NetKarma: a tool for obtaining a provenance-based record of experimentation

Beth Plale, Ph.D., Pervasive Technologies Institute, Indiana University
Chris Small, Global Research Network Operations Center, Indiana University

1 Overview

As computer network experiments increase in complexity and size, it becomes increasingly difficult to fully understand the circumstances under which the experiment was run, particularly when these results are shared for purposes of reproducibility. The ***provenance*** of an experiment is its lineage or historical trace [5] that can capture experiment conditions, time ordering, and relationships within the experiment and across the experiment and infrastructure layer. The GENI Provenance Registry (NetKarma) project, funded in Spiral 2, will provide a tool for capturing the workflow of GENI slice creation, topology of the slice, operational status and other measurement statistics and correlate it with the experimental data.

The tool, NetKarma, will allow researchers to see the exact state of the network and store configuration of the experiment and its slice. The provenance of the data will be stored and visualized through a data portal. The provenance data can be used by the researcher to analyze their data, allow for the suspension and resumption of an experiment and provide a single reference to find the details and data collected in an experiment. NetKarma bridges the gap between the GENI control plane infrastructure and the data collected in the experiments themselves or in GENI instrumentation such as the GENI Instrumentation and Measurement Systems (GIMS). The purpose of integrating the data and control plane allows researcher to expeditiously find the most relevant information collected.

The 3 year plan of development of the tool will draw from the successful Karma provenance collection and representation tool (http://dataandsearch.org/provenance) to start and iterate successively over the 3 years by engaging and working with the different layers of the GENI infrastructure, and by engaging the community in continual feedback for improvement. In the first year we will focus on the experimental tools, and by the end of year 1 intend to deliver a working version of the tool that collects and stores provenance information from the experiment tools. In Year 2 we will focus attention on the Control plane and Measurement Plane layers, and GMOC. In Year 3 refinements will be made, the tool will be hardened, the user interface refined, and educational/tutorial materials developed. We will strive to be responsive to community needs throughout the project.

The remainder of the whitepaper discusses the relevance to GENI of provenance collection in Section 2. Section 3 discusses details of integration with GENI. Section 4 gives an example of how a researcher would interact with the proposed tool.

2 GENI Relevance

Provenance data and its collection can benefit users of GENI in several ways as follows:

Experiment Planning, Deployment and Execution – our tool will integrate with the "GENI Control Prototype" based on the PlanetLab implementation. It can contribute to the ease of use of the control plane infrastructure by collecting provenance about researcher experiments, including the conditions of an experiment at both experiment and substrate levels, making this provenance information available for experiment planning and reproducibility. In order to gain a deep understanding of an experiment, it is necessary to know the knowledge of the slice topology, resources used, utilization of the resources and any error conditions. The attributes of an experiment collected by NetKarma could be used in part to create an "Experiment Specification Language" to describe not only what happened inside the slice but also the fundamental structure of the slice over the substrate and any event that occurred in the substrate that may have affected the experiment. This functionality, which we will explore, is similar to that described in section 5 of the "Lifestyle on a GENI Experiment" document.

Experiment **Sunset – provenance data about an activity can** facilitate its archival by collecting and organizing historical traces of experiments at runtime, and this has been identified as a key aspect of experiment sunsetting [3].. As our tool supports the Open Provenance Model (OPM) [4] for describing provenance information, it is capable of representing associations including causal and temporal relationships. Provenance information combined with measurement data would allow any future researcher to expand and analyze existing work. NetKarma would also allow for temporary suspension of a slice where resources are released to the GENI community. Since a detailed technical description of the resources and topology of the slice is contained in the provenance data, the same researcher or a future researcher can recreate the slice and resume the experiment easily.

Measurement and Instrumentation - The GENI System Overview calls for a high level of measurement instrumentation in GENI (section 3.6). NetKarma will provide a mechanism to correlate measurement data with the conditions under which the data was collected. A SPARQL query interface or equivalent will support inspection of data provenance and inspection of the structure and state of an experiment framework. For example, a query could be constructed to examine characteristics about the experimental data collected during a substrate outage or examine all experiments that used over 100 end nodes.

A provenance "graph" capturing a single experiment is uniquely identified globally unique ID and is immutable. Graphs can be annotated. Thus a graph is a single reference to the experiment, providing authors with a single reference to the entire set of data collected including documentation or logs. The use of a single reference point also has the side effect of identifying all papers that have used the GENI infrastructure. Simply by the fact that a paper contains the NetKarma "reference" shows that the paper used data collected in the GENI Infrastructure.

Outreach – Involving non-traditional users of network experiments is a goal of GENI. Through a portal interface, NetKarma will provide a location where newcomers to GENI, including graduate and undergraduate students, can browse prior experimental results. This bootstrapping model will enable newcomers to be productive more quickly.

3 Intergration with GENI

Creation of slices, experiment control tools, and authentication creates a wealth of data describing the state of the experiment. Dedicated measurement infrastructure and operational views into the substrate add additional information. The NetKarma project hopes to take advantage of the large amount of data provided by other pieces of the GENI Infrastructure in order to develop a picture of the state of each slice linked to any experimental data created. NetKarma will integrate the various sources of information into a coherent whole linked to the experimental results.

3.1 Provenance Capture

The NetKarma system will support the push/pull gathering of event information from multiple sources. The collection mechanisms supported by the tool will eventually include:

1. *Experimental Tool Data Gathering* – NetKarma will integrate with Experiment Control tools, such as RAVEN and GUSH, to capture communication and any internal representation of an experiment. We will investigate means to selectively capture information about data generated in the slice and about logs added by the experimenter.
2. *Control Plane Data Gathering* – Capture representations used internally by the control plane frameworks to represent experiments. This includes information about each element of the substrate, topology, and contact information about the experiment.
3. *Measurement Plane Data Gathering* – Gather references to measurement data captured in the measurement plane.

In the first year of the NetKarma project work will initially focus on integration of the experimental tool provenance.

There are two levels of provenance information to be captured by the NetKarma system.

- "Low-level" messages
- "High-Level" conceptual representations of the slice and the experiments

"Low-Level" messages are the traces of the actual communications being passed internally and between the GENI framework components. These include messages to create the slice, start and stop processes inside the slice or alarms for errors on the substrate. Traces of these messages will be collected and aggregated into a raw stream that can be processed at a later time.

"High-level" conceptual representations are aggregates of the message data generated for the component's own internal uses. Experimental control tools may have a representation of which processes were executed, in which order, and on what components. Control Plane frameworks will have knowledge of the researcher who owns a particular slice and what resources are used and how they are connected. It is expected that many of the GENI experimental framework components will have these internal representations in order to achieve their roles.

NetKarma will be agnostic on what high-level formats it can support. NetKarma will be able to support a wide variety of formats because it is just providing the linkage between the various formats and the experimental result. Following the GENI concept of federation, many of the internal representations may not need to be stored in a central source. Instead, NetKarma will provide a reference to data generated and stored inside the GENI clearinghouses and other long-term storage locations such as the GENI Instrumentation and Measurement System.

3.2 Provenance Representation

The provenance information model [1] we will use in NetKarma contains two levels: a registry level and an execution level. The registry level has similarities to the registry as used in Service-Oriented Architectures (SOA). It records the metadata about process objects and data, which are available for use in an execution sequence. It also serves to capture any known structure of an experiment, such as the sequence of execution. It supports composite relationships. The execution level models instances of the registry level and records the executable information of method invocations and data products used or generated by each invocation. Service, method invocation, and data product in the execution level are, respectively, instances of abstract service, abstract method, and abstract data product in the registry level. A method can be invoked by a service (represented by “service method invocation”), a method, or a client (represented by “client method invocation”). The client is an entity that initiates workflows or services; it could be a user or a workflow engine.

The NetKarma concept of a registry and execution levels maps very well with the concept of “low-level” message data and “high-level” conceptual representation capture. Message traces and other lower level data products are stored in the registry level as defined in the NetKarma model. “High-level” conceptual representations, such as topology schemas, from the GENI framework are comparable to the SOA registry level where the topology data product specifies the state of the slice at any point in time. It is also possible to infer a higher-level product from the message traces to generate a execution sequence, topology or other instances of a higher-level representation.

It will also possible to add the time dimension to GENI conceptual representations when these are missing. As currently formulated, many GENI frameworks do not store previous state of the slice or experiment. Control planes do not store a record of the exact configuration. Experiment control tools do not have a complete record of the execution sequence. While this is likely to change for many of the tools and frameworks, NetKarma allows for situations where there is no state recorded by other GENI framework elements. By capturing traces and tracking changes to high level representations a complete record of the conditions can be captured and stored even when all of the components do not support features such as time ordering.

3.3 Handles

Each experiment instance has a unique handle that not only refers to the experiment but also refers to all collected data including related provenance data. NetKarma will integrate with the Digital Object Registry to create a single reference to refer to a data “bundle”. This bundle will be automatically created for each experiment containing references to provenance data, measurement data and any other kind of data or

annotation that can be connected to an experiment. A single Digital Object Identifier (DOI) will be used to refer to the data bundle and allow for any user with the DOI reference to look up all relevant information about the experiment on the NetKarma portal. Users would use the DOI reference as they do today to find digital object such as raw data measurements. It would facilitate making all data collected on GENI accessible to other researchers, assuming proper privacy and security controls.

The use of a single reference would allow other researchers to quickly and easily see all the relevant data associated with an experiment. It also has the side effect of having any paper that refers to data generated on the GENI Infrastructure be easily identified as making use of the GENI infrastructure.

4 Experiment Lifecycle using NetKarma

In this section we describe a hypothetical GENI experiment and its lifecycle. We examine how elements of the GENI Infrastructure interact and the value the NetKarma registry provides.

An experimenter wants to deploy an experiment on the GENI infrastructure. Using provisioning tools they create a slice with programmable nodes and virtualized circuits to connect the nodes. NetKarma captures a trace of all messages sent by the provisioning tool using hooks embedded in the provisioning tools themselves or by capturing messages as they are communicated to the control planes. These messages are captured in their raw format and any internal representation used by the provisioning tools and the control plane framework. A handle is generated linking the provenance data and information about the experimenter. Views of the data are available in the NetKarma portal.

As the researcher conducts their experiment in the slice, information is added to the NetKarma registry. This can be done automatically from information gathered by the experiment process tools or can be added manually by the experimenter. Additional data and annotations can be manually added through the NetKarma portal. Any data collected in GENI instrumentation, such as GIMS, will also be automatically stored and correlated.

Once the experiment is finished, the researcher can refer to the provenance and any measurement data with a single DOI handle that will point to the data collected in the experiment. Another researcher discovering the reference in a paper or browsing the NetKarma portal can examine the entire data set, recreate the slice and rerun the experiment.

References

1. Bin Cao, Beth Plale, Girish Subramanian, Ed Robertson, and Yogesh Simmhan, Provenance Information Model of Karma, IEEE 2009 Third International Workshop on Scientific Workflows (SWF'09), July 2009.

2. Dennis Gannon, Beth Plale, Marcus Christie, Yi Huang, Scott Jensen, Ning Liu, Suresh Marru, Sangmi Lee Pallickara, Srinath Perera, Satoshi Shirasuna, Yogesh Simmhan, Aleksander Slominski, Yiming Sun, Nithya Vijayakumar, 2007: Building Grid Portals for e-Science: A Service Oriented Architecture. *High Performance Computing and Grids in Action*, IOS Press - Amsterdam, Lucio Grandinetti editor.

3. Global Environment for Network Innovations, Lifecycle of a GENI Experiment, GENI-SE-SY-TS-UC-LC-01.0, January 2009.

4. The Open Provenance Model (v1.01). Moreau, L. (Editor), B. Plale, S. Miles, C. Goble, P. Missier, R. Barga, Y. Simmhan, J. Futrelle, R. McGrath, J. Myers, P. Paulson, S. Bowers, B. Ludaescher, N. Kwasnikowska, J. Van den Bussche, T. Ellkvist, J. Frieire, P. Groth, Technical Report, Electronics and Computer Science, University of Southampton, 2008. http://eprints.ecs.soton.ac.uk/16148

5. Yogesh L. Simmhan, Beth Plale, and Dennis Gannon, A survey of data provenance in e-Science, *ACM SIGMOD Record,* Vol. 34, No. 3, September 2005.

6. Yogesh Simmhan, Beth Plale, and Dennis Gannon, 2008: Karma2: Provenance Management for Data Driven Workflows. *International Journal of Web Services Research*, IGI Publishing, Vol 5, No 2, 2008.