Infrastructure for Active and Passive Measurements at 10Gbps and Beyond

Best Practice Document

Produced by the UNINETT-led working group on network monitoring

Author: Arne Øslebø (UNINETT)

April 2015

Document No:	GN3plus-NA3-T2-UFS142
Version / date:	April 2015
Original language	English
Original title:	“Infrastructure for active and passive measurements at 10Gbps and beyond”
Original version / date:	Version 1.0; 29 August 2014
Contact:	campus@uninett.no

UNINETT bears responsibility for the content of this document. The work has been carried out by a UNINETT led working group on network monitoring as part of a joint-venture project within the HE sector in Norway.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 605243, relating to the project 'Multi-Gigabit European Research and Education Network and Associated Services (GN3plus)'.

Table of Contents

Executive Summary

UNINETT has been working on active and passive monitoring for many years and has been operating a large scale monitoring infrastructure with up to 30 probes since 2005. During this time we have gained a lot of experience in deploying and operating a monitoring infrastructure. In this document we want to give an overview of the advantages of having a passive and active monitoring infrastructure, what kind of software that can run on it and the recommended hardware to use.

1 Introduction

Setting up an infrastructure for active and passive measurements can be very useful for monitoring the network. It can be used for both performance and security monitoring and it can also be a very good tool for debugging network problems. A monitoring probe is usually a commodity hardware server and for passive monitoring either a specialised monitoring card or a commodity NIC is used.

The difference between active and passive monitoring is that **active monitoring** actively generates network traffic and measures the results while **passive monitoring** passively captures and monitors the existing network traffic.

2 Use cases

There are many use cases for an active and passive infrastructure and in this section we provide a short description of some of them.

2.1 Round trip measurements

Round trip measurements can measure both delay and packet loss in the network. Normal ping sending ICMP messages can be used to measure in a full mesh between all the monitoring probes. In a large scale infrastructure, a random delay should be used before starting the measurements, to avoid all the monitoring probes sending packets at the exact same time.

Long term storage of the measurements is recommended, as this makes it possible to see how the delay in the network changes over time.

2.2 One-way delay measurements

An alternative to round trip measurements is to carry out one-way delay measurements. If the frequency of sent packets is high enough, then it can be used for measuring micro outages in the network due to buffer congestions in the routers. It can also be used for measuring the routing convergence time. When a link goes down in the network, the measurement probes detect packet loss and by measuring how long it takes before packets are received again, it is possible to measure the time it takes for the network to reconfigure the routing.

2.3 Multicast measurements

Monitoring multicast can be done by tools like multicast beacon[1] that sends packets to a specific multicast group. All the measurement probes will join this group and report statistics about the received packets.

If there are official multicast services in the network, like multicast TV, then another measurement method can be for the monitoring probes to join the official groups and measure the data that it receives. This can provide a detailed statistics about join time, packet rate, packet loss, etc.

[1] http://sourceforge.net/projects/multicastbeacon/

2.4 Throughput measurements

A common complaint by end users is bad throughput performance on the network. Very often this is related to a problem on the end systems or local networks. By performing regular throughput measurements between the measurement probes, it is possible to measure and document that the network between the measurement probes is problem-free.

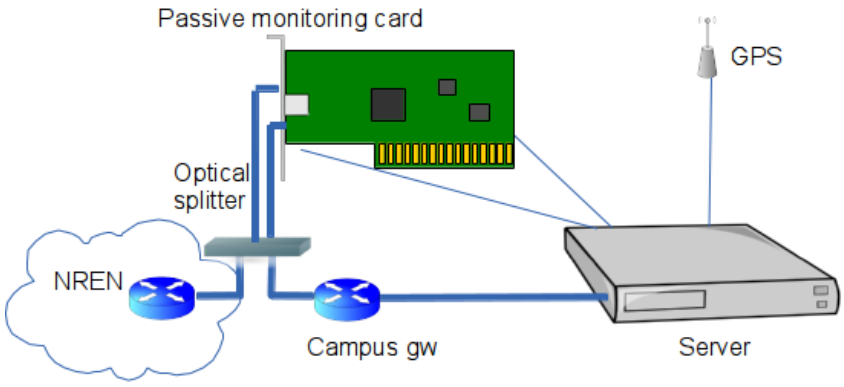
2.5 Categorising network traffic

To get a better overview of what the network is being used for and where the traffic goes, it can be very useful to categorise the network traffic based on the application that generates it. This can be done by doing passive monitoring of the network. To see where the traffic goes, it is possible to use geo-location to map the IP addresses to both AS numbers and country code.

2.6 IDS system

For security monitoring it can be useful to use the monitoring probes for running an Intrusion Detection System (IDS). Running a full system can put a very high load on the monitoring probe, disturbing other measurement tasks. By careful tuning of the IDS system it is however possible to scale a system to handle 10Gbps load. One option can be to just monitor outgoing traffic. This can be used for detecting infected hosts on the campus network.

3 Equipment and installation



Monitoring probes will typically be installed so that the passive monitoring card listens to traffic between the NREN and the campus router. It is also possible to install monitoring probes in the backbone network. This can help to make it easier to locate network problems, but the increasing speed of backbone network links makes this a lot more expensive compared to installing probes at the edges.

The figure above shows the general setup for a monitoring probe. You have a commodity server with a dedicated passive monitoring card in addition to the built in NIC that is used for active measurements. A GPS antenna can be used for accurate time synchronisation and an optical splitter is used for connecting the passive monitoring card.

Instead of using an optical splitter, it is also possible to use port mirroring in the router to capture network traffic. The disadvantage of this is loss of proper timing information of the original packets. If the probes are only used for security monitoring, this is not a problem, but if it is used for QoS monitoring then an optical splitter is necessary.

3.1 Server

High speed throughput tests and passive monitoring of high speed networks at 10Gbps and beyond requires a lot of CPU power. The only way to scale the passive monitoring is to take full advantage of the multi-core technology of modern CPUs. So a high-end server with a fast multi-core CPU should be used. A minimum of eight cores should be available, as well as support for hyperthreading. The CPU should also have as high a frequency (i.e. clock speed) as possible.

A second CPU can also be a good idea if multiple passive monitoring applications will be active at the same time. Even if it is believed that a second CPU is not needed, it is recommended to buy a server that at least has a second CPU socket that can be used if needed.

The amount of memory depends a lot on what kind of applications are being used. If the monitoring probe is used for IDS or other passive monitoring tools that keeps a lot of state, then 32GB is considered the absolute minimum. For full IDS at 10Gbps then more than 32GB of memory is needed.

The amount of disk space needed on a monitoring probe depends heavily on the use cases. For 10Gbps passive monitoring we recommend doing most of the monitoring in real-time, so that only high level reports or alerts are stored on disk. This does not require fast disks or a lot of storage (by modern standards), although using tcpdump or a similar tool to store full packet dumps can put a strain on both disk performance and storage capacity. If a requirement is to be able to store full packet capture to disk an array of high speed SSD disks are needed. In most practical situations, tcpdump will be used with a filter so that only a small subset of the total traffic if stored to disk. In this case most normal disks will suffice when it comes to both performance and capacity.

It is recommended to use Linux as an operating system as most monitoring tools have been developed and tested on Linux. It is also possible to use one of the BSD types of operating systems, but it should then be verified that a proper driver for the selected passive monitoring card is available.

3.2 NIC for active measurements

When throughput tests are only done on 1Gbps speeds the built-in Network Interface Controller (NIC) of the server will most likely be sufficient. If tests are done on higher speeds, then a dedicated NIC PCI card is recommended. At the time of writing cards based on the Intel 82599 controller, like the Intel X520 series of cards, are good choices.

3.3 Monitoring card

For passive monitoring cards it is possible to use either a commodity NIC or specialised cards designed for passive monitoring. A third option is also cards or servers based on the Tilera CPU.

3.3.1 Commodity NIC

For doing passive monitoring on 1Gbps or 10Gbps, it is a good option to use a good quality commodity NIC. If you require 100% packet capture even when the link is fully saturated with minimum size packets, then specialised hardware is necessary. For realistic measurements with average packet sizes, even at full 10Gbps, then a commodity NIC is good enough.

At the time of writing, cards based on the Intel 82599 controller like the Intel X520 series of cards, are a good choice. The advantage of this card is that it has built in support for hash based load balancing

for up to 16 cores per interface. This means that all packets that have the same source and destination IP address and port number, are sent to the same buffer. Packets are relatively evenly distributed among the available cores.

While the newest version of the standard driver for the Intel 82599 controller in Linux provides relatively good performance, there are modified drivers available that increase performance when using the NIC for passive monitoring. The most stable version is PF_RING[2] with DNA-enabled driver. This is a commercial product, but free to use by non-profit academic institutions. Looking at the recommended cards for PF_RING is a good method for finding the best commodity NIC for passive monitoring.

With PF_RING it is also possible to run multiple applications at the same time where all get access to the captured packets using zero copy mechanisms to keep performance up.

While the performance of Intel X520 type of cards is good, it is not able to capture 100% of the traffic with minimum packet sizes. The table below shows the results of some performance measurements carried out on an Intel X520-SR2 card, capturing packets of minimum packet size. This is on a dual port NIC capturing packets on both ports at the same time:

Gbps	Mpps	CPU Load(%)	Packet drop (%)
0.7	1	1	0
3.3	5	4	0
6.7	10	7	0
10.2	15	13	0
13.9	20	18	0
16.8	25	23	0
20	29.8	31	3.2

As we can see from this table, at 29.8Mpps we get a 3.2% packet drop and 31% of the CPU is used just to capture packets. If, however, we use a more realistic distribution of packet sizes, then the Intel X520 has no problem capturing all packets without packet drop and without using too much CPU:

Gbps	Mpps	CPU Load(%)	Packet drop (%)
17.3	**5**	**7**	**0**
20	**6.5**	**9**	**0**

2 http://www.ntop.org/products/pf_ring/

3.3.2 Specialised monitoring card

One limitation of using a commodity NIC for capturing packets is that it is not possible to connect a GPS antenna to the NIC for accurate timestamps. For general purpose measurements used for monitoring the network, a GPS antenna is usually not needed. If the measurement infrastructure will be used for research work then accurate timestamps might be needed and a specialised monitoring card must be used so that a GPS signal can be connected.

A specialised monitoring card is currently also needed if you are going to be monitoring at speeds of more than 10Gbps or if it is critical to capture each and every packet at full theoretical packet rate.

A specialised monitoring card is usually developed using a field-programmable gate array (FPGA) so that the capabilities of the card can be changed. The advantage of cards like this is that they are capable of doing a lot of processing on the card without involving the main CPU of the server. Some of them can even generate full IPFIX (IP Flow Information Export) flows on the card.

The main disadvantage of specialised cards is the cost. They are a lot more expensive than off-the-shelf commodity NICs.

Some vendors of specialised monitoring cards are INVEA-TECH[3], Emulex[4] and Napatech[5].

3.3.3 Tilera

One alternative to FPGA-based cards are cards or servers based on the Tilera CPU. The Tilera CPU is a specialised CPU that uses a large number of cores to scale the processing power of the CPU. It also has instructions that are designed especially for processing of network packets and it has specialised high speed I/O that allows for passive monitoring at speeds up to 100Gbps.

The advantage of Tilera is that it runs normal Linux so that it is possible to do your own development on it. This is usually not possible on FPGA based solutions. The challenge is to parallelise algorithms so that they can take advantage of the large number of cores.

The Tilera cards and servers are comparable to specialised monitoring cards in price.

[3] https://www.invea.com/
[4] Sells EndaceDAG cards, http://www.emulex.com/
[5] http://www.napatech.com/

3.4 GPS antenna

A GPS antenna is only needed if completely accurate timestamps on captured packets are needed. For most practical measurements, a good NTP source is good enough. If a GPS antenna is deemed necessary, the challenge of installing antennas should not be taken lightly. Usually server rooms where the monitoring probes are installed are located in the basement while the GPS antenna has to be installed on the roof. This often results in long cable runs.

The antennas should also have a surge protector so that lightning cannot follow the GPS cable from the roof and down into the server room.

Specialised monitoring cards that support GPS take a pulse per second (PPS) signal from the GPS antenna to keep synchronised. This can provide timestamps of captured packets with an accuracy in the nanosecond range.

3.5 Optical splitter

An optical splitter is in principle easy to install. Just insert it between two routers so that part of the light signal can be tapped and sent to the passive monitoring card. For single mode optics, an 80:20 splitter is usually a good choice.

Multimode optics usually have weaker signal strength, so a 60:40 splitter is recommended.

The table below shows the approximate dB loss for different types of splitters.

Split ratio	Live port (dB loss)	Monitoring port (dB loss)
50:50	4.5	4.5
60:40	3.1	5.1
70:30	2.4	8.3
80:20	1.8	8.1
90:10	1.0	11.5

3.6 Equipment summary

This is a summary of recommended equipment for a passive and active monitoring probe at the time of writing.

CPU	Fast (>2.5GHz), minimum 8 cores with hyperthreading
Memory	Minimum 32GB
Disk	Usually no special requirements. For full packet capture to disk a big array of high speed SSD disks is needed.
NIC	For 1Gbps, use built in NIC. For 10Gbps use a NIC based on the Intel 82599 controller
Passive monitoring card	A NIC based on Intel 82599 controller or look at recommended cards for PF_RING DNA.
Optical splitter	80:20 for single mode, 60:40 for multimode
GPS	Usually not needed

4 Software

In this section we will give some recommendations on different open source software that can be used on the monitoring infrastructure. There are a lot of research projects developing various types of monitoring software and new tools often appear. The problem with many of these tools is that as soon as the project is finished, the tools become “abandonware”. The tools are also often not of a production quality as they were just implemented to test a particular research thesis.

In this section we will therefore list software that are considered relatively stable, but when a new monitoring framework is put into production it is recommended to do search to see if new tools are available. CAIDA also maintains a list of possible software tools for doing various types of network measurements[6].

4.1 Flow monitoring

There are many applications and tools available for doing flow monitoring, but the two most commonly used are YAF[7] and Nprobe[8]. Both of them are stable and well supported by the developers. They provide support for both NetFlow and IPFIX and support deep packet inspection to categorise network flows based on the application generating the traffic.

They are both released under the GPL license although the developers of Nprobe require payment before you can download the source code. It is free for researchers and academic institutions.

Nprobe also has several plugins that allows for more detailed analysis for certain protocols like DNS, BGP, HTTP, SIP and RTP.

UNINETT has used Nprobe to develop Appflow[9]. Appflow categorises network traffic based on the application that generates it. It also shows which AS number and country the traffic from each application comes from and goes to. The application supports real time aggregation of flow records which makes it highly scalable, both when it comes to network speed and the number of monitoring probes.

[6] http://www.caida.org/tools/
[7] http://tools.netsa.cert.org/yaf/index.html
[8] http://www.ntop.org/products/nprobe/
[9] https://tnc2014.terena.org/getfile/1738

4.2 IDS system

The two main open source Intrusion Detection Systems (IDS) are Snort[10] and Suricata[11]. Both of them have a large user base and are capable of scaling to high speeds. Careful tuning of the systems and the number of rules is important if the goal is to monitor multi-gigabit of traffic.

4.3 Throughput testing

The *de facto* standard for measuring maximum TCP and UDP throughput in a network is the tool Iperf[12]. Iperf allows for tuning various parameters and UDP characteristics and then provides detail about achieved bandwidth, delay jitter and packet loss. It is based on a client-server paradigm where an Iperf client connects to an Iperf server and then performs throughput tests.

To carry out scheduled Iperf tests there is a tool called BWCTL[13] that acts as a wrapper around Iperf.

4.4 Multicast measurements

Dbeacon is a multicast beacon where all monitoring probes that are running it monitor the beacon's reachability and collect various statistics such as packet loss, delay and jitter. A web page presents the results in a grid showing all the results between the different beacons. It supports both IPv4 and IPv6.

4.5 One-way delay

The OWAMP[14] (One-way ping) tool is an implementation of the OWAMP protocol[15] and measures one-way packet delay and loss across Internet paths. The advantage of one-way delay measurements compared to the more traditional round trip measurements is that it makes it easier to isolate the direction in which congestion is experienced.

[10] https://www.snort.org/
[11] http://suricata-ids.org/
[12] https://github.com/esnet/iperf
[13] http://software.internet2.edu/bwctl/
[14] http://software.internet2.edu/owamp/
[15] RFC4656, http://tools.ietf.org/html/rfc4656

5 Managing and monitoring the monitoring probes

To make the deployment of a large scale monitoring infrastructure as easy as possible, all monitoring probes should use the same hardware and run the same software. Installation should be made as automatic as possible with only a few local changes for each probe. To achieve this, it is strongly recommended to use a configuration management and orchestration tool. There are many tools available like Puppet[16], CFEngine[17] or Salt[18]. All of these tools are open source, but enterprise versions with support are also available.

Since both throughput testing and high speed passive monitoring put a very high load on the servers, it is important do basic system monitoring of all the servers. Any standard system monitoring tool can be used. Some candidates are Zabbix[19], Nagios[20] or Icinga[21].

[16] http://puppetlabs.com/
[17] http://cfengine.com/
[18] http://www.saltstack.com/
[19] http://www.zabbix.com/
[20] http://www.nagios.org/
[21] https://www.icinga.org/

Glossary

AS	Autonomous System. Within the Internet, an AS is a collection of connected IP-routing prefixes under the control of one or more network operators with a clearly defined routing policy to the Internet.
BGP	Border Gateway Protocol
DNS	Domain Name Server
FPGA	Field-Programmable Gate Array
GPL	General Public Licence
ICMP	Internet Control Message Protocol
IDS	Intrusion Detection System
NIC	Network Interface Card
NREN	National Research and Education Network
OWAMP	One-Way Ping
QoS	Quality of Service
RTP	Real-Time Transport Protocol
RTP	Real-Time Transport Protocol
SIP	Session Initiation Protocol

Complete BPDs are available at http://services.geant.net/cbp/Pages/Home.aspx

campus-bp-announcements@terena.org