OAK RIDGE National Laboratory

Distributed Computing and Data Ecosystem (DCDE)

Connecting DOE Facilities Together for Seamless Science

Mallikarjun (Arjun) Shankar, Ph.D.
Group Leader, Advanced Data and Workflow, NCCS
CADES Director
Oak Ridge National Laboratory
shankarm@ornl.gov

Co-led with Eric Lancon (BNL)
ASCR PM: Richard Carlson

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

U.S. DEPARTMENT OF ENERGY

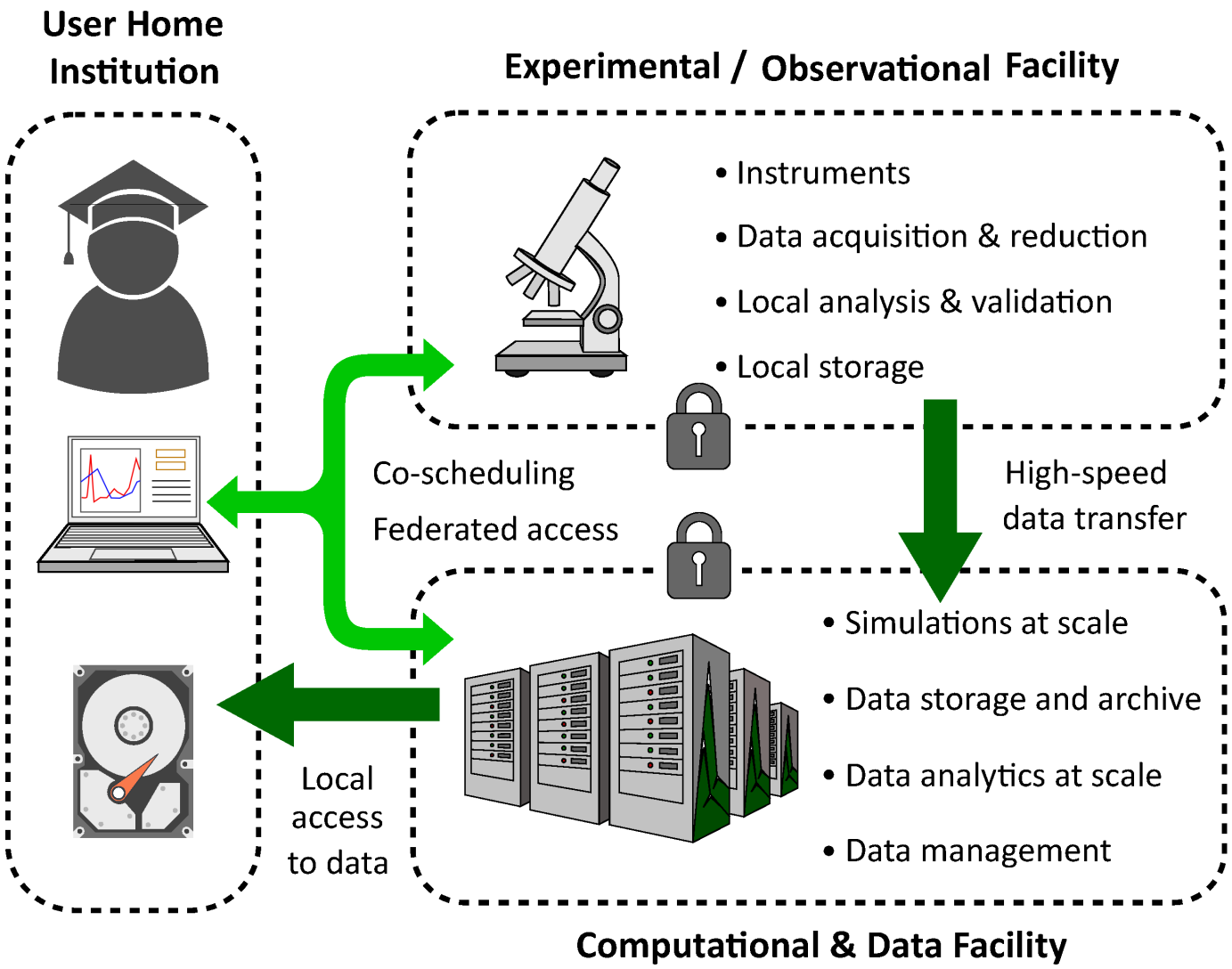
Outline

- Emerging context for DOE science
- Future Laboratory Computing - Working Group
  - DCDE report
  - Pilot project and lessons learned
  - SC19 demo
- Connecting facilities together
  - A focus on federated access management
  - Technical and policy aspects

OAK RIDGE National Laboratory

Emerging Context for DOE Science

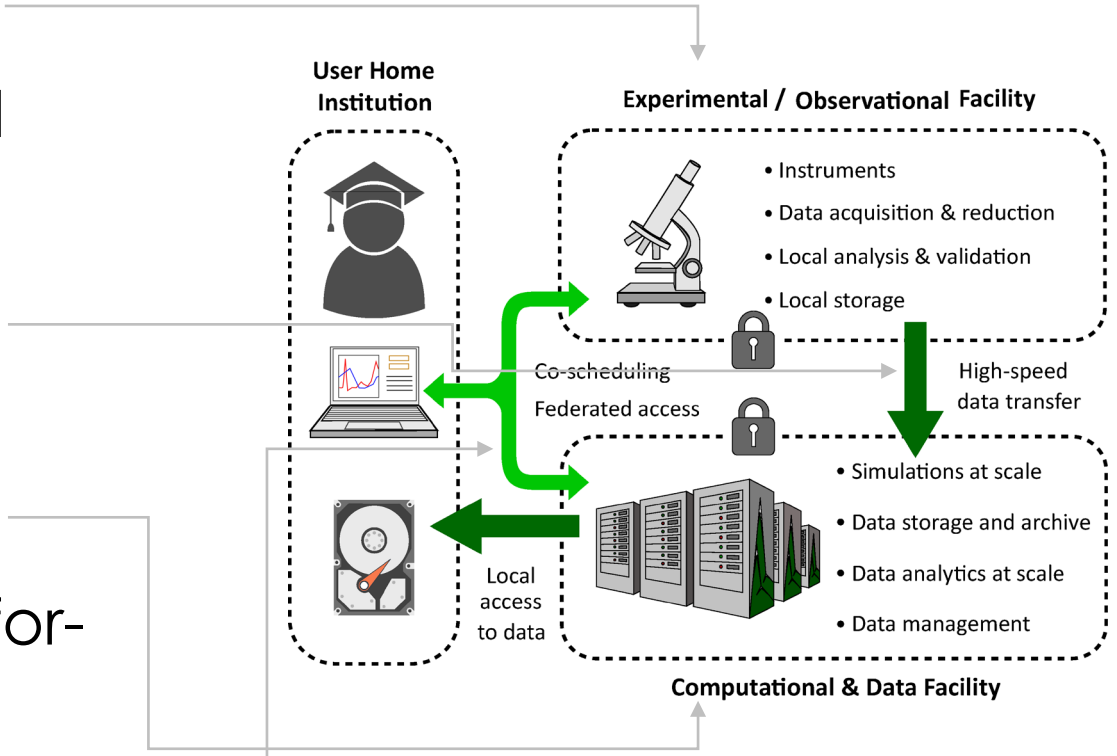
Connecting Facilities: A Cross-Facility Design Pattern



Policy Considerations when Federating Facilities for Experimental and Observational Data Analysis, M. Shankar, et al., *Handbook on Big Data and Machine Learning in the Physical Sciences, 2020, Eds. S. Kalinin and I. Foster, http://doi.org/10.1142/9789811204579_0018*

Policy Considerations

- Experimental/Observational Facility Data Management
  - Metadata Representation, Volumes and Reduction
- Data Movement
  - Streaming, Store and Forward, Staging
- Computing Facility Policies
  - Allocation by scale, domain, hardware-for-application, heterogeneity
- End-to-End
  - User access, portability, co-scheduling, governance



OAK RIDGE National Laboratory

Future Laboratory Computing Working Group (FLC-WG) Activities

Next set of slides include several adapted from DOE/SC/ASCR PM Rich Carlson's presentations to ASCAC (January 2020) and the National Laboratory CIOs (5/7/2020), and the DCDE Pilot Demo @ SC19

FLC-WG Concept and Goals

- ASCR has a long history of conducting research and supporting operations in Middleware, Grid, and higher-level Services to form Distributed Science Infrastructures
- Operation of these infrastructures has been historically been performed by an individual Science domain (i.e., ESG - Climate, LHC – High Energy Particle Physics)
- A Pilot project built upon the success of the Future Lab Computing – Working Group to pilot the use of laboratory resources using a federated Identity service to access those resources
- Federating DOE/SC facilities as they continue to generate, process, analyze, and archive more data will significantly increase the value and usability of those facilities

FLC-WG Initiated in 2017 and Reported Back

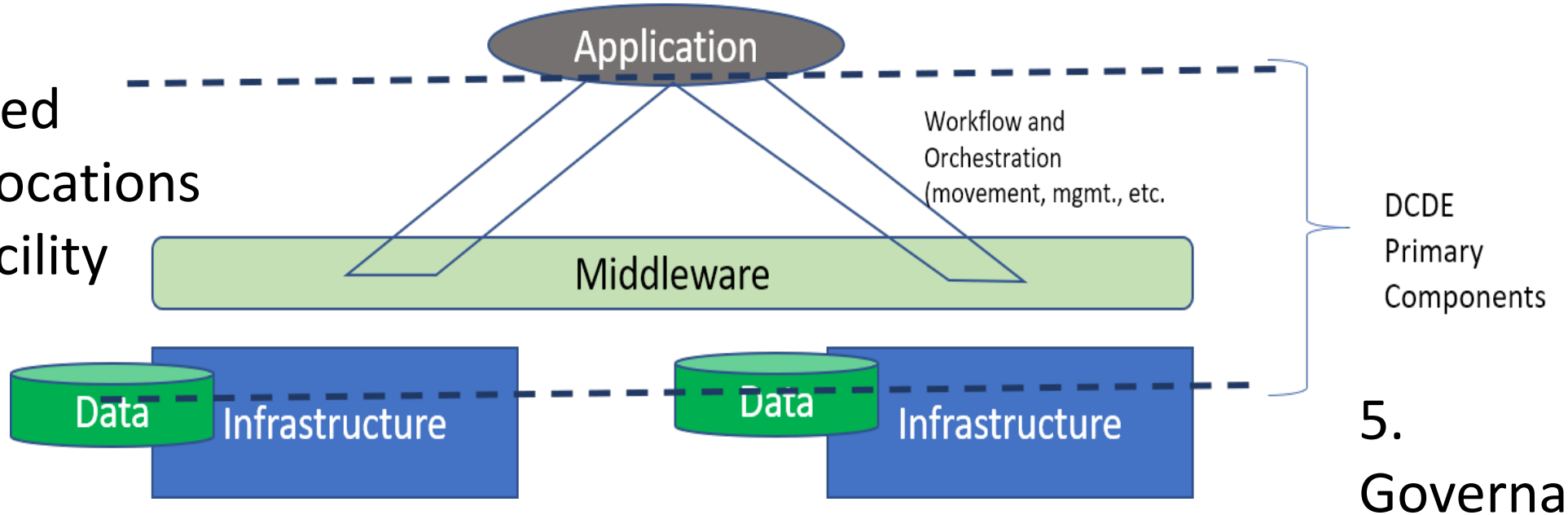
- DOE/SC Laboratories provide computing/storage resources to lab staff, researchers, and visiting scientists
- Demands on these resources are increasing
- Labs have the capability to leverage decades of research to create modern Distributed Computing and Data Ecosystems (DCDE) to meet the current and future demands of DOE scientists
- ASCR constituted Future Laboratory Computing Working Group (FLC-WG). Met through 2018 and delivered report with findings.
- DCDE pilot established for FY2019 fleshes out the key components and documents procedures to establish the infrastructure.

FLC Working group report (2018): ***Background and Roadmap for a Distributed Computing and Data Ecosystem***, https://doi.org/10.2172/1528707

DCDE Components

1. Seamless user access

2. Coordinated Resource allocations and cross-facility workflows



3. Data storage, movement, and dissemination for distributed operations

4. Variety, Portability: through virtualization, and containers, etc.

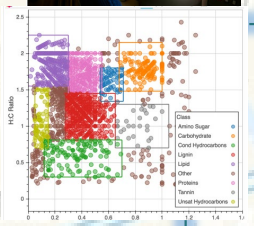
5. Governance and Policy Structures

DCDE Pilot – The Art of the Possible

- Team across ANL, BNL, LBNL, ORNL, and EMSL
  - Goal is to deploy, not develop, existing tools and services
  - Integration with LCRC@ANL, SDCC@BNL, CADES@ORNL, and EMSL@PNNL as domain driver
- Services used:
  - AuthN/AuthZ: InCommon, CILogon and COManage
  - Globus
  - Application and Containers
  - Jupyter notebook and Parsl workflow

Distributed Computing and Data Ecosystem (DCDE) Demo Overview

1. Cowley@PNNL



2. Dong@BNL

CoManage Federated ID Mapping

Jupyter Hub

Parsl Invocation

3. Murphy-Olson@ANL

Jupyter Hub

Parsl Invocation

Jupyter Hub

Parsl Invocation

4. Maheshwari@ORNL

Jupyter Hub

Parsl Invocation

5. Chawla@LBNL

Jupyter Hub

Parsl Invocation

Challenges and Lessons

- Federated IdM remains clunky and a critical challenge
- Firewall and tunneling issues are a recurring obstacle
- HPC access: need to translate identities to run as a user on a unix system
- Workflow tools from notebook interface still need to integrate seamlessly with infrastructure

Pilot to Production

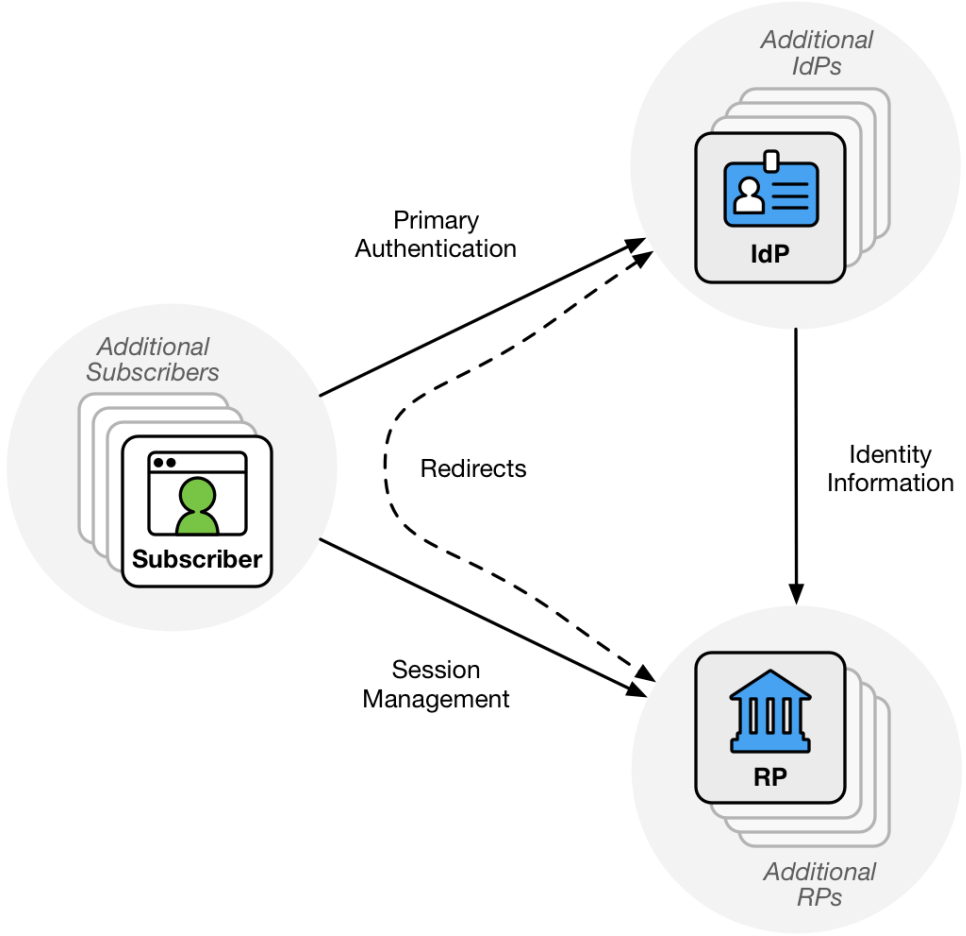
- Goal: leverage the existing lab and facility activities to create a complex wide solution encompassing
- A comprehensive service with commonly agreed upon schemes will allow each resource owner to define the identities and attributes needed to access their physical resource
  - Generate a production level Federated IdM service based on pilot labs
  - Integrate ASCR facilities into this federation
  - Integrate other SC labs into this federation
  - Integrate other SC facilities into this federation
- Resolve open policy issues
  - What attributes are required by a Resource Provider?
  - How will Federated IDs map to local accounts (multiple options)?
- Subsequent service additions: performance tuning, workflows, etc.

OAK RIDGE National Laboratory

Federated Identities across the SC complex

Current activities in the DCDE Team

Federation Design Pattern



Adopting NIST language (refinements of Authn/Authz):

IAL refers to the identity proofing process.

AAL refers to the authentication process.

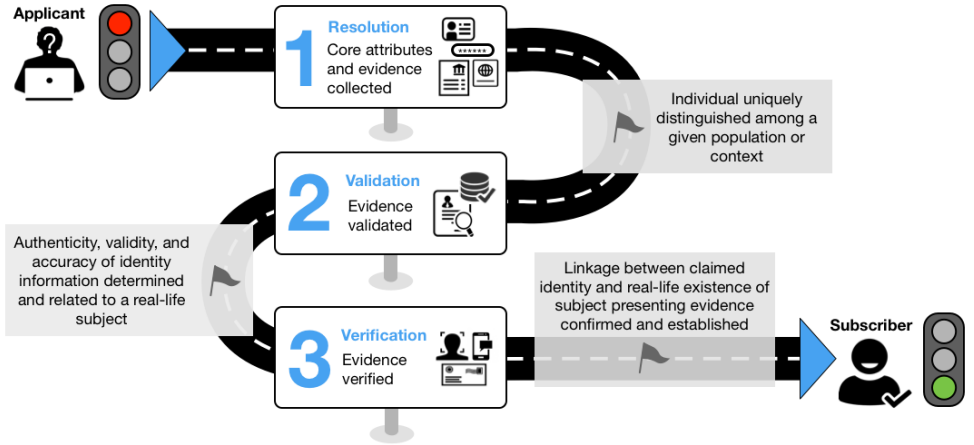
FAL refers to the strength of an assertion in a federated environment, used to communicate authentication and attribute information (if applicable) to a relying party (RP).

https://pages.nist.gov/800-63-3/sp800-63c.html

IAL, AAL, FAL Category Overview

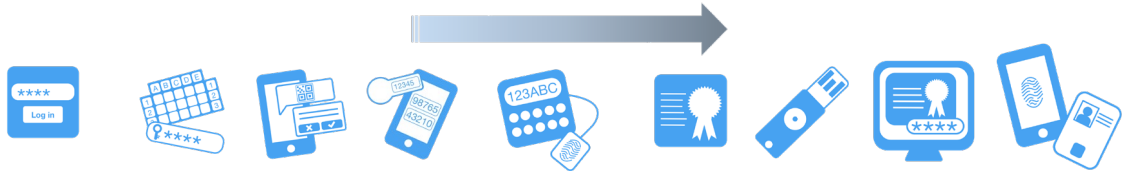
IAL

IAL	Requirement
1	No requirement link to real-life ID
2	Evidence supports the real existence
3	Physical presence is required for identity proofing.



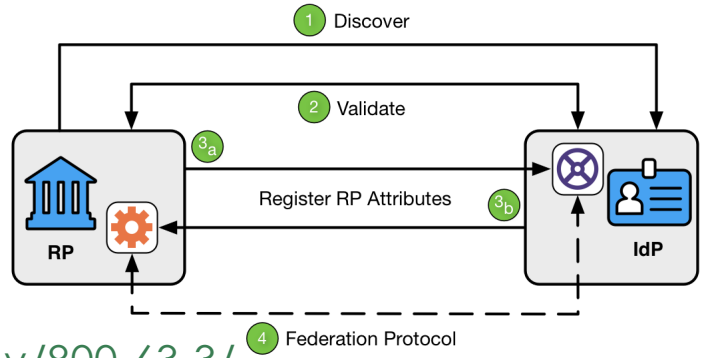
AAL

AAL	Requirement
1	AAL1 provides some assurance that the claimant controls an authenticator bound to the subscriber's account.
2	AAL2 provides high confidence..
3	AAL3 - very high confidence



FAL

FAL	Requirement
1	Bearer assertion, signed by IdP.
2	Bearer assertion, signed by IdP and encrypted to Relying Party (RP).
3	Holder of key assertion, signed by IdP and encrypted to RP.



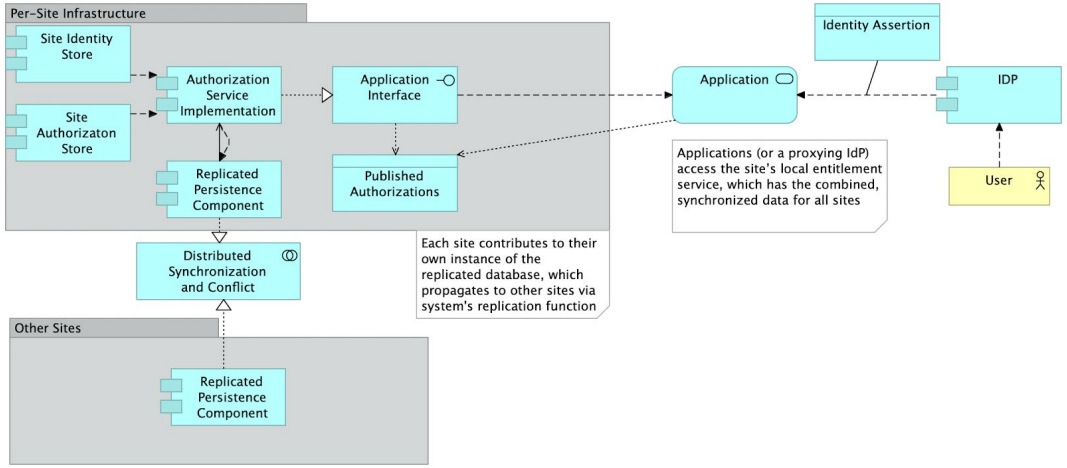
NIST Special Publication 800-63 https://pages.nist.gov/800-63-3/

Addressing the Technical Design

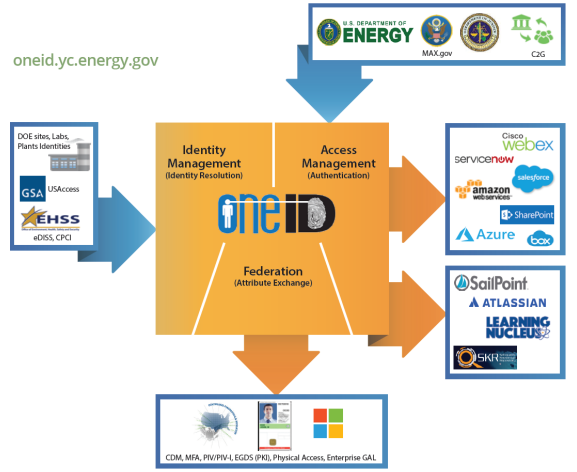
- Information Store Design
  - Central Store: a centrally managed service contains all information (identity and attributes) needed to make decisions. All users and resources query this service.
  - Application Driven Service: each lab maintains an attribute service that maps attributes to identities. Every application queries all lab servers to build a full list of attributes that associate with that identity.
  - Distributed Database: each lab maintains an instance of a distributed database which may be replicated across sites. Queries to any instance will return complete set of attributes for an identity.

  Also influenced by derivatives of AARC Blueprint.

- Exploring DOE OneID approach, including bridging to InCommon, etc.



Example of distributed database concept. Courtesy: Pete Friedman, ANL



LLNL-PRES-773800
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Addressing the Policy Issues: *E Pluribus Unum*

- Attributes
  - Each lab requires multiple attributes acting in effect as a derivative CSP
  - Minimal set of requirements need to be defined
  - Non-lab facility user requirements need to be defined
- Trust zones must do no harm, and allow individual laboratory overrides

Information required to get a computing account for non lab employees	AMES	ANL (LCRC)	ANL(ALCF)	ANL(LAB)	BNL	JL
Site ID number						
Online Cybersecurity training						
Online Computer use agreement						
First Name						
Last Name						
DoB						
Citizenship						
SSN						
email						
Gender						
Business phone						
Home address						
Shipping Address						
Home phone						
Country of Birth						
City of Birth						
Highest degree						
Degree Year						
Affiliation (home institute)						
Position (at home institute)						
Lab host Name						
Lab host Department						
Research type conducted at Lab						
Type of visa (if applicable)						
CV						
Emergency contact						
IALlevel 1						
IALlevel 2						
IAL level 3						

Summary

- Federated Identity Management is a key enabling service to foster scientific discovery
- The DCDE pilot project demonstrated that IdM services are ready for full scale deployment within the DOE/SC lab complex
- While some policy and trust issues need to be resolved, there are significant benefits to creating and using a federated IdM service
- DCDE is developing a design document that can be used to implement a SC wide federated IdM service

Acknowledgements: DOE/ASCR PM: Rich Carlson
DCDE Team: R. Adamson, A. Adesanya, W. Allcock, M. Altunay, R. Bair, D. Cowley, M. Day, S. Dong, D. Dykstra, P. Friedman, S. Fuess, K. Heffner, B. Holzman, K. Hulsebus, M. Karasawa, E. Lancon, B. Lawrence, S. Maerz, E. Moxley, J. Neel, D. Murphy-Olson, K. Maheshwari, A. Shankar, C. Snavely, T. Throwe