COMBINE: COmputational Modeling of BIg NEtworks

Report of Contributions

https://indico.fnal.gov/e/combine

Contribution ID: **0** Type: **not specified**

Large Scale Network Simulation Methods

Tuesday, 11 September 2012 09:58 (12 minutes)

Large Scale Network Simulation Methods

Dr. Riley has been working in the field of large-scale network simulation since his PhD thesis under the direction of Dr. Richard Fujimoto and Dr. Mostafa Ammar. In that work he was the first to show that distributed simulation techniques using conservative synchronization protocols could be applied to the popular ns-2 network simulation tool which resulted in the developemnt and release of "pdns".

However, just breaking a large topology into different logical processes proved to be insufficient for achieving larger simulations. He discovered that routing table models consumed large amounts of systems memory, growing in proportion to N-squared (where N is the total number of network nodes). This problem was alleviated using a novel approach for packet routing called "NIx-Vector" routing, which was applied to the ns-2 distributed simulation and resulted in successful experiments with more than 100,000 nodes.

The lessons learned in the distributed ns-2 work were then applied to a new network simulation environment developed by Dr. Riley called the "Georgia Tech Network Simulator" (GTNetS), which was designed from the outset for scalability and efficiency. Using a large number of CPUs on the Pittsburgh Supercomputer Center, GTNetS was successfully demonstrated to execute a network topology of more than 1 million network elements. This was the first ever simulation experiment with packet-level detail that exceeded the 1 million node threshold.

More recently, Dr. Riley is co-PI on the new ns-3 network simulator development effort. Again, ns-3 has been designed to support distributed simulation using conservative synchronization protocols. The inter-process communications in ns-3 is performed using the ubiquitous MPI message passing interface. Researchers at the Army Research Lab in Aberdeen Maryland have used ns-3 to model networks of more than 100 million network elements, clearly two orders of magnitude larger than has been possible previously.

Primary author: Dr RILEY, George (Georgia Institute of Technology)

Presenter: Dr RILEY, George (Georgia Institute of Technology)

Contribution ID: **3** Type: **not specified**

Internet map and simulation

Tuesday, 11 September 2012 13:05 (12 minutes)

Many networking problems and solutions are distributed and collaborative. To investigate them we need a realistic, detailed model of the Internet and ideally a simulator that would work with that model to simulate problems/solutions of interest at just the right level of granularity. In my prior research I've worked on creating detailed models of Internet routing, address space, traffic, communication patterns and vulnerability distribution so I could evaluate worm and spoofing defenses. In all cases accurate models made a world of difference, but getting them took months to years. Such work is an overhead for researchers that do not focus on Internet mapping/simulation - they can either lose time on it and pass up research opportunities or they can adopt naive models and reach wrong results.

My future research consists of two parts: an Internet map that can easily integrate data from multiple sources and export it and a distributed network simulator that can interact with the exported Internet model in a customizable manner, and whose level of simulation granularity and details can also be customized. I've done some preliminary work on hard questions, such as how to combine data of different granularity, how to infer missing data or fix incomplete data, how to achieve simulation at different levels of granularity, etc.

Primary author: Dr MIRKOVIC, Jelena (USC/ISI)

Presenter: Dr MIRKOVIC, Jelena (USC/ISI)

Contribution ID: **4** Type: **not specified**

Fermilab Network Modeling & Simulation Efforts

Tuesday, 11 September 2012 14:48 (12 minutes)

We have two previous projects that are related to network modeling and simulation. In these projects, we used the modeling & simulation methodology to study and analyze network functions and performance.

(1) The Modeling Process and Analysis of Virtual GMPLS Optical Switching Routers

Generalized multi-protocol label switching (GMPLS) has emerged as a very promising protocol technology for the next generation optical networks. GMPLS successfully combines the best features of IP and optical networks in terms of quality of service (QoS), privacy, flexibility and scalability. GMPLS introduces enhancements to the existing IP routing and signaling protocols by supporting not only networks that perform packet switching (IP), but also networks that perform switching in the time (TDM), wavelength (DWDM), and space domain (circuit switching). This project designed and implemented a modeling tool for analysis of GMPLS optical switching routers (GOSR). A model of the GOSR has been built using OPNET modeling and simulation platform, in lieu of a prototype. The virtual model contains all the necessary GMPLS functions of an optical backbone router. The virtual model of the GOSR has the capability of giving a more integrated and realistic simulation on wavelength routing, wavelength assignment, wavelength switching, dynamic label switching path (LSP) setup and tear down, and blocking mechanism of GMPLS light paths. The OPNET process modeling methodology was used to develop the virtual GOSR models. The simulation results obtained include the blocking rate, OSPF-TE bandwidth analysis, and CPU utilization. The modeling environment developed in this project provides a simulation platform for further development and future enhancement of GMPLS protocols, routing protocols, and optical switching router implementations.

(2) Development of Modeling Cyber-attack Scenarios for JTRS IW Threat Analysis Using OPNET Modeler

The Joint Tactical Radio System (JTRS) program intends to develop a family of communications devices that support reliable multi-channel voice, data/imagery, and video communications for battlefield communications. The JTRS devices use Software Defined Radio (SDR) technology based on the Software Communications Architecture (SCA). Ultimately JTRS will extend the information infrastructure supporting Network-centric Warfare to the last "tactical" mile, enabling the warfighter to gain access to real-time information. Under the JTRS program the role of the radio changes from network "terminal" to network node. This role includes behaviors similar to a router in a packet-switched network based on Internet protocols. Therefore, these network nodes will be susceptible to the types of attacks generally referred to as "cyber-attacks" that generate information processing faults. In this project we investigated classes of cyber-attacks with respect to the feasibility of modeling attack behaviors in OPNET Modeler, as well as the development of an attack library. The purpose of the development was to validate and demonstrate the effectiveness of the JTRS mobile nodes under a wide range network attacks. We specifically developed generic models for ICMP-based Distributed Denial of Service (DDOS) and attacks oriented to MANET protocols. The design objectives were structured around a concept of a "fault insertion node" that can incorporate fault functionality based on pluggable "fault modules" with configurable interfaces. A fault insertion node is capable of generating the necessary communication fault insertion frames generated as a function of fault modules hosted within the node. Fault insertion node modules were developed with the Ethernet data link interface; IEEE 802.11b Ad-hoc interface; and a capability to incorporate other link interfaces.

Primary author: Dr WU, Wenji (Fermilab)

Co-author: DEMAR, Phil (Fermilab)

Presenter: Dr WU, Wenji (Fermilab)

Contribution ID: **5** Type: **not specified**

High Performance Network Modeling in the US Army MNMI (Mobile Network Modeling Institute)

Tuesday, 11 September 2012 11:13 (12 minutes)

The Mobile Network Modeling Institute (MNMI) at the Army Research Laboratory (ARL) has been established to apply High Performance Computing (HPC) tools and techniques to improve the Army's ability to study large-scale, mobile, ad-hoc networks. Current modeling technologies offer a wide range of radio models, but fail to provide scaling and performance at an acceptable level of fidelity. Due to the complexity of the military waveforms involved, it is critical to accurately model the entire network stack with specific attention to the ad-hoc routing protocols and RF-propagation effects of the environment. This involves significant computational load that can exploit parallelism of supercomputers to achieve a sufficient scaling and performance improvements over current serial methods.

The Army wishes to study the performance of networks for a variety of reasons including for research, deployment planning, procurement, and to improve "Network-Centric Warfare". This involves a combination of pure simulation, as well as emulation and live experimentation. Abstract models usually associated with simulation can be replaced with high fidelity implementations of network protocols and applications which sometimes share a code base with real systems. Constructed environments are used to combine live systems with network simulation and emulation providing scaling to live tests that maintain a true user experience while enabling the network and attached devices to experience a more complete exercise.

The MNMI has been working with the ns-3 open source network simulator to exercise its scaling and performance capabilities at the ARL DoD Supercomputing Resource Center (DSRC) [http://www.arl.hpc.mil]. Experiments on cluster platforms have demonstrated scaling of simple network simulations to the order of 300 million simulated nodes. Performance analysis of various network configurations with ns-3 has produced promising results on the order of 400,000 packet receive events per wall-clock time using hundreds of processors. Other novel techniques are being researched such as using co-simulation via distributed shared memory (coupling of orthogonally-decomposed codes), a real-time distributed scheduler, and off-loading computationally expensive operations to Data Parallel Processors (e.g. GPUs). The MNMI continues to work with its partners and the ns-3 community to evaluate, improve, and extend the capabilities and performance of the simulator.

Network emulation work in the MNMI is centered on the Extendable Mobile Ad-hoc Network Emulator (EMANE) software which is developed by Cengen [http://labs.cengen.com/emane] and funded by the Army Research Lab and Naval Research Lab. With EMANE, the physical and link layers are modeled in software while the rest of the network stack is executed natively in live systems. This is usually instantiated as virtual machines running on top of simulated network interfaces in a cluster. Unmodified applications and operating systems can be run over network models allowing user interaction and traditional network measurement tools to be used. With recently-acquired, dedicated hardware, emulations of more than 5000 nodes are anticipated while maintaining real-time RF propagation calculations.

Both ns-3 and EMANE have been used to stimulate live exercises aimed at evaluating emerging communications technologies. The C4ISR-Network Modernization exercises held at Ft. Dix, NJ offer a "sand box" to test individual communications technologies and Systems-of-Systems network architectures. Limited resources are available for live asset testing, so virtualization plays a key role in developing a realistic environment for product evaluation. OneSAF (semi-automated force modeling software) is traditionally used to populate the battlefield with virtual entities but has limited fidelity in its communications modeling. The MNMI has used its resources and capabilities to augment the OneSAF presence to construct an interactive virtual network environment, giving the systems under test a more realistic network and application loading. This work has led to the

development of a real-time distributed scheduler for ns-3 which allows live simulations such as these to scale using distributed computing resources.

Future plans for the MNMI include the porting of additional military radio waveforms to the simulation and emulation platforms of choice as well as additional performance enhancements. We are currently pursuing the coupling of multiple simulation engines which will allow each to decompose large problem sets independently to take advantage of parallelism inherit to their roles. For example, a forces modeling code may decompose a large simulation along organizational boundaries for parallel processing, while an RF propagation code will benefit from decomposing the problem geographically. The engines will be coupled together via Distributed Shared Memory for sharing data and an RTI for coordinating time.

Primary author: Mr RENARD, Kenneth (Army Research Laboratory)

Co-author: Mr CLARKE, Jerry (Army Research Laboratory)

Presenter: Mr RENARD, Kenneth (Army Research Laboratory)

Contribution ID: **6** Type: **not specified**

Evolution of Networking Research into a Science

Tuesday, 11 September 2012 11:52 (12 minutes)

Evolution of Networking Research into a Science

Introduction
Building COmputational Models for BIg NEtworks will consist of developing a complex set of models, both in the vertical and the horizontal space as defined in the call for abstracts. Among the large number of variables that must be considered, a major component is network traffic – how it has evolved over the years, what drives it, and how can we predict traffic patterns, if at all, and thus build future models for network traffic. Understanding the characteristics of network traffic, modeling traffic and studying the effects of different models of traffic workloads on network and application performance has been a major thrust of our research. We have mainly studied traffic models as a means to study better experimental methods for networking research. Our broader and longer-term vision is to help evolve networking research from its current adhoc status into a science.

Background
Experimental methods for networking research have evolved almost ad-hoc over the last decade. That there is no consensus among the research community on best practices in empirical research is only part of the problem. The fundamental problem that calls to question the results of any empirical research is this – as a nascent field, empirical networking has yet to develop into a science. In the physical sciences, experimentation and evaluation of new ideas has been honed so that experiments conducted by one research group can be repeated and tested by another. Due to lack of such coherent standards in experimental methods, this is simply not possible in networking research today. Even within the wider field of computing for example, there are well-known benchmarks for testing new processors against existing ones. In most fields, benchmark standards test new inventions. For example, if Intel develops a new processor, several benchmarks test the new processor to demonstrate that it performs better than an existing one. This is possible because there are agreed-upon scientific processes for experimentation and evaluation in that field.

```
For over a decade now, networking researchers have built small and large research testbeds usin
```

Despite the significant advances made in development and deployment of very large testbeds and the software framework for building different topologies and reserving these resources to run large-scale experiments, there is still little understanding of what constitutes best experimental practices. There are no agreed-upon scientific methods for running experiments, no benchmark workload models, no network emulation standards. The networking research community lacks a coherent, shared view of best practices for experimental methods for networking research. Even today, there are no agreed-upon research methods or standard practices for maintaining traffic datasets, generating traffic, emulating network characteristics, or designing and running experiments. Hence while many researchers propose new and improved protocols for improving our cyber-infrastructure, real progress and deployment of the best protocols is slow.

Research Questions
However, computer networking, as a nascent field with explosive growth, woefully lacks such benchmarks and standards. Establishing such benchmarks remains a challenging research endeavor, and it forms the central motivation for our research. We need to develop models for several components of experimentation which include workload modeling, network path emulations, network topologies, measurement methodologies, and determining which performance metrics best describe the outcome of experiments. Before we develop any models, however, there has to be some agreement in the scientific community of certain standards.

Several research projects have chipped away at this problem of lack of scientific methods for net-

working research for some years now, but the field is fragmented at best. It is time to take this research to the next level by integrating the research of experimental methods in networking in a fundamental way to create a holistic view.

Some of the research questions that we're grappling with lately have been the following:
– Can we develop a set of necessary and sufficient conditions for running a successful experiment for empirical networking research? For example, topology of testbed, duration of experiment, number of runs for each experiment, source of input traffic, mix of traffic, workload model, network emulation, and measurement methodology. How can we build models if we do not have comprehensive understanding of how each component works and how they all interact to form this complex system?
– Could we classify experiments into sets, based on the input, the output, or the goal?
– Can we develop scientific methodology for calibration of a network for experimentation? We already do this in the lab, but can we generalize calibration for any network?
– Run the same experiments using different sources of traffic to determine if there are some invariants in traffic generation and experimentation and what are the variants?
– How can we reproduce and validate experiments conducted by a different research team?
– How can we develop "experimental methods in networking research" into a science?
Given our research interests and the work we have done in this area, we completely appreciate the dire need to get out of the reductionist stance that we, as a community, have taken in the so-far pseudo-scientific investigation of networking as a field. It brings to mind the blind men and the elephant, and an evolving elephant at that. We have decided that the problem is so complex that we shall simply take one slice of it and analyze that slice independently rather than deal with the very difficult problem of creating holistic models to solve problems in this space. Yet, we do remain skeptical about building such models because networks seem to be evolving so rapidly and they are such complex systems. However, it is all man-made and we are driving its evolution. So surely, if we employ large enough resources to this problem and go at it systematically and globally, we should be able to create reliable and evolving models for networks?
My interests and expertise thus form a slice of the holistic modeling that this workshop seeks to attempt to motivate. My own motivation in attending this workshop would be to lend my expertise and collaborate with this group so we might be able to work toward constructing such large-scale models for networks that would encompass all components emulating the vertical and horizontal spaces while being able to evolve over time.

References
– J. Aikat, S. Hasan, K. Jeffay, and F. D. Smith, Towards Traffic Benchmarks for Empirical Networking Research: The Role of Connection Structure in Traffic Workload Modeling, IEEE MASCOTS (Modeling, Analysis, and Simulation of Computer and Telecommunication Systems), Washington DC, August 2012.
– J. Aikat, S. Hasan, K. Jeffay, and F. D. Smith, Discrete-Approximation of Measured Round Trip Time Distributions: A Model for Network Emulation, GENI Research and Education Experiment Workshop 2012 (GREE12), Los Angeles, CA, March 2012.
– J. Aikat, K. Jeffay, and F. D. Smith, Experimental Methods for Networking Research and Practice, SIGCOMM 2011 Education Workshop, Toronto, Canada, August 2011.

Primary author: Dr AIKAT, Jay (UNC-Chapel Hill)

Co-author: Dr JEFFAY, Kevin (UNC-Chapel Hill)

Presenter: Dr AIKAT, Jay (UNC-Chapel Hill)

Contribution ID: **7** Type: **not specified**

Architecting and Operating Energy-Efficient Networks

Tuesday, 11 September 2012 13:56 (12 minutes)

Primary author: MONGA, Inder (Lawrence Berkeley National Lab)

Presenter: MONGA, Inder (Lawrence Berkeley National Lab)

Contribution ID: **8** Type: **not specified**

A new paradigm for network simulations; model-checking meets event simulation

Tuesday, 11 September 2012 13:30 (12 minutes)

For large-scale system simulations, two main components need to be developed; the problem model, and the simulation model. For basic sciences, the models always reflect the physical phenomena, and the challenges arise from numerically implementing those models in computational environment, and then verifying the physical phenomena. In networking, or the Internet in particular, the modeling problem is still an open question. Finding the correct interaction model between network components is an open research question, especially with the increase of complexity and types of components. The simulation component of different models can then follow.

We propose using formal analysis to first model large networks (the model), then leverage temporal model-checking approaches to simulate the dynamic network behavior over time (the simulation). In doing so, we propose two stages for network modeling:
- Configuration Modeling: This covers policies (routing, security, QoS, …) as well as topology information and high level applications. Developing models across communication layers while taking care of topology will integrate both vertical and horizontal understanding of network operations.
- Network State Modeling: This covers dynamic behavior of physical network components, including link behavior (capacity, connectivity, quality, etc)

The Model

We have used model-checking for network configuration analysis for performing large-scale network verification.
We will leverage our experience to extend the models, add more network components, integrate real-time processing and enable large-scale network simulation. Also, a vital addition is distributed analysis of such models. This section provides a summary of our previous work on model-checking as a means for network analysis.

The problem of model checking a multi-faceted system (i.e., multilayer multi-device network) can be broken down into two main subproblems: 1) How to merge these heterogeneous systems into one monolithic framework, and 2) what is the system state upon which we can define transitions and build the model-checker?

When it comes to modeling a multi-component system, the problem of finding a middle ground, or a common representation becomes extremely important. Utilizing highly specialized data structures, or representation model will work for one layer but will break the other (or at least becomes highly inefficient). To address this problem, one can revert to one of the basic structures: sets/collections, Boolean expressions, formal grammar description, etc. Choosing a grammar description will change the focus of the work from model-checking the system to designing a more complete (and more complex) language. This will never be flexible or efficient enough for large scale analysis. Using basic sets while is very flexible, it is far from being scalable without using a symbolic representation. Boolean expressions comes as a plausible solution providing both: simple set-like operations, as well as having many very efficient practical implementations.

The first step in analysis/modeling the network will start by digesting all the information and policies of multiple device types and compile them into basic expressions. Every predicate a network device defines can be simply written as a Boolean expression. The problem now becomes one of defining the variables and labels upon which such expressions are built. It is important to mention that using such a generic representation enables the complete separation of device specifics and syntax from the actual analysis as long as the settings, policy and status got mapped into a Boolean expression.

The other part of the problem is modeling the system/network state. Let us start with a domain specific assumption: packets move through the network faster than the network configuration and layout can change. This assumption will lead us to define the system state from a packet-centric perspective rather than a network one. In other words, the state space defined for the model checker is composed of all possible packets (i.e., packet types, header values, etc) and their status as they travel across the network (i.e., current location of the packet, quality of service received so far, whether it is encrypted/tunneled or not, etc).

In our previous work, we compiled large numbers of devices with heterogeneous types into a single state machine. The states are defined as explained above, and the transitions drawn between them are defined by the topology, hardware capacity, and network/service policies. While the number of states is intractable, they are efficiently represented symbolically. Also, the valid transitions from a state to state are defined collectively via symbolic representation.

Fine tuning the model for performance was possible by exploiting efficient encoding of network data into the used Boolean variables as well as tweaking the order over which we build the expression trees using binary decision diagrams (BDD). We managed to concurrently model a few thousand devices of different types and crossing multiple layers, answer security and reachability queries, and add updates to the model in efficient and scalable manner.

The Simulation

The formal model (described above) can be used to answer queries on states reflecting packet transitions, or locations. Constraints can be defined for specific location, domain, or time modality. The query and its response reflect a snapshot of the network operations, whether temporally or spatially. We will take this static evaluation one step further, and evaluate continuously over time, while changing multiple constraints. The new constraints can be modeled to represent network dynamic conditions, configuration changes, …
We look at the problem as integrating discrete-event simulation with model-checking, where events are steps in time where the model needs to be evaluated.

So, our network simulator will start with the configuration (topology, routing, security…), and move along state transitions given network constraints (flow values, link changes, …). In simulation modeling, the network operation needs to be monitored and tracked over time, without restrictions. For this, only an initial state needs to be identified, and the simulation will track the model response at each time step.

When large network traces are available, those could be used in replay mode to trigger model tracking. Knowing end-to-end flow information from offline traces (CAIDA, …), model constraints can be changed incrementally as new flow information becomes available from the traces. With those constraints, the model can answer the queries (as described above), and each time-based snapshot (query result) gives a snapshot of the network state which corresponds to the simulation outcome for this time step.

Configuration and network state models can be changed to simulate what-if scenarios using the same traffic traces. Most changes can be applied very effectively on to a model without rebuilding except the directly affected part. The same idea can be used to apply the effect of external phenomenon. For example, one can model power/link outage, massive interference causing packet getting transmitted in error, excessive volumes of cross traffic, etc by merely tweaking few transitions or invalidating some of the states. This opens the door for many applications from disaster recovery planning to resource allocation and optimization.

Challenges

Several opportunities exist to enable large-scale simulation with formal modeling. For large networks, modeling all layers with diverse parameters can render the model unmanageable. Building the model is the most expensive operation, and parallel processing can enable fast model generation.
Parallel processing can be used to: parallelize model-checking platforms (formal modeling domain,

not here), or parameter selection and tuning (e.g. variable ordering), or decentralization of the simulation.
It is worth noting that current non-parallelized implementation can build the model for multi-layer configurations of 5K devices in less than 30 minutes [1,2]. While this seems satisfactory, it required manual tweaking of model building parameters (mainly variable ordering of BDDs and field encoding mechanisms). For a more general approach we have to automate this process and this requires, in turn, serious parallelizing of the model fine-tuning as well as the model construction operation itself.

Another source of complexity to the system stems from our need to model a more dynamic background status of the system. In other words, to model a realistic cross traffic, and actual network load, we need to 1) use actual sources and available network traffic traces, and 2) approximate these in a way that keep the model feasible to manage and analyze. Our prior work [4, 7] on traffic analysis gives us the ability to pinpoint the places to cut down traffic data without losing overall load fidelity.

References:

[1] Ehab Al-Shaer, Wilfredo Marrero, Adel El-Atawy, Khalid Elmansor, "Network Configuration in A Box: Towards End-to-End Verification of Network Reachability and Security", In the 17th IEEE International Conference on Network Protocols (ICNP'09), Princeton, New Jersey, USA, 2009.
[2] Adel El-Atawy, Taghrid Samak, "End-to-end Verification of QoS Policies", (NOMS'12), Maui, Hawaii, USA, April 2012.
[3] Alan Jeffrey and Taghrid Samak "Model Checking Firewall Policy Configurations", IEEE International Symposium on Policies for Distributed Systems and Networks (Policy 2009) 20-22 July 2009 – London, UK
[4] Taghrid Samak, Dan Gunter, Valerie Hendrix, "Scalable Analysis of Network Measurements with Hadoop and Pig", 5th Workshop on Distributed Autonomous Network Management System (DANMS), co-located with NOMS 2012.
[5] Taghrid Samak and Ehab Al-Shaer, "Synthetic security policy generation via network traffic clustering", The 3rd Workshop on Artificial Intelligence and Security, AISec, in conjunction with ACM/CCS 2010, ACM, October 2010
[6] Taghrid Samak, Adel El-Atawy and Ehab Al-Shaer, "Towards Network Security Policy Generation for Configuration Analysis and Testing", Workshop on Assurable & Usable Security Configuration (SafeConfig), Colocated with ACM CCS 2009, Chicago, USA, November 9, 2009
[7] Adel El-Atawy, Taghrid Samak, Ehab Al-Shaer and Hong Li, "On Using Online Traffic Statistical Matching for Optimizing Packet Filtering Performance", In the 26th Annual IEEE Conference on Computer Communications (INFOCOM'07), Anchorage, Alaska, USA, May 2007.

Primary authors: Dr EL-ATAWY, Adel (Google Inc.); Dr SAMAK, Taghrid (Lawrence Berkeley National Laboratory)

Co-author: Mr GUNTER, Daniel (Lawrence Berkeley National Laboratory)

Presenter: Dr SAMAK, Taghrid (Lawrence Berkeley National Laboratory)

Contribution ID: **9** Type: **not specified**

Modeling Expected and Anomalous Performance of End-to-end Workflows in Extreme-scale Distributed Computing Applications

Tuesday, 11 September 2012 14:35 (12 minutes)

Abstract Purpose:
In this extended abstract, we present our broad vision of research activities that are needed to model expected and anomalous performance of end-to-end workflows in extreme-scale distributed computing applications used within the DOE communities. In addition, we present a brief summary of our current DOE-funded studies to detect and diagnose uncorrelated as well as correlated network anomaly events within PerfSONAR measurement archives collected over large-scale network topologies across multiple domains.

Application Workflow Agendas:
The next-generation of high-performance networks such as the "ANI 100Gbps network testbeds" [1] and "hybrid packet/circuit-switched network testbeds" [2] are being developed in DOE communities. They are critical for supporting research involving large-scale distributed Petascale and Exascale science experiments and their data analysis, and also cloud computing initiatives in the DOE community such as the "Magellan" [3]. They cater to the increasing network demands of distributed computing application "inherent workflow agendas" that are described in workshop reports [4] [5]. An example workflow agenda relating to bulk file transfers from research instrumentation sites can be seen in the LHC data transfers from Tier-0 to Tier-1 and Tier-2 sites. An example agenda relating to data sharing amongst worldwide collaborators for replicating results, and refining conclusions can be seen in the LHC Tier-2 site collaborations. An example agenda relating to multi-user remote instrumentation steering and visualization relates to the remote access of PNNL Confocal microscopes in GTL project. An example agenda relating to remote analytics for real-time experimentation can be seen in the ITER inter-pulse data analysis using simulation codes at remote supercomputer centers.

DOE Networking and User Expectations:
The next-generation DOE networks provide two major advantages compared to today's networks: (i) high bandwidth capacity levels that deliver extreme-scale raw throughput performance for bulk file transfers, and (ii) very low latency levels or packet serialization times that deliver high-quality user experience for remote instrumentation steering, visualization and interactive-analytics. Consequently, the various DOE-supported distributed computing application users are generating a combination of both bandwidth-intensive and latency-sensitive traffic flows on the order of scales that have never been seen before. Given the substantial infrastructure investments to provide the high-performance networking capabilities, the networks will need to function in a manner that meets the high application performance expectations of the users. Examples of user expectations could include: (a) moving a Terabyte of LHC data within 6 hours between international collaborator sites, (b) smooth remote steering of the PNNL Confocal microscope that generates 12.5 Gbps high-definition video stream per camera to deliver "at-the-instrument" user experience for multiple geographically dispersed remote users, and (c) a west-coast remote user experiencing reliable performance over long time-periods when manipulating simulation codes and their graphical user interfaces pertaining to 2 to 3 Gbytes ITER inter-pulse data being transferred and analyzed every 20 minutes at NERSC supercomputer nodes.

Need for Novel Characterization and Modeling Strategies:
To ensure such robust functioning of next-generation networks, unique traffic flows need to be characterized and modeled to understand the user, application and network interplays. In the same context, the host and network device technologies supporting these extreme-scale applications are in their early stages of development to support 10-100Gbps application traffic flows, and will in-

troduce performance bottlenecks that need to be detected, localized and resolved. The lessons to-be learned from such bottleneck detection testing will drive the design, development, deployment and monitoring of future 10-100Gbps, and beyond - supporting host and network technologies. Even more importantly, users/operators of the extreme-scale applications will need to have "expectation-management" tools that enable them to model, analyze and visualize if their inherent workflow agendas are performing as expected or are anomalous (particularly if they are faulty), given the infrastructure resources (e.g., instrument, compute nodes, storage, network circuit) being co-scheduled to meet their application demands. If the anomalies are benign and cross expectation boundaries, it will still be beneficial for users/operators to be notified about such changes. The expectation and change notifications could be instantaneous or could be in the form of daily or weekly trends that highlight facts such as for e.g., typically noon – 4pm on Tuesdays and Thursdays, paths of interest for the user tend to be congested due to flash crowd behaviors in recurring LHC experiments, or due to any other extreme-scale application traffic flows that increase the co-scheduling loads over the shared infrastructure of a DOE community.

Potential Characterization and Modeling Strategies:
We are envisioning research and development activities that aim to meet the above needs of extreme-scale: (i) user-and-application, as well as application-and-network interplay characterization and modeling involving bandwidth-intensive and latency-sensitive traffic flows, (ii) fault detection and localization by analyzing performance measurements across end-to-end host and network devices, and (iii) users' workflow agenda performance "expectation-management" modeling and analysis tools that extend familiar and widely-adopted middleware software interfaces (e.g., Pegasus Workflow Management System [6], Netlogger [7], perfSONAR[8], NetAlmanac [9]). These three activities should build upon each other, and our hope is that they will ultimately provide the DOE community: mathematical-models of network requirements for extreme-scale DOE user applications; fault detection and localization framework leveraging latest advances from regression analysis, model learning and constraint satisfaction theory; openly-available tools for extreme-scale application modeling and simulations, and real-network workflow agenda performance measurements. Further, we remark that existing models of application performance in the DOE community are mostly network Quality of Service (QoS) centric and focused on bulk file transfer application performance. There is a dire need to fill the dearth of knowledge in the DOE community regarding performance issues and modeling of user-and-application, as well as application-and-network interplay when considering mixtures of bandwidth-intensive and latency-sensitive traffic flows that will dominate the next-generation DOE networks.

When considering large mixtures of bandwidth-intensive and latency-sensitive traffic flows, and the nature of next-generation network paths, the measurement data volumes from sampling will be substantially large, and the user expectations of application performance will be considerably high. Consequently, there is a need to explore advanced statistical analysis coupled with effective visualization techniques for modeling user-and-application, as well as application-and-network interplay to detect bottleneck phenomena. There are already several tools that have been developed such as Pathload [10], Pathchar [11], Iperf/BWCTL [12], NDT[13], NPAD [14] to diagnose common bottlenecks such as duplex-mismatch, network congestion and hop mis-configurations along a path. However, the next-generation DOE networks will support extreme-scale distributed computing application "agendas" that have workflows involving user actions that are both bandwidth-intensive and latency-sensitive to communicate with multiple remote resources and collaborator sites. To detect bottlenecks in such application workflow agendas, novel user Quality of Experience (QoE) performance metrics and agenda-exercising tools need to be developed that are aware of user-and-application and application-and-network interplay issues, and bottleneck phenomena that may be very different from the phenomena seen in today's networks.

We believe that agenda-exercising tools that will need to be developed should be suitable for online monitoring and resource adaptation in production environments to maintain peak performance. It will be inevitable for the tools to be able to query and leverage the existence of perfSONAR measurement archives (and possibly demand the creation of new kinds of perfSONAR measurement archives) along network paths so as to reinforce analysis conclusions about network bottleneck causes. False alarms from such tools without proper reinforcement mechanisms through perfSONAR measurement archives could lead to undesirable mis-configuration of expensive resources.

Hence, the agenda-exercising tools should be developed as interoperable (e.g., perfSONAR framework compliant) middleware software that can be leveraged by resource adaptation frameworks such as the ESnet OSCARS [15].

Prior Anomaly Detection and Diagnosis Research Results:
In our current research grant from DOE ASCR titled "Sampling Approaches for Multi-domain Internet Performance Measurement Infrastructures to Better Serve Network Control and Management", our preliminary results show how temporal and spatial analysis of latency and throughput performance data along with route information over multi-domain networks that were obtained using PerfSONAR web services can help in better understanding of the nature, locations and frequency of anomalous events, and their evolution over time. We are using network cliques modeling and evolution characterization techniques, and metrics (e.g., common hop to common event ratio; location affinity, event burstiness), adopted from fields such as Social Networking. We have been able to analyze "uncorrelated anomaly events" found in worldwide PerfSONAR measurement data sets, and also "correlated anomaly events" found in PerfSONAR measurement data sets between the various DOE national lab network locations. We believe our preliminary results are a major step towards modeling, end-to-end monitoring, troubleshooting and intelligent adaptations of workflows to ensure optimum user QoE in extreme-scale distributed computing applications.

References

[1] ANI 100G Network Testbed - https://sites.google.com/a/lbl.gov/ani-100g-network

[2] DOE ASCR Research – Next-generation networking for Petascale Science - http://www.sc.doe.gov/ascr/Research/NextGe

[3] Magellan DOE Cloud Computing Initiative - http://magellan.alcf.anl.gov

[4] Workshop on Advanced Networking for Distributed Petascale Science: R&D Challenges and Opportunities. April 8-9, 2008.

[5] Workshop on Science-Driven R&D Requirements for ESnet, April 23-24, 2007.

[6] Pegasus Workflow Management System – http://pegasus.isi.edu

[7] D. Gunter, B. Tierney, B. Crowley, M. Holding, J. Lee, "NetLogger: A Toolkit for Distributed System Performance Analysis", Proc. of IEEE MASCOTS, 2000.

[8] A. Hanemann, J. Boote, E. Boyd, J. Durand, L. Kudarimoti, R. Lapacz, M. Swany, S. Trocha, J. Zurawski, "PerfSONAR: A Service Oriented Architecture for Multi-Domain Network Monitoring", Proc. of Service Oriented Computing, Springer LNCS 3826, pp. 241-254, 2005. http://www.perfsonar.net

[9] ESnet Netalmanac – http://code.google.com/p/net-almanac

[10] C. Dovrolis, P. Ramanathan, D. Morre, "Packet Dispersion Techniques and Capacity Estimation", IEEE/ACM Transactions on Networking, Volume 12, Pages 963-977, December 2004.

[11] A. Downey, "Using Pathchar to Estimate Internet Link Characteristics", Proc. of ACM SIGCOMM, 1999.

[12] A. Tirumala, L. Cottrell, T. Dunigan, "Measuring End-to-end Bandwidth with Iperf using Web100", Proc. of Passive and Active Measurement Workshop, 2003 - http://dast.nlanr.net/Projects/Iperf

[13] Internet2 Network Diagnostic Tool (NDT) – http://www.internet2.edu/performance/ndt

[14] M. Mathis, J. Heffner, P. O'Neil, P. Siemsen, "Pathdiag: Automated TCP Diagnosis", Proc. of Passive and Active Measurement Workshop, 2008.

[15] ESnet OSCARS - http://www.es.net/oscars

[16] C. Logg, L. Cottrell, "Experiences in Traceroute and Available Bandwidth Change Analysis", Proc. of ACM SIGCOMM Network Troubleshooting Workshop, 2004.

[17] R. Wolski, N. Spring, J. Hayes, "The Network Weather Service: A Distributed Resource Performance Forecasting Service for Metacomputing", Journal of Future Generation Computer Systems,

Volume 15, Pages 757-768, 1999.

Primary author: CALYAM, Prasad (The Ohio State University)

Presenter: CALYAM, Prasad (The Ohio State University)

Contribution ID: **10** Type: **not specified**

Global Rendezvous Scale Issues for an Information-Centric Future Internet

Tuesday, 11 September 2012 14:09 (12 minutes)

The Past
In the 1990's we got involved with Internet traffic characterization [A, D] and in particular traffic-flow profiling [B], as well as Multicast and Continuous Media Dissemination [C, J, K, L] and Wireless Internet Multimedia Communications [F, G, H]. Those investigations involved vertical understanding (for example, continuous media dissemination feasibility and contributing issues throughout the layers), horizontal understanding (for example, considered long Internet paths with one or more wireless links at various points in the path and its impact on performance [F, G, H]), and large-scale computational modeling and analysis (for example, required the use of real-time sampling of the data to cope with its rate and size [E]) of the Internet or networks in general.
• Internet Traffic Characterization: We introduced the concept of IP flows and used it to characterize real Internet traffic at various levels of granularity. The notion of IP flows provides a flexible tool for bridging the gap between the connectionless/stateless networking model of the Internet's (inter)network layer and the connection-oriented/stateful model more appropriate for some applications (e.g., packet video).
• Multicast and Continuous Media Dissemination: We have contributed on various aspects of multicast protocols and multimedia multipoint communications. IP Multicast for point-to-point and wireless networks with mobility has been investigated. We have also developed efficient multimedia dissemination techniques that support heterogeneity (in both the terminals and network paths) and effective congestion control in packet switching networks using hierarchical coding of continuous media, such as real-time video.
• Wireless Internet Multimedia Communications: The Internet protocols were designed with wireline networks in mind and perform rather poorly in wireless environments. We contributed to the understanding of the problem and the awareness of the community, in addition to proposing a framework to address them in a realistic, effective, general, and efficient way.
The Future
It has now been realized for long that the Internet has evolved from an internetwork for the pairwise communication between end hosts to a substrate for the delivery of information. The users are increasingly concerned with the content they are accessing (or contributing), rather than the exact network end point providing it. This major shift has resulted in the emergence of a series of new technological enablers for the efficient delivery of content, ranging from application layer solutions (e.g. CDNs), to proposals for new, clean-slate designs for the Future Internet based on the Information-Centric Networking (ICN) or content-centric networking paradigm ([U-V], [1],[2],[3]).
In all these efforts, the act of locating the desired content in the network (e.g., through name resolution) has been regarded as an increasingly challenging task, facing serious scalability and complexity concerns. The huge volumes of available content in the Internet, especially with the advent of user generated content, have resulted in a correspondingly enormous name space challenging even the management of meta-data information and the act of locating the desired content. Considering that the current amount of unique Web pages as indexed by Google is greater than 1 trillion [4] and that some billions [5] of devices, ranging from mobile phones to sensors and home appliances are joining the network to offer additional content or information, we could be safely speaking that an ICN has to manage a number of Internet Objects (IOs) in the order of 1013. (Other

studies raise this estimate to 1015 [6].) At the same time the large size of the Internet ecosystem adds to the scalability concerns, since the need for efficiently locating the desired content spans several thousand networks (more than 35K ASes reported in the latest CAIDA trace set), with hundreds of thousands of routers. Moreover, the vast number of (mobile) end host devices, is not only contributing to the huge volume of content, but also to a considerably high volume of requests for content.

Though major research efforts have been devoted to building highly scalable name resolution systems, locating information in the current (and future) Internet is further facing significant complexity challenges. The current Internet landscape is a mosaic of different actors. A multitude of different producers of content, ranging from simple end users to large content providers, is offering large volumes of content. Each content provider may require the establishment of different access rights and privacy policies. The provided content must be discovered and delivered across a multitude of distinct networks under different administrative authorities, often following complex routing policies dictated by economic parameters. At the same time, the emergence of large CDNs introduces a further layer of complexity by allowing the replication and caching of content at several parts of the internetwork, driven by end user demand. In addition, different types of access networks (e.g., ADSL, Wi-Fi, 3G/LTE, 4G…) and end user devices (tablets, smartphones, laptops, etc.) introduce further complexities for the adaptation of content according to the current context.

It therefore becomes evident that locating the desired content in the current (and anticipated future) Internet is a task that has many dimensions that call for careful consideration. Until now most research efforts on this challenge were focused on particular aspects, often investigating a limited subset of the involved parameters in isolation e.g., using simplified inter-domain topologies, conducting small scale simulations, neglecting aspects such as content replication, etc. Hence, a horizontal understanding is required taking into account the entire set of the aforementioned aspects and the interactions among them. A series of important questions cannot be answered unless a holistic view on this landscape is taken e.g., how does the heterogeneity of the Internet impact on the mechanisms employed to locate the content? What is the impact on the performance of a name resolution system? How would the exchange of information (meta-data) between different actors affect the operation of such a system in terms of reachability of content?

In order to gain a better understanding of this issue, we need to simultaneously model a series of practical aspects and features stemming exactly from this diversity. Namely, we have to model aspects such as: (1) the generation of new content in the Internet, (2) the temporal evolution of the popularity of the different types of content available, (3) the locality characteristics of end user requests, (4) the (current) content replication/caching policies of CDNs, (5) both the inter-domain and intra-domain level topology characteristics, (6) (Inter-domain) Routing policies, (7) the implications introduced by wireless networks i.e., content tailored for mobile devices, smart phones, (8) the implications introduced by the Internet of Things (IoT) e.g., high volumes of information, geographical characteristics, access patterns, and socio-economic aspects.

This complicated set of interacting issues and models, is expected to impact the investigation of the various issues emerging from the interaction of the respective actors, contributing to the realistic investigation of currently available, as well as new mechanisms for locating content in the Internet. For instance, today it is difficult to assess the potential benefits and pitfalls stemming from the establishment of a synergy between CDNs, content providers, and ISPs expressed via the exchange of meta-data related to the discovery of the closest replica of some content.

Some first steps in these investigations have been undertaken in research projects we have or are participating, but we are only at the very beginning ([1, O-T]).

References in comments, below.

Primary author: Prof. POLYZOS, George (MMlab/AUEB)

Presenter: Prof. POLYZOS, George (MMlab/AUEB)

Contribution ID: **11** Type: **not specified**

Scale+Fidelity+Speed+Integration=Necessary+Possible in Big Network Simulations

Tuesday, 11 September 2012 11:26 (12 minutes)

We are in violent agreement with the vision and outlook of the COMBINE workshop. It is clear that effects such as feedback and rapidity of change are becoming so pronounced in global computer networks that modeling and simulation of big networks is both a timely need as well as a formidable challenge.

We are convinced that simulation is the third leg of network science, just as it has been accepted so in physical sciences recently. However, capture of scale, fidelity, integrated capture of complexity and similar challenges underlie the realization of simulation as a first-class scientific principle. In relation to this view, our team's expertise and interests have beeen in taking network modeling and simulation tools to the extreme along these important dimensions *simultaneously* – largest scale (multi-million node scenarios), highest fidelity (virtual machines, packet-level models, packet-fluid hybrids), and realistic/active behaviors (combined with other man-made phenomena, such as electric, vehicular, and other grids).

Our team's research has been driven by the understanding that (a) network effects at scale are very difficult to capture sans simulations, (b) feedback effects dominate to such a degree that fidelity in simulations is difficult to be abstracted, (c) the effects of interest when viewed in the overall system-level, at nation- or global-scale, are best captured by large-scale simulations and analyses alone.

Scale: We have dealt with the scale aspect in the past, by advancing techniques to sustain large-scale parallel simulations of network models on some of the largest supercomputing installations (MASCOTS'03). We executed multi-million node simulations (using pdns) on a Pittsburg Supercomputing Center machine (the largest of that time) as part of Georgia Tech team (with Fujimoto and Riley) on a DARPA NMS project in 2002-03. These runs still stand today as some of the largest packet-level network simulations to date.

Complexity: Our multi-million node computer network worm propagation simulations reported in ACSAC'04 still represent some of the largest, high-fidelity cyber security simulations with respect to scale. We are convinced that the effects of scale are best captured by simulations of this size, to uncover unforeseen effects or emergent phenomena.

Fidelity: We are currently uncovering the issues, and devising solutions, to increasing the model fidelity to the extreme, using virtual machines (VMs) as surrogates for either end-hosts or intermediate routers or both. While VMs have been employed in network simulations in recent times, a new challenge remains to be solved, namely, the incorporation of a first-class notion of virtual time into the VMs. Native VM schedulers are a gross mismatch to VM-based network simulations. We recently showed that using native schedulers (which are typically optimized for throughput) in fact can give wrong answers from simulations. With funding from Army Research Laboratory and others, we are developing new virtual time-ordered schedulers (PADS'11, MASCOTS'12) that are indispensible for large-scale high-fidelity network simulations in the future when executed on many-core host platforms of supercomputing-scale configurations.

Related: Our parallel discrete event simulator, musik, is being developed to suit the vision of very large-scale simulations in the future, and has now matured to the point of being the only simulation engine in the world to be tested on the largest extant supercomputing platforms (Jaguar, 216K+ cores), with multiple discrete event applications exercised as proofs of concept (epidemiology, radio signal, vehicular transportation, etc.)

We are highly interested in interacting with the network modeling and designing experts at the workshop and sharing ideas towards meeting the grand vision painted in the call, namely, a community analagous to that of the climate simulation community, for network science.

Primary author: Prof. PERUMALLA, Kalyan (Oak Ridge National Laboratory)

Presenter: Prof. PERUMALLA, Kalyan (Oak Ridge National Laboratory)

Contribution ID: **12** Type: **not specified**

Computationally Modeling High-Speed Scientific Networks

Tuesday, 11 September 2012 14:22 (12 minutes)

Computationally Modeling High-Speed Scientific Networks
Jun Yi, Venkatram Vishwanath, and Rajkumar Kettimuthu
Mathematics and Computer Science Division, Argonne National Lab
{jyi, venkatv, kettimut}@mcs.anl.gov

1, The Need of Computational Modeling for Scientific Networks Scientific experiments (e.g., climate modeling and prediction, biomedical imaging, geosciences, and high-energy physics) are expected to generate, analyze, and distribute data volumes on the order of petabytes. These experiments are critically dependent upon advanced high-speed networks to move their enormous data between local and remote computing and storage facilities. These experiments usually have a wide range of networking requirements and characteristics, e.g., bulk data transfer (high bandwidth, loss-less), climate visualization (large bandwidth, less jitter), and real-time data analysis and decision making (high bandwidth, low latency, and loss-less). Moreover, these scientific data flows usually traverse multiple different networks (shared WAN, dedicated circuit-based WAN, and LAN) and are transferred using different protocols for better performance (For example, GridFTP or parallel TCP over high-bandwidth
large-delay networks, UDT over high-latency networks). The heterogeneity of networks, flows, and protocols and the ever-increasing traffic volume challenge the scientific network planning, operation (e.g., troubleshooting and configuration), and design to satisfying the heterogeneous requirements of scientific experiments.

Here we present two use cases that will benefit from the computauional modelling of end-to-end large scale networks. DOE's Advanced Photon Source (APS) user facility at Argonne National Laboratory provides the Western Hemisphere's most brilliant X-ray beams for research. It is projected to generate 100+ terabyte per day within the next year. As data volumes increase, experiments may have to use external supercomputing facilities to process the data in real-time, which will necessitate additional fast WAN transfers. The Office of Science at DOE projects to support such high speed transfers as APS and comes across a challenging question naturally: how to evolve from current high-speed network. Adding hardware capacity and improving software efficiency are two solutions. However, where to add these hardware capacity and how/where to improve software efficiency cost-effectively need scientific, not intuitive or hypothetical, answers. A computational model of scientific networks can answer the question precisely: we can experiment on various network configurations with the data communications requirements that the Office of Science at DOE collects yearly from scientists, and choose the one that satisfies their requirements with the lowest cost or lowest evolution effort.

Globus Online and Globus GridFTP provide high-throughput, reliable, and secure big data movement service, but the throughput in most cases, is still far from the physical network capacity. To further improve the throughput of a GridFTP transfer, an end-to-end approach can help to identify the bottleneck and optimally choose routing and transportation protocols and parameters along the path, which, however, can not be fully effective until we have deeper understanding of interactions among various data flows and between the flows and network. With a computational model, the network configuration based on current status can be modeled and experimented to locate performance bottlenecks and choose the optimal configurations from the vast configuration space in real-time.

2, The Challenges of Computational Modeling for Scientific Networks
However, it is challenging to computationally model scientific networks, even with the massive computing, networking, and storage capacities provided by today's supercomputers, grids, and

clouds. These challenges originate from the demands of computational models:
[1] Scalability. The model should be able to represent networks of various scales. The execution of the model should fully harness the resources of future exascale computing facilities.
[2] Accuracy. The model should be as accurate to the real world as possible. It should accurately predict performance results (e.g., throughput, latency, transfer completion time, resource usage) under various traffic uncertainties with appropriate levels of computation loads within a certain time limit.
[3] Composibility/Extensibility. Models of two interconnecting regional
networks should be able to easily compose to form a larger network without knowing the internal details of each model.

Among all of these challenges, the scalability challenge is the most critical one from our perspective. If a network model can not take full advantage of the underlying massive parallel computing capacities, it can not produce accurate results within a certain limited time, not even to mention experiments on a larger composed networks.

Existing network modeling techniques will not be effective considering the sheer scale and complexity of scientific networks. The discrete packet-level event simulation method usually uses a central discrete event scheduler to schedule the time-stamped events such as packet and timer expiration. This centralized scheduler becomes both the communication and computation bottleneck at scale. For example, 1M events per network processing entity (e.g., network reception interrupt handler) per second and 1M vertical or horizontal network processing entity for a network of medium scale will generate 1T events to be delivered, processed, synchronized, and persisted in a second. Assuming 256 bytes for each event packet, 256T bytes per second traffic will be generated, which will overwhelm current and forthcoming aggregated network capacities of supercomputers and grids (several terabytes per second). Moreover, event handlers must be executed in ascending timestamp order to preserve the semantics of the physical network. Further, an executing event handler may generate new events, which should be queued temporally as well. This method is not scalable computationally due to the implicit fact that events are processed serially.

The pivot issue that prevents network models from harnessing the massive computing/communication capacity lies mainly in the process/event synchronization method. On the one hand, massive amount of data and control information flows among network elements and any patterns of dependency may exist and change dynamically (e.g., a single event may have a ripple effect over the entire network). The complicated inter-dependency naturally requires a serial execution of events, which exacerbates the modeling performance when combined with large scale network models. For example, the tardiness of a single element may extremely waste computing resource as a whole (e.g., many entities wait for the completion of a tardy entity at a certain lock step) and slow down the entire model. On the other hand, the capacity of today's supercomputers, grids, and clouds can only be fully harnessed by programs exhibiting massive parallelisms. The execution of a serial program makes no much difference on a desktop computer or supercomputer.

Existing event/process synchronization methods in the distributed system literature is not efficient at scale. Most existing synchronization methods only deals with logical (not timeliness) dependency. Just until recently, parallel discrete event/process simulation was merely an academic research topic. There are basically two methods to impose the correct temporal order of distributed event execution: conservative and optimistic methods. By a conservative method, only a safe event can be executed (if a process contains an event E1 with timestamp T1 and the process can determine that is impossible to receive an event with a smaller timestamp, then the process consider that executing event E1 is safe without violating temporal constraints). Unsafe events must be blocked. Consequently, most events processings are blocked in most time. By an optimistic method, the temporal relationship between events can be broken but a detection and recovery mechanism is added: whenever the incorrect temporal order of events is detected a rollback mechanism is invoked to recover. However, deadline deadlocks are frequently formed and hard to detect in a complicated and large network and therefore large computation is spent on expensive deadlock detection and recovery. Moreover, the running pace of events in the model rarely match those of the physical network and the violation of temporal relationships exists everywhere (due to pervasive and complicated dependency among network elements) and thus most computation is spent

on computation rollbacks.

3, Our Basic Idea and Its Challenges
We propose a hierarchical method for modeling scientific networks. The basic idea is to organize temporal relationships, event delivery, and work flow execution hierarchically to reduce communication overhead, increase the parallelism of event processing, and dynamically balance event processing workload in a synchronization-aware manner. The entire model is organized as a tree with (either vertical or horizontal) network elements as leaves. Each interior node comprises a temporal synchronizer (TS), a workload distributor (WS), and an event forwarder (EF). For example, multiple network processing elements (e.g., routing, queue, and transport elements) of a networked computer can serve as leaves under the same parent. Multiple TSs, WSs, and EFs corresponding to multiple interconnecting physical network elements can serve as children of a common interior node corresponding to an autonomous system.

The hierarchy allows disjoint subtrees to run in parallel provided that the difference of their simulation times is no greater than the communication latency between them. A TS is responsible for synchronizing the event processing within its subtree. It uses the existing tree-barrier method to solve the communication bottleneck problem, e.g., it reduces temporal communication load and latency since a few messages cross the hierarchy can progress the entire model to next time step. Moreover, it avoids time-consuming and complicated event synchronization deadlock detection and recovery by the conservative and the optimistic synchronization methods.

An EF is responsible for forwarding data and control events between interconnect (either vertical or horizontal) network elements, which further reduces communication workload since absolutely majority of control and data flows within the same subtree. Moreover, since EFs forward all timestamped events, they are able to accurately estimate the needed resource by any subtree in any specific time range, which facilitates distributing workload to the underlying distributed and parallel execution environment. A WS is responsible for distributing event processing workload within its subtree (e.g., to accelerate event processing if it lags behind other subtrees). We do not use optimistic event synchronization methods to seek opportunistic event independence at the cost of expensive computation rollbacks, which we believe can only accelerate turnaround time of a model of scientific networks to a limited extent. Alternatively, we resort to synchronization-aware, hierarchical, and dynamic load balance approach, where sluggish subtrees will be timely allocated sufficient resource to keep pace with the remainder of the model.

However, we still face many challenges to pursue this approach. We list a few of them:
[1] Configurations. How to decide the size, height, and descendents of subtrees is important to the efficiency of this approach. We intend to adaptively configure the tree and subtrees according to the network element interaction pattern of the physical network and the capacity and workload of the underlying parallel and distributed execution environment.
[2] Workload distribution. Distributing the event processing within a subtree appropriately into concrete execution entities (e.g., threads within the same process, threads/processes distributed across multiple cores or heterogeneous machines, etc.) greatly affects the running time of the model due to changing communication latency and workload distributions in each simulation time-step. A dynamic workload- and synchronization-aware scheduling framework is needed.
[3] Fault tolerance. TSs, EFs, and WSs will be single points of failure and redundancy mechanisms are needed to recover from failures. We intended to use existing replication technique to increase the robustness of this approach.
[4] Group communication. TSs need to broadcast timestamps within theirs respective subtrees. Effective group communication mechanisms will extremely improve the efficiency of this approach and therefore is worth investigating.

Primary authors: Dr KETTIMUTHU, Rajkumar (Mathematics and Computer Science Division,

Argonne National Lab); Dr VISHWANATH, Venkatram (Mathematics and Computer Science Division, Argonne National Lab)

Presenters: Dr KETTIMUTHU, Rajkumar (Mathematics and Computer Science Division, Argonne National Lab); Dr VISHWANATH, Venkatram (Mathematics and Computer Science Division, Argonne National Lab)

Contribution ID: **14** Type: **not specified**

Simulation of Large (>10K node) Computer Networks

Tuesday, 11 September 2012 10:11 (12 minutes)

1. INTRODUCTION

Predictive analysis of cyber risk and performance is one of the major gaps in cyber analytics.[1] Understanding how a specified mission-critical application will execute in a network context, characterizing the potential impact of network threats on critical applications, and predicting the effect of proposed defensive actions are critical capabilities for a risk-based cyber strategy. The Livermore Lab has embarked on a multi-year effort to develop a large-scale realistic network simulation capability. Specifically, we are developing computer network simulations for realistic networks derived from real and synthetic network maps, and which incorporate real hardware and geographic constraints, at enterprise (10K node) and above scale, and generate traffic from realistic traffic models matched to observed data. In this abstract we describe our approach and specific applications areas of interest.

Network simulation has been an active area of work since the 1960's,[2] resulting in a broad set of both commercial[3, 4] and open-source[5] tools. Network simulation is based on discrete event simulation[6]–the most basic event is the sending receiving of a network packet. Nodes in the simulation can be host computers, which create and receive packets, and routers, which forward packets on the route to their destination host. The simulators generally implement full TCP/IP network protocol stacks over physical models for wired and wireless RF communication links. Network simulators are generally used in the development of new network technologies–new routers, protocol variations, congestion control algorithms, etc. In these applications simulation of networks with hundreds of host computer and routers is adequate and there is little motivation to extend simulations to much larger networks. Most existing efforts are limited to modest scale (few hundred nodes), unrealistic network models,[7] and unrealistically simple on|off traffic models.[8]

For our intended applications existing network simulators are limited in three regards:

• Host behavioral models are unrealistically simple.[8] To reproduce behaviors seen in real networks we will need more sophisticated user models representing more complex activities like Web surfing, e-mail interchanges, and peer-to-peer file interchange.

• There has been little effort to scale network simulations to even the enterprise network level. A few demonstrations of parallelized network simulation have been performed at Georgia Tech[9] and at the Army Research Lab[10] but these efforts have barely begun to explore the area. For example, little is known about optimal cluster configurations or effective mapping of simulated nodes and communication links to physical compute nodes.

• There has been little systematic validation of the simulations outside the narrow range of detailed network technology applications noted above. In particular the ability of simulations to produce statistically realistic network behaviors at enterprise scale and above is completely unexplored. This is exactly the performance space of interest for mission assurance applications.

To focus our research efforts, we have identified three application areas: enterprise networks, mission-critical applications, and worldwide routing. In this abstract we summarize our approach to modeling enterprise networks and understanding the scaling issues involved.

1. RESEARCH GOALS AND CAPABILITIES

Our research goals are centered on understanding the capabilities and limitations of network simulation. In the application areas we want to focus on the following questions:

• Can we reproduce the statistics of observed behaviors at scales from enterprise-level networks up to the global Internet? What model fidelity is needed to produce a given behavior? What level of abstraction can we get away with?

• Can we integrate models at different scales to achieve high fidelity and large scale, e.g, virtualized nodes and networks around nodes of interest, while using more abstract packet-level simulations at the largest scales?

• What are the limits to scaling network simulations with current tools?

• Can we predict the response of the network to changes in topology or dynamics?

In addition to utilizing Livermore's significant high-performance computing resources, we will take advantage of several other existing research programs at the Lab.

The Livermore Laboratory has ongoing efforts in understanding network topology and services, analysis of live traffic capture, host-based behavior tracking, and data analysis on large graphs. Our network mapper, which provides highly detailed descriptions of real networks and services, in combination with host-based measurement and live traffic capture and analysis, provide an unprecedented source of validation data for realistic behavior models and associated traffic generators.

We have surveyed and evaluated existing network simulation frameworks, opting to begin with ns3.[5] To date we have developed an XML-based network description language to describe the simulation topology and applications, and generate the simulation code automatically. We have outlined a statistically driven model to generate realistic behavior. We have identified a series of test problems for each of the application areas described. These test problems are typically simplified versions of the ultimate application scope, based on published work, so we have a point to validate against.

1. ENTERPRISE NETWORK APPLICATION

An enterprise network consists of ~10K nodes, with most nodes in trees attached to a core clique of fully connected central routers. The background for this network is the rest of the Internet, connected to the core routers (through an edge router) by a very small number of links, typically only one, with a second backup connection. The combination of fully connected core routers and few links to the larger Internet gives these networks a definite sense of inside and outside. Traffic flow is dominantly between internal hosts, but with significant Internet traffic.

We plan to couple results from current maps of realistic networks, including the Lab, with behavioral data from our traffic capture and host-based behavior projects. The overall objective is to model enterprise networks with realistic traffic generators, and measure the range of variability of realistic networks given constraints from mapping data.

There are many tools for mapping enterprise networks,[11-15] and some simulation studies of performance. We believe that quantifying errors in mapping, generating realistic traffic, and multi-scale network modeling are all new.

There are a number of specific tasks required. We have developed the capability to convert a network map into a simulation topology, complete with specification of the variety of traffic-generating applications to be simulated on each node. We are currently studying the limitations in simulating 10K node networks where most hosts are actively generating traffic. We will be studying how to create ensembles of network models consistent with the mapping input data, and developing metrics to quantify performance from the ensembles. We will also create multi-scale models to study fidelity issues.

In conclusion, we are developing capability to simulate realistic networks, derived from real and synthetic network maps at enterprise (10K node) and above scale, and generate traffic from realistic traffic models matched to observed data. We aim to understand the capabilities and limitations of large-scale network simulations, with demonstrated applications in cyber security, global network situational awareness, performance modeling and prediction.

1. ACKNOWLEDGMENTS

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

1. REFERENCES

2. J. M. McConnell, Vision 2015: a globally networked and integrated intelligence enterprise, July 2008 (Director of National Intelligence, 2008).
3. Modeling and Tools for Network Simulation, edited by K. Wehrle, M. Günes, and J. Gross (Springer, New York, 2010).
4. A. Varga, The OMNET++ discrete event simulation system in European Simulation Multiconference ESM'2001 (Prague, Czech Republic, 2001), https://labo4g.enstb.fr/twiki/pub/Simulator/SimulatorReferenc meth48.pdf.
5. A. Varga and R. Hornig, An overview of the OMNeT++ simulation environment in the 1st international conference on Simulation tools and techniques for communications, networks and systems (ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Marseille, France, 2008).
6. ns-3 Collaboration, The ns-3 network simulator (Washington, 2011), Vol. 2011, http://www.nsnam.org/.
7. R. Fujimoto, Parallel and Distributed Simulation Systems, (John Wiley & Sons, 2000).
8. L. Li, et al., A first-principles approach to understanding the internet's router-level topology, SIGCOMM Comput. Commun. Rev. 34, 4, 3 (2004).
9. E. K. Çetinkaya, et al., A comprehensive framework to simulate network attacks and challenges in IEEE Second International Workshop on Reliable Networks Design and Modeling (RNDM'10) (Moscow, 2010).
10. C. D. Carothers, D. Bauer, and S. Pearce, ROSS: A high-performance, low-memory, modular Time Warp system, Journal of Parallel and Distributed Computing 62, 11, 1648 (2002), <Go to ISI>://000179497400003.
11. J. Clarke, et al., The Network Interdisciplinary Computing Environment (US Army Research Laboratory, 2011).
12. Lumeta, Lumeta - Global Network Visibility (2011), Vol. 2011, http://www.lumeta.com/.
13. AdRemSoft, NetCrunch (2011), Vol. 2011, http://www.adremsoft.com/netcrunch/.
14. Nmap.org, nmap (20111), Vol. 2011, http://nmap.org/.
15. Q. Software;, PacketTrap (2011), Vol. 2011, http://www.packettrap.com/network/index.aspx.
16. Solarwinds, LANSurveyor (2011), Vol. 2011, http://www.solarwinds.com/products/LANsurveyor/.

Primary author: Dr BARNES, JR., Peter (Lawrence Livermore National Laboratory)

Co-authors: Dr JEFFERSON, David (Lawrence Livermore National Laboratory); Dr COLON, Domingo (Lawrence Livermore National Laboratory); Dr BRASE, James (Lawrence Livermore National Laboratory); Dr HORSLEY, Matthew (Lawrence Livermore National Laboratory); Dr SOLTZ, Ron (Lawrence Livermore National Laboratory); Dr NIKOLAEV, Sergei (Lawrence Livermore National Laboratory)

Presenter: Dr BARNES, JR., Peter (Lawrence Livermore National Laboratory)

Contribution ID: **15** Type: **not specified**

Opening remarks

Tuesday, 11 September 2012 08:30 (15 minutes)

Primary authors: Mr CARLSON, Rich (Internet2); CARLSON, Rich (DOE)

Presenters: Mr CARLSON, Rich (Internet2); CARLSON, Rich (DOE)

Contribution ID: **16** Type: **not specified**

Workshop agenda and objectives

Tuesday, 11 September 2012 08:45 (15 minutes)

Primary author: DOVROLIS, Constantine (Georgia Tech)

Presenter: DOVROLIS, Constantine (Georgia Tech)

Contribution ID: **17** Type: **not specified**

Keynote

Tuesday, 11 September 2012 09:00 (45 minutes)

Primary author: Dr FOSTER, Ian (ANL)

Presenter: Dr FOSTER, Ian (ANL)

Contribution ID: **18** Type: **not specified**

Challenges in the computational modeling of big networks

Tuesday, 11 September 2012 09:45 (12 minutes)

Primary author: Dr FOSTER, Ian (ANL)

Presenter: Dr FOSTER, Ian (ANL)

Contribution ID: **19** Type: **not specified**

Challenges of Multi-Scale Network Modeling and Analysis

Tuesday, 11 September 2012 11:00 (12 minutes)

Primary author: NICOL, David

Presenter: NICOL, David

Contribution ID: **20** Type: **not specified**

Grand Challenge: Leveraging Extreme-Scale Supercomputers for Modeling the Human Sustainability Network

Tuesday, 11 September 2012 11:39 (12 minutes)

Presenter: CAROTHERS, Chris

Contribution ID: **21** Type: **not specified**

TBA

Primary author: BAGRODIA, Rajive

Presenter: BAGRODIA, Rajive

Contribution ID: **22** Type: **not specified**

Predicting Global Failure Regimes in Complex Information Systems

Tuesday, 11 September 2012 13:43 (12 minutes)

Presenter: MILLS, Kevin

Contribution ID: **23** Type: **not specified**

Modeling Large-scale Networks with Flow Graphs

Tuesday, 11 September 2012 15:01 (12 minutes)

Presenter: KISSEL, Ezra