DataNet Federation Consortium

Challenges and Mitigation,
Future Plans and Activities
(including Milestones and Timelines)

Reagan Moore and Mary Whitton
UNC-CH

National Science Foundation Cooperative Agreement: OCI-0940841

DFC DataNet FEDERATION CONSORTIUM

Topics

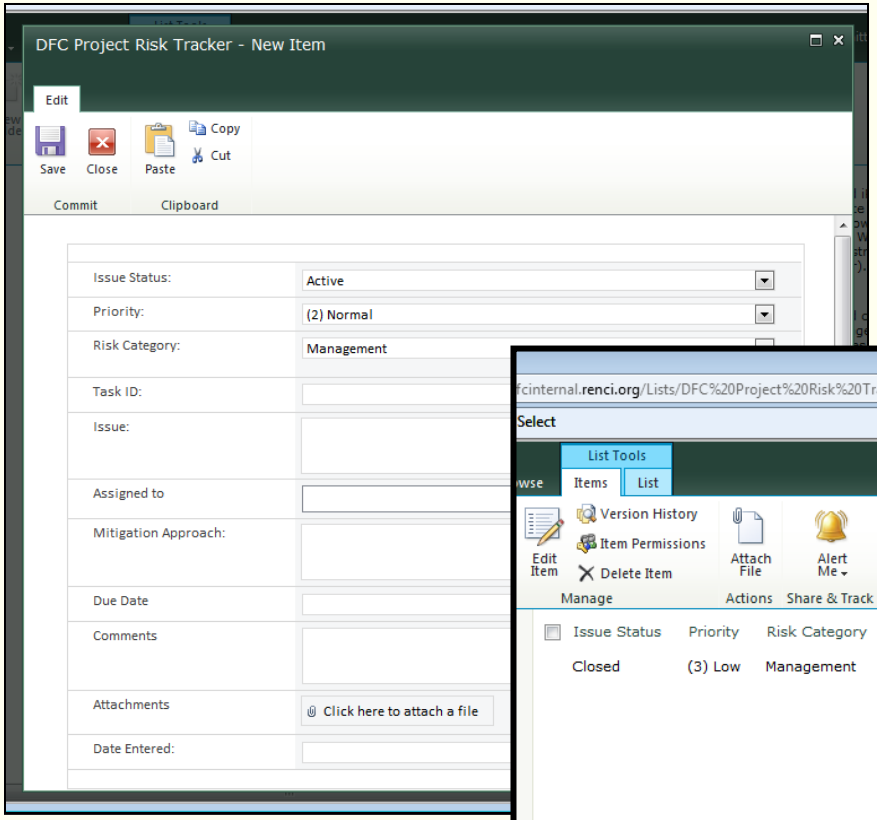
- **Challenges, Risks, and Mitigation**
  - Managerial
  - Technical
- **Technology Futures**
- **Milestones and Timelines**
  - 3 science collaborators
  - 3 communities of practice
  - Demonstration

DFC DataNet FEDERATION CONSORTIUM

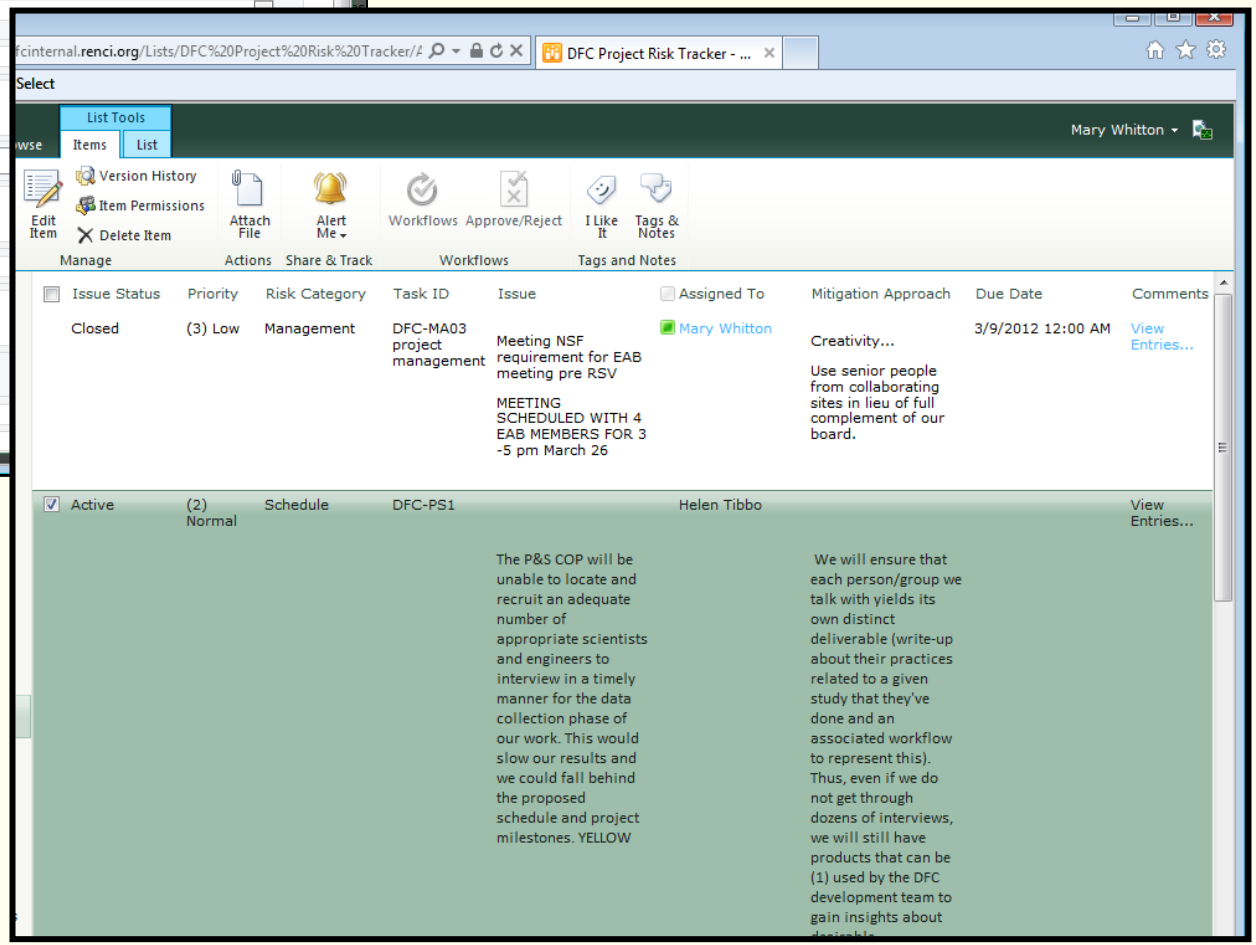
Risk Tracking

- **Project Manager is in charge of tracking**
- **Risk list reviewed at weekly management team meetings**
  - High risk areas-weekly
  - Moderate and low—bi-weekly
- **Risks entered and tracked in SharePoint**

SharePoint Risk Entry Form



SharePoint Risk Report



Managerial Risks (1)

- **PI no longer available**
  - Executive Committee proposes new PI through Change Process
  - Vice Chancellor for Research chooses new PI as part of UNC institutional commitment and forwards recommendation to NSF for approval
- **Project Manager no longer available**
  - UNC-CH advertises for and hires new person
  - Interim support from other RENCI personnel

Managerial Risks (2)

- **Disagreement between collaborators**
  - Procedures for conflict resolution
    - Executive Committee : PI : UNC-CH Office of Sponsored Research
- **Sub-contractor lack of progress**
  - Procedures for identifying causes and remediation
    - Project Manager : Executive Committee : PI : UNC-CH OSR
  - Revisions to statement of work, deliverables, and budget
    - Change process. Executive Committee : UNC-CH : NSF Program Manager

DFC DataNet FEDERATION CONSORTIUM

Managerial and Technical Risk: Too few experienced personnel

- **Expertise transfer to next generation**
  - Risk: Replacement of expert developers (NSF SDCI)
  - Risk: Acquisition of production level expertise for E-iRODS development (RENCI)
  - Risk: Steep learning curve for new hires
  - Risk: Growth of open source community expertise
  - Management(1): Schedule a series of multi-day SW architecture and code reviews
  - Management(2): Work with collaborating groups on tutorials, documentation
  - Management(3): Collaborate on third book in a series on use of Policy-based Data Management (iRODS)

Technical Risk: Dependencies (1)

- **DFC technology dependent on availability of continually advancing iRODS software**

  NSF grant OCI-1032732 "SDCI Data Improvement: Improvement and Sustainability of iRODS Data Grid Software for Multi-Disciplinary Community Driven Application" runs through November 2013

  – Risk: Open Source community will not be not large enough to advance iRODS software without availability of funded developers
  – Management: Grow the developer community by selling them on benefits of DFC and of their participation in open-source development community
  – Management: Continue to seek grant or contract funding for base iRODS development

Technical Risk: Dependencies (2)

- **Federation with National Climatic Data Center: Hydrology and Marine**
  - Risk: Establishment of mechanisms to deposit and access climate data records at NOAA center
  - Management: Bring issue to NSF's attention early for possible MOU between NOAA and NSF on data sharing
  - Management: Support meetings and tutorials at NCDC on data grid technology

Technical Risks Related to Features

- **Risk: (Resources) Unable to integrate new data resources**
  - Mitigation: Write wrappers for integration using multiple back-end access mechanism
- **Risk: (Users) New user authentication is needed**
  - Mitigation: Add new authentication to iRODS PAM module
- **Risk: (Methods) Unable to integrate a new function or method**
  - Mitigation: Extend web-services oriented access for new functions
- **Risk: (Workflow) Workflow in non-supported language**
  - Mitigation: Extend iRODS to mediate with new workflow system (using Workflow Orchestration Server)
- **Risk: (Policies) Unable to formulate policy in iRODS R language**
  - Mitigation: Extend cross-language micro-services to new policy languages (as done for Python)
- **Risk: (Federation) Unable to access another data architecture**
  - Mitigation: Write Soft-link extensions and Resource Driver Extensions

Technical Risk: Security (PEP sec. 10)

- Security audits by the University of Wisconsin. Participate in an upcoming security meeting at University of Wisconsin
- Coordinate a security training session at UNC-CH by Barton's group
- Submit a supplemental request to NSF under the SDCI project for extended analyses
- Use the theorem proving code developed by Hao Xu to verify assertions about the code. An example is the handling of symbolic links within the rule engine

Technical Risk: Transition to Production System

- **Release of Production software**
  - Risk: User communities with mission critical data management needs will not adopt academic software so DFC market penetration is low.
  - Risk: Adoption of academic software outside of DFC and academic domain is low, increasing the challenge for sustaining the software.
  - Risk: User communities with less technical skills and IT support will not adopt academic software so DFC market penetration is low.

Technical Risk: Transition to Production System

- **Release of Production software**
  - Management 1: Work with RENCI to transfer technology and secure development funding for Enterprise iRODS (E-iRODS) to ensure timely release
  - Management 2: Work with RENCI to develop a iRODS to E-iRODS transition plan and risk analysis that ensures DFC needs are met by E-iRODS releases
  - Management 3: Incorporate E-iRODS materials in DFC outreach, marketing, and tutorial materials

E-iRODS: Production version of iRODS

- High Availability, Reliability, Accessibility, Security, Scalability, and Modularity
- Targeted at academic, federal, and industry users
- Support and training packages
- Binary and source code releases
- Compatibility with iRODS
- Open-source, commercialization requires licensing

DFC DataNet FEDERATION CONSORTIUM

E-iRODS: Key Aspects

- **Focus on Code Hardening & Testing**
  - – Defensive Programming , Static Analysis Tools, Peer code reviews, Refactoring of code, Continuous Testing, System Level Testing
- **Refactoring to increase Modularity**
  - –Pluggable core components to allow user customization of capabilities without re-compilation.
- **Additional documentation and support**
  - – *iRODS Server installation packages* for supported platforms: RPM, DEB, MSI, etc.
  - –remote ICAT configuration, automated MySQL installation
  - –*Comprehensive offline Administration Guide*
  - –*Configuration Cookbook*: Rule configurations for well known use cases

DFC DataNet FEDERATION CONSORTIUM

E-iRODS 3.0 Beta: March 1, 2012

- Initial release based on iRODS 3.0
- Hardened binary release of iRODS
  - Passes continuous integration with back-ported bug fixes from community trunk
  - Packaging and signing: initially RPM and DEB
- Future releases
  - Tracks community code
  - Integration of subsequent releases done after code review, refactoring, analysis of impact, and test suite updates.
  - Agile product backlog with E-iRODS team review determines release set
  - Less frequent release schedule
- Certification
- Documentation
- Subscription Support Contracts

DFC DataNet FEDERATION CONSORTIUM

Programmatic Sustainability

- **Programmatic sustainability**
  - Interoperability between major NSF initiatives (DataNet, EarthCube, XSEDE)
  - Interoperability between national grids (OSG, ESG, DFC, SEAD, DataOne, Terrapop)
  - Interoperability between federal agencies (NOAO-CLASS, NASA-NCCS, EPA, …)
  - Interoperability for international collaborations (EUDAT, Max Planck Institute, International Neuroinformatics Coordinating Facility, CoopeUs)

Technical Sustainability (1)

- **Technical Sustainability through technology transfer**
  - Enterprise version of iRODS, E-iRODS, from RENCI
  - Integration of DFC technology with DDN storage controllers
    - Integrate Processing with storage
    - Enables automated feature detection
  - Vendor supported systems integration
    - Distributed-Bio development of genomics data grids

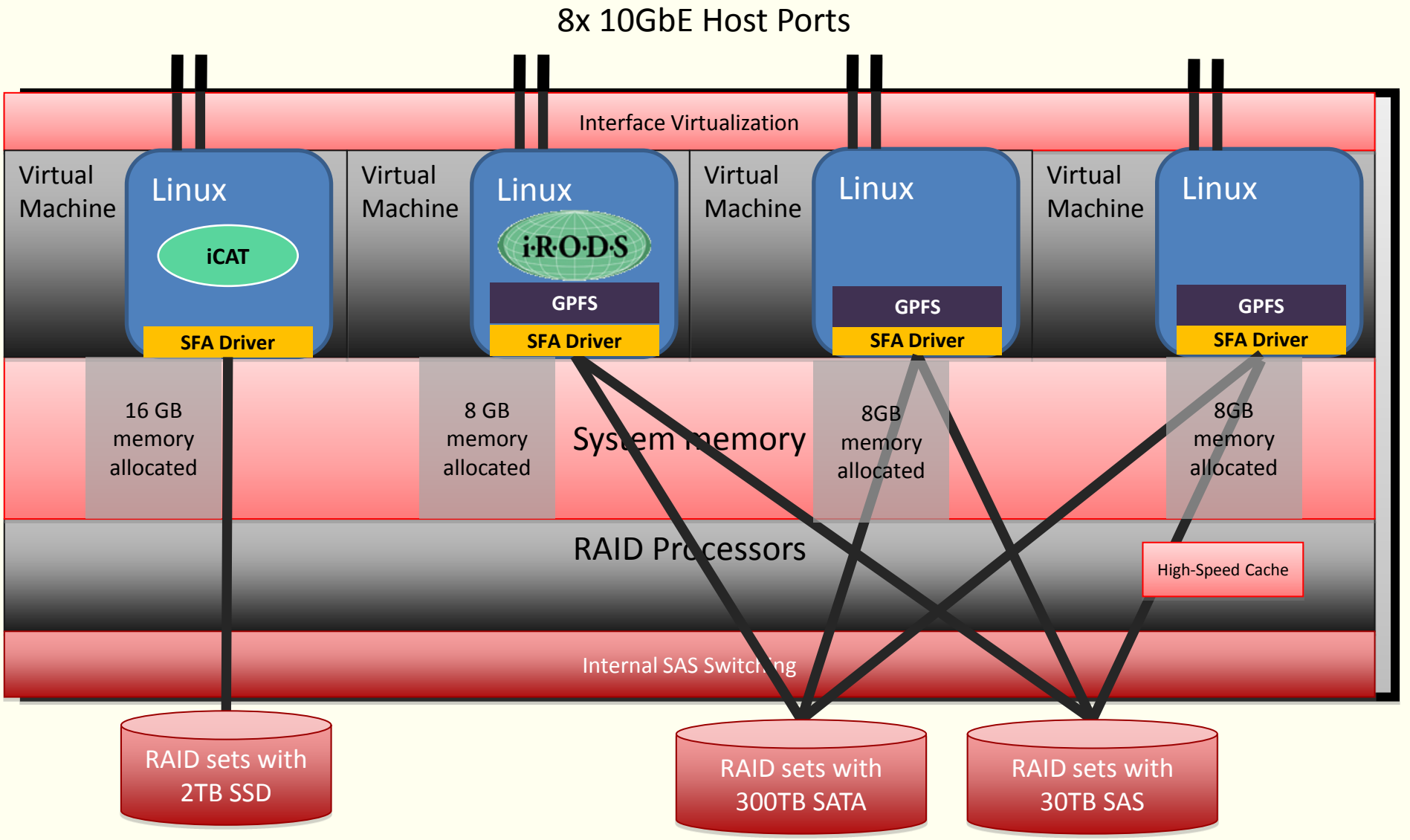
Technology Futures (1)

- **Incorporation of new technology in iRODS**
  - E.g., upgrade from Generic Security Service API for authentication to Pluggable Authentication Modules
  - Use of BOOST libraries to enable porting to Windows
  - Via E-iRODS contributions to iRODS, modern software management environment (git, gforge, hudson, maven)
- **Evolution of collections**
  - Collection life-cycle stages
  - Evolution of data management regulations

Technology Futures (2)

- **Creation of scalable data management system**
  - Automation of management policies within storage controller
  - As add storage, add management enforcement
- **Automation of information extraction**
  - Feature extraction based on policies within storage controller
  - Automated indexing of data based on features

Processing Embedded with Storage (DDN)



Technology Futures (3)

- **Characterization of domain knowledge**
  - Preserve workflows that implement domain analyses
  - Build reference procedures that transform data into information (feature detection)
- **Unified representation of data and analysis procedures (OOI initiative)**
  - Characterization of data through field representations (interpolation functions in space and time)
  - Characterization of analysis procedures based on manipulation of field representations

Milestones and Timelines

Milestones: Hydrology-SC and Inst. Env.

		Quarter							
		1	2	3	4	5	6	7	8
1	✓ Install and verify transfer to/from iRODS grid	X	X						
2	Load grid: existing pre-processed data; new raw data			X	X	X	X	X	X
3	Develop iRODS microservice that runs VIC model				X				
4	Verify VIC results run within iRODS					X			
5	Verify link to NCDC with data transfers				X	X	X		
6	Develop iRODS workflow for pre-processing data: own and NCDC precipitation data (up to 10 kinds)					X	X		
7	Run VIC over different precipitation models and compare to each other and other prediction sources						X	X	X
8	Develop workflows for pre-processing other data types: min and max air temperature, wind speed								
9	Via DFC, access data sources such as NASA								
10	Share high res model results with other communities: drought, river discharge, soils								

Milestones: UCSD-OOI

		Quarter							
		1	2	3	4	5	6	7	8
1	✓ Federate OOI's data grid with dfcmain hub (public data)			X					
2	Demonstrate interoperability between authentication methods				X				
3	Demonstrate data transfer to (an arm of) NOAA					X			
4	DFC transfer data from OOI's integrated ocean network using OOI protocol						X		
5	Procedures and test of stream representation of data						X		
6									
7									
8									
9									
10									

Milestones: Drexel-Engineering

		Quarter							
		1	2	3	4	5	6	7	8
1	✓ Install data grid and basic data management policies		X	X					
2	Survey engineering requirements for CIBER-U	X	X	X					
3	Design engineering repository based on DFC			X	X				
4	Integrate MediaWiki interface for access to engineering repository				X	X			
5	Integrate format registry for engineering data				X	X	X		
6	Automate format transformations					X	X		
7	Develop and demonstrate engineering repository						X	X	
8									
9									
10									

Milestones: Technology Development

		Quarter							
		1	2	3	4	5	6	7	8
1	Survey partners for software requirements	X	X	X					
2	Deploy and Test DFC Federation Hub Prototype		X	X	X	X	X		
3	Federate DFC Hub with Science & Engineering Partners		X						
4	Integrate domain specific systems (NetCDF, OOI API)			X		X	X		
5	Develop interoperability mechanisms (Authentication)			X		X			
6	Design and develop Social Media DataBook			X	X	X	X		
7	Develop workflow support & apply to domains				X	X	X		
8	Demonstrate infrastructure							X	
9									
10									

Milestones: Policies and Standards		Quarter							
		1	2	3	4	5	6	7	8
1	✓ Publish an extensive review of the related literature	X	X						
2	✓ IRB for interviews submitted and approved		X						
3	Conduct Initial Interviews of DFC “water” scientists		X	X	X				
4	Identify and interview 10 non-DFC “water” scientists (2 hours/per interview); analyze interviews for common workflows and policies			X	X	X			
5	Test common workflow w/DFC community & iterate				X	X	X	X	X
6	Test elicited policies w/DFC community & iterate				X	X	X	X	X
7	Report/Share/publish results of “water” workflows				X	X	X	X	
8	Repeat 4-8 above, but for non-DFC Engineering			X	X	X			
9	Report analyzing Hydrology vs. Engineering, publish					X	X	X	X
10	Comparing, analyzing, and writing up analysis of the policies to international policy standards						X	X	X
11	Taking the Grand Challenges and design policies from the challenges, combining lessons learned from 1-10.						X	X	X

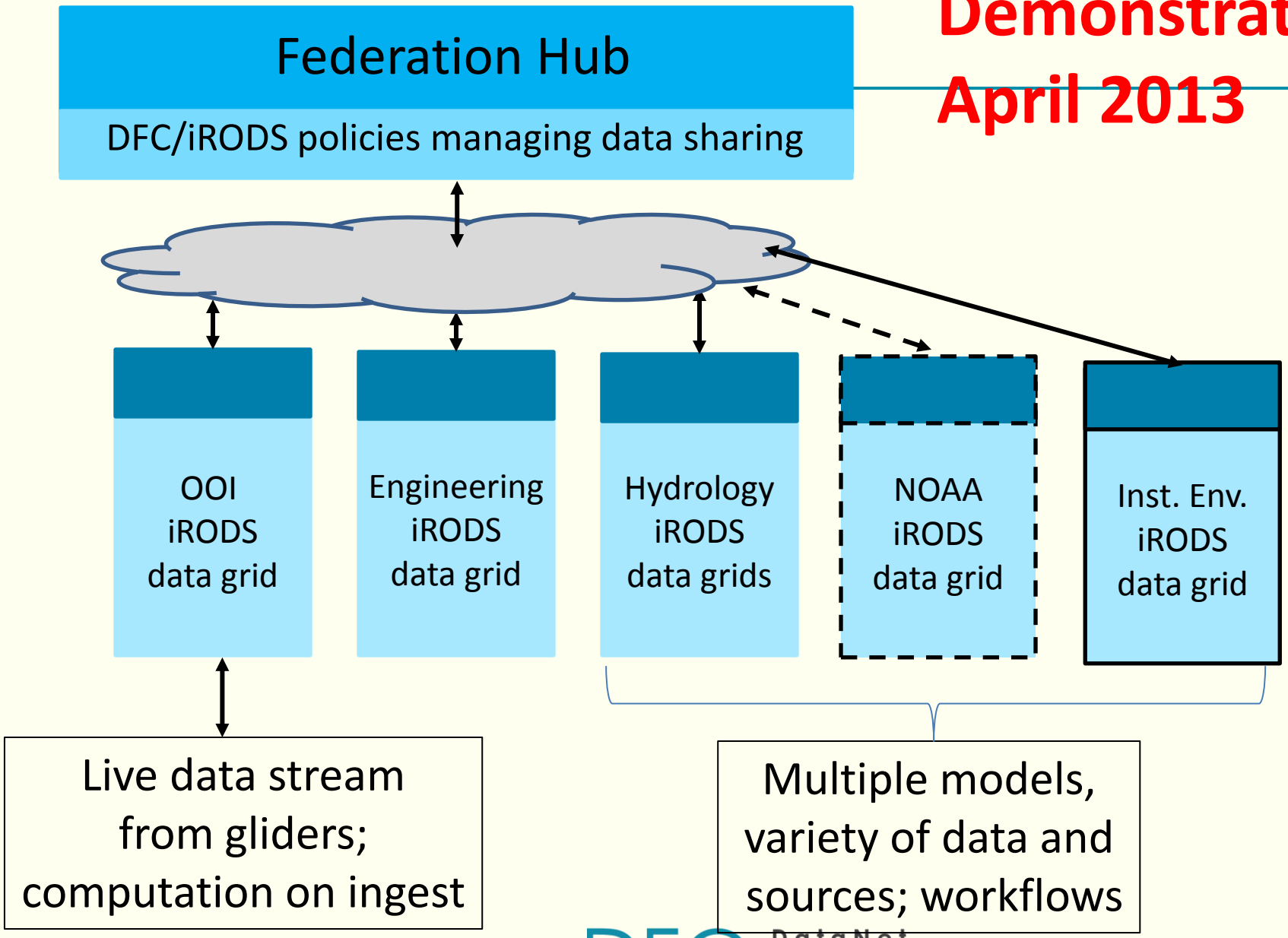
Milestones: Facilities and Operations

		Quarter							
		1	2	3	4	5	6	7	8
1	✓ Help set up iRODS grids at HE&M sites	X	X						
2	✓ Establish two DFC hubs at RENCI	X	X						
3	Help implement and verify federation of sites		X	X					
4	Coordinate installation of new iRODS releases			X		X		X	
5	Formalize system management processes: change management, system status			X					
6	Establish gForge project to track issues and bugs			X					
7	Launch email help list; compile responses into FAQ			X	x	x	x	x	x
8	Develop materials: manuals, tutorials			X	x	x	x	x	x
9	Tutorials for user groups (fall, spr at iRODS User Mtg)					X	X		
10	Develop plan for wider system deployment and deployment of E-iRODS based DFC			X	X	X	x	x	x
11	Test E-iRODS for support of DFC capabilities						X	X	X

Milestones: Demonstration

		Quarter							
		1	2	3	4	5	6	7	8
1	Demonstration feature definition spirals			X		X			
2	Technology insertion spirals				X		X		
3	Test domain specific demo functionality			X	X	X			
4	Integrate and test integrated systems on DFC facilities (Hydrology (including NCDC), Engineering, OOI)				X	X	X		
5	Develop and test demo of E-iRODS based DFC system					X	X		
6	Refine and polish demos						X		
7									
8									
9									
10									

Demonstration April 2013



DFC DataNet FEDERATION CONSORTIUM

Questions?

E-iRODS: Enhanced Testing

RENCI E-iRODS Testing Environment

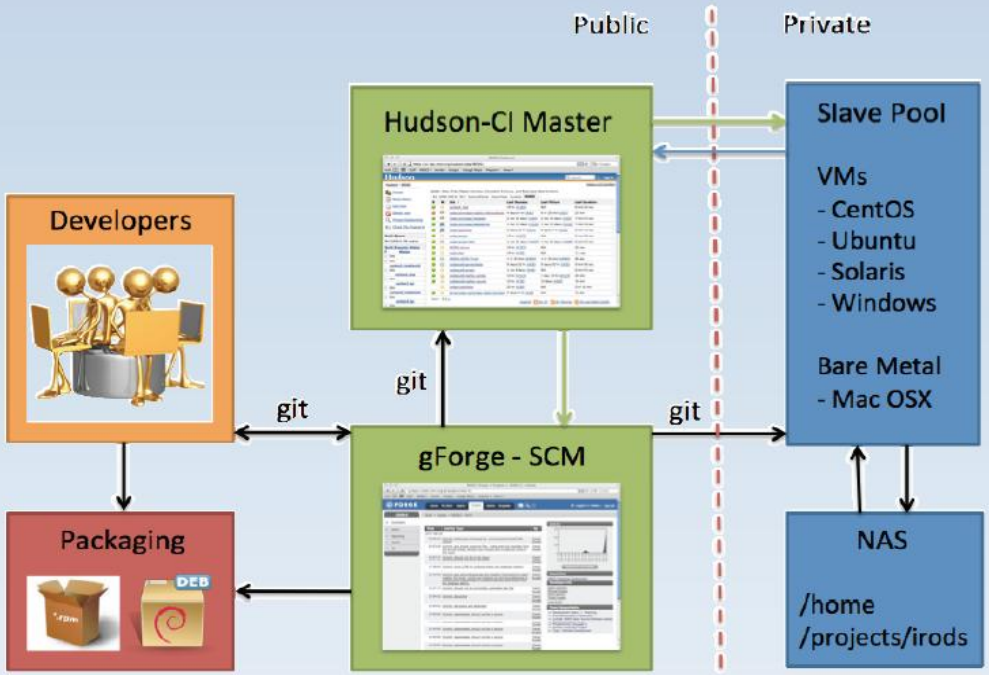
Hudson launches task on Slave Pool

Slave Pool runs script to "deploy a Gridbundle"

- Gridbundle – topological definition of an n-zone iRODS network (json)

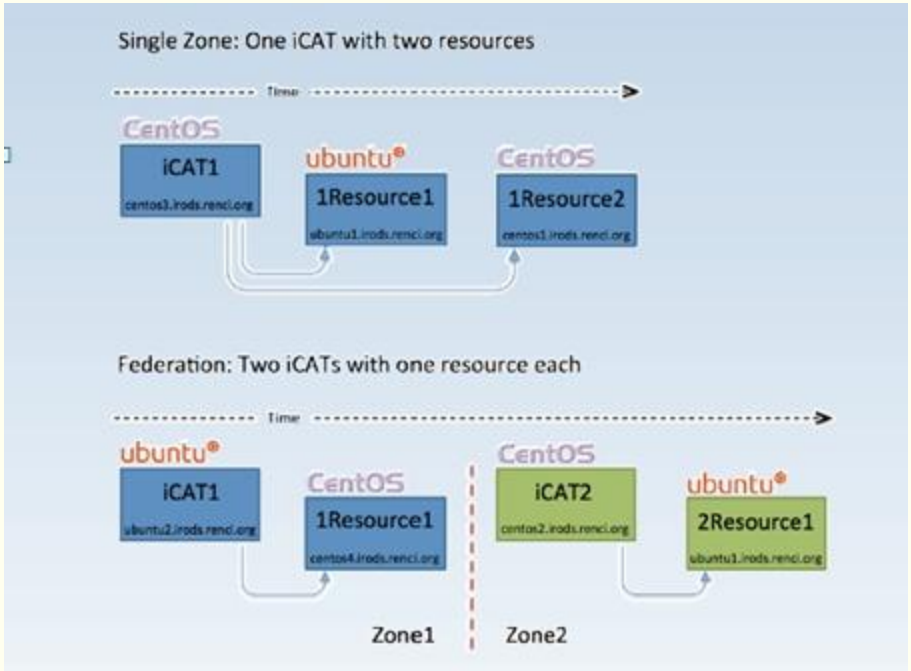
Tests run against the resulting "live grid"

- Automated and/or manual testing
- Aggregates test results from various machines



- Automated system testing
- Frequent code pushes to repository
- Continuous Build & Test
- Comprehensive test suites that target deployment scenarios
- Reduces risk and staff time

E-iRODS: Testing Framework

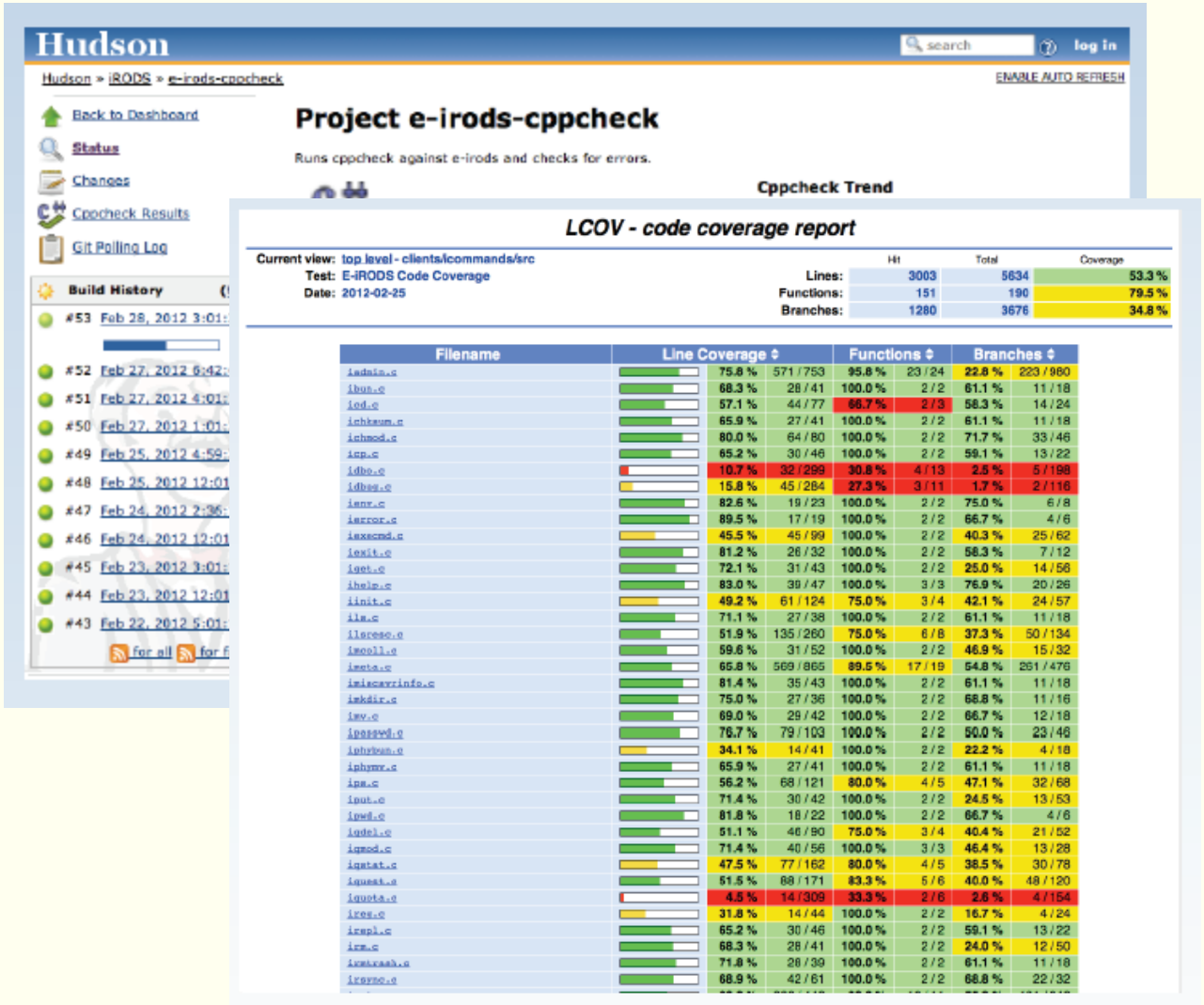


- Predefined test scenarios that combine:
  - Server characteristics
  - Grid topologies
  - Features to test

- Automated tests of distributed scenarios
  - Virtual network and virtual computers run on NSF Genie infrastructure.
  - Rapid creation/take down of deployments to test
- IP and hostname stitching to automate distributed tests
- Scripted tests and expected outcomes

DFC DataNet FEDERATION CONSORTIUM

E-iRODS: Testing Framework



- Packaged releases
- Releases certified on platforms and topologies
- 100% server side API testing

DFC DataNet FEDERATION CONSORTIUM