Craig Stewart, Indiana University

SciGaP



SciGaP Role and Goals: Improve sustainability by converging on a single set of hosted infrastructure services

Cyberinfrastructure Layers



Color Coding

Airavata components

Complimentary gateway components

Dependent resource provider components

SciGaP

Apache Airavata Components

Component	Description
Airavata API Server	Apache Thrift-generated server skeletons of the API and data models; directs traffic to appropriate components
Registry	Insert and access application, host machine, workflow, and provenance data.
Orchestrator	Handles experiment execution request validation, scheduling, and decision making; selects GFAC instances that can fulfill a given request
GFAC	Manages the execution and monitoring of an individual application.
Workflow Interpreter	Execute the workflow on one or more resources.
Messaging System	WS-Notification and WS-Eventing compliant publish/subscribe messaging system for workflow events