Table of Contents

Cover Page 1
Table of Contents 2
Table of Contents 2
Project Summary 1
1 Background and Significance 2
1.1 The Opportunity 2
1.1.1 Sample Use Cases 3
1.2 Our Proposal: Overview 3
1.3 The State of the Art: Other Systems 4
1.3.1 Existing Software 4
1.3.2 Emerging Technologies 5
1.3.3 Experience and Competence 6
1.3.4 Other Related Projects 8
2 Research Design and Method 9
2.1 Standard Schemas and APIs 9
2.2 Troubleshooting and Analysis 10
2.3 Information and Capability Discovery 10
2.4 Adaptive Monitoring 11
2.5 Ultra High-Speed Networking 11
2.6 Test Scheduling 11
2.7 Authorization and Policy Issues 11
2.8 Active Versus Passive Monitoring 12
3 Implementation and Deployment Plan 12
3.1 ESnet Deployment 13
3.2 UltraScienceNet Deployment 13
3.3 Security Deployment 13
3.4 IPv6 Deployment 13
3.5 Sample Grid Deployment 14
4 Deliverables 15
4.1 Publication of ESnet Monitoring Data 15
4.2 Develop Standard Schemas and APIs 15
4.3 System Integration 15
4.4 Deployment 16
4.5 Improved Troubleshooting and Tools 16
4.6 Improved Best Practices 16
5 Milestones 17
6 Connections 18
7 Project Management 19
8 Conclusions 19
9 Literature Cited 21

10 Budget and Justification **24**
10.1 LBNL Budget Explanation 55
10.2 SLAC Budget Explanation 56
10.3 Internet2 Budget Explanation 57
10.4 U. of Delaware Budget Explanation 57
10.5 Other Support of Investigators 59
11 Facilities **64**
11.1 LBNL Facilities 64
11.2 SLAC Facilities 64
11.3 Internet2 Facilities 65
11.4 University of Delaware Facilities 66
11.5 PSC Facilities 66
12 Biographies of Key Personnel **68**
13 Appendix **80**

Project Summary

The demands of data intensive science and the growth of high-capacity connectivity have led to the increased need for tools to measure, test, and report network performance. However, while a single administrative domain might deploy a single network measurement infrastructure, multiple administrative domains are unlikely to do so. The success of Grid computing will depend on the existence of an interoperable federation of network measurement infrastructures. Measurement and Analysis for the Global Grid and Internet End-to-End Performance (MAGGIE) is an initiative to allow sharing of network monitoring data. MAGGIE will accomplish this by defining and implementing standard schemas and protocols, and by developing a common framework for constructing federations of existing network monitoring infrastructures.

Current Grid applications typically use a much smaller percentage of available network bandwidth than is reported or expected. While application developers often see the network as the problem, network engineers typically point to host issues or poor application design. Network monitoring services are needed to resolve these questions and to verify whether the network is in fact the source of the problem.

Previous attempts to address these performance problems have lead to the development of a collection of basic tools that can be used to measure specific network variables (e.g., available bandwidth, one-way delay) and several early stage network measurement infrastructures that use these tools to monitor specific portions of the global Internet. These infrastructures typically cover a single administrative domain and/or a specific application community and are designed to provide highly skilled network engineers with the ability to manage their network.

While it is realistic to believe that an administrative domain will deploy a single network measurement infrastructure, it is unrealistic to believe that multiple administrative domains will all deploy the same measurement infrastructure. It is important that a multi-domain, multi-infrastructure measurement environment be explored, and that a mechanism be created to tie together a heterogeneous collection of monitoring systems. To date, little or no effort has been expended to allow existing measurement tools and infrastructures to work together and/or share data with peer infrastructures.

MAGGIE will develop the software tools and procedures that are needed to allow this data sharing to occur. It will also integrate information from multiple sources and infrastructures to enable improved understanding and diagnosing of network problems. We will contribute to, extend, and implement standards and protocols proposed by the Global Grid Forum's network monitoring and security working groups to ensure inter-operability between these existing measurement infrastructures.

MAGGIE combines the skills and talents of DOE Lab and University network researchers with ESnet and Internet2 engineers and researchers. The inclusion of production network operations staff ensures that we will have access to all publishable monitoring data on each of these networks, and will be able to deploy additional monitoring tools and services where needed. We also have close ties to the DOE UltraScienceNet project, and will work closely with them to deploy a monitoring system across the UltraScienceNet network. This will be critical to the success of the UltraScienceNet, as on an experimental network such as UltraScienceNet, easy access to a wide range of monitoring data will be critical to ensure the network is operating as expected, and to understand how applications behave on such a network.

As Grid computing communities use multiple network infrastructures to link all members of the community together, data from multiple measurement infrastructures must be combined to obtain a full picture of end-to-end performance issues. MAGGIE will operate over both production networks (ESnet and Abilene) with new experimental testbed networks (UltraScienceNet) allowing this full picture to be developed.

1 Background and Significance

Current Grid applications typically use a much smaller percentage of available network bandwidth than is reported or expected. While application developers often see the network as the problem, network engineers typically point to host issues or poor application design. Network monitoring services are needed to resolve these questions and to verify whether the network is in fact the source of the problem.

Finding and fixing Grid application configuration and performance problems is a challenging task. Problems can manifest themselves in many ways, making it difficult for experienced network engineers and application developers to understand what is going wrong. For example, an application requiring extensive network I/O could incur performance impacts due to slow disk access, large network delays, or poor programming techniques. Determining the problem and finding the solution require both a deep understanding of all possible problem areas and the ability to mine and correlate network, system, and application data sources. These capabilities are provided by Network Monitoring Services.

1.1 The Opportunity

In response to the demands of data intensive science and the growth of high capacity connectivity, the research community has developed a large selection of tools to measure and test network performance within local, metropolitan, and wide-area networks. These basic tools measure specific network variables (e.g., available bandwidth, one-way delay), and several early-stage network measurement infrastructures typically allow highly skilled network engineers to manage a single administrative domain or a specific application community.

However, while a single administrative domain might realistically deploy a single network measurement infrastructure, multiple administrative domains are unlikely to do so. U.S.-based DOE scientists will use multiple networks (e.g., ESnet, UltraScienceNet, and Abilene) to meet their day-to-day network needs, it is important that a multi-domain, multi-infrastructure measurement environment be explored, and that a mechanism be created to tie together a heterogeneous collection of monitoring systems. However, to date, little or no effort has been expended to allow existing infrastructures to work together and share data with peer infrastructures. This functionality will be critical to the Grid's success in enabling geographically separated scientists to work together effectively as a team.

CAIDA and Internet2 recently sponsored a workshop[1] that identified the following components as necessary to create a federation of network measurement infrastructures. MAGGIE will address each of these in working toward the goal of federated, cross-domain network measurement infrastructure.

- A standard protocol (schema) for requesting tests and collecting results of tests
- A mechanism to discover network monitoring resources (measurement nodes and databases)
- A mechanism for inter-administrative domain authentication and authorization
- A standard protocol for coordinating and running tests between measurement frameworks
- A set of requirements for performance tools (or wrappers there around) to be employed in a measurement framework

Measurement and Analysis for the Global Grid and Internet End-to-End Performance (MAGGIE) is an initiative to allow sharing of network monitoring data. MAGGIE will accomplish this by defining and implementing standard schemas and protocols, and by developing a common framework for constructing federations of existing network monitoring infrastructures.

Such an interoperable federation is essential for troubleshooting Grid applications, and for efficient utilization of interlinked networks. Without such an infrastructure, Grid applications developers will continue to focus on the network as the source of performance problems, and network engineers will continue to look to some other network segment for end-to-end performance problems.

Many network misconfigurations today result in low throughput, but go undetected because something else (such as TCP, the end-host, etc.) is the bottleneck. As TCP problems are fixed, and as the network speeds continue to

[1] Performance Measurement Architecture Workshop 2003. This material is based in part on work supported by the National Science Foundation under Grant No. ANI-0314723.

increase, misconfigured networks will be even more common. End-to-end network monitoring across multiple networks and multiple administrative domains is the only way to detect such problems.

A common framework for network monitoring will lead to a better understanding of what tools and techniques are useful to Grid applications developers as well as to network engineers, and will result in fewer redundant tests being run. In addition, MAGGIE will engender a better understanding of the authorization issues inherent in such a federated approach and how these map to the "virtual organizations" (VOs) being formed in the Grid community.

MAGGIE will also, in the process, foster collaboration between existing measurement infrastructure projects, enabling researchers to identify their most useful features, encouraging the sharing of ideas and code, and enabling each infrastructure to improve more quickly than it would if the individual teams were working alone.

MAGGIE combines the skills and talents of DOE Lab and University network researchers with ESnet and Internet2 engineers and researchers. The inclusion of production network operations staff ensures that we will have access to all publishable monitoring data on each of these networks, and we will be able to deploy additional monitoring tools and services where needed.

In addition, we have close ties to the DOE UltraScienceNet, and will work closely with them to deploy a monitoring system across the UltraScienceNet as well. This will be critical to the success of the UltraScienceNet, as on an experimental network such as UltraScienceNet, easy access to a wide range of monitoring data will be critical to ensure the network is operating as expected, and to understand how applications behave on such a network.

1.1.1 Sample Use Cases

The federation of network monitoring infrastructures proposed for MAGGIE will satisfy the following use cases:

***Use Case 1*: Network Troubleshooting.** Often, achievable throughput across a given path is much slower than the available bandwidth that is reported or expected. Network monitoring is needed to discover the reason for the problem, which network segment has the problem (host, interface, link, router, switch, autonomous domain, etc.) and whom to contact to correct it. Many end-to-end network paths include unmonitored network segments, especially close to the end user's desktop computer. To cover these unmonitored segments, the infrastructure will incorporate monitoring tools that use host-based performance data and bidirectional tests to the end host.

***Use Case 2*: Capacity Planning and Auditing.** Network monitoring data is necessary for capacity planning purposes, so that resources are available when needed to prevent bottlenecks, and so that measurable service level agreements are put in place and their requirements are met.

***Use Case 3*: Grid Scheduling.** Network monitoring data is required by Grid data management middleware when selecting the best source(s) from which to copy replicated data. Either raw historical data or predictions of future end-to-end path characteristics between the destination and each possible source might be used for this purpose. Accurate predictions of the performance obtainable from each source requires a history of measurements of available bandwidth (both end-to-end and hop-by-hop), latency, loss, and other characteristics important to file transfer performance.

***Use Case 4*: Grid Troubleshooting.** Imagine that a Grid job which took 10 minutes to run yesterday is taking 60 minutes to run today, using the same set of hardware. Why? Is the problem the hosts, disks, or networks? Grid Application users and developers often blame the network, when in fact the problem may be elsewhere. Network monitoring data (as well as host and application monitoring data) is needed to determine the source of the problem.

***Use Case 5:* Evaluation of experimental networking components.** Network monitoring data is needed for evaluation of new protocols for very high-speed networks, new network interface card (NIC) features (i.e., TCP Offloading Engines (TOE)), large MTUs, and so on.

1.2 Our Proposal: Overview

MAGGIE will develop a federation of test and analysis frameworks that will allow application users and network engineers to determine how well they are using available network capacity and to diagnose problems that degrade network utilization. This federation will be designed for a broad range of network capacity settings, including production networks like ESnet and Abilene, and very high end networks like the DOE UltraScienceNet.

MAGGIE will port the tools and technologies used to measure and monitor today's gigabit networks to tomorrow's multi-gigabit networks. The project will also develop the software tools and procedures needed to share network performance and monitoring data across multiple independent measurement systems and administrative

domains. The monitoring data will be available both in simple formats for end users and detailed formats for network engineers.

We will contribute to, extend, and implement standards and protocols proposed by the Global Grid Forum (GGF) network monitoring and security working groups to ensure inter-operability between these existing measurement infrastructures. This approach recognizes that current and future global networks will be built by peering relationships between independent network administrative domains.

Finally, MAGGIE will encourage the close collaboration needed to identify the most useful features of each infrastructure. As a result, each infrastructure will improve faster than if the individual teams worked alone. The combination of interoperable measurement infrastructures scalable to multiple domains and better tools will make the network more useful for all DOE scientists.

MAGGIE's federation of network monitoring infrastructures will be:

- Interoperable. It will be possible to request performance measurement data and run regularly scheduled or on-demand active measurement tests so that performance measurement data can be obtained for every link in the end-to-end path. Performance measurement data will be available in a common format, enabling performance analysis tools to work with the entire federation of network measurement infrastructures.
- Useful for applications. It must provide information that includes available bandwidth, achievable throughput, delay, jitter, and routing needed for applications such as data placement, Grid scheduling, and Grid troubleshooting.
- Useful for network engineers. It must provide information needed to help planning and to identify, isolate, report and solve problems such as misconfigured or underpowered routers, switches, and hosts.
- Useful for DOE scientists and end users. It must provide a mechanism to allow end users to be able to predict what to expect, and to perform basic troubleshooting steps with the ability to forward the results to the proper network engineer so the problem can be resolved. It must also identify basic network configuration errors that are known to cause performance problems.
- Easy to deploy and configure. Adding new monitoring hosts to a monitoring infrastructure must not be too difficult, otherwise there will not be enough participating hosts to be useful.
- As unobtrusive as possible. Passive monitoring should be done whenever possible; active monitoring only when necessary. Active monitoring data should be stored when feasible, and stored active monitoring data should be fetched when available.
- As accurate and timely as possible. Care must be taken to ensure that the best data source is used. The infrastructure should automatically determine if it is better to retrieve archived data or gather new data to answer a performance question. Care must also be taken to ensure that competing measurements do not undercut each other's accuracy.

1.3 The State of the Art: Other Systems

MAGGIE will evaluate and, when appropriate, take advantage of existing software and emerging technologies, and build on previous work by project participants as well as other related projects, as described in this section.

1.3.1 Existing Software

There are a number of existing tools and services for network monitoring. See [Infra] for a comparison of those most relevant to the current proposal. Most, such as AMP [AMP], NIMI [NIMI], Surveyor [Surveyor], E2E piPEs [PIPES], and RON [RON] have focused only on use cases 1and 2 (see section 1.1.1). This proposal will address all five use cases. MAGGIE will evaluate and/or use a combination of the following tools and systems to build the federation of network monitoring infrastructures and demonstrate its extensibility:

- **NIMI**. Provides a secure management infrastructure, allowing authenticated and authorized tool execution and evaluation [NIMI].
- **IEPM-BW/PingER**. Provides initial data analysis components, a platform for testing high-performance transports and applications, and monitoring system fault management mechanisms [IEPM-BW].
- **NTAF.** Includes Web Services-based data publication mechanisms that use the GGF-recommended format [NTAF] [NMWG].

- **NetLogger.** Provides tools to transport monitoring results and load them into a relational database [NETLOG].
- **NDT**. Provides Web100-based network diagnostic functions and detailed analysis reports for desktop/laptop computers [NDT].
- **E2E piPEs**. Provides a framework for scheduled and on-demand testing services between network measurement points; already deployed throughout the Abilene network.
- **ESnet Performance Center**. Provides a means of performing tests between ESnet core routers and end host systems [EPC].
- **Scriptroute**. Provides a simple-to-use ability to define and launch lightweight on-demand network monitoring tools without the need for passwords or credentials [Scriptroute].

Monitoring Tools. Many tools (e.g., ping, traceroute [Traceroute], iperf [Bwctl] [Iperf], owamp [Owamp], pathrate [Pathrate], Pathload [Pathload], ABwE [ABwE], netest [Netest], and NWS [nws]) already exist that perform specific network monitoring tasks. In many cases, these tools will need extensive modification (or replacement) to work effectively at the new higher speeds.

1.3.2 Emerging Technologies

In addition to the software technologies described above, there are some emerging technologies that we plan to take advantage of. These include peer-to-peer overlay networks for information discovery, and several of the Grid Services standards now being defined by the GGF. A major benefit from the MAGGIE proposal will be to foster collaboration between these measurement infrastructure projects, enabling researchers to identify their most useful features and encouraging the sharing of ideas and code. In this manner, each infrastructure will be improved faster than if the individual teams worked alone. It should be noted that the participants in this proposal are actively involved in the development and deployment of most of the above-mentioned existing tools, services, standards, and protocols. As such, they are uniquely qualified to integrate these existing projects into the proposed federation of network monitoring infrastructures.

1.3.2.1 Peer-to-Peer Technology

We will explore the use of Peer-to-Peer (P2P) technology for network monitoring. One of the important aspects of P2P computing is that the system becomes more powerful and more useful as new peers join [Gribble]. Because P2P systems are decentralized, the robustness, availability, and performance of the systems grow with the number of peers. The diversity of the system also scales, as new peers can introduce specialized data (e.g., information on new network paths) that the system was previously lacking. Decentralization also helps eliminate control issues, as trust is diffused over all participants in the system. The need for administration is greatly diminished, since there is no dedicated infrastructure to manage. In particular, we will look into using P2P technology for resource and information discovery. This is discussed in more detail in section 2.3.

1.3.2.2 Web / Grid Services

The GGF is working on a set of specifications to extend Web Services standards such as SOAP and WSDL in a way that is useful for building "Grid Services." By writing Grid Service wrappers for existing network monitoring infrastructures, we can provide a common request-response interface to a variety of network monitoring infrastructures.

Several members of the MAGGIE team are active members of the GGF's Network Measurement Working Group (NMWG). The NMWG has produced a GGF draft recommendations document that defines a classification hierarchy for network measurements, and defines a standard naming convention for network monitoring data [NMWG-name]. The NMWG is currently defining a request-response schema for network monitoring using XML schema [XML]. The schema will also allow Grid services to handle requests for on-demand tests from authorized users. E2E piPEs, NTAF, and NLANR's Advisor are using early drafts of the NMWG schemas natively, and other projects have voiced an interest in speaking the same test/data request/response "language". MAGGIE will work to deploy, test, and continue to refine the NMWG schemas. This will provide a common protocol for both running tests and collecting results from many existing monitoring systems.

There are also a number of emerging standards related to authorization issues. These include federated identity management via the Security Assertion Markup Language (SAML) [SAML], which is a framework for exchanging authentication and authorization messages. Additionally, WS-Agreement and the newly defined Web Services

Resource Framework (WS-RF) provide powerful mechanisms for service negotiation given local authorization policies. MAGGIE will explore the use of these standards to provide a federation-wide approach for discovery of network monitoring resources in other administrative domains and a federation-wide approach to AAA to allow for trusted inter-measurement domain requests for tests and data.

1.3.3 Experience and Competence

In this section we describe some previous work that we will build upon in this proposal. While our primary task is to federate these separate infrastructures, we also expect to use this close collaboration to strengthen each infrastructure. We will analyze the strengths and weaknesses of each of these technologies, and determine how each may be improved by incorporating ideas and functions from peer infrastructures.

1.3.3.1 NIMI

National Internet Measurement Infrastructure (NIMI), developed at PSC, is a software system for building network measurement infrastructures. A NIMI infrastructure consists of a set of dedicated measurement servers (termed NIMI probes) running on a number of hosts in a network, and measurement configuration and control software, which runs on separate hosts. A key NIMI design goal is scalability to potentially thousands of NIMI probes within a single infrastructure; as the number of probes increases, the number of available measurable paths increases via the N-squared effect, potentially allowing for a global view of the network.

A fundamental aspect of the NIMI architecture is that each NIMI probe reports to a configuration point of contact (CPOC) designated by the owner of the probe system. There is no requirement that different probes report to the same CPOC, and, indeed, there generally will be one CPOC per administrative domain participating in the infrastructure. However, the NIMI architecture also allows for easy delegation of part of a probe's measurement services, offering, when necessary, tight control over exactly what services are delegated.

The architecture was designed with security as a central concern: All access is via X.509 certificates. Each NIMI probe is configured by its CPOC to allow differing levels of access to particular sets of resources. The owner of the probe can thus determine which certificate has what type of access by controlling via policy what resources are accessible to particular certificates.

The sole function of a NIMI probe is to queue requests for measurement at some point in the future, execute the measurement when its scheduled time arrives, store the results for retrieval by remote measurement clients, or ship the results to a remote Data Analysis Client (DAC), and delete the results when told to do so. An important point for gaining measurement flexibility is that NIMI does not presume a particular set of measurement tools. Instead, the NIMI probes have the notion of a measurement "module," which can reflect a number of different measurement tools. Currently, these measurements include httperf, ping, mtrace, traceroute, TReno, and zing (a generalized "ping" measurement). Note that including other active measurement tools on selected probes is simple, as the tool, along with its wrapper script to provide a minimal common interface, is automatically propagated to the NIMI probes by their respective CPOCs.

From their work on NIMI, PSC brings expertise in secure, reliable network measurement systems to the MAGGIE team.

1.3.3.2 NTAF, NetLogger, and Scishare

As part of the Net100 project, LBNL has developed a framework for running network test tools and storing the results in a relational database, which we call the Network Tool Analysis Framework (NTAF). NTAF manages and runs a set of prescheduled network testing tools and sends the results to a database for later retrieval. Recent results are cached and can be queried via a client API. The goal of the NTAF is to make it easy to collect, query, and compare results from any set of network or host monitoring tools running at multiple sites. The basic function performed by NTAF is to run tools at regular intervals, plus or minus a randomization factor, and send their results to a central archive system for later analysis. For example, *iperf* can be configured to run for 20 seconds every 2 hours, plus or minus 10 minutes, to a list of hosts.

NTAF also contains an early prototype of the GGF NMWG data publication schema, allowing any client that understands this schema to query for test results. The results for all NTAF tests are converted into NetLogger events. NetLogger provides an efficient and reliable data transport mechanism to send the results to a relational database event archive. For example, if the network connection to the archive goes down, NetLogger will transparently buffer monitoring events on local disk, and keep trying to connect to the archive. When the archive becomes available, the events buffered on disk will be sent automatically. More details are available in [NETLOG]. Each NTAF-generated

NetLogger event contains the following information: timestamp, program name, event name, source host, destination host, and value. Using a standard event format with common attributes for all monitoring events allows us to quickly and easily build SQL tables of the results. More details are in [NTAF]. We will use the data publication component of NTAF. MAGGIE will use NetLogger to reliably transfer monitoring data to one or more instances of the NetLogger Archive.

LBNL has also developed a distributed system for secure information sharing called Scishare [SCISHARE]. Scishare allows one to store and manage information on local facilities while sharing it with remote participants. The system design follows the peer-to-peer model. Each participant designates a set of items they wish to share within the system. Peers are able to search for items by sending a query to the network. The network delivers this query to the other peers, which then run the query against the data they have designated to share. Scishare includes a messaging framework called P2PIO [P2PIO]. Any client can use P2PIO to query the P2P network, and to retrieve the corresponding result set in an iterative manner. MAGGIE will explore the use of P2PIO to address the information discovery problem.

From their work on NetLogger and NTAF, LBNL brings the following expertise to the MAGGIE team: schema design, distributed systems troubleshooting, and monitoring data archives and publish/subscribe interfaces, and P2P-based information discovery systems.

1.3.3.3 E2E piPEs

The Internet2 E2E piPEs system is a decentralized measurement framework designed to scale to multiple administrative measurement domains. As architected, it consists of modules responsible for analysis, discovery, network health monitoring, authorization and authentication, scheduling, testing, and storing. The piPEs framework permits the scheduling of on-demand and regularly scheduled active measurement tests, including bwctl/iperf (throughput, loss, jitter), owamp (one-way latency), and traceroute tests employing IPv4 and IPv6, and is extensible to other tools as well. Currently deployed on the Abilene backbone and a small number of campuses, work is just beginning on mesh-based discovery of network measurement resources, non-primitive authorization and authentication systems, and configurable policy management of permitted extra-administrative requesting of regular and on-demand performance tests. With MAGGIE, we will add an access and authentication system interoperable with (if not identical to) the NIMI measurement framework. We will also rework our web services to be compatible with the revised GGF NMWG schemas and employ the network resource service developed as part of the MAGGIE project.

From their work on bwctl, owamp, and the other E2E piPEs components, Internet2 brings the following expertise to the MAGGIE team: distributed system operations over national backbone networks, data archive services, distributed system service scheduling, and group-level security development.

1.3.3.4 NDT

The Network Diagnostic Tester (NDT) is a client/server-based tool that can identify common network configuration errors and performance problems with desktop computers. Its primary purpose is to allow users to self-test their desktop computer. The built-in test engine downloads a Java applet to the desktop computer, eliminating the need to preload software onto the desktop before testing can begin. The applet communicates with a server process to perform a series of tests that actively probe for specific configuration and performance problems. Once testing is completed, a built-in analysis engine combines measured values, Web100 Kernel Instrumentation Set (KIS) values, and calculated values to determine what, if anything is wrong with the desktop computer and the local network infrastructure. The analysis engine converts the network conditions into easy-to-understand diagnostic messages for the end user. A simple good/bad message is printed, with drill-down capabilities that allow the user to retrieve as much or as little data as required. Finally, the user is allowed to email a complete set of test results back to an administrator allowing that administrator to understand what happened and what problems the user is facing.

From their work on the NDT system, Internet2 brings the following expertise to the MAGGIE team: diagnostic methods for identifying network configuration problems, analysis methods with responses tailored for different user communities, and intelligent monitoring of desktop computers.

1.3.3.5 IEPM-BW/PingER

The IEPM-PingER project provides low impact (by default < 100bits/s on average for each monitor-remote host pair monitored), widely deployed (over 500 remote hosts in over 100 countries are monitored from over 30 monitoring hosts in 13 countries), regular, active, end-to-end ping based measurements. Since the ubiquitous Internet

ping facility is used, no accounts or credentials are needed on the remote hosts. The measurement data is archived, analyzed, reported in graphical and tabular format, and made publicly available via the web. Information is available going back nine years. With the large number of monitor-remote hosts pairs (over 3500), PingER has developed techniques for aggregating the results by affinity groups such as HENP experiments, Grid VOs, network communities, communities interested in developing countries and the Digital Divide, world regions, top-level domains, etc. PingER links are hierarchical rather than full mesh, in order to more closely match the needs of the communities served by the monitoring sites. PingER is used to provide Round Trip Time (RTT), losses, derived throughputs, jitter, etc., and the data is downloadable for further analysis.

The IEPM-BW project is complementary to IEPM-PingER in that IEPM-BW provides more network-intrusive, detailed regular end-to-end active measurements to provide a better understanding of high-performance paths, for a few tens of well-connected HENP, Grid, and Network sites. Currently, there are over 40 remote (server) hosts monitored from about ten monitoring (client) hosts. As a consequence of focusing on high-performance measurements, over 50% of the remote hosts are connected with Gbps links and the remainder are connected at 100Mbps. IEPM-BW has also been deployed in specialized ultra-high speed testbeds such as SuperComputing and iGrid. The infrastructure is robust and software only, yet deliberately simple, to enable quick deployment, with an emphasis on measuring with multiple network and application tools such as ping, iperf, traceroute, bbftp, GridFTP, ABwE, and traceroute. In addition, IEPM-BW provides analysis, presentation, archiving and limited prediction. Authentication between the clients and servers is based on ssh keys. Besides providing regular measurements with chosen tools, the infrastructure is also used to evaluate new tools such as Pathload, bbcp, Qiperf, and new TCP stacks such as FAST [Fast] and HS-TCP [HS-TCP]. The IEPM-BW results have also been incorporated into a tool for correlating Internet performance changes and route changes to assist in trouble shooting from an end-user perspective.

From their work on IEPM-BW and PingER, SLAC brings the following expertise to the MAGGIE team: network troubleshooting and analysis, and experience in designing and running network monitoring infrastructures.

1.3.3.6 ESnet Performance Center

ESnet Performance Centers (EPCs) are high-speed Unix-based hosts located at ESnet hub sites. These machines are connected to the corresponding ESnet core router via Gigabit Ethernet. Designated ESnet users will be able to access the EPC machines by way of a web interface. These users can run network tests from any Performance Center to any other Performance Center, or to the host machine at an ESnet site where they are running their web browser.

ESnet Site network personnel can run fairly high bandwidth tests to both the closest EPC and to the EPC closest to the final destination of their data and derive useful network information from their testing. The EPCs are an ESnet-wide shared resource, and are used only for debugging. They are not yet incorporated into any sort of automated testing systems. There are software “locks” that prevent multiple users from accessing the same Performance Center simultaneously, so sharing the resource is important.

From their work on the EPCs, ESnet brings expertise in network troubleshooting and in running a production network to the MAGGIE team.

1.3.4 Other Related Projects

A large number of network monitoring systems have been developed, all of which have a slightly different focus. MAGGIE will not design and implement yet another monitoring infrastructure, but will focus on interoperability issues between existing systems. This proposal envisions the creation of a federation of measurement frameworks, beginning with the several measurement frameworks under development by the participants in this proposal. It is hoped that this set of interoperable measurement frameworks will serve as a "kernel" around which a global federation of monitoring infrastructures will inevitably grow.

1.3.4.1 Advisor

The NLANR Network Performance Advisor [Advisor] is a single application that integrates the measuring, analyzing, and displaying of network performance statistics. The Advisor enables the writing of the analysis and display portions by providing a platform to allow easy integration of any number of network diagnostic tools, combined with the ability to uniformly query the results of these tools. It will ship with a network performance analysis tool and a network debugging utility aimed for network engineers, and knowledge about a number of diagnostic tools, including ping, ifconfig, iperf, AMP, Surveyor, and the Web100 suite of tools. Due to the Advisor’s

design, new analysis and display tools will be easy to write, and new network diagnostic tools will be straightforward to integrate. The Advisor distinguishes itself by the ability to display and analyze an extremely broad set of network statistics, due to its ability to integrate any network diagnostic tool. As a proof of concept and to gain experience with Advisor, the NLANR and Internet2 teams have successfully demonstrated the integration of Advisor PMCs and E2E piPEs measurement nodes and the importation of both data sets into the Advisor GUI using an early version of the NMWG schema.

1.3.4.2 MonALISA

The MonALISA [MonALISA] framework provides a distributed monitoring service system using JINI/JAVA and WSDL/SOAP technologies. Each MonALISA server acts as a dynamic service system and provides the functionality to be used by other services or clients that require such information. The goal is to provide the monitoring information from large and distributed systems to a set of loosely coupled "higher level services" in a flexible, self-describing way. This is part of a loosely coupled service architectural model to perform effective resource utilization in large, heterogeneous distributed centers. The framework can integrate existing monitoring tools and procedures to collect parameters describing computational nodes, applications, and network performance. As a proof of concept and to gain experience with MonALISA, we are in the process of integrating some of the IEPM and E2E piPEs results into MonALISA.

1.3.4.3 AMP

The Active Measurement Project (AMP) currently has about 130 monitors collecting data. Most of these monitors are located at NSF-funded HPC sites. All monitors are connected together, forming a full mesh, and they gather three measurements—round trip time, loss rate, and topology info— on a regularly scheduled basis. Each monitor may also run on-demand throughput tests. Efforts are already underway to enable the AMP monitors to respond to data requests using the GGF NMWG schema.

1.3.4.4 Scriptroute

The Scriptroute system uses a script programming language to facilitate the implementation of measurement tools and the coordination of measurements across servers. For example, traceroute can be expressed in Scriptroute in tens of lines of code and tasks can be combined across servers in hundreds of lines. For security, sandboxing and local control over resources are used to protect the measurement host, while rate-limiting and filters that block known attacks are used to protect the network. Further, because network measurements often send probe traffic to random Internet hosts and administrators sometimes mistake measurement traffic for an attack, a mechanism is provided to allow sites to block unwanted measurement traffic.

1.3.4.5 Monitoring Tools

Many existing tools (e.g., ping, traceroute, iperf, owamp, pathrate, pathload, AbwE, and netest) already perform specific data-collection tasks. Some of these tools are well suited for deployment in meshes of regularly scheduled tests sharing scarce resources (e.g., owamp), while others (e.g., iperf) require the development of resource allocation and scheduling daemons for test arbitration (e.g., the role bwctl plays for iperf). As needed, MAGGIE will develop wrappers similar to bwctl for other useful tools.

2 Research Design and Method

MAGGIE will bring together several current and previous measurement framework projects, enable them to share best practices, and create a network measurement federation through which the measurement frameworks become interoperable. This federation is expected to induce other measurement framework projects to join the federation by creating a set of interfaces and components that ease the burdens of achieving interoperability and by reaching a critical mass of measurement points and coverage of networks. The reduced burden of interoperability will then be well worth the effort of achieving it. In particular, it will allow us to provide a common monitoring interface for ESnet, Internet2, and UltraScienceNet network monitoring data.

2.1 Standard Schemas and APIs

Exchanging information between multiple network monitoring infrastructures requires standards for describing network monitoring tools and results. Emerging Web services technologies such as WSDL, SOAP, XML Schema,

and UDDI provide the tools to make interoperability feasible, but much work remains to be done. This includes defining common data descriptions for both active and passive monitoring data, common data attribute dictionaries, and common query formats. Standard mechanisms are needed to locate appropriate monitoring data providers, as well as standard schemas, publication mechanisms, and access policies for monitoring event data.

The GGF Network Monitoring Working Group has completed a document that categorizes network measurement data, and is working on a document to define a standard schema to describe network monitoring tools and results.

MAGGIE participants will continue to contribute to and lead GGF efforts to define standard APIs and protocols for network monitoring, and work with other GGF groups to define standards for authorization.

2.2 Troubleshooting and Analysis

Using the potentially overwhelming amount of monitoring data to track down problems is itself a problem. How does one close the "wizard gap," enabling ordinary users to achieve results comparable to experienced network engineers? How do we automate the discovery of common configuration and performance problems so that experienced network engineers can concentrate on unusual or new problems? MAGGIE researchers will address each of these problem areas, perhaps the hardest research problems included in the scope of this proposal.

New tools are needed so that users and engineers can quickly spot anomalous behavior or conditions that affect E2E performance. These tools must provide predictive and alerting functionality, with easy drill-down capabilities to correlate various measurements such as *traceroute,* Reverse Path Tree (RPT) [Scriptroute], and available bandwidth. Wider access to and integration of tools such as Network Diagnostic Tester (NDT) are also needed to enable easier and quicker detection of misconfigured hosts or network connections and performance problems. Tools such as those used in IEPM are needed to quickly detect, analyze, and report on failures (e.g., unreachable hosts, failed or hung processes or tools, etc.) within the monitoring infrastructures themselves.

MAGGIE will extend existing analysis and visualization tools and develop new ones, and apply them to multiple infrastructures.

2.3 Information and Capability Discovery

Several network measurement infrastructure issues relate to information discovery, these include:

- Finding archived data sites
- Finding the end-to-end path segments between the communicating hosts
- Determining which measurement tool to use and when to use it
- Finding monitoring hosts at or near the desired path

A hierarchical system such as the Globus MDS [MDS] or Scriptroute's *tinydns* [TinyDNS], or a peer-to-peer overlay can be used to find the data archive sites, measurement tools, or some passive monitoring sites. These methods rely on registering hosts, and the services they provide, in a database. Clients make calls to a well-known database server (e.g., DNS, LDAP) to find the correct service or server. The client can then contact the archive server to retrieve the necessary data, or it can contact the active monitoring host to request a test.

For active monitoring one would like to find monitoring hosts that are close to the end-points of interest. One possible solution is to use "discovery packets," which would work like the SCNM [SCNM] "activation packets." Discovery packets would be special UDP packets that would travel the end-to-end path. Routers along the way could send a copy of these packets out a specified port that contains the monitoring host, and the monitoring host could then reply with an "I'm here" message.

However, this will only work if the monitoring host is actually on the same path. To find monitoring hosts that are only "near" the path is more difficult. One possibility is to use a peer-to-peer overlay that includes the ability to keep track of "close" neighbors, such as the *Pastry* system [Row01], which uses a "proximity metric" to locate peers that are nearby. Another solution is to use a method such as Global Network Positioning (GNP) [Ng02], a proposed technique for estimating Internet latency between points. GNP estimates latency using multidimensional mappings derived from measurements between each point and special landmarks. Finally, the IP anycast service provides an expanding search capability that allows hosts to contact one of several servers that provide identical services. Thus, clients can find the closest server, where "closest" is based on network topology.

MAGGIE will explore these problems, leveraging existing Grid solutions and exploring new P2P-based solutions. We will develop grid-based mechanisms and procedures that allow independent measurement domains to share location information.

2.4 Adaptive Monitoring

A series of measurements can *react* to the results or data of previous measurement runs and, depending on how configured, can initiate a different measurement tool designed to probe a different metric within the network. For example, a ping study could suddenly find that one of its destination hosts is unreachable, and immediately initiate a traceroute to determine the last accessible hop in the path to the host. Upon noticing the available bandwidth drop significantly, related traceroutes could be evaluated for changes, and NDT could be run to see if there is a common network misconfiguration. Upon ping's failure, a TCP port-scanning tool could be used, if one suspected ICMP filtering, as an alternate form of determining connectivity. Moreover, this procedure can also be used to alleviate a suddenly congested network from more intrusive monitoring. For example, if congestion is observed, bwctl/iperf can be halted while a less intrusive diagnostic tool (Web100, for instance) is run to isolate the location of the congestion.

MAGGIE will develop diagnostic flow charts and software modules that can automate adaptive monitoring techniques. The flow charts will determine which additional test(s) should be performed when a primary test fails.

2.5 Ultra High-Speed Networking

Many of the existing monitoring tools have never been used in ultra high-speed environments (at >1 Gbps) such as UltraScienceNet. It is well known that many packet pair dispersion network tools that attempt to determine the available bandwidth or capacity in a path are likely to fail in high-speed environments. For example, there can be problems with packet timing if the Network Interface Cards (NICs) coalesce interrupts or use a TCP offloading Engine (ToE), both of which are common on today's 10GE NICs. It is also well known that host computer issues such as NIC drivers, CPU clock rate, and memory and I/O bus rates affect end-to-end performance. In addition, the framing used in OC192 (10 Gbps) links and beyond may cause timing problems. These factors must be considered when evaluating the results produced by measurement tools that use packet pair dispersion techniques. As ultra high-speed networks approach and exceed 10 Gbps rates, it becomes essential that these common tools be evaluated to determine their operational limitations.

MAGGIE will work with UltraScienceNet researchers to deploy monitoring systems on this network and analyze/evaluate/validate existing tools in this environment.

2.6 Test Scheduling

Many active network probes interfere with each other, and therefore must be scheduled to minimize measurement bias. We intend to employ token-based mutual exclusion techniques as part of the federation-level command and control infrastructure. Early work introduced a hierarchy of control that can be assimilated into a cohesive whole [Swany02]. Building on hierarchical token-passing schemes presented in the literature, we will develop a scalable approach for multi-domain measurement control.

Mechanisms are needed to federate various measurement infrastructures into a logically cohesive system. This will involve abstracting the test scheduling information from the various systems so that they may be "controlled" in a consistent fashion. Given that most current monitoring systems provide internal scheduling, part of this effort will involve what can be termed meta-scheduling. This is required to ensure that intrusive probes do not interfere with one another. While this is generally dealt with within a single measurement system, the notion of disparate, but coordinating, monitoring infrastructures complicates matters. Thus, a measurement federation should provide state synchronization as well as multi-domain measurement coordination (subject to resource policy and access control). Potential solutions include the Grid community's WS-Agreement [WS-AG] specification, which is designed to handle co-scheduling of Grid resources.

MAGGIE will provide a control and synchronization adaptation layer based on the emerging Grid Services standard.

2.7 Authorization and Policy Issues

Security issues dealing with tools usage and authorization policy issues are an important part of any measurement infrastructure. The administrator needs to ensure that steps are taken to prevent unauthorized use of the active

measurement devices to generate excessive amounts of traffic, effectively shutting down the network. Passive monitoring devices and monitoring data archives also need to be protected to ensure that they are not co-opted into revealing sensitive data.

Network Monitoring systems have a wide range of authorization policies. Systems that only do regularly scheduled testing such as PingER have a completely open policy, and require no authentication to access the results. Other systems, such as NIMI, which run active tests on demand, require a strong authentication and authorization policy mechanism.

Authorization techniques are an active research area, and a number of groups are working on this problem. These include the GGF's OGSA-Authorization working group, which is developing a standard for a WS-RF authorization service. This standard will specify a SAML query and response protocol for authorization requests and attribute assertions. It will also define a basic vocabulary for user attributes that can be used to permit access. Other work includes the web services community (e.g., federated identity management [FED]) and the IETF (e.g., AAA [RFC2903][RFC2903]).

MAGGIE researchers have experience in this area. NIMI uses (K)X.509 certificates to authenticate users via mutually accepted Certificate Authorities, and Akenti [AKENTI] to describe policy for those users on what measurements they may use, and what limitations those measurements may have. E2E piPEs is investigating the use of Internet2's Shibboleth project [Shibboleth] to provide a basis for identification and authentication of users and "entities", and a flexible mechanism to express authorization policy based on "attributes" of that entity. These attributes could include "membership" in a given group or site.

In MAGGIE, these mechanisms will be expanded to develop procedures and policies that enable these autonomous policy systems to effectively interoperate.

2.8 Active Versus Passive Monitoring

An active measurement tool is defined to be intrusive when its average probing traffic rate during the measurement process is significant compared to the available bandwidth in the path. It is desirable to perform passive monitoring, using tools such as NetFlow [NETFLOW], NetraMet [NetraMet], SNMP [SNMP], SCNM [SCNM], Magnet [Gardner], and Web100 [WEB100] whenever possible, as this is usually the least intrusive form of monitoring. However, passive monitoring data is often not available or has been sanitized for reasons of privacy, undermining its value. Meanwhile, some types of information, such as achievable throughput can only be obtained by active monitoring. In other cases, more resources are required to extract the relevant passive information rather than to simply make a direct active measurement, and it is therefore very important to monitor the bandwidth used by the measurement tools relative to the total available bandwidth.

Using Web100 instrumentation data, as is done by quick iperf [Qiperf], can make active tools less intrusive. Many available bandwidth measurement tools have parameters that trade lower accuracy for less intrusiveness, but finding the acceptable level of intrusiveness versus accuracy is difficult.

MAGGIE proposes to research mechanisms that will compare and integrate both active and passive monitoring results, within and between measurement domains, to provide the most complete and least intrusive monitoring system possible.

3 Implementation and Deployment Plan

A federation of interoperable network measurement infrastructures will be deployed over both of the major U.S. backbone networks—ESnet and Abilene—used by the DOE science community. In addition, new ultra high-speed testbeds—such as UltraScienceNet [UltraScienceNet], UltraLight [UltraLight], and UKLight [UKLight]—will be used to evaluate the performance of new and existing monitoring tools, especially at high speed. In return, this monitoring will be critical to understand the achievable performance of these testbeds, troubleshoot problems, and influence planning and use of future testbeds.

Efforts have already begun to develop and deploy individual measurement infrastructures (i.e., E2E piPEs, IEPM, NDT, NetLogger, NIMI, NTAF). These measurement infrastructures currently cover a large portion of the production networks (ESnet, Abilene), national peering points (Gigapops), and end sites (DOE Labs and universities). We will leverage our previous experience to ensure that major peering points (StarLight) and end sites (CERN, FNAL, SLAC, NERSC) that are important to the DOE science mission are covered by one of our measurement infrastructures and that these measurement infrastructures are interoperable.

Initial deployment will be based on the current 40 IEPM, 11 piPEs, 6 ESnet Performance Center, and 35 NIMI sites, currently being deployed or operated by the PIs. By using these existing infrastructures we can rapidly begin to tackle the major inter-domain interoperability problems identified in this proposal. We also expect to grow these infrastructures with donated and procured resources as more sites and institutions see the benefits of this collaborative approach. Apart from a small number of test systems for MAGGIE participants, MAGGIE does not directly require the purchase of any new hardware. We will use existing monitoring systems where they are available. In cases where hardware is not available or is outdated, we expect the site to provide the monitoring platform, and will work with the site to determine the best platform to purchase.

3.1 ESnet Deployment

We will deploy monitoring hosts at one or more ESnet core routers locations which will perform regularly scheduled end-to-end tests to a select set of key end-user sites, such as CERN, SLAC, NERSC, FNAL, ORNL, BNL, and ANL. We ensure that this monitoring covers paths that include the Internet2 and GEANT networks. This monitoring data will be very useful for analyzing end-to-end performance of critical data paths such as CERN to FNAL, and for Grid scheduling purposes across these networks.

We will also explore the use of the NIMI infrastructure on ESnet monitoring hosts to enable on-demand measurements, using NIMI's more sophisticated security and authorization policy mechanisms.

3.2 UltraScienceNet Deployment

DOE is developing UltraScienceNet, an ultra high-speed testbed network that will allow network researchers, as well as middleware and application developers, to experiment with optical network technologies. On a new, experimental network such as UltraScienceNet, easy access to a wide range of monitoring data will be critical to ensure the network is operating as expected, and to understand how applications behave on such a network.

The MAGGIE researchers have a strong working relationship with the ORNL UltraScienceNet testbed operators and work closely with networking staff at all of the Labs that will be connected to UltraScienceNet. We will deploy network monitoring components at various points in the testbed and will provide the ORNL UltraScienceNet team with monitoring displays that will assist them in network management and operation. This monitoring infrastructure will also be linked to the ESnet and Abilene production networks, allowing comprehensive monitoring of the end-to-end network path. Finally, we will evaluate the standard measurement tools to determine their strengths and weaknesses in the new high-speed environment.

3.3 Security Deployment

Several security issues must be addressed to ensure that measurement infrastructure is not co-opted for malicious purposes and cannot expose sensitive data to unauthorized users. The tools currently used by various team members have three basic approaches to prevent misuse. One approach is to limit the amount of data any specific tool can generate (e.g., PingER sends < 100 bits/sec per measurement pair, and NDT limits tests to 10 sec in each direction). A second approach is to wrap the raw tools (e.g., iperf) such that each tool is under direct control of a scheduling system (e.g., *bwctl*). The scheduler ensures that only one copy of the tool is operational at any point in time, and may stop and start a tool, preventing failed tools from continuing to pump data into the network. Finally, the access control policies will prevent the individual measurement servers from gaining unauthorized access, thus preventing the servers from being used to launch DOS or DDOS style attacks on the network.

This defense-in-depth approach ensures that the measurement infrastructure components do not expose the network operators or user sites to unacceptable risks.

3.4 IPv6 Deployment

All of the DOE and Internet2 backbone network routers support dual stack routing infrastructures. Thus, IPv4 and IPv6 protocols run natively over ESnet and Abilene, respectively. Several of the measurement tools (e.g., PingER) deployed within our existing measurement frameworks operate with both IPv4- and IPv6-based hosts. MAGGIE will expand network monitoring frameworks that do not yet report IPv6 statistics to report both IPv4 and IPv6 statistics. This monitoring will make IPv6 migration testing of DOE science applications and middleware easier for scientists and application developers.

3.5 Sample Grid Deployment

Figure 1 illustrates how a federation of multi-domain measurement infrastructures might be deployed in a typical Data Grid environment. Each backbone network has a set of monitoring hosts that form a backbone measurement domain that regularly runs a suite of scheduled tests, continuously monitoring network paths to a set of key sites. If a problem is detected, the appropriate network operator is automatically notified to resolve it. Each site has a set of monitoring hosts that forms a site measurement domain that regularly runs a suite of scheduled tests to ensure the local site infrastructure is operating properly. One of these monitoring hosts will be located at or near the site LAN/WAN gateway. Monitoring software will also be installed on the storage hosts and the compute hosts (e.g., the front end node of a cluster) to ensure that end-to-end monitoring of applications can be accomplished.

Consider a case in which a Grid job has data stored at site 1, but the user only has permission to run the job at site 2 or 3. A Grid Scheduler must determine whether to run the job at site 2 or 3 based on network monitoring information. Knowing the available bandwidth between network monitoring hosts A, E, and F is not enough. The scheduler needs to know the disk-to-disk achievable throughput between the storage service at site 1 and the storage services at sites 2 and 3.

Host A will begin the analysis by making a discovery query over ESnet, Abilene, site 2, and site 3 measurement domains to determine if the data exists that will answer this question. If not, host A may request an on-demand test to gather the data. Host A will also query its internal site 1 measurement database to determine the local storage system's performance characteristics. If host A is unable to determine the end-to-end performance characteristics of sites 2 and 3, it requests an on-demand test from the local storage system to the remote compute clusters. The results are passed back to the scheduler so the user's job can be executed.

Note that passive monitoring at the end systems may best solve this problem. If the Grid middleware that made transfers between the storage system and computer cluster published the results of previous transfers, then this information would make new on-demand testing unnecessary. Kernel-level monitoring information from Web100 could augment or replace this middleware-derived data. MAGGIE will work with Grid middleware developers to determine the best way to address these issues.

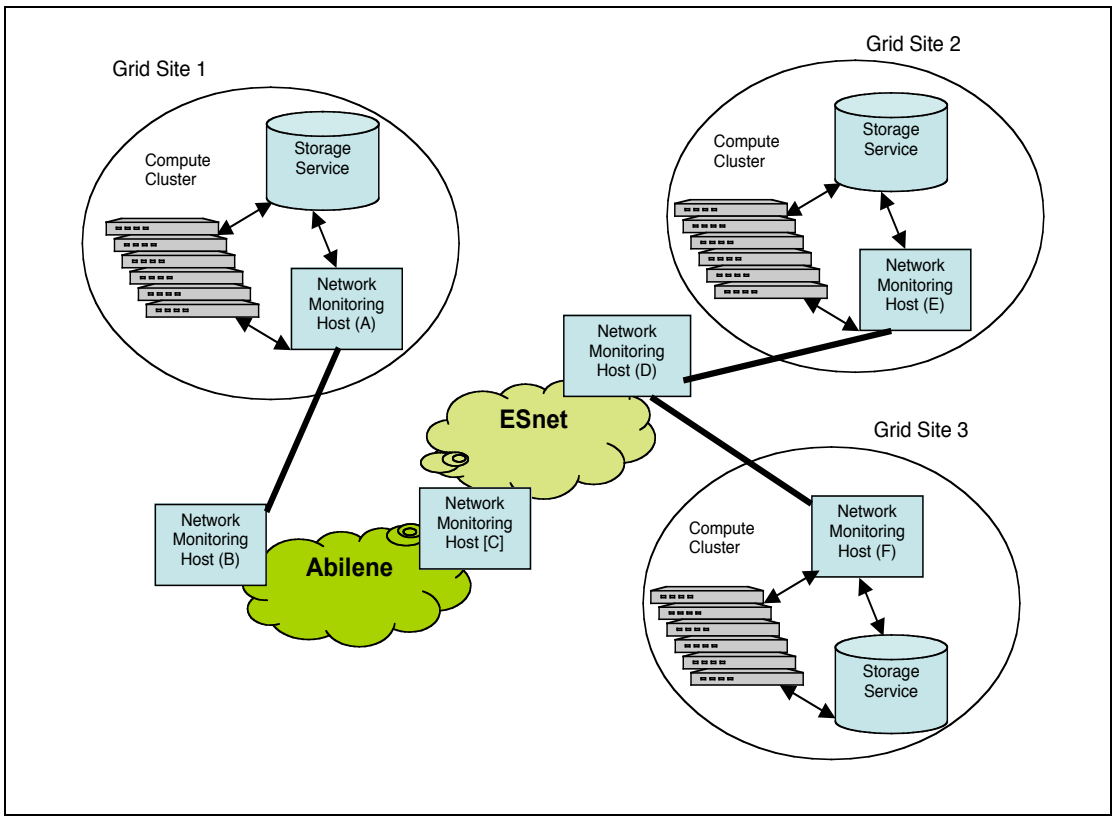

Figure 1: Sample Grid Deployment

4 Deliverables

MAGGIE will deliver an interoperable measurement federation across multiple administrative domains that include Abilene, ESnet, and UltraScienceNet, and probably others as well. We break the MAGGIE deliverables into the following broad areas, described in detail below. These areas are:

- Publication of select ESnet monitoring data
- Definition of standard schemas, protocols, and APIs
- Integration of these standard schemas, protocols, and APIs with existing monitoring systems
- Deployment of federated system across ESnet, Internet2, DOE UltraScienceNet, and on several production Grids
- Development of improved troubleshooting and data analysis tools
- Improved "Best Practices" enhancing existing tools and systems as needed

It is important to note that the MAGGIE proposal, if funded, does not exist in a funding vacuum. Some of the contributing institutions will also devote some of their own resources from existing funded projects to the achievement of these goals. Existing funding such as Internet2's contribution to the E2E piPEs project is devoted toward the extension and maintenance of these existing measurement infrastructures; not towards creating a federation of infrastructures. Our hope with MAGGIE is to create a common set of protocols that allows current and future projects to stand on each other's shoulders, and not each other's toes.

4.1 Publication of ESnet Monitoring Data

ESnet collects a large amount of monitoring data, such as backbone utilization data provided at http://www1.es.net/realtime-stats/. However, ESnet users have requested that even more data be made available, such as current and peak flow utilization values and that this information be made available via a standard Web Services API. ESnet is also currently working on a mechanism for publishing some per-flow statistics. We will determine what ESnet monitoring data can be made public (due to privacy concerns at some ESnet sites), and determine the most useful way to publish this data.

4.2 Develop Standard Schemas and APIs

Working with the GGF NMWG, MAGGIE will define the following:

- Standard schemas: These include schemas for requesting tests, collecting results, discovering services, database queries, and negotiating AAA requirements. (LBNL, Internet2, SLAC)
- Standard APIs: We will design and develop a Grid application and middleware client API to make it easy for applications such as a Grid resource broker or GridFTP to query the monitoring service for data. We will work with the Grid applications and middleware users and developers to gather requirements and design this API. (LBNL)

4.3 System Integration

A number of tasks will be required to incorporate these standard protocols and APIs into existing systems. Specific deliverables towards this goal include:

- Design and Implement a Grid Service interface for ESnet EPC, E2E piPEs, NIMI, and IEPM/PingER, that uses the GGF NMWG schema to launch network probes and retrieve the results. (All)
- Passive Monitoring Service: Design and develop a Web-services-based front end to publish passive data collected from SNMP, NetFlow, and Web100. (LBNL and SLAC)
- Archive Service: Integrate NetLogger with existing systems to collect data for the existing monitoring systems, and deploy the NetLogger SQL-based monitoring data archive service. Design an API to handle typical queries to the archive service. (LBNL)
- Discovery Service: Design, develop, and deploy a service to locate network monitoring nodes, data, and archives. (LBNL and PSC)

- Develop Grid Service interface components to MAGGIE federation-level services, including abstracting and proxying AAA from specific systems into the Grid services space and developing token-based control and mutual exclusion protocol for the meta-system. (U. Del)
  - Access, Authentication, and Configurable Policy Engine: Develop and deploy across the E2E piPEs project an access, authentication, and configurable policy engine. (Internet2 and PSC)

4.4 Deployment

Our deployment plan is described above in section 3. Some specific deployment deliverables include:

- Deployment of regularly scheduled tests to and/or from the Esnet core, and a mechanism to publish this data (ESnet, SLAC)
- Publication of various passively collected ESnet monitoring data (ESnet, LBNL)
- Assist UltraScienceNet network engineers with the deployment of passive and active monitoring data systems for UltraScienceNet (SLAC, ESnet)
- Deployment of monitoring "federation" software across ESnet, Internet2, and UltraScienceNet. (All)

4.5 Improved Troubleshooting and Tools

A federation of network monitoring infrastructures greatly increases the ability to perform troubleshooting and analysis across multiple domains. A number of tools and techniques exist or are in progress, but will have to be enhanced to take advantage of new sources of data. Additionally, a number of network monitoring tools will need to be improved to be useful in next generation networks such as UltraScienceNet. Some specific tasks include:

- Visualization: Design and develop tools to assist in visualizing performance problems, gathering extra relevant information from various infrastructures and preparing reports for relevant people. (SLAC)
- Misconfiguration detection: Evolve and extend the NDT analysis tool and expand upon the in-progress integration of the NDT tool into E2E piPEs and IEPM-BW to aid in the detection of common network configuration problems. (Internet2)
- Integration of results from passive and active tools: Compare and contrast measurements made by passive and active measurement tools to understand their agreement and appropriate use. (LBNL, SLAC)
- Automated Event Detection: Develop new automated event detection algorithms to detect significant step changes in end-to-end network performance (e.g., available bandwidth or RTT) and apply filters to the alerts to reduce noise from repeated alerts. (SLAC)
- First/Last Mile Analysis: Develop a custom packet generator tool to create packet trains uniquely suited for diagnosing first and last mile problems to the end host. (Internet2)
- Integration with new proposed tools: Evaluate and, where appropriate, integrate and assist in deployment of new tools such as those developed by the NSF–funded "Effective Diagnostic Strategies for Wide Area networks" and the DOE-proposed "Pythia: Automatic Performance Monitoring and Problem Diagnosis in Ultra High-Speed Networks." (All)

4.6 Improved Best Practices

MAGGIE will identify best practices among the existing measurement frameworks, report on lessons learned, and work to implement them across all deployed frameworks. Specific deliverables include:

- Adaptive Measurement: Design a mechanism to allow researchers to dynamically adjust a particular measurement based on the results of a previous run. (PSC)
- End Host Analysis: Develop a set of modules that integrate with the federation of measurement frameworks and allow diagnosis of first and last mile problems (which are expected to be the most common origin of performance problems for most users). (Internet2)
- Data Analysis: Compare active vs. passive monitoring data, compare available bandwidth with achievable throughput measurements, analyze intrusiveness of measurements. (All)

5 Milestones

The following is a breakdown of the above deliverables by topic and year-by-year.

Year 1:

- Define Standard Schemas, Protocols, and APIs
  - Finalize request/response schemas for monitoring data. (through the GGF NMWG)
  - Finalize archival database schemas. (LBNL, Internet2, U Del)
  - Define the security/trust model for inter-domain testing. (PSC, LBNL, Internet2)
  - Identify developers of Grid applications and middleware that can make effective use of network monitoring data, and work with them to define APIs to the monitoring data. (LBNL)
- System Integration
  - Develop proof of principal implementations of request/response schema for various monitoring systems, evaluate usability; identify interoperability issues. (SLAC, Internet2)
  - Design a discovery service to locate network monitoring tools and archives. (LBNL and PSC)
  - Design and develop Grid services interfaces to passive monitoring data from SNMP, NetFlow and Web100. (SLAC, LBNL)
  - Define semantics of federated measurement meta-scheduler. (U. Del)
  - Design and implement mechanisms to support adaptive measurement. (PSC)
- Deployment
  - Coordinate with ESnet site administrators (ESCC sub-committee). (SLAC, LBNL)
  - Install monitoring hosts in the ESnet core for regularly schedule testing (ESnet)
  - Coordinate with Internet2 user community (Joint Tech's sub-committee). (Internet2)
  - Identify deployment sites for monitoring hosts. (LBNL, Internet2, SLAC)
  - Coordinate with UltraScienceNet administrators (SLAC, ESnet)
  - Begin deploying monitoring tools on ESnet and Internet2 (ESnet and Internet2)
  - Deploy NetLogger monitoring data archive service (LBNL)
- Development of Improved Troubleshooting and Data Analysis Tools
  - Start collecting data for troubleshooting analysis and begin tool evaluation. (SLAC, Internet2)

Year 2:

- Define Standard Schemas, Protocols, and APIs
  - Refine request/response schemas for monitoring data. (through the GGF NMWG)
  - Refine the security/trust model for inter-domain testing. (PSC, LBNL, Internet2)
  - Finalize the client API for access to monitoring data. (LBNL)
- System Integration
  - Refine implementations of request/response schema for various monitoring systems, evaluate usability; identify interoperability issues. (SLAC, Internet2)
  - Implement a discovery service to locate network monitoring tools and archives. (LBNL, PSC)
  - Implementation of client API for requesting monitoring data (LBNL)
  - Refine the Grid services interface to passive monitoring. (SLAC, LBNL)
  - Work with Grid application developers to integrate access to network monitoring information into steering the Grid application. (LBNL)
  - Initial implementation of meta-scheduling overlay. (U Del)
  - Refine mechanisms to support adaptive measurement. (PSC)
- Deployment
  - Continue to identify deployment sites for monitoring hosts. (LBNL, Internet2, SLAC)
  - Continued coordination with ESnet site administrators. (SLAC, LBNL)
  - Deploy tools for access to passive network monitoring information. (SLAC, LBNL)
- Development of Improved Troubleshooting and Data Analysis Tools
  - Design and develop automated event detection for data extracted using Grid services from monitoring infrastructures. Evaluate effectiveness. (SLAC)
  - Design and develop command-line interface to tool for end-host analysis. (Internet2, SLAC)

- Compare, contrast, and validate passive tools versus active measurement network monitoring tools. (SLAC, LBNL)
- Evaluate new software and tools, work with tool developers; and select the best for our needs. (SLAC, LBNL, Internet2)

Year 3:

- Define Standard Schemas, Protocols, and APIs
  - Final standardization (through the GGF) of the request/response schemas for monitoring data. (through the GGF NMWG)
  - Finalize the client API for accessing monitoring data. (LBNL)
- System Integration
  - Continue to refine implementations of request/response schema for various monitoring systems, evaluate usability; identify interoperability issues. (SLAC, Internet2)
  - Refine the discovery service to locate network monitoring tools and archives. (LBNL and PSC)
  - Refine the implementation of meta-scheduling overlay. (U Del)
  - Continue to work with Grid application developers to integrate access to network monitoring information into steering the Grid application. (LBNL)
  - Refine and deploy mechanisms for adaptive measurement. (PSC)
- Deployment
  - Continue to identify deployment sites for monitoring hosts. (LBNL, Internet2, SLAC)
  - Continue deployment of new monitoring tools on ESnet, Abilene, and UltraScienceNe (SLAC, ESnet, and Internet2)
  - Document APIs, toolkits, identify follow on needs, work with operational folks provide technology transfer and to ensure sustainability of tool utilization, etc. (all)
  - Make presentations on progress and plans to ESCC, Internet2, DoE, GGF etc. Coordinate with others including the European efforts and PPDG. (All)
  - Encourage deployment of chosen tools in other promising infrastructures such as MonALISA and PlanetLab. (All)
- Development of Improved Troubleshooting and Data Analysis Tools
  - Refine automated event detection for data extracted using Grid services from monitoring infrastructures. Evaluate effectiveness. (SLAC)
  - Refine tools to assist in visualizing and pin-pointing performance problems. (SLAC, Internet2)
  - Develop tools to provide effective filtering of event alerts. (SLAC)
  - Continue to evaluate new and tools, and select the best for our needs. Identify improvements and new needs and communicate to developers. (Internet2, SLAC, LBNL)

6 Connections

We have very close connections with the HENP community, including CERN, SLAC, the EU DataTag [DataTag], and the EU DataGrid [EDG] (now called EGEE [EGEE]) projects, which will enable this project to make a large impact on the ability for physicists to get data to and from CERN, SLAC, FNAL, and BNL. We also have strong ties to other high-energy physics projects such as the Particle Physics Data Grid (PPDG), Grid2003 [Grid2003] and the SLAC-led BaBar [BaBar] project. We will work with these groups to determine the requirements and APIs needed by the applications community, and work with them to ensure the right types of data are being collected and from the right network paths.

We have close relationships to a number of other user communities as well, including Fusion, Earth Sciences, visualization, and several Grid portals groups. We are involved with several of the current DOE SciDAC projects. We meet with these groups regularly at Global Grid Forum and other Grid related meetings, and will help as many groups as possible utilize the network monitoring data we will publish to develop “network aware” applications and middleware.

We also have very close working relationships with Globus [Globus], pyGlobus [PYGL], and pyGridware [PYGR] developers, as well as the bbftp [Bbftp] and bbcp [Bbcp] developers, and will work with them to ensure the network monitoring data is useful to them, and understand how they can use it for problems such as replica selection and Grid resource scheduling. We also have close connections with the DOE Science Grid SciDAC project, and will test and operate this network-aware Grid middleware on the DOE Science Grid. We will work with the Caltech

MonALISA team to provide easy access via Grid Services to data from multiple infrastructures. As a proof of principle, we have already started providing MonALISA with access to IEPM and E2E piPEs data.

Through the participation of the E2E piPEs project and the EPC project, the participants in this proposal already have measurement frameworks deployed on Abilene and ESnet and can guarantee that those major backbone networks will adopt the MAGGIE interoperability approach. Moreover, given the participants' contacts with the UltraScienceNet community, we will extend the MAGGIE interoperability approach to that network as well. Finally, members of the MAGGIE team are already working closely with ongoing efforts within the NLANR DAST Advisor project, the DANTE GN2 JRA1 project, and the MonALISA project, suggesting that most major measurement framework projects will be open to adopting the MAGGIE project's constructs, greatly increasing MAGGIE's value to the DOE community.

Since SLAC will be an early UltraScienceNet site, and LBNL is closely associated with ESnet operations, this will enable early testing and close cooperation when deploying monitoring infrastructures to and testing tools on UltraScienceNet with UltraScienceNet developers to ensure that the measurement infrastructure components we develop will operate in their extreme network environment.

Finally, we will work closely with the proposed project from CAIDA, titled "Pythia: Automatic Performance Monitoring and Problem Diagnosis in Ultra High-Speed Networks," to deploy and test new tools and analysis techniques over ESnet, Abilene, and UltraScienceNet.

7 Project Management

As currently envisioned, the MAGGIE project includes both group-wide and individual-institution components. Milestones related to the former category will ultimately be the responsibility of the entire group, with each principal investigator having a single vote in group-wide decisions. However, one principal investigator will be selected by the principal investigators to manage the overall project deliverables and serve as a flywheel for regular communications and group interactions. Milestones in the latter category will be the responsibility of the principal investigator associated with the institution. However, each principal investigator will be expected to provide quarterly updates to the other principal investigators outlining progress made and changes in plan.

Communication among the MAGGIE project participants (principal investigators and other parties working on the project, directly funded or not) will be via a variety of forums, as appropriate for different phases of the project.

- Conference Calls: The principal investigators will hold regular (at least biweekly) conference calls to discuss project progress. Quarterly updates will be staggered and usually held during such calls. Each principal investigator will designate a second who will attend conference calls in their absence.
- Web: Each institution will maintain an up-to-date task list and status for all activities related to the MAGGIE project. These web pages will also include all relevant contact information (phone, email, instant message handle, etc.) for all associated with the project. Principal investigators will be expected to update such pages at least every two weeks.
- The principal investigators will meet in person at least three times per year for a full day. Whenever possible, such meetings will be held in conjunction with appropriate conferences (e.g., GGF meetings, Internet2 meetings, etc.). The goal of such in-person collaborations will be review of progress made and group design of important architectural components.

The principal investigators will meet at least once per year with DOE program administrators to review annual progress and chart out the path going forward in the year to come.

8 Conclusions

The demanding needs of data intensive science and the associated high capacity connectivity have driven the research community to develop a large selection of tools to measure and test network performance within a country/state, across continents and trans-oceanic paths. An interoperable federation of network measurement infrastructures is needed in order to run the tools and collect the results. Furthermore such an infrastructure is critical to achieve a functioning Grid and enabling geographically separated scientists to effectively work together as a team.

As noted in the introduction to this proposal, the Performance Measurement Architecture Workshop 2003 identified five (5) major areas of research that need to be investigated to successfully develop a usable network

measurement infrastructure. MAGGIE will address each of these issues in order to reach its goal of federated, cross-domain network measurement. MAGGIE will foster collaboration between existing measurement infrastructure projects, enabling researchers to identify their most useful features and encouraging the sharing of ideas and code. In this manner, each infrastructure will be improved faster than if the individual teams worked alone.

9 Literature Cited

- [ABwE] J. Navratil and R. L. Cottrell, *ABwE: A practical approach to available bandwidth estimation,* PAM2003, available at http://moat.nlanr.net/PAM2003/PAM2003papers/3781.pdf
- [Advisor] http://dast.nlanr.net/Projects/Advisor/
- [Akenti], M.Thompson, A. Essiari, S. Mudumbai. *Certificate-based Authorization Policy in a PKI Environment* ACM Transactions on Information and System Security, Nov 2003. http://www-itg.lbl.gov/security/Akenti/
- [AMP] http://amp.nlanr.net/AMP/
- [BaBar] http://www.slac.stanford.edu/BFROOT/
- [Bbcp] http://www.slac.stanford.edu/~abh/bbcp/
- [Bbftp] http://doc.in2p3.fr/bbftp/
- [Bwctl] http://e2epi.internet2.edu/bwctl/
- [EDG] http://eu-datagrid.web.cern.ch/eu-datagrid/
- [EPC] https://performance.es.net/
- [EGEE] http://egee-intranet.web.cern.ch/egee-intranet/gateway.html
- [DataTag]: http://datatag.web.cern.ch/datatag/
- [Fast] http://netlab.caltech.edu/FAST/
- [Fed] "Federation of Identities in a Web Services World" A joint whitepaper from IBM Corporation and Microsoft Corporation July 8, 2003, Version 1.0
- [Gardner] M. Gardner, W. Feng, M. Broxton, G. Hurwitz, and A. Engelhart, "Online Monitoring of Computing Systems with MAGNET," *IEEE/ACM Symposium on Cluster Computing and the Grid (CCGrid'03)*, May 2003.
- [GGF] http://www.gridforum.org
- [Globus] http://www.globus.org/
- [GlobusMDS] http://www.globus.org/mds/
- [Grid2003] http://www.ivdgl.org/grid2003/
- [Gribble] Steven Gribble, Alon Halevy, Zachary Ives, Maya Rodrig, and Dan Suciu. *What can databases do for peer-to-peer*? In WebDB Workshop on Databases and the Web, 2001.
- [HS-TCP] http://www.icir.org/floyd/hstcp.html
- [IEPM-BW] http://www-iepm.slac.stanford.edu/bw
- [Infra] http://www.slac.stanford.edu/grp/scs/net/proposals/infra-mon.html
- [Iperf] http://dast.nlanr.net/Projects/Iperf/
- [MDS] http://www.globus.org/mds/mds2/
- [MonALISA] http://monalisa.cacr.caltech.edu/
- [NETARCHD] J. Lee, D. Gunter, M. Stoufer, B. Tierney, *Monitoring Data Archives for Grid Environments*, Proceeding of IEEE Supercomputing 2002 Conference, Nov. 2002, Baltimore, Maryland, LBNL-50216.
- [Netest] http://www-didc.lbl.gov/NCS/netest/
- [NETFLOW] http://www.ipdr.org/documents/ipfix/infomodel/draft-ietf-ipfix-info-01.txt
- [NETLOG] D. Gunter, B. Tierney, K. Jackson, J. Lee, M. Stoufer, *Dynamic Monitoring of High-Performance Distributed Applications*, Proceedings of the 11th IEEE Symposium on High Performance Distributed Computing, HPDC-11 , July 2002, Edinburgh, Scotland, LBNL-49698.

- [NetraMet] http://www.caida.org/tools/measurement/netramet/
- [NDT] R. Carlson, *Developing the Web100 based Network Diagnostic Tool (NDT)*, 2003 Passive and Active Measurement Workshop, San Diego, CA, April 2003.
- [Ng02] T. E. Ng and H. Zhang. *Predicting Internet network distance with coordinates-based approaches*. In *IEEE INFOCOM*, 2002.
- [NIMI] http://www.ncne.org/research/nimi/, and V. Paxson, A. Adams, and M. Mathis. *Experiences with NIMI.* In Passive & Active Measurement (PAM2000).
- [NMWG] http://www-didc.lbl.gov/NMWG/
- [NMWG-name] http://www-didc.lbl.gov/NMWG/docs/draft-ggf-nmwg-hierarchy-02.pdf
- [NTAF] T. Dunigan, M. Mathis and B. Tierney, *A TCP Tuning Daemon*, Proceeding of IEEE Supercomputing 2002 Conference, Nov. 2002, Baltimore, Maryland, LBNL-51022.
- [NWS] http://nws.cs.ucsb.edu/
- [Owamp] One way latency measurements. Available at http://e2epi.internet2.edu/owamp/
- [Pathload] M. Jain and C. Dovrolis. *End-to-end available bandwidth: measurement methoidolgy, dynamics and relation with TCP throughput.* In ACM SIGCOMM, 2002.
- [Pathrate] C. Dovrolis, P. Ramanathan, and D. Moore. *What do packet dispersion techniques measure?* In IEEE INFOCOM, 2001
- [Pchar] B. Mah. *Estimating bandwidth and other properties*. In Internat Statistics and Metrics Analysis Workshop, 2000
- [PingER] http://www-iepm.slac.stanford.edu
- [PIPES] http://e2epi.internet2.edu
- [PYGL] http://dsd.lbl.gov/gtg/projects/pyGlobus/
- [PYGR] http://dsd.lbl.gov/gtg/projects/pyGridWare/.
- [Qiperf] A. Tirumala, R. L. Cottrell, and T. Dunigan. Measuring end-to-end bandwidth with iperf using Web100. PAM 2002. Available at http://moat.nlanr.net/PAM2003/PAM2003papers/3801.pdf
- [RON] D. Anderson, H. Balakrishnan, M. F. Kaashoek, and R. Morris. Resilient overlay networks. In *SOSP*, 2002. [Row01] A. Rowstron and P. Druschel. *Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems*. In Proc. IFIP/ACM Middleware, November 2001.
- [RFC2903] Generic AAA Architecture, C. deLaat et. al. Aug 2000.
- [RFC2904] AAA Authorization framework, J. Vollbrecht et. al. Aug 2000.
- [SAML] http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security
- [SCNM] D. Agarwal, J. M. González, G. Jin, B. Tierney *An Infrastructure for Passive Network Monitoring of Application Data Streams*, 2003 Passive and Active Measurement Workshop , San Diego, CA, April 2003
- [scishare] K. Berket and D. Agarwal, *Enabling Secure Ad-hoc Collaboration*, Proceedings of the Workshop on Advanced Collaborative Environments, Seattle, WA, June 22, 2003. LBNL-52895. http://www-itg.lbl.gov/P2P/file-share/
- [p2pio] K. Berket, A. Essiari, D. Gunter, W. Hoschek, *P2PIO Protocol Specification Version 0.2*, http://dsd.lbl.gov/firefish/p2pio-spec/spec.pdf
- [Scriptroute] Neil Spring, David Wetherall, and Tom Anderson. *Scriptroute: A Public Internet Measurement Facility*
- USENIX Symposium on Internet Technologies and Systems (USITS), 2003. Available at http://www.cs.washington.edu/research/networking/scriptroute/papers/scriptroute.pdf
- [Shibboleth] http://shibboleth.internet2.edu/