# Orchestrating Distributed Resource Ensembles for Petascale Science

**Final Report, RENCI, University of North Carolina at Chapel Hill**
**ASCR DE-SC0005286, October 2010 — November 2013**

Ilya Baldin, Anirban Mandal, Paul Ruth, Yufeng Xin
RENCI, University of North Carolina at Chapel Hill
{ibaldin, anirban, pruth, yxin}@renci.org

## Executive Summary

Distributed, data-intensive computational science applications of interest to DOE scientific communities move large amounts of data for experiment data management, distributed analysis steps, remote visualization, and accessing scientific instruments. These applications need to orchestrate ensembles of resources from multiple resource pools and interconnect them with high-capacity multi-layered networks across multiple domains. It is highly desirable that mechanisms are designed that provide this type of resource provisioning capability to a broad class of applications. It is also important to have coherent monitoring capabilities for such complex distributed environments.

In this project, we addressed these problems by designing an abstract API, enabled by novel semantic resource descriptions, for provisioning complex and heterogeneous resources from multiple providers using their native provisioning mechanisms and control planes: computational, storage, and multi-layered high-speed network domains. We used an extensible resource representation based on semantic web technologies to afford maximum flexibility to applications in specifying their needs. We evaluated the effectiveness of provisioning using representative data-intensive applications. We also developed mechanisms for providing feedback about resource performance to the application, to enable closed-loop feedback control and dynamic adjustments to resource allocations (elasticity). This was enabled through development of a novel persistent query framework that consumes disparate sources of monitoring data, including perfSONAR, and provides scalable distribution of asynchronous notifications.

### Accomplishments and Highlights

Some of the major accomplishments of the project are the following.

- We developed a simple API, enabled by semantic resource descriptions, for applications to provision resources from networked cloud platforms. Applications were able to dynamically provision compute, storage, and network resources across multiple domains. The resource provisioning API [11] is used on the ExoGENI testbed.
- We used two representative, data-intensive, application classes - Hadoop/MapReduce and Scientific workflows, to study the effectiveness of provisioning them on networked clouds. We evaluated their performance on Networked Infrastructure-as-a-Service (NIaaS) platforms. The results of these evaluations were published in [8, 7].
- We experimented with provisioning a large-scale scientific workflow using our networked cloud testbed augmented with Hopper, a Supercomputer at a DOE Leadership Class Facility, NERSC. The use case was a Solar Fuels scientific workflow from the DOE Energy

Frontier Research Center (EFRC) at UNC, Chapel Hill. The workflow used resources from North Carolina linked to DOEs Hopper supercomputer at NERSC in California. The workflow was under the control of the Pegasus workflow management system and started execution on a Condor pool built out of cloud resources in NC and completed with a massive 3K-way MPI job on Hopper by transferring data over a dynamic QoS-provisioned link between NC resources and NERSC. We demonstrated this work at Supercomputing 2011, and presented it in [9].

- We created a capability for persistent queries for perfSONAR and other measurement sources to support closed-loop feedback control for application performance monitoring and future resource provisioning. We developed a software for supporting persistent queries - the Persistent Query Agent (PQA) that enables federated performance monitoring by interacting with multiple aggregates and performance monitoring sources. Using a distributed, scalable, publish-subscribe framework, it sends triggers asynchronously to applications/clients when relevant performance events occur. The PQA software is available for download. The capabilities were demonstrated at SC'12 and SC'13. The details about the tool are in a technical report [6].
- We developed a capability for closed-loop feedback control using persistent queries on monitoring data. We demonstrated this as a part of a SCInet Network Research Exhibition demonstration at Supercomputing 2013.

# 1 Project Activities

## 1.1 Provisioning and Evaluating Data-Intensive Applications on Networked Clouds

One of the goals of the project was to develop a simple API for applications to provision resources from networked clouds, and to assess the performance of representative science applications on networked clouds. We selected two different classes of data-intensive scientific applications - Hadoop/MapReduce and Scientific workflows, to study the effectiveness of provisioning them on networked clouds and evaluating their performance on Networked Infrastructure-as-a-Service (NIaaS) platforms.

### 1.1.1 Hadoop/MapReduce

We designed, implemented, and evaluated a system for on-demand provisioning of Hadoop clusters across *multiple* cloud domains. The Hadoop clusters were created "on-demand" and are composed of virtual machines from multiple cloud sites linked with bandwidth-provisioned network pipes. The prototype used an existing federated cloud control framework called Open Resource Control Architecture (ORCA) [2, 5], which orchestrates the leasing and configuration of virtual infrastructure from multiple autonomous cloud sites and network providers. ORCA enables computational and network resources from multiple clouds and network substrates to be aggregated into a single virtual "slice" of resources, built to order for the needs of the application.

We ran several experiments to evaluate our provisioning approach, as described in [11], and the performance of data-intensive MapReduce applications on the provisioned resources. The experiments examined various provisioning alternatives by evaluating the performance of representative Hadoop benchmarks and applications on resource topologies with varying bandwidths. The evaluations examined conditions in which multi-cloud Hadoop deployments pose significant advantages or disadvantages during Map/Reduce/Shuffle operations. Further, the experiments compared multi-cloud Hadoop deployments with single-cloud deployments and investigated Hadoop Distributed File System (HDFS) performance under varying network configurations.

The results showed that networked clouds make cross-cloud Hadoop deployment feasible with high bandwidth network links between clouds. As shown in Figure 1, performance for some benchmarks degraded rapidly with constrained inter-cloud bandwidth. MapReduce shuffle patterns and certain Hadoop Distributed File System (HDFS) operations that span the constrained links were particularly sensitive to network performance. Hadoop's topology-awareness feature can mitigate these penalties to a modest degree in these hybrid bandwidth scenarios. Additional observations, as shown in Figure 2, showed that contention among co-located virtual machines is a source of irregular performance for Hadoop applications on virtual cloud infrastructure.

The details regarding this work have been published in [8].

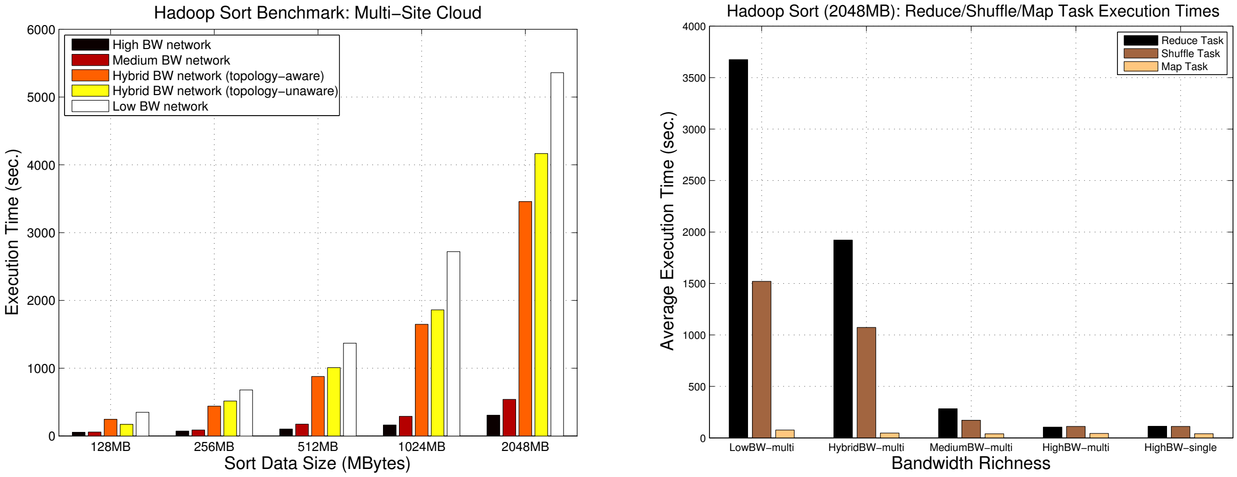

Figure 1: Multi-site Cloud with different provisioned bandwidths (Sort) and Reduce/Shuffle/Map average times

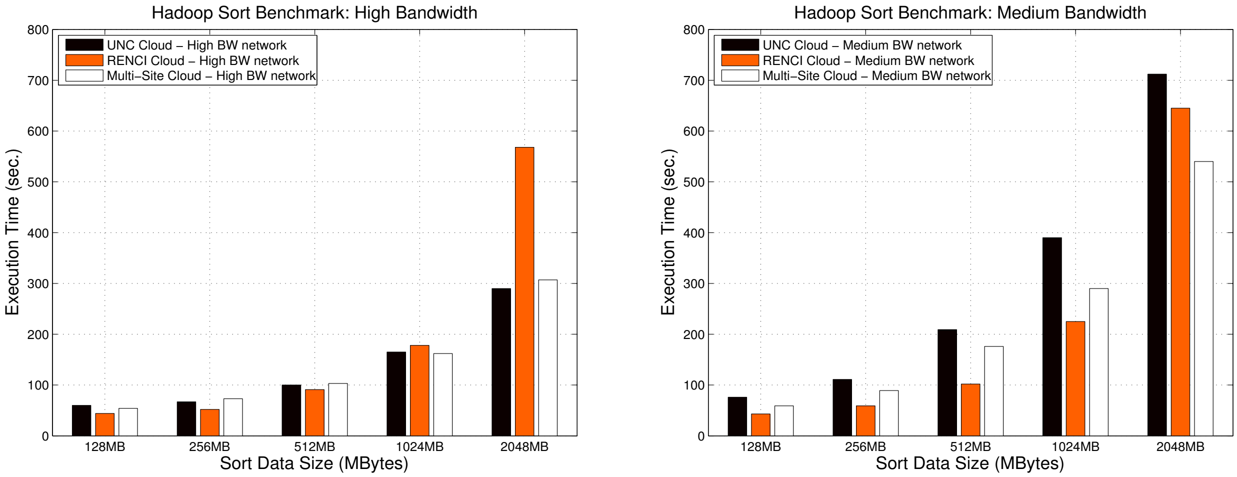

Figure 2: Hadoop Sort Benchmark: High and Medium Bandwidth

### 1.1.2 Scientific Workflows

### Solar Fuels (PSOCI) Workflow

We experimented with provisioning large-scale scientific workflows using our networked cloud testbed augmented with a Leadership Class Facility, NERSC. One of our proposed used cases was a Solar Fuels scientific workflow, the source of which was the DOE Energy Frontier Research Center (EFRC) at UNC, Chapel Hill.

In this case, ORCA was used to execute a complex computational workflow simulating the atomic behavior of a solar fuel catalyst. The application used resources from North Carolina linked to DOEs Hopper supercomputer at NERSC in California. The workflow was under the control of the Pegasus workflow management system and started execution on a Condor pool built out of cloud resources in NC and completed with a massive 3K-way MPI job on Hopper. ORCA created a slice of resources consisting of a collection of cloud resources from multiple sites in NC and a dynamic QoS-provisioned link between NC resources and NERSC composed of VLAN/MPLS segments from BEN, NLR Framenet, and ESnet. The segments were requested by ORCA through OSCARS and Sherpa respective APIs and stitched together by ORCA using a Cisco switch it controls at the StarLight facility in Chicago, where NLR and ESnet have points of presence.

This work was demonstrated at the SuperComputing 2011 conference in Seattle, and was also presented in [9].

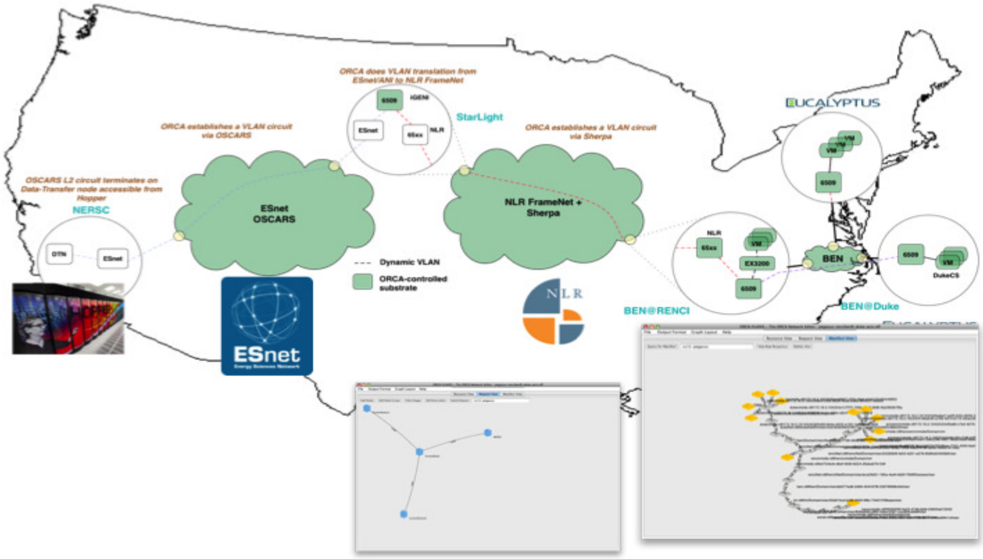
ORCA does VLAN translation from ESnet/ANI to NLR FrameNet
6509
iGENI
ESnet
65xx
NLR
StarLight
ORCA establishes a VLAN circuit via OSCARS
ORCA establishes a VLAN circuit via Sherpa
EUCALYPTUS
VM
VM
VM
6509
OSCARS L2 circuit terminates on Data-Transfer node accessible from Hopper
NERSC
DTN
ESnet
ESnet OSCARS
NLR FrameNet + Sherpa
Dynamic VLAN
ORCA-controlled substrate
NLR
65xx
VM
EX3200
6509
BEN
6509
VM
DukeCS
BEN@RENCI
BEN@Duke
ESnet
Energy Sciences Network
NLR

Figure 3: Provisioning Solar Fuels Workflow on Networked Cloud and DOE Hopper Supercomputer

### "Montage" Workflow Evaluation - I/O Aware Network Management

In this work, we evaluated the performance of scientific workflows on networked cloud systems with particular emphasis on evaluating the effect of provisioned network bandwidth on application I/O performance. The experiments were run on ExoGENI [1], a widely distributed networked infrastructure as a service (NIaaS) testbed. ExoGENI orchestrates a federation of independent cloud sites located around the world along with backbone circuit providers. The evaluation used a representative data-intensive scientific workflow application called Montage. The application was deployed on a virtualized HTCondor environment provisioned dynamically from the ExoGENI networked cloud testbed, and managed by the Pegasus [3] workflow manager.

The results of our experiments showed the effect of modifying provisioned network bandwidth on disk I/O throughput and workflow execution time. As shown in Figure 4, the marginal benefit as perceived by the workflow reduces as the network bandwidth allocation increases to a point where disk I/O saturates. There is little or no benefit from increasing network bandwidth beyond this inflection point. The results also underscored the importance of network and I/O performance isolation for predictable application performance, and are applicable for general data-intensive workloads. Insights from this work will also be useful for real-time monitoring, application steering and infrastructure planning for data-intensive workloads on networked cloud platforms.

The details regarding this work have been published in [7].

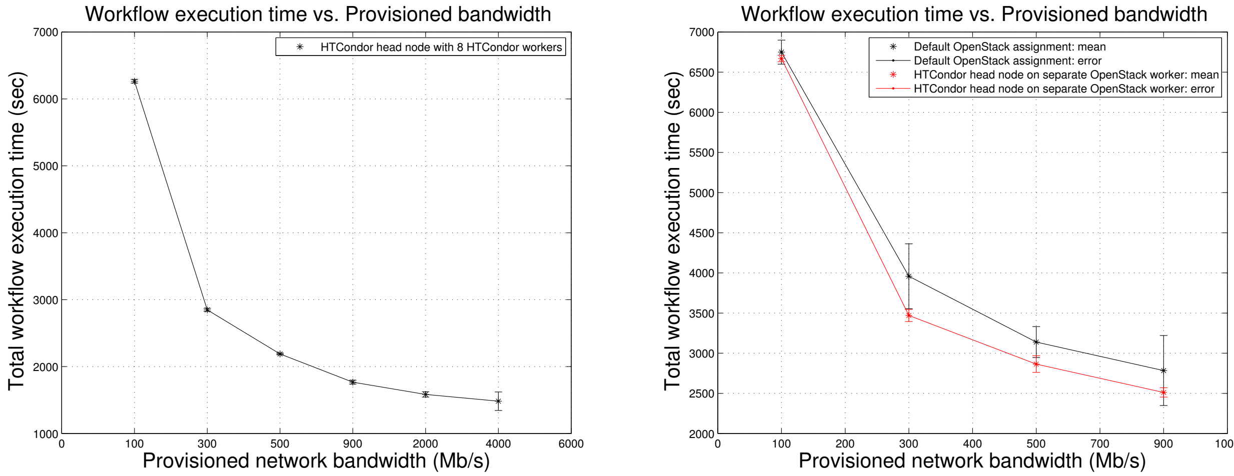

Figure 4: Workflow execution time vs. provisioned network bandwidth (FIU rack and NICTA rack)

## 1.2 Persistent Queries and Monitoring

Another goal for the project was to create a capability for persistent queries for perfSONAR [10] and other measurement sources to support closed-loop feedback control for application performance monitoring and future resource provisioning.

It is essential for distributed data-intensive applications to monitor the performance of the underlying network, storage and computational resources. Increasingly, distributed applications need performance information from multiple aggregates, and tools need to take real-time steering decisions based on the performance feedback. With increasing scale and complexity, the volume and velocity of monitoring data is increasing, posing scalability challenges. In this work, we have developed a software for supporting persistent queries - the Persistent Query Agent (PQA) that provides real-time application and network performance feedback to clients/applications, thereby enabling dynamic adaptations.

PQA enables federated performance monitoring by interacting with multiple aggregates and performance monitoring sources. Using a distributed, scalable, publish-subscribe framework, it sends triggers asynchronously to applications/clients when relevant performance events occur. The applications/clients register their events of interest using declarative queries and get notified by the PQA. PQA leverages a complex event processing (CEP) framework called Esper [4] for managing and executing the queries expressed in a standard SQL-like query language called the Esper Event Processing Language (EPL). Instead of saving all monitoring data for future analysis, PQA observes performance event streams in real-time, and runs continuous queries over streams of monitoring events. We have designed and architected the PQA system, as shown in the following Figure 5.

This work was demonstrated at the SuperComputing 2012 conference in Salt Lake City. The details are in a technical report [6]. This paper is under under submission to a special issue of the IEEE Communications magazine.

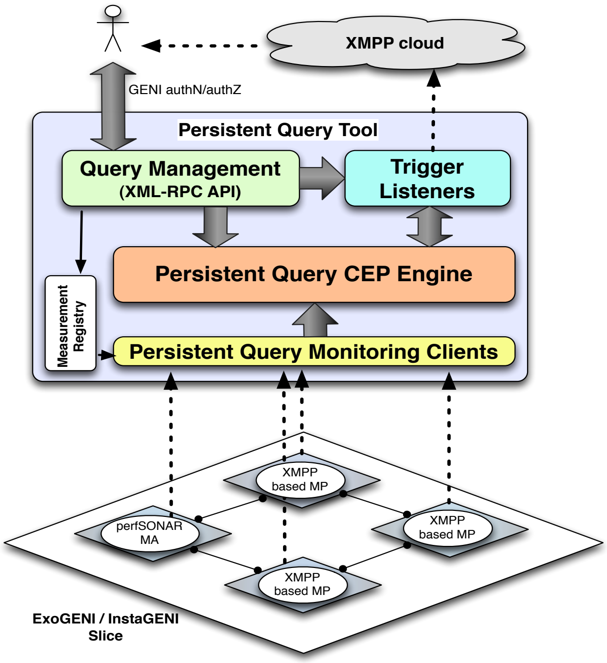

Figure 5: Persistent Query Agent (PQA)

## 1.3 Closed-loop Feedback Control

We developed a capability for closed-loop feedback control using persistent queries on monitoring data. We demonstrated this capability as a part of a SCInet Network Research Exhibition demonstration at Supercomputing 2013. The demonstration utilized several new features available on ExoGENI including multi-point broadcast networks, sliverable storage, stitchport on-ramps, and a closed-loop monitoring and control mechanism to execute Montage, a data intensive Pegasus scientific workflow application. The slice included dynamically provisioned connections over I2, ESnet, NLR and BEN, as well as a connection over SCInet to a data storage host located in RENCIs SC'13 booth.

We demonstrated dynamic scaling of a virtualized HTCondor cluster based on measurements of idle job queue length. The architecture is shown in Figure 6. We used persistent queries to consume the measurements and send asynchronous triggers to a "autonomic, dynamic monitoring and control" agent when the idle job queue length exceeded a threshold decided by the application scientist. The “IdleJobs” metric and an EPL query expressing the above constraint was registered with PQA. When the “autonomic, dynamic monitoring and control” agent received the triggers from PQA, it initiated scaling actions (growing or shrinking the virtualized HTCondor cluster) by sending modification requests to ExoGENI, the resource provisioning system.

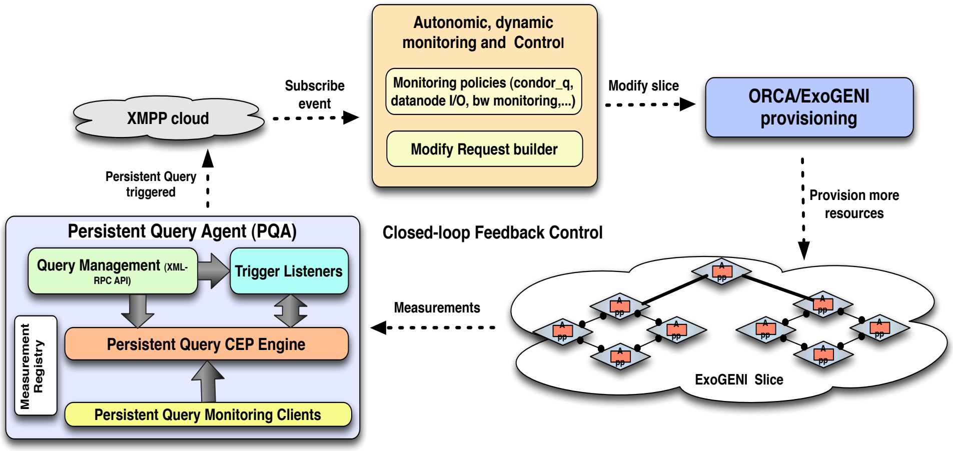
Autonomic, dynamic monitoring and Control
Monitoring policies (condor_q, datanode I/O, bw monitoring,...)
Modify Request builder
XMPP cloud
Subscribe event
Modify slice
ORCA/ExoGENI provisioning
Persistent Query triggered
Provision more resources
Persistent Query Agent (PQA)
Closed-loop Feedback Control
Query Management (XML-RPC API)
Trigger Listeners
Measurement Registry
Persistent Query CEP Engine
Persistent Query Monitoring Clients
Measurements
ExoGENI Slice

Figure 6: Dynamic monitoring and closed-loop feedback control using persistent queries

# 2 Research Products

## 2.1 Publications and Presentations

1. A. Mandal, P. Ruth, I. Baldin, Y. Xin, C. Castillo, M. Rynge, and E. Deelman. Evaluating I/O aware network management for scientific workflows on networked clouds. In Proceedings of the Third International Workshop on Network-Aware Data Management, NDM ’13, pages 2:1-2:10, New York, NY, USA, 2013. ACM.

2. P. Ruth, A. Mandal, Y. Xin, I. Baldine, C. Heerman, and J. Chase. Dynamic network provisioning for data intensive applications in the cloud. In IEEE 8th International Conference on E-Science (e-Science), 2012, pages 1-2, 2012.

3. A. Mandal, Y. Xin, I. Baldine, P. Ruth, C. Heerman, J. Chase, V. Orlikowski, and A. Yumerefendi. Provisioning and evaluating multi-domain networked clouds for Hadoop-based applications. In IEEE Third International Conference on Cloud Computing Technology and Science (CloudCom 2011), pages 690-697, 2011.

4. A. Mandal, I. Baldine, Y. Xin, P. Ruth, and C. Heerman. Enabling persistent queries for cross-aggregate performance monitoring. Technical Report TR-13-01, Renaissance Computing Institute, 2013, http://www.renci.org/wp-content/uploads/2013/04/TR-13-01.pdf.

5. Y. Xin, I. Baldine, A. Mandal, C. Heermann, J. Chase, and A. Yumerefendi. Embedding virtual topologies in networked clouds. In Proceedings of the 6th International Conference on Future Internet Technologies, CFI ’11, pages 26-29, New York, NY, USA, 2011. ACM.

6. I. Baldin, A. Mandal, Y. Xin, P. Ruth, C. Castillo, J. Chase, M. Rynge, and E. Deelman, Dynamic Monitoring and Adaptation of Data Driven Scientific Workflows Using Federated Cloud Infrastructure. Presentation and Demonstration at SCInet Network Research Exhibition at SC'2013, Denver, CO.

### 2.2 Software

We have developed a software for supporting persistent queries for perfSONAR and other measurement sources to support closed-loop feedback control for application performance monitoring and future resource provisioning. The name of the software is Persistent Query Agent (PQA). It is available for download by checking out from SVN repository at
https://code.renci.org/svn/drops/trunk/pSPersistentQuery

### 2.3 Website

The project website is https://code.renci.org/gf/project/drops/ . All publications, software, and reports are archived on this website.

## References

[1] I. Baldine, Y. Xin, A. Mandal, P. Ruth, A. Yumerefendi, and J. Chase. Exogeni: A multi-domain infrastructure-as-a-service testbed. In *8th International ICST Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities (TRIDENT-COM 2012)*, 2012.

[2] J. Chase, L.Grit, D.Irwin, V.Marupadi, P.Shivam, and A.Yumerefendi. Beyond virtual data centers: Toward an open resource control architecture. In *Selected Papers from the International Conference on the Virtual Computing Initiative (ACM Digital Library)*, May 2007.

[3] E. Deelman, G. Singh, M.-H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, A. Laity, J. C. Jacob, and D. S. Katz. Pegasus: a Framework for Mapping Complex Scientific Workflows onto Distributed Systems. *Scientific Programming Journal*, 13:219–237, 2005.

[4] EsperTech. http://www.espertech.com, 2013.

[5] D. Irwin, J. S. Chase, L. Grit, A. Yumerefendi, D. Becker, and K. G. Yocum. Sharing Networked Resources with Brokered Leases. In *Proceedings of the USENIX Technical Conference*, June 2006.

[6] A. Mandal, I. Baldine, Y. Xin, P. Ruth, and C. Heerman. Enabling persistent queries for cross-aggregate performance monitoring. Technical Report TR-13-01, Renaissance Computing Institute, 2013, http://www.renci.org/wp-content/uploads/2013/04/TR-13-01.pdf.

[7] A. Mandal, P. Ruth, I. Baldin, Y. Xin, C. Castillo, M. Rynge, and E. Deelman. Evaluating i/o aware network management for scientific workflows on networked clouds. In *Proceedings of the Third International Workshop on Network-Aware Data Management*, NDM '13, pages 2:1–2:10, New York, NY, USA, 2013. ACM.

[8] A. Mandal, Y. Xin, I. Baldine, P. Ruth, C. Heerman, J. Chase, V. Orlikowski, and A. Yumerefendi. Provisioning and evaluating multi-domain networked clouds for hadoop-based applications. In *IEEE Third International Conference on Cloud Computing Technology and Science (CloudCom 2011)*, pages 690–697, 2011.

[9] P. Ruth, A. Mandal, Y. Xin, I. Baldine, C. Heerman, and J. Chase. Dynamic network provisioning for data intensive applications in the cloud. In *E-Science (e-Science), 2012 IEEE 8th International Conference on*, pages 1–2, 2012.

[10] B. Tierney, J. Boote, E. Boyd, A. Brown, M. Grigoriev, J. Metzger, M. Swany, M. Zekauskas, Y.-T. Li, and J. Zurawski. Instantiating a Global Network Measurement Framework. Technical Report LBNL-1452E, Lawrence Berkeley National Lab, 2009.

[11] Y. Xin, I. Baldine, A. Mandal, C. Heermann, J. Chase, and A. Yumerefendi. Embedding virtual topologies in networked clouds. In *Proceedings of the 6th International Conference on Future Internet Technologies*, CFI '11, pages 26–29, New York, NY, USA, 2011. ACM.