Networking and Information Technology
Research and Development

The government seeks individual input; attendees/participants may provide individual advice only.

Middleware and Grid Interagency Coordination (MAGIC) Meeting Minutes[1]
July 1, 2020, 12-2 pm ET

Virtual

Participants

Valentine G. Anantharaj (ORNL)	Miron Livny (UW-Madison)
Kathy Austin (TAM)	David Martin (ANL)
Eric Burger (OSTP)	Deep Medhi (NSF)
Devin Casey (NARA)	Tom Morton (DoD CIO)
Richard Carlson (DOE/SC)	Michael Nelson (Carnegie)
Dhruva Chakravorty (TAMU)	Mike Norman (UCSD)
Michael Corn (UCSD)	Donald Petravic (NCSA)
Ewa Deelman (ISI)	Steve Petruzza (Utah)
Martin Doczkat (FCC)	Arcot Rajasekar (UNC)
Sharon Broude Geva (UMich)	Stefan Robila (NSF/CISE/OAC)
Joanna Grama (Vantage)	Birali Runesha (UChicago)
Dan Gunter (LBL)	Arjun Shankar (ORNL)
Chin Guok (ESnet)	Alan Sill (TTU/OGF)
Margaret Johnson (NCSA)	Valerie Virta (NIH/CIT)
Joyce Lee (NCO)	Sean Wilkinson (ORNL)
Jason Lopez (Xodiak)	

Proceedings

This meeting was chaired by Richard Carlson (DOE/SC) and Stefan Robila (NSF).

Michael L. Norman, UC San Diego, Cloudbank: Managed Services to Simplify Cloud Access for Computer Science Research and Education-

NSF funded (5M/5years); Set of managed services to simplify cloud access. Architected to scale up and scale out; currently, opt in. Joint project among UCSD, UWashington, UC Berkeley

Project at intersection of stakeholders

- NSF- entity that would facilitate cloud use for CISE researchers; Cloud resources must be IDC free
- CISE researchers and educators – help educate about cloud and with preparing NSF proposal
- Cloud providers: want to be paid by NSF, not the converse

Interviewed researchers: address pain points of researchers and educators using public cloud

[1] Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program.

- Issues: time and hassle; account and financial management issues
- Intended to be like commercial institution that deals with funds on deposit to spend on public cloud
- How address – managed services (e.g. consulting, onboarding/set up, tools, unified reporting)
    - Cloud agnostic approach- multi-cloud cost monitoring and optimization tool integrated into portal – provides back end allowing PI with multiple clouds/accounts to view through dashboard
    - Engage with cloud reseller (Strategic Blue) – bundles small requests into bulk request to cloud providers and pass on to researchers
- Cloudbank workflow- cloud funds (for cloud resources) to CloudBank, which renders services to PI (provisions PI account and access to PI' research group). Begin research in cloud
    - Launching August 1 – support for AWS, Microsoft Azure, Google, and IBM later)
    - Researchers' accounts in selected cloud
    - Cloud usage reported to CloudBank (consumed by portal) and to Strategic Blue (financial back office)
    - Funds stay in CloudBank and paid out on usage basis; can optimize cloud spend and provide flexibility for researcher (in case change clouds during project)
    - Discussion – CloudBank has 750k in funds to purchase resources for NSF researchers

Initial focus of CloudBank: CISE Solicitations

1) Smart and Connected Communities and 2) Cyber-Physical Systems

- If receive award,
- Designed to scale up
- More funds can be put into bank

3 main efforts:

1) user Portal, (UCSD)
2) Education & training (UWashington with UC Berkeley)
3) Financial management – Strategic Blue and UCSD

Team: External advisory board, cloud providers, NSF, UCSD/SDSC, UWashington, UC Berkeley, Strategic Blue

Portal

- Friction removed from NSF program officers, PIs.
- Where users learn about cloud, receive proposal preparation assistance and account management
- Cloud Resource Catalog – service categories
- Provider tools – pre-built platform
- 3 tier model for user support – 1) basic help with portal, 2) Proposal support, research, 3) deeper engagements

Timeframe

Year 1 – near end

August 1, 2020 – On track for transition to production operations. Early user testing underway.

- Will scale up More NSF programs want to participate

Future CloudBank Center of Excellence – community-based; sharing best practices and leverage cloud providers' organized events and existing programs (XSEDE, Campus Champions, etc) www.cloudbank.org

Discussion

- If researcher's institution doesn't have BAA, etc. Initially, UCSD affiliate. Potential future
- CloudBank owns accounts; performs continuous cloud monitoring
- Cloud providers' loose credentialing procedure (e.g., stores credentials in unencrypted format)
-

Valerie Virta, NIH STRIDES (Science & Technology Research Infrastructure for Discovery, Experimentation and Sustainability) Initiative

Background

Lower barriers to state of the art data storage and computational capabilities, including training and education (to access AI/ML). Current partners: AWS and Google

Implementation of NIH strategic plan for data science.

- Constructing data infra., modernizing data ecosystem, improving data management, analytic and tools; workforce development and ensuring stewardship and sustainability of big data resources (esp. on Cloud).
- Benefits
    - To researchers : favorable pricing and access to cloud and professional services/training
        - Available, scalable, pay as go, growing tools/services, agile workflows to test prototypes
        - Consistent and centralized NIH identities and login credentials
        - Difficulties when NIH institutes and centers (IC) wish to use cloud (acquisition vehicles, budgeting, growing prototype capabilities). Number and maturity of projects vary by IC
        - STRIDES offering standardized approach has raised more needs and opportunities for managing NIH systems and data
    - To NIH:
        - centralized cloud environment provides more data protections/ security
        - connect multiple projects and collaborations;
        - granular reporting on usage and spend
        - dedicated secure connectivity for data/file transfer
        -

Training critical – on cloud platforms (including technical/infra and applied scientific training) for decisionmakers, system engineers, researchers, etc, can lead to certifications

Participants: 18 NIH ICs, 148 programs, 230 accounts. 71PB stored. >20M compute hours. $40M contributed

Onboarding

Intramural: AWS fully on boarded programs (122); Google fully on boarded programs (65)
Extramural: AWS enrolled institutions (14); Google enrolled institutions (8)

Operationalizing Cloud for NIH-Managed Data

- Goal: Develop robust cloud capabilities with standardized cybersecurity controls and extensive authority to operate (end of this month).
- Centralized NIH research cloud environment important otherwise, multiple ICs would have to create own ATOs –not necessarily align.

- NIH-wide ATO can be customized for different projects, but still have security controls. Create optimized cloud environments in partnership with cloud vendors; consistent approach (IAM services and federated login)
- NIH-Managed Cloud Environments – Service layers
    - Vendors: manage compute, storage, networking, databases
    - NIH manages Shared tech services:- IAM, automation of processes (account), reusable templates, security controls, high-speed network connections
    - Cloud infra and core services: choose specific cloud services, operating system, application, configure system, provide own data
    - Overall architecture: NIH (architecture standards, SSO, connection to cloud, security, construction of ATO on overall environment)

Funding Model

- NIH (Internal): ICs funds via STRIDEs through Other Transaction Agreements (OTA) – outside of grants, cooperative agreements and contracts. Requires justification
- Centrally funded: shared technology service costs
- Extramural: for STRIDES eligibility, must be NIH-funded

Success Stories for a Future of interconnected Data Sets

Folks working within given projects can pull data from multiple efforts. Success stories:

- Sequence Read Archive (NLM National Center of Biotechnology information for the cloud). Largest accessible dataset in world on cloud
- Kids First – Data Resource Portal of clinical and genetic data – almost completed; creating fire-compatible space
- PII-Secured AWS Computing Environment (PACE) –leverages AWS cloud; stores sensitive data and uses ML in collaboration with NLM researchers. Protected data used for scientific research

Discussion

- Cost: modest discount, but still costly. How to reduce costs – Multiple strategies. When cost of cloud compared to on premise HPC, but growing digital divide. Many schools lack access to in premise environment. Research software engineering environment – gov program addressing high cost issue; fund program to address it. Bring costs in-house -More funding and development of in house workforce
- NSF CC Star (CC*) program – build capabilities on campuses to reduce dependence on outside help. Look to CC*

Devin J. Casey, Implementation Lead, CUI Oversight, NARA, Controlled Unclassified Information Program

Information security program (gov not doing well):

1) Prevent unauthorized access to information (factor: lack of federal standard)
2) Facilitate access to those who need access for lawful government purpose (tended to overprotect block sharing among agencies and with nonfederal agencies)

Define CUI:

- Defined sensitive, unclassified information needing protection: laws, regulations or gov. wide polices calling for the protection of the information
- Re-assigned responsibility to appropriate entities within an agency.

- Ensure protected information is appropriately marked.
- Ww.archives.gov/cui

CUI Registry: categories of CUI (e.g., privacy, tax, law enforcement, intelligence, financial, export control)

Information security reform – agencies already protecting information.

- Standardizes markings (adopt CUI marking and clarifies what to protect.
- Standardizes protections/physical security requirements (controlled environment and locking barrier)
- Physical destruction (blend of existing policies and practices; multi-step process)
- Minimum security baseline for information
- Promotes inter-agency sharing in timely manner; CUI information protected by program meeting CUI standards

Defines what and how we protect information for executive branch (32 CFR 2002)

- CUI Basic – not require specific protections
- CUI Specified – specific protections (unique markings, enhanced physical safeguards, limits on who can access information)

Federal Acquisition Regulation (FAR (FY19) Case for CUI – draft phase; will come out for public comment

- Work with GSA to create CUI FAR case; standardize how CUI applies to non-Fed entities
- Not required for nonexecutive branch unless required by law/regulation or included in contract
- Based on existing DFAR 7012 case – but goes further in discussing training
- Create standard form for all unclassified, but sensitive information (from marking, handling to cybersecurity)
- CUI quarterly stakeholder update (sign up through CUI blog https://isoo.blogs.archives.gov)- followed by adhoc stakeholder meetings to answer questions about case; increase understanding

Legacy Information and Markings

- Legacy information not automatically CUI; agencies determine what legacy information qualifies as CUI. Protect pursuant to existing or previous contract.

Implementation Process – has begun

- DoD using some CUI requirements; CUI practices and legacy practices will exist at the same time.

CUI requirements have not yet been mandated. Falls on entity with contractor agreement with gov to ensure compliance with CUI requirements (where Cui being create, store, process or transmit).

Controlled Environments – Limits access to authorized holders

- Assessing physical environments – review and understand controlled environment training on CUI website
- Assess electronic environments – establish electronic barriers to limit access; need to review frequently

DoD implementation and related efforts: links: https: www.acq.osd.mil/cmmc/ or osd.pentagon.ousd-intel-sec.mbx.dod-cui@mail.mil. www.Dodprocurementtoolbox.com

Discussion

- FedRamp moderate or equivalent
- CUI marking requirement – more flexible than classified; bulk marking (cover); banner markings/click through, User access agreements/training.
- For complex environments, came up with a standard for metadata marking for CUI
- Entity in contract with government, is responsible
- Executive Branch – information security reform in next year. Once policies are published, program agendas follow.
    - Annual reporting – most agencies (pre-COVID) predicted publishing their policies prior to end of CY20.
    - Then agencies train employees and then information marked as CUI. Most agencies will wait until NARA publish FAR case to help standardize CUI for nonfederal entities. Will then see industry using CUI markings when performing information protection or creation in support of gov. entities.
    - FAR case was to come out for public comment a month ago. GSA will publish unified agenda (NARA blog post will come out) – timeline and NARA will do webinar.

Potential speakers

Cloud efficiency; network capacity, workforce development, national research cloud and cloud economy (workflows, jobs in cloud vs. inprem; develop continuum of campus computing or national computing)

- Kevin Thompson (NSF OAH: CC* program)–campuses funded to develop own campus capabilities, so not have to depend on outside help. Agencies' activities to support research community.
- Currently, sorting out what cloud, national facilities and campus facilities are good for.
- How labs are handling CUI
- Workforce Development – open topic;
    - OAC/NSF - significant investments in workforce development. CaRC- much activity in CI workforce development. Turning cloud resources into portable capabilities (Miron)
- National Research Cloud – what thinking, where going and impact on NSF, NIH and DoD thinking (Dhruva)
- Cloud economy – spot market interesting. How is software being developed for AI/ML to leverage spot pricing? Economies can be leveraged from software and user perspective (Dhruva)
- If seeing push to facilitating institutions making use of cloud (Commercial and non-commercial). How see network capacity issue; bigger infrastructure question? (Sharon). Very broad issue, but will arise as go more into the cloud. Kevin Thompson?
- software approaches – how is ML being thought about; leverage strengths of cloud.
- Not just optimizing cloud, but best practices for delivering onprem resources. Alan Sill (TACC speaker on cost-savings measures)
- HPC – optimizing hybrid environments
- JET – engage with JET on network access activities

Next meeting: August 5 (12 pm ET). Discuss tasking – take topics and determine strategy for next year; continuum of computing; multi-month and single months.