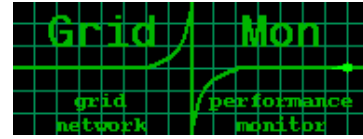


Grid Network Performance Monitoring (UK e-Science, 2004 All Hands Meeting)



Mark Leese (m.j.leese@dl.ac.uk) and Robin Tasker (r.tasker@dl.ac.uk)
CCLRC Daresbury Laboratory, Warrington, Cheshire WA4 4AD
<http://gridmon.dl.ac.uk/>

Abstract:

Writing about the virtues of Grid computing for the UK e-Science community would be a superfluous exercise. What cannot be underestimated however is the role the underlying network will play in the Grid's effectiveness.

Without network performance data, Grid middleware and applications cannot optimise their performance by adapting to changing network conditions, networks cannot be debugged for efficiency, and the Grid cannot support the measurable SLAs required for the "utility computing" model. To address some of these issues, a UK e-Science Grid network performance monitoring project began in June 2002.

Progress has been presented at each All Hands Meeting, and this continues for 2004 with an update focusing on the international Grid network monitoring work the UK has contributed to in the last year, via the GGF. Namely, development of XML schemas for communicating with network monitoring systems, and beginning proposals for delivering network functions as Grid services, with the end goal of transforming the network into a manageable Grid resource.

Some consideration is also given to work planned for the GridPP2 programme: development of a network diagnostic engine, a troubleshooting tool for use in Grid/ Network Operations Centre environments.

Glossary:

BW	Bandwidth to the World	OGSI	Open Grid Services Infrastructure
CCLRC	Council for the Central Laboratories of the Research Councils	piPEs	performance initiative Performance Environment system
EDG	European Data Grid	R-GMA	Relational Grid Monitoring Architecture
EGEE	Enabling Grids for E-science in Europe	RPC	Remote Procedure Call
GGF	Grid Global Forum	SLA	Service Level Agreement
GOC	Grid Operations Centre	SLAC	Stanford Linear Accelerator Centre
HPC	High Performance Computing	SNMP	Simple Network Management Protocol
IEPM	Internet End-to-end Performance Monitoring	SOAP	Simple Object Access Protocol
LDAP	Lightweight Directory Access Protocol	UDDI	Universal Description Discovery Integration
NCSA	US National Centre for Supercomputing Applications	W3C	World Wide Web Consortium
NDT	Network Diagnostic Tool	WP	Work Package
NGS	National Grid Service	WSDL	Web Service Description Language
NOC	Network Operations Centre	WSP	Web Service Provider
NREN	National Research & Educational Network	WSRF	Web Services Resource Framework
		WXS	W3C XML Schema
		XML	eXtensible Mark-up Language

monitoring, and briefly describe previous GridMon work.

Introduction

This paper summarises the work undertaken by the 'GridMon' Grid network performance monitoring project within the last 12 months.

To put the work in context, we will first consider the motivation for Grid network performance

Motivation

Network performance monitoring is crucial to the Grid. The resulting data is required for:

- Debugging networks for efficiency, an essential step for those wishing to use data intensive applications.

- Grid middleware and applications, for optimising their performance by adapting to changing network conditions (including the ability to be “self healing”).
- Supporting the Grid “utility computing” model, via measurable SLAs.

The concepts and practice of network monitoring are well understood and are widely used to identify problems, quantify performance and set expected levels of service. Monitoring for the Grid however is a special case, and must be given special treatment. As discussed in more detail below, debugging networks for efficiency now has much greater importance, while publication to middleware and SLA support are new concepts.

Debugging networks for efficiency

Network performance data has always been important in this process. The difference in the Grid arena is that it becomes an essential step for those wishing to use data intensive applications:

- Projects with large data sets, e.g. high energy and particle physics experiments, radio astronomy, and medical applications.
- High-bandwidth dependant projects, including real-time remote visualisation applications, such as the TRICEPS demonstration mentioned later in this section.

It is essential because simple over-provisioning of networks is not greatly aiding the end user. This also explains Grid network monitoring’s focus on end-to-end performance.

Publication to middleware and Grid applications

Making network performance data available to middleware and Grid applications is a significant new development. The very idea that performance data published to man and machine will allow the middleware and applications to achieve optimum performance by adapting to changing network conditions is a major driver for this work. It also relates to the Grid’s much publicised self-healing capability.

An adaptive Grid could take many forms, such as an application varying its transport strategy by tuning TCP parameters, or a more distributed case involving middleware, described below.

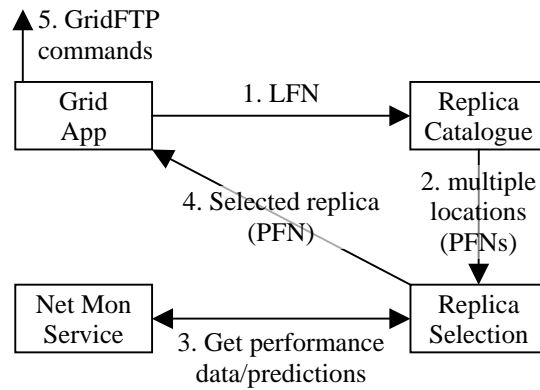


Figure 1: Replica Selection

File replication is a proven technique for improving data access. It involves distributing multiple copies of the same file across the Grid, the idea being that when a Grid resource such as a computing element requires the file, it will have access to a choice of copies, some of which may be more desirable to use than others.

A replicated file has a Logical File Name (LFN) which maps to one or more PFNs (physicals). A Replica/Replication Manager is responsible for replication issues, including maintaining the mapping between logical and physical filenames, and deciding what and where replicas should exist (normally based on recent usage patterns). The Replica Manager includes a Replica Selection Service which uses network performance data (from somewhere) to find the “best” replica. “Best” could be defined in several ways. The most obvious definition of “best” would be the “quickest obtainable”, but it could just as equally be that which will have the least impact on other network users, or the geographically closest. The principle is illustrated in figure 1.

Support for measurable SLAs

The Grid as “utility computing”, with its associated SLAs, may currently have little application in the academic world. Interest in the commercial world however is strong, and may be of significance to the DTI, enthusiastic backers of the e-Science programme.

In recognition of the importance of network monitoring as a whole, it is already defined as a key role of the proposed UK GOC.

As a final comment for this section, a clear and recent example of the importance of the network to the Grid is the real-time remote visualisation

TRICEPS HPC Challenge demonstration given at Super Computing 2003 [1]. UK e-Science projects played a major role in the demonstration, and it received the Award for Most Innovative Data-Intensive Application. Providing the required networking was however, a non-trivial, non-automated exercise. Further, much data belonging to the computational stage of the activity did not travel over a production network.

Previous GridMon Work

From inception, the GridMon project has aimed to establish a basic, UK-wide network monitoring infrastructure that could be extended and enhanced to provide performance data for:

- a. debugging networks
- b. publication to Grid middleware
- c. and to a lesser extent, potential SLA support

A basic infrastructure exists, with a presence established at each of the original e-Science Centres. Monitoring is performed by a kit of tools installed on a suitable computing node at each centre, with regular tests performed between all centres. A mesh of monitoring is thus created, allowing each centre to build a picture of the quality of its links to all other centres.

Performance data is stored locally on those nodes, and is published to interested humans via a web interface, shown in figure 2. Users are presented with a clickable UK map, which leads to a form, which in turn leads to plots of performance data.

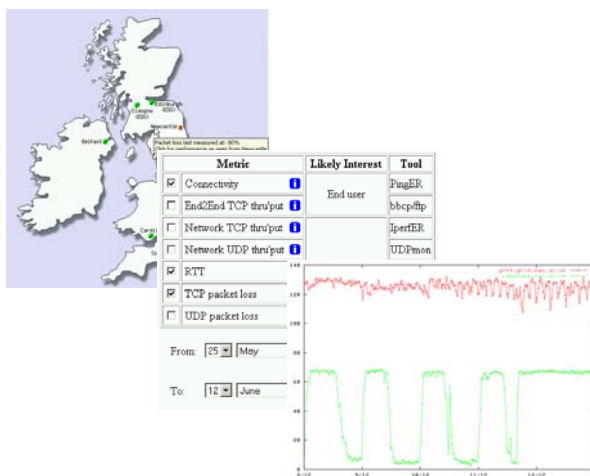


Figure 2: GridMon web interface

Given sufficient backing, this infrastructure is well poised to expand its coverage into the GridPP2 and NGS project domains.

GridMon is now in its second phase, publication of network performance data to Grid middleware. The following sections describe work carried out in this area within the last year.

Publication to Grid Middleware

During the lifetime of the project, various methods of publishing data to the Grid middleware have been touted, including LDAP and R-GMA. While LDAP may have enjoyed early prevalence in US projects, and R-GMA in European equivalents, the popularity of these technologies is waning, and there is a growing movement towards the use of web and Grid services.

Web services are essentially “online” applications accessed via XML messages. Grid services are essentially web services with some Grid specific add-ons and pre-requisites. The use of XML messaging allows web services to interface with each other, and Grid services to similarly interact.

When new technologies are developed, there is the inevitable temptation to quickly adopt them without considering their ‘true’ value, either to maintain your cutting-edge status or simply because everyone else is doing the same. In this case however, web and Grid services do offer real benefits. Most notably that application-application communication is far more open and direct.

The relevance of this hopefully becomes clear if for a moment we consider GridMon as a general network monitoring service, accessible by other services and applications. We can reason that its potential clients are those shown in figure 3. Note that they are both numerous and varied:

- Grid middleware and end-user software (Grid applications)
- other network services, e.g. a network cost function, which may for example estimate how long a file will take to retrieve from another site
- network administration software, as used by human administrators in Grid or network operations centre environments
- automated test systems, such as the Internet2 piPEs system [2].

- corresponding monitoring services in other administrative network domains.

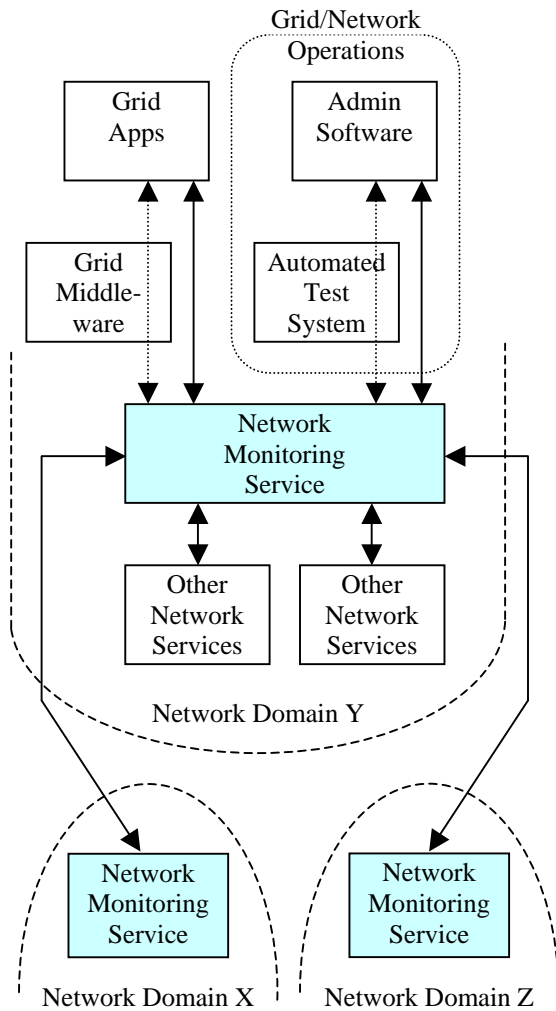


Figure 3: network monitoring service clients

Expanding on the last point, use of web and Grid services will lead to much easier integration of differing monitoring architectures, allowing systems to use functionality and data provided by others. The possibility of GridMon being able to share data with the monitoring efforts of Internet2 [2] in the US, and European NRENs via DANTE [3] should not be lost on the reader.

Had GridMon struck out on its own to develop web and Grid services interfaces to its data, then the work may well now be complete. However, global collaboration is the driving force behind the Grid, and it is therefore imperative to follow international standards. As a result, GridMon is

actively involved in network monitoring standards work with the appropriate body, the GGF.

GGF NM-WG [4]

To assist in data portability, the GGF Network Measurement-Working Group has proposed a network measurement classification system [5]. The NM-WG “hierarchy document” describes a set of network characteristics together with a classification hierarchy, both aimed at Grid applications and services. Application of the hierarchy will facilitate creation of common schemata for describing network monitoring data. Using a standard classification for measurements will maximise the portability of data.

NM-WG are also engaged in XML schema work, to which GridMon has been a lead contributor. This will be discussed, following a short introduction to web services.

The basic web service architecture is shown in figure 4. A client will search a UDDI registry for a service that is of interest. Searches can be performed based on business name, service name or a service category. To make initial contact with a service, the client is given the URL of the service’s WSDL document. This XML document describes the methods (functions) that the service has made available, and how the client should interact with them. Once the client has retrieved the WSDL document it can start using the service, via XML RPCs and XML messages encapsulated in SOAP messages.

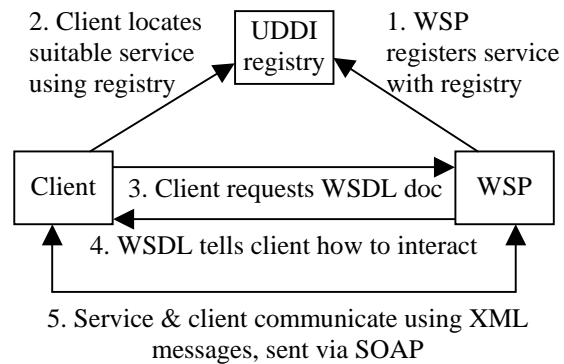


Figure 4: web service architecture

For simple implementations, the results of using services can be returned as simple data types, such as strings, as they would with other RPC implementations. The only difference here is that results are encapsulated in SOAP. This is

acceptable for simple transactions, but isn't at all practical when dealing with large and complex datasets or situations where the service could return differing amounts and/or types of data. Enter the schema, a self-describing method of representing data. The self-describing nature makes it easier to share data between clients and services that are capable of parsing schemas (being flexible about what data they can send and receive).

Work has begun on producing XML based network monitoring schemas, spearheaded by the NM-WG [4] and based on their own hierarchy document.

As illustrated in figure 5, in fully interactive systems, clients will be able to request historic data, future or on-demand tests or predictions (as popularised by such monitoring tools as the Network Weather Service [7]). Results can then be returned. All request and result messages can be formatted using standardised schemas, a truly powerful combination.

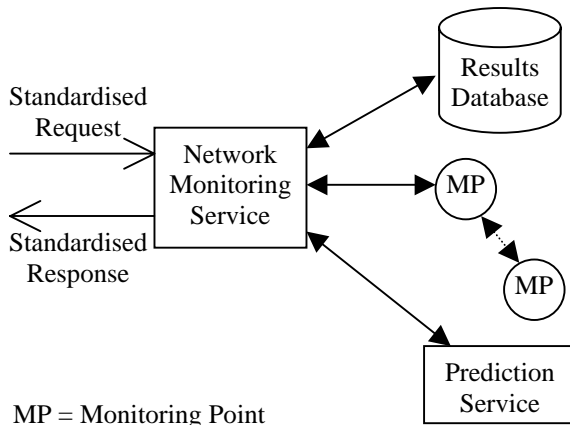


Figure 5: request-response mechanism

Although a slight aside, it should be noted that the prediction service does not predict how the network will react to certain traffic, but gives predictions of what bandwidth will be available, what RTT will be etc. Middleware and Grid applications can then make decisions based on those predictions.

Some may ask "How can future performance be predicted if there is no control over what the network will be doing in the future (via control of the underlying network pipes)?" Figure 6 provides a response.

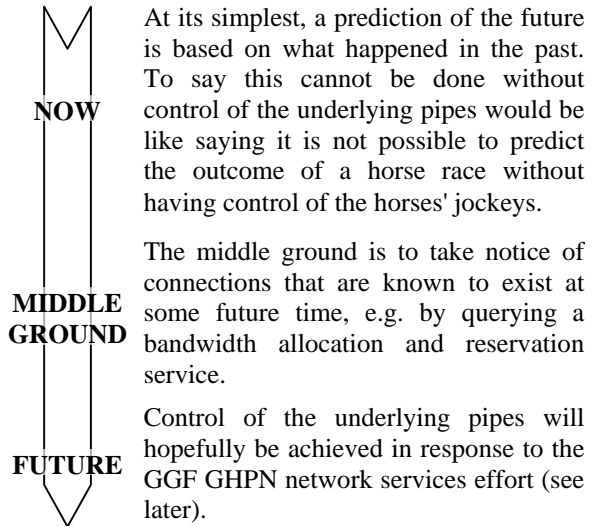


Figure 6: Prediction service options

SNMP [8] is concerned with the low-level management of network devices. The NM-WG request-response schemas relate to higher-level software (web services) which are more flexible, more powerful, and are a better fit with the Grid environment.

Discussing the schemas in detail is beyond the scope of this paper. However, it worth considering what types of request can be made. Three textual examples are given below. They highlight the flexibility of requests, most notably that requests can range from very simple, to very complex.

Simple: Give me the last value for available bandwidth to host X. I'm only interested in something in the last 30 minutes. If there isn't any data, don't run any new tests.

More Complex: Give me the last value for available bandwidth to host X. I'm only interested in something in the last 30 minutes. If there isn't any data, run an on-demand test, using

- between two and four TCP streams
- a 4MB TCP buffer size
- the iperf tool if available, pathchar as a second choice, and any other bandwidth tool as a third

Historical Query: Give me a maximum of 20 results for available bandwidth to host X from the last 24 hours. If there are more than 20 matches,

give me the 20 values closest to the start of the period. Report what parameters were used in the tests, and provide a mean of the results.

Referring back to figure 5, it should also be noted that network monitoring services are not currently required to provide:

- access to past data, and
- support for on-demand/future tests, and
- predictions

What features systems should support is a matter for separate discussion. The important step at this stage is making available a powerful yet flexible request-response mechanism that many will be happy to use, thus simplifying the sharing of data and interoperation of network monitoring systems.

For those interested in implementation details:

- Schemas have been produced in RelaxNG format, a schema language of OASIS, a web service focused consortium for structured information standards. Its advantage over the more common WXS schema languages from the W3C (also implemented) are better readability, and better support for modularisation.
- Trial software making use of the schema work, albeit using early WXS versions of the schemas, has been produced at SLAC, Georgia Institute of Technology and NCSA.
- UK investigation is also taking place into the use of Document/Literal SOAP binding, as opposed to the more common RCP/Encoded. There are some concerns over the ability of RCP/Encoded messages to be easily validated against their parent schema.

Further work for the NM-WG includes investigation into capability discovery, that is, how clients can discover the capabilities of monitoring systems available to them. For example, what measurements can a service make, and what data already exists?

Unfortunately, it is difficult to add further detail to this section, the changes in web and Grid services being fast paced. As we write, we await to see the full impact of recent developments in web and grid services, namely WSRF, a modification of the GGF OGSi standards in conjunction with the web services community.

GGF GHPN-RG [9]

GridMon has contributed to the group's initial "network services" document [10]. This defines requirements for various network sub-services, which could be combined to form a holistic network service, allowing the network to be treated as a Grid resource. This in turn produces a simple stack as shown in figure 7. With computing nodes, storage elements and the interconnecting network as resources, management can be easier and end-to-end.

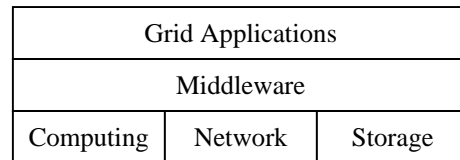


Figure 7: GHPN stack

The document was a brainstorming exercise, designed to whet peoples' appetites. It is now being edited using more formal methods, starting with the addition of use cases.

The proposals will include a network monitoring service, possibly publishing direct to middleware, and GridMon should take this into account.

International Exposure

Through international collaborations such as those summarised above, GridMon has good exposure to other monitoring initiatives, including those of Internet2 and DANTE. This is a UK e-Science project, but it does not exist in a vacuum, and is evolving to show the best way to carry out monitoring, based on the best techniques and technologies from around the world.

Indeed, there is potential for the project and the UK to gain further, by collaborating in a proposed GGF research group specifically aimed at answering questions such as "what network characteristics are most useful to Grid Applications?"

Diagnostic Engine

The arrival of the Grid has prompted more network performance tests to be run than ever before, with the resulting data often stored for later use. In addition, the Grid is pushing forward the design and implementation of systems that will be able to perform tests without human

intervention. This makes possible the development of automated test tools, which can localise and identify a variety of network faults and inefficiencies by leveraging historic data and/or running and requesting on-demand tests.

Such diagnostic tools can vary greatly, not least by placing their emphasis on analysing different aspects of performance. One tool may concentrate on TCP performance for example, while another focuses on efficient routing. There are two predominant types however, based on the user group at which a tool is aimed:

End-user – tools are used by network end-users when they discover or suspect that their network connectivity is not functioning at it should. An example would be the NDT (Network Diagnostic Tool) [11] tool originally produced at Argonne National Laboratory, and now being further developed by Internet2.

Clients access the service via a Java applet embedded in a web page loaded from the NDT server. Once initiated, tests are run between the applet and the NDT server, a Linux machine with a Web100 enabled kernel. Results are then displayed to the user.

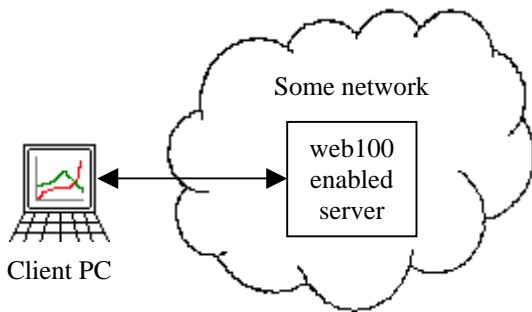


Figure 8: NDT

Although an excellent system in many respects, the NDT may not be suitable for all applications. Firstly, it assumes that the human client is aware of the address of a relevant NDT server. Secondly, no use is made of historic data, and thirdly, the NDT focuses on TCP performance alone.

Another tool worthy of note is the “Detective” [12] of SURFNet, the Dutch NREN. As with the NDT, the tool communicates with servers in the network, unsurprisingly SURFNet in this case. The system is novel in that it requires the client to

install software on their work station, and it allows TCP and UDP throughput tests to be run on-demand.

Grid Operations Centre – tools are used by experienced network staff to debug the network and correct faults. This is the approach to be adopted by GridMon for the GridPP2 [13] programme. This particular diagnostic tool will be developed in response to the requirements of GOCs and the EGEE project, with reference to the aforementioned Internet2 work. No other requirements are set, but the emphasis is likely to be on staff requesting in-depth tests when they suspect faults or inefficiencies have arisen, perhaps in response to fault “tickets” raised with the a GOC helpdesk.

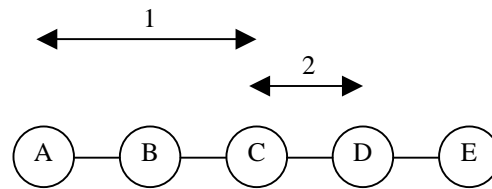


Figure 9: divide and conquer;
GridPP2 diagnostic engine

It is likely that any such system will build upon a deployed network of monitoring and/or test nodes. Consider figure 9. Let us assume we have detected a problem on the A-E network path, and that we have access to a network monitoring service at each node. Tests could be run between A and C. If that path is detected as “okay” (and assuming that the fault isn’t a cumulative problem) you could surmise that the fault lies between nodes C and E. So tests could then be run between C and D. If the problem does not lie there, it could be assumed to exist between nodes D and E.

The GridPP2 work is only just beginning. As a result, it has only been possible to highlight the potential usefulness of such a tool, and the potential avenues to be explored in its creation.

Work will begin with a study of existing tools, and a requirements capture exercise, with plans to leverage similar efforts scheduled within the EGEE JRA4 “Development of Network Services” group [14].

Conclusion

The last year of the project has gone well. GridMon has been involved with the work of various international groups, and in some cases has helped lead relevant efforts. This not only ensures that the UK is well represented in such activities, but also provides the preparatory work for GridMon to evolve into a “best of breed” monitoring solution, building on the work of the GGF, Internet2, SLAC and others, acknowledged leaders in their respective fields.

The next 12 months will see continued contributions to such Grid networking efforts, with a focus on bringing a well defined deliverable from the NM-WG schema activity. This will allow a GGF NM-WG compliant interface to developed for the GridMon infrastructure.

It is also hoped that GridMon’s scope can widen to become an infrastructure for GridPP2 and the NGS also.

As demonstrated, the diagnostic engine work has obvious benefits. It should also provide useful knowledge and proof of concepts that can be fed back into contributing projects such as Internet2’s piPEs initiative and EGEE. All these strands of work are being carried out because they will prove to be genuinely useful, rather than being the proving of a technology. The future therefore, is again bright.

Acknowledgements

The work described here is supported by core e-Science funding, and by the GridPP-2 project. The initial GridMon infrastructure was closely coordinated with WP7 of the EDG project [15], and benefited from the IEPM work at SLAC[16], and multicast work at Manchester e-Science[17].

References

1. Transcontinental RealityGrids for Interactive Collaborative Exploration of Parameter Space (TRICEPS):
<http://www.sve.man.ac.uk/Research/AtoZ/RealityGrid/TRICEPS-SC03.pdf>
2. piPEs:
http://e2epi.internet2.edu/E2EpiPEs/e2epipe_index.html
3. Dante inter-domain performance monitoring:
<http://www.dante.net/tf-ngn/perfmonit/>
4. GGF NM-WG:
<https://forge.gridforum.org/projects/nm-wg>
5. B. Lowekamp, B. Tierney, L. Cottrell, R. Hughes-Jones, T. Kielmann, and T. Swany. *A Hierarchy of Network Performance Characteristics for Grid Applications and Services*, Global Grid Forum, 19 June 2003:
<http://www.didc.lbl.gov/NMWG/docs/draft-ggf-nm-wg-hierarchy-00.pdf>
6. R.J. Allan, D. Chohan, X.D. Wang, M. McKeown, J. Colgrave, and M. Dovey. *UDDI and WSIL for e-Science*, Grid Support Centre, 2002:
<http://esc.dl.ac.uk/Papers/UDDI/uddi/uddi.html>
7. Network Weather Service:
<http://nws.cs.ucsb.edu/>
8. SNMP, RFC 1157, IETF:
<http://www.ietf.org/rfc/rfc1157.txt?number=1157>
9. GGF GHPN-RG:
<https://forge.gridforum.org/projects/ghpn-rg>
10. G. Clapp, T. Ferrari, D.B. Hoang, T. Lavian, M.J. Leese, P.D. Mealar, I. Monga, V. Sander, and F. Travostino. *draft-ggf-ghpn-netservices-1.0*, Grid Global Forum, February 2004:
<http://forge.gridforum.org/projects/ghpn-wg/>
11. Network Detective Tool:
<http://miranda.ctd.anl.gov:7123>
12. SURFNet Detective:
http://detective.surfnet.nl/en/index_en.html
13. GridPP2, the Grid for UK Particle Physics:
<http://www.gridpp.ac.uk/>
14. EGEE JRA4, Development of Network Services: <http://egee-jra4.web.cern.ch/EGEE-JRA4/>
15. EDG WP7, Network Services:
<http://ccwp7.in2p3.fr/>
16. IEPM-BW: <http://www-iepm.slac.stanford.edu/bw/>
17. miperfer: <http://www.csar.cfs.ac.uk/staff/daw/>