GA–A26397

NETWORK QUALITY OF SERVICE FOR MAGNETIC FUSION RESEARCH

FINAL REPORT TO THE
U.S. DEPARTMENT OF ENERGY
for the period
SEPTEMBER 15, 2004 through SEPTEMBER 14, 2007

by
M. QIAN and D.P. SCHISSEL

DATE PUBLISHED: APRIL 2009

GENERAL ATOMICS

DISCLAIMER

GA–A26397

NETWORK QUALITY OF SERVICE FOR MAGNETIC FUSION RESEARCH

FINAL REPORT TO THE
U.S. DEPARTMENT OF ENERGY
for the period
SEPTEMBER 15, 2004 through SEPTEMBER 14, 2007

by
M. QIAN and D.P. SCHISSEL

Work supported by the
Department of Energy under
DE-FG02-04ER25614

GENERAL ATOMICS PROJECT 30227
DATE PUBLISHED: APRIL 2009

GENERAL ATOMICS

1. INTRODUCTION

Complex and sophisticated computational support is becoming a more important part of many different scientific research fields. As magnetic fusion research continues to grow in size and complexity, the pressure for more computational power continues to intensify. Recent advancements in CPU and storage technologies have been great, resulting in an increased need for network improvements. Even though there have been important upgrades in bandwidth, the large scientific data movements along with large amount of real time audio and video data are able to stress present scientific networks. This problem is unlike some contemporary cases where increasing the number of resources can be an easy solution. Because of the infrastructure nature of the network, growth can be limited at times and at different locations.

In this project, we want to explore the existence of data prioritization during magnetic fusion experimental operation and use congestion-management and congestion-avoidance techniques to provide more predictable performance and more effective utilization of the limited bandwidth. The Quality of Service (QoS) implementation is based on the Differentiated Services (Diff-Serv) architecture, an emerging standard from the Internet Engineering Task Force (IETF). Specifically, the two-year project sought to deploy, to an operating magnetic fusion experiment (DIII-D), a robust and reliable network QoS capability for routine usage during experimental operations to perform quasi real-time data analysis on computer resources not located on the Local Area Network (LAN). Such a capability could lead to, for example, the usage of a massive supercomputer such as at Seaborg at NERSC or the National Leadership Computing Facility at ORNL to support quasi real-time analysis of fusion experimental data; something that today is not even considered within the experimental fusion community.

During experimental operation of magnetic fusion devices, the science is performed in realtime and pseudo real-time with a very high cost. Therefore, any computational work, data generation, and data movement that supports experimental operation needs to have a high priority. The ability to rapidly provide detailed data analysis including the ability to compare simulations with experimental data can result in higher quality experiments and therefore is highly valued.

For the present QoS study, data movement was provided by MDSplus and computational data analysis by EFIT and TRANSP. MDSplus, developed jointly by MIT, LANL, and the IGI in Padua, Italy, is by far the most widely used data system in the international fusion program. Based on a client/server model, MDSplus provides a hierarchical, self-descriptive structure for simple and complex data types and is currently installed and used in a variety of ways by about 30 experiments, spread over 4 continents. The result is a *de facto* standard that greatly facilitates data sharing and collaborations across institutions. EFIT is used to

calculate the external and internal magnetic field topology of fusion plasmas. The code TRANSP is used for time dependent analysis and simulation of tokamak plasmas and is used worldwide including as a computational service on FusionGrid.

The aim of this project is to test the concept of marking certain data packets that emanate from MDSplus marked as high priority so that they receive QoS network treatment rather than the standard best effort routing. A successful proof-of-concept will allow new MDSplus clients to be written that will allow scientists with the proper authority to mark certain data streams as high priority.

This project had a close working relationship with the On-Demand Secure Circuits and Advance Reservation System (OSCARS) Project and the Network Project at NERSC where the computer science research for on-demand provisioning of guaranteed bandwidth over ESnet is being performed.

2. EXAMINING POTENTIAL NETWORK PERFORMANCE INCREASES

A computer dedicated to network testing was installed on a private VLAN directly attached to the DIII-D core Cisco 6500 that is directly connected to the Juniper M7i ESnet border router. The DIII-D Cisco 6500 is composed of a Cisco 6509 Catalyst switch with a Firewall Services Module (FWSM) blade running in transparent mode, a Multilayer Switch Feature Card (MSFC2) performing routing functions, a 48 port 10/100/1000 Ethernet switch, a supervisor card (Sup2) and the Switch Fabric Module (SFM). A computer for network testing was also been identified and configured at NERSC.

A test Label Switched Path (LSP) was then established between DIII-D and NERSC for initial QoS testing. Only traffic from the specific test computers (source/destination IP addresses) was injected into the LSP. Since the LSP was using the production DIII-D network, tests were of short duration and performed where practical later at night.

The initial QoS testing concept was as follows. Two identical network streams were initiated through the LSP tunnel with one being tagged for Best Effort (BE) and the other being tagged for Expedited Forwarding (EF). The EF traffic runs from the network test machine to NERSC while the BE traffic runs from different computer that is in the DIII-D production computing cluster. When run separately the two streams obtained approximately the same network throughput as is expected. When run together, the EF stream should have priority, and since the two streams together occupy more than the available bandwidth, the BE stream should see a reduction in network throughput. Utilizing the network utility Iperf these tests were run for both UDP (Table 1) and TCP (Table 2) traffic. As can be seen, the results are as expected. Figure 1 illustrates the temporal history of the TCP streams. Here, the BE stream was started first and shows the variety of throughput rates as is expected on a production network from a production computer. When the EF stream was initiated, the BE rates drop some but more importantly, the EF stream retains a very high constant value. Thus, the ability to tag network traffic as high-priority combined with the ability to give expedited forwarding between DIII-D and NERSC was verified.

With the success of the initial LSP testing with Iperf, the next step was to attempt to demonstrate the same QoS behavior with fusion data being transmitted via the MDSplus data management and acquisition system. MDSplus is by far the most widely used data system in the international fusion program. Based on a client/server model, MDSplus provides a hierarchical, self-descriptive structure for simple and complex data types, and is currently installed and used in a variety of ways by about 30 experiments, spread over 4 continents. MDSplus is also used to securely transfer data on FusionGrid. Thus, the demonstration of QoS via MDSplus will indicate that any fusion data transmitted via MDSplus should be able to be given priority. With the ubiquity of MDSplus in the fusion community, this becomes a very powerful capability.

Table 1
Utilizing UDP, a Comparison of Best Effort versus Expedited Forwarding for Two Single Streams and Two Competing Streams[(a)]

	Best Effort (Mb/s)	Expedited Forwarding (Mb/s)
Single Stream	98.4	95.3
Competing Streams	48.5	95.4

[(a)] As expected, when the two streams compete, the EF stream is given priority with the resulting decrease in BE traffic due to overall network bandwidth limitations.

Table 2
Utilizing TCP, a Comparison of Best Effort versus Expedited Forwarding for Two Single Streams and Two Competing Streams(a)

	Best Effort (Mb/s)	Expedited Forwarding (Mb/s)
Single Stream	47.1	76.8
Competing Streams	31.2	76.5

[(a)] As expected, when the two streams compete, the EF stream is given priority with the resulting decrease in BE traffic due to overall network bandwidth limitations.

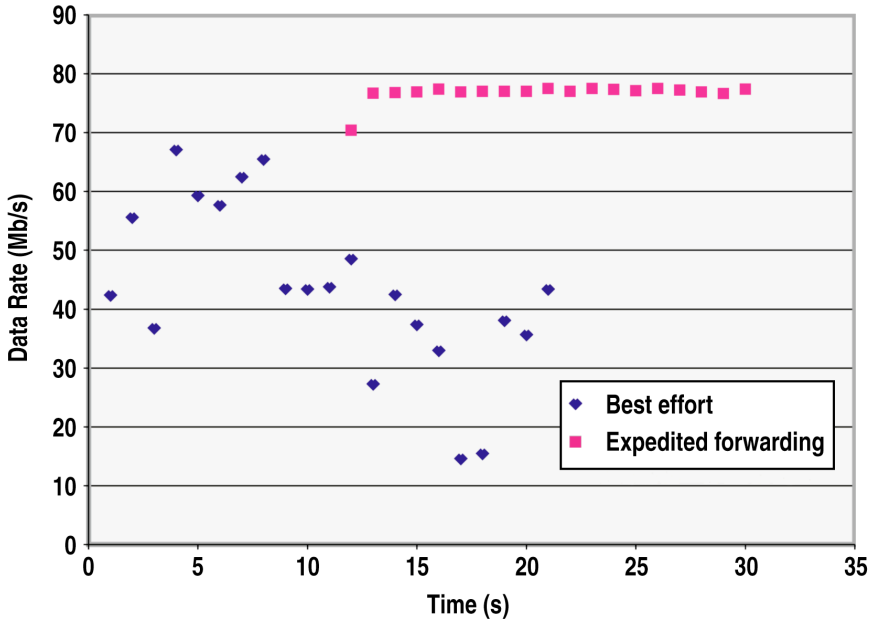

Fig. 1. Utilizing TCP, a graphical comparison of Best Effort (BE) versus Expedited Forwarding (EF). The first 12 s shows the BE traffic with considerable variation due to changes in usage of the production network. At 13 s into the test, the EF traffic stream is turned on and when it reaches its final value at 14 s remains at the 78 Mb/s level for the remainder of the test. The constant rate of the EF traffic combined with the reduction in the BE traffic indicates that the network QoS system was working.

A DIII-D C-based utility for measuring MDSplus data throughput was used for the LSP testing. The client code makes multiple serial requests for a 10 MB size of piece data from the MDSplus server and the average time is used to calculate network throughput. The DIII-D test network machine was configured as an MDSplus server and the NERSC network test machine was configured as an MDSplus client. The test was run where the MDSplus testing utility was run on the NERSC machine to fetch data from the DIII-D MDSplus server. When the traffic out of DIII-D's ESnet border router was small (~10 Mb/s), an average network throughput of 23 Mb/s was observed. Utilizing a 130 Mb/s UDP blast from DIII-D to NERSC resulted in the MDSplus throughput dropping by an order of magnitude (3.2 Mb/s). When this same MDSplus traffic was given priority (Expedited Forwarding), the network throughput climbed back up to the maximum value of 23 Mb/s. Thus MDSplus data transfer with network QoS was demonstrated.

3. TESTING THE QoS RESULTS WITH THE DIII-D COMPUTATIONAL CODE EFIT

The purpose of this test was to demonstrate that a fusion code could be run off-site utilizing network QoS and therefore be able to support the pseudo-real-time needs of an operating tokamak. This test builds on previous ones where the project demonstrated the ability to transmit fusion data over the WAN using QoS utilizing the MDSplus data management system. The fusion code EFIT, that calculates the shape of the fusion plasma, was installed at NERSC along with an MDSplus client. A test Label Switched Path (LSP) was established between the NERSC EFIT machine and the main production DIII-D MDSplus repository by the OSCARS project. This path utilized the 155 Mb/s OC-3 network between GA and NERSC that is no longer used for production traffic.

The running of the EFIT code was triggered from DIII-D as it would be during normal tokamak operations. The code read the small quantity of input data from the main DIII-D MDSplus repository and calculated the plasma shape as a function of time. Output of the EFIT code, along with other data to increase the overall transmission time, was then written back to the MDSplus repository at DIII-D. Since the OC-3 network was not handling any production traffic, we were able to completely congest the network to study the beneficial effect of the QoS environment.

Test results indicate that QoS using the LSP tunnel was able to restore network transmission rates back to pre-congestion levels. With a single data stream and no TCP send and receive window tuning, an application-level baseline of 4 MB/s was achieved. Congestion was added to the network to reduce this number down to 1.9 MB/s. Utilizing QoS, network throughput was restored even with congestion back up to the 3.9 MB/s level (Fig. 2). These Byte numbers at the application level although they appear low, are not unreasonable for a single non-tuned stream over a high latency WAN.

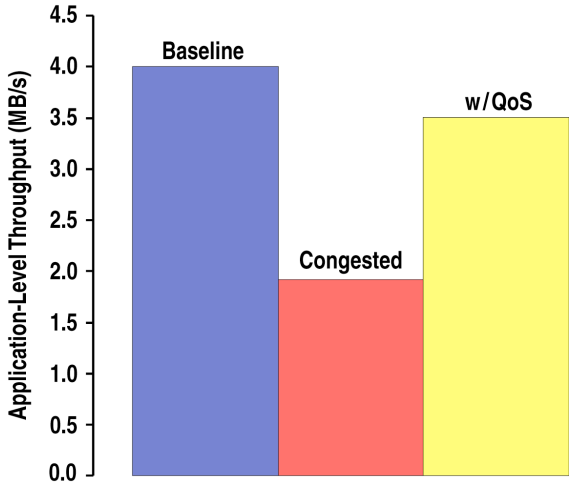

Fig. 2. Application-level throughput increased to nearly pre-congestion levels with QoS was enabled.

4. MODIFICATION OF MDSPLUS

The MDSplus data acquisition and management system has become the standard for data acquisition and management in experimental fusion research. Overtime, it has become an important data structure used in numerous computational analyses codes. Extending MDSplus with QoS capability is addressing an infrastructure capacity of the underlying data structure. Therefore, the successful modification of MDSplus that adds a QoS capability can have an impact well beyond the DIII-D National Fusion Facility.

A version of MDSplus was modified for usage in a network QoS environment. Specifically, the section of MDSplus code that communicates with the transport layer was modified to change the value of the 6-bit Differential Services Code Point (DSCP) field. An arbitrary, but not too high value, of "000110" was chosen to mark data packets emanating from MDSplus as high-priority. This version of MDSplus can only transmit data with the DSCP field set to "000110" and was created for testing purposes only.

The QoS-enabled MDSplus was successfully demonstrated between the DIII-D National Fusion Facility in San Diego, CA and the MIT Plasma Science and Fusion Center in Cambridge, MA. In this test, a QoS-MDSplus server was launched on a Dell PowerEdge 2850 2u server at DIII-D, and a local Linux node at MIT acted as the client. The testing commenced with the MIT client requesting data and the DIII-D MDSplus server responding by sending QoS marked data packets. MPLS tunnel was established between these to nodes by the OSCARS project that would only accept network packets with the DSCP field set to the predetermined value. These packets left the DIII-D MDSplus server with their DSCP field set, transitioned through the local DIII-D router to DIII-D's ESnet boarder router where they entered the MPLS tunnel. Once in the tunnel the data successfully made it over ESnet to the MIT LAN and finally to the local client with the DSCP still correctly set.

5. REAL-TIME NETWORK TESTING OF MDSPLUS WITH QoS IMPLEMENTATION

With the MDSplus modifications accomplished, additional tests were designed and launched to study the detailed network traffic characteristics between DIII-D and MIT. On the PowerEdge machine, two different MDSplus servers were run simultaneously. One was the modified QoS version while the other was a non-modified version. A local Linux node at MIT acted as the client and the same MPLS tunnel was established by the OASCRS project. The test commenced with the MIT client making two different data request from the MDSplus servers at DIII-D. One request went to the non-modified server and the other to the modified server that would send data packets marked for QoS delivery. The two servers responded as expected by sending regular data and QoS marked data packets with only the QoS marked packets running on the OASCRS network circuit.

The surprising result was that with varying network traffic conditions existing between DIII-D and MIT, there was no significant difference in data transfer performance between the two types of data packets. This was in contradiction with previous results between DIII-D and NERSC. At first, the condition was believed to be associated with some of the Cisco routers that were being traversed over ESnet. Later when replacement routers were updated, the network performances of these two tests were still very similar. This result holds under both condition of normal network traffic and when the network was artificially heavily congested.

6. CONTROLLED LAN TEST

During fusion experimental operation, raw data is acquired, analyzed and then stored in data repositories. The data acquisition and management process at the DIII-D National Fusion Facility involves a great variety of systems. A test network that reflects the production network mixture was designed in order to evaluate the QoS LAN network portability and functionality. This controlled, separated local network was set up at GA and allowed testing that will not interfere with normal DIII-D tokamak operation. The testing LAN consists of: the Dell PowerEdge 2850 2u server, a Linux laptop, a Macintosh and a Windows computer for local nodes, all connected through the Cisco Catalyst 3560G switch.

Significant knowledge and valuable experience was gained during installation and testing of this heterogenic system. Software related topics including but not limited to installation of MDSplus software, server environment interactions and configurations, varies local machine configurations, settings for the Cisco switch and varies other useful software tools.

Testing on this network showed some increase in data transfer performance. At the switch end, a variety of configuration experiments were carried out. Different classification options of the data packets were tried based on IP or Layer 2 MAC ACLs to class definition of groups of packets. Policy was attached to each traffic class in the case of choosing layer 2 classification option. Also extensively studied was the different treatments based on queue and schedule in varies situations where resource contention existed. Different configuration of resource usage limits within reasonable ranges produced up to 12% increase in network performance. This study allowed us to evaluate the QoS capability before the possible costly extension to the production GA LAN that serves the DIII-D Tokamak.

7. NETWORK QOS TESTING USING TRANSP DATA

The TRANSP computational service is a time dependent tokamak transport analysis code developed at the Princeton Plasma Physics Laboratory. The code provides a means to combine data from tokamak experiments yielding a picture of the processes that accounts for the energy transport and heating in the tokamak plasmas. With the increase of computational power, scientists are looking into even more timely and detailed feedback during experiments. The data intensity and time dependent characteristics make TRANSP an attractive and challenging candidate for deployment of the QoS technology.

A prototype web-client was created for the scientists to set up their between-pulse TRANSP computation. This setup can be expanded as required to address further computational controls that the scientists require. Obtaining the require TRANSP input data was automated and interfaced to the existing Data Acquisition and Automated Analysis System for the appropriate MDSplus dispatching phase signals. This work was made easier by utilizing the current input data creation process through the AUTOTRANSPPRE code to generate ufiles in flat file system. This between-pulse process was also integrated to the existing TRANSP FusionGrid service.

Timing for a TRANSP run, without any user input from the web front-end, starts with preparing input data, packaging the prepared data at DIII-D, sending the data and request for a run TRANSP to PPPL, and getting the service complete notice from PPPL. Towards an end of a regular tokamak operation, the total time for an automated simple TRANSP run can takes approximately 3000 s. This time breaks down to about 150 s to the point where the full request reaches PPPL, 2400 s of CPU-time, and an additional average 450 s of wall-time clock time. The CPU time at PPPL may vary greatly due to the quality of the data from the experiments, but the overhead time posted for the wall time both at PPPL and DIII-D are fairly consistent. Taking into account that the starting point of this process is after certain MDSplus dispatching phase, the completion of between-pulse TRANSP might run into the following pulse since the time between pulses is approximately 1200 s. Nevertheless, providing feedback even several pulses behind can still provide valuable information for the DIII-D experiments. Also another concurrent project is aiming in reducing the time it takes at the PPPL end so it can be a true between-pulse TRANSP.

The back end storage method can be easily changed into loading the prepared ufiles into MDSplus tree structure and therefore link to another concurrent project which is designed to change the method of TRANSP service requests in order to reduce the total time needed. When the integration of these two projects is completed, extending TRANSP's QoS capability can be translated into MDSplus's QoS functionality.

In analyzing and designing the between-pulse TRANSP system, the understanding had gone beyond a single project and address many aspects of middleware that interfaces

experimental data with large analysis codes. In utilizing and fitting different existing services and projects together, the end product is something more coherent and can have not only computational impact but also more structural frame reference for broader computational services for variety of fusion research activities.

8. SUMMARY

Computational analysis in fusion research places an increasing demand on the network infrastructure. In this project, both LAN and WAN QoS technology was examined in the fusion experimental setting. Vertically, we went through how to design and implement the QoS functionality from user interface to application level, to different system layers, to the transport layer. Horizontally, the project covered from per-hop design and implementation to path behavior testing, from a single application to multi-services integration, from one functional point to the data infrastructure/process issues. We observed different performance increases due to different data traffic priorities, QoS services and network bandwidth allocations under stressed and normal network traffic condition. We also acquired timing information during a normal tokamak operation for running an automated between-pulse TRANSP service.

Under some cases, the QoS technology worked as expected. In other cases, it did not. Clearly, as the expanse of the desired usage increases across the WAN there is the potential for older hardware or other reasons why the technology will not perform as expected. Since the cost of running fusion experimental facilities is so high, and therefore the cost per plasma pulse is very high, the tolerance for failed systems is rather low. Therefore, in the area of QoS, it seems appropriate to start small, even in the LAN setting, and growing the size of the installation from that point onwards. Unfortunately, the largest future benefit of QoS technology to fusion science will be addressing the WAN issue as the amount of increasing international collaboration during machine operations is only increasing. This will be especially true as the two new Asian tokamaks, KSTAR in South Korea and EAST in China, come into operation and the U.S. is expected to play a large role in assisting their startup.

This QoS project provided an excellent test bed for how such prioritization technology might be used in magnetic fusion experimental research. The knowledge and lessons learned will help direct further research in creating a production service including a potential US contribution to ITER.

ACKNOWLEDGEMENT

This work supported by the U.S. Department of Energy under DE-FG02-04ER25614.