Storage and bandwidth requirements for passive Internet header traces

Jörg Micheel[1,2], Hans-Werner Braun[1] and Ian Graham[2]

{joerg,hwb}@nlanr.net
{ian,joerg}@cs.waikato.ac.nz

[1]NLANR MOAT, SDSC, UCSD
10100 John Hopkins Dr
92093-0505 La Jolla, CA

[2]The University of Waikato
WAND, Department of Computer Science
Private Bag 3105
Hamilton, New Zealand

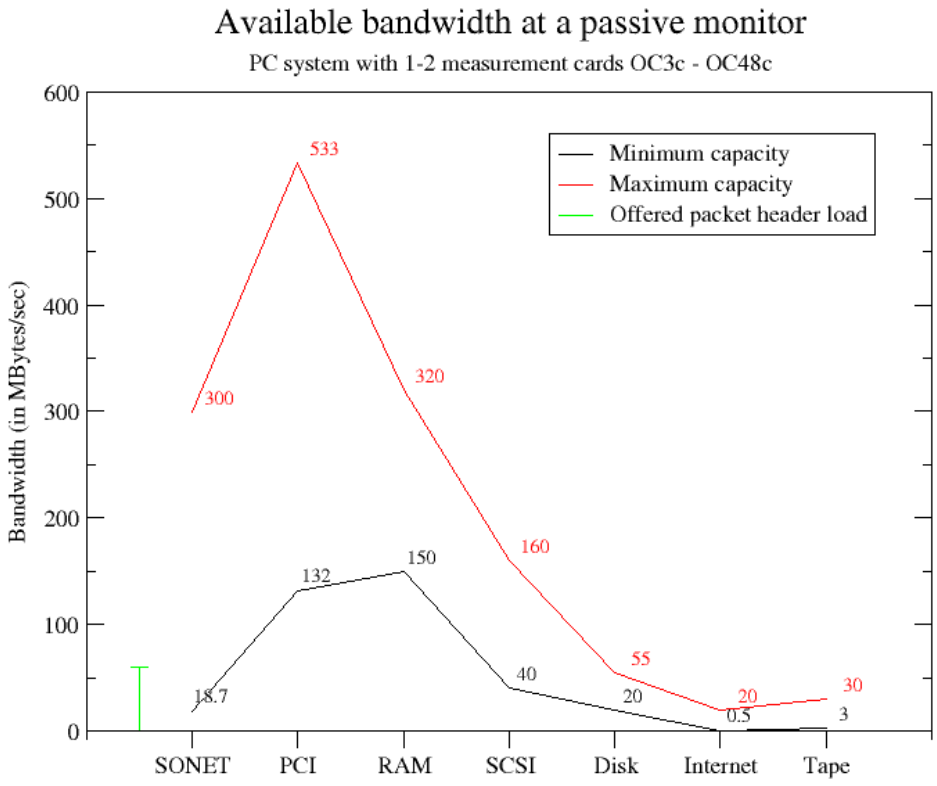
Introduction

In this paper we discuss the storage and bandwidth requirements for data gathered by passive Internet measurements. We consider passive network measurements utilising high-precision time-stamping as the most accurate means to deliver an a view of today's packet network dynamics.

Passive Internet header traces are of high interest to the research community, to better understand the Internet organism, and to guide the development of tools and technologies that meet the demands of applications and end users. Similar passive methods are also used for operational purposes, to track network performance issues, or examine conformance to Service Level Agreements.

The drawback of precision is an enormous amount of data initially produced by the data collection units, in this paper further referred to as network monitors, passive monitors, or simply monitors. The issue of data reduction is central to the approach of a distributed infrastructure of passive monitors.

As the Internet grows it becomes increasingly difficult to accurately track its behaviour. At the packet level, growth happens in two major directions: increased connectivity (more links) and increased capacity (quadrupling of backbone link bandwidth every 18 months to two years). This growth currently appears to happen faster than the growth of other



computer technologies such as CPU speed, cache, memory, disk and tape performance and a drop in price/increase in capacity for all of those. For example, the addition of a single busy OC12c monitor (AIX) to the NLANR PMA infrastructure doubled the total amount of data collected daily from the 17 monitors.

A corollary to the problem is that monitors are connected to the central trace repository by the same Internet links being monitored. It is sensible that the transfer of network measurement data should impact the total amount traffic on those links only by a margin. An obvious bottleneck is the access link to the central data repository.

In the following sections we discuss in detail the steps of data collection and reduction as implemented by the NLANR PMA project and WAND group monitors. Most of the issues discussed in this paper apply to other passive measurements; for instance, based on CISCO NetFlow, while topics related to data collection hardware do not.

Above we depict an overview of a distributed measurement architecture with central trace data repository, such as implemented with PMA [PMA].

Subsystem	Bandwidth (Mbytes/sec)	Comments
OC3c SPE	18.72	
OC12c SPE	74.88	
OC48c SPE	299.52	
PCI bus 32@33[1]	133	Peak rate [PCI95]
PCI bus 64@33	266	Peak rate
PCI bus 32@66	266	Peak rate
PCI bus 64@66	533	Peak rate
FIFO 64@50	400	In and out
SRAM QDR 16@333	666	Peak rate
DRAM PC DIMM 64@133	800	Approximately
RAMbus 64@300	2400	Peak rate
RAMbus 64@400	3200	Peak rate
CPU bus	1000	Peak rate
CPU 32bit read/write to RAM	150-320	Measured
Internet access links	0.5-20	Varies a lot
PC NIC 100BaseTX	12.5	Peak rate
PC NIC Gig Ether	125	Peak rate
SCSI-3 bus	160	Peak rate
Disk drive (IBM 75GXP IDE)	37	Sustained
Disk drive (IBM 73LZX SCSI)	30-58	Sustained
DDS-4 tape drive (HP C5686A)	3	Sustained, no compression
LTO Ultrium tape (HP C7400A)	20-40	Sustained, no compression

Table 1. Bandwidth of selected computer and network systems

We distinguish between loss free and lossy reduction methods, or compression and sampling, respectively. Another form of data reduction is computing of higher-level parameters from the trace data.

The kinds of analysis are not always known at measurement time, which is why parameter-computation is not feasible. The packet header

[1] In this paper the abbreviation X@Y indicates a bus interface X bits wide and clocked with a frequency of Y MHz.

data and timestamps must be preserved. Time-based sampling is possible, but trace duration and schedules must be balanced, as we discuss below.

Network measurement cards

Network measurement cards provide for the physical layer attachment of the network monitor to the link being observed. The data is retrieved as packets and simple functions are performed, typically packet arrival timestamping and the discarding of the packet payload, leaving only various levels of packet header information.

Dag measurement cards currently store packet headers in fixed-size 64 bytes records, which include 8 bytes for a timestamp. For burst of small packets, this may result in an increase of the data rate relative to link speed. For bi-directional links a pair of cards must be used, which will share the same PCI bus. In order to cope with temporary shortage of bandwidth, a reasonable amount of FIFO memory must be used to not cause the loss of measurement data. On the Dag3.2 cards a total of about 8000 packet headers can be buffered, covering the minimum of 3 milliseconds worth of minimum sized packets at OC12c links.

On average, the reduction of packets to IP headers will result in a reduction of data. The ratio depends on the average packet length, which is a function of the measurement point and observation time. Below we report a typical average packet size for selected monitors within the WAND and PMA infrastructure:

Site	Average	10-percentile	90-percentile
NZIX	230	167	301
Auckland	380	310	468
NASA AIX	440[2]	433	452
Colorado COS	590	578	612

Table 2. Average packet size for selected measurement points

Considering a fixed size record of 64 bytes, the packet header capture provides for a 4:1 to 9:1 reduction of data relative to current link load.

[2] Correlates well with [AIX005].

Network monitors

Network monitors deploy one or two network interface cards to gather data from the link and, via temporary buffers in host memory, to store the trace data onto local disks. For this process to work reliably, the data rate as produced by the network measurement cards must not exceed the bandwidth of the PCI bus(ses). Further, since trace data gets buffered by memory and is being accessed by both the network measurement cards and the intelligent disk subsystems (SCSI), the memory and memory bus bandwidth implemented must be at least twice the peak data rate to be supported. Very often, the PCI bus is shared among the network cards and the SCSI subsystem, imposing even more stringent conditions. A large buffer (typically 128 Mbytes to 512 Mbytes) is used to buffer short-term bursts and allow the data to drain to disk with a more sustainable average rate.

Site	Packets (in packets/sec)		Peak measurement data rate (in Kbytes/sec)
	Day	Night	
Auckland	1000	250	64
NZIX	3500	1000	224
COS	20000	*unknown*	1300
AIX	110000	50000	7000

Table 3. Data load as produced for selected measurement points

Online data repository

Today's SCSI disk systems offer peak bandwidth exceeding 20 Mbytes/sec. This bandwidth is equivalent to the raw data rate present on a single (unidirectional) OC3c network link. However, this data rate cannot be sustained across the entire drive to track-to-track seek time. Very often, a standard filesystem implementation is utilised on the disk, which via cylinder group layout policies trades disk fragmentation and random small file access time against the raw read/write performance of the drive, restricting the data streaming performance much further. The PMA project reserves a dedicated SCSI drive and DMA channel per network interface card. As discussed before, the offered packet header load on actual PMA systems is much less and so configurations with two network capture cards on a busy OC12c links are known to work reliably (AIX).

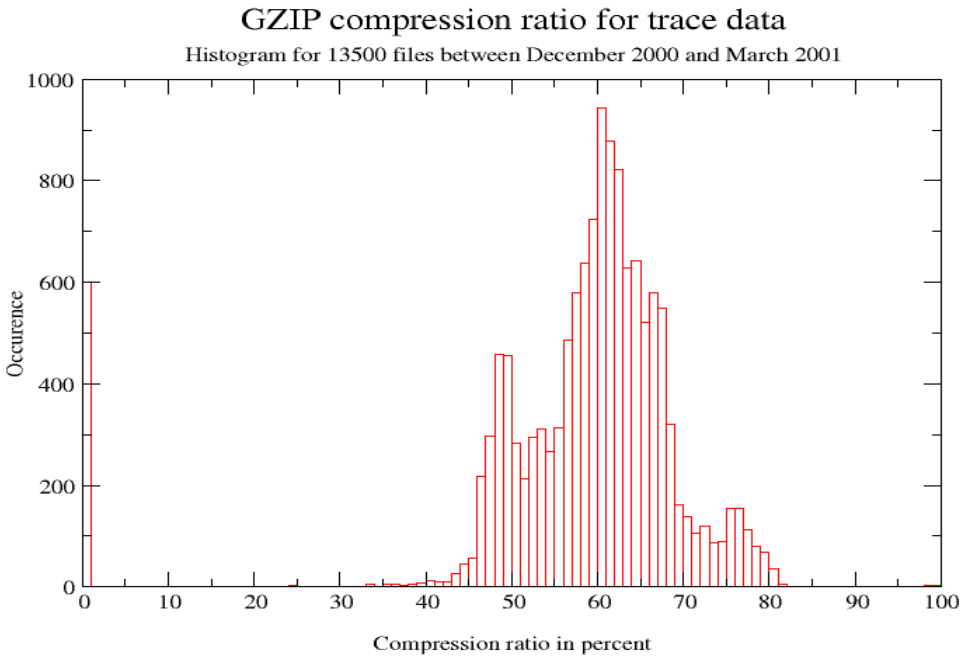


Figure 2. Histogram of gzip compression ratio for 13500 trace files

Online data processing

The current PMA monitors implement an 8 x 90 second per day stratified random sampling policy [NAI00]. The tradeoffs for sampling are discussed in [Claffy93]. This sampling method is very efficient from the point of view of data reduction: data is collected for a total of 12 minutes per day, yet long term trends and variations can be observed. Plotting the 12 minutes against 24 hours we achieve a 7200:1 reduction. However, this method has drawbacks. It is typically difficult to do flow analysis with a 90 second snapshot. Within PMA we have decided to complement the current 24 x 7 sampling with a selected number of longer-term observations.

As data is only collected for 90 seconds per every three-hour window, the remaining time is used to post process the collected traces. In order to preserve scarce Internet bandwidth and disk storage capacity, traces are compressed with a standard UNIX file compression utility `gzip`. The efficiency of compression varies; typical ratios are between 50% and 75%. The average ratio is about 58%. As compression speed is limited [Barcelona], this compression must be done to the collected traces rather than to the incoming data stream from the network interface cards.

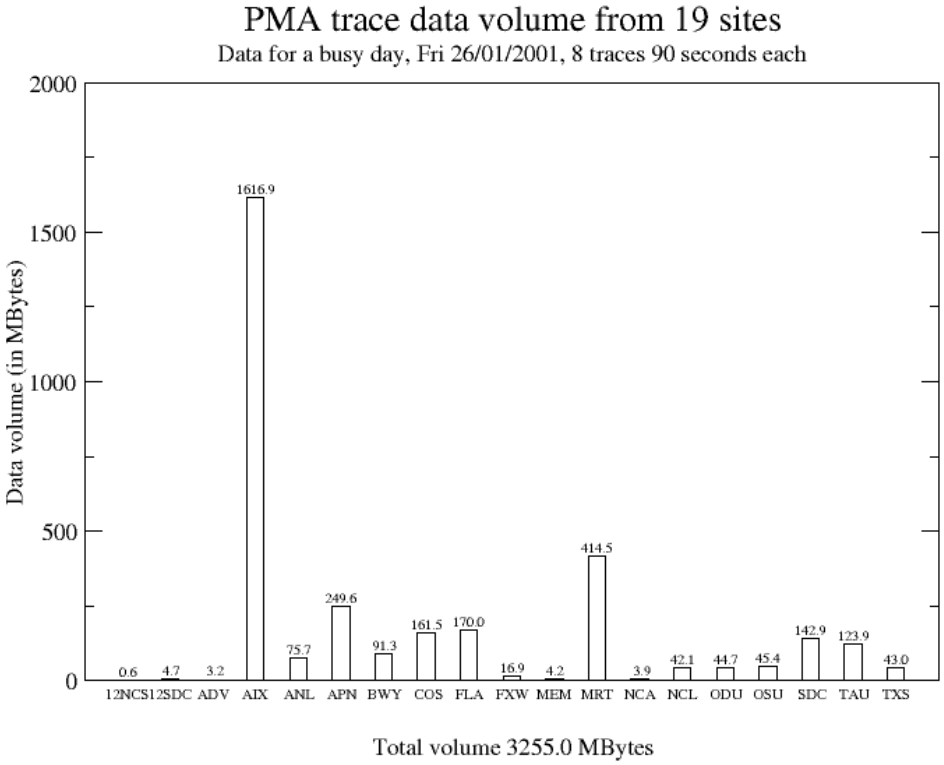
PMA also implements the computation of a selected number of parameters from the trace file, such as packet size distributions, flow sizes and length', throughput summaries and more [Datacube]. This results in a set of uncompressed ASCII log files of about 560 Kbytes per day per monitor. In comparison to the `gzip` compressed trace data file, this overhead is minimal. Depending on the trace data volume, this reduction of trace data to a few parameters has efficiency between 30:1 and 3000:1, unless the trace volume is negligible.

For WAND group monitors with low-bandwidth link load it is possible to implement a different data capture policy. As discussed in [Micheel01], the data rates for the Auckland university access links are low enough to do on-the-fly data compression with `gzip` level 1. The data volume produced per day is approximately 2GB. Deploying today's largest hard drives, uninterrupted header capture for up to 6 weeks is possible. The archiving and analysis is done on the central trace data server.

Network transfer to repository

Within the data path from the network to the data repository, the transfer via Internet is the first severe choke point. For PMA monitors attached to HPC networks (vBNS and Abilene) peak data transfer rates of a few Mbit/sec are possible, but this is comparably low to the data rates delivered from header capturing.
Figure 3. Daily trace data volume from 19 sites

As an example, we pick a recent trace taken at Colorado State university (COS): it takes 1 hour to compress 1/2 hour trace file, it takes 2-3 hours to transfer an 1/2 hour compressed trace file.



For the WAND group monitors, an Internet charging scheme per MByte of transferred data makes network transfers completely prohibitive. Here, old style postal shipment of tapes and disk drives is required.

A bottleneck for the network transfer is the access link to the central data repository. Here, the varying time for data compression and analysis of the trace data at the monitors helps stretching out the delivery of trace data over a larger interval and thus limiting the amount of incoming data at any one time.

Data repository and offline data storage

The data repository is build on regular hard disk drives, so everything discussed for the monitors with respect to bandwidth applies here. A real problem is that the repository holds trace data for the combined monitors for a longer period of time, say 3 months, including a sample of older data. With the current set of monitors and the 8x90 scheduling policy the project is expecting to exceed 1 Tbyte of data per year.

In a recent study by the WAND research group [Daley01] we have investigated possible techniques to improve the compression ratio for Internet packet header traces. The idea is to use knowledge about the structure of packet headers and traffic dynamics to significantly improve the compression ratio beyond the one currently achieved by `gzip`. We are optimistic that we can push header trace compression beyond 90% (10:1) in the future.

Discussion

Method	Reduction
IP headers only	4:1 – 9:1
GZIP compression	2:1 – 4:1
Time-based sampling 8x90	7200:1
Parametric sampling	30:1 – 3000:1

Table 4. Efficiency of different data reduction methods

Conclusion

By analysing the data path of passive Internet header traces collected from a distributed set of monitors we have identified that there is sufficient capacity for processing of trace information at the monitor itself. The choke point in terms of bandwidth is the Internet connection from the monitor to the central data repository; the choke point for data storage is the central data repository itself.

We have shown that both loss less and lossy data reduction methods can and should be applied. Sampling is the most effective way of data reduction, however, a generic data

collection infrastructure puts restrictions on sampling, so improving data compression efficiency is a viable alternative to widen the bandwidth and storage choke points.

References

[AIX005] Sean McCreary, K Claffy: Trends in Wide Area IP Traffic Patterns: A View from Ames Internet Exchange, ITC Specialist Seminar on IP Traffic Modelling, Measurement and Management, September 2000, http://www.caida.org/outreach/papers/AIX0005/

[Claffy93] K.C. Claffy, G.C. Polyzos, H.-W. Braun: Application of Sampling Methodologies to Network Traffic Characterisation, Proceedings of ACM SIGCOMM '93, May 1993. ftp://oceana.nlanr.net/papers/sigcomm.sampling.ps.gz

[Daley01] Nicholas T Daley: Compression of Dag card output, Supervisor John Cleary, WAND group technical report, February 2001. http://wand.cs.waikato.ac.nz/~joerg/nickreport.html

[Datacube] http://moat.nlanr.net/Datacube/

[Micheel01] Jörg Micheel, Ian Graham and Nevil Brownlee: The Auckland data set: an access link observed, 14th ITC specialist seminar, April 2001. http://wand.cs.waikato.ac.nz/wand/publications/barcelona-2001.pdf

[NAI00] Hans-Werner Braun: Measurement and network analysis activities, October 12th, 2000. http://moat.nlanr.net/Presentations/NAI/sld023.htm

[PMA] http://moat.nlanr.net/PMA/

[PCI95] Tom Shanley, Don Anderson: PCI System Architecture, Third Edition, Mindshare Inc, Addison Wesley Publishing Co. 1995. ISBN 0-201-40993-3