Terapaths: A QoS Collaborative Data Sharing Infrastructure for Petascale Computing Research -II

DWMI: Datagrid Wide Area Monitoring Infrastructure

PI: Les Cottrell, SLAC

Goals

Todays's data intensive sciences, such as High Energy Physics (HEP), need to share large amounts of data at high speeds. This in turn requires high-performance, reliable end-to-end network paths between the major collaborating sites. In addition end-users need long and short-term forecasting for application and network performance for planning, setting expectations and trouble-shooting. To enable this requires a network monitoring infrastructure between the major sites.

The main goal of the DWMI project is to build, deploy and effectively learn how to use an initially relatively small but rich, robust, sustainable, manageable network monitoring infrastructure focused on the needs of critical HEP experiments such as Atlas, BaBar and CMS. A characteristic of these experiments is a hierarchical tiering of sites. The major data sources (accelerator sites such as CERN, FNAL or SLAC) are tier 0, tier 1 sites are major data re-distribution centers for a region (e.g. a major HEP data center in each of France, Italy, Germany, the UK and US etc.), tier 2 are major collaborator sites (typically major university sites such as Caltech), tier 3 are smaller collaborators etc. The idea is that the raw experimental data is replicated from the tier 0 to tier 1 sites, where it is analyzed and made available to higher tiered sites. To match this architecture, DWMI needs to be deployed at tier 0, tier 1, and a few tier 2 sites. The measurements at each of these sites will then be configured to provide regular end-to-end network performance measurements and analysis to its collaborator sites.

The sub-goals of the DWMI project are:

- Make contact with and work with HEP tier 0, 1 and 2 sites to deploy, configure, exercise, use and evaluate the Internet End-to-end Performance Monitoring BandWidth (IEPM-BW) toolkit/infrastructure.
- Evaluate, recommend and integrate network measurement tools (probes) and determine their applicability. In particular:
  - Evaluate the challenges and effectiveness of making network measurements for Quality of Service (QoS) enhanced paths including paths that need reservations.
  - Explore tools that will work for future higher speed (e.g. > 1 Gbits/s) networks and dedicated network paths.
- Develop and integrate techniques for providing network performance forecasts from the network measurements.
- Develop and integrate techniques for automatic detection of significant, persistent changes in network performance events.
- Develop and integrate effective techniques for generating and managing alerts from events, including gathering and providing extra information concerning the events(e.g. tracroutes, host parameters etc.)
- Provide access to the results for researchers, and provide standard web services access to IEPM-BW data for applications such Grid middleware replica selection.
- Integrate IEPM_BW features (e.g. vizualization, analysis, event detection) with other infrastructures such as MonALISA, AMP etc.

Activities

- We have successfully installed the infrastructure at BNL, CERN, Caltech, FNAL, SLAC and Pakistan, the latter to see how to use the infrastructure for lower performance grid wannabe sites. Other sites in consideration are: UMich, FZK, DESY.
- We have developed and put into production Management tools for automation and robustness, including:
  - Installation and update kits;
  - Measurement and reporting of unreachable particpating hosts
  - Documentation, including a Program Logic Manual
  - Database of site, host, location, contact, OS, cpu, test parameters ...
  - Analysis of logs to detect anomalies, malingering tasks
  - Utilities for adding/updating a host, probe etc...
- We have evaluated the optimum measurement tools/probes (optimized for traffic, accuracy, coverage of metric space) for active end-to-end monitoring. Based on this, we now support: ping, traceroute, pathload, pathchirp, abwe/abing, bbcp, bbftp, GridFTP, iperf, and thrulay. We have also evaluated pipechar. In production, we we use a selection of probes based on the quality of the path being measured, for example for high-performance critical paths we use heavyweight tools such as iperf and thrulay to measure achievable throughput and bbftp for file transfer, for fragile paths we may only use ping, traceroute and possibly a lightweight packet pair bandwidth estimation technique.

  We have studied and reported on limitations using current active end-to-end measurements in future high-speed networks. As a result of this we are exploring the effectiveness of using passive (e.g. Netflow) tools to augment or even replace some of the active measurements.

- We have developed an effective traceroute visualization toolkit to enable one to look at multiple traceroutes simultaneously and drill down to more detailed information. We are working with AMP to integrate this with their project.
- For event detection we have developed, published and integrated a step change detection algorithm. It has been successfully applied to several metrics with different measurement repetition frequencies, including RTT, available bandwidth and achievable throughput. It is now in regular use to generate email alerts for network adminstrators. In the last month, since we turned on the email alerting, we have had 4 detected significant, persistent alerts that we have carefully studied and reported on. Of these:
  - One was caused by a fan failure in a DWDM multiplexer between Stanford and the CENIC PoP in Sunnyvale;
  - A second was caused by loss of fibre connectivity to BNL;
  - A third by a denial of service attack that impacted the firewall performance of a site in the UK;
  - The fourth is under investigation.

  All of thse events were automatically reported within a few hours of the onset, as opposed to in the past where at least one event went un-noticed for several weeks and caused a drop by a factor of 5 in performance.

  Given the success and experience of these alerts, we are working on developing tools to gather more information to report to the network administrator. For the future we are also developing techniques (using the Kolmogrov-Smirnov technique) to enable finding the end of an event (i.e. when the network performance recovers. We are also evaluating other event detectors including the use of neural networks and Principal Component Analysis (PCA) to enable simultaneouly evaluating multiple metrics and paths.

- We have developed and are now integrating a long-term forecasting technique that takes into account seasonal (e.g. diurnal and weekly) variations. As part of the integration we will also make the forecasting tool more general purpose so it can be applied against data from other monitoring infrastructures.
- In preparation for evaluating QoS at BNL we worked with ESnet to evaluate the impact and use of the ESnet , see OSCARS Results. Our next steps will be to set up the measurements for for the QoS project at BNL.

Impact to specific DoE Science applications

Improved network understand and expectations together with more quickly discovering and reporting network problems is critical to all network based applications. The DWMI project's deployment of the IEPM-BW infrastructure focused on the needs of the DoE supported LHC, BaBar, CDF and D0 HEP experiments provides an evolving and practical basis for improved networking.

Synergy developed with DoE application developers to facilitate technology transfers

We are collaborating with groups at CERN, BNL, FNAL and Caltech to install, configure and put into use the IEPM-BW measurement toolkits. We have set up a network of contacts at the monitoring and monitored IEPM-BW sites. When we receive alerts and deem them of interest, we communicate with our contacts at the relevant sites to alert them to the problem and to better understand it.

We have made contact with the Open Science Grid (OSG) community's Wilko Kroeger to explore how to assist them with their network monitoring needs.

We have and will continue to work with the ESnet OSCARS project to assist in monitoring the effectiveness of QoS, and to help specify the requirements for monitoring (e.g. to provide persistent requests, and a program to program API to the scheduler. We are also working closely with Dantong Yu and the BNL Terapaths project to provide monitoring and support for the QoS services.

IN addition to working with DoE develkopers, are in regular contact with developers of monitoring infrastructures and tools funded by other agencies. In particular we are working closely with Internet2 to evaluate and improve thrulay and more closely integrate perfSONAR, and with the NLANR AMP developers to integrate the traceroute analysis and visualization. We are working closely with Iosif Legrand and others at the Caltech HEP group and CERN to integrate the IEPM and PingER measurements into MonALISA. We are also evaluating whether to include PingER and/or IEPM-BW as part of the Virtual Development Toolkit.