Progress Review – 26 February 2004

Project Title: **INCITE: Edge-based Traffic Processing and Inference for High-Performance Networks**

Project Type: SciDAC

PIs: **Richard Baraniuk** (Rice University), **Wu-chun Feng** (LANL), **Les Cottrell** (SLAC)

Co-PIs: **Edward Knightly, Robert Nowak, Rolf Riedi** (Rice University)

Websites: incite.rice.edu, www-iepm.slac.stanford.edu

This report overviews progress in the first 2.5 years of the INCITE (InterNet Control and Inference Tools at the Edge) Project, a joint research program between Rice University, Los Alamos National Laboratory (LANL), and Stanford Linear Accelerator Center (SLAC). INCITE focuses experts from the fields of high-speed networking, high-performance computing (supercomputing and computational grids), statistical signal processing, and applied mathematics towards the goal of analyzing, modeling, and improving high-speed network services based solely on edge-based measurement at hosts and/or edge routers.

The Need for Edge-based Traffic Inference and Processing

Computer networks, distributed applications running on clusters, and computational grids are extremely complex and difficult to analyze. Moreover, optimizing their performance requires that end-systems have knowledge of the internal network traffic conditions and services. Without special-purpose network support (at every router), the only alternative is to indirectly infer dynamic network characteristics from edge-based network measurements. There is a great need for analysis, modeling, inference, and control tools capable of predicting network and application behavior based on limited measurements and network support. The tools must adapt the level of detail to fit the task and must deliver information in real time at guaranteed levels of accuracy and reliability.

To date, a number of proposals have been made for new methodologies for network modeling and control. Unfortunately, existing approaches are either ad hoc in nature, overly restrictive (requiring special-purpose network support), or lack rigorous and quantifiable performance bounds applicable to real-world networks. Thus, we still face a fundamental and daunting task — transforming modern high-speed inter-networks into manageable and predictable systems. The key to meeting this great challenge is the development of new data-driven theory and tools for modeling, characterizing, and controlling vast and continually evolving networks, all within the context of a scalable and easily deployable network core.

INCITE's Goals and Objectives

The *INCITE (InterNet Control and Inference Tools at the Edge) Project* is developing on-line tools to characterize and map host and network performance as a function of space, time, application, protocol, and service. In addition to their utility for *trouble-shooting* problems, these tools will enable a new breed of applications and operating systems that are *network aware* and *resource aware*. Launching from the foundation provided by our recent leading-edge research on network measurement, multifractal signal analysis, multiscale random fields, and quality of service, our effort consists of three closely integrated research thrusts that directly attack several key networking challenges of DOE's SciDAC program (see Figure 1):

Thrust 1: **Multiscale traffic analysis and modeling techniques**
Thrust 2: **Inference and control algorithms for network paths, links, and routers**
Thrust 3: **Data collection tools**

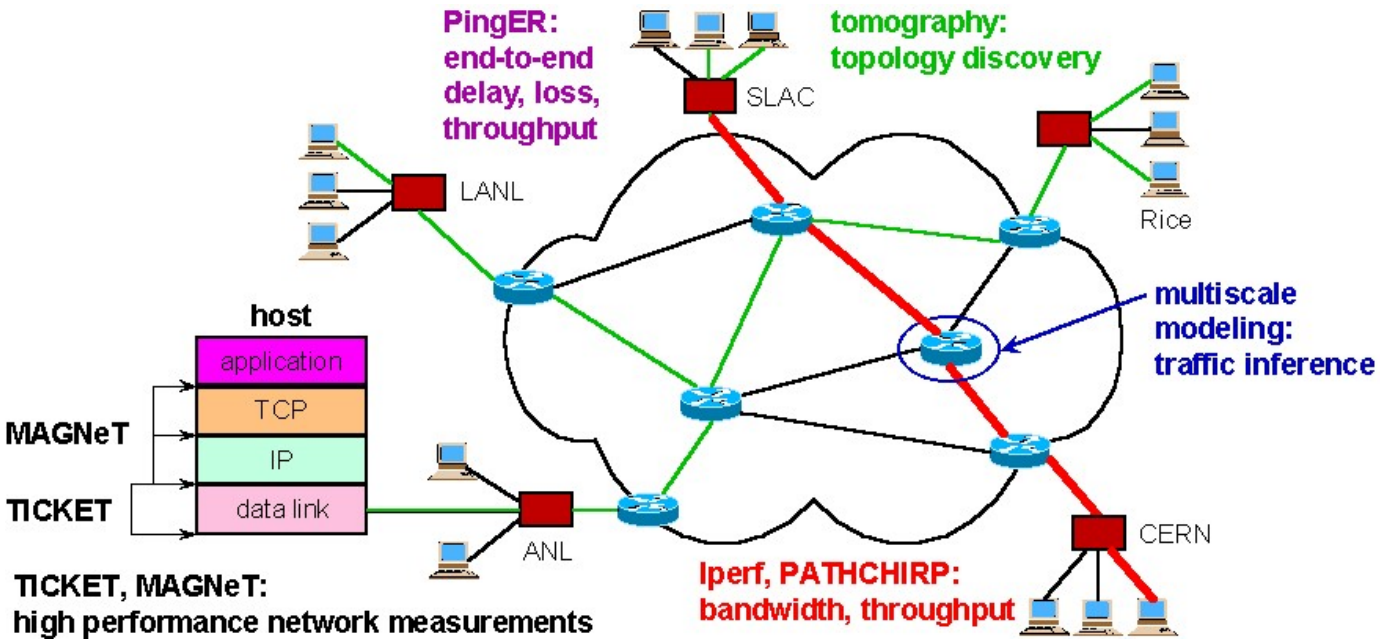

Figure 1: Relationships between INCITE's three research thrusts.

Technical Challenges and Progress to Date

Thrust 1: Multiscale traffic analysis and modeling

We are designing novel models that capture the multiscale variability and burstiness of high-speed network traffic. Leveraging *wavelets* and the powerful theory of *multifractals*, we are integrating model fitting, synthesis, and prediction into one unified statistical framework. Our new models, designed to match salient traffic characteristics at a prescribed level of abstraction, offer unprecedented realism while remaining analytically tractable, statistically robust, and computationally efficient. Using multifractal models, we are studying how large traffic flows interact and distribute their burstiness. Furthermore, we are investigating, analyzing, and characterizing the (adverse) modulation TCP/IP places on *application-level* traffic.

Progress to date: Burstiness in high-speed network traffic increases the queueing at routers and degrades performance. We have developed two powerful traffic models, the ***multifractal wavelet model*** (MWM) and the ***alpha/beta model***, that accurately and efficiently capture the statistical burstiness of high-speed traffic. In the alpha/beta decomposition, we have used connection-level information to identify that a very small percentage of connections cause nearly all of the bursts in traffic traces. The rest of the connections aggregate into simple a fractional Gaussian noise process (see Figure 2). We feel this discovery could have far-reaching implications in a number of areas, including modeling, queuing, scalability, and understanding the effect of network topology on traffic, synthesis, and inference (probing schemes, intrusion or hot spot detection). As of late 2003, the MWM model has been available in an open-source distribution from the incite.rice.edu web page. These tools are currently being integrated into the ns-2 network simulator, and a complete open-source analysis toolbox will be released soon. We also plan to explore using traffic modeling/synthesis tools to provide accurate performance predictions for computational Grids under realistic network conditions.

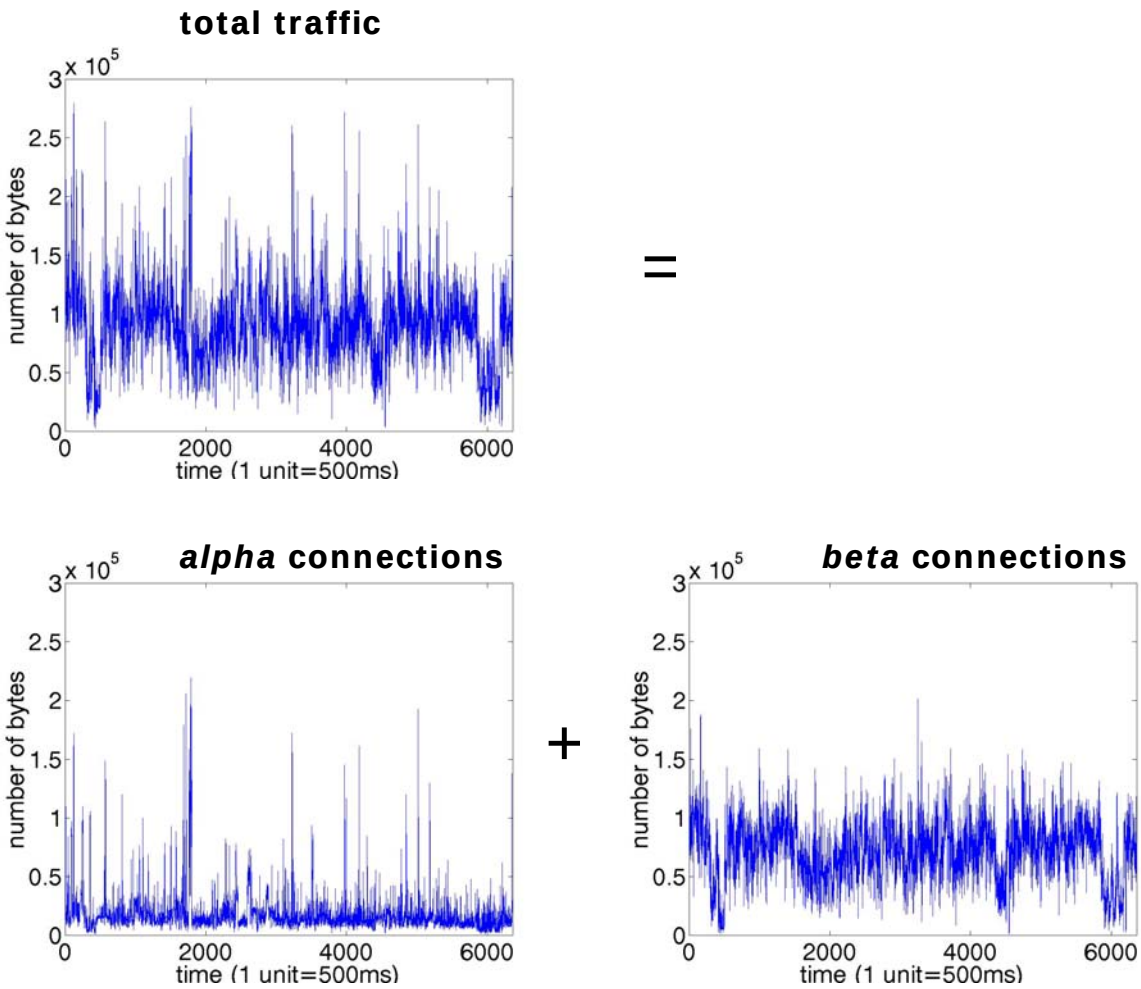

Figure 2: Connection-level traffic decomposition into an aggressive alpha component and residual beta component.

Thrust 2a: End-to-end path modeling and inference

Using multifractal traffic models, we have developed new theory and methods for understanding and inferring network-relevant properties of end-to-end paths and connections, especially their dynamics. This is the first step in making applications that are *network-aware*. In contrast to current approaches to path modeling, we explicitly model the causes of queuing latency and loss by introducing a model for competing packets. Our approach to infer the competing cross-traffic load utilizes an innovative exponentially spaced probing sequence of *packet chirps* that balance the trade-off between overwhelming the network with probes and obtaining statistics rich enough for accurate estimates. Passive monitoring theory is providing a novel means to analyze and model the *interaction of large network flows* that multiplex through queues along a common path.

Progress to date: Our first edge-based probing tool, ***pathChirp***, is based on a chirp packet train with exponentially decreasing packet spacings. Chirps probe the path at a wide range of probing rates using few packets, allowing accurate estimates of available bandwidth based on the self-induced congestion principle while introducing only a light load on the network. Our test results with pathChirp in simulation as well as on the real Internet have been encouraging; in particular we have been able to accurately estimate the available bandwidth on a high-speed link using about 10x fewer packets than the state-of-the-art tool *pathload* and many times fewer than *iperf*. Figure 2 illustrates pathChirp's ability to track changes in available bandwidth on the CAIDA Gigabit testbed; observe how the pathChirp estimate rises and falls in proportion to the introduced cross-traffic. Our paper on pathChirp received the *Best Student Paper Award* at the *Passive and Active Measurement Workshop* in April 2003. pathChirp is under test at a number of labs and is being integrated into the ns-2 network simulator. Integration into PingER and continuous monitoring of key Internet paths is

planned for the near future. Open-source software is available at spin.rice.edu/Software/pathChirp. Current and future work includes developing a version of pathChirp for locating the tight link (link with lowest available bandwidth) both in space and over time.

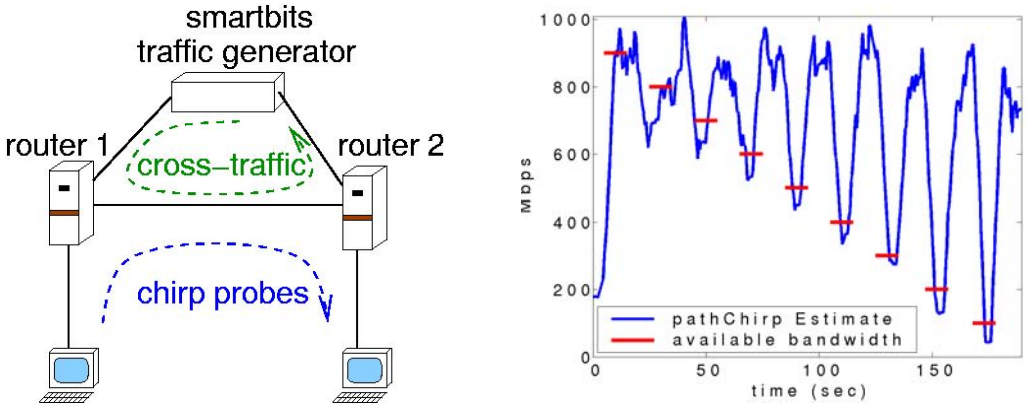

Figure 3: pathChirp provides an accurate estimate of bandwidth on the CAIDA Gigabit testbed with only lightweight probing load.

Our second edge-based probing tool, ***AbwE***, is designed to estimate the available bandwidth on very high-speed links (Gbit and higher) using efficient packet-pair dispersion. ABwE provides measures of the current path capacity, the cross-traffic and available bandwidth. During the last period we: extended its range of applicability to paths with lower performance, longer RTTs and higher loss rates; created a command line version; improved error recovery; and added RTT reporting. The new version of this estimator is called `abing` and includes a client and server. `Abing` was officially released as public available software in December 2003 at the workshop organized by NLANR and CAIDA at SDSC La Jolla, California. At the workshop, we presented a talk on "*What we have learned from developing and testing ABwE"* (*http://www.caida.org/outreach/isma/0312/slides/jnavratil.pdf*) that was evaluated by the workshop organizer as the best talk at the workshop. We described there the principals of our methods and discussed our results which we can achieve on our monitoring paths. We focused also on the problems discovered on high capacity lines (OC-12, OC-48 or OC-192) including packet reordering, coalescence, framing etc. The study of the effects obtained from several "problematic" paths operating with new technology devices will need more time. Understanding these effects will have a crucial impact on packet pair dispersion methods and for future research in this field. In our talk we also demonstrated the advantages of such a light weight (each forward and reverse bandwidth measurement requires sending only 40 * 1450 Byte packets) tool for network administration and showed several examples in which we can detect networking hidden problems.

The new features of `abing` allow us to monitor the paths to sites with large RTTs and large (<=10%) packet losses. To test our tool in such conditions we deployed `abing` at 30 PlanetLab sites and started regular monitoring of paths to Pakistan, Australia and China. The addition of RTT reporting assists in discovering structural changes and we plan to use this as a source of information for the Tomographical analysis solved in Task 3 of INCITE. The new version of `abing` also allows us to monitor paths which utilize wireless connections. It operates in a fashion to the well known `ping` command. We have versions for Linux, Sun and MS Wiindows. In the future we are planning to develop an IPv6 version of `abing`. The current version of `abing` is available from http://www-iepm.slac.stanford.edu/tools/abing/vers.1.1.1/

Thrust 2b: Network tomography

Resolving network behavior on a link-by-link or subpath basis is crucial to numerous networking problems, including performance prediction and diagnostics, control, dynamic routing, localization of "hotspots," selection of alternate paths, and attack detection. Network tomography refers to methods for estimating performance (e.g., loss rates, congestion) on internal links or subpaths using only external measurements made at the edge of the network. Thus, network tomography provides more than just an indication of end-to-end performance, it can detect where congestion is occurring inside the network.

Progress to date: ***Network Delay Tomography***. The substantial overhead of performing internal network monitoring motivates techniques for inferring spatially localized information about performance using only end-to-end measurements. In this paper, we present a novel methodology for inferring the queuing delay distributions across internal links in the network based solely on unicast, end-to-end measurements. The major contributions are: (1) we formulate a measurement procedure for estimation and localization of delay distribution based on end-to-end packet pairs; (2) we develop a simple way to compute Maximum Likelihood Estimates (MLEs) using the Expectation-Maximization (EM) algorithm; (3) we develop a new estimation methodology based on recently proposed nonparametric, wavelet-based density estimation method; and (4) we optimize the computational complexity of the EM algorithm by developing a new fast Fourier transform implementation. Realistic network simulations are carried out using network-level simulator ns-2 to demonstrate the accuracy of the estimation procedure.

Merging Inferred Network Topologies. Knowledge of network topology is useful for understanding the structure of the Internet, for developing and testing new protocols, and as prior information to network tomography algorithms. Building on existing techniques for inferring a single-source tree topology using end-to-end measurements, we address the problem of merging multiple tree topologies. We develop a multiple source active probing methodology and statistical framework for testing where the paths from two sources to two receivers branch at a common internal node. This information can then be used to determine where portions of the tree topology from one source to a set of receivers overlap with the tree topology from a different source to the same set of receivers. The algorithm uses a novel random probing structure and packet arrival order measurements that are easy to make. As a result, we do not require precise time synchronization among the participating hosts. We have performed experiments over the Rice University LAN as well as over a small test bed of hosts scattered around the Internet to verify that our methodology is versatile and robust.

Thrust 2c: New protocols for large file transfers over high-speed links

Service prioritization among different traffic classes is an important goal for the future Internet. Conventional approaches to solving this problem consider the existing best-effort class as the low-priority class, and attempt to develop mechanisms that provide "better-than-best-effort" service. We have been exploring the opposite approach, and are designing a new distributed algorithm to realize a low-priority "background" service (as compared to the existing best effort) from the network endpoints. To this end, we are developing ***TCP Low Priority*** **(TCP-LP)**, a distributed algorithm whose goal is to utilize only the excess network bandwidth as compared to the "fair share" of bandwidth as targeted by TCP. The key mechanisms unique to TCP-LP congestion control are the use of one-way packet delays for congestion indications and a TCP-transparent congestion avoidance policy. We are also developing ***HSTCP-LP*** (High-Speed TCP Low Priority), a high-speed TCP stack whose goal is to utilize only the excess network bitrate (bandwidth) as compared to the "fair-share" of bitrate as targeted by other TCP variants. By giving a strict priority to all non-HSTCP-LP cross-traffic flows, HSTCP-LP enables a simple two-class prioritization without any support from the network. It enables large file backups to proceed without impeding ongoing traffic, a functionality that would otherwise require a multi-priority or separate network.

Progress to date: The key mechanisms unique to TCP-LP and HSTCP-LP congestion control are the use of one-way packet delays for early congestion indications and a novel TCP-transparent congestion avoidance policy. Our results have shown that: (i) HSTCP-LP is largely non-intrusive to TCP traffic; (ii) both single and aggregate HSTCP-LP flows are able to successfully utilize excess network bandwidth; moreover, multiple HSTCP-LP flows share excess bandwidth fairly; (iii) substantial amounts of excess bandwidth are available to low-priority class, even in the presence of "greedy" TCP flows; and (iv) response times for interactive flows are dramatically improved when bulk flows use HSTCP-LP.

Our implementation of HSTCP-LP is derived by modifying the Linux-2.4-22-web100 kernel, which by default uses the HSTCP stack. An extensive set of Internet experiments on fast-production networks were performed. In the majority of the experiments, they launched flows from SLAC (Stanford, CA) to UFL (Gainesville, FL), as well as from SLAC to UMICH (Ann Arbor, MI). Our experiments showed that HSTCP-LP utilizes only 4.5% of the bitrate in this scenario, thus confirming its low-priority nature. Aleksandar Kuzmanovic, Rice University doctoral graduate student of Edward Knightly collaborated this summer with Les Cottrell from SLAC on extensive HSTCP-LP experimentation. The HSTCP-LP source code is available at: http://www.ece.rice.edu/networks/TCPLP/.

Denial of Service attacks are presenting an increasing threat to the global inter-networking infrastructure. While TCP's congestion control algorithm is highly robust to diverse network conditions, its implicit assumption of end-system cooperation results in a well-known vulnerability to attack by high-rate non-responsive flows. We are investigating a class of low-rate denial of service attacks that, unlike high-rate attacks, are difficult for routers and counter-DoS

mechanisms to detect. Using a combination of analytical modeling, simulations, and Internet experiments, we showed that maliciously chosen low-rate DoS traffic patterns that exploit TCP's retransmission time-out mechanism could throttle TCP. Moreover, as such attacks exploit protocol homogeneity, we studied fundamental limits of the ability of a class of randomized time-out mechanisms to thwart such low-rate DoS attacks.

Thrust 3a: Active network measurement and testing

After testing and validation, our algorithms for path inference and network tomography will be incorporated into SLAC's *PingER* measurement tools. Data from test measurements will be made publicly available by adding new metric selection(s) to the *PingER* time series tables. We also hope to integrate these algorithms with a proposed ESnet NIMI infrastructure. This will enable multiscale traffic analysis, path modeling, and tomography on high-performance paths between ESnet sites and ESnet collaborators.

Progress to date: Changes in network topology can result in large changes in performance. To study this we have extended the IEPM-BW measurements to include the forward and reverse routes at 10-minute intervals. By over-plotting the times of significant route changes on time series plots of Round Trip Time (RTT), available (measured by `abing`) and achievable bandwidth (measured by `iperf`, `GridFTP` and `bbftp`) we can spot whether a change in bandwidth performance is correlated with a route change. Figure 4 is an example of such a plot for the path from SLAC to Caltech from October – November 2003. During this period CENIC replaced many of the older CalREN2 networks that resulted in performance changes for this path.

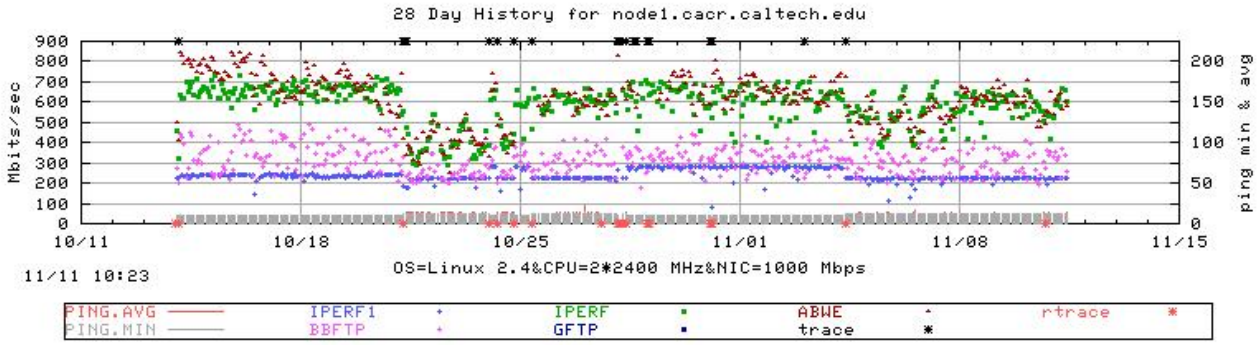

Figure 4: Time series plot with route changes indicated. The asterisks along the top axis indicate the forward traceroute changes. The asterisks along the bottom axis indicate reverse route changes. Note the correspondence between throughput changes and forward route changes.

Study of such plots reveals that there are several causes for bandwidth changes including: diurnal and weekend changes in load and thus congestion; changes caused by loads caused by long high-performance bulk-data transfers (e.g. for HENP data replication applications); changes caused routing changes. The latter are extremely important since they can be caused by mis-configurations and can go un-noticed for many days. In one case we identified a route change that caused a factor of six reduction in bandwidth between SLAC and Caltech, went un-noticed for a month, and was caused by incorrect advertising of routes between two Internet Service Providers (ISPs). To automate the process of correlating the bandwidth changes with route changes, we analyze the available bandwidth to detect significant step changes in the available bandwidth. Table 1 below shows preliminary information on the frequency of bandwidth changes being related to route changes for 32 paths from SLAC to hosts in Europe, Canada, the US and Japan. It can be seen that few route changes (<4%) result in a change in the available bandwidth, whereas most (65%) of the step bandwidth changes are associated with route changes.

Table 1. Summary of Route and Throughput Changes for 11/28/03 - 2/2/04.

Location (# nodes)	# route changes	# with bandwidth increase	# with bandwidth decrease	# bandwidth changes	# bandwidth change with route	# bandwidth change w/o route
Europe(8)	370	2	4	10	6	4
Canada & U.S. (21)	1206	24	25	71	49	22[1]
Japan (3)	142	2	2	9	4	5

Often a routing change will affect multiple routes. To assist in visualizing such effects we developed a tabular display of time of day versus remote host, showing the unique route number for each time and host. An example of such a table is shown in Figure 5. If the route number changes it is displayed in color, otherwise a period (.) is displayed to conserve space. From this web page we can also display the route associated with each route number and when it was last seen, the routes seen for a selected host for that day, as well as routes formatted for easy web viewing and in a format suitable for sending to ISPs.

When such changes occur, we want to know where along the route the problem occurs. The tomography tools being developed by INCITE are still in the research and testing phase. Therefore we developed a relatively simple tool to display topology maps, that takes as input the `traceroutes` and provides a map of the network routes from the measurement point to chosen remote hosts for selected times. Users can select the host(s) and time(s) on a web page such as shown in Figure 6, and view the associated topology maps. The maps are in the form of tree-graphs with the nodes and edges being colored by ISP, changes are in black, "mouseover" displays the node name, and one can select a node from which to start the route, or display a single route by selecting the end node.

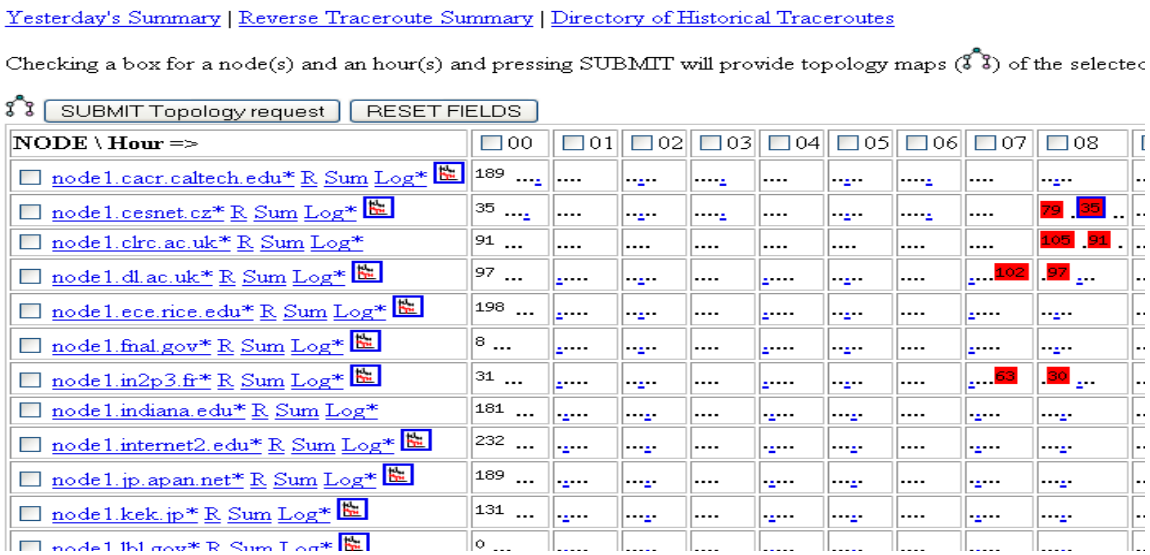

Figure 5: Screen shot of part of a traceroute summary web page with summary table.

Figure 6 shows several typical situations. In the simplest cases (top two maps) one or more hops are skipped. In more complicated cases when the route is changed in the beginning (as in the bottom right map) nearly the whole path is new and many remote hosts can be affected. In such situations we can expect a higher likelihood of bandwidth and RTT

[1] Note that 9 of these throughput changes are regular variations on Friday nights due to regularly scheduled large data transfers on one of the target networks.

changes. In the bottom right map, the path changes from CENIC/Abilene (left-hand path) to ESnet/Abilene. The new the route was not very stable and therefore there are multiple routes seen in the selected time window.

Sometimes, from the map, it is difficult to distinguish what is is a significant (in terms of performance) route change and what is not. Therefore we also review the network performance changes directly in the form of available bandwidth and RTT. In the case illustrated in the bottom right map of Figure 6, the route change had a large influence on the bandwidth to SLAC. Since the network is still working, the normal user may be only vaguely aware that performance is much worse. In some cases even the NOC (Network Operation Center of an ISP) personnel may not spot such problem and it can last for several days or weeks until reported. Fig. 4 shows the time series of the available bandwidth measured between SLAC and the remote host (Caltech) on October 9th, 2003, the effects of the route changes are clearly visible at around 14:00.

For a more complete understanding of the example in Figure 7, we present in Figure 8 a map showing these routing changes. The original path (before 14:00) is via Abilene (left branch) with 5 intermediate hops (part in Abilene and part in the Caltech local network). At 14:00 the routing changed to CENIC (right branch) and the number of hops substantially increased. The CENIC routers didn't have an advertised route to CALTECH and so routing used a "backup" route through LosNettos which had a bottleneck capacity of only 100 Mbits/s. The joining point for both paths (old and new) was one of the border routers at CALTECH. When the situation was fixed (17:00) the routing was set back to the original path. The reason for this situation was just a mis-configuration of the BGP (Border Gateway Protocol) system.

The maps become much more complex when there are more routing changes for the selected hosts within the selected time window. We can see this situation on Figure 6 bottom right, where there are 4 changes on the path to the University of Florida. In some cases the "changes" are trivial and may be caused by a router not responding. To assist in identifying such trivial changes we identify the node in the map as "busy".

We will present this work on PAM2004 workshop help in France. More details are available in the paper at http://www.pam2004.org/papers/285.pdf.

The measurement on paths where the capacity is below 155Mbits/s is reasonably well understood and the results are nearly identical for all packet dispersion methods (e.g. `pathchirp` and `pathload`) and also fit well to performance measurements done by `iperf`.. We are therefore focusing our studies on the behavior of ESnet, Abilene/CENIC and testbed paths with higher speed paths. We also plan to make more detailed comparisons between the packet dispersion tools developed in the INCITE project (`pathchirp` and `abing`) on paths with high capacity (>= OC-12) bottlenecks. The results will be published in the near future as a paper at CCCT'04 in Austin, TX.

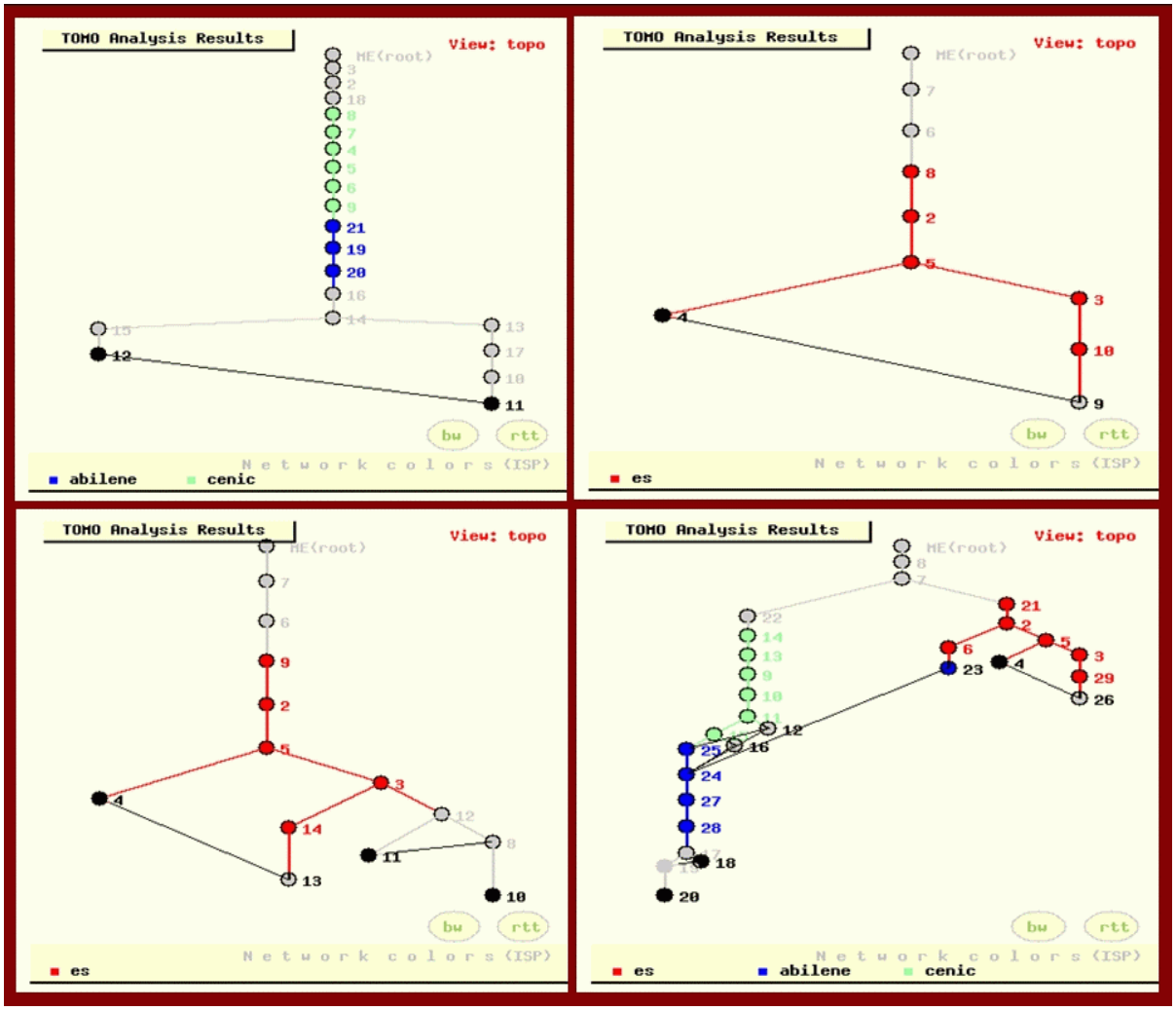

Figure 6: Screen shots showing examples of topology maps.

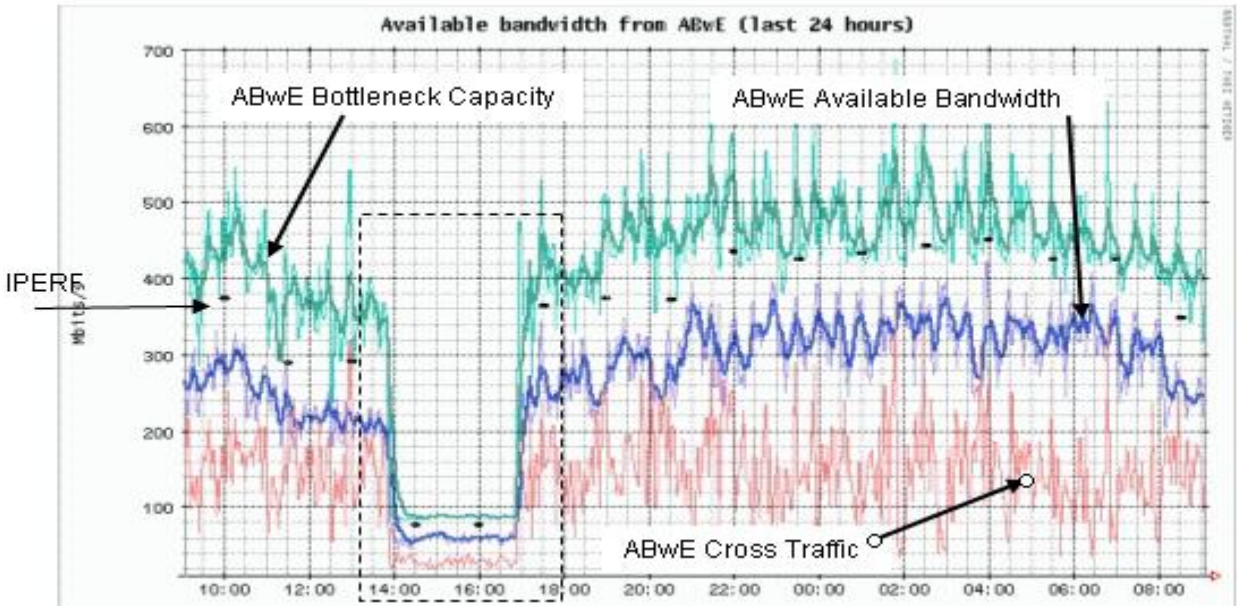

Figure 7: Plot of ABwE (available bandwidth) measurements and corresponding IPERF (achievable bandwidth) measurements. The y axis is the bandwidth in Mbits/s and the x-axis is the time in hours.
The frame area (between 14:00 and 17:00) represent the situation caused by routing changes.
The network work but the bandwidth dropped from ~300Mbits/s to 60Mbits/s.
Since the drop was relatively long, it was detected by both measurements methods (ABWe and Iperf)

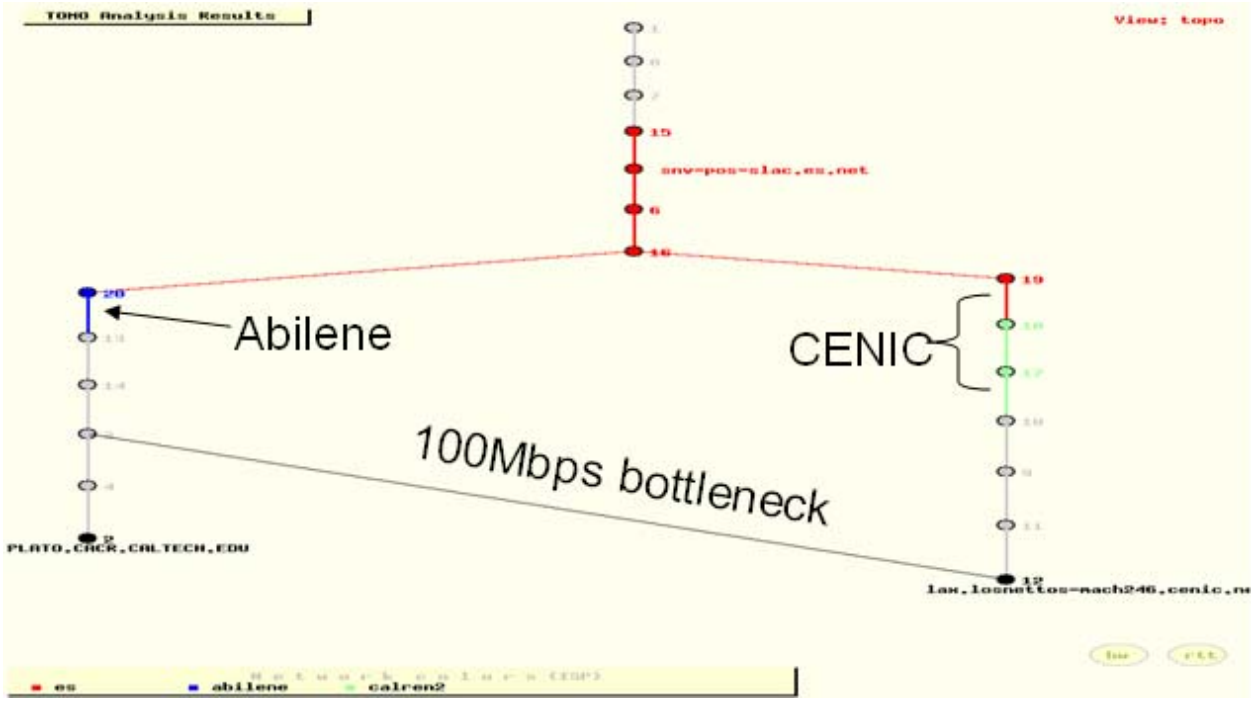

Figure 8: Traceroute tree-path showing routing change on the path from SLAC to CALTECH.

Thrust 3b: Passive application-level traffic measurement

INCITE's monitoring tools include *MAGNET* (Monitoring Apparatus for General kerNel-Event Tracing), *MUSE* (MAGNET User-Space Environment), and *TICKET* (Traffic Information-Collecting Kernel with Exact Timing). Together they act as a *software network oscilloscope* that can measure (capture packets) at different points in a host, cluster, or network, from the application to data link layer. MAGNET and MUSE permit applications and developers to obtain detailed information about the environment on a host and enable new *resource-aware* applications that adapt to changes in their environment (load balancing when needed, sensing when a node's resources are scarce or are bottlenecked, and so on). MUSE monitors without requiring modification or re-linking of applications. TICKET serves as a high-speed "tcpdump" replacement.

Using our sophisticated statistical traffic analysis tools, we are investigating, analyzing, and characterizing network traffic that is generated at the *application* level, that is, *before* the traffic is adversely modulated by TCP/ IP. While tools such as *PingER* provide information about traffic as it appears on the network, they say nothing about what the protocol stack does to the application-generated traffic before it enters the network. As no such tool currently exists, we have developed and will deploy a "Monitor for Application-Generated Network Traffic" (MAGNeT). MAGNeT, in conjunction with *PingER*, serves two purposes. First, it allows us to more closely examine, characterize, and model how the protocol stack modulates network traffic (since we will have traces of traffic before and after modulation by the protocol stack), and will likely provide insight as to how next-generation protocols must be designed to support the requirements of SciDAC end-user applications. Second, it will provide a library of traces for the network research community to use as real inputs to test these new, next-generation protocols or protocol enhancements. Currently, researchers must make do with unrealistic experimental conditions using infinite-file input distributions or distributions derived from traffic already modulated by the protocol stack.

Progress to date: Due to the tremendous feedback from application users to generalize MAGNeT to monitor *any* kernel event in a cluster or grid environment, we embarked on a new version of the software, appropriately renamed Monitoring Apparatus for Generalized kerNel-Event Tracing (MAGNET with a big "E"). As in the previous version of MAGNeT, we re-used our implementation of a highly-accurate timestamp generator as well as our performance profiler to understand the impact of MAGNET on the OS kernel. We then performed the following research and development:

- Designed and implemented a more efficient MAGNET kernel patch that is capable of exporting *any* kernel event (e.g., CPU context-switching, file I/O, memory paging, etc.) via a kernel/user-space shared memory region.
- Profiled the performance of the new version of MAGNET and demonstrated a 50%-100% improvement over the old version, which in turn was able to monitor traffic less obtrusively than even the ubiquitous `tcpdump` tool, which does *not* scale to gigabit speeds.
- Created the MAGNET User-Space Environment in order to allow applications to appropriately filter the plethora of information that the kernel generates, e.g., monitoring four events in the network subsystem creates about 10 MB of data in one second. See Figure 8 below.
- Released an alpha version of MAGNET and MUSE.
- Initiated integrative activities with GridFTP (ANL) within Globus to enable network-aware applications.
- Demonstrated a crude prototype at SC2003 that integrated SvPablo/Autopilot (UIUC/NCSA) with MAGNET as a sensor to enable the *live* remote monitoring of a Globus application running over a grid. See Figure 9 for an architectural diagram of the SC2003 prototype and Figures 10 and 11 for sample screenshots of the bandwidth used in ScaLAPACK as a function of time and the number of context switches in ScaLAPACK as a function of time, respectively.

ScaLAPACK is a set of specially formulated routines for solving systems of linear equations on distributed memory architectures. One characteristic of this computation is that the amount of work (and hence, the amount of data transferred between nodes) decreases as a function of time. In ScaLAPACK, messages exchanged between the processors becomes progressively smaller as the matrix is reduced, an effect that can be clearly seen in Figure 10. This figure shows periods of intense communication (spikes) followed by periods of near quiescence. As the amount of data that needs to be transferred between nodes decreases, the bounding envelop for maximum transfer rate also decreases. Relative to Figure 11, ScaLAPACK demonstrates a similar decreasing "feast-or-famine" pattern for context switches per second; however, the pattern is not as clearly as defined because context switches per second are affected both by computational and by communication. During periods of intense computation, other processes (such as system daemons) execute, causing context switches to occur. Even without other processes, each decision to continue the execution of a compute-bound process increases the number of context switches by one.

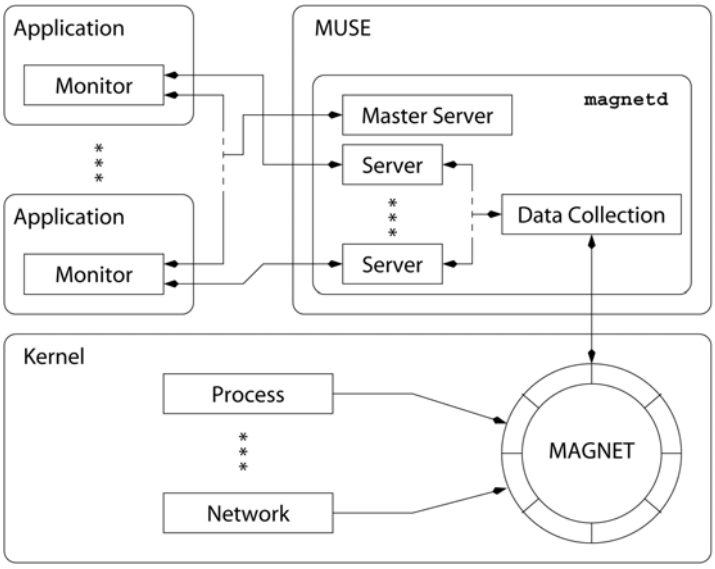

Figure 8: MAGNET / MUSE architecture.

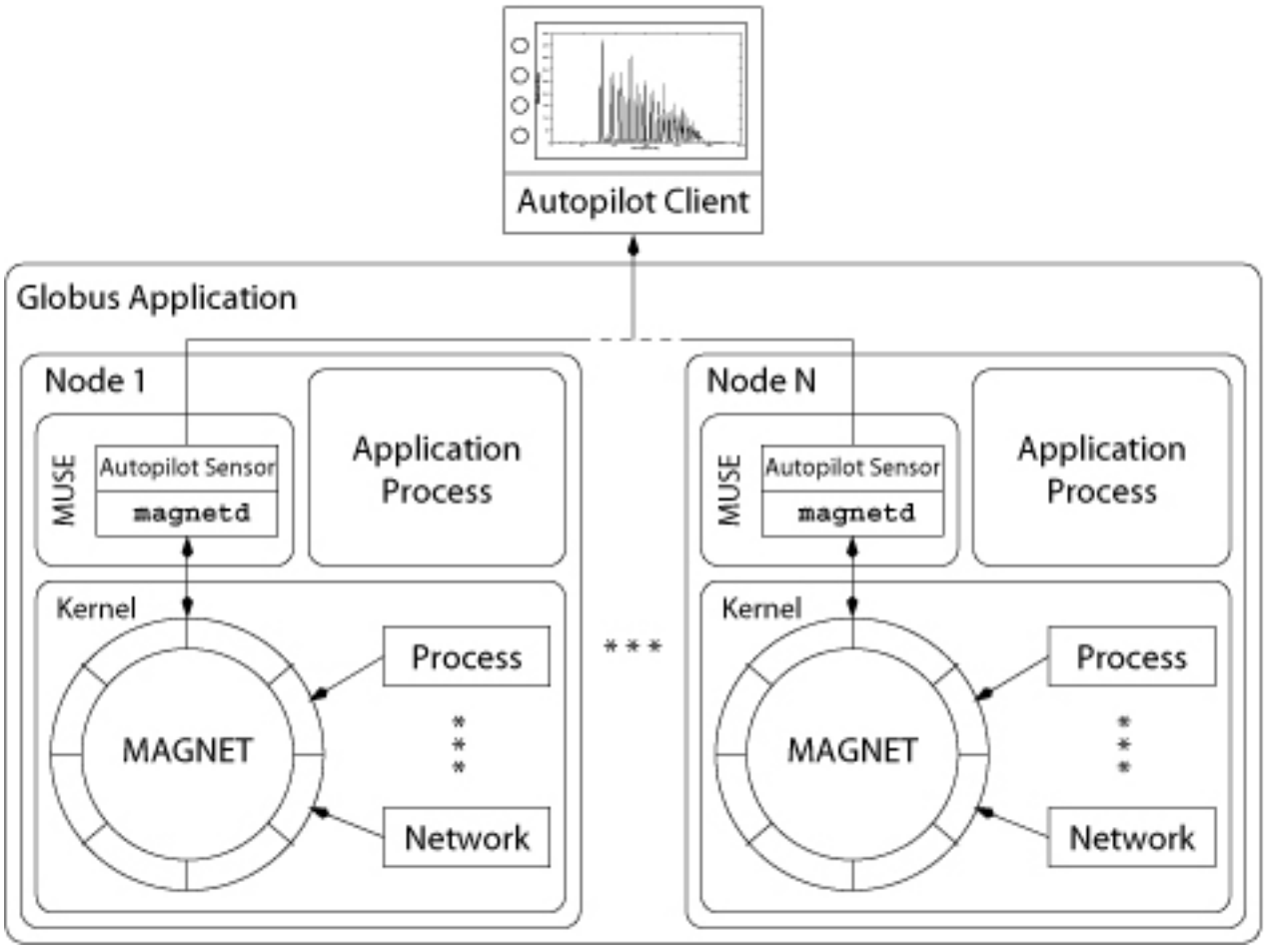

Figure 9: An integrated architecture for Autopilot, MAGNET, and Globus

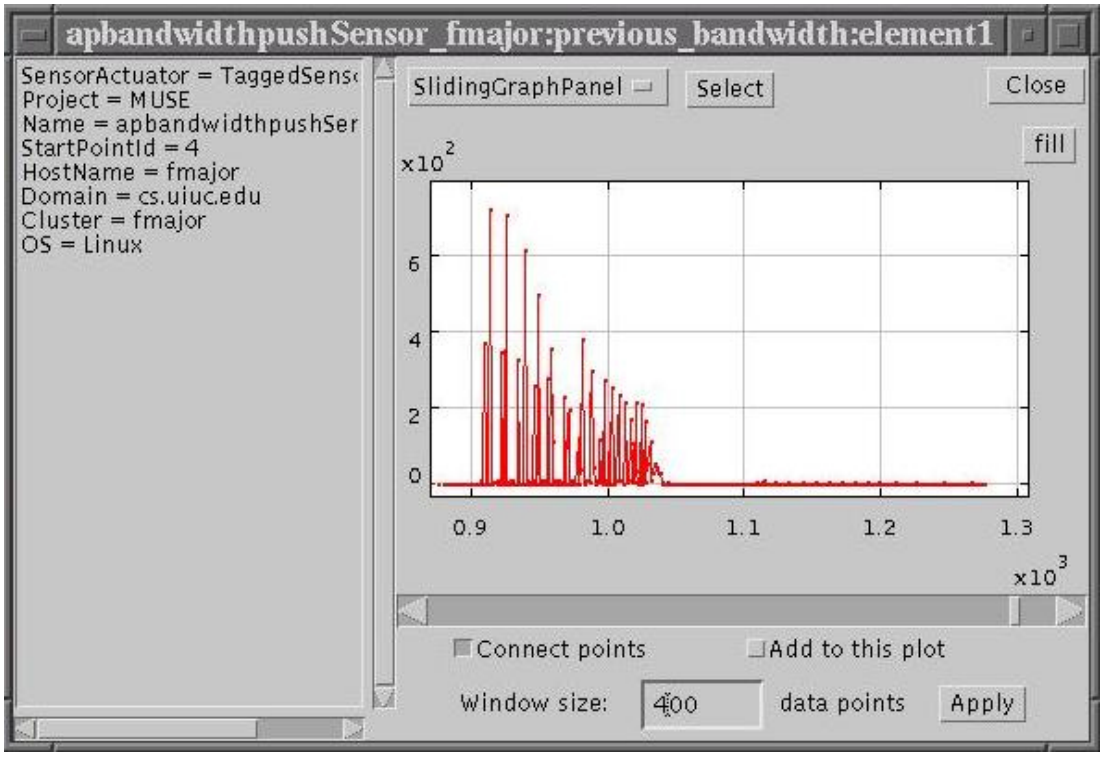

Figure 10: Bandwidth used on ScaLAPACK (via an Autopilot GUI).

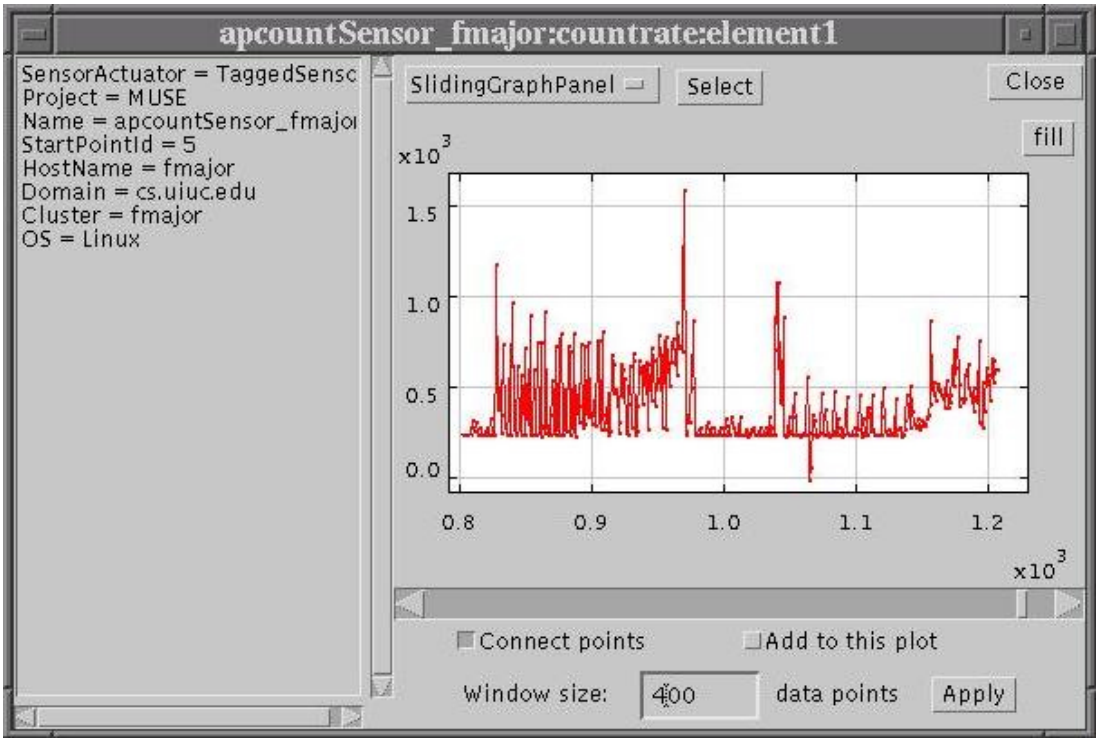

Figure 11: Context-switch rate on ScaLAPACK (via an Autopilot GUI).

To simultaneously address the issues of scalability, performance, and cost (something that no existing monitor currently does), we propose TICKET: Traffic Information-Collecting Kernel with Exact Timing. TICKET is a non-intrusive, passive monitor that combines efficient, commodity-based software with a novel but commodity-based hardware architecture to monitor traffic at gigabit-per-second speeds and beyond. Furthermore, it can achieve this performance at a fraction of the cost of commercial offerings, i.e., as much as 100 times cheaper than commercial offerings. Thus, TICKET's low cost and superior performance will result in a price-performance ratio that is orders of magnitude better than anything currently available, as shown in Table 2.

Table 2. Comparison of passive performance monitors.

Product	Price	Performance	Price / Performance
`tcpdump`	$1,000	400 Mbps	$2.50 / Mbps
Commercial offerings	~$200,000	2000 Mbps	$100.00 / Mbps
TICKET	$2,000	3000 Mbps	$0.67 / Mbps

Note: Although the software for tcpdump and TICKET are free, the former requires a PC while the latter requires two PCs.

In addition, and perhaps most importantly, TICKET can take measurements that are five *orders* of magnitude finer than `tcpdump` and eight *orders* of magnitude finer than commercial offerings. Because of these attributes, we have been using this tool to collect aggregate traffic traces of the LANL backbone for the purpose of building new, but usable, network traffic characterization models. Our first such model, to appear this month at the world-renown *38th Annual Conference on Information Sciences and Systems (CISS'04)*, will present a deterministic characterization of network traffic for average performance guarantees. Future work will examine an alpha-beta modeling of the traffic with Rice University (slated for April 2004) as well as a new traffic characterization that will facilitate a tractable analysis of probabilistic quality-of-service guarantees.

SciDAC and DOE Relevance

One SciDAC's Holy Grails for data grid applications such as the PPDG is to make the applications and operating systems more *network aware* and *resource aware*. Here we briefly overview how INCITE's tools provide a powerful toolkit for attaining these goals. We expect that as we improve our tools, new applications will constantly emerge.

Optimizing large file transfers

Many SciDAC applications, including automated data replication, web caching, and streaming, will require large file transfers for which current protocols like TCP are ill-suited. Thanks to real-time network awareness from our efficient available bandwidth and tomography tools, applications such as GridFTP, bbcp bulk file copy/transfer, or our TCP-LP will be able to make reasonable initial settings of TCP parameters such as maximum window size and number of streams. This information can also be reported to higher-level software or to the user to enable more global decisions. We plan to integrate our INCITE tools with the Globus MDS and similar mechanisms. We plan to demonstrate a proof-of-concept by adding network steering to a high throughput application such as bbcp or GridFTP. Another strategy that high-throughput SciDAC applications may find particularly attractive will be the ability to determine whether there is an advantage in splitting up a copy, for example, copying the first half of the file from site *B* and the second half from site *C*. This in turn requires knowing whether the bottleneck hop is shared by both paths, in which case there is no advantage to splitting up the file copy in this way. We will to infer and provide this information with our tomography tools.

Latency sensitive applications and resource scheduling

Controlling transmission latency is crucial in high-performance Grids and applications like remote visualization and application tuning. Moreover, in a Grid environment, deciding where and when data should be processed requires a scheduler that makes intelligent trade-offs between communication cost and processing cost. INCITE's tools will provide accurate on-the-fly estimates and predictions of latency, jitter, and available bandwidth to these applications. Indeed, researchers at ANL, UIUC, and NCSA (among others) are already using MAGNET+MUSE.

Network monitoring and troubleshooting

A major concern for SciDAC grid applications is the ability to trouble-shoot problems, including excessive latencies, congestion, routing faults, and anomalous traffic. Users and applications need information on the round trip time, losses, routing path, level of utilization of critical links, and also the achievable throughput to partner sites. Further they need not only the current status (which they can measure) but also whether it is normal or not, and if this situation is likely to be stable for several hours or the whole day, etc. Hence, each monitoring system must make measurements, archive them, and provide predictions. Currently, using iperf and other network intensive tools (such as the bbcp, bbftp and GridFTP file transfer tools), we can measure some of this information, but due to their excessive load on the network, we obtain only a very coarse resolution (measurements only every 60-90 minutes). INCITE's light-weight tools will be able to accurately capture this information at all levels, from applications to network, continuously in real-time – when they are actually needed by users and applications.

Milestones and Deliverables

The INCITE project is producing four classes of deliverables: (1) new theory and technologies, (2) software simulations of these technologies, (3) experimental testing and reference implementations, and (4) publications in the form of talks, technical reports, and papers. In particular, we expect to produce the following:

Multiscale traffic analysis toolbox

We will complete development and release an open-source toolbox for analyzing high-speed network and Grid traffic traces using wavelets, multifractals, and the alpha/beta traffic decomposition (Spring 2004).

End-to-end path modeling and inference toolbox

We will complete validation and testing of *pathChirp* and *AbwE*, integrate them into the PingER infrastructure, and explore applications in Grid computing (Summer/Fall 2004).

Network tomography toolbox

We will complete validation and testing of *NeTomo*, integrate it into the PingER infrastructure, and explore applications in Grid computing (Summer/Fall 2004).

New protocols for large file transfers over high-speed links

After final validation and testing, we will release an open-source toolkit for HSTCP-LP (Spring 2004). We will also explore and develop new protocols that tune their transmission parameters according to estimates from our on-line path inference tools (Summer/Fall 2004).

Passive application-level traffic measurement

We will accelerate the development and testing of MAGNET+MUSE at the expense of TICKET. A new version of MAGNET and an initial prototype of MUSE will be completed, tested and released (Summer 2004). In addition, ongoing collaborations with the following projects will be pursued: (1) Globus (Argonne) and SvPablo/Autopilot (NCSA/U. Illinois), (2) Supernova Science Center (UCSC/LANL/LLNL/U. Arizona), (3) Earth System Grid II (Argonne/NCAR/LLNL), and (4) a vertically-integrated project with the scientific visualization and climate modeling teams at LANL.

Opportunities for Follow-On and Collaborations

We have been working closely with the IEPM-BW developers at SLAC who have close ties with the PPDG (the IEPM-BW is an unfunded PPDG collaborator). Through the SLAC PI of INCITE, we will make the PPDG more aware of the INCITE work as it progresses and comes closer to a tool that the PPDG can use. Also we will be working with major SciDAC and Grid sites to encourage the deployment of traceroute servers and the traceping route measurement, recording and archiving tools so that we will be able to report more data via our topology/tomography tools.

We are also more closely collaborating with other SciDAC funded projects such as the CAIDA bwest project developing the pathrate and pathload bandwidth estimation tools and the LBL project developing netest-2 and pipechar. As we create archival data we will also work with the Globus and PPDG teams to make INCITE data available via MDS and other similar tools.

Current INCITE users include: Globus, Particle Physics Data Grid Collaboratory Pilot, Scientific Workspaces of the Future, SciDAC Center for Supernova Research, TeraGrid, Transpac at Indiana U., San Diego Supercomputing Center, ns-2, Telecordia, CAIDA, Autopilot, TAU, the European GridLab project, IEPM-BW, CAIDA, LBL, ns-2 project, JavaSim, J9, ONR, Sprint, AT&T, Telecordia, Internet2,

Project Management

Strong management is crucial in a distributed project like INCITE. We have set up an INCITE web site at incite.rice.edu and an archive of notes from meetings between INCITE collaborators. We have regular face-to-face meetings of the collaborators and many telephone meetings. Several Rice graduate students have spent internships at SLAC and LANL for technology transfer; further internships at SLAC and LANL are planned (including one to LANL in April 2004).

Recent Publications

W. Feng, G. Hurwitz, H. Newman, S. Ravot, R. L. Cottrell, O. Martin,F. Coccetti, C. Jin, X. Wei, S. Low, "Optimizing 10-Gigabit Ethernet for Networks of Workstations, Clusters, and Grids: A Case Study," *IEEE/ACM SC 2003*, November 2003.

V. Ribeiro, R. Riedi, R. Baraniuk, "Optimal Sampling Strategies for Multiscale Models with Application to Network Traffic Estimation," *IEEE Statistical Signal Processing Workshop*, September 2003.

M. Gardner, W. Feng, M. Broxton, G. Hurwitz, and A. Engelhart, "Online Monitoring of Computing Systems with MAGNET," *IEEE/ACM Cluster Computing and the Grid* (CCGrid'03), May 2003.

M. Gardner, M. Broxton, A. Engelhart, W. Feng, "MUSE: A Software Oscilloscope for Clusters and Grids," *IEEE International Parallel and Distributed Processing Symposium* (IPDPS'03), April 2003.

V. Ribeiro, R. Riedi, R. Baraniuk, J. Navratil, R. L. Cottrell, "pathChirp: Efficient Available Bandwidth Estimation for End-to-End Paths," *Passive and Active Measurement Workshop (PAM)*, March 2003. (winner of best student paper award)

J.Navratil, R.L. Cottrell, "AbwE: A Practical Approach to Available Bandwidth Estimation," *Passive and Active Measurement Workshop (PAM)*, March 2003, moat.nlanr.net/PAM2003/PAM2003papers/3781.pdf.

M. Rabbat, "Multiple Sender Network Tomography," *MS Thesis*, Department of Electrical and Computer Engineering, Rice University, 2003.

A. Keshavarz-Haddad, "Effects of Traffic Bursts on the Network Queue," *MS Thesis*, Department of Electrical and Computer Engineering, Rice University, 2003.

M. Rabbat, R. Nowak, and M. Coates, "Network Tomography and the Identification of Shared Infrastructure," *Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, November 2002.

P. Abry, R. G. Baraniuk, P. Flandrin, R. Riedi, and D. Veitch, "The Multiscale Nature of Network Traffic: Discovery, Analysis, and Modeling," *IEEE Signal Processing Magazine*, May 2002.

M. Coates, A. Hero, R. Nowak, and B. Yu, "Internet Tomography," *IEEE Signal Processing Magazine*, May 2002.

S. Sarvotham, R. Riedi and R. Baraniuk, "Connection-level analysis and modeling of network traffic," *Proc. IEEE/ACM Internet Measurement Workshop*, San Francisco, CA, November 2001.

Y. Tsang, M. Coates and R. Nowak, "Network Delay Tomography", in IEEE Transaction of Signal Processing in Networking, Aug. 2003, Volume 51, Issue 8, pp. 2125-2136

M. Coates, M. Rabbat, and R. Nowak, "Merging Logical Topologies Using End-to-end Measurements," in Proceedings of ACM SIGCOMM Internet Measurement Conference, Miami Beach, Florida, October 2003.

M. Gardner, W. Deng, T. S. Markham, C. Mendes, W. Feng, and D. Reed, "A High-Fidelity Software Oscilloscope for Globus," *GlobusWORLD 2004*, San Francisco, CA, January 2004.

M. Gardner, W. Feng, M. Broxton, G. Hurwitz, and A. Engelhart, "Online Monitoring of Computing Systems with MAGNET," *IEEE/ACM Symposium on Cluster Computing and the Grid (CCGrid'03)*, May 2003.

M. Gardner, M. Broxton, A. Engelhart, and W. Feng, "MUSE: A Software Oscilloscope for Clusters and Grids," *IEEE Parallel & Distributed Processing Symposium*, April 2003.

W. Feng, M. Gardner, and J. Hay, "The MAGNeT Toolkit: Design, Evaluation, and Implementation," *Journal of Supercomputing*, Vol. 23, No. 1, August 2002.

J. Navratil, R. L. Cottrell, "A Practical Approach to Available Bandwidth Estimation," presented in *PAM'03 (Passive & Active Measurement Workshop)*, La Jolla, CA, pp. 1-11, April 2003.

R. L. Cottrell and C. Logg, "Overview of IEPM-BW Bandwidth Testing of Bulk Transfer Data, *SLAC-PUB-9202*, July 2003.

A. Kuzmanovic and E. Knightly, "TCP-LP: A Distributed Algorithm for Low Priority Data Transfer," *Joint Conference of the IEEE Computer and Communications Societies (IEEE Infocom'03)*, April 2003.

A. Kuzmanovic, E. Knightly, and R. L. Cottrell, "HSTCP-LP: A Protocol for Low-Priority Bulk Data Transfer in High-Speed High-RTT Networks," *Second International Workshop on Protocols for Fast Long-Distance Networks (PFLDnet'04)*, February 2004.

A. Kuzmanovic and E. Knightly, “Low-Rate TCP-Targeted Denial of Service Attacks (The Shrew vs. the Mice and Elephants),” *Proc. of ACM Sigcomm '03)*, August 2003.

A. Kuzmanovic and E. W. Knightly and R. L. Cottrell. “Bulk Data Transfer in High- Speed High-RTT Networks,” *Proceedings of the 2nd International Workshop on Protocols for Fast Long-Distance Networks*, February 2004.

V. Ribeiro, R. Riedi, and R. G. Baraniuk, “Optimal Sampling Strategies for Multiscale Models with an Application to Network Traffic Measurement,” *IEEE Statistical Signal Processing Workshop*, St. Louis, September 2003.

V. Ribeiro, R. Riedi, and R. G. Baraniuk, “ Spatio-Temporal Available Bandwidth Estimation for High-Speed Networks,” *ISMA 2003 Bandwidth Estimation Workshop (Best)*, CAIDA, La Jolla, CA, December 2003.

V. Riberio, R. Riedi, R. G. Baraniuk, “Spatio-Temporal Available Bandwidth Estimation with pathchirp,” *ACM SIGMETRICS*, 2004.

V. Ribeiro, R. H. Riedi, and R. G. Baraniuk, “Multiscale Queuing Analysis of Long-Range-Dependent Network Traffic,” submitted to *IEEE Transactions on Networks*, 2004.

S. Sarvotham, R. Riedi, and R. G. Baraniuk “Connection-level network traffic modeling: From network topology to traffic dynamics,” *Computer Networks Journal*, 2004 (invited).

S. Ayyorgun and W. Feng, “A Deterministic Characterization of Network Traffic for Average Performance Guarantees,” *38th Annual Conference on Information Sciences and Systems (CISS’04)*, March 2004.