BNL WIDE AREA DATA TRANSFER FOR RHIC AND ATLAS: EXPERIENCE AND PLAN

M. Chiu, W. Deng, B Gibbard, Z. Liu, S. Misawa, D. Morrison, R. Popescu, M. Purschke, O. Rind, J. Smith, T. Throwe, Y. Wu, D. Yu
Physics Department
Brookhaven National Lab

Abstract

We describe two illustrative cases in which Grid middleware (GridFtp, dCache and SRM) was used successfully to transfer hundreds of TeraBytes of data between BNL and its remote RHIC and ATLAS collaborators. The first case involved PHENIX production data transfers to CCJ, a regional center in Japan, during the 2005 RHIC run. Approximately 270 TB of data, representing 6.8 billion polarized proton-proton collisions, was transferred to CCJ using GridFtp tools. The second case involved transfers between the ATLAS Tier 1 center at BNL and both CERN and the US ATLAS Tier 2 centers, as part of the ATLAS Service Challenge (SC). The work described in this paper demonstrated the current level of maturity of Grid tools being used by large physics experiments to satisfy their data distribution requirements.

INTRODUCTION

LHC and other HENP (High-Energy and Nuclear Physics) experiments, such as RHIC and BaBar, are breaking new ground, both in terms of the amount and complexity of data involved, but also in the size and global distribution of the collaborations themselves. This leads us to the fundamental challenge that must be addressed for LHC-scale physics: enabling collaborative data sharing.

The Problem: Enabling Collaborative Data Sharing and Computing Leveraging

Particles collisions at increasingly large energies have provided rich and often surprising insights into the fundamental particles and their interactions. Experimentation at increasing energy scales and sensitivity, along with the greater complexity of measurements, have necessitated a growth in the scale and cost of the detectors, and a corresponding increase in the size and geographic dispersion of scientific collaborations as well as in the volume and complexity of the generated data.

The Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory has hundreds of physicists from around the world using RHIC to study what the universe may have looked like in the first few moments after its creation. The largest collaborations today, such as CMS and ATLAS [3] which are building experiments for CERN's Large Hadron Collider (LHC) program, each encompass ~2,000 physicists from 150 institutions in more than 30 countries. Current and future HENP experiments face unprecedented challenges in terms of: (1) the *data-intensiveness* of the work, where the data volume to be processed, distributed and analyzed is now in the multi-PetaByte (10^{15} Bytes) range, rising to the ExaByte (10^{18} Bytes) range within a decade; (2) the *complexity* of the data, particularly at the LHC where rare signals must be extracted from potentially overwhelming backgrounds; and (3) the *global extent* and multi-level organization of the collaborations, leading to the need for international teams in these experiments to collaborate and share data-intensive work in fundamentally new ways.

Grid Computing for HENP

Unprecedented data processing requirements outpacing the capacity of any individual computing sites, along with the prevalence of cost-effective high speed networks, gave rise to the new era of a collaborative computing paradigm: Grid. RHIC and ATLAS Grid enabled infrastructures are distributed continentally and internationally. When completed, the Grid will provide Petaflops of computing power distributed over many sites, and manage and store PetaBytes of data per year. The cargo airplane used for data transfer between collaborators is being replaced by a network operated by grid transfer tools. In this paper, we provide two examples of data transfers in the RHIC and LHC programs:

- The first case involved PHENIX production data transfer to CCJ, a regional center in Japan, during the 2005 RHIC run. Approximately 270 TB of data, representing 6.8 billion polarized proton-proton collisions, was transferred to CCJ using GridFtp tools [1].
- The second case involved transfers between the ATLAS Tier 1 center at BNL and both CERN and the US ATLAS Tier 2 centers, as part of the ATLAS Service Challenge (SC) [2].

PHENIX DATA TRANSFER

During the 2005 RHIC run, a substantial fraction of the RHIC computing facility was being used to process data from the previous year. The new proton-proton data was therefore moved to CCJ to utilize a computing resource comparable to the PHENIX in-house computing cluster. The p+p event size is 40 KBytes and the event rate is 4 Khz, giving a data rate of 160 MByte/second. The beam uptime factor is ½. Therefore, the aggregate data transfer rate is up to 80 MByte/second. BNL and

CCJ set up the following infrastructure for the data transfer, as shown in Figure 1. BNL set up 6 dual-CPU on-line data acquisition servers, each of which was connected to eight SCSI drives via a raid controller. CCJ setup 4 dual-CPU servers, each of which had 4 TeraByte SATA disks. Servers at both sides had gigabit network cards. A GridFtp server and required DOE certificates were installed at both ends. The local network was reconfigured and tuned to route data directly from the online data acquisition system to the BNL public network, thus avoiding the use of tape storage as an intermediate buffer and preserving the scarce resource of tape I/O bandwidth. We tuned both the file system and the network TCP stack for the end hosts. We found that the built-in file system (EXT3) degraded under the parallel data transfers required by a long round trip network transfer. XFS was tuned along with the OS kernel to reach the required disk I/O performance before the actual data transfer.

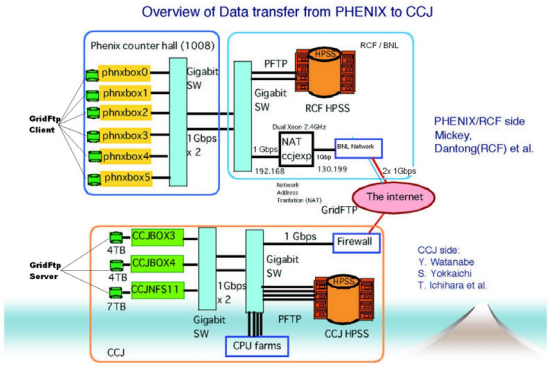

Figure 1: PHENIX Data Transfer Infrastructure.

The 2005 RHIC run ended on June 24th. A transfer speed of 60 MB/s was achieved around the clock, sufficient to keep up with the incoming data stream from the detector. The peak data transfer rate was 100 MB/second, as shown in Figure 2. About 270 TeraBytes were moved along the PHENIX internal network, RCF LAN, BNL campus LAN, DOE Esnet, Inter-continent SINET, CCJ LAN, and finally stored in the CCJ HPSS system, as shown in Figure 3.

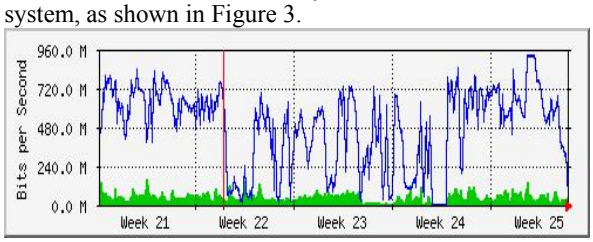

Figure 2: Data Transfer Rate from BNL During the Last Month of the 2005 RHIC Run

When several ESnet and SINET outages happened during the run, the data transfer was rerouted to alternate paths. The problems were promptly discovered and resolved by on-call personnel and network engineers. Because of large disk caches at both ends, no data were lost due to the few hours of network outages.

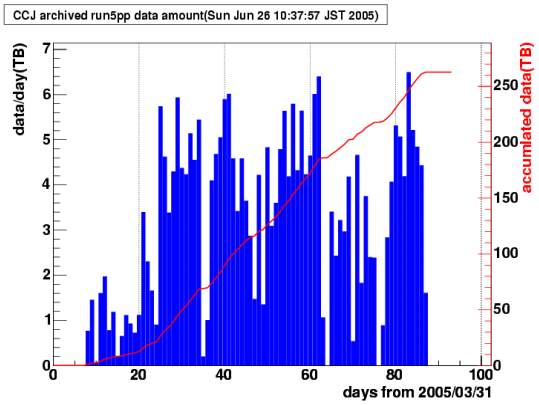

Figure 3: PHENIX Data Transfer Daily Rate

BNL LHC SERVICE CHALLENGE

Before the LHC ramps up to full production level, a series of service challenges is being performed to validate the Grid based infrastructure role in satisfying the intensive requirements of data volume and complexity generated by LHC. This role includes transferring raw data from on-line detector systems at CERN to Tier 1 sites in several continents with progressively higher data rate, and processing and analyzing the increasing data volume with the involvement of even more sites. The service challenges will help the infrastructure managers identify and solve problems of stability, scalability and performance. The first two challenges focused on sustained reliable file/data transfer between LHC Tier 0 sites and Tier 1 sites for extended periods of time in a production-like environment. The third one concentrated on verifying services, building on the Grid infrastructure to provide enough capacity, stability and scalability for a range of physics processes and their corresponding software. The final service challenge will be finished six months prior to LHC data taking, with a target capacity of twice the nominal rate specified in the Computing TDR [3] in order to buffer the peak data rate, backlog due to partial infrastructure failure, and even provide a temporary sharing facility to other peers which have failures.

In this section, we discuss how we were able to demonstrate 150 MB/s wide area data transfer rates using the SC infrastructure with our dCache configuration.

Service Challenges 1 & 2

We used four file transfer servers with a 1 Gigabit WAN network connection to CERN. The performance and throughput challenges (70~80 MB/second disk to disk) were met. We enabled data transfer between dCache/SRM and CERN SRM at OpenLab.

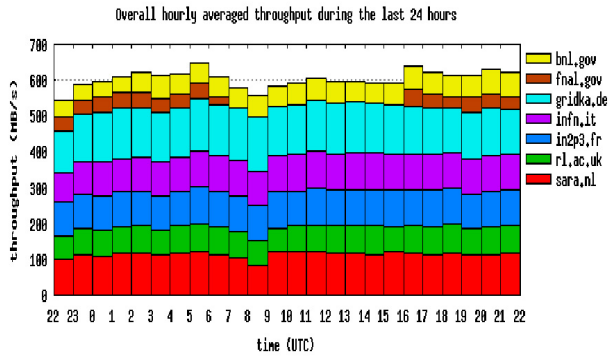

Figure 4: BNL Service Challenge 2 Data Transfer as Shown in GridView.

Many components needed to be tuned. The performance was not stabilized 24/7. The long 120 ms network RRT high packet dropping rate prevented us from utilizing the 1 Gbps BNL WAN connection. We had to use multiple TCP streams and multiple file transfers to fill up the network pipe. We found that our data transfer servers had sluggish parallel file I/O with EXT2/EXT3 and many processes in "D" states. When we increased the number of file streams, the performance of the file system became worse. We switched to the XFS file system and saw improvement. We still needed to tune file system parameters and hard drivers to optimize the overall data transfer performance.

The I/O scheduling algorithm in RedHat Linux could not provide optimized scheduling for disk read/write and network TCP processing. We observed a lack of parallelism between these two loosely coupled processes: the disks sat idling as they waited for the CPU to process data from network stack to memory. Network transfer was slowed again when data was flushed to disk. Host level monitoring showed that servers were close to stalling and could not process any other jobs even though the CPU should have been decoupled from I/O operation.

In summary, for Service Challenge 1 and 2, when network traffic was close to the bottleneck limitation, the contention among several applications significantly impacted the effective bandwidth utilization. The best effort IP packet switch network did not scale with the number of applications sharing the network. We observed that throughput was only 60~70% of bandwidth limitation and it continued to decrease when more applications were added into network contention.

Service Challenge 3

Service Challenge 3 includes: a throughput phase to test the data storage facility and underlying network, and a service phase to exercise the ATLAS offline software to generate, store, access, and distribute data in files with the real life sizes expected in ATLAS production after 2007. The goals for BNL were: 1) to provide reliable file transfer, data management and placement services while efficiently managing and utilizing our available resources, such as Network, dCache storage system, and tape storage, 2) to integrate BNL Tier 1 storage resources into the data flow generated by the ATLAS production system.

We used the existing BNL USATLAS production dCache system for the service challenge. USATLAS dCache/SRM supported the SRM 1.1 interface that was required by SC3. dCache servers involved in SC were one SRM door, four GridFTP doors, eight write pools. HPSS was as tape backend for dCache[5].

BNL SC3 disk throughput sustained a data transfer rate of 120 MB/second for a week, followed by the SC3 Tape data transfer which started on July 23 with approximately 80 MByte/second, as shown in **Figure 5**. The green line represents data coming from CERN, and the blue line represents data migration into HPSS at BNL. **Figure 6** shows that CERN stably transferred data to all Tier 1 centres at a rate of 1 GByte/second during a one day rerun, during which BNL contributed 90M Byte/second.

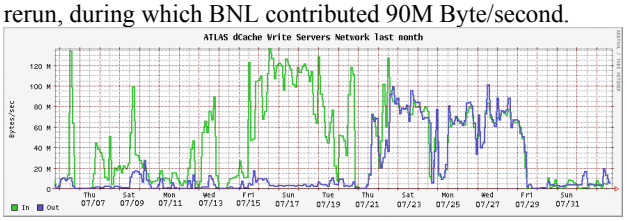

Figure 5: BNL Service Challenge 3 Data Transfer Plot for July 2005 as shown in Ganglia Monitoring System

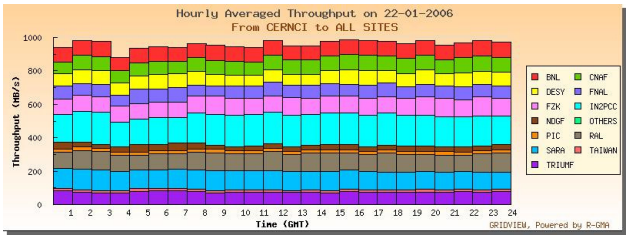

Figure 6: Service Challenge 3 Rerun

ATLAS ran a distributed production exercise that generated about 6 TBytes of raw, AOD and ESD per day - 10% of the expected production rate - and used the new version of the Distributed Data Management (DDM) [4] system (DQ2) to send to all ATLAS Tier 1 sites. **Figure 7** and **Figure 8** show the BNL participation in the service phase, demonstrating good peak data rates and reasonable aggregate data volumes in this exercise.

During SC3, we gathered significant experience in network tuning, dCache performance improvement and problem diagnosis. When everything was running smoothly, BNL got a very good data rate of above 100 MByte/second. The middleware (FTS) was stable but still needed improvement in terms of compatibility: we discovered that it frequently blocked the data transfer channel when communicating with dCache/SRM due to protocol mismatch. We needed to improve operation, i.e. proactive monitoring and timely problem reporting, to prevent further performance and functionality deterioration. We achieved the best performance among the dCache sites which participated in the ATLAS SC3 service phase. 15 TB of data was transferred to BNL. Sites using CASTOR SRM showed better performance.

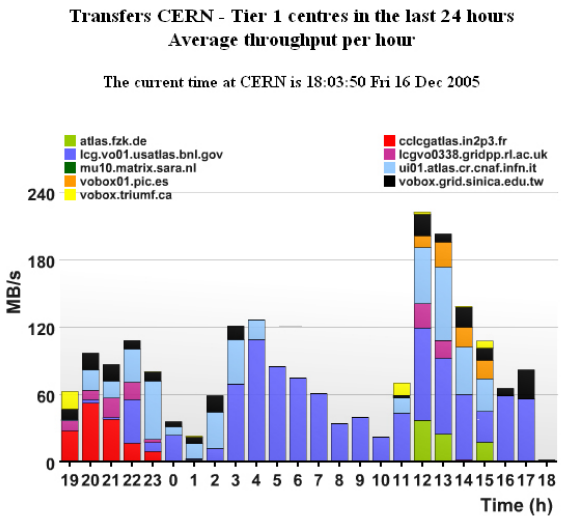

Figure 7: Data Transfer in the Last Day of SC3 Service Phase.

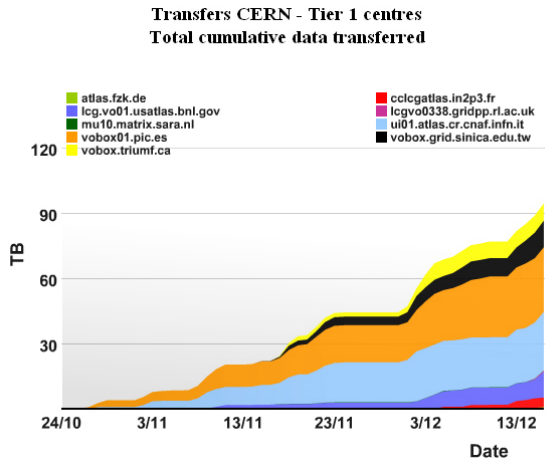

Figure 8: The Aggregated Data Volume During the SC3 service phase

CONCLUSION AND FUTURE WORK

The BNL to CCJ data transfer was the first time that such a magnitude data transfer was sustained over the period of an entire RHIC run. All raw proton-proton event data were delivered by Grid tools to collaborators a few hours after their generation, regardless of the collaborators' physical distances from BNL, providing equal opportunity for all participants. It provided valuable lessons for the following BNL LHC service challenges, which had higher data transfer rates and involved Tier 0 at CERN and three USATLAS Tier 2 sites. Future work will focus on applying this dCache/SC experience to large-scale RHIC data transfers and improving the stability and performance of data transfers as the BNL WAN and LAN backbone is upgraded to 2x10 Gbps bandwidth by the end of February 2006.

PHENIX RUN 2006 Data Transfer

In 2006, RHIC will again run polarized protons, beginning in early March. The total data volume is expected to be marginally higher than last year, ~300 TB, with potentially higher peak transfer rates. Once again, a copy of this data will be transferred to CCJ through an infrastructure similar to that described above. The introduction of SRM into the transfer process is expected to add additional robustness.

Service Challenge 4 Plan

The BNL data management infrastructure for SC4 will be connected to LHCOPN which provides a dedicated 10 Gbps network connection between BNL end systems and a CERN egress router. The throughput goal is to establish stable data transfer at a speed of 200 MBytes/second to BNL dCache and equivalent data migration speed from dCache to BNL HPSS. The dCache/SRM will be shared by LHC SC4 and USATLAS production. Its capacity for data writing and caching will be increased to 17 TeraBytes, which is 24 hours worth of storage at the planned SC4 data rate. The services used in SC3 will be upgraded and added to the BNL operator web page to ensure 24/7 monitoring and problem reporting. We will deploy and strengthen the necessary monitoring infrastructure based on ganglia, nagios, Monalisa, and LCG-RGMA.

ACKNOWLEDGEMENTS

We would like to thank Yasushi Watanabe, Takashi Ichihara, Satoshi Yokkaichi to ensure data received at CCJ. We would like to thank Jerome Lauret to generously loan us tape resource for service challenges. We would also like to thank Jamie Shiers, Miguel Branco, James Casey, David Cameron, Maarten Litmaath, Gavin Mccance, Kaushik De, Patrick McGuigan, Saul Youssef, Jim Shank, Rob Gardner, Andrew Zahn, Frederick Luehring for Tier 0 and Tier 2 coordinations.

REFERENCES

[1] M. Purschke, etc, “PHENIX experiement uses Grid to transfer 270TB of data to Japan” CERN Courier September 2005

[2] I. Bird, L. Robertson, and J.Shiers: Deploying the LHC Computing Grid-The LCG Service Challenge, https://uimon.cern.ch/twiki/pub/LCG/TalksAndDocuments/SC-IEEE.pdf.

[3] ATLAS Computing Group, ATLAS Computing TDR http://doc.cern.ch//archive/electronic/cern/preprints/lhcc/public/lhcc-2005-022.pdf

[4] M. Branco: Distributed Data Management (DDM): https://uimon.cern.ch/twiki/bin/view/Atlas/DDM

[5] B. Gibbard, Z. Liu, R. Popescu, O. Rind, etc, “Large scale, grid-enabled, distributed disk storage systems at the Brookhaven National Lab RHIC/ATLAS Computing Facility”, CHEP06, Mumbai, India, Feb. 2006.