

Final Technical Report

DOE award number: DE-SC0002350

Name of the recipient: University of Virginia

Project title: Resource optimization in hybrid core networks with 100G systems

Principal investigator: Malathi Veeraraghavan

Period covered by the report: Sep. 1, 2009 - Feb. 28, 2013

Executive summary: [Discuss (i) how the research adds to the understanding of the area investigated; (ii) the technical effectiveness and economic feasibility of the methods investigated; and (iii) how the project is otherwise of benefit to the public. The discussion should be written in terms understandable by an educated layman]

Research-and-Education Network (REN) providers have deployed both connectionless IP-routed and dynamic circuit networks. The objective of this project is to determine how these two networks can be integrated in a manner that offers both resource optimization and user-performance improvements. This will allow an integrated packet/circuit network to scale better to multi-Terabits/sec capacities as will soon be required to support the high-performance networking needs of DOE scientists.

The first solution investigated, Hybrid Network Traffic Engineering System (HNTES 1.0), was an online method for identifying long-duration flows arriving on the IP-routed network, and then setting up a dynamic circuit to redirect these flows to the high-speed circuit network. We changed our focus to size from duration after recognizing that long-duration flows may not be high in the rate dimension. Instead, high-rate flows are often generated by data transfer nodes dedicated to moving large-sized datasets, and high-rate flows are the ones that can have adverse effects on delay-sensitive flows. Furthermore, the need to port mirror packets to a server cluster in HNTES 1.0 for online detection of heavy-hitter flows could be cost-prohibitive.

Our second solution, HNTES 2.0, is an offline solution in that it identifies address prefixes of completed high-rate, large-sized flows by analyzing NetFlow records (which are created by IP routers by sampling packet headers), and then configures firewall filters in ingress routers to direct packets of these flows on to circuits that are traffic-engineered and QoS controlled specifically for these flows. This solution achieves dual goals of reducing the adverse effects of high-rate, large-sized (α) flows on delay-sensitive flows while simultaneously ensuring that these flows enjoy high throughput.

The results of our work have been adopted by ESnet. First, in response to our finding that policing α flows is detrimental to their throughput, ESnet created a new service class for these flows in their router configurations. Second, our α -flow identification algorithm has been extended to characterize α flows, and is being integrated into a graphical user interface for ESnet's usage. It will allow network engineers and scientists interested in network performance to visualize the size, duration and average rate experienced by these flows. Third, we showed that the offline HNTES solution can be highly effective in directing α flows to separate queues based on a detailed traffic analysis of NetFlow records obtained from ESnet routers.

The project benefitted the public by developing methods for providers to offer Internet services in a cost-effective manner. Also, the high-speed networking services supported by this research will enable new scientific discoveries in fields such as high-energy physics, climate studies, Earth sciences, genomics, etc. Two project Web sites, 13 publications, 3 technical reports, 12 presentations, and several software programs resulted from this work. This report provides the list of publications, presentations, and software programs. As for student education and training (human resources), this DOE project supported one PhD, three Masters, and three undergraduate students. Also, a new SBIR award was made recently to commercialize HNTES.

Comparison of the actual accomplishments with the objectives of the project: The project objectives stated in the proposal were as follows. Given that dynamic-circuit (DC) networks are being deployed as a complement to existing IP-routed networks, design a networking solution in which DC networks are leveraged to optimize resource utilization across the combined hybrid networks and improve user performance, and demonstrate this solution on research testbeds.

Objectives	Accomplishments
Traffic analysis	Obtained and analyzed NetFlow traffic from ESnet. Our analysis showed that large-sized, high-rate flows are low in frequency, but when they occur, they can have adverse effects on real-time flows.
Hypothesis testing	Proved the hypothesis that alpha flows are repeatedly generated between the same data transfer node pairs with static public IP addresses; this allowed for the design of an offline HNTES in which address prefixes from completed α flows can be used in firewall filters set in ingress routers to automatically redirect future α flows to traffic-engineered QoS-controlled paths.
Design and implement HNTES	Software modules consisting of flow monitoring module, IDC interface module, offline flow analysis tool, monitored flow data base with interface, were prototyped in HNTES1.0. This version of HNTES targeted online flow identification and redirection. After recognizing the costs of such a solution, we designed and implemented HNTES2.0, an offline version. The code for α -flow characterization is being transitioned to ESnet.
Test the effects of live path switching	We designed and ran experiments to determine the conditions under which out-of-sequence TCP segments occur, and in turn, cause throughput drops. This was important for HNTES1.0 in which live flows could be switched to circuits configured along different paths from the default IP-routed paths.
Implement and test software to use circuits	Router-configuration software was implemented and tested on the DOE LIMAN testbed. After a series of experiments, we showed that policing α flows could be detrimental to their throughput; instead a simple two-queue scheduling scheme in which α flows are directed to a separate queue is sufficient to achieve the dual goals of reducing adverse effects of α flows on delay-sensitive flows, while ensuring high throughput for α flows. This work led to the addition of a new QoS class in the deployed configuration used in ESnet routers.

Table 1: Objectives and accomplishments

Project activities [Summarize project activities for the entire period of funding, including original hypotheses, approaches used, problems encountered and departure from planned methodology, and an assessment of their impact on the project results. Include, if applicable, facts, figures, analyses, and assumptions used during the life of the project to support the conclusions.]

Year 1 (Sept. 2009 - Aug. 2010):

Our focus was on **dynamic circuits**, and hence on online flow detection, live circuit setup, and live flow redirection (“live” implies that actions are taken after the flow is detected). Also, the heavy-hitters were chosen to be **long-duration flows**.

- Netflow data analysis:
 - Since OSCARS circuit setup delay was on the order of 1 minute, only long-duration flows could use circuits because of the focus on dynamic circuits. Therefore, we developed algorithms and implemented statistical data analysis programs to analyze Netflow data to find long-duration flows. Internet2 Netflow data was analyzed because ESnet Netflow data was unavailable to us.
- HNTES 1.0 design:
 - A high-level hybrid network architecture document, and a detailed Hybrid Network Traffic Engineering Software (HNTES) design document, describing the components shown in Fig. 1, were completed and posted on the project Web site as listed under Technical Reports in the Products section.

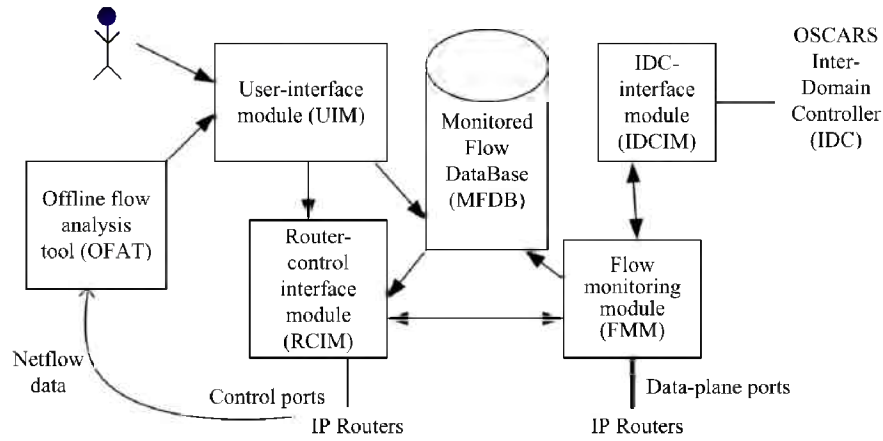


Figure 1: HNTES1.0 Architecture

- HNTES 1.0 software development: Different modules of the HNTES architecture were prototyped. This design called for online flow detection, online circuit setup, and online flow redirection (see Fig. 2).
 - Step 1: Identifiers of long-duration flows found through Netflow analysis, executed by the Offline Flow Analysis Tool (OFAT) module of the HNTES software (see Fig. 1), are stored in an SQL Monitored Flow Data Base (MFDB), and IP routers are configured by the Router Control Interface Module (RCIM) to port mirror packets from all of these flows to the Flow Monitoring Module (FMM) (see Fig. 1).
 - Step 2: When a packet from one of these monitored flows arrives at an IP router, it is port mirrored by the router to the HNTES server, and is thus received by the FMM (this is “live” online detection). The FMM then sends a message to the IDC Interface Module (IDCIM) to initiate circuit setup (see Fig. 1). The latter communicates with an OSCARS IDC (this is “live” circuit setup).
 - Step 3: After successful circuit setup, the IDC configures a policy based route (PBR) at the IP router to redirect packets for the live-detected flow to the newly established circuit (this is “live” flow redirection).
- HNTES 1.0 testing on DOE Advanced Networking Initiative (ANI) testbed:
 - We executed experiments on the DOE ANI testbed to test modules of the HNTES 1.0 software.
 - We demonstrated the code with video recordings (posted on the project web site) in an Oct. 2010 review with the DOE program manager. An interesting set of experiments tested the possibility of out-of-sequence packets when flows are redirected from IP-routed paths to circuits. When rate limiting is applied to the IP-routed path to lower the link rate to below that of circuit emulated path, then packets did arrive out of sequence, causing TCP throughput to drop. Three duplicate acknowledgments will cause the TCP sender to flag this event as a lost segment and hence half the congestion window size. This points to a need for careful selection of circuit rates before redirection of packets on live flows.

Year 2 (Sept. 2010 - Aug. 2011):

We changed focus from long-duration flows to **large-sized (elephant) flows**, and from dynamic circuits to **static circuits**, the reasons for which are explained below.

- Identified issues with the HNTES 1.0 solution and changed focus to size dimension:
 - Heavy-hitter science flows on ESnet and DOE laboratory networks are primarily of the large-sized variety. Furthermore, simulation studies showed that long-duration flows with low-to-moderate rate circuits (which are sufficient for low-latency applications such as remote visualization) have smaller effects on general-purpose flows than large-sized flows.

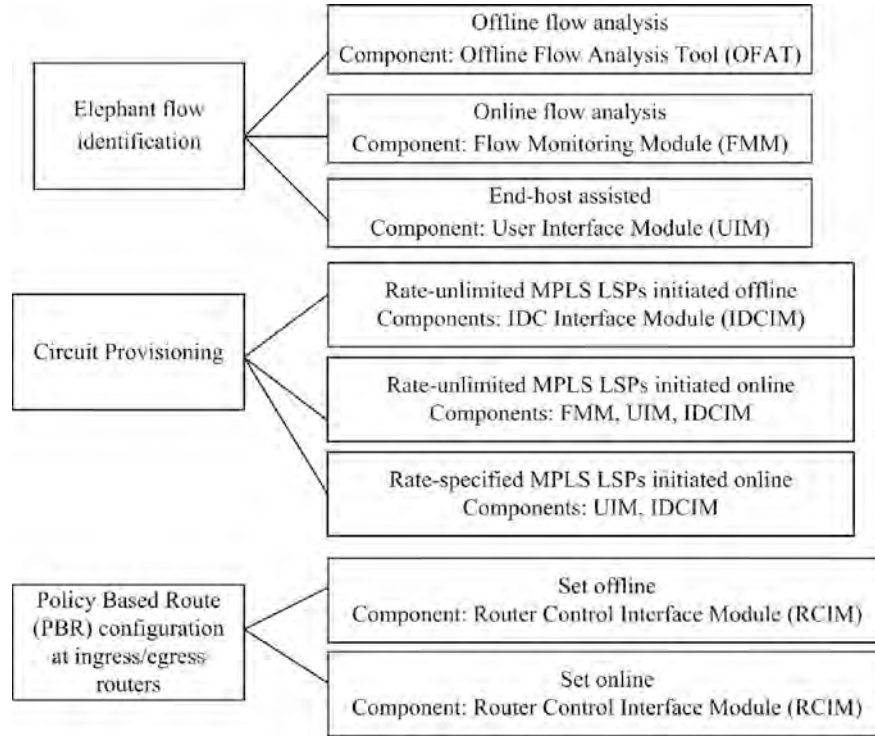


Figure 2: Tasks performed by HNTES

- Experiments conducted between hosts located at ESnet Points of Presence (PoPs) showed that 100 MB transfers take only 2 seconds each, and 1 GB transfers complete in 10 seconds each. While the threshold for a file to qualify as “large-sized” could be considerably bigger, nevertheless, the transfer delays could be smaller than circuit setup, which takes on the order of minutes. Therefore, the solution considers the use of static circuits, i.e., pre-provisioned circuits.
- The Year 1 focus was on creating one circuit per science flow (fine granularity) since the goal was to provide science flows rate-guaranteed circuits. In year 2, our focus is on isolating elephant flows from general-purpose flows by moving them off IP-routed networks to circuits. For this goal, individual per-flow circuits are not required. Instead, all elephant flows between the same ingress-egress ESnet routers can be grouped together onto one Label Switched Path (LSP). This is a coarse grained approach for using circuits for science flows. Such an approach is feasible with static circuits.
- Flow Monitoring Module (FMM) in HNTES 1.0 is impractical because it needs to process headers from all mirrored data packets for heavy hitter flows in real time, and as link speeds increase, this becomes infeasible without a high-performance computing platform. Requiring the deployment of such platforms to serve one or more ESnet routers could be cost prohibitive. Therefore, online (live) flow detection solutions that require the examination of even just the headers of data-plane packets are not desirable.
- Innovative approach to use rate-unlimited static LSPs:
 - If static circuits used between ingress-egress router pairs of ESnet or any other network are made to carry aggregate flows for science flow isolation reasons, the question is whether these circuits should be rate limited. For example, if each ESnet router is to be connected to its other routers via static circuits, in today’s structure, each 10 Gbps link has to be divided by some large number, such as 50 if there are 41 other sites and at least 9 border routers that peer with other providers, making each circuit rate quite low. The science flows will suffer from low throughput if such an approach is adopted. On the other hand, if all LSPs could be allowed to enjoy full link bandwidth (i.e., there is no rate limiting), then science flows would receive the best throughput possible

making it counter-productive to isolate each science flow in its own (necessarily lower-rate) LSP. This solution not only moves science flows off the IP-routed links, which means they no longer negatively impact the performance of general-purpose flows, it simultaneously allows science flows to receive the best throughput possible without micro-management of circuit network bandwidth. What is sacrificed is throughput variance, especially if the science-flow loads are moderate to high. This solution is feasible with the MPLS technology through the use of rate-unlimited static LSPs.

- Preliminary HNTES 2.0 design: Combines size dimension with static LSPs
 - The offline version of all three tasks shown in Fig. 2 will be executed, i.e., (i) offline flow analysis to determine large-sized flows, (ii) a periodic (e.g., nightly) run to determine if any new static LSPs are required (if all elephant flows recorded on a day are already in the database then no circuits will be required), and if a new circuit is required, the IDCIM software would contact the IDC to set up circuits, and (iii) a periodic (e.g., nightly) run of the RCIM to add policy based routes to redirect any newly identified flows to appropriate LSPs. Therefore, none of the three steps: flow identification, circuit setup, and PBR configuration are “live.” Further details are provided in HNTES 2.0 Design document, which is listed under Products in this document.
- Design algorithms for flow size estimation
 - Flow size estimation from Netflow data: We ran experiments to determine whether the size of a flow can be estimated accurately from Netflow data, which consists of sampled packet headers (e.g., ESnet routers execute 1-in-1000 packet sampling) in order to identify large-sized flows.
- Study of the question “why move elephant flows off the IP-routed network”
 - Experimental studies across ESnet to see the effect of elephant flows on link loads (passive SNMP measurements): The next section describes our findings.
 - Simulations: We executed simulation studies to test the hypothesis that under moderate-to-high loads, elephant flows consume an unfair share of link bandwidth thus effecting mice flows. This is because in TCP’s congestion control algorithm, the sender window grows in size until it reaches the bandwidth-delay product (BDP) size. In high-speed networks, especially over long-distance paths (high round-trip time), only elephant flows are sufficiently large for the sending window to grow to the BDP size. Mice flows complete before this size is reached, which means the effective throughput of mice flows is less than the bottleneck link rate.
 - PerfSONAR active measurements analysis: We also analyzed PerfSONAR One-Way Ping (OWAMP) data to determine if there are delay surges that can be attributed to router buffer buildups and found the presence of such surges.

Assessment of impact: significant results

- Flow size estimation experiments: We executed an experimental study across ESnet to test the conditions under which the hypothesis that “a flow size from sampled Netflow records can be multiplied by 1000 (since the ESnet Netflow sampling rate is 1 in 1000 packets) to estimate actual flow size” holds. We ran GridFTP transfers with files of known sizes, and collected Netflow data to test the hypothesis. We expected the hypothesis to be true for large files, and false for small files. The hosts *anl-diskpt1* and *lbl-diskpt1* located at ANL and LBL, respectively, were used to run the GridFTP server and client. There were eight ESnet IP routers in between these hosts, with 10 Gb/s links between the routers. A workflow script was created to automate the steps involved in this experiment. First a firewall counter was read at an en route IP router. Next the GridFTP transfer was initiated. When the transfer completed, the firewall counter was re-read to verify the transferred file size in bytes. From the GridFTP logs stored at the server, the data ports used for the TCP connections were obtained. Netflow data was collected from two of the routers. All flow records corresponding to the GridFTP transfer were filtered out of the Netflow records using the five-tuple identifier, and the size of the transfer was estimated using the 1000 factor multiplier on the total size reported by the flow records for the data connection. The size-accuracy ratio is defined to be the ratio of the Netflow estimated size and actual file size. For each file size (100MB, 1GB, and 10 GB), multiple runs were executed since the packet sampling at the

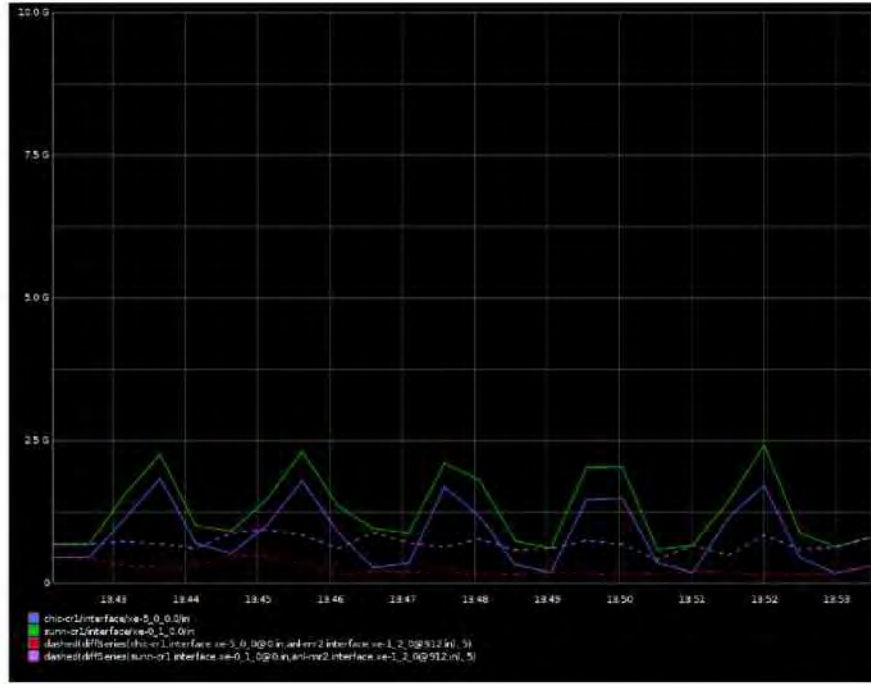


Figure 3: SNMP link loads (solid: with GridFTP transfers; dashed: without these transfers)

router makes the size-accuracy ratio a random variable. The results obtained are shown in Table 2. For all three file sizes, the sample mean shows a size-accuracy ratio close to 1, and more interestingly, the standard deviation is smaller for larger files. As these experiments were run across ESnet, traffic loads could be quite different when each of the runs was executed.

Size	NetFlow records obtained from Chicago ESnet router		NetFlow records obtained from Sunnyvale ESnet router	
	Mean	Standard deviation	Mean	Standard deviation
100 MB	0.949	0.2780	1.0812	0.3073
1 GB	0.996	0.1708	1.032	0.1653
10 GB	0.990	0.0368	0.999	0.0252

Table 2: Size-accuracy ratio for files for different sizes (sample size: 50)

- Experimental study of the effect of elephant flow on link loads: To check whether elephant flows have an effect on link loads, SNMP measurements (which averages bits sent/received over 30 sec periods) were obtained for the time interval in which five 10 GB GridFTP transfers were executed in the experiments described above. Figure 3 shows the SNMP link load plots while the GridFTP transfers were executed. The y-axis is the link load from SNMP measurements. The first marked horizontal line is 2.5 Gbps. Link capacity is 10 Gbps. The x-axis indicates time vertical lines separated by 1 minute. The solid lines show SNMP link loads on two router interfaces (on the Chicago and Sunnyvale ESnet routers) including the five experimental 10 GB GridFTP transfers, and the dashed lines represent the rest of the traffic load. The increases in link loads caused by the five experimental 10 GB transfers are clearly evident.

Year 3 until end (Sept. 2011-Feb. 2013):

- GridFTP log analysis: GridFTP servers collect usage statistics that record the size, time, duration and other information about each transfer served. These logs served as an invaluable resource to determine

	NCAR-NICS		SLAC-BNL	
setup delay	1 min	50 ms	1 min	50 ms
$g = 0$	0.12% (2.14%)	87.09% (89.33%)	1.95% (39.41%)	52.58% (89.68%)
$g = 1$ min	56.87% (90.54%)	92.89% (98.04%)	12.54% (78.38%)	93.56% (99.73%)
$g = 2$ min	62.16% (90.71%)	94.59% (98.17%)	15.93% (85.49%)	94.47% (99.85%)

Table 3: Percentage of sessions suitable for using VCs (percentage of transfers)

whether scientific file-transfer (high-rate) flows last long enough to allow for the setup of online virtual circuits (i.e., when circuit setup action is initiated after identifying an α flow at an ingress router) given that current OSCARS IDC circuit setup delay is on the order of 1 min. In other words, if transfer times are only a small multiple of setup delay or even shorter than setup delay, then the usage of circuits is an inefficient solution. With permission from NERSC (due to privacy considerations), Globus provided us NERSC GridFTP transfer logs for one month, Sep. 2010. Similarly, we collected logs from NCAR for a three-year period (2009-2011), and with login access on the NERSC and SLAC GridFTP servers, we were able to download usage statistics logs directly from these servers (with permission from NERSC and SLAC personnel). From these logs, information about transfers on four specific paths were isolated: NCAR-NICS (2009-2011), SLAC-BNL (Feb 10 - Apr 26, 2012), NERSC-ORNL (Sept 2010), and NERSC-ANL (Mar 4 - Apr 22, 2012). Our analysis focused on characterizing the size, duration, and throughput of GridFTP transfers in production environments, to determine the suitability of using dynamic virtual circuits (VCs) in an online HNTES. Shell scripts, `awk` scripts, and R programs were implemented to extract information from the raw GridFTP transfer logs. The output files from these programs were imported into a PostgreSQL database for easier filtering and analysis. As scientists often move large numbers of files within “sessions” rather than as single file transfers, we developed a heuristic and implemented an R program to group transfers into sessions from the log files, and focussed on size and duration of sessions, not individual file transfers. For throughput, we continued using the per-transfer measure because session throughputs could be lower if some of the individual transfers within a session had lower throughput, and we are interested in finding the maximum throughput achieved by these data transfers. While single file transfers are typically short, session sizes are much larger making the overhead of dynamic VC setup negligible even under a high throughput assumption, e.g., the third-quartile of observed transfer throughput. Finally, we analyzed the SLAC-BNL GridFTP logs to determine the split between the number of transfers that use multiple parallel TCP flows vs. those that use a single TCP flow.

Our findings are as follows: **(i)** Our GridFTP analysis showed that scientists do indeed move lots of files in sessions rather than single large files. Of the two datasets analyzed, NCAR-NICS data set consisted of 52,454 transfers, while SLAC-BNL dataset consisted of 1,021,999 transfers. If we set the threshold of maximum allowed gap g between transfers for inclusion within the same session to be 1 min, then there are only 211 sessions and 10,199 sessions for the two datasets, respectively. A session is considered “suitable” for dynamic VCs if the duration is at least 10 times the setup delay even under high throughput (the factor 10 is set somewhat arbitrarily). Assuming the current VC setup delay of 1 min, 56.87% of NCAR-NICS sessions, and 12.54% of SLAC-BNL sessions, would be suitable for dynamic VCs. It is interesting to point out, that these two numbers correspond to over 90% and 78% of all transfers in their respective datasets. Table 3 shows the percentage of sessions suitable for VCs for the NCAR-NICS and SLAC-BNL transfers. Furthermore, if VC setup delay can be decreased to 50 ms (e.g., with hardware implementations of signaling protocols), over 92.9% of NCAR-NICS sessions and over 93.5% of SLAC-BNL sessions would be suitable for dynamic VCs. **(ii)** The highest observed GridFTP transfer throughput is ~ 4.3 Gbps. Furthermore, on all four paths for which data was analyzed, transfers occurred at rates as high as 2.5 Gbps. Thus, these transfers do consume a significant portion of link capacity, which is typically 10 Gbps on these paths, and hence these high-rate, large-sized flows have the potential for adding delays to packets of other flows. While it was previously known that scientific transfers can reach high rates, the value of this analysis lies

in providing the actual observed rates. (iii) Our analysis of the SLAC-BNL dataset determined the percentage of transfers that use single- vs. multiple TCP flows. Of the 1,021,999 transfers, 864,762 (84.615%) transfers consisted of multiple (more than one) parallel TCP streams (a vast majority used 8 streams). A paper on this work was published in the Super Computing (SC2012) conference, and has been listed in the Products section of this document.

- NetFlow records analysis: In order to implement an offline HNTES, we needed to test the following hypothesis: Most high-speed data transfer nodes have static IP addresses, and α flows are created repeatedly between the same source-destination subnets. The basis for this hypothesis is that scientists typically execute their simulations on the same supercomputing centers, and hence we expect them to transfer data between the same two clusters. If the hypothesis is true, the offline prefix identifier based AFTES scheme will be effective in identifying and directing α flows to traffic-engineered, QoS-controlled paths. To test this hypothesis, we carried out traffic analysis of NetFlow records collected from ESnet.

We implemented an offline analysis tool (OFAT) and provided this tool to ESnet. Chris Tracy, ESnet, our collaborator, executed the OFAT code against seven months' Netflow data (May - Nov. 2011) collected at a provider edge (PE) router. ESnet ran scripts to anonymize the IP addresses before providing UVA the data for further analysis.

In this analysis, the methodology used was to aggregate α NetFlow records corresponding to a given source-destination pair (/32 or /24 IP address prefixes) on a per-day (aggregation interval was set to 1 day) basis to create α prefix flows. The number of bytes in each α prefix flow is referred to as α bytes. The total time within each aggregation interval in which at least one of the constituent α flows of an α prefix flow experienced an α -interval is referred to as α time of that prefix flow.

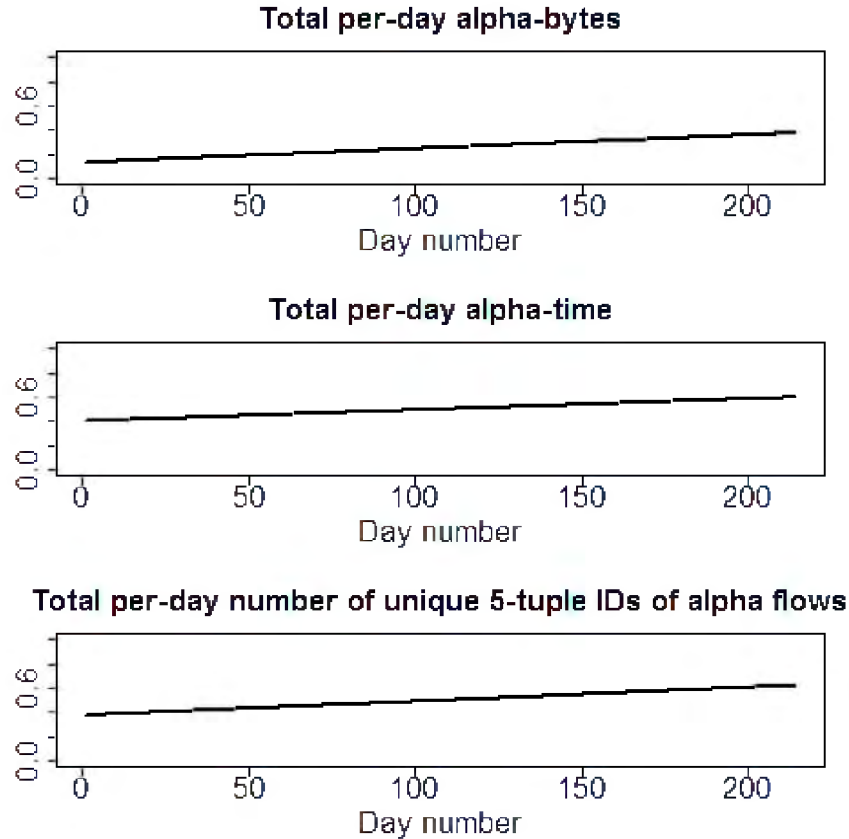


Figure 4: Binary quantized cumulative data plus a smoothing spline function with $df=2$

First, we characterized the α bytes and α time of α prefix flows as shown in Fig. 4. The increasing trend

Aging parameter (days)	/24	/32
7	82%	67%
14	87%	73%
30	91%	82%
214	92%	86%

Table 4: Percentage of α -bytes that would have been redirected and isolated across whole 214-day period

within the 7-month period is interesting to observe. Next, we defined a term called effectiveness as follows: *Effectiveness* is defined as the percentage of α -bytes that would have been redirected to traffic-engineered paths and isolated from the general traffic had AFTES been deployed. Table 4 shows the aggregate percentage of α -bytes that would have been directed across the whole observed time period of 214 days corresponding to different values of the aging parameter, A . Usage of /24 prefix IDs was more effective than the /32 prefix IDs. This is to be expected since data transfer nodes are often cluster computers with IP addresses in the same subnet. With new installations of high-speed data transfer nodes, previously unseen /32 prefix identifiers would have been covered by /24 identifiers. If rules are never deleted from the firewall filter, for the /24 case, 92% of α -bytes would have been isolated from β flows across the 7 months.

We also quantified the cost of this offline HNTES solution. By using /24 or /32 address prefix ID based firewall filter rules, β flows whose addresses match these prefixes will get directed to the traffic-engineered, QoS-controlled paths created for α flows, and consequently face degraded performance. However, our data analysis shows that a majority of such β -flow packets are from file-transfer applications and not delay-sensitive interactive applications. This is likely because most scientific computing/data transfer nodes are located in separate subnets from those used to connect general-purpose user hosts.

- QoS configuration experimental studies: A set of experiments were executed on the LIMAN testbed. This testbed offered us the opportunity to login to the Juniper routers and experiment with various configurations. The goals of these experiments were to gain an understanding of the adverse effects of high-rate α -flows on general-purpose flows, and to quantify the effects of different QoS mechanisms supported by the routers. The **diskpt** nodes on the LIMAN testbed were used for these experiments. Figure 5 shows the experimental setup. The IP routers, named **newy-tb-rt1-1** and **bnl-tb-rt-2**,

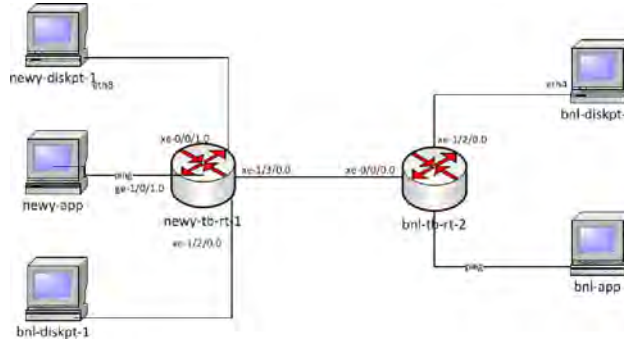


Figure 5: DOE ANI LIMAN testbed experimental setup

were Juniper MX routers running JunOS 10.2. The **diskpt** nodes were high-performance Linux hosts. The link between the routers and the links to the **diskpts** were 10 Gb/s Ethernet (10 GbE), while the interfaces to the **app** hosts were 1 GbE. The application used was **nuttcp**. Three flows were created all destined to **newy-diskpt-1**: (i) a fixed-rate (3 Gbps) **nuttcp** UDP flow from **bnl-diskpt-1**, (ii) a **nuttcp** TCP flow from **bnl-diskpt2**, and (iii) a stream of repeated **pings** from **bnl-app**.

The purpose of this experimental study was to evaluate the effects of different scheduling and policing mechanisms to achieve two goals: (i) reduction in delay and jitter of real-time sensitive flows that share

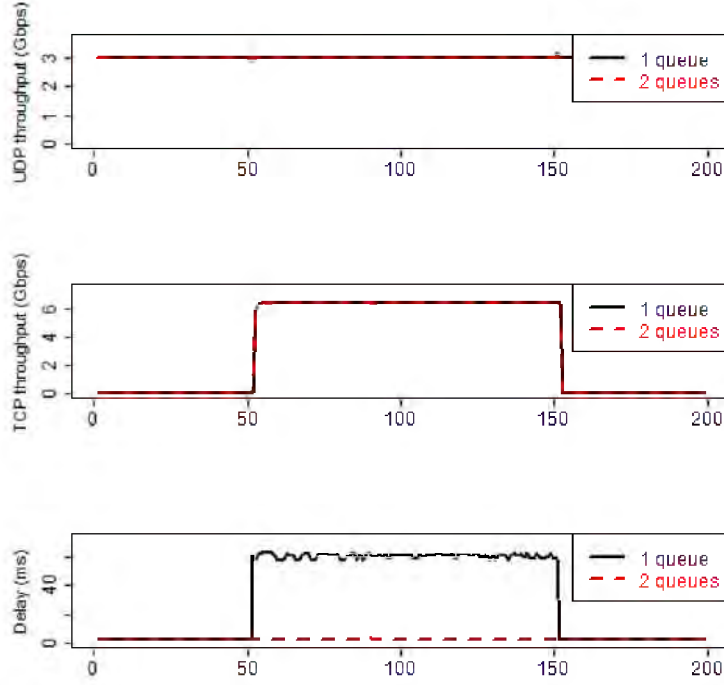


Figure 6: The x-axis is time measured in seconds; the top graph shows that the UDP rate is 3 Gbps in both the 1-queue and 2-queues configurations; the middle graph shows the TCP flow throughput; the bottom graph shows the delays experienced in the ping application.

the same interfaces as α flows, and (ii) support for high-throughput α -flow transfers.

Our *key findings* are as follows: **(i)** With the current widely deployed best-effort IP-routed service, which uses first-come-first-serve (FCFS) scheduling on egress interfaces of routers, the presence of an α flow can increase the delay and jitter experienced by audio/video flows (see the 1-queue configuration graphs of Fig. 6). **(ii)** This influence can be eliminated by configuring two virtual queues at the contending interface and redirecting identified α flows to one queue (an α queue), while all other flows are directed to a second queue (a β queue), as seen in the 2-queue configuration graphs of Fig. 6. **(iii)** If α flows use the dynamic circuit service offered by providers such as ESnet and Internet2, the currently configured policing mechanism will direct in-profile packets to a higher priority science-data queue, and out-of-profile packets to a lower priority scavenger-service (SS) queue, which in turn, may have adverse effects on throughput. The reason of this degraded α -flow throughput is that the separation of in-profile and out-of-profile packets to different queues can cause out-of-sequence arrivals at the TCP receiver, which triggers TCP’s fast retransmit/fast recovery congestion algorithm. Fig. 7 shows that in all the 3-queue configurations, where out-of-profile packets are sent to the scavenger-service queue, TCP-flow throughput drops significantly when the UDP flow rate is 1 Gbps or higher. The percentages shown for the α , β , and ss queues represent the allocated service rates. Since some of the plots overlap, we have summarized the mean TCP-flow throughput in Table 5.

In summary, a simple 2-queue solution, which separates out α -flow packets to a separate queue, is sufficient to meet the dual goals of low-delay impact on real-time flows, and high throughput for α flows. Policing α flows leads to a reduction in throughput, and is hence undesirable.

This work led to the addition of a new QoS class in the configuration of ESnet routers.

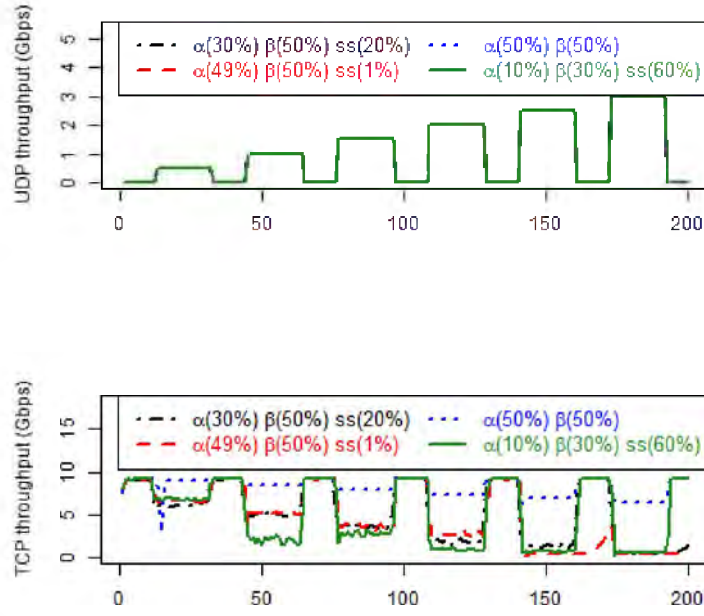


Figure 7: The x-axis is time measured in seconds; the top graph shows the on-off mode in which the UDP rate was varied; the lower graph shows the TCP flow throughput under the four configurations.

UDP rate (Gbps)	α -flow throughput (Gbps)			
	Percentages for 2-queues (α , β) and 3-queues (α , β , SS) configurations			
	(50,50)	(49,50,1)	(30,50,20)	(10,30,60)
0	9.12	9.09	9.07	9.12
0.5	8.92	6.62	6.06	6.83
1	8.43	5.22	5	2.12
1.5	7.94	3.78	3.67	2.82
2	7.44	2.7	1.93	0.92
2.5	6.95	0.33	1.38	0.69
3	6.46	0.34	0.38	0.61

Table 5: α -flow throughput under different background loads (UDP flow rate) and QoS configurations

Products Identify products developed under the award and technology transfer activities, such as: a. Publications (list journal name, volume, issue), conference papers, or other public releases of results. b. Web site or other Internet sites that reflect the results of this project; c. Networks or collaborations fostered; d. Technologies/Techniques; e. Inventions/Patent Applications, licensing agreements; and f. Other products, such as data or databases, physical collections, audio or video, software or netware, models, educational aid or curricula, instruments or equipment.

A. Publications (list journal name, volume, issue); conference papers; or other public releases of results.

- Z. Yan, C. Tracy, M. Veeraraghavan, “A hybrid network traffic engineering system,” IEEE HPSR, June 24-27, 2012.
- Z. Liu, M. Veeraraghavan, Z. Yan, C. Tracy, J. Tie, I. Foster, J. Dennis, J. Hick, Y. Li and W. Yang, “On using virtual circuits for GridFTP transfers,” IEEE SC2012, Nov. 10-16, 2012, Salt Lake City, UT.
- Z. Yan, M. Veeraraghavan, C. Tracy, C. Guok, “On How to Provision Quality of Service (QoS) for Large Dataset Transfers,” The Sixth International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ 2013), April 21 - 26, 2013 (won best paper award; submitted paper in Dec. 2012)
- Tian Jin, Chris Tracy, Malathi Veeraraghavan, Zhenzhen Yan, “Traffic engineering of high-rate large-sized flows,” IEEE HPSR 2013, July 8-11, 2013 (submitted paper in Feb. 2013).
- Z. Yan, M. Veeraraghavan, C. Tracy, C. Guok, “In How to Provision Virtual Circuits for Network-Redirected Large-Sized, High-Rate Flows,” International Journal on Advances in Internet Technology, vol. 6, no. 3 & 4, 2013, http://www.iariajournals.org/internet_technology.
- M. Veeraraghavan and Zhenzhen Yan, “Interaction between Applications and the Network,” Proc. of OFC2011, March 6-10, 2011, Los Angeles, CA.
- A. Jukan, M. Veeraraghavan, and Masum Hasan, “Hybrid networking: Evolution toward combined IP and dynamic circuit services”, Guest Editorial, IEEE Communications Magazine, Special Issue, May 2011.
- H. Bhuiyan, M. McGinley, T. Li, and M. Veeraraghavan, “An in-depth cross-layer experimental study of transport protocols over circuits,” in the Proceedings of IEEE ICCCN2010, Aug. 2-5, 2010.
- M. Veeraraghavan, M. Karol, G. Clapp, “Optical Dynamic Circuit Services,” IEEE Communications Magazine, vol.48, no.11, pp.109-117, November 2010.
- Malathi Veeraraghavan and David Starobinski, “A routing architecture for scheduled dynamic circuit services,” In Proceedings of the Re-Architecting the Internet Workshop (ReARCH '10), ACM, Philadelphia, PA (refereed workshop), Nov. 29-30, 2010.
- Malathi Veeraraghavan, Jie Li, Martin Reisslein, “A strawman proposal for future diverse internets,” The 16th IEEE Symposium on Computers and Communications (ISCC2011), June 28-July 1, 2011.
- Jie Li and Malathi Veeraraghavan, “A Reliable Message Multicast Transport Protocol for Virtual Circuits,” 4th International Conference on Communications, Mobility, and Computing (CMC 2012), 21-23 May 2012, pp. 119-123.
- Jie Li, Matthew Manley, Malathi Veeraraghavan, Steve Emmerson, “Analysis and selection of a network service for a scientific data distribution project,” 4th International Conference on Communications, Mobility, and Computing (CMC 2012), 21-23 May 2012, pp. 124-127.
- Technical report: Hybrid Network Traffic Engineering Software (HNTES), Feb. 28, 2010, <http://www.ece.virginia.edu/mv/research/DOE09/documents/deliverables/feb2010/mv-HNTES-final.docx>
- Technical report: A Hybrid Networking Architecture, Feb. 28, 2010, <http://www.ece.virginia.edu/mv/research/DOE09/documents/deliverables/feb2010/hybrid-arch-final.docx>

- Technical report: M. Veeraraghavan, Z. Yan, C. Tracy, Hybrid Network Traffic Engineering System (HNTES) 2.0 Design Document, May 15, 2011, <http://www.ece.virginia.edu/mv/research/DOE09/documents/deliverables/may2011/HNTES2.0-design.pdf>

B. Web site or other Internet sites that reflect the results of this project.

- UVA Hybrid Networking Project web site: <http://www.ece.virginia.edu/mv/research/DOE09/index.html>
- Wiki site: <http://venividiwiki.ee.virginia.edu/mediawiki/index.php/HNTES>

C. Networks or collaborations fostered.

- In addition to working closely with Chris Tracy, ESnet, we co-authored a conference paper with Chin Guok. We interacted with the other researchers at ESnet, including Brian Tierney, Inder Monga, Eric Pouyoul, and Jon Dugan. For the GridFTP work, we worked with Jason Hick, NERSC, Yee-Ting Li, SLAC, John Dennis, NCAR, and Ian Foster, ANL.

D. Technologies/Techniques.

- Our algorithms are part of documents posted on the project Web site under Documents.

E. Inventions/Patent Applications

- None.

F. Other products, such as data or databases, physical collections, audio or video, software or netware, models, educational aid or curricula, instruments or equipment.

Presentations:

- Hybrid Network Traffic Engineering DOE Projects Kick-off meeting, Batavia, IL Sep. 28, 2009.
- Resource Optimization in Hybrid Core Networks with 100G Links, ESnet J.Techs Spring Meeting Salt Lake City, UT, Feb. 4, 2010.
- Hybrid network traffic engineering, NSF PerfSONAR workshop Arlington, VA July 8-9, 2010.
- Resource Optimization in Hybrid Core Networks with 100G Links, DOE Projects Annual Review, Berkeley, CA Oct. 5-7, 2010.
- Heterogeneous networking, NSF FutureHetNets2011, March 23-26, 2011, Mountain View, CA.
- Analysis of Internet2 PerfSONAR data, Spring Internet2 Member Meeting, 18-20 April 2011, Arlington, VA
- Hybrid network traffic engineering system (HNTES), Presentation to ESnet, Virtual (attendees were in Berkeley, CA), June 23, 2011.
- Hybrid network traffic engineering system (HNTES) Project 1, DOE PI Meeting, BNL, Long Island, Jan 12-13, 2012.
- Hybrid network traffic engineering system (HNTES), ESCC Meeting at Joint Techs Baton Rouge, LA Jan. 25-26, 2012.
- Hybrid network traffic engineering system (HNTES), Annual DOE PI meeting for the ASCR Network & Middleware Bethesda, MD Mar. 1-2, 2012.
- Hybrid network traffic engineering system (HNTES), Internet2 Member Meeting, Arlington, MA, April 23-25, 2012.
- Traffic engineering and GridFTP log analysis, 2013 Winter ESCC Meeting at Joint Techs, Jan. 17-18, 2013.

Software: HNTES modules, such as the Offline Flow Analysis Tool, IDC Interface Module and others haven been posted on our project Web site.