TeraPaths: A QoS Collaborative Data Sharing Infrastructure for Petascale Computing Research

Bruce Gibbard & Dantong Yu

High-Performance Network Research PI Meeting

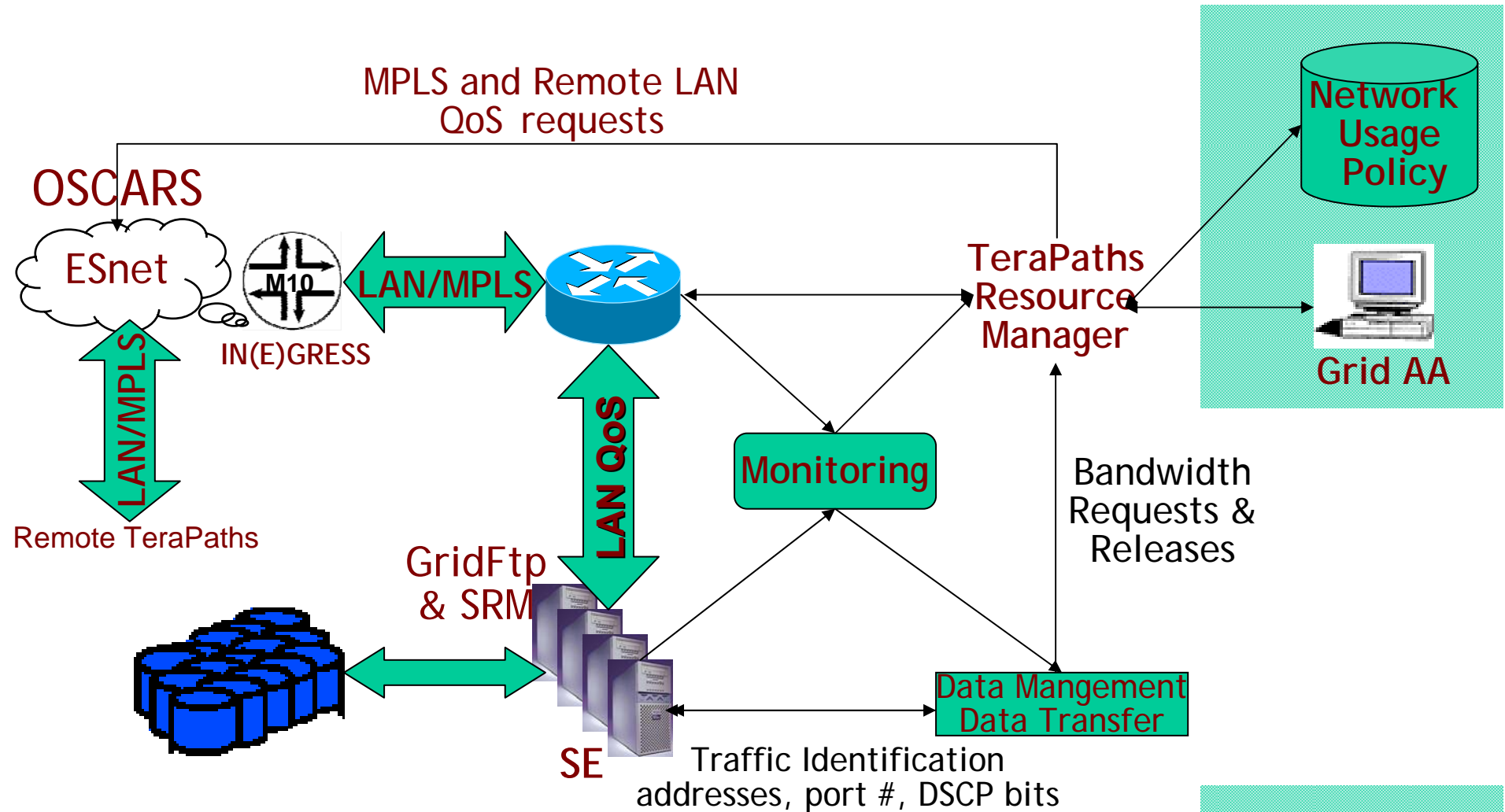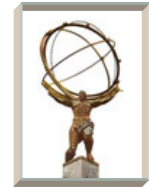**September 28-30, 2005**

Brookhaven National Laboratory

BROOKHAVEN
NATIONAL LABORATORY

Outline

❆ **Introduction to TeraPaths.**

❆ **Proposed Primitive Infrastructure.**

❆ **Last Year's Accomplishments:**

    ❑ End to End QoS Test, Deployment and Verification.

    ❑ Software Development Activities.

❆ **Current Status Summary.**

❆ **Future Works: Near Term Plan and Long Term Plan.**

What Is TeraPaths?

- This Project Will Investigate the Integration and Use of LAN QoS and MPLS Based Differentiated Network Services in the ATLAS Data Intensive Distributed Computing Environment As a Way to Manage the Network As a Critical Resource.

- The Collaboration Includes BNL and University of Michigan, and Other Collaborators From OSCAR (ESNET), Lambda Station (FNAL), and DWMI (SLAC).

Proposed Infrastructure

MPLS and Remote LAN
QoS requests

OSCARS

ESnet

M10

IN(E)GRESS

LAN/MPLS

LAN/MPLS

Remote TeraPaths

LAN QoS

GridFtp
& SRM

SE

Monitoring

TeraPaths
Resource
Manager

Network
Usage
Policy

Grid AA

Bandwidth
Requests &
Releases

Data Mangement
Data Transfer

Traffic Identification
addresses, port #, DSCP bits

Second/Third year

Progress

❋ Enabled LAN QoS Inside BNL Campus Network Along the Path From Gridftp Servers and SRM Into the Border Routers.

❋ Tested and Verified MPLS Paths Between BNL and LBL, SLAC (Network Monitoring Project), FNAL. MPLS/QoS Path Between BNL and UM for Super Computing 2005 Is Under Construction.

❋ Set up Network Monitoring Tools Provided by DWMI, Install Other Monitoring Tools When Necessary.

❋ Integrated the LAN QoS With MPLS Paths Offered by OSCAR.
  ❑ Verified LAN QoS Capability by Injecting Regular Traffic and Prioritized Traffic in LAN. (BNL ATLAS and ITD)
  ❑ Verified WAN MPLS/LAN QoS Bandwidth Provisioning (BNL ATLAS and UM)
  ❑ Did Integration Test and Documented Experience. (BNL ATLAS and UM)

❋ Verified the Impact Between Prioritized Traffic and Best Effort Traffic. Learned the Effectiveness and Efficiency of MPLS/LAN QoS and Its Impact to Overall Network Performance.
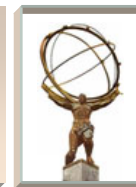
Prioritized Traffic v.s. Best Effort Traffic

Network QoS with Three Classes: Best Effort, Class 4 and EF

Legend: Best Effort — Class 4 — Express Forwarding — TOTAL — Wire S
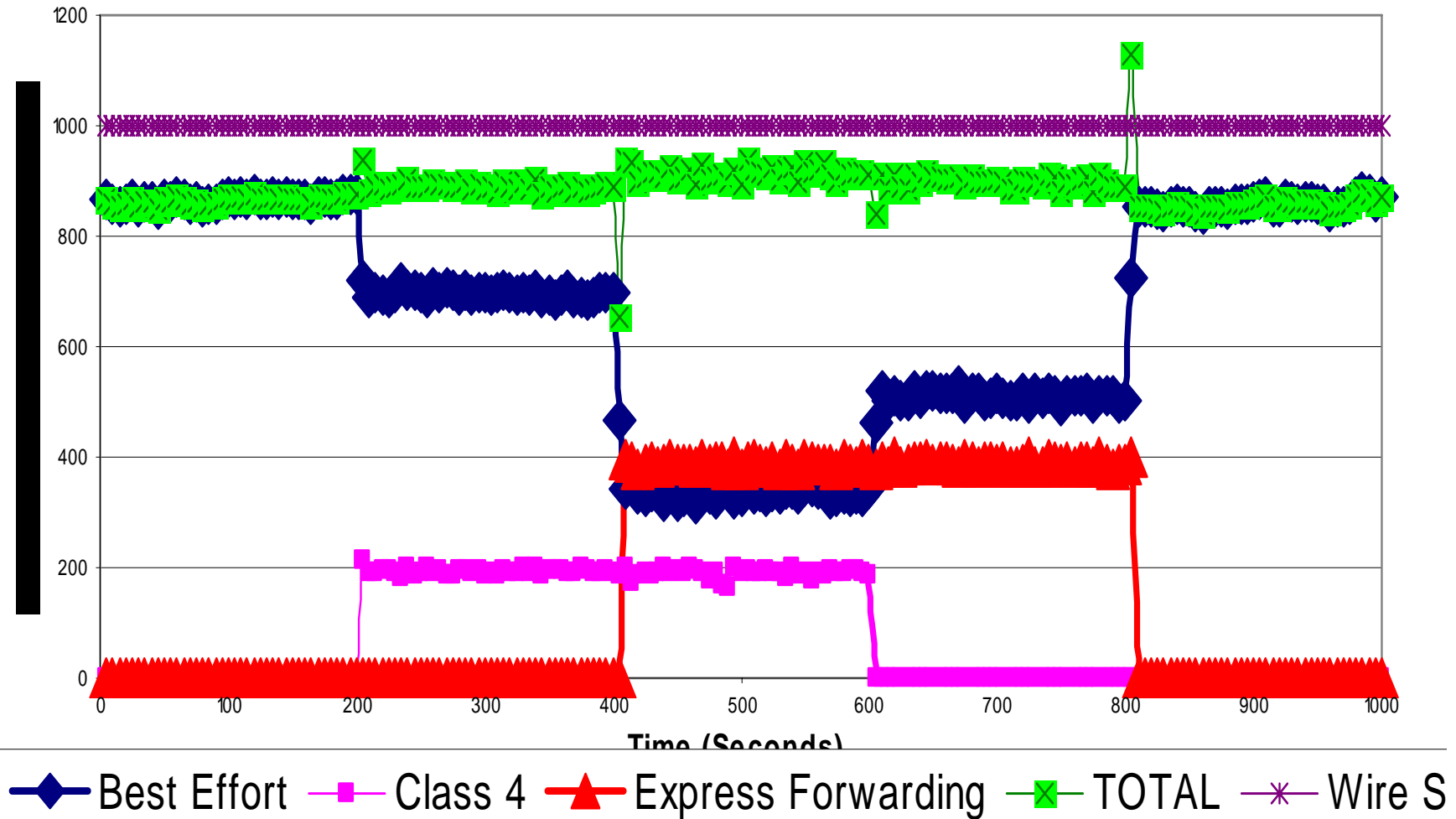
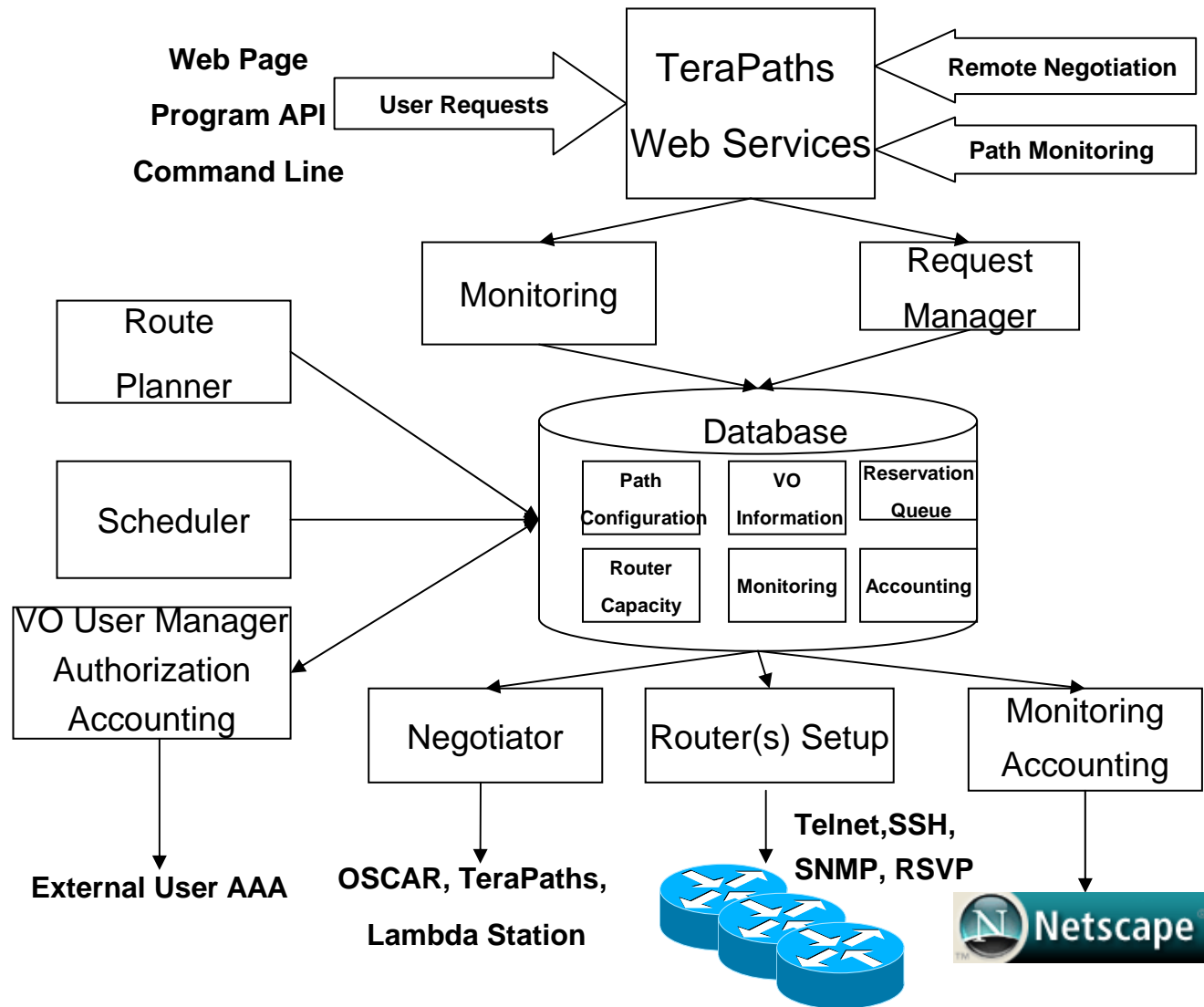# Software Development: Automate the MPLS/LAN QoS Setup

❋ **Implement Automatic/Manual QoS Request System**
  - ❑ Web Interface for Manually Inputting QoS Requests.
  - ❑ Access the Capability From a Program (C, C++, Java, Perl, Shell Scripts, Etc.).
  - ❑ Work With a Variety of Networking Components.
  - ❑ Work With the Network Providers (WAN, Remote LAN).
  - ❑ Provide Access Control and Accounting.
  - ❑ Provide Monitoring of System.

❋ **Design Goals: To Permit the Reservation of End-to-End Network Resources in Order to Assure a Specified "Quality of Service"**
  - ❑ User Can Request Minimum Bandwidth, Start Time, and Duration
  - ❑ User Will Either Receive What They Requested or a "Counter Offer"
  - ❑ Must Work End-to-end With One User Request

TeraPaths System Architecture

Web Page

Program API

Command Line

User Requests →

TeraPaths Web Services

← Remote Negotiation

← Path Monitoring

Monitoring

Request Manager

Route Planner

Scheduler

VO User Manager
Authorization
Accounting

Database

| Path Configuration | VO Information | Reservation Queue |
| Router Capacity | Monitoring | Accounting |

Negotiator

Router(s) Setup

Monitoring Accounting

External User AAA

OSCAR, TeraPaths, Lambda Station

Telnet,SSH, SNMP, RSVP

Netscape®

Typical Usage

❋ User (or User Agent) Requests a Specified Bandwidth Between 2 Endpoints (IP & Port) for a Specified Time and Duration.

❋ Functionalities:

- ❑ Determines Path & Negotiates With Routers Along Path for Agreeable Criteria
- ❑ Reports Back to User

❋ User Accepts or Rejects (Time Out)

❋ The System Commits the Reservation End-to-end

❋ The User Indicates the Start of the Connection & Proceeds.

Software Solution

❄ **The Software Components Are Being Realized As "Web Services"**

- ❑ Each Network Hardware Component (Router/switch) in the Path Will Have a Web Server Responsible for Maintaining Its Schedule and for Programming the Component As Needed.
  - ⌘ Example – at BNL, a Web Service Makes a Telnet Connection to a Cisco Router and Programs the QoS Acls.
  - ⌘ Others May Use SNMP or Other Means
- ❑ For Other Network Provisioning Services That Have Their Own Means of Doing This, a Web Service Will Provide a Front End.

❄ **Web Services Benefits**

- ❑ The Services Are Typically Written in Java and Are Completely Portable.
- ❑ Accessible From Web or Program
- ❑ Provide an "Interface" As Well As an Implementation. Sites With Different Equipment Can Plug & Play by Programming to the Interface.
- ❑ Highly Compatible With Grid Services, Straightforward to Port It Into Grid Services and Web Services Resource Framework (WSRF in GT4).

The Web Services Provided

❆Reservation Reserve(reservation)

- ❑ The Argument Is the Requested Reservation, the Return Value Is the Counter Offer.

❆Boolean Commit(reservation)

- ❑ Can Commit to the Reservation or Let It Time Out.

❆Boolean Start(reservation)

- ❑ At Start Time, the User Must "Pick up the Ticket". This Is When the Routers Are Actually Reconfigured

❆Boolean Finish(reservation)

- ❑ To Be Polite & Possibly Release the Resources Early

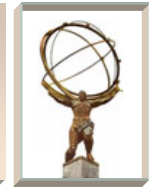❆Reservation[] Getreservations(time, Time)

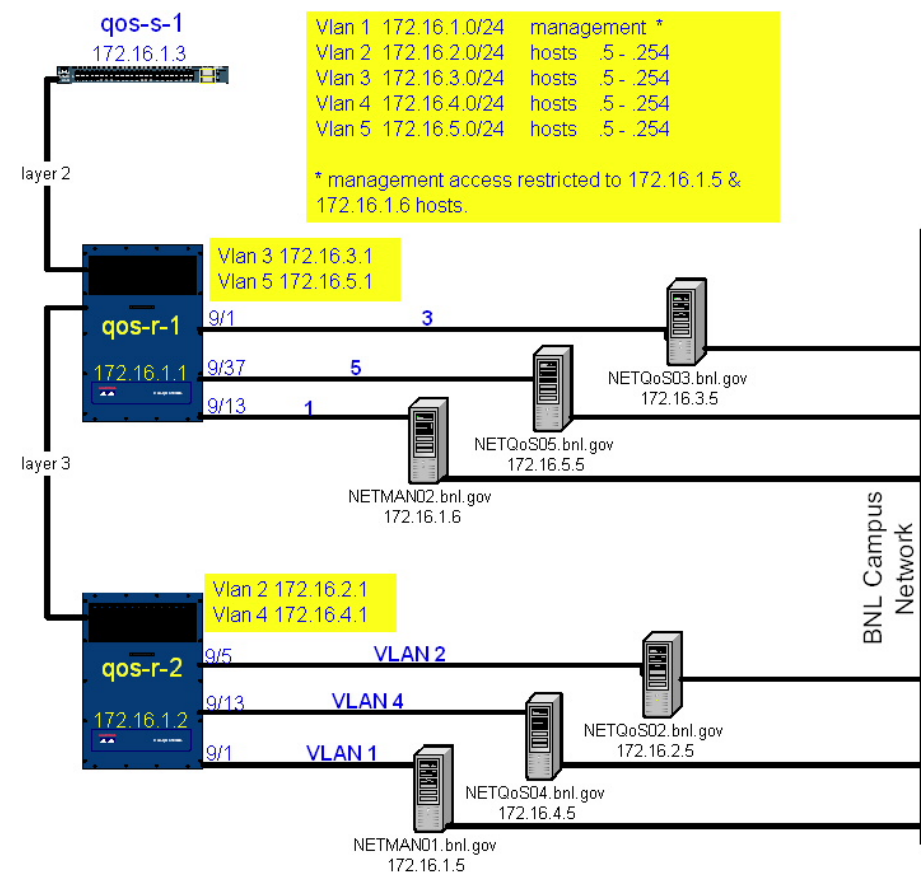- ❑ Gets the Existing Reservations for a Particular Time Interval

❆Int[] Getbandwidths()

- ❑ Provides a List of Bandwidths for the Managed Component

Summary of Progress

※ **A Full Featured LAN QoS Simulation Testbed Was Created in a Private Network Environment: Consists of Four Hosts and Two CISCO Switches (Same Models As Production Infrastructure)**

※ **Implemented Mechanisms to Classify Data Transfers to Multiple Classes.**

※ **Software Development Well Underway and Demo Version Is Running on Our Testbed.**
Http://netman01,02.Bnl.Gov:8080/tn

※ **Collaborate Effectively With Other DOE Projects.**

QoS Testbed

qos-s-1
172.16.1.3

Vlan 1 172.16.1.0/24   management *
Vlan 2 172.16.2.0/24   hosts .5 - .254
Vlan 3 172.16.3.0/24   hosts .5 - .254
Vlan 4 172.16.4.0/24   hosts .5 - .254
Vlan 5 172.16.5.0/24   hosts .5 - .254

* management access restricted to 172.16.1.5 & 172.16.1.6 hosts.

layer 2

Vlan 3 172.16.3.1
Vlan 5 172.16.5.1

qos-r-1
172.16.1.1
9/1   3
9/37   5
9/13   1

layer 3

NETQoS03.bnl.gov
172.16.3.5

NETQoS05.bnl.gov
172.16.5.5

NETMAN02.bnl.gov
172.16.1.6

Vlan 2 172.16.2.1
Vlan 4 172.16.4.1

qos-r-2
172.16.1.2
9/5   VLAN 2
9/13   VLAN 4
9/1   VLAN 1

NETQoS02.bnl.gov
172.16.2.5

NETQoS04.bnl.gov
172.16.4.5

NETMAN01.bnl.gov
172.16.1.5

BNL Campus Network

BROOKHAVEN
NATIONAL LABORATORY

Office of Science
U.S. DEPARTMENT OF ENERGY

Recent Goal, Milestones and Work Items

- ❆ **Implement Bandwidth Scheduling System.**
  - ❑ Current: First Come, First Serve for LAN Bandwidth.
  - ❑ Software System Integration With Oscar MPLS Scheduler.
  - ❑ Information Collector for Dwmi. (BNL ATLAS and SLAC).

- ❆ **Integrate TeraPaths/QoS Into Data Transfer Tools Such As File Transfer Services (FTS) and Reliable File Transfer.**

- ❆ **SC2005 Demo of Data Transfer Between BNL and UMICH, FNAL and CERN and Utilize Wan MPLS and LAN QoS. (BNL, UMICH, FNAL). ATLAS Data Management (DMS) ➔ Data Transfer Tools (FTS)➔ Data Channel Manager ➔ TeraPaths Bandwidth Reservation ➔ Monitoring and Accounting.**

Extend Scope and Increase Functionality of Prototype Service

❊ Inter-network Domain QoS Establishment, Dynamically Creating, Adjusting of Sub-partitioning QoS Enabled Paths to Meet Time Constrained Network Requirements.

- ❑ **Create Site Level Network Resource Manager for Multiple VOs Vying for Limited Wan Resource.**

- ❑ **Provide Dynamic Bandwidth Re-adjusting Based Resource Usage Policy and Path Utilization Status Collected From Network Monitoring (DWMI).**

- ❑ **Network Scheduling Research and Improve Efficiency of Network Resource.**

- ❑ **Integration With Software Products From Other Network Projects: OSCAR, Lambda Station, and DWMI.**

- ❑ **Improve User Authentication, Authorization, and Accounting.**

- ❑ **Goal: to Broaden Deployment and Capability to Tier 1 and Tier 2 Sites, Create Services Which Will Be Honored/adopted by CERN ATLAS/LHC Tier 0.**