SANDIA REPORT

SAND2006-4025
Unlimited Release
Printed July 2006

Modeling the 10-gigabit Ethernet ASC WAN

Lawrence F. Tolendino and Jason S. Wertz

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2006-4025
Unlimited Release
Printed July 2006

Modeling the 10-gigabit Ethernet ASC WAN

Lawrence Tolendino
Network Systems Design & Implementation

Jason Wertz
Advanced Networking Integration

Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-0806

Abstract

In recent years, modeling and simulation has played an increasingly important role in the maintenance of the nuclear stockpile. The Advanced Simulation and Computing (ASC) program continues to support and encourage the development of a modeling and simulation infrastructure to make these goals a reality. The Distance Computing Network has been making make the ASC resources available to users throughout the tri-lab environment for over five years. This network relies on the Transmission Control Protocol/Internet Protocol (TCP/IP) protocol suite to provide high performance and reliable communications. Understanding TCP/IP operation in this unique environment is critical. Software modeling has been used to analyze current network performance and predict the effect of proposed changes. Recently the network architecture was radically changed and the software model had to be changed as well. Whereas the original network was based on 2.5 gigabit per second ATM links, the redesigned network is comprised of 10-gigabit Ethernet links arranged as a 3-node ring. Therefore, a new software model was needed to continue to predict the performance of proposed changes and allow engineers to experiment with new network applications without the risk of interfering with critical operations.

Intentionally Left Blank

Table of Contents

Table of Figures

Intentionally Left Blank

Introduction

In recent years, modeling and simulation has played an increasingly important role in the maintenance of the nuclear stockpile. The interest in modeling and simulation of nuclear weapon performance was motivated by the Nuclear Test Ban Treaty and the desire to maintain a safe and effective nuclear weapon stockpile. The Advanced Simulation and Computing (ASC) program continues to support and encourage the development of a modeling and simulation infrastructure to make these goals a reality. From an infrastructure viewpoint the ASC effort has been focused on the supercomputers, visualization systems, and assorted support systems primarily located at the three main weapons laboratories: Los Alamos National Laboratory (LANL), Lawrence Livermore National Laboratory (LLNL), and Sandia National Laboratories (SNL). Part of the implementation strategy for these expensive hardware systems was to create an environment for sharing the resources throughout the weapons complex. The Distance Computing Wide Area Network (DisCom WAN) is part of the infrastructure that makes resource sharing possible.

The first incarnation of the network provided dedicated, high-speed bandwidth between the three primary weapons laboratories while utilizing the capabilities of SecureNet for administration and control.

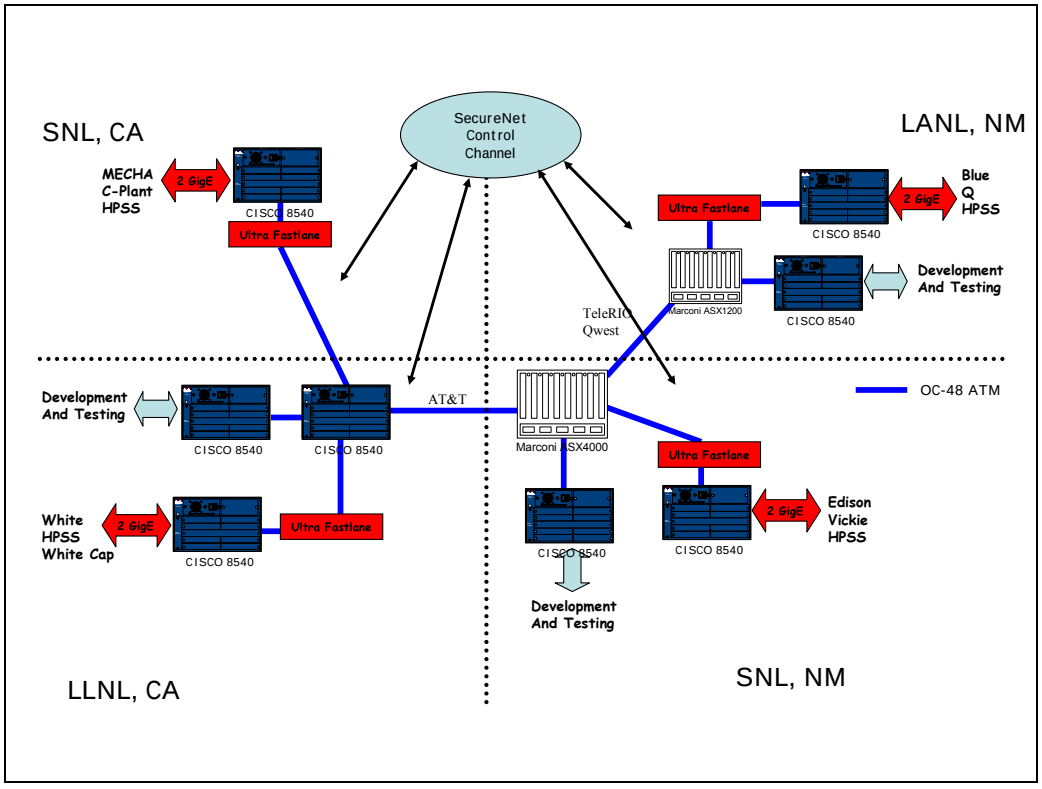

Figure 1: DisCom WAN Circa 2003

The DisCom WAN successfully provided users with remote access to ASC resources for over five years. The continual development and evolution of the WAN represented a challenge to the laboratories' staff because the WAN had taken on a critical production role in the ASC program. Making changes to the production network required great care and planning. Therefore, software modeling had been developed to analyze current network performance and predict the effect of proposed changes, see SAND2003-3760.

It was expected that the developing ASC tools and techniques would require increased network performance as they matured and future network upgrades would provide that increased bandwidth. The original network link (OC-48, 2.5 gigabits per second) between New Mexico and California was provided by a commercial carrier, AT&T, under a 5-year contract. As the contract was nearing its end, a new request for quotation was issued to continue to provide the link between the New Mexico and California laboratory locations. This request for quotations produced some surprising results because the cost for bandwidth and the carrier technologies available had evolved over the 5-year period of the initial contract.

Once the responses were evaluated it became clear that it would be possible to cut the cost of the network links drastically while increasing the link bandwidth to 10 gigabits per second. After careful evaluation by the three laboratories it was decided to contract for two links between New Mexico and California and configure the network as a ring. The ring configuration would provide the laboratories with increased performance and reliability at a reduced cost over the previous contract.

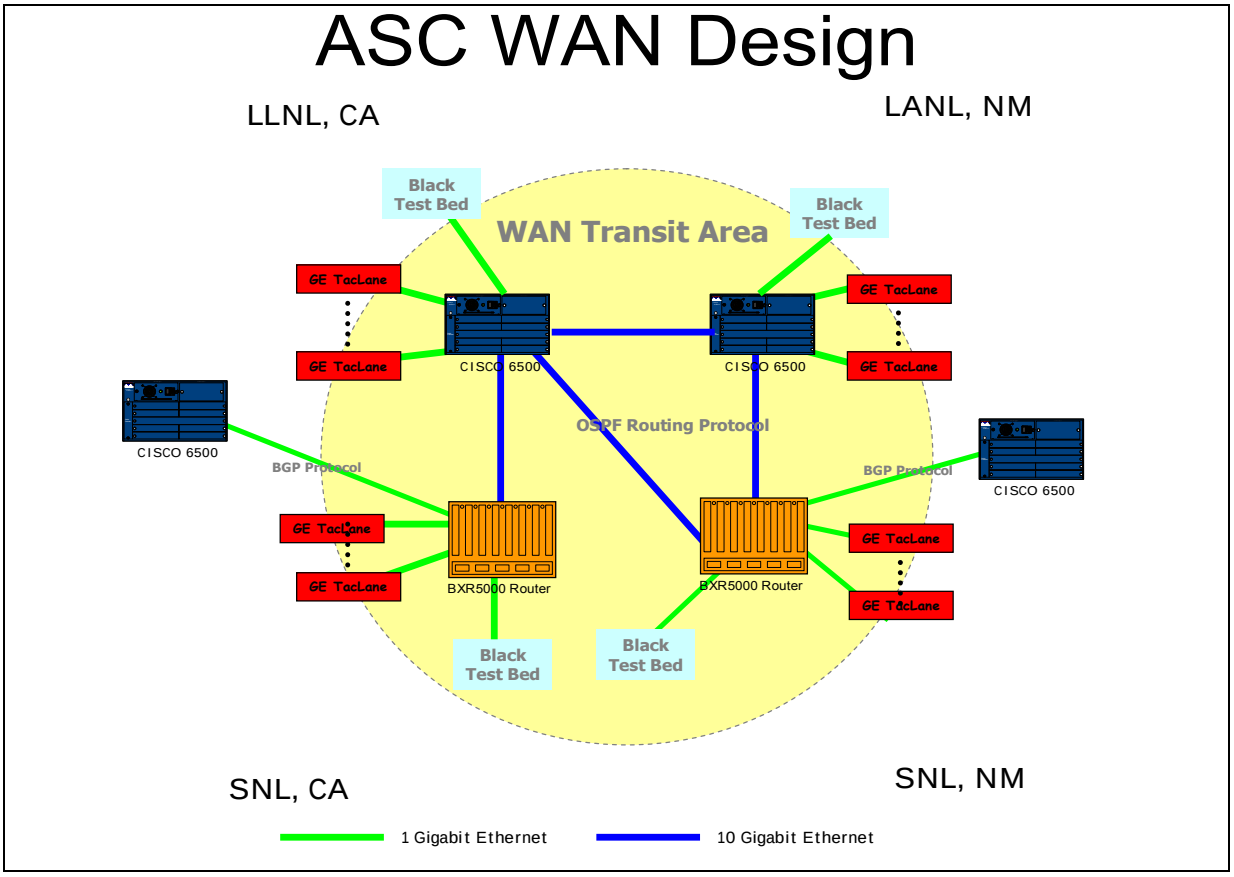

Figure 2: ASC WAN 2006

The decision to reconfigure the WAN as ring was a major architectural change as is illustrated in Figure 2. Instead of two simple point-to-point links with fixed routing, the network architecture is now a ring which utilizes 10-gigabit Ethernet (10GbE) links and the dynamic Open Shortest Path First (OSPF) routing protocol. The old DisCom WAN was based on OC-48 ATM links and ATM based encryptors while the new ASC WAN is comprised of 10GbE links configured as a ring and equipped with 1 Gigabit IP encryptors.

This network architecture change mandated a major update to the software model of the network. There are two reasons for this software update. First, the ring architecture and dynamic routing capability require a rewriting of the code to keep the model accurate. Second, the shear increase in bandwidth and robustness will encourage the migration of additional applications to the ring network. In order to continue to predict network performance as the network applications evolve requires an update to the software model.

The remainder of this paper reports on the changes made to the network model and initial results obtained. Throughout the rest of this paper, the first WAN that served the ASC community for five years will be referred to as the DisCom WAN while the new ring network will be referred to as the ASC WAN.

Opnet Modeling Experience

The Opnet Modeler product was used to model the individual links that made up the first Discom WAN. It was not necessary to model the entire network at once as the network functioned as independent links with fixed routing paths. These simple models were configured with models of routers, stand in models for encryptors, and links with appropriate bandwidth and propagation delays. The following figure shows one of the "simple models" that was used to evaluate the impact of two applications sharing a link between two ASC sites.

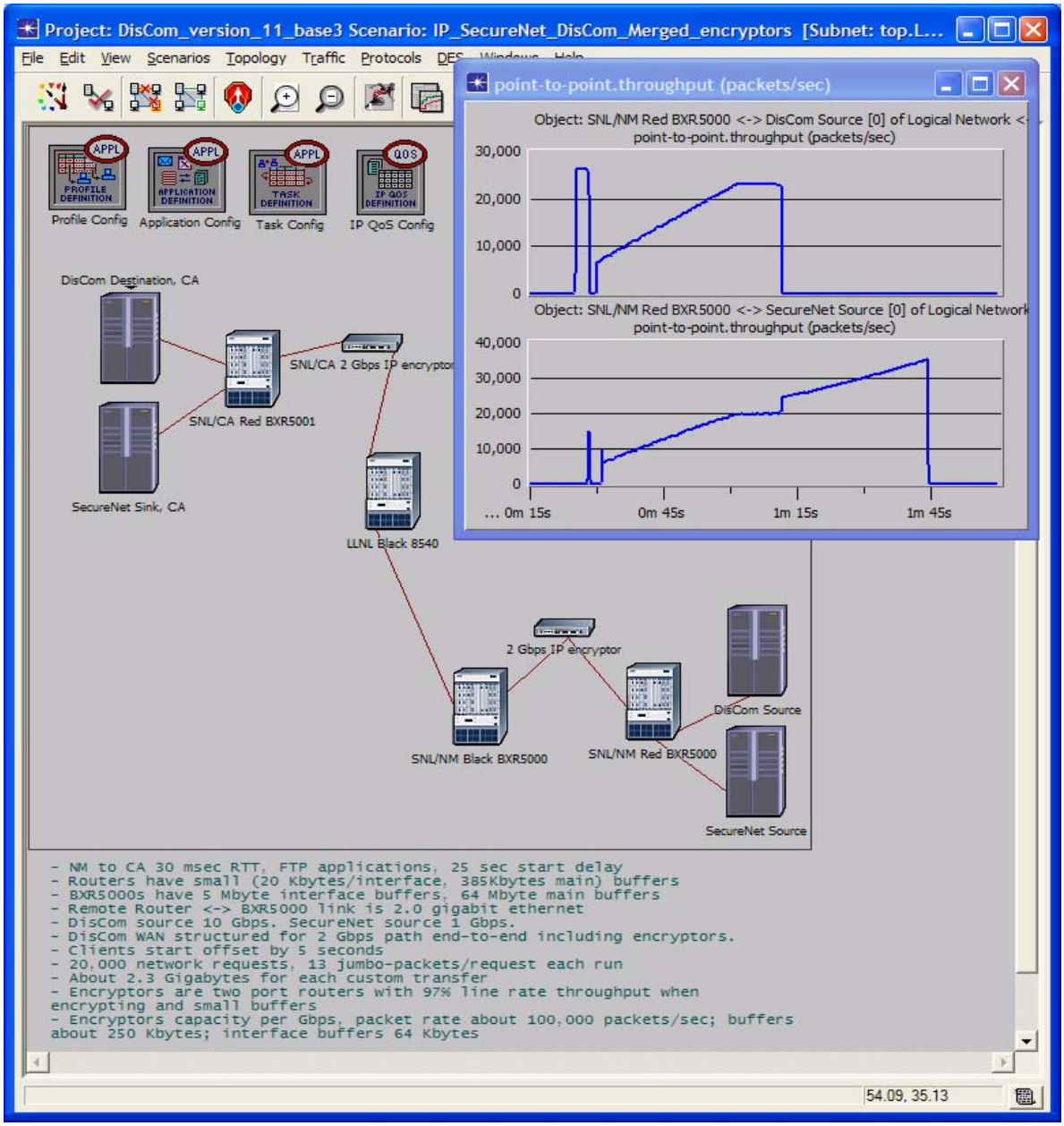

Figure 3: Link Model for DisCom WAN

This "simple" link model incorporated the key architectural elements of the DisCom WAN.

- Link throughput limited to 2 Gigabits per second,
- Fixed routing,
- Representative round trip times of 25 to 30 msec,
- Appropriate TCP software stacks and relevant TCP parameter settings (such as buffer sizes),
- Support for jumbo Ethernet frames (9,000 bytes), and
- The correct number of network elements.

The simple link model of the DisCom WAN provided performance estimates that could be verified by analyzing actual network performance. For instance, the size of the buffers required to achieve high performance in the model (at least 3.75 Megabytes for the TCP buffer) were validated in the operational network.

The benefits of having confidence in the model include the ability to play what if games with the network without impacting network operations. For instance;

- What if SNL/NM - SNL/CA SecureNet traffic was routed over the DisCom links? Would there be a noticeable impact on ASC data movements?
- What if HighSpeed TCP was implemented on the ASC hosts? Would HighSpeed TCP improve data movement performance?
- What if new routers were installed in the network that did not have large data buffers? Are large data buffers in the network routers necessary for good performance?

Modeling the data flows for SecureNet carried on the data links that serve DisCom data movements showed that the two could coexist, sharing the bandwidth fairly.

Our colleagues at Kansas State University created a model element for a variant of TCP, called HighSpeed TCP. Using this model element we can determine its effectiveness in the DisCom WAN environment without impacting actual network operations. Initial testing indicates that there would not be significant benefit to using the HighSpeed variant of TCP.

The size of the data buffers in the network routers proved to be a critical determining factor in the network performance. This was found to be especially significant in the routers that provided service between 10 Gigabit per second data paths and 1 Gigabit per second data paths for data circuits with large round trip times as exist between NM and CA (~ 30 msec). Without sufficient buffering data throughput dropped precipitously because of dropped packets and the behavior of the TCP congestion control algorithms.

New Opnet Model Design

The model of the ASC WAN was much more challenging than the previous link based model. No longer could each of the links be modeling separately since the ring network's operation changes dynamically based on the condition of the links or other network elements.

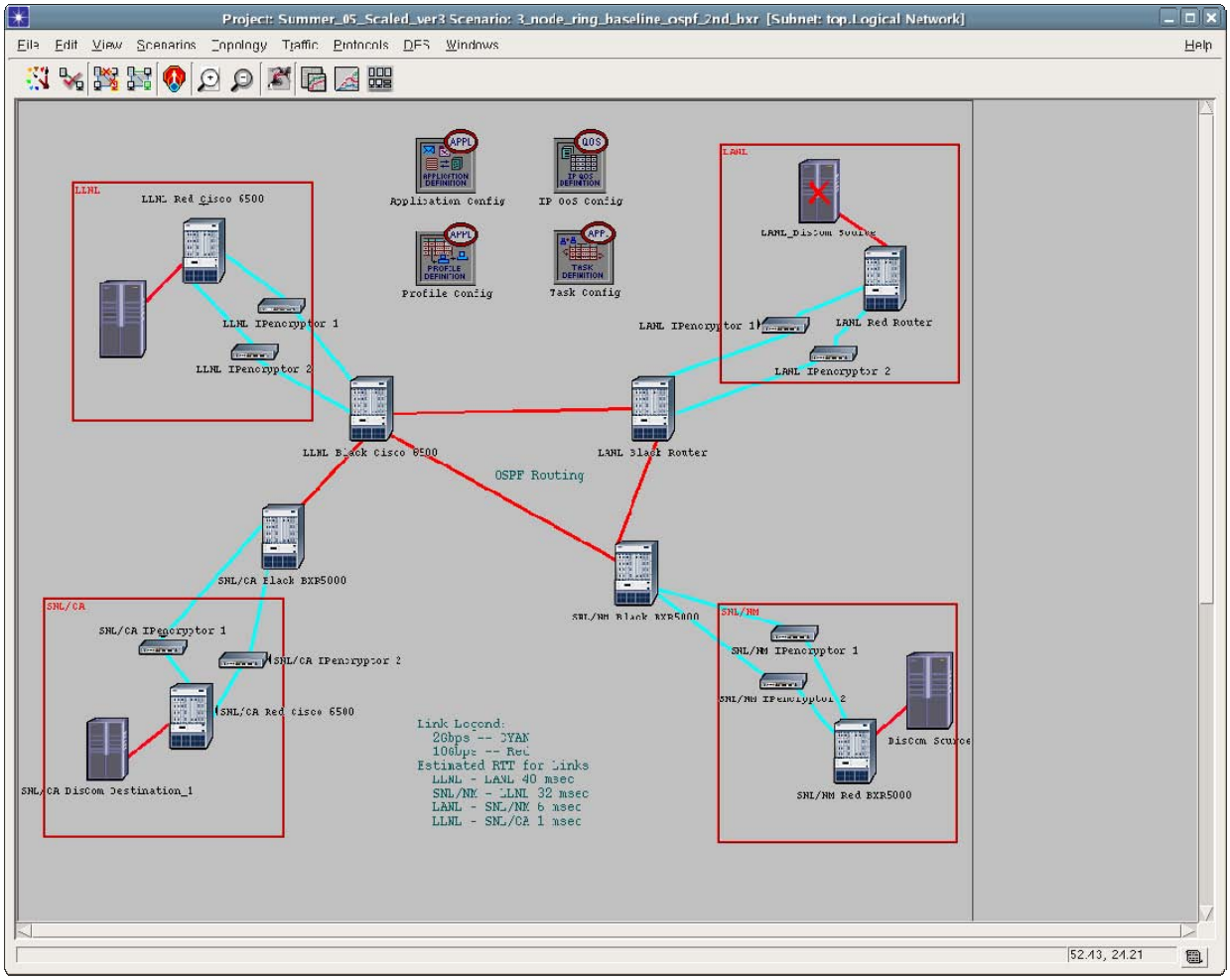

Figure 4: Opnet Model of ASC WAN Ring Network

The new Opnet model incorporated all the features of the previous model including Ethernet custom model elements such as jumbo frame support and the HighSpeed TCP model. To the set of features present in the simpler link model of the WAN, the Open Shortest Path First (OSPF) dynamic routing protocol was added to the model capabilities when the ASC WAN was modeled. In addition, the new model was configured with the actual network addresses used in the ASC WAN. All the data physical data paths are also included in the model as are all the basic network elements at each of the four sites. These elements include the paired IP encryptors and red and black routers at each of the four sites. In addition, the characteristics of the model elements such as the routers buffers were modified to more accurately reflect the type of router at each location. Including all these details in the model certainly increased the fidelity of the simulation. However, the memory used in this model grew to a very impressive size, easily outgrowing a PC configured with 1.5 gigabytes of RAM.

Indeed the 32-bit version of the Opnet Modeler application proved to be inadequate for this discrete event simulation model. The 32-bit application address space was just too small for this model with all its detail and would crash when modeling network operation lasting more than a minute or two. Development and execution of the discrete event model of the ASC WAN had to be moved to a host that supported 64-bit memory addressing. After considering several alternative operating systems, the Opnet modeling effort was transferred to a 64-bit Linux based PC with 4 Gigabytes of RAM installed. The 64-bit Linux system with sufficient RAM proved to be sufficient to run the model for any reasonable time interval.

As implemented on the 64-bitLinux system, the model features the following capabilities.

- Link throughput 10 Gigabits per second,
- Dynamic OSPF routing,
- Representative round trip times of 32 to 45 msec,
- Appropriate TCP software stacks and relevant TCP parameter settings (such as buffer sizes),
- Support for jumbo Ethernet frames (9,000 bytes),
- Optional HighSpeed TCP module,
- Parallel IP encryptors sharing the data transfer load, and
- The correct number of network elements.

Since the 64-bit Linux system that hosts the simulation model has a dual core processor, the Opnet parallelized execution option was examined. Parallelized execution adds considerable complexity to program execution since multiple events can be modeled simultaneously. That being the case, one must be very careful that such events do not have interdependencies. Only events without interdependencies can be modeled simultaneously without compromising the model's results. The primary variables for parallel simulations are the number of processors used and the 'parallel event execution time window' (PEETW). The number of processors was set at '2', which corresponded to the dual core CPU. This left the PEETW as the only real variable in the system. The PEETW is the period of time where each processor performs an independent task. When the PEETW timer is ended, the task results from the processors are combined and new tasks are assigned to each processor. If the PEETW is too short, the simulation spends its time syncing up the processes and the simulation takes longer to run. If the PEETW is too long, the processes don't sync often and the simulation runs quickly, but with potentially inaccurate results. Since the various tasks that can be assigned to the two processors are dependant on each other, the PEETW must be optimized to get the best results in the least amount of time.

Opnet ASC Ring Model Results

There are two main types of results for this model. There are feature test results and model test results. Feature tests are ones that test specific aspects of the model, such as OSPF load sharing or parallel simulation. Model tests are ones that test many or all aspects of the model at once.

One of the most important features of the new model was OSPF dynamic routing. OSPF provides the ASC WAN with fault tolerance and Equal Cost Multi Path (ECMP) routing capability. To test fault tolerance in the ring, simply fail a link on any of the three main border routers. What is expected is that traffic will be routed around the failed link. That is exactly what happened in the model. Figure 5 shows the link between LANL and LLNL. The blue line represents the traffic on the link when all the other links are up. The red line represents the traffic on the link when the link between Sandia/NM and LLNL fails. The traffic from the Sandia link is shifted to the LANL link when the Sandia/NM to LLNL links fails.

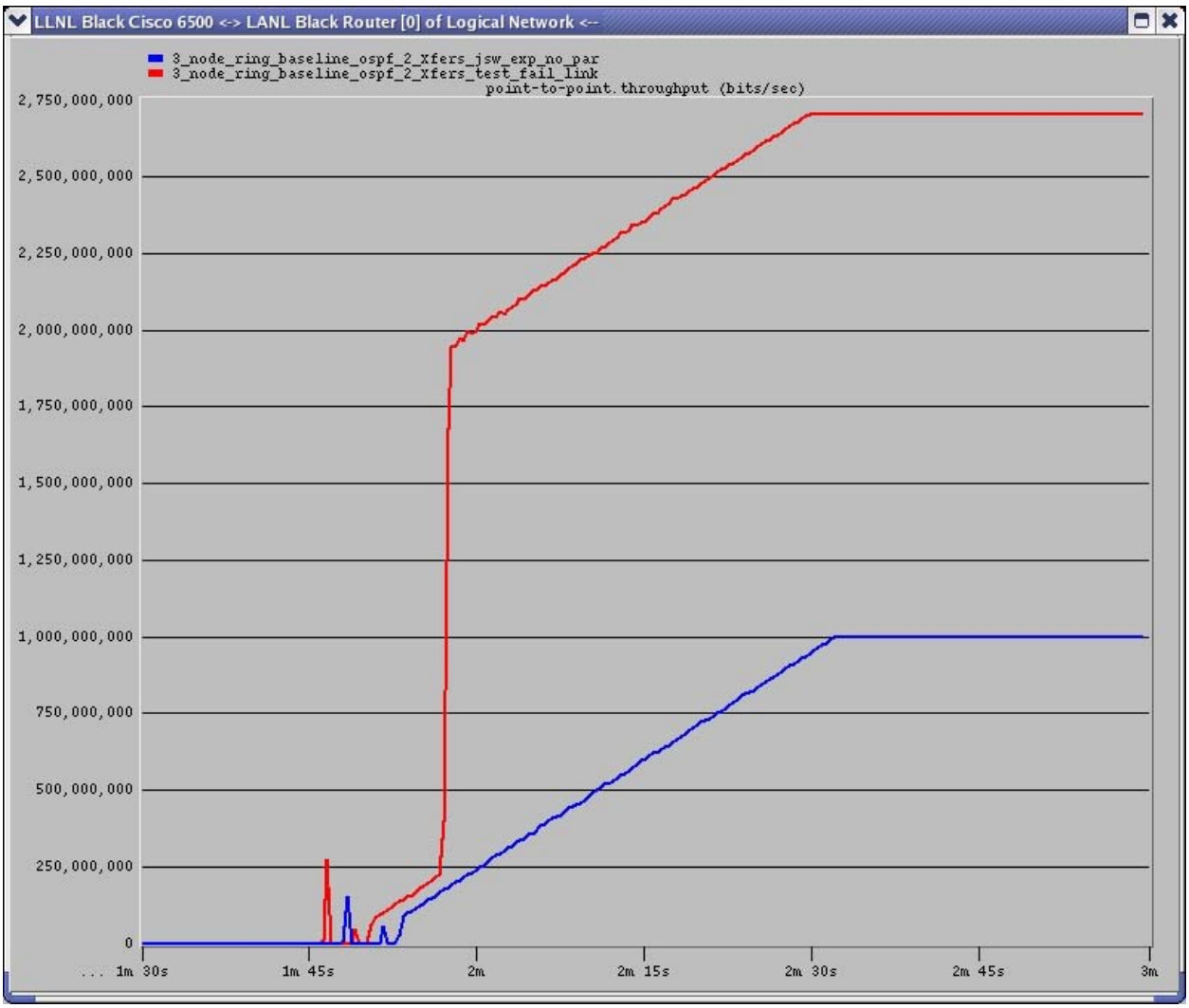

Figure 5: LANL to LLNL link, with and without failed link

The ECMP capability of the model can be demonstrated by running a single end-to-end flow across from the LANL source to the LLNL destination. The traffic passes across parallel encryptors at each end. If ECMP is working, the traffic will split almost evenly across the two encryptors. The model exhibits this behavior. Figure 6, shows the two LANL encryptors, each sending roughly half of the traffic flow.

Figure 6: Two encryptors sharing a load

Several attempts were made to implement a parallel simulation in Opnet. The results were not very good. The PEETW was adjusted across a wide range of values from 0 seconds (meaning no parallelization), to 1 second. The results were compared to a non-parallel simulation. The main factors in the comparison are the accuracy of the results and the simulation time. It was found that in order to get reasonable simulation results, the PEETW had to be set to be approximately .000001 seconds or less. Unfortunately, with such a short PEETW, the parallel simulation actually takes longer to run than a non-parallel simulation. The reason this happens is because the simulation spends a large amount of time syncing results between processors and not much time actually running the model. The conclusion that was drawn from these tests is that parallel simulations are not an effective way to run the model at this time.

For testing the full model, two main traffic flows were used (See Figure 7). The first was a flow from Sandia/NM to Sandia/CA. The second was a flow from LANL to LLNL. Both flows were generic TCP/IP traffic sent at approximately 2 Gbps. The flows both go through the LLNL Black Cisco 6500, but that router is big enough to handle both flows without slowing either down. The results from this are simply that the model predicts similar data rates (approx. 1.9 Gbps) as the real ASC WAN has been observed to have under ideal conditions.

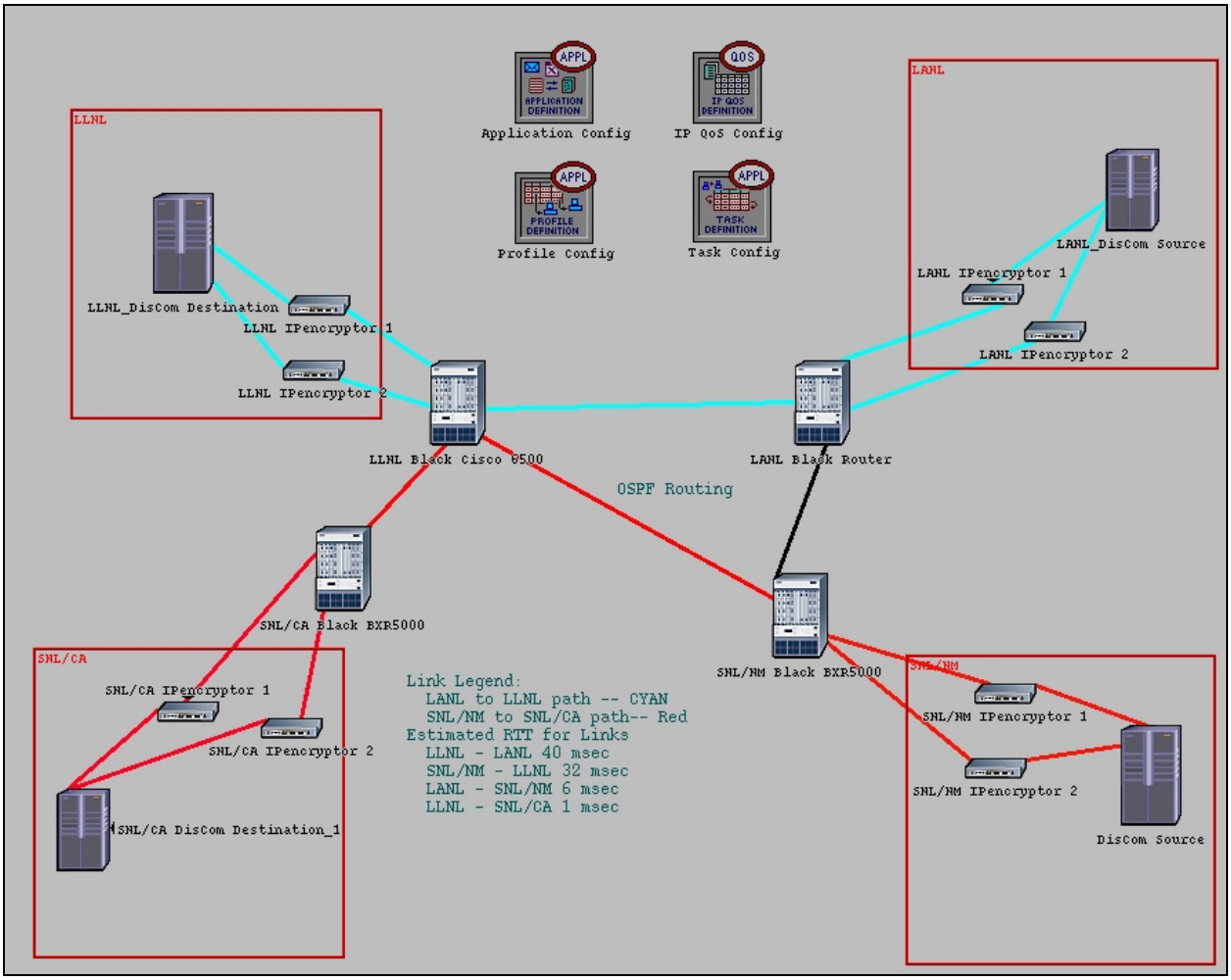

Figure 7: Data paths used to test model

A more interesting case for the ASC WAN model to look at is when additional real-time flows such as high levels of VOIP or streaming video are added. This demonstrates how well the system works under congestive or competitive conditions. As can be seen in Figure 8, adding a simple video stream on top of the data stream causes a serious degradation to the data transfer. It also impacts the video stream by introducing jitter and delay.

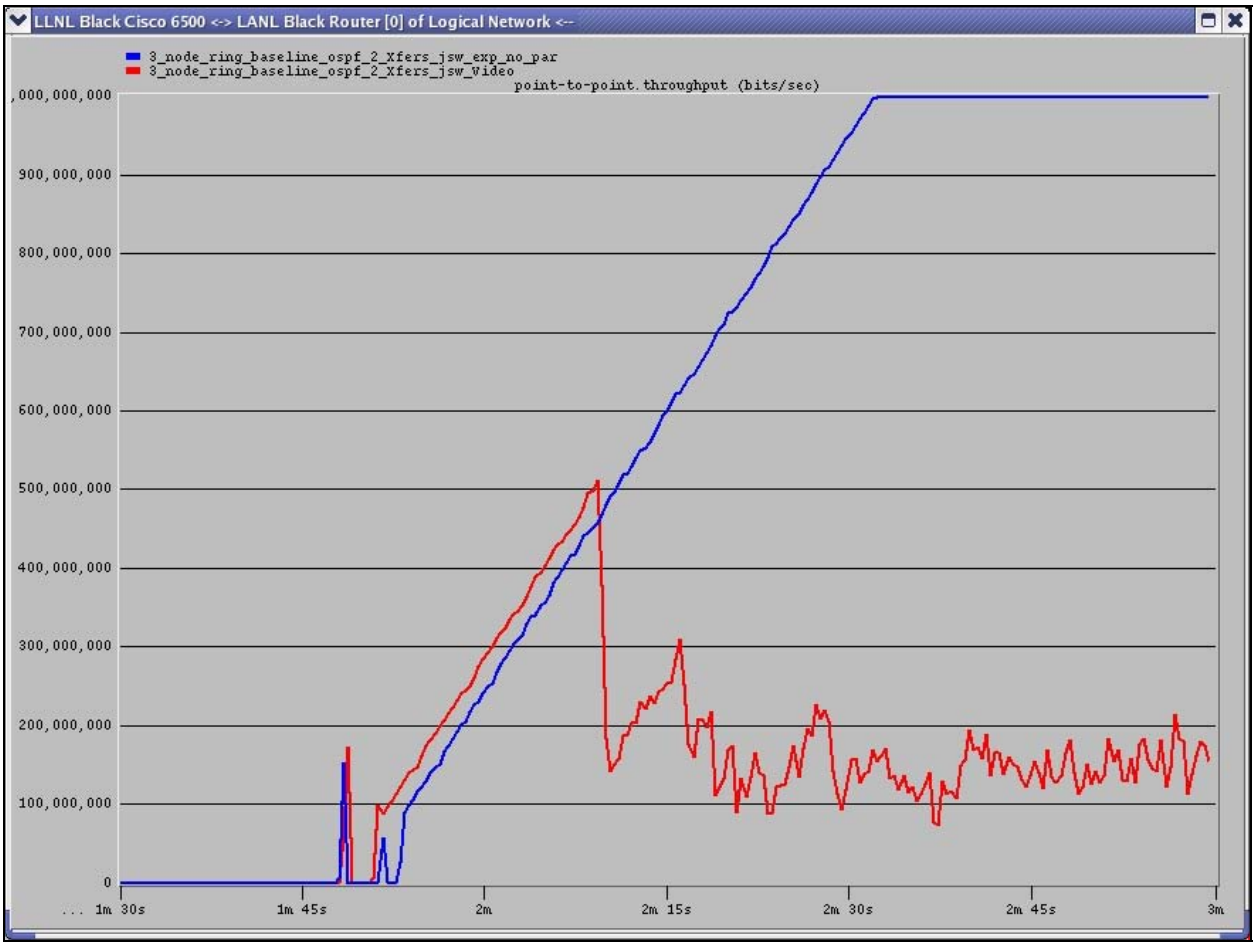

Figure 8: Data stream with and without video stream

--The blue line represents a simple data stream

--The red line is a data stream and a video stream competing for bandwidth.

Conclusion and Future Directions

The development of the ASC WAN model and updating it to support the new architecture has certainly been a worthwhile effort that has provided a useful tool as well as providing insight into the issues surrounding our network environment. The model output has been compared to past model results and the laboratory test results obtained over the last few years. Basic performance measures such as throughput over Gigabit per second data links that span continental distances are consistent with actual network test results. Various ASC WAN operational characteristics such as load sharing, alternate path routing, TCP performance, and router characterization have been reconciled with the Opnet Modeler results giving us confidence in using the model.

One of the first alternate uses of the ASC WAN envisioned by some members of the community is using the WAN as transport for classified video teleconferencing. Initial results show that it will be challenging to combine time critical data streams, such as video conferencing, with the typical ASC data movements. If the two data streams are simply combined without engineering the performance of both data streams are severely impacted. The performance would be unacceptable to both types of customer. It will take careful engineering and the implementation of some form of quality of service to allow this combination to perform acceptably. Fortunately it appears that the ASC WAN model will aid this development.

The model has also been used to examine the impact of combining the SNL/NM to SNL/CA SecureNet data flows with ASC data flows. If both flows are active simultaneously, model results show that the two data flows share the available bandwidth fairly and efficiently. When only one data flow is active, it receives the entire bandwidth available. Therefore, routing some SecureNet traffic of over the ASC WAN is a cost effective action that provides improved performance for SecureNet while helping to justify the investment in the ASC high bandwidth links.

Our collaboration with KSU has produced some useful adjuncts to the basic Opnet Modeler capabilities. The HighSpeed TCP module has allowed the examination of alternative TCP implementations in the ASC WAN environment. Using the module in various scenarios allowed the efficacy of HighSpeed TCP to be evaluated without impacting the production network. Initial results indicate the HighSpeed TCP would not significantly improve throughput in the ASC environment. The collaboration is now pursuing the SCTP protocol to evaluate its potential impact on ASC supercomputing in the WAN environment.

While the model has performed adequately and provided useful results, there are some serious concerns. First the model has apparently reached the limits of the hardware and operating systems. Even though the model looks simple, when it is implemented in a discrete event simulation tool such as Opnet Modeler, there are millions of objects to keep track of. Each packet in each data flow is tracked through each of the network elements and all of their software processes. It is likely that nature the modeling effort will have to change if we wish to make keep the model relevant.

One of the options being considered is to model some of the data flows as bulk activities creating a level of background against which to model the data flow of interest. This option will require some additional learning and a period of development before any attempt is made to incorporate this capability in the ASC WAN model. First, to successfully implement this alternative requires learning how bulk activities are implemented in Opnet Modeler. Next some of the data movements in the ASC WAN will have to be studied and characterized so they can be represented by a bulk model. If these two efforts are successfully, then we may be able to use discrete event simulation on data flows of interest and produce useful results without such using as much memory as required in the past efforts.

Another area of development must be to execute models in parallel. The chip industry is rapidly developing multi-core CPUs and four core units will be available within the year. Initial efforts have not produced noticeable performance increases utilizing the current dual core CPUs. However, our efforts have only scratched the surface of parallelizing the ASC WAN model. Given the imminent availability of four core CPUs and the potential of increased performance, additional efforts must be made in parallelizing the model.

Further development is also warranted in the refinement of some network element models. It is likely that the ASC WAN network will implement four parallel encryptors at each site to improve link throughput. Our current encryptor model is a simple router with appropriate buffer sizing. Opnet indicates that a type one IP encryptor model will be available to government customers with the next release of the product. Upgrading the Opnet Modeler software to include this new model certainly seems worthwhile considering the central importance of the encryptors in increasing network performance.

Distribution

1	MS 0136	G.E. Connor, 4333
1	MS 0630	N.A. Marsh, 4601
4	MS 0788	M.J. Benson, 4334
1	MS 0788	J.H. Maestas, 4334
1	MS 0788	P.A. Manke, 4338
1	MS 0788	V.K. Williams, 4334
1	MS 0788	M.A. Rios, 0788
1	MS 0795	P.C. Jones, 4317
1	MS 0801	R.W. Leland, 4300
1	MS 0801	D.S. Rarick, 4310
1	MS 0801	D.R. White, 4340
1	MS 0805	W.D. Swartz, 4329
4	MS 0806	Len Stans, 4336 (4)
1	MS 0806	J.M. Eldridge, 4436
1	MS 0806	S.A. Gossage, 4336
1	MS 0806	T.C. Hu, 4338
1	MS 0806	C.M. Keliiaa, 4336
1	MS 0806	B.R. Kellogg, 4336
1	MS 0806	J.H. Naegle, 4336
1	MS 0806	T.J. Pratt, 4338
6	MS 0806	L.F. Tolendino, 4334
6	MS 0806	J.S. Wertz, 4336
1	MS 0813	G.K. Rogers, 4312
1	MS 0813	R.M. Cahoon, 4311
1	MS 0823	J.D. Zepper, 4320
1	MS 1393	J.A. Larson, 3820
1	MS 9012	C.T. Deccio, 8949
1	MS 9012	R.D. Gay, 8949
1	MS 9151	K.E. Washington, 8900
1	MS 9151	C.T. Oien, 8940
1	MS 9158	H.Y. Chen, 8961
2	MS 9018	Central Technical Files, 8944
2	MS 0899	Technical Library, 4536