#41

TEN-155 Multicast:
MBGP and MSDP monitoring

Jan Novak

Saverio Pangoli

DANTE IN PRINT is a track record of papers and articles published by, or on behalf of DANTE. An HTML version is available at: http://www.dante.net/pubs/dip/

For more information about DANTE or *DANTE IN PRINT* please contact:

TEN-155 Multicast
MBGP and MSDP monitoring

Jan Novak
Saverio Pangoli

Abstract

The paper concentrates on the monitoring of a production, native multicast network in the inter domain environment, where every domain is one European country. The multicast traffic monitoring is not included as it involves only simple polling of SNMP (Simple Network Management Protocol) variables in the currently standardised IPmroute MIB (Management Information Base). The authors focus entirely on MBGP (Multiprotocol Border Gateway Protocol) and MSDP (Multicast Source Discovery Protocol) monitoring, as there are no applications currently available to perform routine data collection about the operation of these protocols in today's network.

The work described in this paper covers the implementation of a simple monitoring system providing informational data about the worldwide multicast infrastructure and some debugging data to detect certain kinds of network misbehaviour affecting overall performance. The main motivation of this work has been to provide at least simplest operational data about the performance of the MSDP and MBGP protocols and make them available publicly for the use of the Network Operation Centres of the TEN-155 connected networks. Where detailed information could be required, pointers to the DANTE web site are provided.

KEYWORDS: TEN-155, Multicast, MSDP, MBGP, PIM-SM, DVMRP, ATM, Full Mesh.

Jan Novak is a former Network Engineer from DANTE and is now working for Cisco. His email address is <janovak@cisco.com>.
Saverio Pangoli is also a Network Engineer at DANTE; his email address is <saverio.pangoli@dante.org.uk>.

1. Introduction

DANTE introduced during 1999 a native multicast service on the production TEN-155 network as described in detail in [1]. The current topology scheme is presented in Figure 1 below:
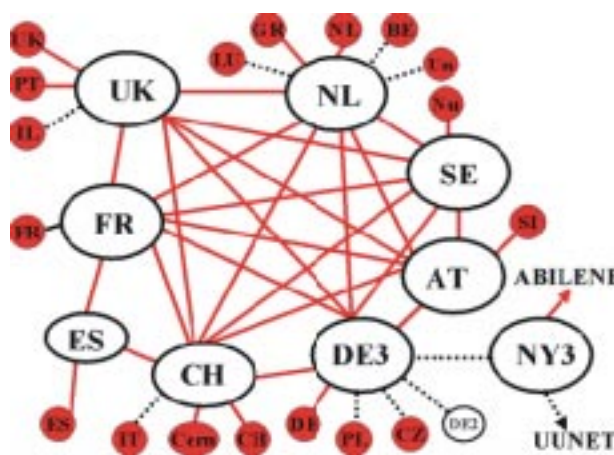

Figure 1. Topology of the current TEN-155 multicast network

The full lines show, where unicast and multicast run together on same ATM/physical infrastructure. Dashed lines are dedicated ATM PVCs or few remaining tunnelled connections.

The whole network is based entirely on PIM-SM (Protocol Independent Multicast - Sparse Mode [2]), MBGP (Multiprotocol Extensions for BGP-4 [3]) and MSDP (Multicast Source Discovery Protocol [4]).

The multicast network has grown further since the status description in [1]. There have been several occurrences of MSDP messages "storms" during the introduction of the multicast service, originating at several places of the world wide MSDP/MBGP infrastructure. In the case of such a storm, CPU usage of the routers can rise by up to 50% and the

MSDP control data rates have been observed at up to 1 Mbit/s of data. The motivation of the work described in this publication is the detection of MSDP/MBGP problems at the early stages. The detailed information about the main detected reasons for MSDP SA messages storms is provided in [5].

The effect of a MSDP SA storm (originated at the particular router) on the router CPU is demonstrated in Figure 2 below.

The monitoring and this publication are divided into two main parts - MSDP and MBGP monitoring. The web access to the monitoring results is available at http:// www.dante.net/mbone/msdp and http:// www.dante.net/mbone/mbgp.

2. MSDP monitoring

2.1. Motivation

As already said, the main motivation for this part of monitoring work is the detection of MSDP messages storms. A short analysis of the data streams and SA messages duplicates generated by MSDP is also provided - this might not be of concern now, but when the multicast usage grows in the general Internet, the MSDP control traffic should be analysed and optimised to lowest possible minimum. The storms are indicated by frequent resets of MSDP TCP sessions but there is no easy operational way to detect the originator of the problem. As a consequence, several tools have been developed, and are publicly available at these sites:

ftp://sith.maoz.com/pub/shep/ MSDPMon.tar
http://www.ncne.nlanr.net/tools/ multicast.html

2.2. The monitoring system

The tools described above are simple implementations of the MSDP protocol, which enable to close a MSDP session from a workstation and dump full MSDP data to a file. This file contains all information necessary to detect MSDP problems and identify the originator. Currently, the monitoring system closes a MSDP session on the backbone router for two minutes in duration every fifteen minutes. DANTE wrote a few PERL scripts using the MSDP file as an input to aggregate and visualise the results. The outputs provided can be divided into two groups:

1) informational/historical data
2) real time/debug data

2.3. MSDP informational data

All the outputs are provided in the MRTG (Multi Router Traffic Grapher - [6]) format and are available at the main page http:// www.dante.net/mbone/msdp.
Information provided:

1) The total rate of MSDP control messages per minute. Typical rates during May 2000 are shown on Figure 3 on the following page. This can be used for a rough estimation of the MSDP control traffic rate. Considering
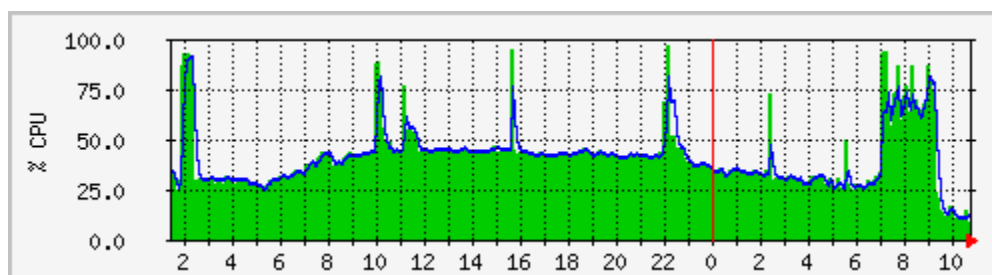

Figure 2. Effect of a MSDP SA storm on a Cisco 7507 router CPU (between 7.00 and 9.30)

the minimum of 96 bits per Source Active (SA) message (a more accurate estimation would have to take into account the number of sources announced by one RP, the number of messages containing data, IP/TCP headers etc.), we get a minimum rate of 1.92 kbit/s of MSDP data in the current worldwide MSDP infrastructure.

2) The number of MSDP SA duplicates and the number of MSDP messages containing data: a duplicate is defined as a message announcing the same source (S) and multicast group (G) but originated by two PIM-SM Rendez-vouz Points (RP).

The count of duplicates is obtained easily as a difference between the number of unique triples (S,G,RP) and the number of unique pairs (S,G) contained in the MSDP data output file.

This measurement provides interesting results - it appears that there are always some duplicates generated, sometimes up to the 30% of the total MSDP SA rate. A closer look at the originators shows, that almost any RP in the current infrastructure originates from time to time some duplicates.

The reason for that is not quite obvious; the duplicates are generated even by RPs with PIM-SM neighbours only.

DANTE tried to attract the attention of Cisco developers to this fact but without success yet. Typical duplicates/data packets rate during May 2000 is shown on Figure 4 below.

3) The number of multicast groups announced and the number of (S,G) pairs (ie. the sum of all sources sending to all groups) in the whole MSDP system: this provides just an idea about the amount of these entries.

It can be used to quantify another recent problem in PIM-SM networks - for different reasons, forwarding states are generated improperly by joins from different networks (functionality failures inside of some PIM domains) as described in [7].

The data gathered from MSDP show active sources in the multicast system. The output from the Cisco multicast routing table shows all forwarding states generated by PIM joins from all the networks. The comparison of these two outputs taken simultaneously indicates the extend of the problem:
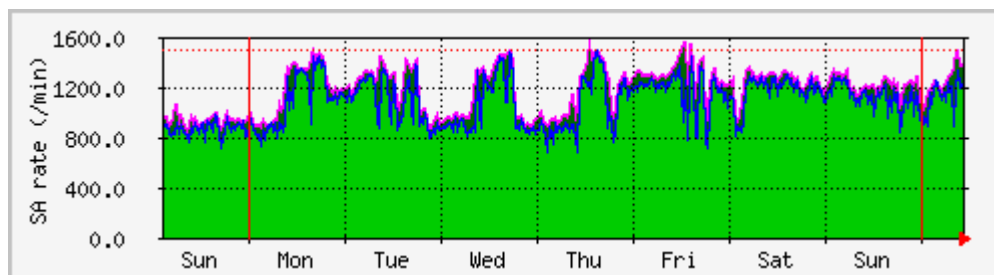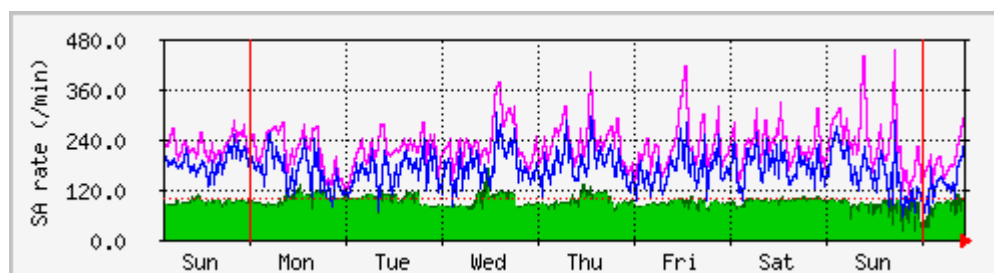

Figure 3. MSDP SA rate per minute

Figure 4. MSDP SA duplicates rate (line above the dark area) and MSDP messages containing data (dark area at the bottom of the graph) per minute

Current number of (S,G) pairs from MSDP data: 1072
Current number of (S,G) pairs reported by a Cisco router : 1683 (DE3 router in Germany, TEN-155 network)

A typical output in May 2000 is shown on Figure 5 below.

4) Number of RPs (originating SA messages) in the current MSDP system: it is collected just for curiosity and historical usage development of MSDP (cf. Figure 6).

TEN-155 has world-wide multicast connectivity via a tunnel from UUNET in the US running MSDP/MBGP and native multicast peering with Abilene e.g. the data collected reflect the overall extend of MSDP usage.

These outputs have been called informational but in fact they provide a first indication of problems - SA rate can be used to indicate MSDP SAs storms, all outputs can be used for a first indication of connectivity failures to the US or in EU.

2.4. MSDP Debugging Data

Debugging almost in real-time (with a delay of 15 minutes) can be done using www.dante.net/msdp/msdp.phtml. These pages provide the following information:

- Rendez-vous Points which generate most of the MSDP traffic - the first ten are displayed. For RPs with more than one MSDP SA message per second a more detailed output is provided listing all (S,G) entries and their rate as generated by the particular RP (outputs are hidden as an URL below the name or the IP address of the particular RP).

- Duplicated (S,G) entries - at the last line of the output the full list of "originals" and duplicated SA messages is provided (it is difficult to distinguish them, if at least one RP network number is not the same as S network number) - the output is hidden as an URL below the number of duplicates in the last row.

This data is archived in text format, only the last hour results are displayed on the web. These pages provide enough information to detect any misbehaviour of some RPs. On the other hand, it does not provide the information to track down if some RP is announcing a SA message correctly. This will be considered as further extension of the system.
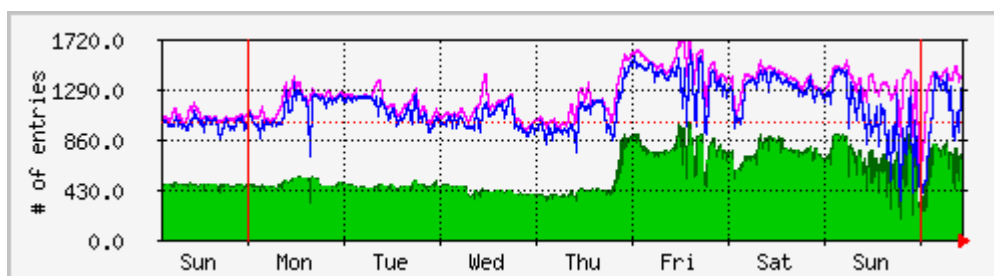A typical output of this page is illustrated in Table 1 shown on the next page.

Figure 5. Number of unique (S,G) entries (line above the dark area) and number of multicast groups in the current world wide multicast infrastructure (dark area of the graph)
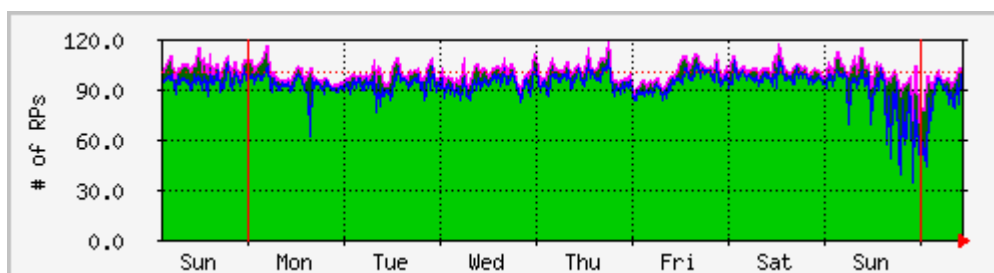

Figure 6. Number of PIM-SM RPs generating MSDP SA messages - May 2000

3. MBGP Monitoring

3.1 Motivation

The initial impulse for this part of the work was in fact the attempt to correlate the rate of MSDP SA duplicates mentioned in paragraph 2.3 to MBGP route flapping. But the problem does not seem to be so simple. In any case, the motivation is to monitor the extent of route flapping in the world wide multicast infrastructure and provide data to identify the sources of problems.

3.2 The Monitoring System

To monitor unicast routing stability DANTE has been using MRTD (Multithreaded Routing Tool Kit, http://www.mrtd.net/) developed by the University of Michigan and MERIT. MRTD does not provide a direct MBGP implementation yet; however the format of MBGP messages differs from unicast BGP4 only in the SAFI (Subsequent Address Family Identifier [3]) field.

To use MRTD for multicast routing, we just modified the definition of the SAFI fields in the source code and recompiled; the result is a multicast-only routing daemon. We installed this modified MRTD daemon on a Sun workstation in our Frankfurt TEN-155 PoP, and set up a multicast peering session with one of the TEN-155 routers, a Cisco 7k.

MRTD was then configured to dump the whole routing table and BGP updates/withdrawals every few minutes. We then wrote several simple PERL scripts to analyse the files and provide useful information. The results of the processing are then transferred via SSH to the DANTE web site. The outputs have similar structure to MSDP - informational and debugging.

3.3 MBGP Informational Data

All the outputs are plotted by MRTG, and are available at the main page, http://www.dante.net/mbone/mbgp/, which provides the following information:

1) The number of multicast routes - this has always been the first and most important

SA rate/min	SA Occurrences	from	RP
	total in 2 mins		
330.5	661	138.18.100.1	fix-west.dren.net
114.5	229	130.1.200.242	
107.5	215	206.190.40.61	lo3.bcstdatxmr01.broadcast.com
73.5	147	144.232.187.198	rp.sprintlink.net
61.5	123	130.240.22.190	rp.net.luth.se
38.0	76	141.142.12.1	charlie.ncsa.uiuc.edu
36.5	73	204.69.199.17	sjck-rp1.cisco.com
34.0	68	193.2.0.76	rarnes2.arnes.si
31.0	62	193.48.1.5	
31.0	62	146.97.248.4	
	=====		
858.0 /min :Sum:	1716		

Total SA messages: 2580 Total SA rate: 1290.0 /min
Total RPs in the MSDP system: 95
Total SAs with data packets: 149
Total number of multicast groups in the MSDP system: 646
Total number of (S,G) entries in the MSDP system: 1244
Total number of duplicate entries in the MSDP system: 434

Table 1. Output for Mon. May 22 12:02:40 BST 2000

parameter in multicast enabled networks; to keep track of the MBGP usage development we keep an historical record of the number of routes.

2) MBGP routes originating Autonomous Systems - similarly to the number of RPs in MSDP/PIM-SM, we record the number of BGP ASs which appear in the routing table as route originators. We compare this number to the number of ASs which have appeared in the last 12 hours as originators of at least one route flap to have some rough idea about the overall routing stability (cf. Figure 7 below).

3) BGP updates and withdrawals - as a more accurate measure of the general routing stability DANTE records number of BGP updates and withdrawals generated by the whole network in 15 minutes intervals, as shown in Figures 8 below and 9 and 10 on the following page.

The first graph shows the overall one-week picture, where several major BGP resets occurred with full routing table update as result (about 8000 routes in May 2000). The last two graphs show the typical rates of up-

dates and withdrawals in otherwise stable network.

Most of these are occasional flaps of some MBGP networks, but we observe also regular participants in flapping, who are monitored separately and will be notified once sufficient history is collected.

3.4 MBGP Debugging Data

Similarly to MSDP, almost real time data is provided at http://www.dante.net/mbone/mbgp.phtml. The output provides a list of updated and withdrawn prefixes together with the AS path as seen by the motoring workstation.

A list of the first one hundred prefixes is provided if there was more than one route flap in the last 15 minutes, otherwise only 25 prefixes are listed.

This page already enabled to detect several cases of pathological behaviour with updates every one-two minutes. A typical output of this page is shown in Table 2 on the next page.
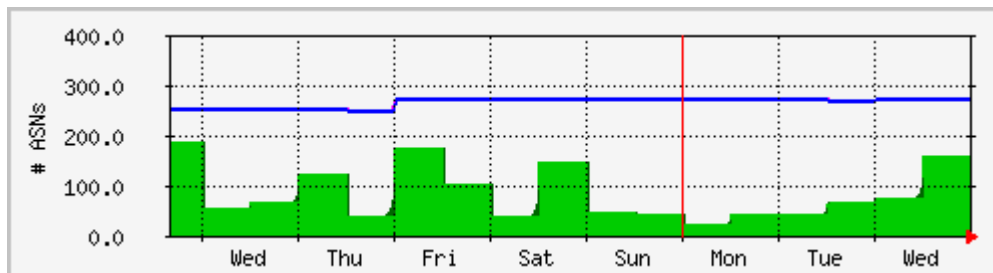

Figure 7. MBGP routes originating ASs - the line above the graph shows the number of ASs originating MBGP routes, the dark area shows the number of ASs originating route flaps during last 12 hours.
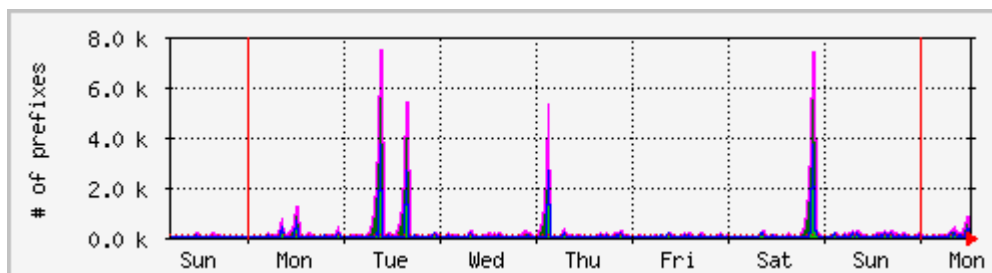

Figure 8. MBGP updates counted over 15 minutes - one day snapshot
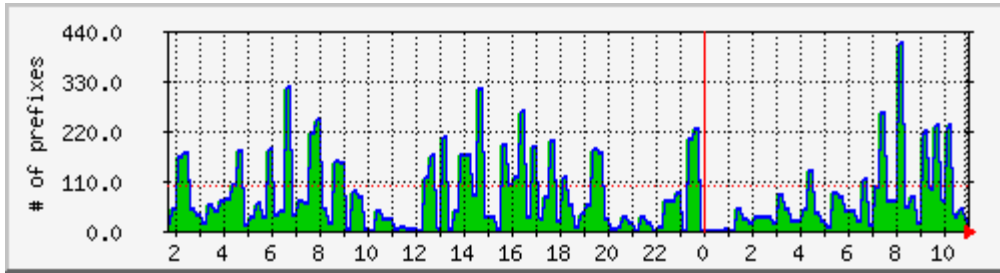
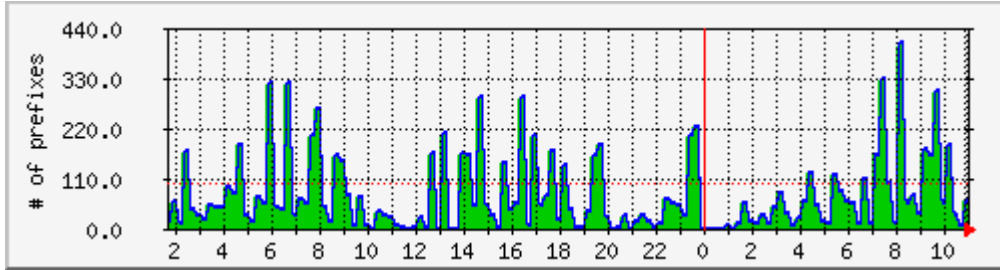Figure 9. MBGP updates counted over 15 minutes - one day snapshot

Figure 10. MBGP withdrawals counted over 15 minutes - one day snapshot

These outputs are not archived, only the last hour's data is available. To provide a history, which is necessary to detect some long-term problems in ASs, a number of other outputs is provided with one month of history:

1) The count of updates originated by the ASs and the originating ASs in the last hour

2) The count of updates originated by the ASs and the originating ASs in last 12 hours

3) Once per 12 hours, these historical files are scanned and the occurrences of ASs in the month's history are counted. If one certain AS appears in at least 60% of all the files (e.g. more than 20 days a month), then this

```
Mon May 22 12:30:02 BST 2000 : Flapping prefixes in last 15 minutes

Current limit on the number of printed lines: 100 Total # of updates: 129 Total # of
withdrawals: 105

BGP updates in last 15 minutes:
===============================
No. of U      prefix              origin        AS path
   3      128.63.14.0/24    INCOMPLETE    6680 8933 9010 704 10888 668
   2      208.171.16.0/20      IGP        6680 8933 9010 704 10888 267 8011
   2      216.93.0.0/17        IGP        6680 8933 9010 704 10888 267 8011
   2      208.246.108.0/22     IGP        6680 8933 9010 704 10888 267 8011
   2      208.137.16.0/21      IGP        6680 8933 9010 704 10888 267 8011


Withdrawn prefixes in last 15 minutes:
======================================
No. of W      prefix
   3      128.63.14.0/24
   2      209.69.224.0/20
   2      130.239.0.0/16
   2      209.176.200.0/21
   2      206.31.58.0/23
```

AS is listed at http://www.dante.net/ mbone/mbgp/month.phtml page and MRTG monitoring of that particular AS is automatically configured and started.

There are several ASs with this behaviour which are monitored.

Recently the comparison of the total number of originated routes (the line at the top of the graph on Figure 10) and flapping routes (dark areas on Figure 11 below) was added - the case where 5000 routes are originated and 50 of them flap (provided they are not the same prefixes all the time) still can indicate healthy BGP behaviour.

This AS originates a stable amount of prefixes (38) and some of them flap from time to time; further investigation is necessary, if the flapping prefixes are always the same. On the other hand, whenever the AS on the Figure 11 starts to originate some prefixes, all of them immediately flap several times and then most probably the next BGP neighbour performs BGP damping, which causes these prefixes to disappear from the routing system for about 1 hour. The behaviour of such an AS, illustrated in Figure 12 below, is clearly pathological and should be corrected.

4. Conclusions

This initial work was born as a bunch of unstructured PERL scripts, but represents a first attempt to find out what is useful and interesting to monitor. Based on this experience new and more flexible tools can be developed and put into routine operation for the use of IP multicast NOCs; in the meanwhile, these web pages are meant as a first aid to the whole TEN-155 multicast community to detect ongoing problems in the backbone. The results collected so far also indicate what can be expected in terms of CPU usage increase when the multicast infrastructure will grow world wide as unicast did.
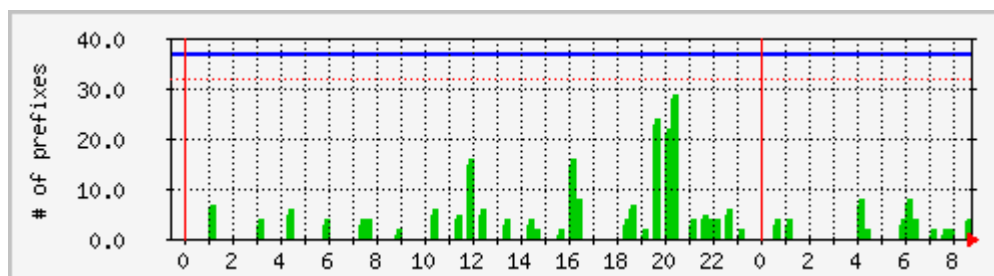

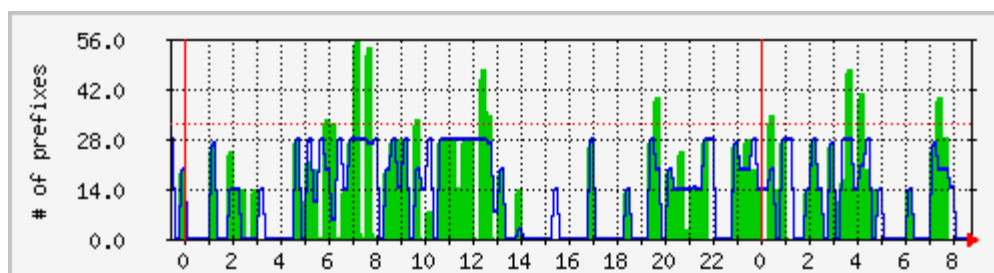Figure 11. One-day snapshot of a "reasonably" flapping AS

Figure 12. One-day snapshot of a "pathological" AS

References:

[1] TEN-155 Multicast - Jan Novak (DANTE), Peter Heiligers (DFN), http://www.dante.net/pubs/dip/40/40.html

[2] Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification - RFC2362

[3] Multiprotocol Extensions for BGP-4 - RFC2283

[4] Multicast Source Discovery Protocol (MSDP) - draft-ietf-msdp-spec-05.txt

[5] MSDP Monitoring page - http://www.dante.net/mbone/msdp

[6] Multi Router Traffic Grapher - http://ee-staff.ethz.ch/~oetiker/webtools/mrtg/mrtg.html

[7] http://www.dante.net/mbone/nop/sm_non.html