



iRODS in the Cloud: SciDAS and NIH Helium Commons



Claris Castillo

RENCI, UNC Chapel Hill



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL



SciDAS



renci

WASHINGTON STATE
UNIVERSITY



Not Scaling up Data Analysis is Not an Option

20th Century



21st Century



Normal veteran (giga-/terascale) and newbie (megascale) users MUST ADVANCE to the peta/exa-scale in this generation. Issues:

- Limited computational skills (What is a C library?)
- Poor use of advanced networks (We need more HDs to mail!)
- Limited access to computational resources (awareness, \$\$\$)
- Unpredictable time to compute result (queue times, queue times, queue times, broken nodes, segfaults, OOM, data geography)
- Missing skillsets (I only know Perl)
- Data must be organized and good stuff deleted (Data policies)

DatAPocalypse Prediction (Genomics):

In 20 years, every CVS, subway, hospital, research lab, public health facility, police station, etc will have a DNA sequencer generating Exabytes of data in aggregate each week.

- *How many bioinformaticists are on the CVS payroll?*
- *How many faculty recruitments failed because campus X research computing resources are stuck in 2015?*
- *How many adverse drug reactions were not predicted because of limited/broken cyberinfrastructure?*



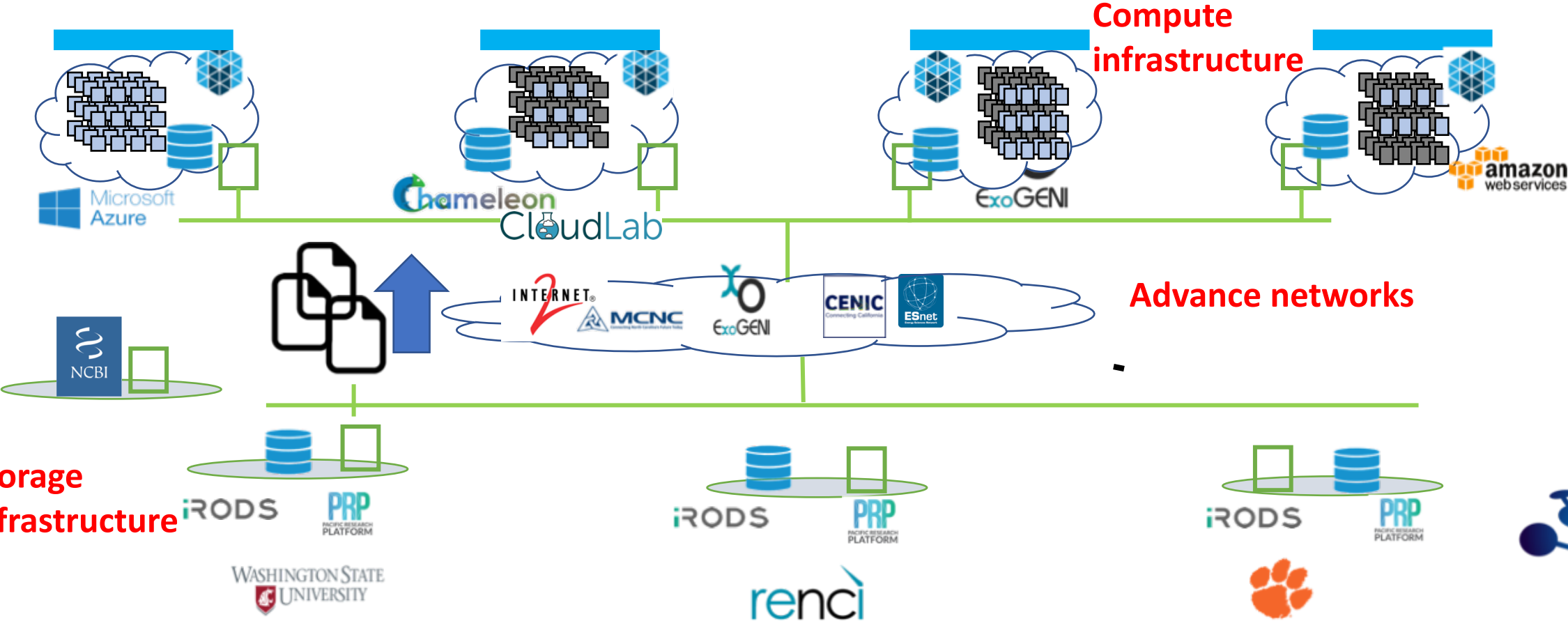
Alex Feltus

WisegEEK.org
www.smartpractice.com

Heterogeneous and Complex CI Ecosystems

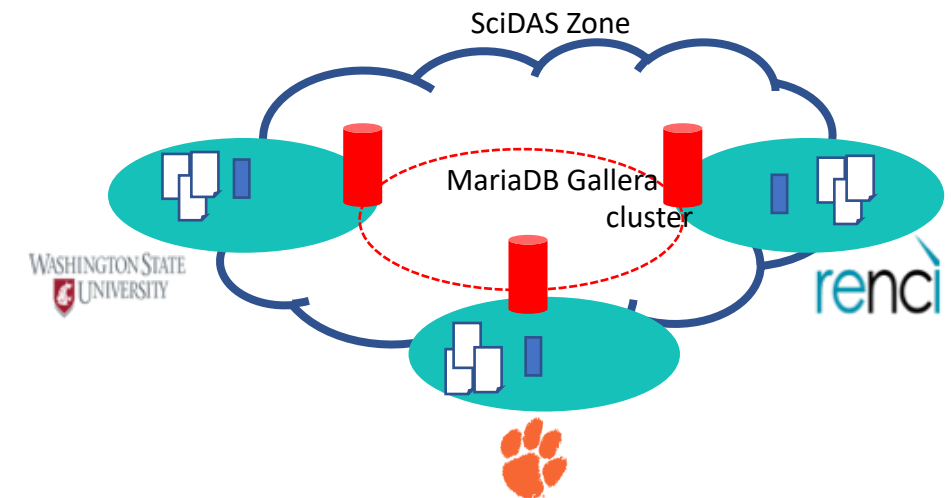
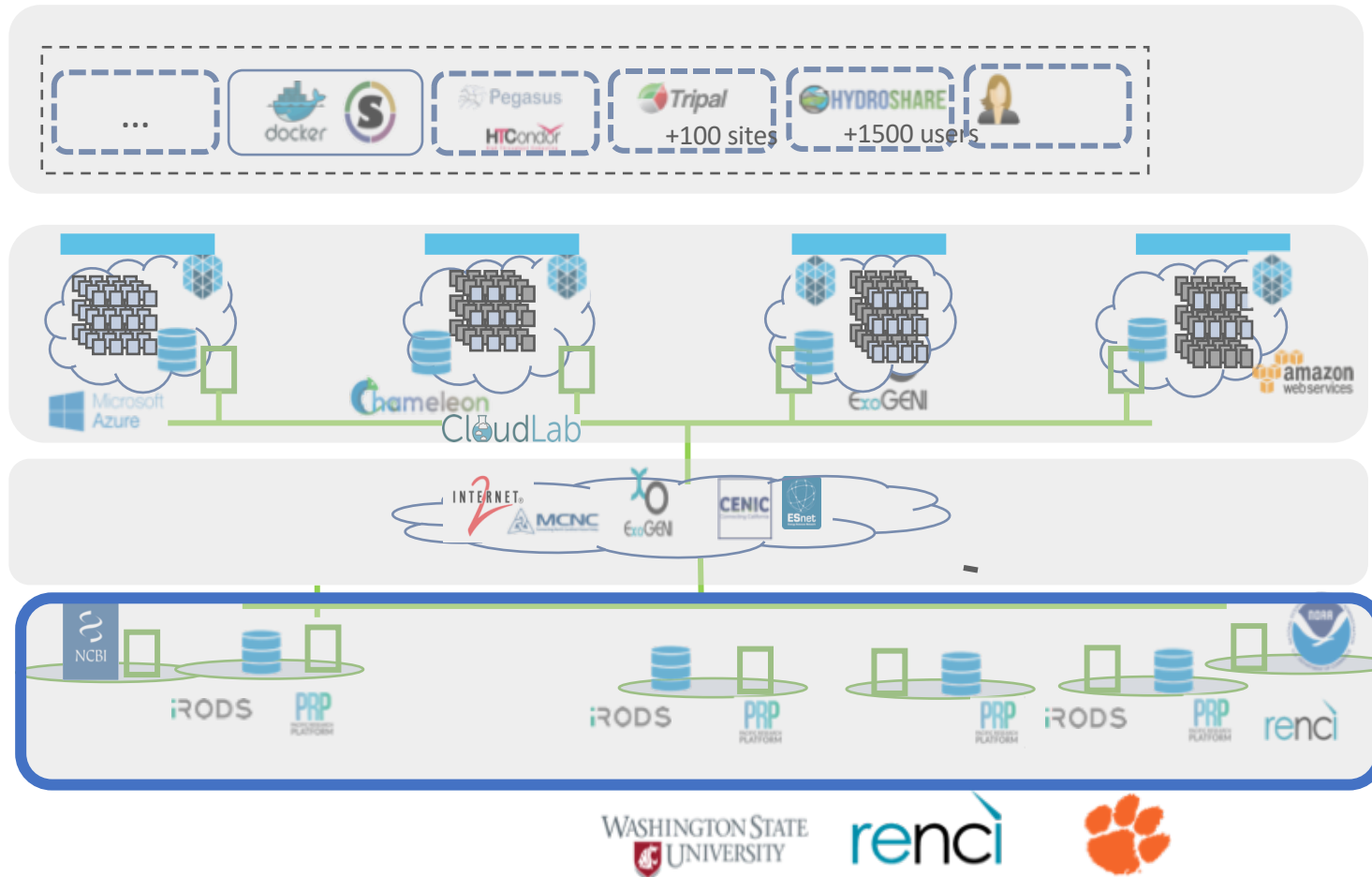


Community data sharing platforms



Commoditization of Cloud computing and the convergence of compute, storage, data and network technologies enables the 'illusion' of a single large computer consisting of widely distributed systems.

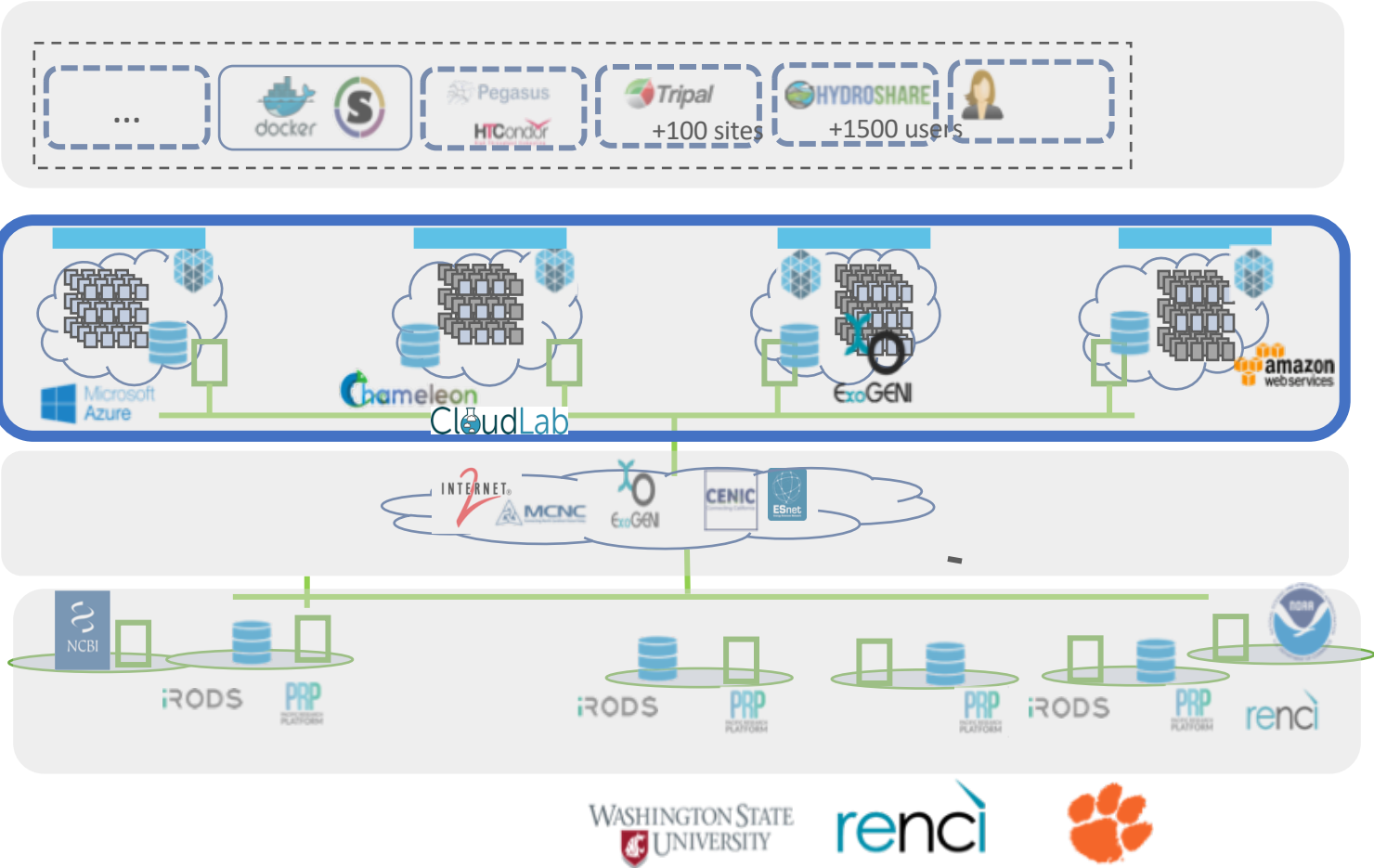
Breakdown: One Layer at A Time -- Data



iRODS team connected iRODS to a MariaDB Galera Cluster to provide a *multi-master, distributed* iRODS catalog over the WAN.

"Distributing the iRODS Catalog: a way forward", M. Stealey, et. al. iRODS User Group Meeting (UGM), Netherlands, 2017.

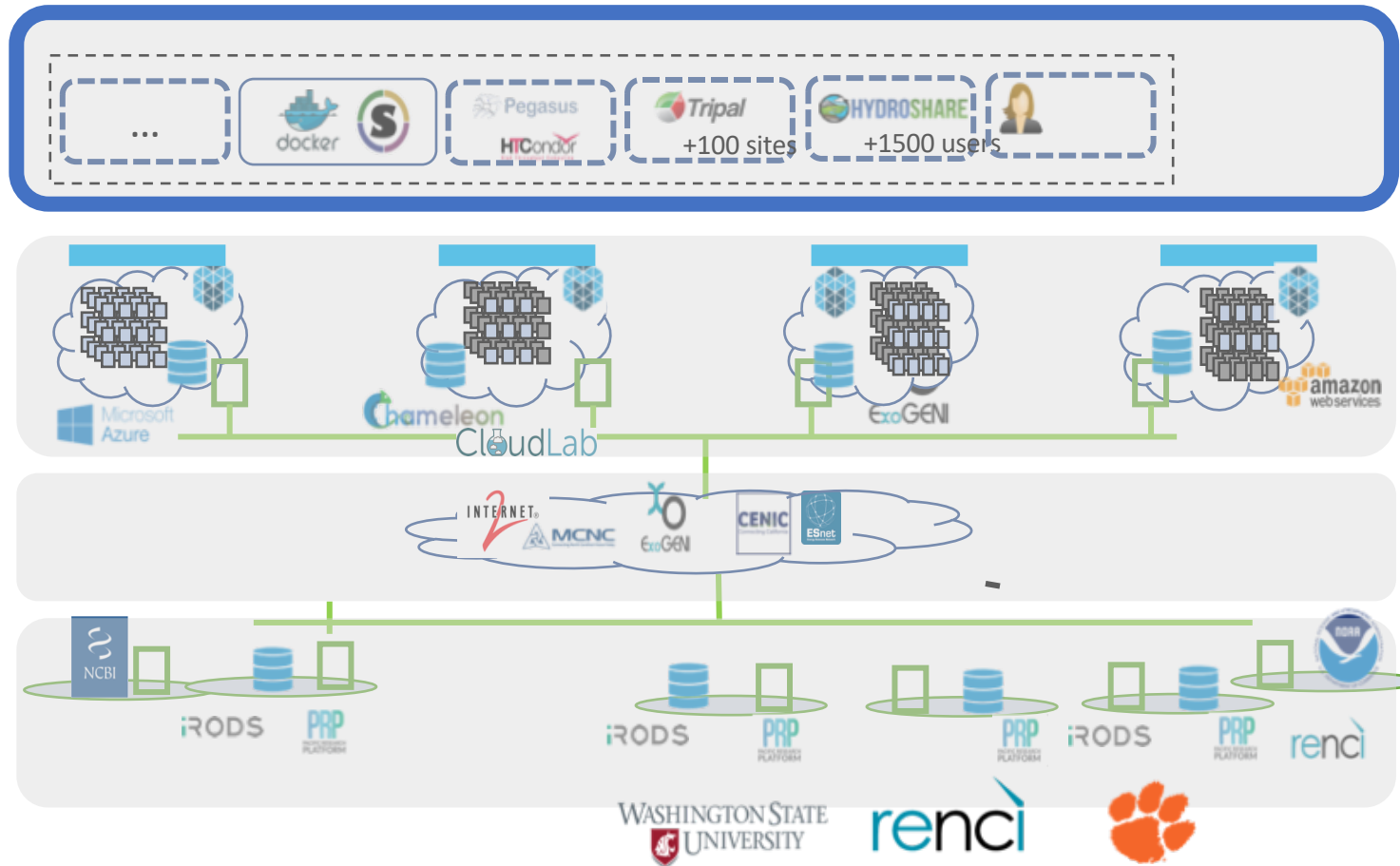
Breakdown: One Layer at A Time -- Compute



Apache Mesos: A layer of abstraction, to utilize an entire *data center* as a single large server



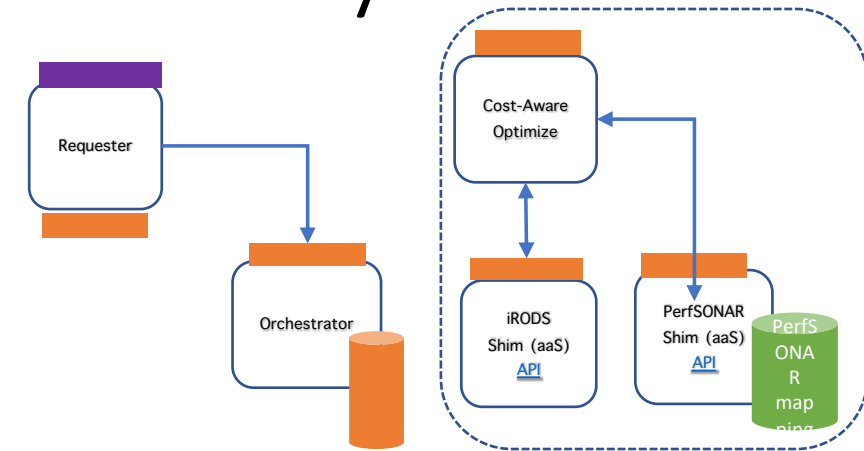
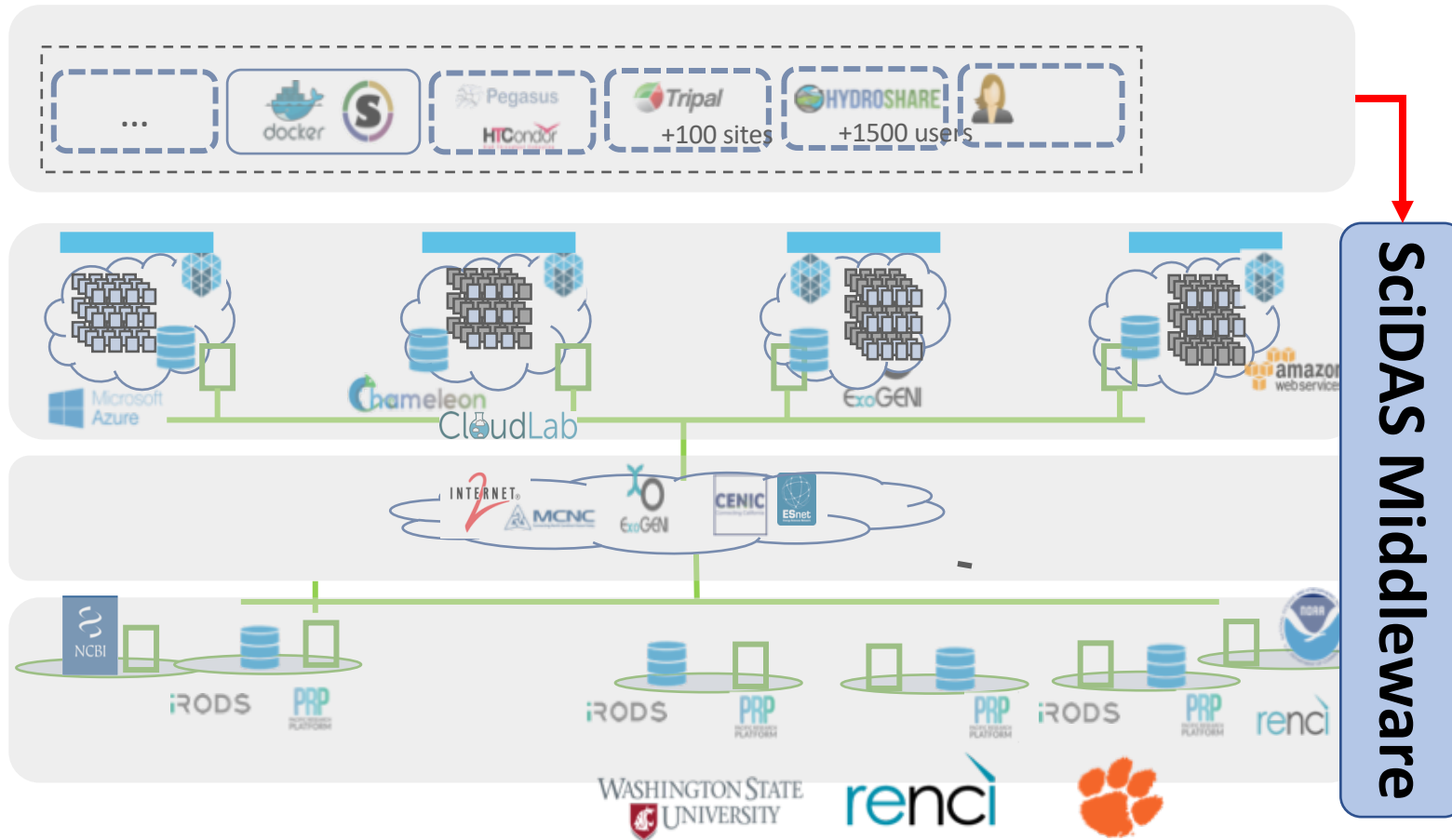
Breakdown: One Layer at A Time – Scientific Tools



Scientific applications will be available in the form of SciApps “virtual appliances” (*NSF CC-ADAMANT, [works15]*)

[works15] *Enabling Workflow Repeatability with Virtualization Support*, Fan Jiang et.al. Workshop on Workflows of Large-Scale Science, Supercomputing Conference (SC15), Austin, Texas, 2015.

SciDAS: Bringing it All Together Into One System

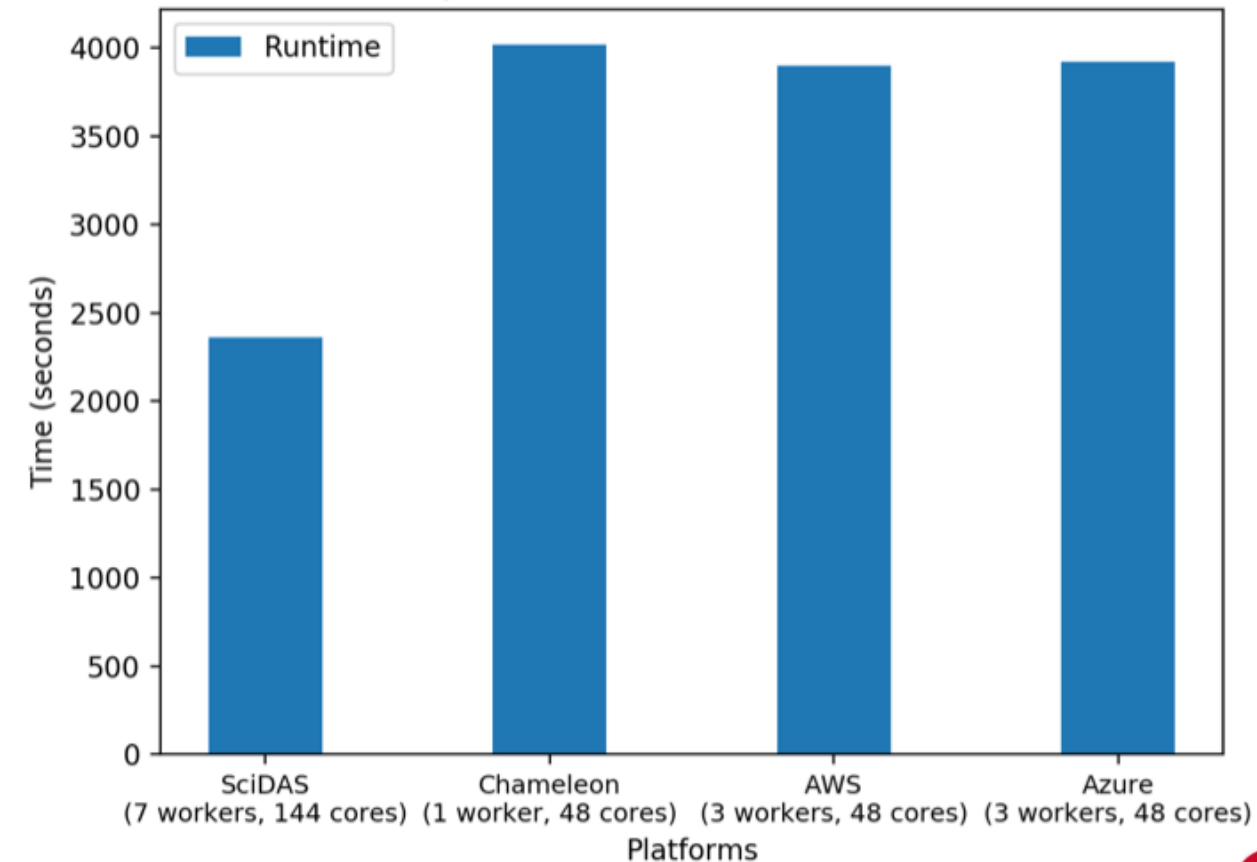


- Network aware placement
 - Optimize for data locality
- Capability aware resource aware placement
 - GPU able nodes
- Authentication and authorization infrastructure
 - CiLogon

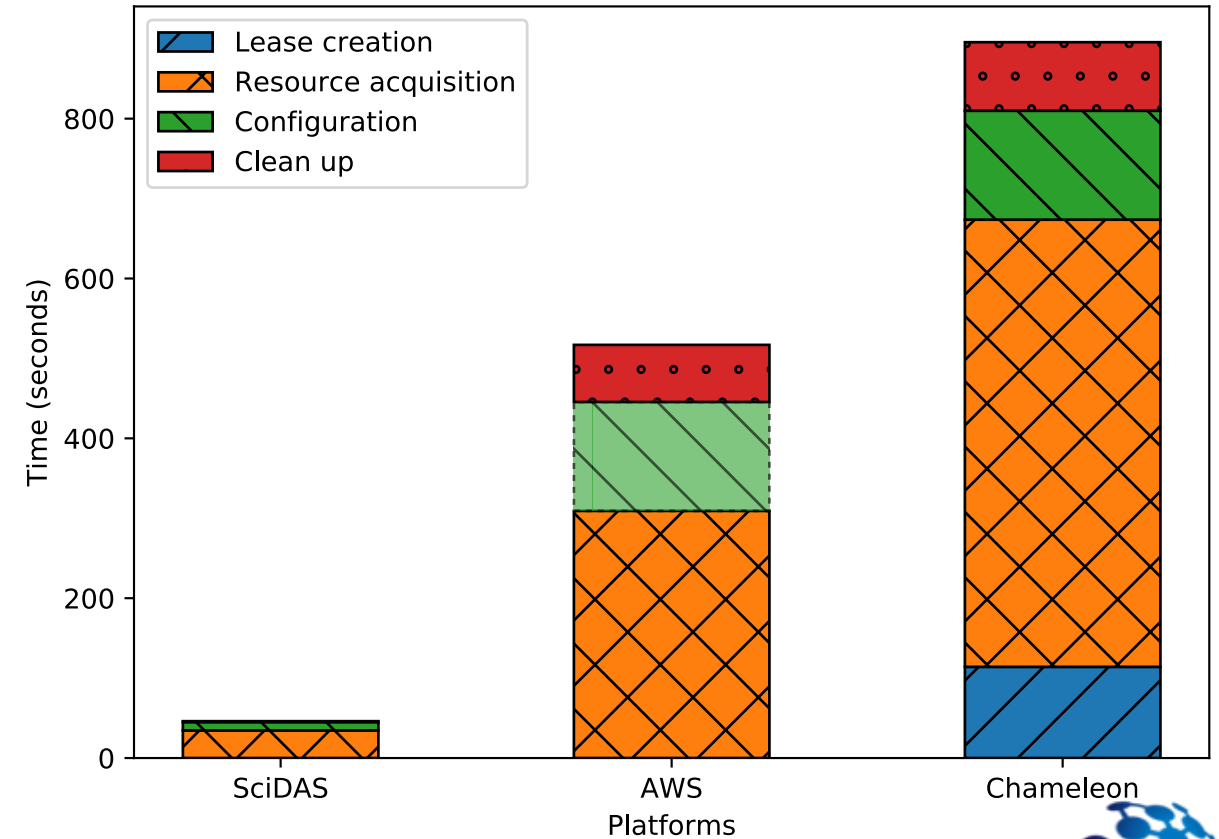
[works15] Enabling Workflow Repeatability with Virtualization Support, **Fan Jiang** et.al. Workshop on Workflows of Large-Scale Science, Supercomputing Conference (SC15), Austin, Texas, 2015.

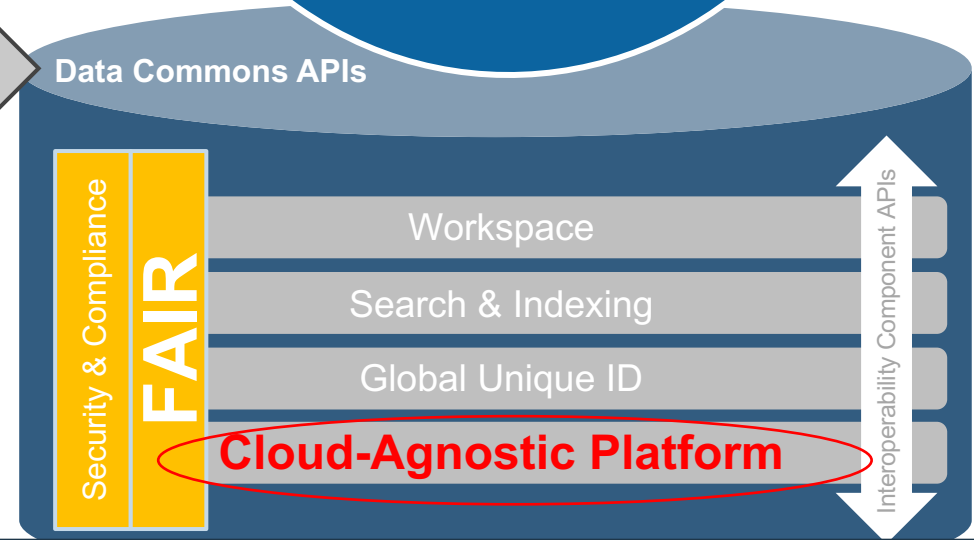
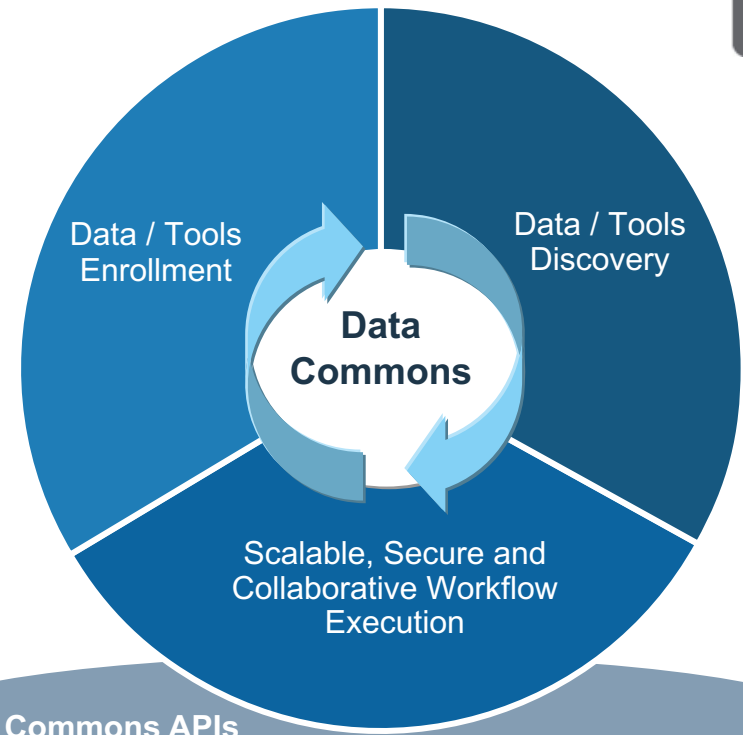
Improving scientific productivity by the numbers

Comparison of KINC Workflow Runtime



Provisioning Time Comparison among Computing Platforms





Helium Commons

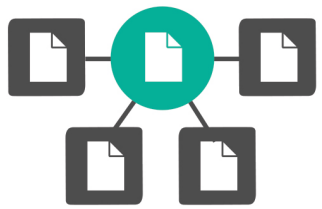
Virtualization system

Metadata to encode rich information


Rule engine programmed with rules to enact policies

Data Federation


DATA VIRTUALIZATION




DATA DISCOVERY



WORKFLOW AUTOMATION



SECURE COLLABORATION





TopMED
MOD
GTEx
...



Bring-Your-Own-Data
Bring-Your-Own-Data-Service

iRODS enables powerful data sharing models in the Commons

BYOD: Cloud storage can be added as storage resources

Data Federation (default): continuous virtual system while retaining control of each endpoint

Extended data collaboration (BYODS): Seamless integration with data hosted on external data services

TopMED
MOD
GTEx
...



Bring-Your-Own-Data
Bring-Your-Own-Data-Service

Thank you!

claris@renci.org