Collaborative Research: SDCI Net:
Policy-driven Large Scale Data Access Framework with Light-weight Performance Monitoring and Estimation (Adaptive Data Access and Policy-driven Transfers - ADAPT)

Award Numbers: 1127101 (USC/ISI), 1127039 (LBNL)

Annual Project Report

July 1, 2013

Project period of
August 1, 2012 through July 31, 2013

Principal Investigators
Ann Chervenak[1], Alex Sim[2]

Project member: Robert Schuler[1], Junmin Gu[2]
Project Website: http://sdm.lbl.gov/adapt/
Project email: adapt@hpcrd.lbl.gov

NSF Program Officer: Kevin L. Thompson

---

[1] University of Southern California, Information Sciences Institute
[2] Lawrence Berkeley National Laboratory

Table of Contents

2 Accomplishments - What was done? What was learned?

2.1 What are the major goals of the project?

Over the next 5-10 years, applications in almost every major scientific domain, including high energy physics, nuclear science, astrophysics, climate modeling, nanoscale materials science, and genomics, are expected to generate exabytes of data. The large and increasing volume of data must be shared by increasing number of geographically distributed collaborators in the world by high-speed networks, and must be moved from the experimental facilities to scientists and to analysis, simulation, and storage facilities.

The overall goal of the Adaptive Data Access and Policy-driven Transfers (ADAPT) project is to facilitate efficient data movement over high-speed networks by supporting find-grained and adaptive data transfer management, passive performance monitoring, and enforcement of site and Virtual Organization (VO) policies for resource sharing, through a general purpose data access software framework.

In this project, we address the following issues with our software framework:

- Passive performance monitoring. To optimize the data accessibility, monitoring information from resources needs to be collected in a way that does not put extra loads on the resources. We will collect passive monitoring information by enhancing data transfer clients to report transfer performance to a measurement archive, and to use the measurement information to generate a simple approximation of the data transfer performance.
- Adaptive data transfer. On high-performance resources, static data transfer properties can cause orders of magnitude performance degradations because of the dynamically changing shared environment. We will implement adaptive transfer management methods for transfers in progress, responding to changes in the throughput performance and in the policies for their use.
- Policy-driven data transfer management. Efficient and optimal data accessibility must be based on a policy that balances user requirements for scalability and end-to-end resource performance. Policy administrators at the Virtual Organization or site level define the overall resource limits between hosts based on their resources (e.g. memory size, bus speed, storage speed) and the available network bandwidth between them. These limits are recorded and interpreted by environment-specific policy rules that determine, along with knowledge of the total resources allocated to other ongoing data transfers and their performances, what the recommended transfer properties limits should be for the client instance. We will develop Policy Service to enforce these policies for data transfer management.

To address these issues in a coordinated and well-defined way, the main objective of the project is to develop and release a general-purpose, reusable and expandable framework for optimizing the performance of data movements over the network for scientific collaborations by supporting adaptive data transfer management, passive performance monitoring, and enforcement of site and VO policies for resource sharing. The framework will be developed based on the existing tools that manage the access and replication of large scientific datasets, with additional capabilities for collecting transfer monitoring information, performing adaptive data transfer management based on both policy and performance constraints, and implementing new policies for VO level data distribution and user level data access. These software tools and framework will be validated on the distributed testbeds before being released to large scale scientific communities under open source licenses. This project enables improving efficient data access by many users and many data requests, as well as data and resource sharing within the user community.

2.2 What was accomplished under these goals?

2.2.1 Major activities

Major activities from year 2 of the project include:

- New implementation of the Policy Service that controls the allocation of streams to data transfer clients

- New version of the adaptive SRM copy data transfer client that interacts with the Policy Service to acquire an allocation of streams for the client and then adapts within that allocation based on transfer performance
- Experimental evaluation looking at tradeoffs from doing adaptation at the Policy Service and at the data transfer client, compared with non-adaptive data clients
- Outreach to science communities in the Open Science Grid collaboration

Our project work includes three components, the Policy Service (PS), Data Movement Service (DMS) and Passive Measurement Archive (PMA), as illustrated in Figure 1. This work builds on several existing components: a Policy Driven Data Management Service, data movement client tools including the SRM client services, and the measurement data archive such as perfSONAR.

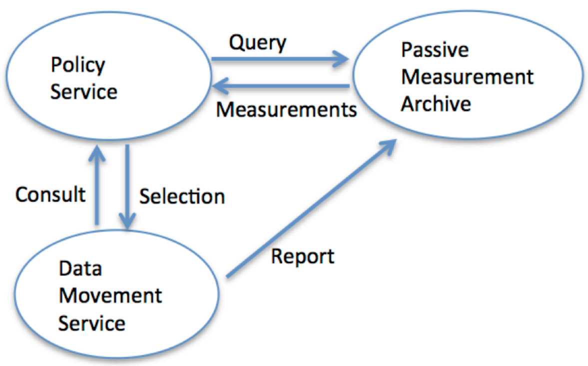


Figure 1: interaction diagram between components

The Passive Measurement Archive (PMA) collects monitoring data passively from the Data Movement Services (DMS), and provides the historical measurements based on the existing archiving services such as perfSONAR. The Policy Service (PS) can make use of this historical measurement information as well as its own specified policies to make recommendations on data transfer properties to the transfer client. The Data Movement Service (DMS) uses information from the PS to manage data transfers and adaptively adjust data transfer properties such as parallelism, concurrency and buffer size.

We designed an open interface between PMA, PS and DMS components that allows other services and projects to benefit from this project. Each service has a modular design so that essential techniques may be deployed and called through libraries by other services and project components. For detailed design approaches, a reference to the 2012 annual report (http://sdm.lbl.gov/adapt/docs/ADAPT-Report-2012-final.pdf) is available[3].

In this second year of the project, we have developed functional enhancements to the component services based on the general use cases in Open Science Grid (OSG). The Policy Service was re-implemented as a RESTful service written in the python language. The adaptive version of the srm-copy data transfer client was modified to interact with the PS; the client acquires a stream allocation from the PS and then adapts the streams per transfer within that allocation based on transfer performance. We also have defined experimental test cases, and collected extensive measurements to analyze and improve the effectiveness of the adaptive methods and policy rules.

2.2.1.1 Component design and functional enhancements

The initial development (phase 1) from 2012 was designed to provide an optimized data transfer experience and easier transition from the current data movement practices in Open Science Grid (OSG) to the new adaptive data movement. The second phase of the project in this year is focused on the service-oriented functionality and generalized framework of the system as illustrated in figures 2.

Figure 2 shows a use case for year 2, where data movement clients interact with multiple storage sites. This is a typical use case in OSG to stage a dataset for analysis or to retrieve a dataset from storage repositories. In phase 2, we have enhanced the data movement client integrated with Adaptive Data Transfer (ADT) library module and implemented and deployed the policy logic as a service. In addition to

[3] http://sdm.lbl.gov/adapt/docs/ADAPT-Report-2012-final.pdf

the functions from each component in phase 1, the specific enhancements have been implemented to the components in the project scope for phase 2, and described in the next Specific objectives section.

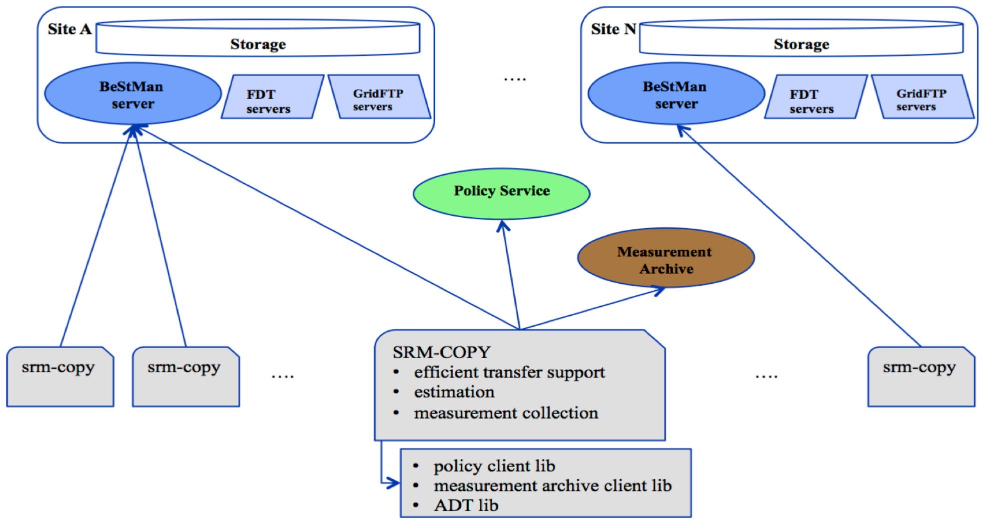


Figure 2: The second phase development and deployment plan as component services

2.2.1.2 Component interactions for the second phase of the project

Figure 3 shows a use case for the second phase of the project (year 2), with PS as an external service and with the external measurement archives (PMA). There can be many source and destination sites, involved in the VO data traffic, and they would affect the transfer performance of each data transfer significantly. The PS has been implemented as a service. The data transfer client, srm-copy, has been enhanced from the Year 1 implementation, and connects to the PS through a web service provided by one of the library modules, Adaptive Transfer Management (ATM). Current transfer statistics from the client tools are archived in PMA.

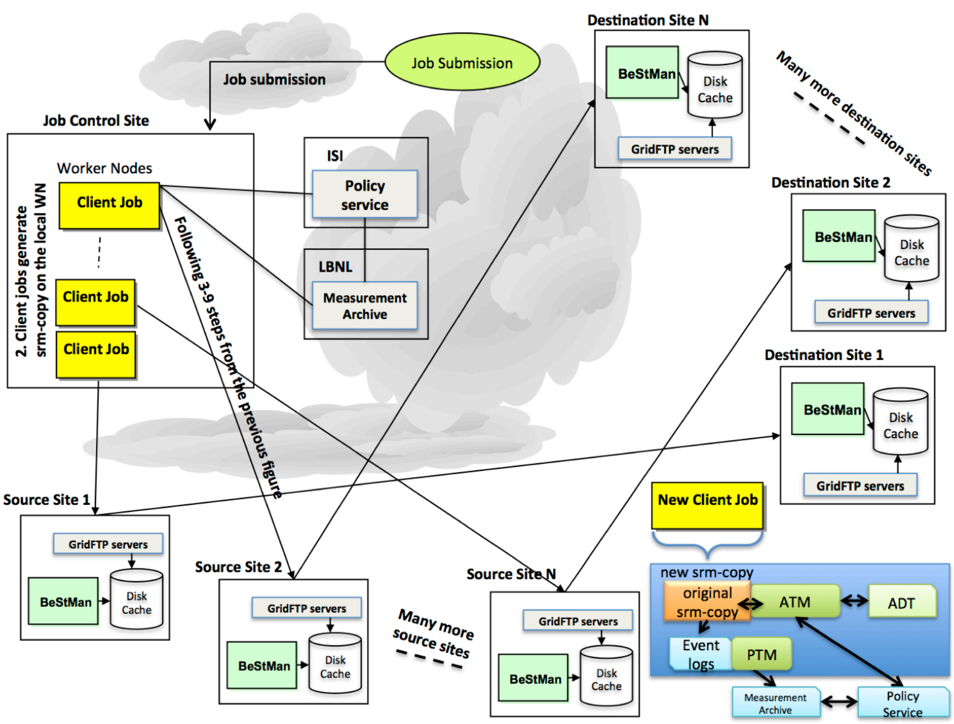


Figure 3: The second phase development and deployment with a use case

2.2.2 Specific objectives

The main goals of this work are to develop functionality for policy-driven allocation of transfer resources and for adaptation of transfer parameters based on measured performance. Next, we discuss in more detail the objectives for each major component.

Adaptive Data Transfer library

The data transfer client performs its transfer adaptation using the Adaptive Data Transfer (ADT) library. This library enables the client to adapt varying environmental conditions to come up with a high-quality tuning for optimal system and network utilization. In this library, a change in the performance is detected, and the adjusted number of concurrent network streams is calculated accordingly. Also, the library provides an interface to the Policy Service for the communication. As a general transfer adaptation library,

- ADT library is integrated into the client tools designated in the project, and will be distributed as a software library under open source license in the next phase.
- ADT library is given a "current" transfer rate, and returns the number of network streams to use, compared to the past adjustments based on the previous research model.
- ADT library is per-process (or per-job) adaptive transfer adjustment module to maximize the data throughput for the job. It has no knowledge of other jobs on the same site or global information on the network.
- ADT library provides an integrated API to the Policy Service so that the client does not need to implement a separate interface to the PS, and uses the transfer parameter advice from the PS in the adaptive transfer adjustment.

The ADT library has been developed as a general transfer adaptation library, and its first version is integrated into the SRM client data movement tool (srm-copy). Its open source release is under way.

2.2.2.1 Data Movement Service

Under the project, we have integrated the ADT library into an existing Data Movement Service (DMS) client: the SRM client tool that is being used in production in Open Science Grid (OSG) and Earth System Grid (ESG). In Year 2 of the project, we implemented enhancements to the srm-copy client tool, including integration with the new Policy Service implementation and more fine-grained control of adaptation based on resources allocations given by the Policy Service. These enhancements and integration show the adaptive data transfer improving the efficiency of data access and resource utilities. In the second phase of the project, the transfer property adaptation algorithm in ADT library has been improved based on the experimental runs in the first phase, and the integration with the centralized Policy Service has been completed. Also, a set of experimental test runs has been designed, and their results have been collected and are being analyzed to improve the effectiveness of the adaptive methods.

2.2.2.2 Policy Service and Library

In Year 2 of the project, we have redesigned and re-implemented the policy logic developed in Year 1 using a RESTful web service approach. This redesign has improved the robustness and scalability of our system and eliminated bugs and race conditions that were discovered in the implementation from year 1. The overall goal of the ADAPT project is to achieve higher throughput for transfers occurring in the environment. Network administrators at the Virtual Organization level define the overall resource limits between hosts based on their resources (e.g. memory size, bus speed, storage speed) and the available network bandwidth between them. These limits are recorded and interpreted by environment-specific policy rules that determine, along with knowledge of the total resources allocated to other ongoing data transfers and their performances, what the actual recommended parameter limits should be for the client instance. As the client performs the transfer, it actively communicates with the Policy Service to receive the most up-to-date advice on transfer parameter limits and informs the Policy Service of its chosen transfer parameters and the current performance.

From the library module that was developed in the first year of the project (phase 1), we have learned the proper policy logic to provide an advice on the number of parallel streams and other transfer rate parameters. The redesigned Policy Service (PS) was implemented in the second phase of the project to provide policy-based transfer advice for file transfers with specified source and destination locations, and to take into account VO policies that place limits on transfer parameters and overall resource usage. We have conducted extensive experiments that seek to isolate policy effects in data transfer performance to study the effectiveness of our policy rules.

2.2.2.3 Passive Measurement Archive

The data movement clients collect the data transfer performance information when the clients transfer files from a data server or replicate a dataset from a source to a destination, and send those collected information to the Passive Measurement Archive (PMA). In the second phase of the project, a perfSONAR based measurement repository service has been set up, integrated and tested as the PMA, so that the historical information can be archived through the common interface.

In the experimental test runs, data transfer performance measurements are continuously logged in a file per client job for simplicity of the tests without the overhead effects to the PMA, and the log format follows the common logging practice guide for Grid logging as defined in http://escholarship.org/uc/item/1jz4k8hd. Usefulness, effectiveness and scalability of the centralized PMA have been studied, and will be tested further as a general transfer log reporting mechanism.

2.2.3 Significant results

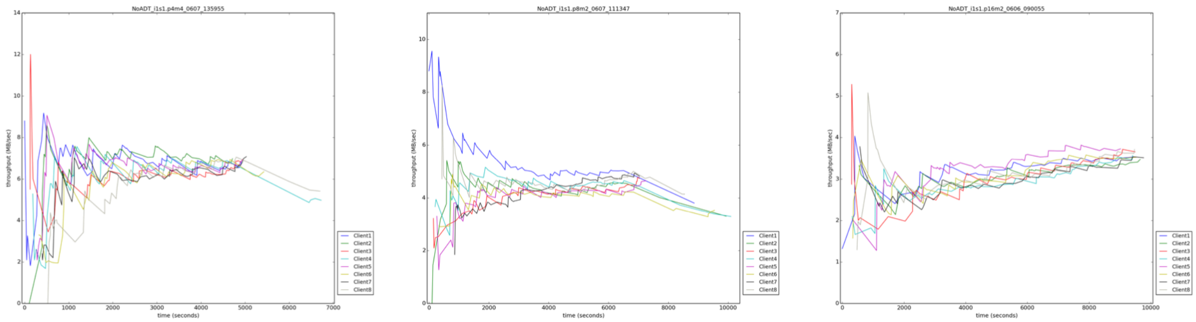
Most of Open Science Grid (OSG) use cases send the data to the OSG resources for data staging before submitting the data analysis jobs. We have collaborated with University of Nebraska at Lincoln (UNL), one of the OSG sites, and the Solenoidal Tracker at RHIC (STAR) experiment at Relativistic Heavy Ion Collider (RHIC) for the experimental test runs. Data are being generated on a large cluster at NERSC/PDSF and are then transferred to UNL. We have run the new enhanced srm-copy jobs on NERSC/PDSF worker nodes, and tested all aspects of the second phase system with GridFTP transfers from NERSC to UNL for data analysis on OSG resources. The use case is similar to the case in Figure 3.

2.2.3.1 Experimental use case and measurements

Data transfer performances with the ADAPT enhancements and the adjusted transfer properties from the second phase of the project have been collected and analyzed in Figure 4-11. Further experiments are still in progress, and we should be able to analyze the effects ADAPT enhancements in greater detail in the coming months.

For each test run, we have limited the following controlled parameters, and varied adaptation rates in clients and policy rule parameters in PS.

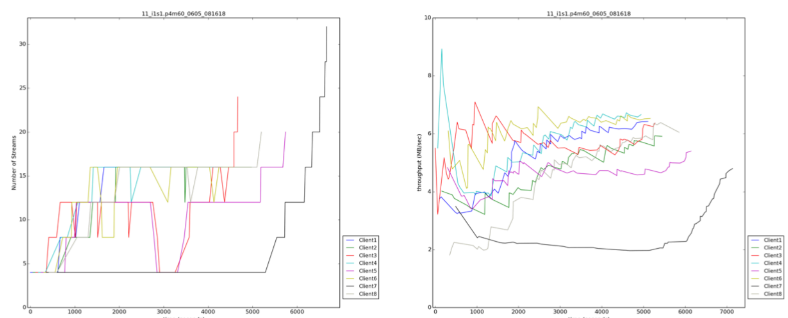
- Number of clients to be 8,
- Each client to transfer 40-50 files ranging from 150MB to 1GB,
- Maximum network streams from the source to destination to be 128,
- Incremental network streams both in clients and PS to be 4,
- Number of parallel streams per file transfer to be 4, and
- Maximum network streams per client to be 32.



(a) Parallel=4, Concurrency=4 (b) Parallel=8, Concurrency=2 (c) Parallel=16, Concurrency=2

Figure 4: The data transfer performance with No ADAPT enhancements.

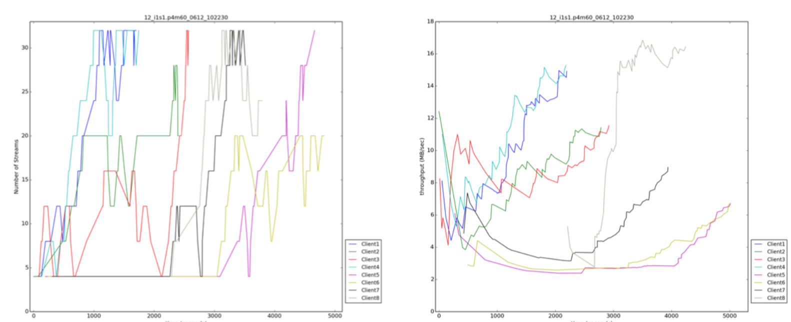
Figure 4 shows that the data transfer performance without any adaptive enhancements. The client jobs took as long as 10,000 seconds in total, shown in (c).



(a) Case 11 run on 06/05/2013 08:16:18

Figure 5: The data transfer performance with ADAPT enhancements. Slow adapt case: parallel streams=4, adaptation delay=after 8 file completions, initial concurrency=1, initial client streams=4, initial PS

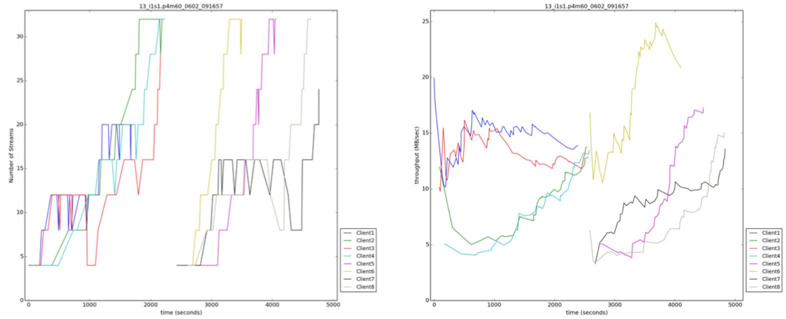
Figure 5 shows a typical adaptation behavior that both client and PS increase the number of network streams to get a better throughput.



(a) Case 12 run on 6/12/2013 10:22:30

Figure 6: The data transfer performance with ADAPT enhancements. Fast adapt case: parallel streams=4, adaptation delay=after 2 file completions, initial concurrency=1, initial client streams=4, initial PS streams=4.

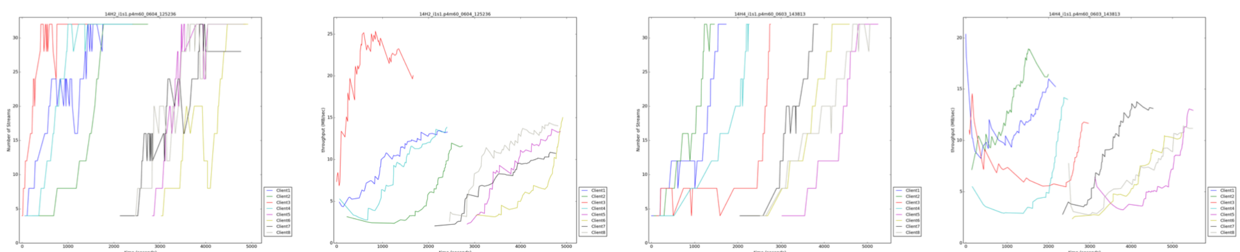
Figure 6 shows that the adaptive behavior responds fast as both client and PS increase the number of network streams to get a better throughput.



(a) Case 13 run on 6/2/2013 09:16:57

Figure 7: The data transfer performance with ADAPT enhancements. Greedy slow adapt case: parallel streams=4, adaptation delay=after 8 file completions, initial concurrency=1, initial client streams=4, initial PS streams=32.

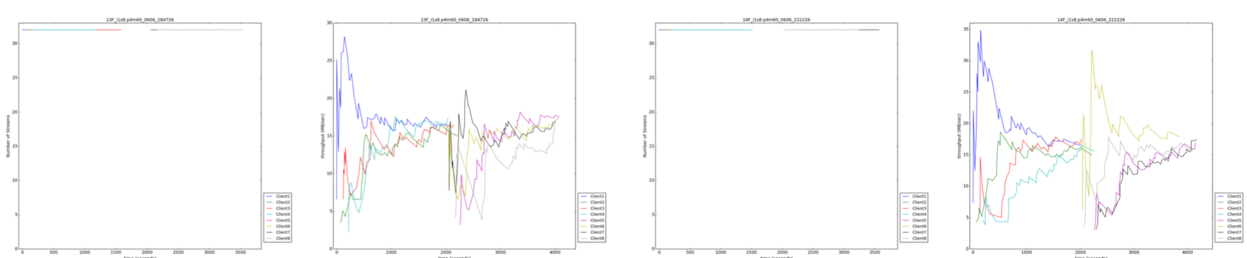
Figure 7 shows the effect of the client-side adaptation. PS allocates the maximum streams for a client job, and the client adjusts the number of network streams accordingly. Because all the available resources that PS manages are initially allocated to the clients, the 5th client waits until they are available.



(a) Case 14 run on 6/4/2013 12:52:36 (b) Case 14 6/3/2013 14:38:13

Figure 8: The data transfer performance with ADAPT enhancements. Greedy fast adapt case: parallel streams=4, adaptation delay=(a) after 2 file completions and (b) after 4 file completions, initial concurrency=1, initial client streams=4, initial PS streams=32.

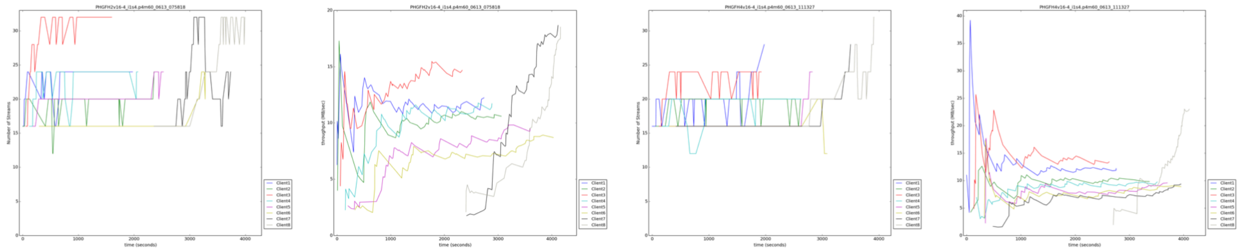
Figure 8 shows the effect of the client-side adaptation with faster adaptation rates than in Figure 7. PS allocates the maximum streams for a client job, and the client adjusts the number of network streams accordingly. Because all the available resources that PS manages are initially allocated to the clients, the 5th client waits until they are available.



(a) Case 15 run on 6/6/2013 18:47:26 (b) Case 15 run on 6/6/2013 22:22:26

Figure 9: The data transfer performance with ADAPT enhancements. Greedy adapt case: parallel streams=4, adaptation delay=(a) after 8 file completions and (b) after 2 completions, initial concurrency=1, initial client streams=32, initial PS streams=32.

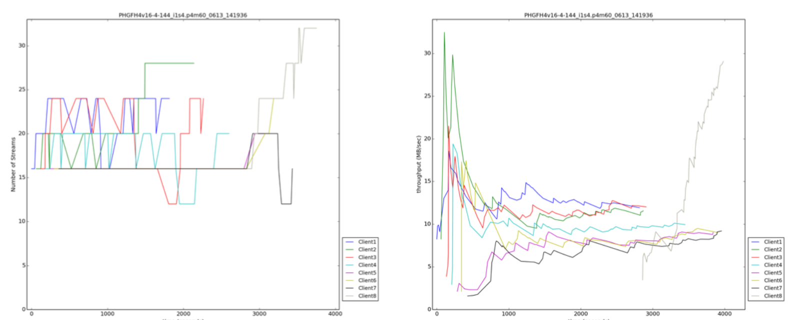
Figure 9 shows the effect of the PS-side adaptation. PS allocates the maximum streams for a client job, and the client allocates the number of network streams according to the PS allocation. Because all the available resources that PS manages are initially allocated to the clients, the 5th client waits until they are available. There is no adaptation behavior showing, as expected.



(a) Case14-2 run on 6/13/2013 07:58:18 (b) Case 14-1 run on 6/13/2013 11:13:27

Figure 10: The data transfer performance with ADAPT enhancements. Greedy adapt case: parallel streams=4, adaptation delay=(a) after 2 file completions and (b) after 4 completions, initial concurrency=4, initial client streams=16, initial PS streams=16.

Figure 10 shows the effect of the PS-side adaptation. PS allocates initially a half of the maximum streams for a client job, and the client allocates the number of network streams according to the PS allocation.



(a) Case 14-3 run on 6/13/2013 14:19:36

Figure 11: The data transfer performance with ADAPT enhancements. Greedy adapt case: parallel streams=4, adaptation delay=after 4 completions, initial concurrency=4, initial client streams=16, initial PS streams=16, maximum network streams from the source to destination=144

Figure 11 shows the effect of the PS-side adaptation with a reservation in the resources. PS allocates initially a half of the maximum streams for a client job, and the client allocates the number of network streams according to the PS allocation.

2.2.4 Key outcomes or other achievements

During the second year of the project, we have developed functional enhancements to the component services based on the general use cases in Open Science Grid (OSG). We reimplemented and deployed the policy logic as a centralized Policy Service, in contrast to the earlier implementation that managed policy locally on each data transfer client. The DMA client, srm-copy was significantly enhanced from the first phase; it connects the PS through a web service provided by Adaptive Data Transfer (ADT) library modules and adapts the number of streams used in data transfers based on the performance of earlier transfers. We also have defined experimental test cases and collected extensive measurements to analyze and improve the effectiveness of the adaptive methods and policy rules. This project has contributed to improving the data accessibility in a large scientific collaboration, providing efficiency in data movement and enforcement of policies for data access and resource sharing.

2.3 What opportunities for training and professional development has the project provided?

The project has provided professional development for three programmers, who have participated in both development and research efforts. These include Junmin Gu, the developer at LBNL who is responsible for the functionality of the adaptive data movement client, and David Smith and Robert Schuler at USC/ISI, who implemented the first and second versions of the policy logic, respectively. This project has provided rich challenges for these developers, advancing their professional skills and development. In addition, a graduate student at USC is doing directed research for Summer 2013 on the project. This will provide his first exposure to graduate level research activities. His initial project has been to help manage, plot and interpret the large amounts of measurement data produced by our experiments.

2.4 How have the results been disseminated to communities of interest?

We participated in the collaboration meetings during the Open Science Grid All Hands Meeting at Indiana University in March of 2013. We presented the results from the first phase, and discussed the further collaboration with sites and VOs. Our goal there was to talk with OSG researchers and infrastructure providers to educate them about our tools and form relationships where our software can be tested on high performance infrastructure. Based on those interactions, we plan to test the latest version of our software on the OSG site at University of Nebraska at Lincoln when our current testing phase is complete.

2.5 What do you plan to do during the next reporting period to accomplish the goals?

Next, we describe our proposed work for the third year of the ADAPT project.

2.5.1 Evaluation and feature improvements

- We will continue testing the functionality, performance and scalability of the ADAPT software between nodes on the Open Science Grid (OSG). Current test runs are performed between NERSC/PDSF and University of Nebraska at Lincoln (UNL).
- We will focus on evaluation and debugging of the software framework, improving the tools based on feedback from users, and release the code as open-source to the community.

2.5.2 Improvements to Adaptive Data Transfer (ADT) library

- Improve the adaptation algorithm and provide user-defined optional features such as an option to decide how often the module communicates with PS to request an updated stream allocation.
- Release ADT library as an open-source library and provide the documentation.
- Provide a reference implementation of a client tool that uses ADT library and adaptation examples.

2.5.3 Improvements to Policy Service (PS)

- Add features to the Policy Service as needed by the users, focusing initially on adding security and improving failure handling.
- Explore additional policies for managing transfer resources. So far, our policies have focused on managing stream allocations between source and destination hosts. We will explore other policies, such as those modifying buffer sizes or that focus on managing throughput. The policies we have studied so far have been focused on maximizing collective throughput for data transfers; we plan to study additional policies that focus on other metrics, such as the

latency for the first file transfers to complete. Such metrics can be important in other context, such as when parts of a large computation are stalled waiting for specific files to arrive.

- Continue experimental evaluation of all policies studied. Measure the effects of the policy service in isolation from the adaptive data transfer client and in combination with that client.

2.5.4 Working with communities

- During the third year of the project, one of our highest priorities will be technology transfer. We will work closely with science communities to help them incorporate adaptive, policy-driven data transfer into their data management workflows. We will start with U. Nebraska at Lincoln and branch out to additional Open Science Grid communities after the software is released.
- We will reach out to the community through workshops/conferences and/or project all hands meetings. We plan to submit a paper in July 2013 on the latest version of our work to the Euro Micro Parallel, Distributed, and Network-Based Processing Conference, with the goal of increasing our work's visibility in the international community.
- We will continue to enhance the srm-copy and Policy Service code and work toward deploying our software within the OSG software stack.

3 Products - What has the project produced?

3.1 Publications

- "*Adaptive Data Transfers that Utilize Policies for Resource Sharing*", Junmin Gu, David Smith, Ann L. Chervenak and Alex Sim, The 2nd International Workshop on Network-Aware Data Management Workshop (NDM2012), Nov 2012.

3.2 Websites

- ADAPT project web: http://sdm.lbl.gov/adapt/

3.3 Other products (Software)

- Adaptive Data Transfer (ADT) library: http://codeforge.lbl.gov/projects/adapt/
- ADT-enhanced srm-copy: http://codeforge.lbl.gov/projects/bestman/, http://codeforge.lbl.gov/projects/adapt/
- Policy Service: https://svn2.misd.isi.edu/repos/policy/trunk/adapt/src/adapt/

4 Participants & Other Collaborating Organizations - Who has been involved?

4.1 What individuals have worked on the project?

Name	Most Project Senior Role	Nearest Person Month Worked
Ann Chervenak	PD/PI	3
Robert Schuler	Developer	2
Junmin Gu	Developer	6
Alex Sim	Co-PI	1
David Smith	Developer	1

4.2 What other organizations have been involved as partners?
Lawrence Berkeley National Laboratory

4.3 Have other collaborators or contacts been involved?
No.

5 Impact - What is the impact of the project? How has it contributed?

5.1 What is the impact on the development of the principal discipline of the project?
This project is focused on computer systems and networking, with the goal of improving the throughput of large data transfers between sites. This use case is typical of the scientific workloads that run in the Open Science Grid. The project has gotten increased visibility in the last year, with interest from participants on the Open Science Grid All Hands Meeting. We are starting to see our paper from the NDM Workshop last year start to appear in the reference lists of papers by submitted by colleagues on other projects. We are encouraged that the ideas that we are exploring in this project are being followed by key players in the larger networking and systems community.

As we complete the current experiments, we plan to distribute the software to the community as open source in the Fall of 2013. We expect to see significant impact on the projects that use our software, and we will collect detailed feedback from the users that will drive further software improvements.

5.2 What is the impact on other disciplines?
The impact on other disciplines will be indirect. By improving data transfer performance, we will enable a range of scientific applications and workflows to run more efficiently on leadership class facilities, national infrastructure such as Open Science Grid, and Clouds.

5.3 What is the impact on the development of human resources?
We will continue to train developers and a small number of students who are working on the project. This project provides exposure to very interesting research and development questions and provides students and developers with experience in running experiments and evaluating the results to drive further software improvements.

5.4 What is the impact on physical resources that form infrastructure?
The goal of this software is to improve the utilization of existing national scale infrastructure such as Open Science Grid and Xcede by maximizing the throughput of data transfers on available infrastructure and reducing the time to transfer large data sets that are needed for computations running on those resources. The software itself is designed to be lightweight and to place minimal burden on the infrastructure, while improving the performance of data transfers already running on that infrastructure.

6 Changes and Problems

6.1 Changes in approach and reasons for change
Nothing to report.

6.2 Actual or Anticipated problems or delays and actions or plans to resolve them
Nothing to report.

6.3 Changes that have significant impact on expenditures
Nothing to report.

6.4 Significant changes in use or care of human subjects

Nothing to report.

6.5 Significant changes in use or care of vertebrate animals

Nothing to report.

6.6 Significant changes in use or care of biohazards

Nothing to report.

7 Special Requirements

Nothing to report.