Digital Object Registry

Corporation for National Research Initiatives

Project Status Report – Sep 1, 2012 through Oct 25, 2012

1. Major Accomplishments

During the first three years, the scope of work on this project was to adopt the Handle System, along with components of the CNRI Digital Object Registry, to create a clearinghouse registry for principals, slices, and/or components in at least one GENI Spiral 1 control framework, capable of supporting limited operations. We have successfully adapted the Digital Object Registry and related technologies to build a GENI Federated Clearinghouse and a Distributed Hash Table for Seattle, aka Million Node GENI, led by Justin Cappos. The scope of work in Year 4 was to discuss, design, and develop a prototype of the Measurement Data Archive (MDA) service. We made these services available by deploying them in production servers with high availability and network bandwidth.

The scope of work in Year 5 is to continue the discussions with the Instrumentation and Measurement (I&M) working group members to standardize the metadata schema used by the MDA service and evaluate the integration of the Handle System into the MDA service developed by the UNC team.

During this reporting period, we continued participating in various GENI activities and programs, including attending GEC 15 held in Texas, and also continued our collaboration with GENI members and System Engineers as part of the I&M Working Group. We:

- Attended the I&M sessions held at GEC 15. Discussed with UNC and BBN team members to identify how the Handle System can be integrated into the iRODS framework that will be used for implementing the MDA service. Discussed with BBN team members to simplify the metadata schema that will be used within the MDA service.
- Continued to make available the prototype of the MDA service.
- Continued to make available the Distributed Hash Table service for the Million Node GENI project, led by Justin Cappos.
- Continued to make available the GENI Federated Clearinghouse service that federates the information from the ProtoGENI clearinghouse and makes that information available via the Digital Object Registry interfaces.

1.a. Milestones achieved: During this period we completed one milestone, S5.a, as discussed below.

- Together with the GENI community, we reviewed the GENI experimental data plan for identifying, archiving, searching and sharing measurement data objects, consistent with current best practices in the research community. Any updates and recommendations will be made during the next few months.

- We reviewed with the GENI I&M community version 2 of the schema for the GENI Measurement Data Object Descriptor (MDOD) that meets the needs of the community with minimum complexity. We will continue discussions to finalize the schema in the next few months.

- We discussed with UNC and other I&M members identifying how the Handle System can be integrated into the MDA service. A plan was identified during the discussions. Details of the plan are discussed in Section 2.

1.b. Deliverables: During this period, we made available the slides we presented during the I&M session at the GEC.

2. Description of Work Performed

2.a. Activities and Findings During This Period

Measurement Data Archive and Handle System

We demonstrated the MDA service in GEC 12 during the poster session and highlighted the project during other I&M related sessions in the same conference event. We divided the MDA service into two sub-services corresponding to the User Workspace and Persistent Archive components. Each of those sub-services provides its own storage and data filtering mechanism via its search capability.

The User Workspace component is an entry point for users (e.g., experimenters, instrumentation researchers, etc.) to store and transfer both measurement data, which could be in a variety of forms (e.g., formatted datasets, raw files, etc.), and metadata describing the measurement data. Users can then curate the data and metadata held in the workspace, e.g., make changes to the files, delete the files, etc.

Data and metadata files managed in the user workspace can be archived for long-term storage in the Persistent Archive. Once data is archived, a persistent and unique identifier is created and assigned to it. Additionally, a persistent reference to the data is provided. Discovery and filtering of the archived data is enabled using a search service.

Both components are based on the Digital Object (DO) Repository software that CNRI has made publicly available. GENI has funded UNC to use iRODS for implementing a production version of the prototype CNRI has built. While CNRI's

prototype is integrated with the Handle System, iRODS is not. During the conference, UNC, BBN, and CNRI participated in a session to evaluate the impact of the Handle System if it is integrated into a GENI deployment of the iRODS framework. In that discussion, we mutually concluded that Handle System would play a key role by offering publicly resolvable identifiers to objects archived within the iRODS framework. Once the Handle System is integrated, users can resolve Handles to access the archived objects.

While a few technical discussions will be held with UNC and CNRI to identify exactly how the Handle System will be integrated, the general consensus was that CNRI would continue to host a Handle server for UNC's use in Year 5. UNC will use the Handle client library to interact with the hosted Handle server and will assign Handles to objects at the time that those objects are archived. UNC will use micro-services within the iRODS framework to enable necessary interaction with the Handle System. At the time Handles are assigned, those micro-services will record URLs to the archived objects within the Handle records. End-users resolving Handles will, then, be redirected to the archived objects.

Measurement Data Object Descriptor

BBN and CNRI participated in a session during the GEC 15 to discuss and simplify the metadata schema identified during Year 4. The goal of this discussion was to identify what metadata elements were absolutely necessary and what metadata elements can be ignored for this version of the metadata schema release. While complex metadata schema might capture more detailed and hence quality metadata, it is usually a deterrent for users who contribute metadata. Simplifying metadata would lower the barrier to entry for contributors. Acknowledging the power of keeping the schema simple, BBN and CNRI have identified the metadata elements within the schema that are important. A few discussions that would follow the GEC 15 will result in the production of a simplified metadata schema that will be used by the UNC in their implementation of the workspace and archive as needed for the MDA service.

2.b. Project Participants

CNRI discussed its project activities with a number of other GENI participants, but all work accomplished this quarter was performed by CNRI alone, or with the cooperation of the I&M members and Harry Mussman. Names and email addresses of CNRI participants are available on the GENI wiki page for the project.

2.c. Publications

No publications were produced this quarter. CNRI gave a presentation on the need for archiving during the I&M session and demonstrated the developed MDA Prototype, the slides of which are available on the GENI wiki page for the project.

2.d. Outreach Activities

Giridhar Manepalli attended the GENI Engineering Conference held in Texas and participated in a variety of discussions with GENI members and System Engineers. He participated in the poster session, discussing the GENI services that CNRI offers. He gave a presentation highlighting the project during the I&M session.

2.e. Collaborations

CNRI continued to collaborate with the I&M working group members to standardize the metadata schema and integrate the Handle System into the MDA service.

2.f. Other Contributions

Production Services

CNRI continued to support the GENI Measurement Data Archive prototype and the GENI Federated Clearinghouse and the Distributed Hash Table services on hardware deployed in a collocation facility that features redundant power and air conditioning units, physical security, etc. A 100Mbps network pipe is dedicated to the machine.