Network Traffic Analysis of the Internet 2 using OSU Flow-Tools

Abhishek Veldurthy
aveldurt@andrew.cmu.edu

Divakar Goel
diwakarg@andrew.cmu.edu

Hazel Diana Mary
hmary@andrew.cmu.edu

Pedro Leon
pgl@andrew.cmu.edu

Abstract
Since its origin in the ARPANET, the Internet has evolved to become a global multipurpose network. Its main participants range from educational, government, research to commercial organizations. These participants have different communication needs and thus generate different types of network traffic patterns. Furthermore, the Internet2, founded in 1996, was designed to provide next-generation production services as well as a platform for the development of new networking ideas and protocols. Therefore, it turns out of interest to analyze the traffic patterns on this new cutting-edge public network. In this paper, we analyze and compare the peering and non-peering traffic of educational and federal organizations. Furthermore, we provide inferential reasoning for bandwidth utilization, routing traffic, multicast and some of the security attacks on specific participants. All the analysis is done using OSU-Flow tools on the sampled flows collected on January 11, 2007.

I. Introduction

We focused on the exhaustive analysis of traffic flowing through the Abilene’s router located in Washington. The general objective is to infer behavioral traffic patterns of the Internet2, based on the analysis and comparison of traffic generated by different kinds of organizations such as universities, research and commercial agencies. Furthermore, the results of this project could help Network Administrators on the capacity planning process and on the detection of security threats.

The chosen participants and the analyzed scenario are shown in table 1 and figure 1, respectively.

<table>
<thead>
<tr>
<th>Type</th>
<th>Connector</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educational</td>
<td>3ROX</td>
<td>CMU</td>
</tr>
<tr>
<td>Educational</td>
<td>3ROX</td>
<td>PSU</td>
</tr>
<tr>
<td>Government/Research</td>
<td>MAX</td>
<td>NSF</td>
</tr>
<tr>
<td>Government/Research</td>
<td>MAX</td>
<td>NASA</td>
</tr>
<tr>
<td>Private/Research</td>
<td>Peer</td>
<td>NREN¹</td>
</tr>
<tr>
<td>Private/Research</td>
<td>Peer</td>
<td>GEANT²</td>
</tr>
</tbody>
</table>

Table 1. Chosen participants

To analyze their respective traffic, Non-peering participants are identified by IP-address range and Peers are identified by the interface they are connected to on the Washington router. Further, because the collected flows have the last 11 bits masked, we had to guarantee that the ip-address ranges were unique for each chosen participants (i.e. possible subnets inside the 11 masked bits were not assigned to another participant on the Internet 2).

Our traffic analysis is focused on answering some specific questions which can be grouped in three types: Traffic patterns of peering and non peering participants; IRP, ERP¹, multicast and DNS traffic; and Network Performance and Security related traffic.

In the following sections, we present graphical results comparing traffic patterns under different conditions (e.g. time, inbound/outbound, peering/non peering, network, transport and application layer levels). We also present detailed results regarding multicast, routing and DNS traffic patterns as well as some comments about the possible reasons why Indianapolis Router was removed from the Abilene’s network.

¹ National Research and Educational Network.
² A multi-gigabit pan-European data communications network, reserved specifically for research and education use.

¹ Interior and Exterior Routing Protocols.
II. Methodology

To perform our analysis we performed the following steps:

- Comprehensive research of the Internet2, including its goals, participant and network infrastructure.
- Extensive use of Net-Flow Tools.
- Use of Perl scripting language to perform traffic time analysis and automate data extraction.
- Consulting of security forums, IANA4 and IETF5 web sites.

III. Peering/ Non peering traffic patterns

We analyzed the protocol wise breakup of total traffic on the Washington Router on an hourly basis.

![Traffic analysis based on Transport layer protocols](image)

The traffic percentages of various protocols are fairly steady with TCP contributing around 87.6%, UDP - 11.9%, ICMP - 0.078%, GRE - 0.133%, ESP - 0.177%, AH - 0.0214% and other protocols less than 0.1%.

The following inferences can be obtained from observing the detailed analysis of the various protocols:

- Traces of IPMP protocol shows that there are some multi-homed hosts which use Washington router for one of their interfaces.
- Some Virtual Router Redundancy traffic can also be seen. Some institutions connected to this router have configured a virtual router to increase the availability for the default gateway that services their subnets.
- Protocol Independent Multicast routing traffic is seen and hence in addition to actual multicast, the hosts are also doing multicasts over the existing unicast framework.
- There was a reasonable amount of security associations in the various flows through these routers which can be established by the presence of authentication headers (AH) and encapsulation security protocols (ESP), commonly used with IPSec.
- There was quite some IP tunneling traffic contributing towards the generic routing encapsulation (GRE) traffic.
- Signaling for resource reservation has been done through the RSVP protocol for unicast and multicast data packets.

![Hourly Traffic Pattern (Washington Router)](image)

The analysis of the actual traffic at the Washington router shows a drastic increase in traffic after 2pm and a peak in traffic between 5-6pm. This high usage trend goes on until about midnight showing an increased usage of bandwidth in the after lunch hours and in the time when users may not be at their work place. The least traffic is observed around 7am to 8am.

a. Non-Peering (Educational and Federal Institutions)

We performed a comparative study of the incoming and outgoing traffic flowing across educational and federal institutions. With educational institutions being represented by Carnegie Mellon University and Pennsylvania State University and federal institutions...
represented by NASA and NSF, the following results were obtained from the network logs.

Despite the fact that our selected federal institutions are as research oriented as the educational institutions, the traffic pattern varied greatly. From Figures 4 and 5 we infer that educational institutions have a lot of incoming traffic possibly due to the incoming data from the various web servers that the students access. Federal institutions on the other hand show a lot of outgoing traffic. This could possibly mean that these institutions have a lot of servers which export data to other federal agencies.

Pennsylvania State University shows higher levels of incoming and outgoing traffic as compared to Carnegie Mellon University. This could possibly suggest higher usage of applications outside the local intranet by PSU as compared CMU or higher levels of research collaborations and data exchange by PSU than CMU.

Table 2 reveals that almost all of the applications utilize either TCP or UDP at the transport layer. Furthermore, TCP is the most popular transport protocol with more than 89% of the input and outbound traffic. Therefore, it seems that the research done by the Internet 2 participants is mainly based on new applications but not on new transport protocols.

<table>
<thead>
<tr>
<th></th>
<th>% TCP</th>
<th>% UDP</th>
<th>% Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inbound</td>
<td>97.25</td>
<td>2.30</td>
<td>0.45</td>
</tr>
<tr>
<td>(Total: 3146 Giga-Octets)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outbound</td>
<td>89.20</td>
<td>10.36</td>
<td>0.44</td>
</tr>
<tr>
<td>(Total: 2930 Giga-Octets)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Inbound/Outbound Transport Layer Peering Traffic

When analyzing application level peering traffic, we found of particular interest the applications used by NREN. Figure 6 depicts NREN’s most popular applications for incoming traffic.

Note that all these applications correspond to traffic which is representative of Microsoft Local Area Networks (e.g. NETBIOS, MS Domain Server, SQL Server, VNC Server, etc.). Figure 6 suggests that NREN is using its Internet link to connect two LANs geographically dispersed. Further, we consider that running these applications on a public network represents a big security hole on NREN as NetBios is a legacy protocol used to transmit files in LANs, which is well known to have a lot of vulnerabilities.

6 We consider the most popular applications as the ones that generate the most flows.

7 http://isc.sans.org/port.html?port=445
On the other hand, when analyzing outbound traffic, we found that more than 90% of the NREN outbound traffic corresponds to PMCP\textsuperscript{8} (Port 3821). Moreover, the remaining outbound traffic is generated by a particular application running on TCP port 4853.

PMCP is a recently created ATSC\textsuperscript{9} Standard used in video broadcasting through data networks. This result suggests that NREN was performing research on this new protocol broadly used on digital television.

c. Peering and Non Peering Traffic

In order to compare transport layer traffic between a non-peer participant and a peer one, we looked at the CMU and GEANT traffic, respectively. Figures 7 and 8 illustrate the behavior of the most popular TCP applications for outgoing traffic.

\begin{table}[h]
\centering
\begin{tabular}{|l|c|}
\hline
Application & Avg. size (octets) \\
\hline
HTTP & 82.48 \\
MSN Messenger & 87.89 \\
HTTPS & 548.71 \\
SMTP & 797.33 \\
TOR, DBGP & 609.80 \\
NNTP & 1008.53 \\
SSH & 926.54 \\
BitTorrent & 356.33 \\
IMAPS & 93.53 \\
Openfire Admin Console & 636.98 \\
\hline
\end{tabular}
\caption{TCP Most Popular Applications (CMU outbound traffic)}
\end{table}

These two figures reveal that HTTP is the most popular application in both of them. Further, in CMU we found MSN Messenger as the second most popular application, even above SMTP (email) traffic. This result demonstrates that real time applications, such as Messenger are gaining more popularity within the Internet Community. Furthermore, though P2P traffic normally utilizes random ports, we were able to find BitTorrent traffic.

On the other hand, in GEANT we found again applications that are commonly used in LANs, such as Active API Server, SSC-Agent and SQL Server. It is important to highlight that these applications are not presented in non-peering traffic.

Finally, Table 3 lists the average packet sizes for the most popular applications used in CMU.
HTTP outbound traffic has the lowest average packet size because this traffic basically correspond to the requests user inside CMU network make to Web Servers on the Internet. Conversely, we also found that the average packet size of HTTP inbound traffic is of the order of 1500 octets. This result is because users usually download files from different Web Servers on the Internet.

As shown in Table 3, Network News Transfer Protocol (NNTP) has the highest average packet size of all the listed applications. This protocol is used by clients and web servers to manage the information posted in newsgroups on the Internet.

The economics behind computer networks predicts the abundance of peering traffic over non-peering traffic. The data which was obtained verified this prediction with peering traffic totaling to about 4 times that of non-peering traffic (inbound + outbound).

The interesting pattern here is that the outbound and inbound peering traffic are quite similar which is the expected behavior for peering relationships.

### IV. IRP, ERP multicast and DNS traffic

On this section we show the routing, multicast and DNS traffic results. To find all the IRP traffic we looked at the IP headers. Traffic corresponding to Internal Routing Protocols should have on the Protocol Field one of the following values: 9 (IGRP), 88 (EIGRP), 89 (OSPF) or 124 (ISIS over IPv4).

We found (600.4 K-Octets) of OSPF traffic. We actually know that Abilene’s routers use ISIS. However, the Internet 2 is a research network. Then, this traffic might be because network administrators were trying to test OSPF as another option to perform internal routing.

On the other hand, we found that the only External Routing Protocol used between Abilene’s Routers and Internet2’ participants is BGP.

In Figure 10, we plot the hourly distribution of BGP Traffic. This traffic remains almost constant throughout the day around 700 K-Octets. This might be because routers using BGP as their routing protocol, constantly exchange Keep-alives messages. Furthermore, we can see that these messages do not generate a big amount of traffic so that they do not affect the performance of the network at all.

Figure 13 summarizes the top 10 participants that generated multicast traffic. This figure shows that one single participant (National Center For Supercomputing Applications) generated almost 50% of all this traffic. We can also notice that this participant did not make many connections (flows) but it transmit a lot of traffic (octets). Moreover, if we look at the data for “others participant” we can notice that the remaining multicast traffic was generated by many participants who actually transmitted small quantities of octets.

Furthermore, during our analysis, we found that 99% of the multicast traffic was UDP. This should be because multicast is usually used with delay sensitive applications such as videoconference or Financial information Services (e.g Bloomberg), in which TCP is not suitable. The remaining 1%, correspond to PIM\(^\text{10}\), OSPF and ICMP traffic. It is important to

\(^{10}\text{Protocol Independent Multicast.}\)
mention that PIM is a signaling protocol between routers configured to support multicast traffic. The transport layer break up for multicast traffic is presented in figure 11.

In Figure 12, we plot the hourly distribution of DNS Zone Transfers. Note that the amount of this traffic is approximately 6000 times higher than BGP traffic. Therefore, if this traffic occurred during the day, it might cause network congestion.

This result suggests the reason why the DNS Zone transfers were scheduled to occur at night time.

Figure 14 depicts the Top Ten participants that generated DNS Zone transfer during this day. This figure reveals that the main responsible of this traffic was Purdue University, who generated more than 90% of it. Further the percentage of flows is nearly zero. This result suggests that this participant generated only one Zone Transfer with a high rate of octets transmitted. In contrast, University of Maryland performed several DNS Zone Transfers but with low rate of transmitted octets.

V. Network Analysis and Performance

As a part of this section, we analyzed some of the interesting observations on the traffic at Washington Router and provided some solutions that would address key issues in capacity planning and bandwidth utilization of the router’s interfaces. This section also proposes some of the possible reasons why Indianapolis router was removed from the Abilene’s Backbone Network.

a. Mapping of traffic at all interfaces of the WASHing router:

As a part of our study we looked the total data transferred through various interfaces of the Washington Router. Figure 16 depicts a complete dissection of the interfaces of Washington router. We observe that the maximum traffic flows through the two backbone routers New York and Atlanta; these have been detailed in the following section. We further analyzed the data flowing in through each interface and its corresponding split up into the output interfaces.

b. Bandwidth Utilization of the backbone links at Washington Router:

The hourly traffic patterns of the two main backbone links WAS-NYC (Interface 113) and WAS-ATL (Interface 49) showed that they together range from 50% to 65% of the total traffic flowing through the Washington router at all hours with maximum traffic at 18:00 and 22:00 hours as shown in Figure 15.
c. Possible reasons for removing Indianapolis router:

While observing the traffic flow at the three backbone links of Indianapolis router namely IND-CHI (Interface 46), IND-ATL (Interface 110) and IND-KAN (Interface 24) we found that the traffic at the interfaces of these links had certain obvious similarities. 82% of the traffic flowing from Indianapolis to Chicago came from Atlanta and Kansas, 87% of the traffic flowing to Atlanta came from Chicago and Kansas and finally 79% of traffic flowing to Kansas came from Atlanta or Chicago.

We could make an inference here that a significant amount of traffic on the Indianapolis router was not from its own peers or participants but from the three backbone routers. Considering the fact that a router at Houston connected Atlanta and Kansas, a backbone link from Atlanta to Chicago and Kansas to Chicago would considerably reduce the load on Abilene to have a backbone router at Indianapolis.
d. Analysis of traffic flowing to Berkeley:

We studied all routers and the interfaces to which traffic to AS25 is directed. Fig. 18. graphically represents the path taken by traffic directed to AS25 at each node. On detailed analysis we found that all traffic for Berkeley (through CalREN) is routed through LOSA Ang even though SNV An is close to Berkeley. As shown in Fig. 19 all traffic from DENV and STTL are first routed to SNV An, which forwards this traffic to LOSA Ang for delivery to CalREN. This may explain why SNV An was removed as a backbone router owing to the number of participants.

<table>
<thead>
<tr>
<th>Router</th>
<th>Output Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATLAn</td>
<td>BACKBONE: oc192 to HSTNng</td>
</tr>
<tr>
<td>CHINng</td>
<td>BACKBONE: OC192 to IPLSng</td>
</tr>
<tr>
<td>DNVRng</td>
<td>BACKBONE: OC192 to SNV An</td>
</tr>
<tr>
<td>HSTNng</td>
<td>BACKBONE: OC192 to SNV An</td>
</tr>
<tr>
<td>IPLSng</td>
<td>BACKBONE: OC-192 to KSCYng</td>
</tr>
<tr>
<td>KSCYng</td>
<td>BACKBONE: OC-192 to DNVRng</td>
</tr>
<tr>
<td>LOSA An</td>
<td>CalREN South</td>
</tr>
<tr>
<td>NYCMng</td>
<td>Interconnect to CHINng</td>
</tr>
<tr>
<td>SNV An</td>
<td>BACKBONE: OC-192 to LOSA</td>
</tr>
<tr>
<td>STTLng</td>
<td>BACKBONE: OC-192 to SNV An</td>
</tr>
<tr>
<td>WASHng</td>
<td>BACKBONE: OC192 to ATLAn</td>
</tr>
</tbody>
</table>

Table 4: Outgoing interface for Berk(AS25) at each router

Figure 18: How traffic flows to Berk (AS25) at each router?

Figure 19: Zoom in of the California region.

VI. Traffic anomalies and Security

a. IP Spoofing:

To identify any IP spoofing attacks carried out on CMU IP address range we analyzed the traffic at different interfaces of the WASHng router that had source IP Address of CMU. The Washington Router primarily uses interface 28 to send and receive CMU’s traffic flows. However we found 30 flows (with the 1 in 100 sampling) with source subnet address in CMU’s range flowing in through interface 113(from NYCMng). We further observed that all the flows had only UDP traffic.

Considering, that UDP does not provide any protection against IP spoofing, (as no sequence numbers / handshakes are required) this observation seems to be a genuine case of IP spoofing. However, if some addresses in CMU’s range are connected via NYC Mng this may be a false positive but considering that there were very less number of flows comprising of only 1 packet
each, this does not look to be genuine for CMU’s range in New York. Moreover, observing traffic to
and from port 53 suggests that DNS flooding and DNS spoofing respectively may have been
attempted.

b. Host Scanning and Port Scanning:
Our study was carried out using the flow-dscan
tool which gives a positive for host scanning if the
count of the length of the destination IP hash chain
exceeds 64 (with IP address masking, we cannot
say which host but can infer results only for the
subnet). Also, port scanning positives are returned
when the bitmap of the destination port number <
1024 per destination IP goes above 64. The data
used for this analysis excluded multicast and www
traffic.

Considering our focus organizations, CMU, PSU,
NASA Goddard and NSF we got the following
results: (No incoming scans were observed)

<table>
<thead>
<tr>
<th>Organization</th>
<th>Outgoing Scans</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMU</td>
<td>28 scans</td>
</tr>
<tr>
<td>PSU</td>
<td>25 scans</td>
</tr>
<tr>
<td>NASA</td>
<td>3 host scans</td>
</tr>
<tr>
<td>NSF</td>
<td>2 host scans</td>
</tr>
</tbody>
</table>

Table 5: Host Scans from institutions.

c. Interesting Observations for Port 22
(SSH) traffic at CMU:
We analyzed the traffic for several of the popular
ports at CMU to study the patterns characteristic to
their default applications. The pattern for SSH
(Port 22) was quite interesting to observe (Figure
20), the requests originating from CMU are
usually less than the responses. But, in-between
hours 15-19 on Jan 11, 2007 there was a surge in
the requests being made.

We believe this was a result of some research
work or a stress testing activity.

VII. Conclusion
During the analysis of the Internet2 network we
were able to compare traffic patterns between
different participants as well as observe the
behavior of different network framework protocols
such as routing, multicast and the DNS protocols.

We found that the most popular transport layer
protocols remain to be TCP and UDP. This allows
us to infer that the research of new protocols is
being done at the application layer. Furthermore,
peering traffic patterns are considerably different
from non-peering traffic. The analysis of the
NREN’s outbound traffic permits us to infer that
NREN was performing research on Programming
Metadata Communication Protocol which is used
to transmit digital television. We further
understand that NREN was doing research on
traditional LAN protocols over public networks
since the use of NetBios, MS Domain and SQL
Server protocols is not common in public
networks.

On the other hand, the increasing development of
web applications has made the HTTP protocol the
most popular in both peering and non-peering
traffic. Moreover, real time applications such as
instant messengers are gaining popularity over the
traditional e-mail. We also found that most of the
UDP traffic is generated by multicast applications. We can infer that this is because real time applications, which are highly delay-sensitive, are common in multicast traffic (e.g., videoconferences).

Finally, our time analysis of the backbone links at the Washington router demonstrates that there were no congestion problems at all. We further can infer that Abilene network administrators perform a constant analysis of the network since they detected that Indianapolis’ router was no longer needed. Also, evidence of good traffic paths was found when analyzing the traffic toward Berkeley. In addition, the security results do not show convincing evidence that security attacks were performed on this day against the chosen participants.

VIII. References

http://www.internet2.edu/
http://www.splintered.net/sw/flow-tools/docs/
http://www.mcnc.org/
http://www.atsc.org/
http://www.ietf.org
http://www.sans.org
http://www.iana.org/assignments/port-numbers
http://www.iana.org/assignments/protocol-numbers/
http://www.geant.net/
http://www.internet2.edu/international/intl_connec
http://about.bloomberg.com/