UH - PACCOM Traffic Analysis

A preliminary look at a lot of data and some proposals

David Conrad
University of Hawaii, Manoa
Information and Computer Sciences Department
2565 The Mall, Keller 304A
Honolulu HI, 96822 USA
(808) 956-5335
davidc@hawaii.edu

Overview

UH/PACCOM Data Collection Project

- Goals
- Methodology
- Preliminary Observations

Analysis of the UH/PACCOM Data Project

- Standardized Data Format Proposal
- UH/PACCOM Data Analysis Proposal
- Data Collection Coordination Proposal

Conclusions

Data Collection

Goals

Allow for the determination of

- Where traffic is coming from and going to
- What services are being used and for how long
- What patterns of usage exist and the periodicity of those patterns

Data Collection

Methodology

- Attach a collector to the bridge between PACCOM routers
- Use a modified tcpdump to collect packet headers
- Archive collected headers

Data Collection

<PACCOM Topology>

Yes, yet another view of the PACCOM topology

This map shows UH's position in this data collection effort

You'll note that the UH - NASA Ames link went from 512 Kbps to T1 during the collection period.

This map indicates the network topology during the collection. It doesn't correspond to reality anymore; after the collection ended, KAIST was moved to Ames.

Anyway, the main point of this map is that UH sits between Japan and Korea and points East

A little more detail on the area in this box...

Data Collection

<Hawaii.Net topology>

This map show the topology of the Hawaii.Net ethernet which connects the PACCOM routers.

Cisco connected KAIST and Nagoya

Menehune connected ICOT, WIDE, and TISN

And imp connected everybody to the US mainland

Chantry is our primary nameserver

Phoenix and Zeus connect the UH LAN to the rest of the world

Lurker is the collection machine

Data Collection

<Traffic Collector Topology>

lurker was multi-homed, existing both on the crossbar ethernet and another ethernet so data could be transfered over the network with no impact on what we were collecting

lurker had a 400 MB disk which had about 300 MB free for data collection.

When lurker’s disk reached somewhere between 50% to 75% full, the data was transfered to net, compressed, and stored on one of two staging disks.

The staging disks were then dumped to Exabyte tape when convenient.

Data Collection

Data collection started on Dec 2, 1991 and ran continuously (more or less) until Jun 2, 1992

Results:

- 40 Exabyte tapes and 13 Optical Disks
- About 90 GBytes of compressed packet headers (about 175 Gbytes uncompressed)

Preliminary Observations

Disclaimer:

- These observations are highly preliminary and their accuracy should not be relied upon too heavily.

General Statistics:

- Approximately 1.4 billion packets
- Approximately 280 GBytes of data
- Over 100,000 IP hosts addressed
- Over 5000 IP networks addressed

Preliminary Observations

<Crossbar>

To menehune: ~ 22% / ~ 27%

To imp: ~ 43% / ~38%

To cisco: ~11.5% / 16%

To UH: ~14.5% / 16%

Preliminary Observations

Router	Packets	%	Bytes of Data	%	Avg
imp	648,104,477	45.206	155,503,433,367	55.213	239.9
menehune	344,381,377	24.021	60,221,936,717	21.382	174.9
UH	266,817,968	18.610	46,319,441,768	16.466	173.6
cisco	144,914,049	10.108	15,978,558,870	5.673	110.2
Totals	**1,404,217,871**	**98.336**	**278,023,370,722**	**98.928**	

Hawaii.Net Sources (sorted on Bytes of Data)

Router	Packets	%	Bytes of Data	%	Avg
imp	651,089,028	45.414	101,308,940,457	35.971	155.6
menehune	321,331,644	22.413	75,808,829,837	26.917	235.9
UH	240,308,993	16.762	46,817,685,421	16.623	194.8
cisco	170,781,439	11.912	46,181,893,088	16.397	270.4
Totals	**1,383,511,104**	**96.501**	**270,117,348,803**	**95.908**	

Hawaii.Net Destinations (sorted on Bytes of Data)

Preliminary Observations

<Ethernet packet distribution pie chart>

As you can see, IP makes up the vast majority of packets and data

Preliminary Observations

Type	Packets	%	Bytes of Data	%	Avg
IP	1,351,204,501	94.238	267,245,864,243	94.885	197.8
DECNet	66,308,217	4.625	13,351,904,836	4.741	201.4
Proteon	11,116,058	0.775	666,963,480	0.237	60.0
Wellfleet	2,836,447	0.198	170,186,820	0.060	60.0
Loopback	1,431,910	0.100	134,599,540	0.048	94.0
Unknown	864,033	0.060	80,766,436	0.029	93.4
DECMOP	56,710	0.004	3,402,600	0.001	60.0
Totals	**1433817876**	**100.000**	**281,653,687,955**	**100.000**	**196.4**

Ethernet Packets by Type (sorted on data)

Preliminary Observations

<IP breakdown Pie chart>

Looking at IP data, we see the following...

UH PACCOM is a little higher in TCP traffic than other WANs which have been studied (usually report about 70% to 80% of the traffic is TCP based).

In the tables that follow, the discrepancy that exists between the Ethernet values and the IP breakdown values is due to a missing IP data set.

Preliminary Observations

IP Breakdowns by Protocol:

Protocol	Packets	%	Bytes of Data	%	Avg
TCP	1,098,813,675	82.300	219,854,217,402	90.885	200.1
UDP	184,045,658	13.785	18,025,698,159	7.452	97.9
ICMP	50,912,144	3.813	3,857,607,261	1.595	75.8
AX.25	1,075,955	0.081	123,136,742	0.051	114.4
IPIP	278,936	0.021	44,251,864	0.018	158.6
Internal	1	0.000	82	0.000	82.0
Unknown	1,031	0.000	57,422	0.000	55.7
Totals	**1,335,127,400**	**100.000**	**241,904,968,932**	**100.000**	**181.2**

IP Protocols (sorted by Bytes of Data)

Note:

- “Unknown” consists of protocols 255 (841 packets/51422 bytes), 254 (85/3060), and 159 (105/2940).

Preliminary Observations

Source	Destination	Packets	%	Bytes of Data	%	Avg
edu	edu	259,532,197	19.656	42,961,186,479	17.975	165.5
edu	jp	124,167,528	9.404	29,830,911,741	12.481	240.2
edu	kr	77,591,654	5.876	22,955,196,298	9.604	295.8
jp	edu	124,186,844	9.405	16,656,519,971	6.969	134.1
jp	com	46,750,125	3.541	9,252,175,757	3.871	197.9
com	jp	40,536,564	3.070	9,048,161,526	3.786	223.2
gov	jp	31,127,213	2.357	5,711,634,670	2.390	183.5
gov	edu	21,959,268	1.663	5,375,550,929	2.249	244.8
kr	edu	61,225,095	4.637	4,701,277,677	1.967	76.8
net	jp	11,532,654	0.873	3,277,698,258	1.371	284.2
Totals		**798,609,142**	**60.483**	**149,770,313,306**	**62.662**	

Top 10 Top Level Domain Source and Destination Pairs (sorted on Bytes of Data)

Preliminary Observations

Domain	Packets	%	Bytes of Data	%	Avg
edu	549901153	48.274	106533365151	52.199	193.7
jp	257499163	22.605	39200417440	19.207	152.2
com	55481378	4.871	13219066704	6.477	238.2
gov	61085942	5.363	12397906846	6.075	202.9
net	36614172	3.214	6739064008	3.302	184.0
Totals	**960581808**	**84.326**	**178089820149**	**87.260**	

Top 5 Top Level Domain Sources (sorted on Bytes of Data)

Domain	Packets	%	Bytes of Data	%	Avg
edu	518263852	45.497	79104538633	38.759	152.6
jp	253957377	22.294	56442906412	27.656	222.3
kr	107964828	9.478	32624074342	15.985	302.2
com	62289413	5.468	11573319488	5.671	185.8
gov	48522451	4.260	4272730674	2.094	88.1
Totals	**990997921**	**86.997**	**184017569549**	**90.164**	

Top 5 Top LevelDomain Destinations (sorted on Bytes of Data)

Preliminary Observations

Transit Traffic

Disclaimer

The following observation should be taken with a grain of salt. They were obtained by looking at the top level domain strings.

Why these values might not mean anything:

- Only about 80% of the IP addresses in the data were resolvable
- Some sites don't follow the domain name convention of putting the country as the top level domain name
- Doesn't take into account transit commercial (.com) and educational (.edu) traffic

Transit Traffic

Observations

- On a random 1 week sample, over 40% of the resolvable traffic both originated from and was destined to a non-US top level domain
- On a random 1.5 month sample, approximately 10% of the resolvable traffic was transit
- Over the entire 6 month data set, only 4.318% of the resolvable traffic both originated from and was destined to a non-US top level domain

Conclusion

More study needed.

Standardized Data Format Proposal

Problem

Many researchers around the Internet are collecting tcpdump-like information. Tcpdump is not quite suitable for this use because:

- Tcpdump is rather wasteful in diskspace
- Tcpdump (by default) collects application layer data
- Host source and destination addresses are available

Solution:

Yet another standard: A common network packet header format

Standardized Data Format Proposal

Based on tcpdump binary savefiles but:

- Packet header + hardware layer information size is reduce from 30 bytes to between 12 and 16 bytes.
- All application data is removed
- All source and destination address are shrouded

Standardized Data Format Proposal

Limitations of Proposed Standard

Since the hardware layer information is compressed, each hardware type must be handled individually.

Hardware types currently supported:

- Ethernet
- FDDI

Packet Types

Since application layer data is removed and addresses are shrouded, each packet type must also be handled individually.

Packet types currently supported:

- IP
- DECNet

A rough draft of this proposal is available. I would appreciate any comments or criticisms you might have.

Data Analysis Proposal

Now that the data has been collected, it's time to do something with it.

Proposed Analyses

- Packet inter-arrival time, size, link layer type, etc. distributions
- More aggregate counts
- Better determination of transit traffic
- IP Service analyses (TCP based, DNS, etc)
- Time Series Analyses

A (very) rough draft of proposed analyses is available. I would appreciate any comments or criticisms you might have

Data Collection Coordination Proposal

This idea behind this proposal originally came from Peter Danzig of USC (danzig%warthog.usc.edu@usc.edu)

Goal:

Organize packet collection efforts and make the results available to network researchers.

Proposal

- A group of people willing and able to perform tcpdump data collection is formed and publicized
- Requests for collection are sent to the group in the form of “when, for how long, and what tcpdump filter to use”.
- If the request is deemed by the group to not violate security or privacy concerns, the request is scheduled.
- The requestor is responsible for supplying whatever archive media (such as 8mm tapes) may be needed.

Conclusions

Data Collection

- Has been performed, and lots of data has been collected
- There doesn’t appear to be that much transit traffic over the long term

Network Data Format Proposal

- Should provide a way of making large packet collections available to researchers without concerns about privacy or security

Data Analysis Proposal

- Plenty of things can be investigated

Data Collection Coordination Proposal

- Should provide a means to coordinate packet collection