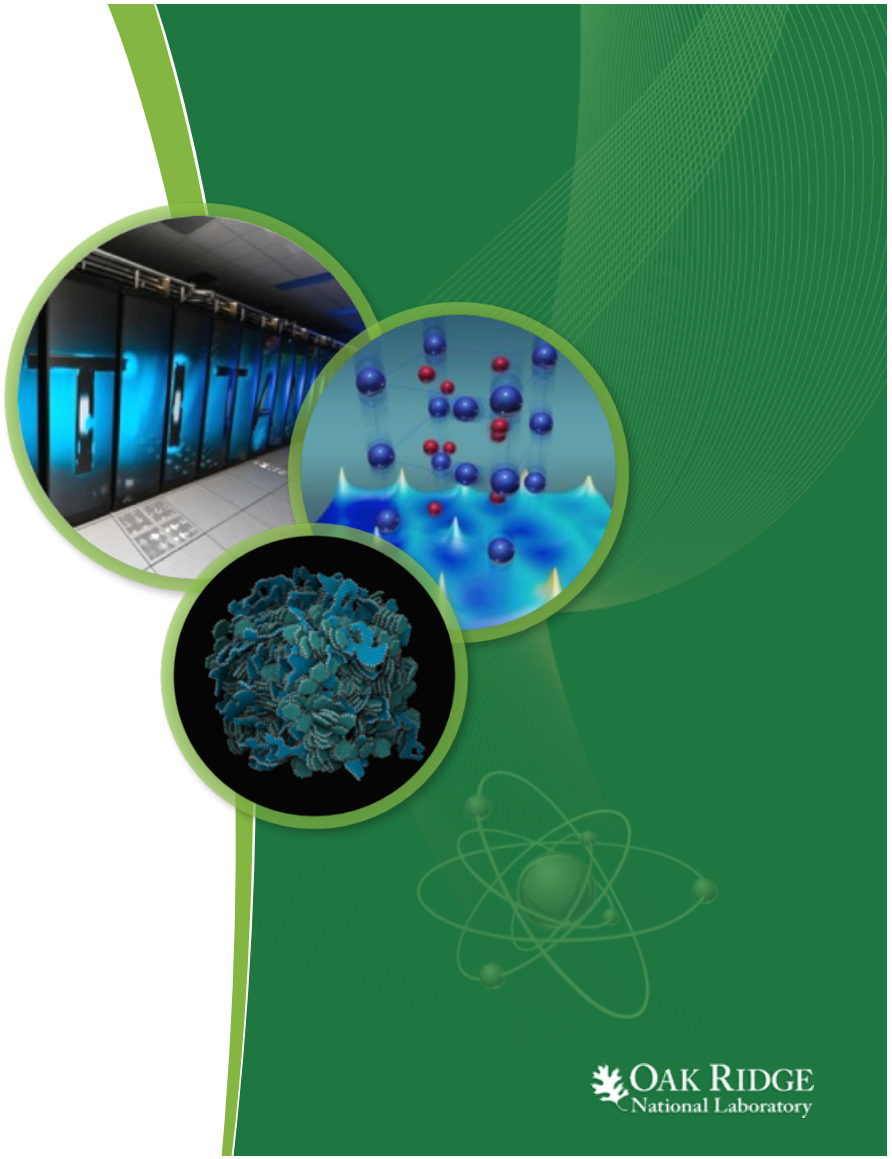
ORNL Science DMZ and Bridging CADES Workflows

Compute and Data Environment for Science (CADES)

Advanced Data Workflow Group

Ryan Prout



ORNL is managed by UT-Battelle
for the US Department of Energy

Goals of The Presentation

- CADES overview
- SDMZ Architecture
- Workflows and Projects
- Future

CADES Resources

- OpenStack Cloud
  - 16 Hypervisors (1,024 VCPU's, 2TB Memory, 20.5 TB Storage)
  - Lustre host aggregate
  - Birthright for the lab
  - Expanding quickly
- Compute
- Storage
  - Lustre, NFS, Scality (Research Data Archive)
- DTNs and SDMZ
- Workflow design and support

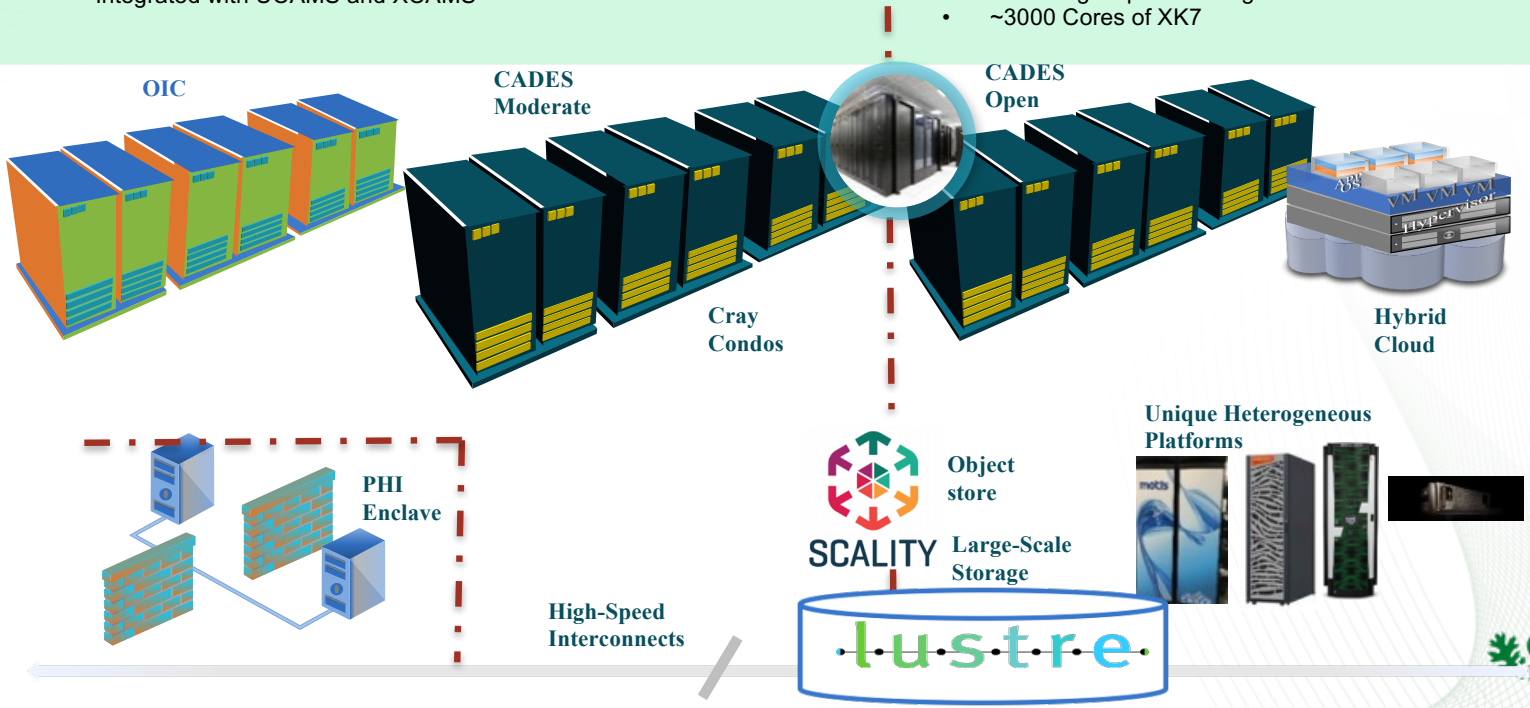
CADES Deployment

.. and several ORNL projects on OIC

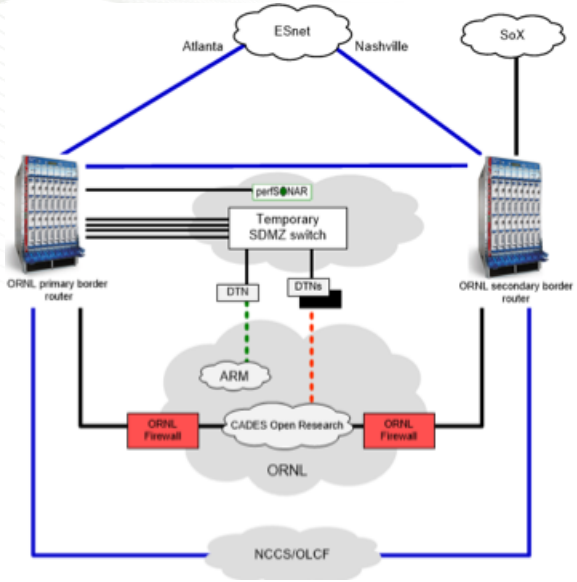
.. and several other smaller projects.

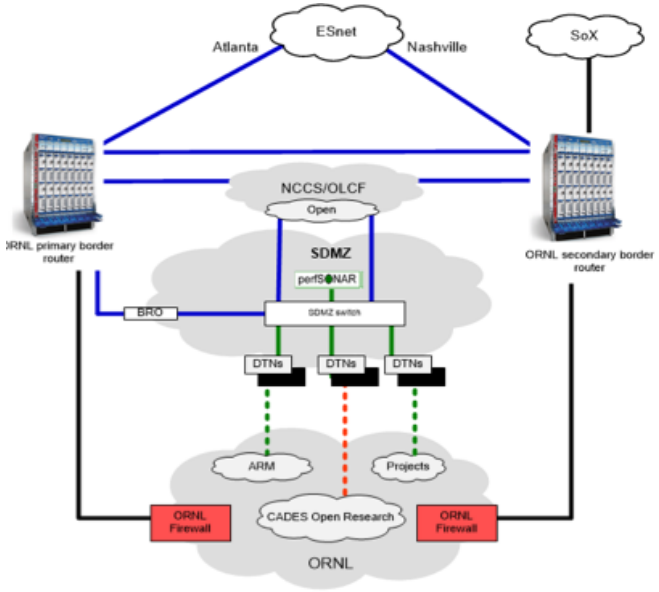
- ~5000 Cores of Integrated Condos on Infiniband
- ~10,000 OIC Cores
- Attested PHI Enclave
- Integrated with UCAMS and XCAMS

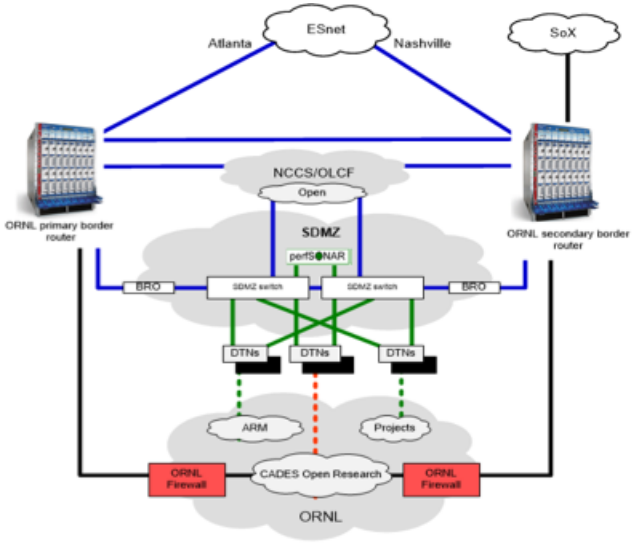
- ~6000 Cores of Integrated Condos on Infiniband
- ~5000 Cores of Hybrid, Expandable Cloud
- SGI UV, Urika-GD/XA: GX
- 5PB+ High-Speed Storage
- ~3000 Cores of XK7



Science DMZ roadmap







OAK RIDGE National Laboratory

ORNL SDMZ Advantages

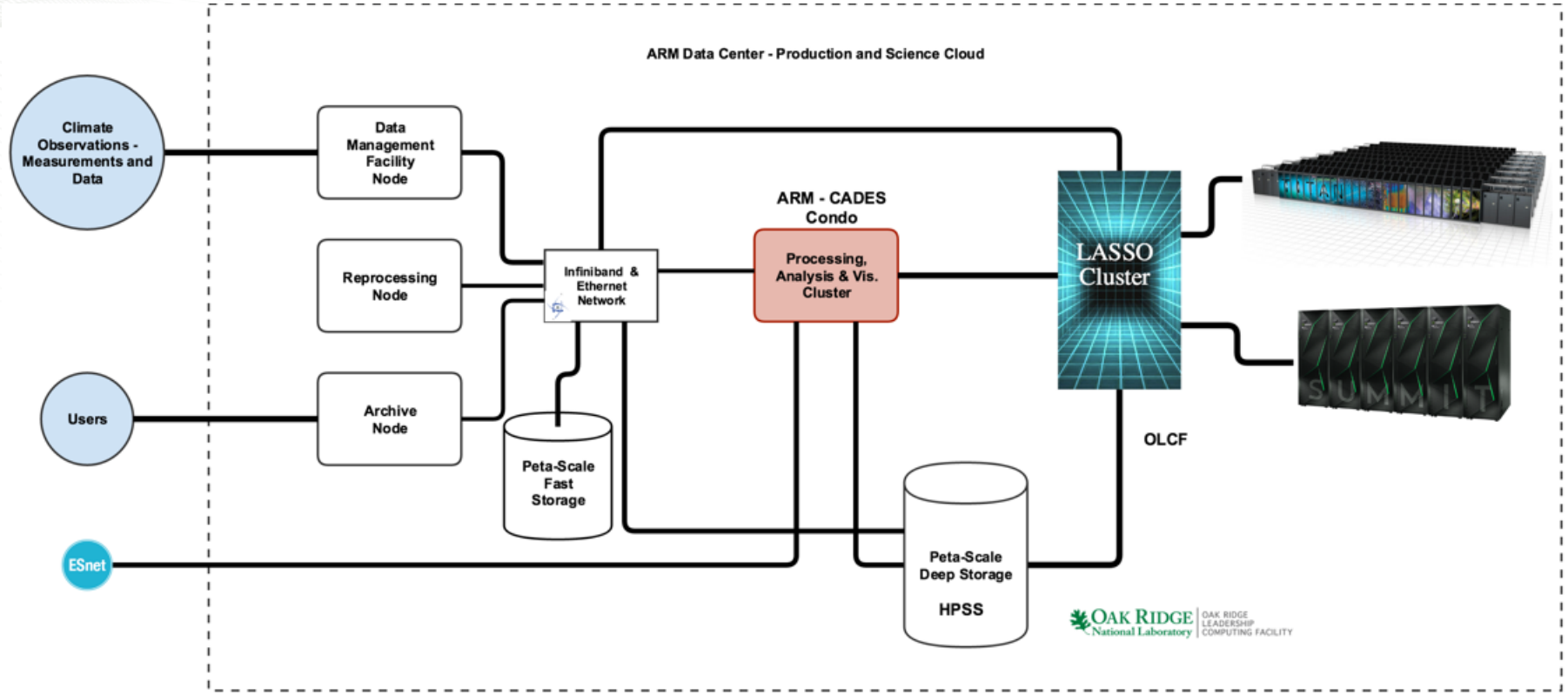
- Create advanced workflows
  - CrossBOW Project
  - ARM<->CADES<->NCCS/OLCF
- High performance
- Scalability
- Internal and External collaborations
- Scientific workflow systems

Bridging workflows through SDMZ

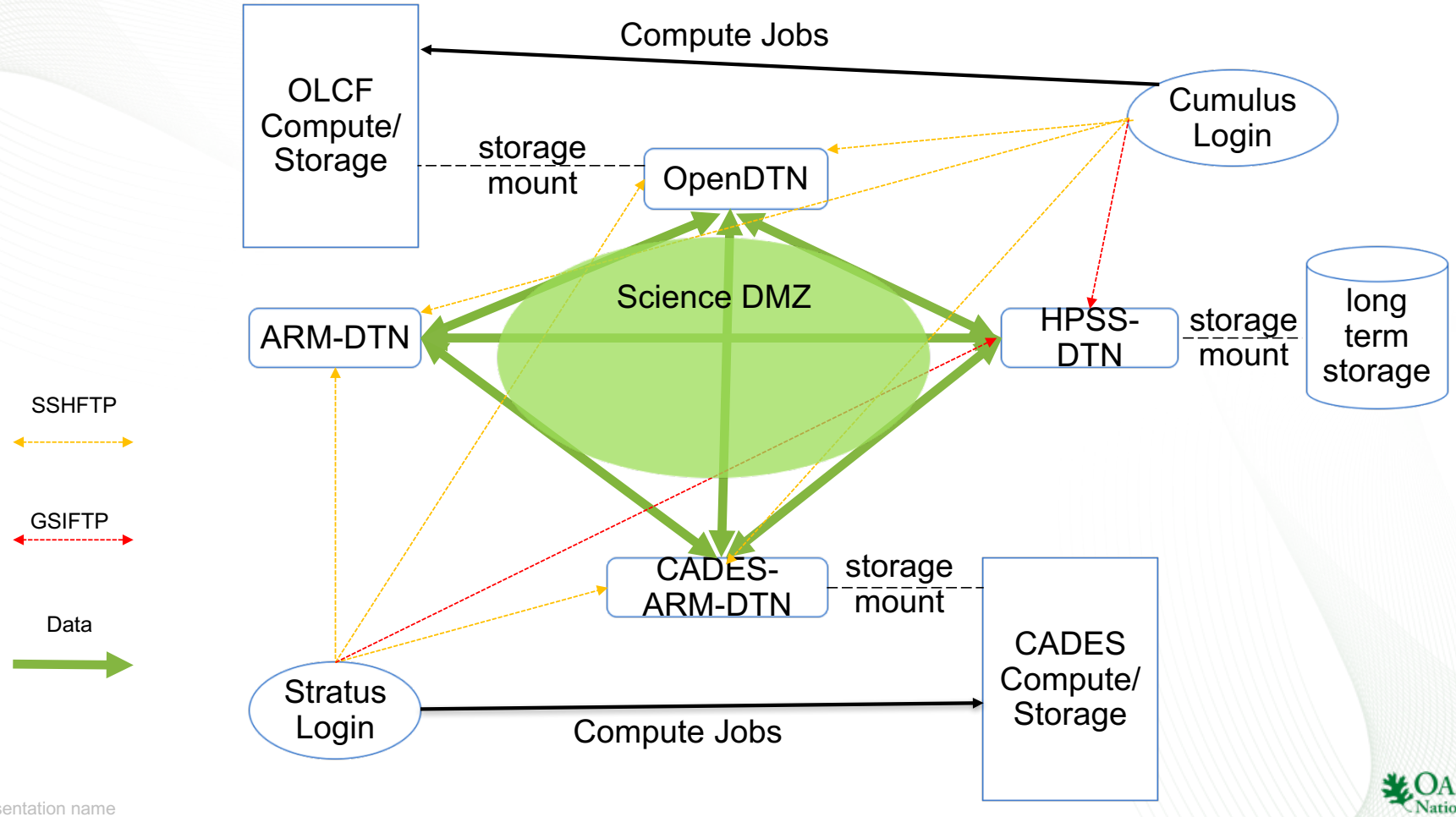
- ARM <-> CADES <-> NCCS/OLCF
  - Atmospheric Radiation Measurement Climate Research Facility
  - Phase I: globus-url-copy for data movement and automation
  - Phase II: Globus APIs and application

- CrossBOW Project (Cross-platform Big Data Operational Workflows)
  - Globus APIs and CKAN server with CrossBOW API
  - Focus on deep learning workflows
  - Challenge of automating and scheduling of analysis tasks
  - https://ramanathanlab.org/cosc526/

ARM Resource Overview



Phase I: ARM Workflow and Connection Types



Phase II: ARM Workflow

- Start working towards the utilization of Globus APIs
- Shared Endpoints
- Integrate workflow portals
- Use the Phase I time to better understand processes and needs

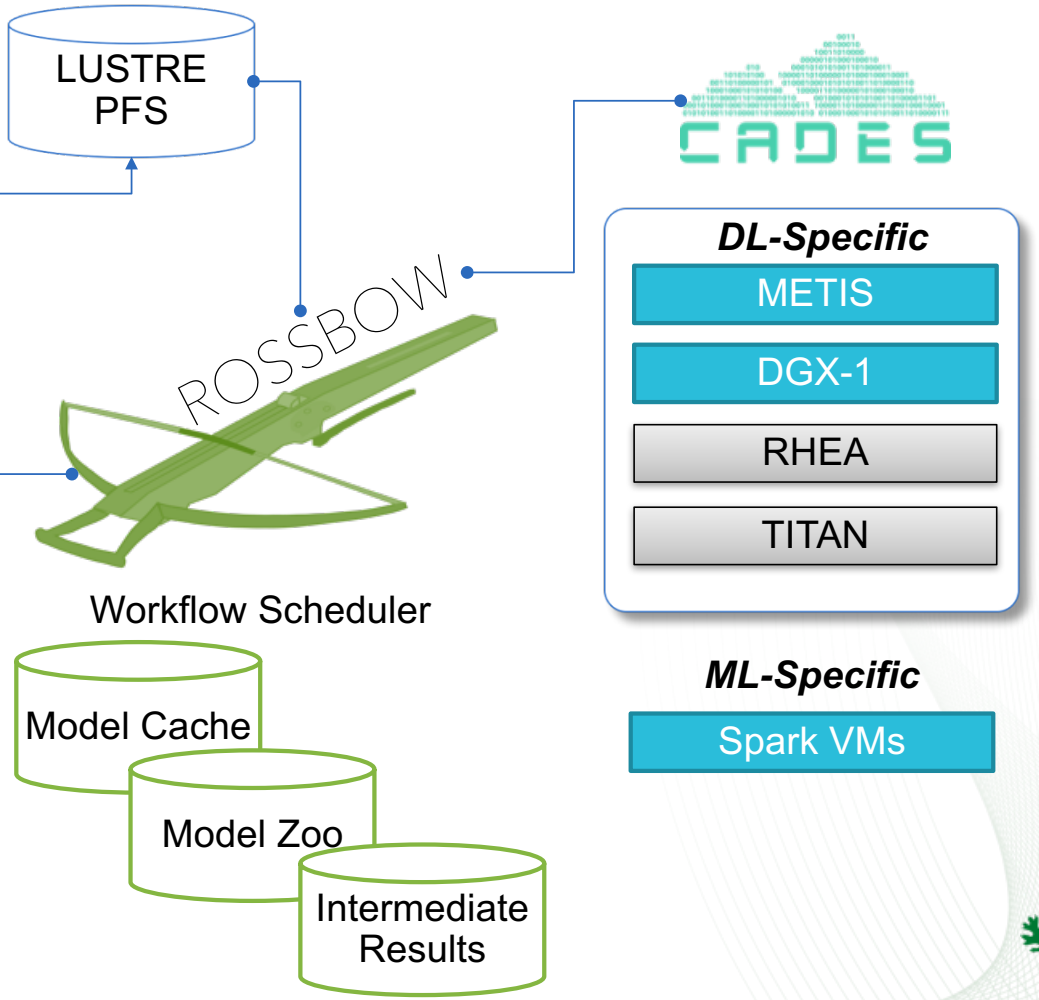
CrossBOW: Cross-platform Big data Operational Workflow

- *File pointers within Lustre file system as URIs*
- *Access based on OLCF / CADES policies*

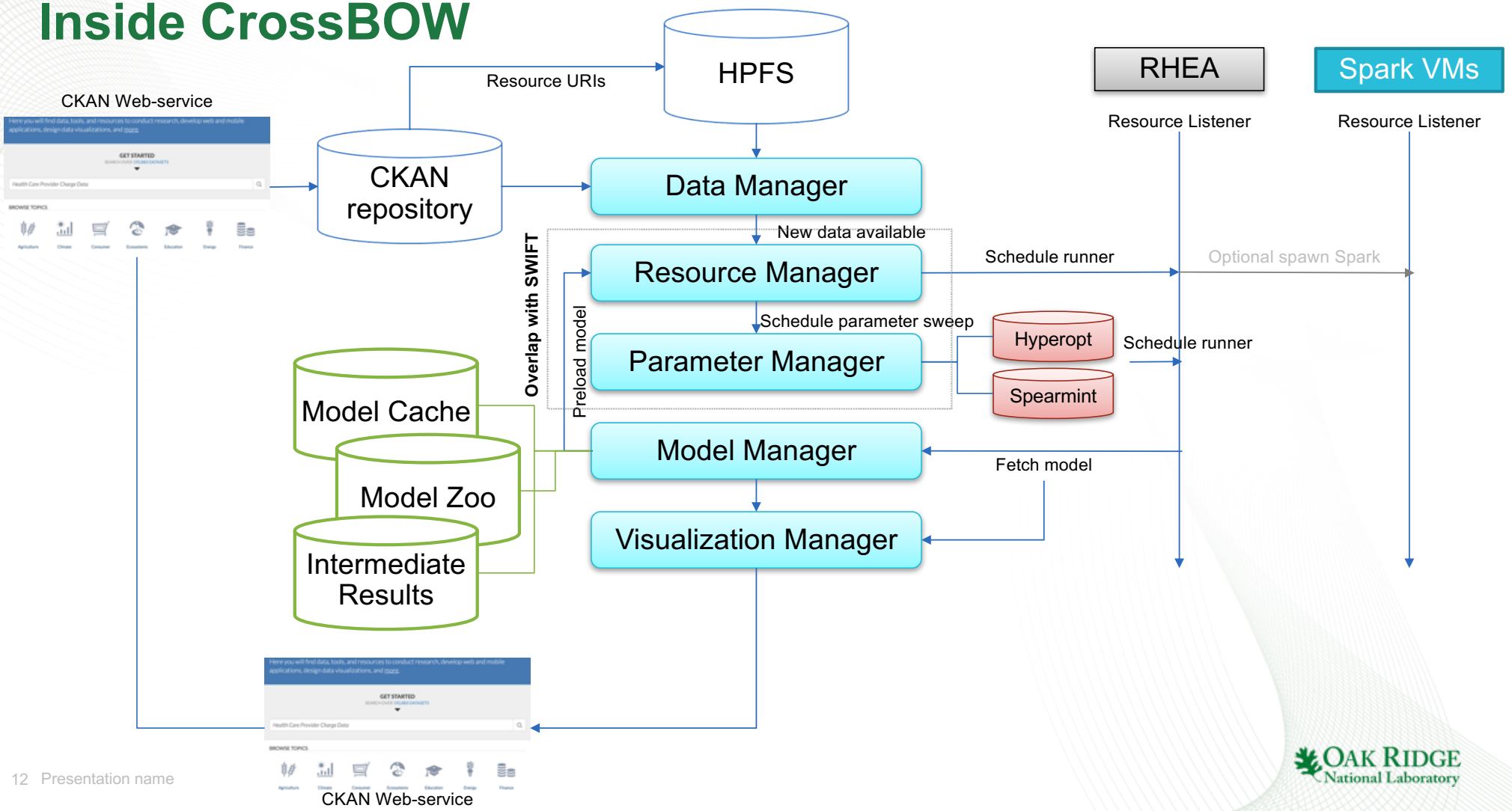
Front-end portal
http://data.ornl.gov

- *Possible connections to ORNL (i.e., registered OLCF) users*
- *XCAMS integration*
- *Scientific datasets*

ORNL LDRD 8279



Inside CrossBOW

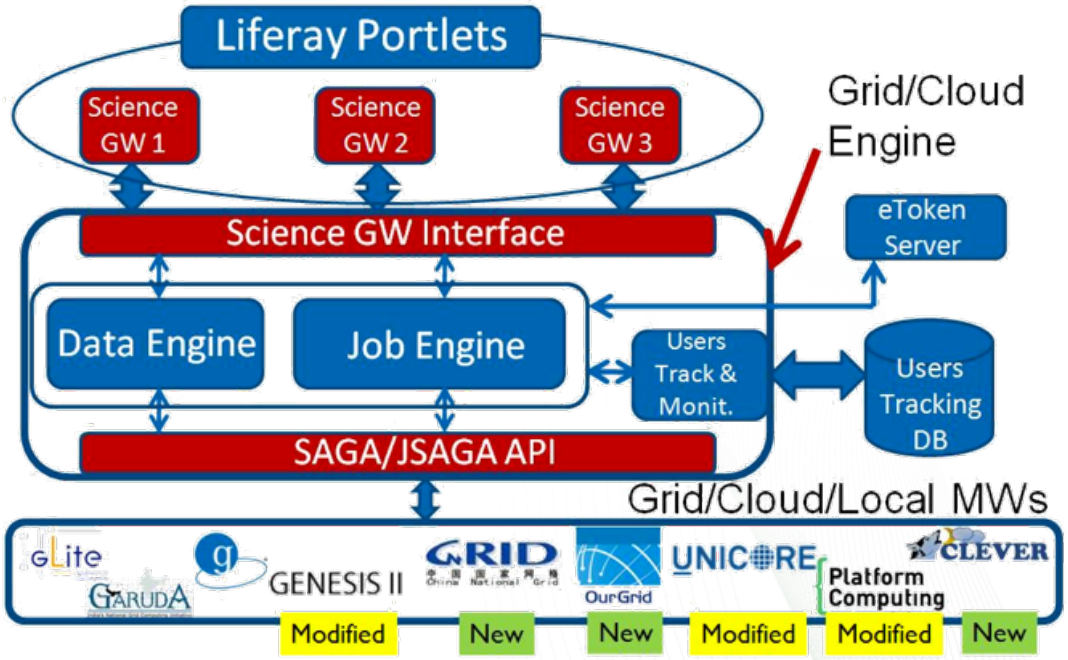


Grid and Cloud Engine

Catania Science Gateway Framework

http://csgf.readthedocs.io/en/latest/grid-and-cloud-engine/docs/index.html

Similar to CrossBOW in the ”Engine” piece



VA Data Transfer – Genomics Research

- Private 10G circuit for data transfer
  - Globus-url-copy between sites (not currently allowed to talk with Globus)
  - Private SDMZ
- Cloud inftrastructure for researchers
  - Big Data – Spark Cluster
  - VMs
  - Science Gateway

Future Work

- “Beef up” ORNL SDMZ infrastructure
- Boost ORNL SDMZ project usage
- Collaboration on SDMZ workflow systems
- Investigate Globus API building blocks and portal integration
- Create abstracted cross infrastructure environment to enable easy workflow automation
- Make Data sharing easy between environments
- Private SDMZ – Medical SDMZ?

Credit to others

Susan Hicks (CADES)
Jason Anderson (OLCF)
Dustin Leverman (OLCF)
Anthony Clodfelter (ARM)
Rob Records (ARM)
Arvind Ramanathan (CrossBOW)