A Workflow-based Intelligent Network Data Movement Advisor with End-to-end Performance Optimization

November 7, 2013

Contents

1 Project Participants

The project team consists of the following participants at at Southern Illinois University, Carbondale (SIUC): 1. Michelle Zhu (PI) 2. Patrick Brown (Ph.D. student) 3. Liudong Zuo (Ph.D. student) 4. Fei Cao (Ph.D. student) 5. Yang Zhao (M.S. student)

The project is being executed in collaboration with the following participants at University of Memphis (UM): 1. Chase Qishi Wu (PI) 2. Xukang Lu (Ph.D. student) 3. Yunyue Lin (Ph.D. student) 4. Daqing Yun (Ph.D. Student) 5. Vivek Varma Datla (Ph.D. student)

2 Summary

Next-generation eScience applications often generate large amounts of simulation, experimental, or observational data that must be shared and managed by collaborative organizations. Advanced networking technologies and services have been rapidly developed and deployed to facilitate such massive data transfer. However, these technologies and services have not been fully utilized mainly because their use typically requires significant domain knowledge and in many cases application users are even not aware of their existence. By leveraging the functionalities of an existing Network-Aware Data Movement Advisor (NADMA) utility, we propose a new Workflow-based Intelligent Network Data Movement Advisor (WINDMA) with end-to-end performance optimization for this DOE funded project. This WINDMA system integrates three major components: resource discovery, data movement, and status monitoring, and supports the sharing of common data movement workflows through account and database management. This system provides a web interface and interacts with existing data/space management and discovery services such as Storage Resource Management, transport methods such as GridFTP and GlobusOnline, and network resource provisioning brokers such as ION and OSCARS. We demonstrate the efficacy of the proposed transport-support workflow system in several use cases based on its implementation and deployment in DOE wide-area networks.

3 Accomplishments on NADMA/WINDMA System Development

3.1 Introduction

Next-generation collaborative eScience applications often generate large amounts of simulation, experimental, or observational data on the order of terabytes or petabytes at present and exabytes in the predictable future. These data sets must be transferred to different topological locations for various purposes such as data sharing, remote visualization, and distributed analysis. Due to the sheer volume, the data transfer at such a scale requires a controlled high-bandwidth network connection, which, unfortunately, is not readily available over shared IP networks. For example, the resource availability on the Internet is subject to constant changes due to concurrent network traffics, therefore providing very little guarantee on transport throughput or dynamics.

To overcome these limitations, the current development of networking technologies has made it possible to provision dedicated connections with reserved bandwidth in high-performance networks. Recently, several high-performance networking projects are under way to develop such network services, which, however, have not found a large user base in broad science community as initially expected mainly because: i) the use of these services typically requires a considerable level of knowledge for network and host configurations that most scientific users often lack; ii) many users are even not aware of the existence of such advanced networking services and resources due to the communication gap between different technical domains. Of course, as domain experts with their own research missions, scientific users should not be expected to explore the availability of special or secret networks and deal with the complexity of network/host configurations in the first place. Ideally, the application or the network itself in the case of software-defined networking (SDN) should be smart enough to make a choice if the user is able to provide some simple information such as a target use case or the expected duration/amount of data flow. However, the reality of the current situation is that a substantial majority of application users are still using old-fashioned tools (e.g. HTTP through a default IP path) that they are familiar with from their empirical studies for their daily data transfer needs.

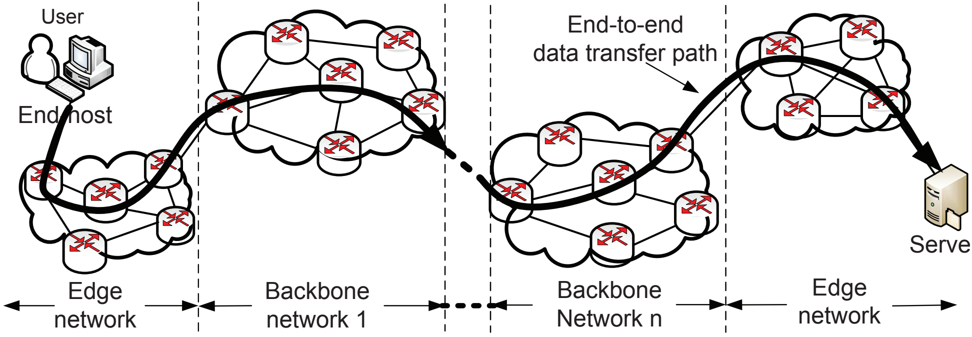


Figure 1: The network infrastructure for wide-area bulk data transfer across multiple domains.

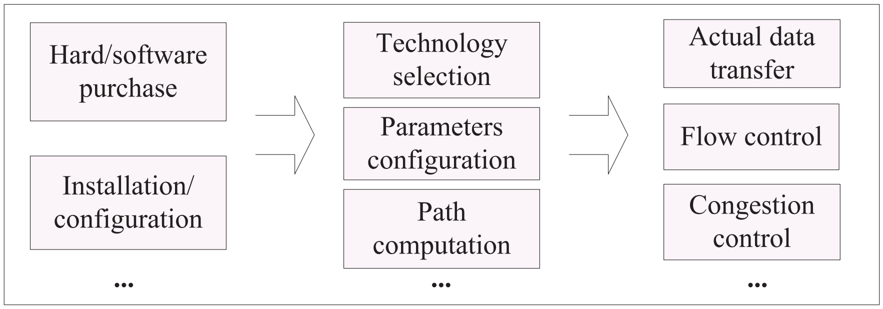


Figure 2: Multiple steps in a data transport solution.

As illustrated in Fig. 1, driven by the continuously expanding scope of scientific collaboration, many eScience applications require wide-area bulk data transfer across multiple domains over different network segments such as edge and backbone networks from a source end node to a destination end node. However, in most cases, we only have insight into or control over a portion of the entire end-to-end path. To meet a specific user request for data transfer, we have to take multiple steps to acquire and deploy the right system hardware/software, select the suitable technologies based on available resources, determine the best data transfer path, and perform the actual data movement, as shown in Fig. 2. Note that system and network resources vary significantly in their type, cost, performance, reliability, and security. For example, an end host might be equipped with multiple network interface cards (NIC) of different speed and cost; OSCARS in ESnet (1; 2; 3) and ION in Internet2 (4) provide different levels of bandwidth provisioning services at different cost and admission rate.

The goal of our project work is to provide users an integrated solution for resource discovery, path composition, and data movement. By leveraging the functionalities of an existing Network-Aware Data Movement Advisor (NADMA) utility (5), we propose a transport-support workflow system to facilitate large data transfer with end-to-end performance optimization. In particular, we introduce a Workflow-based Intelligent Network Data Movement Advisor (WINDMA) utility to augment NADMA to interface with various network services to set up circuits and utilize appropriate transport methods for actual data transfer in different network environments. This new system shares some common functions with the previous NADMA in the service discovery aspect. However, many new features including workflow generation and managements, performance estimation and optimization have been added to WINDMA to enable intelligence in choosing the best possible networking service. Our system operates at the application level and thus does not have the authority to decide on the particular network routing path which is determined by the underlying networking

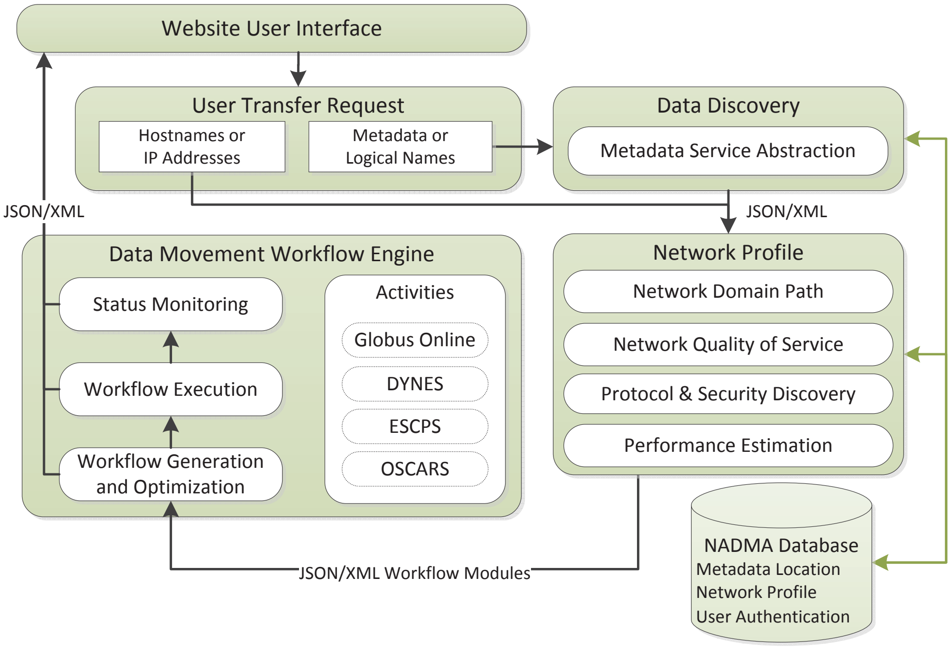


Figure 3: The WINDMA framework: functional components and control flow.

infrastructure of traditional networks. While the current WINDMA implementation seeks to optimize the use of current generation technologies such as dynamic circuit reservation, the modular design and cost model of WINDMA encourage our future exploration into next generation software defined networking, particularly OpenFlow technology, to define the dynamic path based on current traffic.

Within our WINDMA framework, the user only needs to submit a request that describes the data transfer requirements such as the service start and end time, the data source and destination nodes, a desirable bandwidth, or possibly a financial cost limit on the deployment and utility expenses*. These parameters are both typical of the requirements found in dynamic circuit provisioning services as well as useful in the construction of deadline-based transport solutions. Upon the receipt of such a request, our system invokes NADMA functionality to explore the available services and resources in end systems, edge segments, and backbone networks, models them as transport-support workflow modules with quantified parameters, and composes an optimal network path for end-to-end data transfer with different objectives such as cost (financial or technical), delay, and reliability. We implement and test the proposed transport solution in real network environments. The experimental results show that our method can achieve a reasonable accuracy in modeling existing services and improve the performance of data transfer.

The rest of the report is organized as follows. Section 3.2 presents the WINDMA architecture and details its functional components. Section 3.3 presents the transport workflow optimization method. Section 3.4 describes the system implementation and operation procedure. Section 4 presents the design of a fast and simple transfer tool, namely FAST. Section 5 concludes our work.

3.2 The WINDMA Architecture

As shown in Fig. 3, the WINDMA framework consists of a group of distinct components written in python that interact with users through a frontend web interface built on the Drupal (6) content management

*Even though most network services and resources are not free, their financial cost is quite minimal and is often negligible, especially in shared IP networks. Some advanced services such as OSCARS in ESnet and ION in Internet2 are currently free to authorized users, but it is predictable that some accounting components will be integrated into these services in the future.

system. The python components feature data discovery, network profiling, and workflow generation and optimization while remaining independent from each other, communicating input and output using XML and JSON. A website frontend, enabled through Drupal, interprets the outputs and provides a simple web interface that abstracts the interaction of individual components. The independent nature of each component allows for WINDMA to operate as a library of functional components, available to existing systems, or as an independent tool behind a frontend interface as presented in this report.

Each component in the framework is sensitive to the context of its operation, allowing for flexible functionality in both a local and a remote context. The local context allows WINDMA to be distributed as a downloadable client-side tool or library that can operate in a trusted local context with deeper integration and utilization of third-party client-side tools such as data transfer client software. The remote context enables WINDMA to be used in a remote system, such as a website, which may have limited user trust. The automation abilities of WINDMA respect the context of operation and intelligently automate tasks only if the context permits it.

A user data transfer request may be defined as a physical or logical address. The *Data Discovery* component translates any logical data into physical addresses before the request is interpreted by the *Network Profile*r. The Network Profiler discovers the network domain path, quality of service, protocols, and security mechanism capabilities of the transfer request by probing end host resources and networks while querying a central database for known network capabilities. WINDMA forwards this information to a *Data Movement Workflow Engine* component that provides the ability to construct, optimize, execute, and monitor a data movement workflow that is capable of performing the data transfer request. Here we will briefly review the Network Profiler, Data Discovery, and Database components of NADMA that are used by WINDMA and already discussed in (5) and provide a summary of the new Data Movement Workflow Engine of WINDMA while leaving the detailed discussion to Section 3.3.

3.2.1 Network Profiler

To provide accurate advising for data movement in dynamic network environments, it is critical to collect and store status and resource information including network domain topology, provisioning services, and data management and movement protocols, which must be updated in a timely manner. For this purpose, we create and maintain a database that contains information enabling the construction of a network profile. This information includes organization references, hostname and IP address subnets, network domain topology, and network technologies and capabilities of end sites.

The network quality of service is automatically discovered by interacting with the WINDMA database to determine if an end host has access to high-performance network infrastructure and necessary provisioning services that are capable of provisioning dedicated bandwidths. The system also supports scanning end hosts to discover system and network resources, including a variety of transport protocols, absent from the database. The end host resources and networking service technologies that are discovered by the Network Profiler component, such as ESnet OSCARS and Internet2 ION (as enabled by projects such as DYNES (7)), are organized into discrete modules and used by the Data Movement Workflow Engine during workflow construction and optimization.

3.2.2 Data Discovery

Since the physical location of a particular dataset may be unknown to the user, WINDMA provides a data discovery capability whereby the user can query third-party data services by keywords to locate a dataset of interest. This component interacts with metadata services using web service interfaces to query and discover the location of datasets as discussed in (5).

3.2.3 Database

The storage and retrieval of network domain topology, metadata location, authentication mechanisms, and known network capabilities are made possible through an SQL database. As discussed in (5), the database contains a collection of network and protocol information that has been automatically retrieved

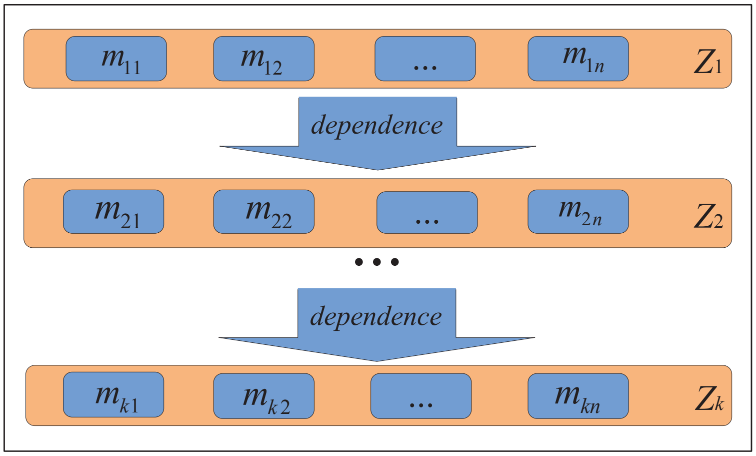


Figure 4: A zone-based transport-support workflow structure.

from known web service sources or manually entered into the database. Unlike NADMA, the database of WINDMA is transportable, enabling WINDMA to be used from both a client-side context as a local utility program and a server-side context as a library to an existing web portal such as Drupal. The SQL database supports both MySQL for server-side contexts as well as SQLite (8) for client-side contexts.

3.2.4 Data Movement Workflow Engine

The resources discovered by the Network Profiler and Data Discovery components are used to generate, optimize, execute, and monitor data movement workflows in the Data Movement Workflow Engine. The resources provided to the engine are modeled as data movement workflow modules. The self-contained engine uses these modules to generate data movement workflows by formulating a workflow optimization problem as discussed in Section 3.3.

The data movement workflow consists of a dependency graph of tasks that must be completed for successful data movement using the subset of modules selected during workflow optimization. This task graph is transformed into a set of activities that fulfill the specific tasks to support workflow execution and status monitoring. The activities utilize client tools and third-party web services to support file transport protocols such as GridFTP and high-performance bandwidth reservation systems. Each activity implements a well-defined interface that allows for interchangeability while abstracting the specific, often messy, operation of the underlying tools and services.

3.3 Transport-Support Workflow Optimization

3.3.1 General Purpose

Emerging high-performance networking technologies have been rapidly developed and deployed to support the transfer of large data sets generated by next-generation scientific applications for collaborative data processing, analysis, and storage. Unfortunately, due to the lack of computer and network knowledge, scientific users have not been able to fully utilize these resources. We model various types of resources discovered by Network-aware Data Movement Advisor (NADMA) in end systems, edge segments, and backbone networks and based on which we formulate Transport-Support Workflow Optimization Problem (TSWOP) considering a comprehensive set of performance metrics and network parameters in different phases including device deployment, circuit setup, and data transfer. We propose an integrated solution to choose an appropriate set of technologies and services to compose the “best” transport-support workflow such that to provide end-users with advice to meet the user’s data transfer requirements.

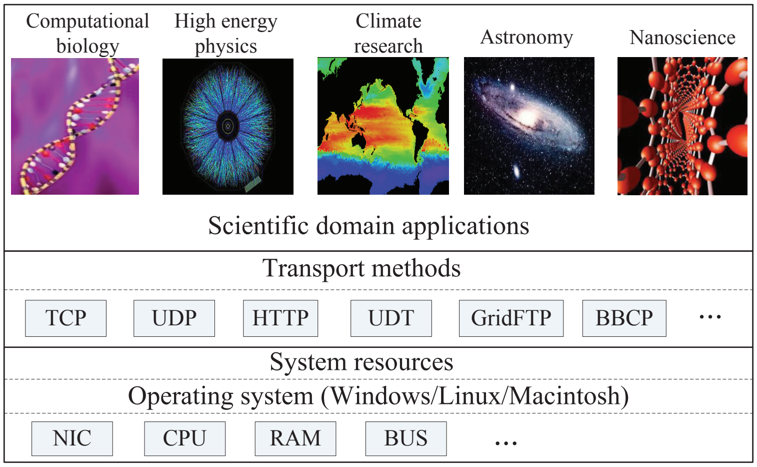


Figure 5: A general structure of end host modules.

3.3.2 Math Models of Transport-Support Workflow Modules

We model the entire data transfer process with a zone-based structure as shown in Fig. 4. The data transfer process is divided into k zones in a logical space, and we categorize those workflow modules with execution dependencies into one of zones. Each module represents a certain type of resource or task that must be used or performed to meet user's data transfer request.

End Host Modules

At the end host, typically three steps across three zones would be invoked during packet receiving process: (i) a data packet arrives at the NIC and generates an interrupt, (ii) the kernel traps the interrupt and reads the packet from the NICs buffer to the transport protocol buffer, and (iii) the transport protocol processes and forwards the packet to the target user applications. Thus the end host modules are modeled within three zones, namely, system resource, transport method, and user application, as shown in Fig. 5.

- The modules in the system resource zone include both hardware such as CPU, network interface card (NIC), and RAM, and system software such as operating systems;
- The modules in the transport method zone include application-layer transport protocols, kernel-level transport protocols, and other network services and resources. In Fig. 5, we model the application-layer protocol HTTP as a module that runs over the kernel-level transport protocol TCP, which is also modeled as a module. We place them in the same zone as both of them are transport protocols providing services to user applications. We would like to point out that our zone-based structure is flexible in that we can further divide the modules in each zone into more subzones as in the case of TCP/IP stack.
- The modules in the user application zone include all user applications that require data transfer services. These applications come from a wide range of disciplines spanning from climate research, nanoscience, astronomy, neutron sciences, high energy physics, computational materials, fusion simulation, to computational biology.

3.3.3 Networking Service Modules

A networking service module could be a technology, a mechanism, or a hardware/software system, which takes the users request as input, performs some predefined routines, and sends back to the user the resources and/or other relevant results under request. The common networking modules provide end users either a default IP or a network provisioning service with guaranteed bandwidth such as OSCARS in ESnet (1) and ION in Internet2 (4). Several common networking service modules are listed in Tab. 1 (9; 10; 11; 12; 13; 14;

Table 1: Networking technologies/services/resources.

Modules	Remark
MPLS	label switching based link-level technology
VLAN	virtual LAN, regardless of the physical location
IP routing	the default IP path
OSCARS	bandwidth reservation within ESnet
ION	bandwidth reservation within Internet2
DYNES	edge network bandwidth reservation for Internet2
DRAGON	using MPLS technology
CHEETAH	circuit-switched
TeraPaths	end-to-end virtual path with bandwidth guarantees
ESCPS	provisioning end-to-end inter-domain dynamic circuits
UCLP	network resources treated as software objects
JGN(2/2plus)	fully-fledged next-generation testbed for research
Geant2	funded by NRENs and EC, testbed for research
HOPI	combining packet- and circuit-switching
GENI	virtual laboratory for exploring future Internet

15; 16; 17; 18; 19; 20). Typically, some of these networking services utilize graph-based algorithms, taking into account the parameters and constraints specified in the Virtual Circuit reservation, such as source and destination endpoints, bandwidth, or VLAN tagging to compute a path for a reservation request.

Since most of these services are not free and even though some advanced services such as OSCARS in ESnet and ION in Internet2 are currently free to authorized users, it could be predicted that some accounting components would be integrated in the near future. Therefore, upon the speficication of user desired services, we could calculate the financial cost according to the charging rules of the services provider. Because end users tend to be greedy, even though the services are not free there is no hard guarantee the desired services would be available for all of user requests. In such case, end user would be informed with the information that the desired services are unavailable.

To compute the data transfer path, the networking modules would take the network topology with capacity information and the current resource reservation as input, return the end host user with desired services together with the financial cost or the services rejection information.

3.3.4 Technical Approach

User Request

A user request R specifies the desired data transfer service such as transfer start time t_s, transfer finish time t_f, source host address H_s, destination host address H_d, and transfer data size DS as well as some data transfer constraints and objectives such as loss rate LR, required bandwidth BW and financial cost upper bound spending on the networking service C_{net}. We use an *m-tuple* $R = (r_1, r_2, , r_m)$ to model a generic user request, where r_i $(1 \leq i \leq m)$ are user-specified parameters that describe demanded services and constraints. We define R as a corpora includes all the parameters that a user may care about. Certainly a specific user request may only involve a subset of parameters, in which case, we can assign 0 or *null* to those parameters that are not under consideration. *Transport-Support Workflow Optimization Problem (TSWOP)*

- **Calculation of Credit:**
  Depending on the modules' properties, the parameters r_i in the user request may be fulfilled at a different degree, which reflects the satisfaction for the data transfer requirement. Our goal is to select

a subset of modules discovered by NADMA across all the zones to meet the user request R as much as possible.

We define a 0-1 vector $X = (x_1, x_2, ..., x_m)$ for a specific module m corresponding to user request R, where x_i is 1 when module m is selected from its zone and it fulfills the user request parameter r_i, x_i is 0 when m is selected but it cannot fulfill r_i. We associate each module with a vector X, which indicates its *credit* of fulfilling the parameters of a user request R. Ideally, we wish to select modules that completely satisfy a user request for all of the objectives and constraints. However, due to the limited networking resources and conflictive parameters that may be specified by scientific domain users who are oftentimes domain experts without sufficient networking knowledge and may not be able to always provide reasonable or realistic user requests, it is generally infeasible to select such modules upon such cases.

For each selected module m, we calculate its credit P_m using the following equation:

$$P_m = f(X_m) = \sum_{i=1}^{m} \alpha_i x_m(i), \tag{1}$$

where α_i is the weight for each parameter r_i in R which is either decided by end user according to the user's importance preference among different parameters or automatically assigned with the same values which indicate the same importance among parameters by our model.

- **Definition of TSWOP:**

  Definition 1. *(TSWOP) Given a user data transfer request $R = (r_1, r_2, ..., r_m)$, its corresponding weight vector $\alpha = (\alpha_1, \alpha_2, ..., \alpha_m)$, and n workflow modules categorized into k zones of the data transfer process, together with a predefined 0-1 credit vector $X = (x_1, x_2, ..., x_m)$ for each module. We wish to select a subset of modules across the k zones such that the user request can be successfully met with the maximal credit computed by Eq. 2,*

$$\begin{aligned} Credit_{\max} &= \max_{\substack{all\ possible\\ module\ selections}} \left(\sum_{i=1}^{k} P_{m_i}\right) \\ &= \max \sum_{i=1}^{k} \sum_{j=1}^{m} \alpha(j) \times X_{m_i}(j) \end{aligned} \tag{2}$$

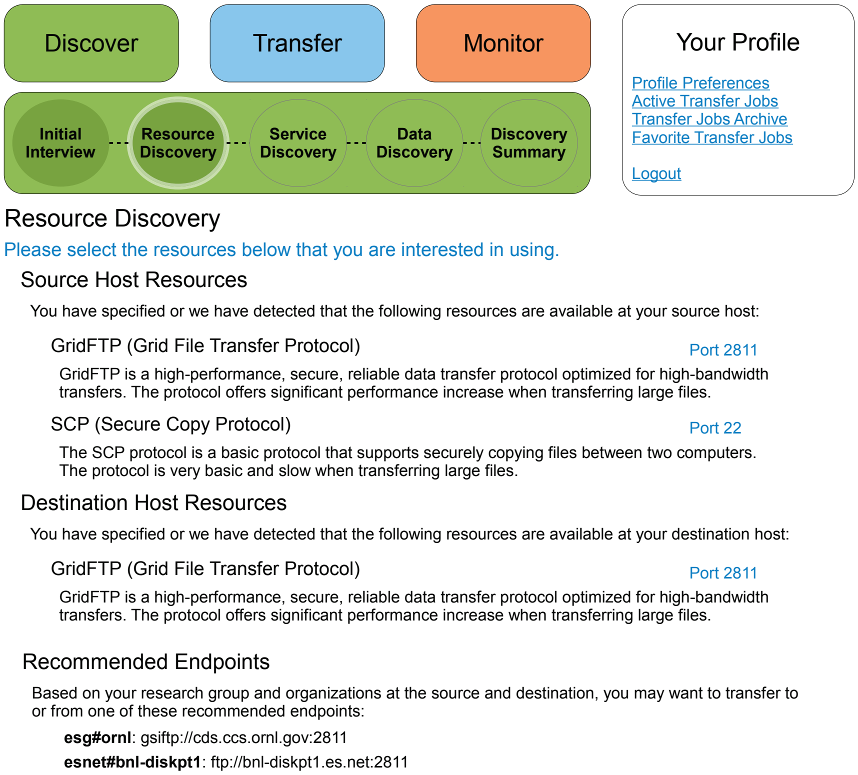
  and under the financial upper bound constraint Cost,

$$\sum_{i=1}^{k} C(m_i) \leq Cost. \tag{3}$$

  in Eq. 3, m_i is the selected module from zone i, and $C(m_i)$ is the financial cost caused by using service m_i.

- **Estimation of X:**

  In our model presented above, given a module it is critical to predefine the parameters in X since they affect the module selection and eventually the data transport performance. However, indeed some of these parameters may be straightforward while others may not. For example, it is relatively easier to determine whether a module could provide a reliable data transfer service than to ensure whether a certain failure rate or loss rate requirement could be satisfied. To make it more clear, let us consider OSCARS as an example, when a user request that ask for a reserved bandwidth is received, OSCARS generates a base topology graph considering the parameters and constraints specified in user's reservation request such as source and destination hosts, required bandwidth or VLAN tagging, in such case, it is easy to decide whether the requirement could be met. While on the other hand, if the user submits a request which has some requirements on failure rate or loss rate, it is not straightforward



Resource Discovery

Please select the resources below that you are interested in using.

Source Host Resources

You have specified or we have detected that the following resources are available at your source host:

GridFTP (Grid File Transfer Protocol) Port 2811

GridFTP is a high-performance, secure, reliable data transfer protocol optimized for high-bandwidth transfers. The protocol offers significant performance increase when transferring large files.

SCP (Secure Copy Protocol) Port 22

The SCP protocol is a basic protocol that supports securely copying files between two computers. The protocol is very basic and slow when transferring large files.

Destination Host Resources

You have specified or we have detected that the following resources are available at your destination host:

GridFTP (Grid File Transfer Protocol) Port 2811

GridFTP is a high-performance, secure, reliable data transfer protocol optimized for high-bandwidth transfers. The protocol offers significant performance increase when transferring large files.

Recommended Endpoints

Based on your research group and organizations at the source and destination, you may want to transfer to or from one of these recommended endpoints:

esg#ornl: gsiftp://cds.ccs.ornl.gov:2811
esnet#bnl-diskpt1: ftp://bnl-diskpt1.es.net:2811

Figure 6: WINDMA protocol discovery summary and selection.

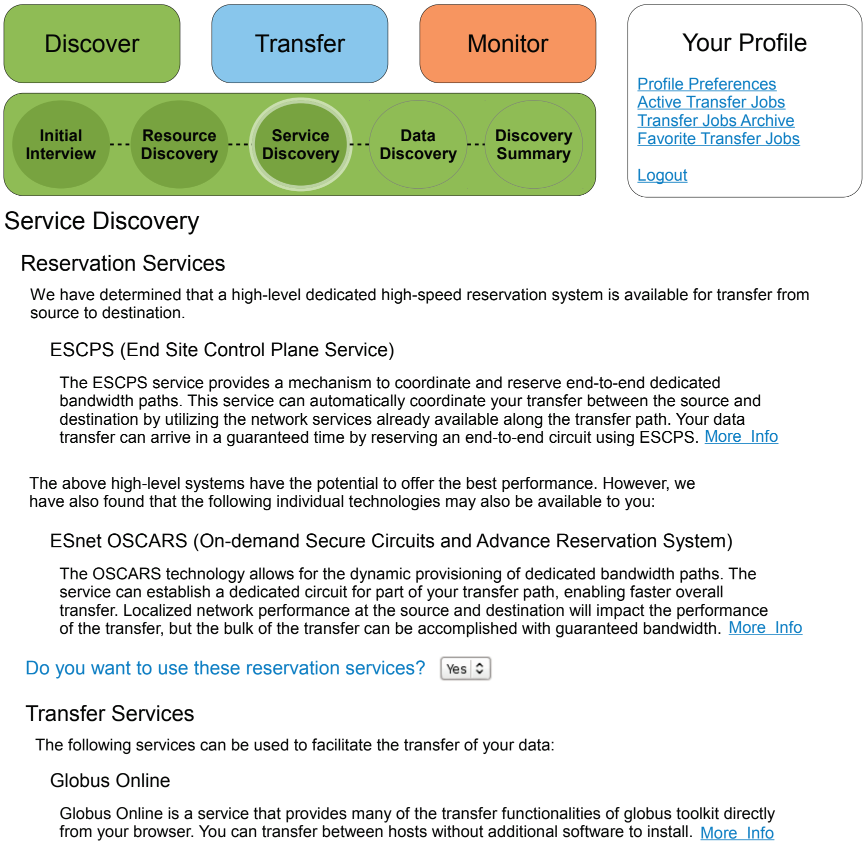
to decide if such requirements could be met. A feasible strategy is to use historical data to estimate and predict the performance of a service, which is used to determine the corresponding parameters in X. The estimation of X need to be further investigated rigorously which would possibly be the major challenge of this model.

3.4 System Implementation and Operation Procedure

The WINDMA transport-support workflow system is exposed through the browser by a web-based frontend built on a LAMP stack, utilizing the popular open source Drupal content management system to provide routine administrative and user functionality. Queries are made against an SQL server to store and retrieve information about known advanced network resource storage and provisioning systems as well as determine network domain capabilities of end host networks. The website offers three ordered phases of Discovery, Transfer, and Monitoring that provide a general means for users to submit data movement requests based on various transport methods, discover network technologies of the end hosts and intermediate networks, and automatically construct and execute optimized data movement workflows. To better illustrate the operation of WINDMA, in each phase we will consider an example case where the user desires to transfer data between a source host and a destination host on the Energy Sciences Network (ESnet), located at Brookhaven National Laboratory (BNL) in New York and Lawrence Berkeley National Laboratory (LBNL) in California.

3.4.1 Discovery Phase

WINDMA requires the entry of the source endpoint and destination endpoint, one of which may be the user's local computer, for the desired data transfer. The two endpoints are sufficient for WINDMA to provide a detailed analysis, but the user may optionally provide additional information about their network or endpoints including network capabilities and supported transfer protocols. Generic information about the nature of the user's end hosts and local area network capabilities are stored in the database as learned



Service Discovery

Reservation Services

We have determined that a high-level dedicated high-speed reservation system is available for transfer from source to destination.

ESCPS (End Site Control Plane Service)

The ESCPS service provides a mechanism to coordinate and reserve end-to-end dedicated bandwidth paths. This service can automatically coordinate your transfer between the source and destination by utilizing the network services already available along the transfer path. Your data transfer can arrive in a guaranteed time by reserving an end-to-end circuit using ESCPS. More Info

The above high-level systems have the potential to offer the best performance. However, we have also found that the following individual technologies may also be available to you:

ESnet OSCARS (On-demand Secure Circuits and Advance Reservation System)

The OSCARS technology allows for the dynamic provisioning of dedicated bandwidth paths. The service can establish a dedicated circuit for part of your transfer path, enabling faster overall transfer. Localized network performance at the source and destination will impact the performance of the transfer, but the bulk of the transfer can be accomplished with guaranteed bandwidth. More Info

Do you want to use these reservation services? Yes

Transfer Services

The following services can be used to facilitate the transfer of your data:

Globus Online

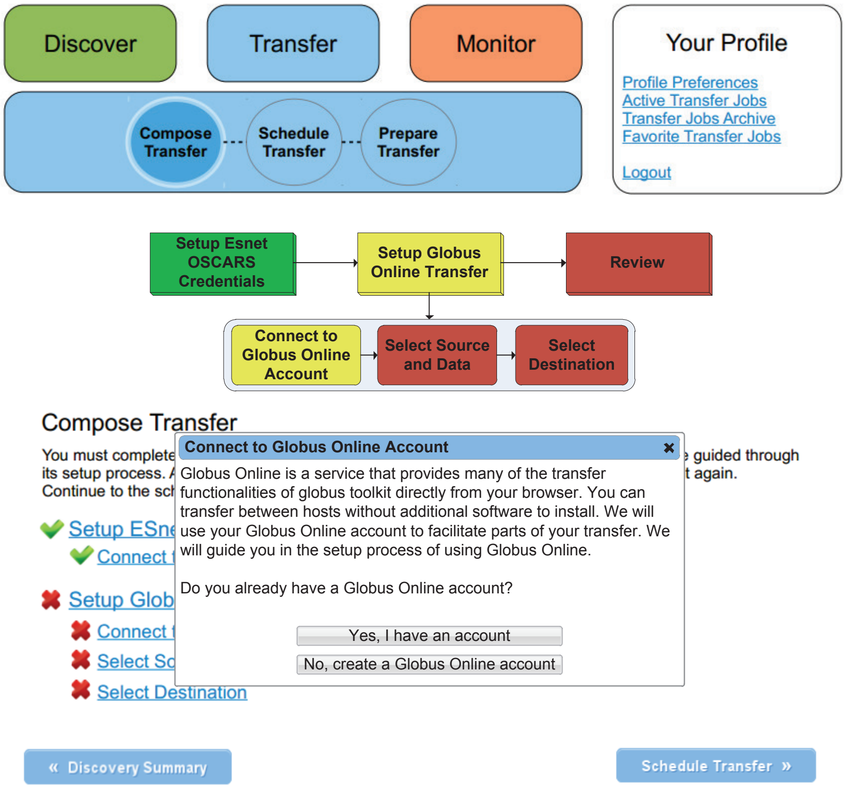
Globus Online is a service that provides many of the transfer functionalities of globus toolkit directly from your browser. You can transfer between hosts without additional software to install. More Info

Figure 7: WINDMA network and circuit reservation service discovery.

information that can be utilized in future discovery processes.

If an endpoint is unknown, the Data Discovery component may be used to locate the desired dataset. The Data Discovery component interacts with metadata services to locate and retrieve physical location information about the desired dataset. The user may query the target metadata service for a dataset of interest by keyword or a dataset identifier. For demonstration purposes, the Data Discovery of WINDMA currently supports the Earth System Grid (ESG) (21) metadata portals. The user submits a keyword query, selects from a list of matched datasets, and WINDMA automatically populates the appropriate endpoint and dataset information.

The endpoints given by the user or discovered by WINDMA as well as optional end host and LAN network information are processed by the website to begin the discovery phase. WINDMA uses this information together with the Network Profile component to discover the network location and organization associated with the source and destination endpoints. The organization and network locations are used to build a network profile. The network domains and transport protocols supported along the possible paths from source to destination are determined in the Network Profile component, yielding a collection of End-host and Network Service modules that represent the capabilities of the end-host local networks as well as the wide-area networks between them. These capabilities, including the availability of high-performance networks and specific protocols, are presented to the user with accompanying information describing each specific technology. The user can use this information to decide which technologies they want to use to enable their data movement or they can allow WINDMA to generate a recommended data movement workflow



Connect to Globus Online Account

Globus Online is a service that provides many of the transfer functionalities of globus toolkit directly from your browser. You can transfer between hosts without additional software to install. We will use your Globus Online account to facilitate parts of your transfer. We will guide you in the setup process of using Globus Online.

Do you already have a Globus Online account?

Figure 8: WINDMA Transfer Phase showing precondition steps guided in a constructed workflow.

using the techniques discussed in Section 3.3.

In our example case, the user specifies the source host at Brookhaven National Laboratory and a destination host at Lawrence Berkeley National Laboratory. WINDMA uses the host information in conjunction with the Network Profile component to discover specific information about each end-host. The combination of our protocol scanning and known ports of the end hosts indicate the availability of SCP and GridFTP transfer protocols, as shown in Fig. 6. The organizations found in WINDMA's database reveal and the characteristics of the physical sites found using WINDMA's network quality of service discovery indicate the availability of high-performance networks at both the source and destination using ESnet OSCARS (1), as shown in Fig. 7.

Based on the discovered information, the user selects the network technologies, including high-performance network systems and specific protocols, to use for the data movement workflow. This selection is submitted to WINDMA's Data Movement Workflow Engine and a data movement workflow is generated by selecting an optimized subset of modules as discussed in Section 3.3. The workflow consists of a dependency graph of tasks that must be performed to enable desired network technologies and carry out the actual data transfer. This task graph can be realized as both a manual set of instructions as well as a template for the automated workflow execution functionality offered by WINDMA in the Transfer Phase. In our example case, the discovered technologies and specific transfer constraints result in a distinct workflow that recommends using OSCARS to facilitate a high-performance transfer over a dynamically provisioned virtual circuit using GridFTP as the transfer protocol.

3.4.2 Transfer Phase

The execution of data movement workflows can be performed directly from the WINDMA website using the Data Movement Workflow Engine component. The task graph workflow constructed in the Discovery phase is realized as a set of jobs to be executed by various transfer tools and third-party services available to WINDMA. A standard activity interface is implemented for different task classes and functions that allow for the execution of interchangeable components in the workflow graph. For example, the task of transferring data using GridFTP can be fulfilled by a Globus Online (22) implementation that exposes an activity interface for a GridFTP transfer. Activity implementations may also be refined by the network and host constraints present in the workflow, ensuring that only the activity implementations that can fulfill this specific workflow are considered.

The standard activity interface allows for implementations to feature both the use of standard client tools as well as third-party web services such as Globus Online for third-party transfer support. The web APIs of dynamic circuit provisioning systems found in high-performance networks also expose the ability to dynamically provision end-to-end dedicated circuits. Each activity uses one or more of these tools or web services to provide a function to fulfill a specific task. The activity itself may contain a series of preconditions that must be satisfied, such as obtaining necessary security credentials for a particular service. WINDMA guides the user in satisfying these conditions from the website.

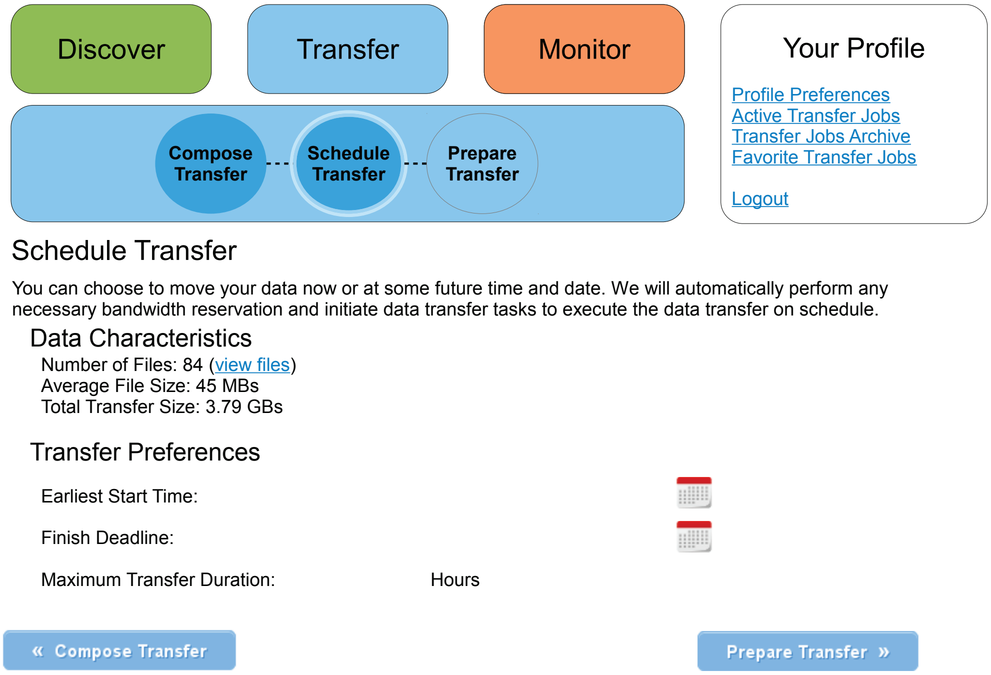
Continuing the example case from the Discovery phase, the proposed transfer from Brookhaven National Laboratory to Lawrence Berkeley National Laboratory produced a task graph that requests an activity for fulfilling a dynamic circuit reservation for a transfer from an ESnet site to another ESnet site using OSCARS. Since this request can be fulfilled through ESnet OSCARS, the OSCARS reservation activity is utilized to expose OSCARS-enabled dynamic circuit reservation to WINDMA utilizing the OSCARS web service APIs. Similarly, the need for a GridFTP transfer can be fulfilled by a Globus Online activity implementation using a layer-3 reservation. Both of these implementations have preconditions of credential gathering and setup that must be completed by the user before workflow execution is possible. WINDMA presents these preconditions as steps in the transfer setup process, as shown in Fig. 8. The user is also provided the capability to specify the data transfer deadline, if needed, as shown in Fig. 9.

In addition to credential management, steps including service authentication, dataset selection, and transfer preferences may be necessary. The user is guided through these steps from the browser as WINDMA uses corresponding tools and web services to complete the preconditions. After satisfying the preconditions of the activities associated with the task graph, the user instructs WINDMA to execute the data movement workflow and the workflow is executed using the associated activities to satisfy the dependencies of the task graph. In our example case, the task graph indicates that file transfer is dependent on dynamic circuit provisioning. Therefore, WINDMA executes the OSCARS activity and waits for its success condition before executing the Globus Online activity.

3.4.3 Monitoring Phase

WINDMA monitors the execution of the workflow using the same activity implementations used to facilitate the Transfer Phase. Activities that support monitoring certain metrics such as transfer speed expose their data to WINDMA through a monitoring interface. The specific activity implements specific monitoring capabilities and WINDMA combines the metrics available from each activity associated with the task graph into a single report. The results are interpreted by WINDMA to extrapolate performance of the current workflow and can be stored as historical indicators of performance of future workflows as discussed in Section 3.3.

The outcome of the workflow execution is similarly monitored by WINDMA. In the event of an error when executing the workflow, the activity that propagated the error determines if the system can be recovered. WINDMA will attempt to resolve the error automatically if possible, otherwise user intervention may be required. For example, the source dataset could be moved on the source machine during a transfer, causing



Schedule Transfer

You can choose to move your data now or at some future time and date. We will automatically perform any necessary bandwidth reservation and initiate data transfer tasks to execute the data transfer on schedule.

Data Characteristics

Number of Files: 84 (view files)
Average File Size: 45 MBs
Total Transfer Size: 3.79 GBs

Transfer Preferences

Earliest Start Time:
Finish Deadline:
Maximum Transfer Duration: Hours

Figure 9: Data transfer deadline specification in WINDMA.

subsequent file transfers to fail. The activity responsible for the file transfer may present this situation as a recoverable error with guided user interaction to recover and continue the execution of the workflow.

The monitoring of the example case is provided by both the OSCARS and Globus Online activities. The OSCARS activity contacts the OSCARS web service to reassure the user of the dynamic circuit provisioning while the Globus Online activity contacts the Globus Online web service to monitor the progress of individual file transfers. The results are appropriately stored in the database for historical reference and displayed to the user.

3.4.4 Collaboration

The web interface of WINDMA facilitates collaboration among the members of a distributed research team. WINDMA takes advantage of the Drupal framework by allowing users to be organized into research groups, enabling collaboration among users and groups through the sharing of data movement workflows. Since the data movement workflows constructed by WINDMA can be represented as a graph of generic tasks, workflows can be shared between users without worry of credential exposure or unintended authorization. Users can share constructed workflows as task graphs with other users. Upon sharing, the task graphs can be mapped into appropriate activities based on the user's preferences or system access capabilities.

In the example case, the user shares the existing data movement workflow stored as a task graph with a new user. WINDMA automatically associates the graph with new activities specific to the new user. If the new user indicates a different set of constraints for her transfer, such as lacking system access to a dynamic circuit provisioning system, WINDMA can appropriately choose the activities that fulfill the workflow using the new set of constraints. In this example, the OSCARS activity would not be used.

4 New Fast and Simple Transfer Tool Design (FAST)

FAST (FAst and Simple Transfer) is a new tool that will enable fast and simple transfers by leveraging existing data transfer technologies such as Globus Toolkit, Globus Online, Fast Data Transfer (FDT) (23), SCP, WGET, OSCARS, DYNES (7), and etc. Using FAST, users can facilitate bulk data transfer between end-to-end connections through simply issuing a simple command-line, just like what wget does, without caring about lower level actions or procedures such as discovering networking resources, setting up the data transfer path, performing the actual data movement.

FAST is not a new networking protocol but a user-friendly software that takes advantage of the tools that are already present on the data transfer end-system. The application scenarios of the first version are the same in which "wget" is currently employed. Specifically, "wget" is capable of executing a "PULL" request from the destination. While some protocols/tools (such as GridFTP/Globus Toolkit) are quite capable of third party transfers, based on our experiences, implementing such abilities alongside other limited tools (such as wget or scp), especially when one desires to estimate the data transfer performance, is very difficult. Thus, server tool of FAST is to be installed on the destination and used in a "PULL" fashion to transfer data from the source to the host by executing our client tool of FAST.

FAST ensures that no matter what, it can fall back to a more traditional or a currently widely used tool such "wget or"scp"which provides scientific users a performance guarantee on some level.

Since some of the services that we want to access require user credentials (certainly they are not all anonymous). While direct username and password credentials exist, they are rare in scientific data stores. There are a couple major systems for credentials:

- OpenID: OpenID is used by some sites (such as Earth System Grid) to gain access to user credentials that can then be used to authenticate and authorize with the data source. In OpenID, one has a provider site (such as google.com even) that to be provided a username and password in order to gain access to credentials (in the form of a cookie or X.509 certificate) that can be used to then access the data from the source.

- MyProxy: Most users do not manage credentials directly but rather gain temporary access to them through a MyProxy-installed server. They will request a short-lived proxy X.509 certificate (typically 12 hours or less) of their real certificate stored with some MyProxy server. Again, this will involve a username and password to access MyProxy to generate the proxy certificate.

The implementation of this tool will employ a collection of languages. In particular, all of the credential is heavily based in Java. There are existing Java programs/libraries for interacting with credential mechanisms. OSCARS also has a simple library for interacting with OSCARS in a Java. FDT is in Java. Globus Toolkit covers a collection of languages including C, Python, Java (jGlobus).

To conclude, the goal of this project is to build a tool that is as simple to use as "wget" but much, much faster. Right now, services like ESG generate simple "wget" scripts for users to download files. We want to build a tool capable of being used in the similar manner, except credential and transfer, we also want to enable circuit reservation and performance estimation:

- Determine available tools, services as a set of modules;

- Acquire credentials for source;

- Estimate Performance;

- Choose modules that provide best performance: typically a reservation component (optional) and transfer component (mandatory);

- Execute modules in order (first reservation and then transfer) to accomplish transfer, providing credentials as necessary.

Based on the previous description, we design the architecture of FAST as shown in Fig. 10. The modules that FAST includes or involves are:

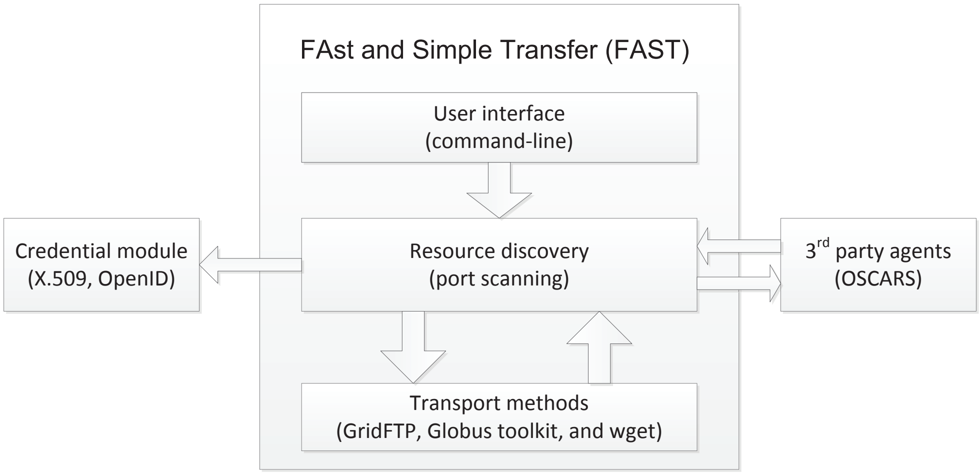


Figure 10: Architecture design of FAST.

- Credential module, for authorization and authentication:
- X.509 certificate;
- OpenID (for ESG users).
- User interface module, just a simple command-line.
- Resource discovery module, exploring available resources both at the end-host and within the network:
- Resource discovery: port scanner (at a minimum, trivial);
- Path establishment: OSCARS (within ESnet).
- Transport module, encapsulations of different transport protocols/methods:
- GridFTP;
- Globus toolkit;
- Wget (considering fallback and by default).
- Performance estimation module (optional): this should be optional that can be decided by the user if it is necessary;
- Log module (optional): keeping log information, historical data and performance evaluation information, and etc, which might be useful for error diagnosis.

5 Conclusion

We proposed a workflow-based transport solution to support bulk data transfer in large-scale eScience applications. By leveraging the resource discovery capability of WINDMA, we constructed cost models for discovered resources and formulated path composition as an optimization problem. Actual data transfer is performed by running underlying transport methods or invoking existing data transfer services. Experimental results show that the proposed transport solution is able to compose an optimal end-to-end network path and carry out efficient data transfer for a given user request. We are currently implementing a light-weight data transfer tool which emphasizes more on the actual fast and simple massive data transfer. The target users will be common domain users who do not have to learn complicated computer and network technology in order to use this tools. Our plan is to introduce this FAST tool to a few DOE scientific groups.

Acknowledgments

This research is sponsored by U.S. Department of Energy's Office of Science under Grant No. DE-SC0002078 with Southern Illinois University at Carbondale and Grant No. DE-SC0002400 with University of Memphis.

References

[1] OSCARS: On-demand Secure Circuits and Advance Reservation System. http://www.es.net/oscars.

[2] C. Guok, D. Robertson, M. Thompson, J. Lee, B. Tierney, and W. Johnston, "Intra and interdomain circuit provisioning using the OSCARS reservation system," in *Proc. of the BROADNETS*, San Jose, CA, Oct. 1-5 2006, pp. 1–8.

[3] Energy Sciences Network. http://www.es.net.

[4] Internet2 Interoperable On-Demand Network (ION) Service. http://www.internet2.edu/ion.

[5] P. Brown, M. Zhu, Q. Wu, and X. Lu, "Network-aware data movement advisor," in *Proc. of the 1st Int. Workshop on Network-aware Data Management, in conjunction with the Supercomputing Conference*, Seattle, CA, USA, Nov. 14 2011.

[6] Drupal. http://drupal.org.

[7] J. Zurawski, E. Boyd, T. Lehman, S. McKee, A. Mughal, H. Newman, P. Sheldon, S. Wolff, and X. Yang, "Scientific data movement enabled by the dynes instrument," in *Proc. of the 1st Int. Workshop on Network-aware Data Management, in conjunction with the Supercomputing Conference*, Seattle, CA, USA, Nov. 14 2011, pp. 41–48.

[8] SQLite. http://www.sqlite.org.

[9] ANI: Advance Network Initiative. http://www.es.net/RandD/advanced-networking-initiative.

[10] UCLP: User Controlled LightPath Provisioning. http://www.uclp.ca.

[11] DRAGON: Dynamic Resource Allocation via GMPLS Optical Networks. http://dragon.maxgigapop.net.

[12] JGN II: Advanced Network Testbed for Research and Development. http://www.jgn.nict.go.jp.

[13] Geant2. http://www.geant2.net.

[14] ESCPS: End Site Control Plane Service. https://plone3.fnal.gov/P0/ESCPS/.

[15] GridFTP. http://www.globus.org/grid_software/data/gridftp.php.

[16] GENI: Global Environment for Network Innivations. http://www.geni.net.

[17] BeStMan Berkeley Storage Manager. https://sdm.lbl.gov/bestman/.

[18] A. Patil, B. Belter, A. Polyrakis, T. Rodwell, M. Przybylski, and M. Grammatikou, "The GEANT2 advance multi-domain provisioning system," in *Proc. of TERENA Net. Conf.*, Catania, Italy, May 15-18 2006.

[19] N. Rao, W. Wing, , S. Carter, and Q. Wu, "Ultrascience net: Network testbed for large-scale science applications," *IEEE Communications Magazine*, vol. 43, no. 11, pp. s12–s17, 2005, an expanded version available at www.csm.ornl.gov/ultranet.

[20] W. Allcock, J. Bresnahan, R. Kettimuthu, M. Link, C. Dumitrescu, I. Raicu, and I. Foster, "The globus striped GridFTP framework and server," in *Proc. of Supercomputing*, 2005.

[21] Earth System Grid (ESG). http://www.earthsystemgrid.org.

[22] Globus Online. https://www.globusonline.org.

[23] FDT: Fast Data Transfer. http://monalisa.cern.ch/FDT.