

# BNL Network Performance Challenge in Grid Environment

## *Project Description*

The unprecedented volume of data already being generated by RHIC experiments at Brookhaven National Lab is demanding new strategies for how the data is collected, shared, transferred and managed. For example, STAR experiment is already collecting data at the rate of over a Tera-Byte/day. The volume of data will increase by a fact of 10 in the coming five years. To manage and process this mount of data is not a trivial task. Many strategies are investigated and implemented. Grid technologies are the commonly adopted strategy by Particle Physics Data Grid (PPDG). These Grid technologies are deployed at many universities, research institutes and DOE national labs that participate PPDG. They will be able to replicate and analyze information from an interconnected worldwide grid of tens of thousands of computers and storage devices. (Reference [1]). Grid Environment requires sustained network to transfer of large amounts of bulk data between collaborating sites with high bandwidth. The Bandwidth Projects to the STAR Collaborating Sites is designed to measure and monitor the current data transfer capabilities to several collaborating sites with Gigabit network links. After we have successful experiment on this project, we will extend the STAR Bandwidth Challenge Project to other RHIC experiments because PHENIX/PHOBOS/BRAHMSz has similar data management requirement to STAR. The CERN LHC experiments expect to collect over ten million Tera Bytes when they turn to full luminosity. The project will benefit LHC experiment as well.

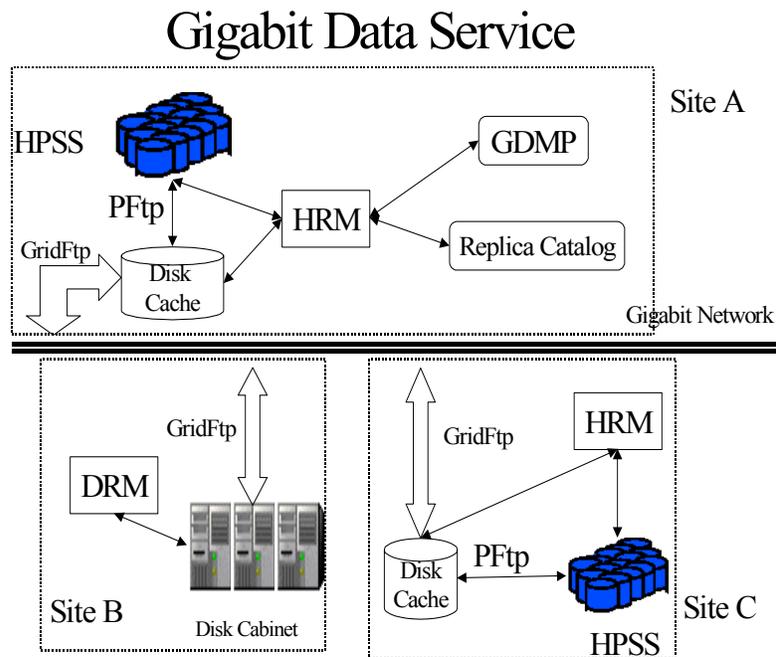
## **Tasks**

1. Achieve Gigabit transfer within local network infrastructure using variously FTP protocols, such vanilla FTP, bbFTP, and GridFtp. We will also deploy several high-level data services in BNL RHIC computing facility. These services are built on top of this ftp transfer protocols, For example, HRM and GDMP. We will also analyze the performance on network routes, switch and firewall in Grid environment.
2. Apply Web 100 (Reference 10) software and tools in the project, conduct real-time monitoring on the Gigabit network link, and research the impact of Web 100 on Gigabit network and high-performance data service.

## **Details of the Project**

Figure 1 shows the system infrastructure. We will set up a data server which is a high performance Linux nodes. It has two Ethernet cards with Gigabit throughput and a fast Ethernet card: one gigabit Ethernet card connects to HPSS mover and NFS file servers in BNL via gigabit link. The other gigabit Ethernet card connects to multiple Linux client hosts. The server host is shown as Site A in the following Figure 1. The second Ethernet card will be used for accepting data request, gathering performance statistics, and reporting the monitoring information in real time and recording the results. This server locates within RCF/ATLAS domain. We will set up several client hosts that have either gigabit Ethernet card or fast Ethernet card. In Figure 1, we show the client hosts, Site B is client with large amount disk storage and Site C is the client with HPSS.

The project works as follows: The GDMP [5] server on Site A keeps track of the files stored at the Site A, Site B and Site C subscribe to Site A. When a new file is ready to publish at Site A, the GDMP server at Site A is responsible to send the new file to all of the subscription sites. The GDMP server communicates with HRM [8], requesting space allocation and file pinning, and making requests for file transfers. The HRM at Site A will contact with the DRM at Site B and HRM at Site A, negotiate with remote these SRMs to allocate space for the files, invoke the file transfer services (GridFtp, BBFtp) to push the new file from Site A to Site B, C. At the initial stage, we will only use the GridFtp, BBFtp to push data between the physical memory at Site A and memory at Site B, C. We will also test the throughput performance of the data transfer among local disks of Site A, B, and C. The final stage of the project will use GDMP and SRM (HRM and DRM) for massive data transfer.



**Figure 1 The infrastructure of STAR Data Services**

## The benefit of this project

Network path performance directly affects the users satisfaction and operating costs. Monitoring, measuring, analyzing and tuning network path performance is a fundamental component of gigabit network operations. We have not conducted such operation in Gigabit Ethernet yet. The complexity of the network infrastructure makes this task extremely hard. To conduct this operation only in local network is not a trivial task because it involves various network path, complex topology, link media, routers, switches and firewalls. We already found the impact of firewall on fast Ethernet with 100 mbps bandwidth. The effect of firewall on Gigabit link is unknown yet. This project can help up to get experiences on Gigabit network environment, identify the problem of the components in network infrastructure, prepare for the data replication service with gigabit throughput between Tier 0 (BNL) and other collaborating sites.

## Background

1. Globus: The Globus project is a multi-institutional research effort that seeks to enable the construction of computational grids providing pervasive, dependable, and consistent access to high-performance computational resources, despite geographical distribution of both resources and users. Reference [2,5]
2. GridFtp: GridFTP is a high-performance, secure, reliable data transfer protocol optimized for high-bandwidth wide-area networks. The GridFTP protocol is based on FTP. It implements a set of protocol features and extensions defined already in IETF RFCs. All these features and extensions are not implemented in vanilla FTP. GridFtp has also added a few additional features to meet requirements from current data grid projects. Reference [3, 4]
3. BBFtp: BBFTP is a piece of software designed to quickly transfer files across a wide area network. BBFTP is divided in two parts: a daemon and several clients. The clients are talking with the daemon on a control connection, and are transferring data through several parallel TCP streams. Reference [11]
4. SRM and HRM: Storage Resource Managers (SRMs) are middleware software modules whose purpose is to manage in a dynamic fashion that resides on the storage resource at any one time. SRM do not perform file movement operations, but rather interact with operating systems, mass storage systems (MSSs) to perform file archiving and file staging, and invoke middleware components (such as GridFTP) to perform file transfer operations. There are several types of

SRMs: Disk Resource Managers (DRMs), and Hierarchical Resource Managers (HRMs). Reference [8]

5. GDMP: The GDMP client-server software system is a generic file replication tool that replicates files securely and efficiently from one site to another in a Data Grid environment using several Globus Grid tools. In addition, it manages replica catalogue entries for file replicas and thus maintains a consistent view of names and locations of replicated files. All kinds of file formats are supported for file transfer. Reference [6,7]
6. Web 100: The Web100 project will provide the software and tools necessary for end-hosts to automatically and transparently achieve high bandwidth data rates over the high performance research networks. The Web100 software suite will endow TCP with better instrumentation. This instrumentation is the foundation for both the TCP auto tuning performed in process-level code, as well as the process-level tools designed to locate bottlenecks within the following major subsystems: the sending application, the sending OS, the Internet path, the receiving OS, and the receiving application. These software and tools will also provide means for performing dynamic, transparent, automatic TCP tuning for user level application. Initially the software and tools will be developed for the Linux operating system, but will be done in a standard, open manner so that they can easily be ported to other operating systems. Currently, the software is distributed as Linux 2.4. 11 kernel patches. Reference [9, 10]

### Network environment we have:

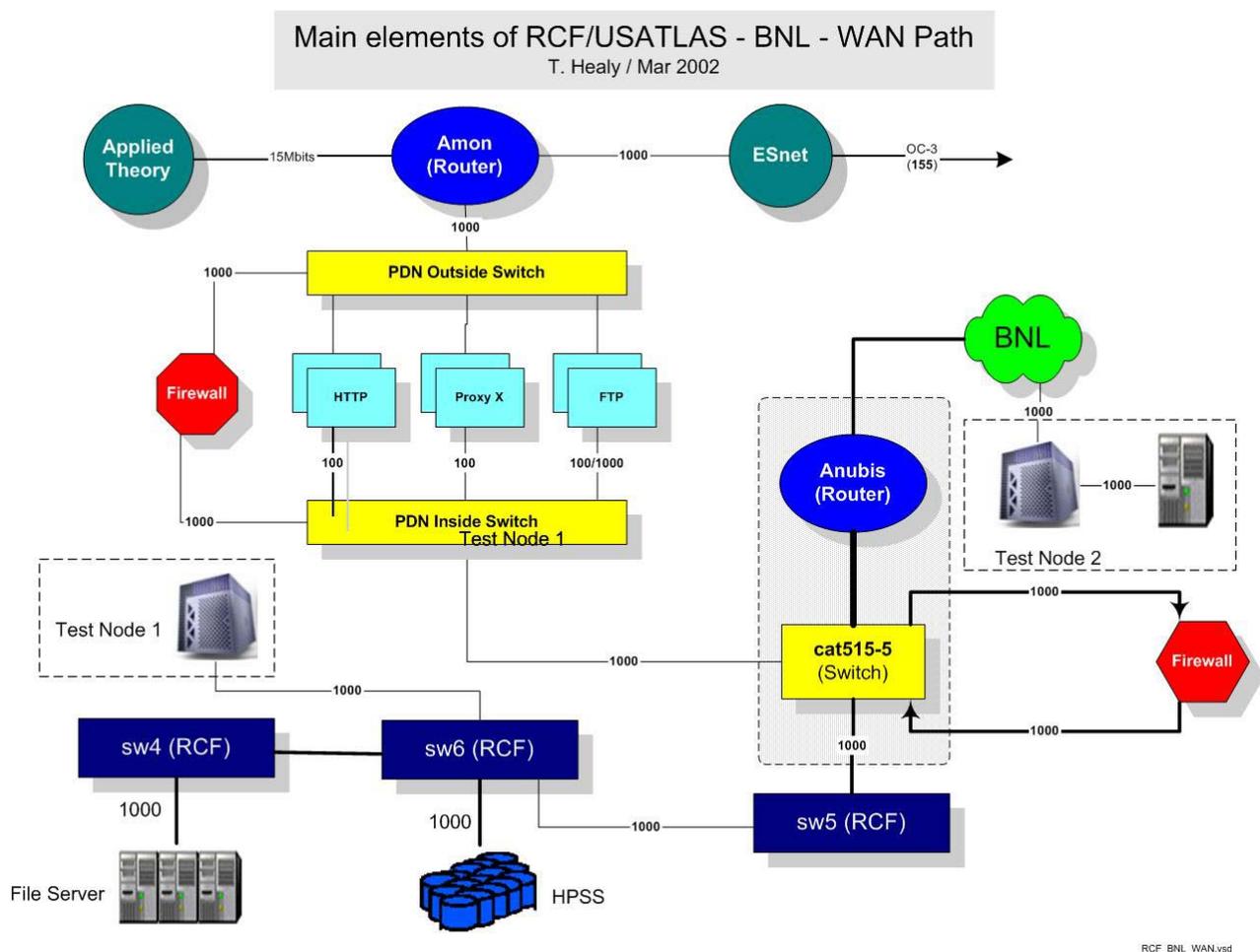


Figure 2 The BNL Local Network Topology

Figure 2 shows the topology of BNL local network. The RHIC / US ATLAS facilities are connected to the switch at the bottom right. These facilities are firewall /proxy isolated from the rest of the BNL site. The firewall is a Cisco PIX 535 that supports Gigabit throughput. BNL recently installed a Cisco 6500 switch at ATLAS domain. Therefore, a gigabit link is established among RCF/ATLAS domain. Brookhaven national lab has a gigabit local backbone. It is shown as the line between Switch CAT-5 and the perimeter firewall in the following figure. There are several gigabit ports available for this project. Test Node 1 shown in Figure 2 is the data server which locates inside the RCF/USATLAS firewall, and Test Node 2 shown in Figure 2 is the data client which is outside of the RCF/USATLAS firewall.

## **End System requirements**

To perform the test, it needs at least one powerful host. Here are the system requirements:

- CPU (dual 1.2 or up)
  - PCI slot: Either 64 bits or 66 MHz bus speed or higher, both of them should be much better.
- Memory requirement: 1G Byte and up. Support parallel tcp streams.
- Storage Requirement.
  - One SCSI Driver, The capacity should be at least 70 GB.
- Dual Link Gigabit Network Cards:
  - a). Optical fiber connection.
  - b). Support Jumbo frame, the maximum MTU should be not smaller than 9000 bytes.
  - c). Checksum offloading from CPU, i.e. Checksum is generated on Network Card.
  - d). The card will gather N frames before sending an interrupt to the CPU.

## **The target systems which satisfy the requirements above and Pricing**

The following system price is \$5834.00.

Parts Name	Part Number	Unit Price (US \$)	Quantity	Total Price (US \$)
New IBM @server xSeries 330 (MXT) (1 U), & PIII 1.26 CPU		\$1729	1	\$1729
SysKonnnect SK-NET SK-5521 Network Adapter	06P3701	\$509	2	\$1,018.00
IBM 73.4 GB 10K-rpm Ultra160 SCSI Hot-Swap SL HDD	06P5756	\$1,099.00	1	\$1099.00
xSeries 1.26 GHz/133MHZ-512KB, PIII Processor (Second CPU)	25P2836	\$999.00	1	\$999.00
IBM 1GB PC133 ECC SDRAM RDIMM	33L3326	\$989.00	1	\$989.00
Total Price				\$5834

## **Reference**

1. SC2001 Bandwidth Challenge Proposal: Bandwidth to the World L. Cottrell, 2001, <http://www-iepm.slac.stanford.edu/monitoring/bulk/sc2001/>
2. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. I. Foster, C. Kesselman, J. Nick, S. Tuecke; January.2002.
3. Data Management and Transfer in High-Performance Computational Grid Environments. W. Allcock, J. Bester, J. Bresnahan, A. Chervenak, I. Foster, C. Kesselman, S. Meder, V. Nefedova, D. Quesnel, and S. Tuecke. Parallel Computing, 2001.
4. Protocols and Services for Distributed Data-Intensive Science. A. Chervenak, I. Foster, C. Kesselman, S. Tuecke. ACAT2000 Proceedings, pp. 161-163, 2000.
5. The Globus Project: A Status Report. I. Foster, C. Kesselman. Proc. IPPS/SPDP '98 Heterogeneous Computing Workshop, pp.4-18, 1998.

6. Heinz Stockinger, Asad Samar, Bill Allcock, Ian Foster, Koen Holtman, Brian Tierney. File and Object Replication in Data Grids, 10th IEEE Symposium on High Performance and Distributed Computing (HPDC-10), San Francisco, California, August 7-9, 2001.
7. GDMP User Guide.
8. Storage Resource Managers: Middleware Components for Grid Storage, Arie Shoshani, Alex Sim, Junmin Gu Nineteenth IEEE Symposium on Mass Storage Systems, 2002 (MSS '02)
9. WEB100: Facilitating Effective and Transparent Network Use.  
[http://www.web100.org/docs/statement\\_of\\_work.php](http://www.web100.org/docs/statement_of_work.php).
10. Web100 Concept Paper. [http://www.web100.org/docs/concept\\_paper.php](http://www.web100.org/docs/concept_paper.php).
11. BBFTP: <http://doc.in2p3.fr/bbftp/>.