Passive Monitoring Challenges on High-Speed Switched Networks

Hamed Haddadi, Lionel Sacks

Department of Electronic & Electrical Engineering

University College London

[hamed@ee.ucl.ac.uk, lsacks@ee.ucl.ac.uk]

Abstract: Emergence of process-intensive and bandwidth-hungry e-Science contexts ranging from materials simulation to physics measurements is expected to result application which need to use several multi-Giga bits per second flows and increasingly higher data rates. This paper covers the challenges in measurement and analysis systems for such networks and describes the architecture of a simulation test bed for storage of network data for long-term and short-term feature extraction and traffic monitoring which are discussed in the UKLight measurement and monitoring project, MASTS.

1. Introduction

High speed optical switching allows for high data rates between several exchange points without the complex routing and IP address and port number analysis. The UK E-Science community, an industrial-academic collaboration initiative, has recently been active in the area of grid networks and high performance computing applications. The purpose of this initiative is to allow researchers throughout UK and Europe to be able to work on multi-site projects generating huge data sets and requiring CPU resources more than those available to a single university or company. Examples of these include the GridPP [1] project at the international particle research laboratory in Geneva, known as CERN, which from year 2007 will begin to produce several peta bytes of data per year, all of which must be made available to physicists' worldwide.

Data volumes like such as these will take a long time to be carried over normal internet links with standard TCP/IP characteristics due to the TCP slow-start and congestion avoidance algorithms. As a result there is need for design and implementation of more efficient protocols for high bandwidth links. An important proposal regarding this matter was the development of UKLight. UKLight is a national facility in UK to support projects working on developments towards 10Gbps optical networks and the applications that will use them. UKLight primarily consisted of optical fibre links between University of London Computer Centre, London Point of Presence, NetherLight [2] in Amsterdam and Starlight in Chicago. The links are now extended to more UK universities and UKLight on certain links carries research and production traffic. UKLight is a switched network which makes it unique in the sense that there is no queuing and routing involved.

From the operator and users' perspective, The monitoring and measurement of network events is an important issue within test networks such as UKLight, where researchers and individuals are allowed to run arbitrary variations of transport, session control and network layer protocols. The UKLIGHT project is envisaged as a platform upon which a rich mix of traffic and data will flow. The users of UKLIGHT are expected to trial many new technologies. For example: ECN, Fast-TCP, Reliable multicast, DiffServ, MPLS, and IPv6 [3].Clearly there is a need for tools to evaluate their effectiveness and assessing their impact on the network as a whole. Equally important is the need for work to enable further research and infrastructure offerings which can be enabled through the provision of a system that allows both the interpretation of new service deployments and third-party access to collected data and characterisation.

2. Monitoring and Measurement of UKLight

Traditionally, flow metrics have been analysed using tools such as NetFlow. NetFlow provides valuable information about network users and applications, peak usage times, and traffic routing [4]. In applications where there is emphasis on packet level data collection, TCP packet trace files are usually collected using variations of libpcap and tcpdump [5]. Tcpdump allows capturing of Ethernet frames at the Network Interface Card (NIC) and stores the data in ASCII format text files.

However these tools are not capable of handling the high data rates and possibly flow rates on UKLight. Many comparative experiments have been carried on using tcpdump on various systems and on a typical 100Mbps line the capture rates declines rapidly after about 60Mbps send rate [6]. However MASTS project's objective is to capture packets at full receive rate and store header data in archive files. The captured files are then archived to allow for different user-defined queries to be applied on the data. This allows statistical analysis on flow arrival rates, job durations and link transfer times. The objective of this exercise is to let the UKLight user community to be able to view the status of the system over a long period of time and observe how the network topology and usage evolves over time.

3. Simulation platform for the monitoring system

The original monitoring architecture proposal for the UKLight is displayed in figure 1.

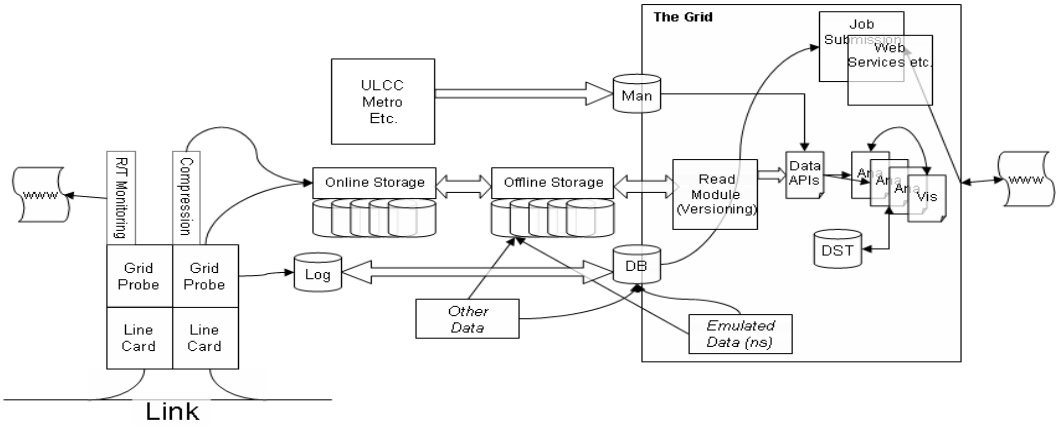

Figure 1: UKLight monitoring architecture

The frames are captured using probes, and stored in temporary online storage which will enable real time monitoring of system performance. These files are further stored in an offline archive for long-term analysis purposes explained in the next section. The location, read number, versions and other details of these files are made available to the community via web services interface.

The packet capture process on 10G links is a challenge for today's hardware and software products. OC192 and OC48 network cards to facilitate 2.5 to 10 Gbps are extremely expensive. High-Speed storage is another important issue. Maximum frame rate of a fully utilised 10Gbps link can be calculated using minimum payload example. The minimum frame payload on Ethernet is 46 Bytes. When there is no congestion or collision (which is the case on a circuit switched network like UKLight) a frame consisting of 72 Bytes with a 9.6 µs inter-frame gap (corresponding to 12000 Bytes at 10 Gbps). The total frame period is hence 12072 Bytes. This is equivalent of 103545 frames per second at high utilisation. At such high frame rate, even storing the 12 bit source and destination MAC address will result into about 10 Mbps hard disk write rate which will lead to more than 820 Giga bits of data every day on a single link.

In order to overcome the high frame rate capture challenge one of the possible s solution is to split the 10Gbps link traffic into 1Gbps streams and feed each one of these links into a probe. These probes can be high-end PCs, with Giga Bit Ethernet card and high storage capacity. The frame headers can be captured in the form of "rolling log files" which can then be pipelined into a Storage Area network. However introduction of this form of splitting creates scenarios which are looked at in the rest of this paper by use of simulation. The data splitting can be done by a "Round-Robin" scheme where packets are equally distributed between the 1Gbps links, or it can be done based on per-link utilisation allocation where links are filled up one by one so if the traffic at certain times is less than 1Gbps only one of the pipes are used. The packets, at the time of splitting, should be time-stamped to a very high precision and stored in the probe archives. Upon re-collection, frames can be read in the form of linked list of 100s or 1000s of frames per iteration from each link. In order to overcome re-ordering saw tooth at the edges of linked lists, after sorting the packets based on time-stamp, 10% of final list members are kept in the buffer and added to the head of the next list. This overcomes most misplacement of packets during the splitting and sorting stage.

The process of capture and store of frame headers even at 1Gbps speed is cumbersome. Researchers at Intel laboratories have been working on a continuous Monitoring project called CoMo [7]. CoMo has been designed to be the basic building block for a network monitoring infrastructure that will allow researchers and network operators to easily process and share network traffic statistics over multiple sites. The architecture of CoMo, as shown in figure 2, is designed to compute and report various performance metrics while sustaining high speed traffic collection. CoMo also provides a query interface to allow users to elicit the system to export the results of the measurement performed. The suitability of CoMo for this project is currently being investigated by the authors and researchers at Loughborough University. The capture rates achieved are up to around 700 Mbps.

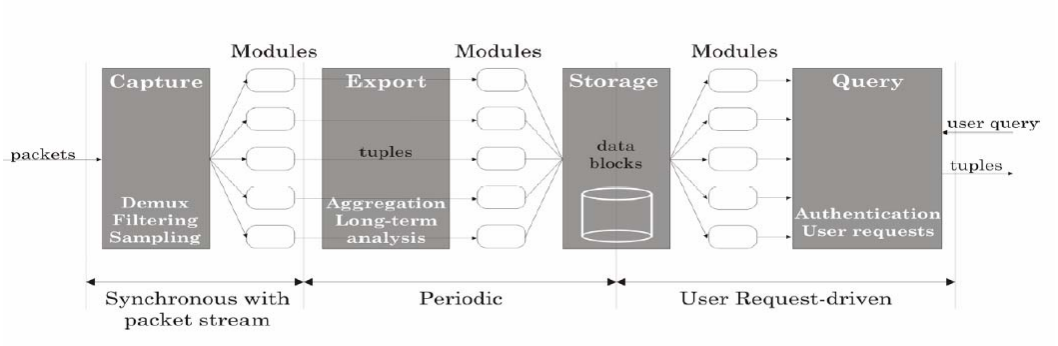

Figure 2: CoMo Architecture (Figure courtesy of Intel Research [7])

The purpose of this exercise is to gather data for statistical analysis. Due to data protection act and privacy issues, it is not always possible to tap network data. However it is required in early stages of the project to be able to test the feature extraction and wavelet compression algorithms, which is the research focus of the authors. A simulation platform for the network and the hardware behaviour of Ciena [8] switches and optical fibre links which are used in the network. Figure 3 displays the architecture of this simulation platform. The simulator is able to reproduce the given traffic characteristics and these are verified using algorithms and graphs. The simulator has been designed to emulate the exact binary characteristics of Ethernet frames, including the header formats, preamble and checksum. IP header data is extracted from the information available in the NS2 trace files and the presentation format precisely follows the IP packet header format as described in IETF Internet Protocol specification RFC 791.

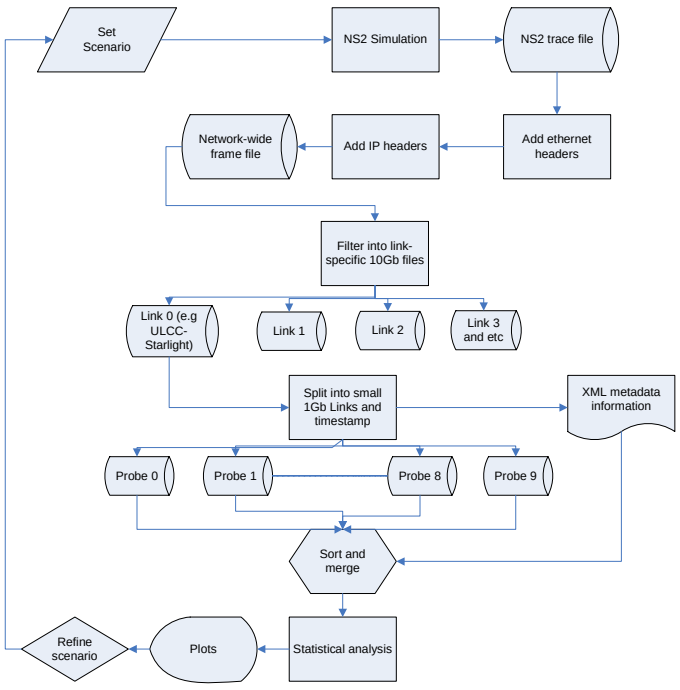

Figure 3: Simulation platform

The traffic scenarios are generated using a network simulator of choice. After testing a few simulators including J-Sim [9] and OmNeT++ [10], authors used NS2 [11] simulator due to its simplicity and wide community of users. The produced trace gives simple traffic information:

```
r 6.043611 0 1 exp 40 ------- 1 0.0 1.2 0 0
```
[time, source, destination, size, flags, nodes, flow ID..]

NS2 generates various forms of traffic distributions like exponential and Pareto. The simulator generates traffic data will enable quite early on to put in place a number of off-line analysis and visualization tools, etc. Each 10G link is broken down to 10 links which are probed and captured. Time-stamping of the packets is done at the switching point where data is split into 1Gbps links. This process will lead to 80[1] archive files. These experiments are carried on with the various possible scenarios to ensure the possibility to analyse the collected data. It is vital to be able to keep a record of these files for each "run" of the probe and capture process. An XML file is produced after the archiving stage of the project which is passed onto the web services interface. User queries can use this XML schema to access the details, location and time properties of various archives within the storage area. The use of this catalogue of archived data for consistent indexing of the data repository enables efficient and effective searches of the data repository Implemented as a relational database.

4. Conclusions and future work

Generation of traffic files has enabled the project to get a head start on feature extraction methods. Due to the massive data sets which will have to be dealt with in this architecture, use of standard tools such as time-series analysis is hardly useful. However using methods such as Wavelet analysis the research community is able to view the large scale characteristics of traffic. The next step on this work is to develop the wavelet analysis method over very large datasets and identify the optimal window size and compression factors for different sets of activities within the network.

Acknowledgment

The authors would like to acknowledge advice and support from Dr Eduarda Mendes Rodrigues and Dr Andrew Moore. This work is supported by EPSRC grant GR/T10503/01.

References

1. “A Grid for Particle Physics - Managing the Unmanageable”, D. Britton, A.T.Doyle, S.L.Lloyd UK e-Science All Hands Conference, Nottingham, September 2004.
2. An Introduction to NetherLight: http://www.netherlight.net/
3. “Case for Support: UKLIGHT Monitoring and Analysis at many Scales”, L Sacks, S Bhatti, D Parish, I Philips, A Moore, R Gibbens, I Pratt, I Graham
4. C. Estan and G. Varghese, "New directions in traffic measurement and accounting," in Proceedings of the 2001 ACM SIGCOMM Internet Measurement Workshop, pp. 75--80, (San Francisco, CA), Nov. 2001
5. TCPDUMP Public Repository: http://www.tcpdump.org
6. “TCPDUMP Subsampling Problem on the Deterlab Testbed”, Soranun Jiwasurat, The Pennsylvania State University
7. “The CoMo White Paper”, Gianluca Iannaccone, Christophe Diot, Derek McAuley, Andrew Moore, Ian Pratt, Luigi Rizzo, Intel Research technical Report, Intel Research, Cambridge, UK
8. “Ciena Multi-Service Optical Switching Core director datasheet”, http://www.ciena.com/files/CoreDirector_PB.pdf
9. “Discrete event simulations with J-Sim”, Jaroslav Kačer, University of West Bohemia, Proceedings of the inaugural conference on the Principles and Practice of programming, 2002 and Proceedings of the second workshop on Intermediate representation engineering for virtual machines, 2002
10. OmNet++ Community site: http://www.omnetpp.org/
11. The Network Simulator - ns-2: http://www.isi.edu/nsnam/ns/

[1] * (4 x 10 Gpbps links x 2 direction on each link x 10 probes attached to each directional link)