The 512K route thing
12 August 2014

World Elephant Day

Newborn Panda Triplets in China

Rosetta closes in on comet 67P/Churyumov-Gerasimenko
The internet apparently has a bad hair day.
What happened?

Did we all sneeze at once and cause the routing system to fail?
Well someone sneezed!
12 August 2014

Verizon Route Leak (AS701)
But route leaks happen all the time

But not from AS701!

– AS701 is a tier 1 ISP
– So very few (noone?) filters what they hear from AS701
– Which means that when AS701 leaks all non-default AS’s (and a few more besides) are likely to hear the route leak

So everybody saw a bunch of routes for a small amount of time...
Minute by Minute ANNces & WDLs

BGP Update Profile for 12 August 2014

Prefix Changes Per Minute (Log Scale)
Second by Second

BGP Update Profile for 12 August 2014 (07:47-08:03 UTC)
07:49.15 Table exceeds 512K
07:49:45 First set of withdrawals
07:49:58 Second set of announcements
07:50:20 Third set of announcements
07:52:04 Fourth set of announcements
07:52:46 Second wave of Withdrawals start
07:58:20: All done
512K is a default constant in some of the older Cisco and Brocade products

Brocade NetIron XMR


CAM partition profiles

CAM is partitioned on the device by a variety of profiles that you can select depending on your application. The available profiles are described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series.

To implement a CAM partition profile, enter the following command:

```
configure terminal

ip routing cam partition profile profile-name
```

where profile-name is the profile name described in Table 47 or 48.

The parameter names used in the table are as follows:

- **Default**: This is the default profile that is used when the device is initialized or reset.
- **IpV4**: This is the IPv4 profile that is used for IPv4 traffic.
- **IpV6**: This is the IPv6 profile that is used for IPv6 traffic.
- **MAC or VPLS**: This is the profile that is used for MAC or VPLS traffic.
- **IPV4 or L2 Inbound ACL**: This is the profile that is used for IPv4 or L2 inbound ACL traffic.
- **IPV4 or L2 Outbound ACL**: This is the profile that is used for IPv4 or L2 outbound ACL traffic.
- **IPV6 Inbound ACL**: This is the profile that is used for IPv6 inbound ACL traffic.
- **IPV6 Outbound ACL**: This is the profile that is used for IPv6 outbound ACL traffic.

The default profile for FC383/FC382 is 512K IPV4 and IPv6 routes. These numbers can be increased to 64K or 128K or 512K as you increase the partition size. The default profile for FC381/PC332 is 128K IPV4 and 512K IPv6 routes. The default profile for FC380/PC331 is 128K IPV4 and 512K IPv6 routes. If you increase the partition size, the default value, 512K, automatically takes up the IPv6 space and vice versa.

### Table 47 CAM partitioning profiles available for Brocade NetIron XMR routers

<table>
<thead>
<tr>
<th>Profile</th>
<th>IPv4</th>
<th>IPv6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Logical size</strong></td>
<td><strong>Logical size</strong></td>
<td><strong>Logical size</strong></td>
</tr>
<tr>
<td><strong>Default Profile</strong></td>
<td>512K</td>
<td>64K</td>
</tr>
<tr>
<td><strong>ipv4 Profile</strong></td>
<td>1M</td>
<td>32K</td>
</tr>
</tbody>
</table>

---

Cisco Cat 6500


Introduction

This document describes how to configure various management interfaces (MI) on Cisco Catalyst 6500 switches that run the Supervisor Engine 720.

Prerequisites

Requirements

Components Used

There are no specific requirements for the document.

Components Used

The information in this document is based on a Cisco Catalyst 6500 switch that runs as a Supervisor Engine 720 with PFC381/PC332, PFC382/PC331, and PFC383/PC330 support modules.

The information in this document is based on the switch software version 4.2(3a) and is subject to change. All the devices used in this document are configured using the default (a) configuration. If required to test, make sure that you understand the potential limitations and constraints.

Problem

As described in the document, PFC381/PC332, PFC382/PC331, and PFC383/PC330 support modules (IPv6 routes and 512K, 128K, 64K) support modules. However, default outputs lack information.

- **MAC or VPLS**: This is the profile that is used for MAC or VPLS traffic.
- **IPV4 or L2 Inbound ACL**: This is the profile that is used for IPv4 or L2 inbound ACL traffic.
- **IPV4 or L2 Outbound ACL**: This is the profile that is used for IPv4 or L2 outbound ACL traffic.
- **IPV6 Inbound ACL**: This is the profile that is used for IPv6 inbound ACL traffic.
- **IPV6 Outbound ACL**: This is the profile that is used for IPv6 outbound ACL traffic.

- **Logical size**: This is the size in bytes of the partition.
- **Logical size**: This is the size in bytes of the partition.
- **Logical size**: This is the size in bytes of the partition.
- **Logical size**: This is the size in bytes of the partition.
- **Logical size**: This is the size in bytes of the partition.
- **Logical size**: This is the size in bytes of the partition.
- **Logical size**: This is the size in bytes of the partition.
- **Logical size**: This is the size in bytes of the partition.

Table 47 CAM partitioning profiles available for Brocade NetIron XMR routers

<table>
<thead>
<tr>
<th>Profile</th>
<th>IPv4</th>
<th>IPv6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Logical size</strong></td>
<td><strong>Logical size</strong></td>
<td><strong>Logical size</strong></td>
</tr>
<tr>
<td><strong>Default Profile</strong></td>
<td>512K</td>
<td>64K</td>
</tr>
<tr>
<td><strong>ipv4 Profile</strong></td>
<td>1M</td>
<td>32K</td>
</tr>
</tbody>
</table>
What happens then?

- Crash and reboot?
- Crash and die?
- Push excess routes to slow path?
- Discard excess routes

Was there any evidence of dropped routes?
Dropped Routes?

Net change of -450 routes
Maybe there's more...
Collateral Damage

Outside of AS701, a further ~2,000 routes were withdrawn between 07:47 and 12:00, but some of these were probably part of the route leak as they appeared to be part of the Verizon enterprise structure. But there were others who were clearly unrelated to Verizon...
Collateral Damage

763 Origin ASes were probably affected

<table>
<thead>
<tr>
<th>AS Pfxs</th>
<th>AS Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>9658</td>
<td>ETPI-IDS-AS-AP Eastern Telecoms Phils., Inc.,PH</td>
</tr>
<tr>
<td>6648</td>
<td>BAYAN-TELECOMMUNICATIONS Bayan Telecommunications, Inc.,PH</td>
</tr>
<tr>
<td>23498</td>
<td>CDSI - COGECODATA,CA</td>
</tr>
<tr>
<td>21332</td>
<td>NTC-AS OJSC &quot;Vimpelcom&quot;,RU</td>
</tr>
<tr>
<td>27882</td>
<td>Telefonica Celular de Bolivia S.A.,BO</td>
</tr>
<tr>
<td>30036</td>
<td>MEDIACOM-ENTERPRISE-BUSINESS - Mediacom Communications Corp,US</td>
</tr>
<tr>
<td>131222</td>
<td>MTS-INDIA-IN 334,Udyog Vihar,IN</td>
</tr>
<tr>
<td>46805</td>
<td>CACHED - CachedNet LLC,US</td>
</tr>
<tr>
<td>45664</td>
<td>LBNI Liberty Broadcasting Network Inc,PH</td>
</tr>
<tr>
<td>8402</td>
<td>CORBINA-AS OJSC &quot;Vimpelcom&quot;,RU</td>
</tr>
<tr>
<td>55465</td>
<td>TTT-AS-AP TT&amp;T Co,TH</td>
</tr>
<tr>
<td>18025</td>
<td>ACE-1-WIFI-AS-AP Ace-1 Wifi Network,PH</td>
</tr>
<tr>
<td>22363</td>
<td>PHMGMT-AS1 - Powerhouse Management, Inc.,US</td>
</tr>
<tr>
<td>15085</td>
<td>IMMEDION - Immedion, LLC,US</td>
</tr>
<tr>
<td>50710</td>
<td>EARTHLINK-AS EarthLink Ltd. Communications&amp;Internet Services,IQ</td>
</tr>
<tr>
<td>21284</td>
<td>VIVODI-AS ON S.A.,GR</td>
</tr>
<tr>
<td>23606</td>
<td>BELLTELECOM-PH-AS-PH Bell Telecommunication Philippines,PH</td>
</tr>
<tr>
<td>7018</td>
<td>ATT-INTERNET4 - AT&amp;T Services, Inc.,US</td>
</tr>
<tr>
<td>50576</td>
<td>KRASNET-UA-AS Krasnet ltd.,UA</td>
</tr>
<tr>
<td>16058</td>
<td>Gabon-Telecom,GA</td>
</tr>
<tr>
<td>13188</td>
<td>BANKINFORM-AS TOV &quot;Bank-Inform&quot;,UA</td>
</tr>
</tbody>
</table>
But then it happened again!

BGP FIB Size

Verizon Leak!

A second time!
So maybe we should broaden the question... Was the AS701 Route Leak the problem?

Or was the FIB growth passing 512K entries the problem?
There is no Routing God!

There is no single objective “out of the system” view of the Internet’s Routing environment.

BGP distributes a routing view that is modified as it is distributed, so every eBGP speaker will see a slightly different set of prefixes, and each view is relative to a given location.

When we look at some of the route collector sites we see a variance of ~20,000 routes across the routing peer set.
The RouteViews View

300,000 now
Zooming in

12 August

10 October

512K
For most networks...

(probably including yours)

It's likely that your router’s routing table has yet to pass over the 512K point

(Except for the occasional route leak of course)

So there is still some time to check if you can cope with a steady-state default free routing table with more than 512K entries
For most affected networks, it was the AS701 route leak that tipped them over the edge on the day.

However, passing through 512K routes in the IPv4 routing table is inevitable.

When? And what’s next?

How quickly is the routing table growing?
20 years of routing the Internet

- 2001: The Great Internet Boom and Bust
- 2005: Broadband to the Masses
- 2009: The GFC hits the Internet
- 2011: Address Exhaustion

1994: Introduction of CIDR
IPv4 in 2014 - Growth is Slowing (slightly)

• Overall IPv4 Internet growth in terms of BGP is at a rate of some ~9%-10% p.a.

• Address span growing far more slowly than the table size (although the LACNIC runout in May caused a visible blip in the address rate)

• The rate of growth of the IPv4 Internet is slowing down (slightly)
  – Address shortages?
  – Masking by NAT deployments?
  – Saturation of critical market sectors?
IPv6 BGP Prefix Count

V6 BGP FIB Size

World IPv6 Day
IPv6 in 2013

• Overall IPv6 Internet growth in terms of BGP is 20% - 40 % p.a.
  – 2012 growth rate was ~ 90%.

If these relative growth rates persist then the IPv6 network would span the same network domain as IPv4 in ~16 years time.
What to expect
BGP Size Projections

• For IPv4 this is a time of **extreme uncertainty**
  • Registry IPv4 address run out
  • Uncertainty over the impacts of any after-market in IPv4 on the routing table

which makes this projection even more speculative than normal!
V4 - Relative Daily Growth Rates
V4 - Relative Daily Growth Rates
## IPv4 BGP Table Size predictions

<table>
<thead>
<tr>
<th>Year</th>
<th>Linear Model</th>
<th>Exponential Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 2013</td>
<td>441,172 entries</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>488,011 entries</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>540,000 entries</td>
<td>559,000</td>
</tr>
<tr>
<td>2016</td>
<td>590,000 entries</td>
<td>630,000</td>
</tr>
<tr>
<td>2017</td>
<td>640,000 entries</td>
<td>710,000</td>
</tr>
<tr>
<td>2018</td>
<td>690,000 entries</td>
<td>801,000</td>
</tr>
<tr>
<td>2019</td>
<td>740,000 entries</td>
<td>902,000</td>
</tr>
</tbody>
</table>

*These numbers are dubious due to uncertainties introduced by IPv4 address exhaustion pressures.*
V6 - Relative Growth Rates

![Graph showing relative growth rates over time.](image-url)
V6 - Relative Growth Rates
V6 - Relative Growth Rates
<table>
<thead>
<tr>
<th>Year</th>
<th>IPv6 BGP Table Size predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 2013</td>
<td>11,600 entries</td>
</tr>
<tr>
<td>2014</td>
<td>16,200 entries</td>
</tr>
<tr>
<td>2015</td>
<td>24,600 entries</td>
</tr>
<tr>
<td>2016</td>
<td>36,400 entries</td>
</tr>
<tr>
<td>2017</td>
<td>54,000 entries</td>
</tr>
<tr>
<td>2018</td>
<td>80,000 entries</td>
</tr>
<tr>
<td>2019</td>
<td>119,000 entries</td>
</tr>
</tbody>
</table>
IPv4 BGP Table size and Moore's Law

Moore's Law

BGP Table Size Predictions
IPv6 Projections and Moore's Law

The graph illustrates the growth of IPv6 BGP table size predictions over time, comparing different models against Moore's Law. The y-axis represents IPv6 BGP table size, and the x-axis represents dates from January 2006 to January 2018. The graph shows varying trends for different models, with Moore's Law depicted as a significant growth projection.
BGP Table Growth

• Nothing in these figures suggests that there is cause for urgent alarm -- at present
• The overall eBGP growth rates for IPv4 are holding at a modest level, and the IPv6 table, although it is growing rapidly, is still relatively small in size in absolute terms
• As long as we are prepared to live within the technical constraints of the current routing paradigm it will continue to be viable for some time yet
BGP Updates

• What about the level of updates in BGP?
• Let’s look at the update load from a single eBGP feed in a DFZ context
Announcements and Withdrawals

Daily BGP v4 Update Activity for AS131072
Convergence Performance

Average Convergence Time per day (AS 131072)
IPv4 Average AS Path Length

Data from Route Views
Updates in IPv4 BGP

Nothing in these figures is cause for any great level of concern ...

– The number of updates per instability event has been constant, due to the damping effect of the MRAI interval, and the relatively constant AS Path length over this interval

What about IPv6?
V6 Announcements and Withdrawals

Daily BGP v6 Update Activity for AS131072

- Withdrawals
- Announcements
- Total
- BGP FIB Size
V6 Average AS Path Length

Data from Route Views
Problem? Not a Problem?

It’s evident that the global BGP routing environment suffers from a certain amount of neglect and inattention.

But whether this is a problem or not depends on the way in which routers handle the routing table.

So let’s take a quick look at routers...
Inside a router

- Backplane
- Line Interface Card
- Switch Fabric Card
- Management Card

Thanks to Greg Hankins
Inside a line card

- Packet Buffer
- DRAM
- TCAM
- *DRAM
- Network
- PHY
- CPU
- Backplane
- Media
- FIB Lookup Bank

Thanks to Greg Hankins
Inside a line card

Packet Buffer

DRAM

TCAM

*DRAM

Network

PHY

FIB Lookup Bank

Thanks to Greg Hankins
The interface card’s network processor passes the packet’s destination address to the FIB module.

The FIB module returns with an outbound interface index.
FIB Lookup

This can be achieved by:

– Loading the entire routing table into a Ternary Content Addressable Memory bank (TCAM)

or

– Using an ASIC implementation of a TRIE representation of the routing table with DRAM memory to hold the routing table

Either way, this needs fast memory
The entire FIB is loaded into TCAM. Every destination address is passed through the TCAM, and within one TCAM cycle the TCAM returns the interface index of the longest match. Each TCAM bank needs to be large enough to hold the entire FIB. TTCAM cycle time needs to be fast enough to support the max packet rate of the line card.

TCAM width depends on the chip set in use. One popular TCAM config is 72 bits wide. IPv4 addresses consume a single 72 bit slot, IPv6 consumes two 72 bit slots. If instead you use TCAM with a slot width of 32 bits then IPv6 entries consume 4 times the equivalent slot count of IPv4 entries.
The entire FIB is converted into a serial decision tree. The size of decision tree depends on the distribution of prefix values in the FIB. The performance of the TRIE depends on the algorithm used in the ASIC and the number of serial decisions used to reach a decision.
# Memory Tradeoffs

<table>
<thead>
<tr>
<th></th>
<th>TCAM</th>
<th>ASIC + RLDRAM 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access Speed</td>
<td>Lower</td>
<td>Higher</td>
</tr>
<tr>
<td>$ per bit</td>
<td>Higher</td>
<td>Lower</td>
</tr>
<tr>
<td>Power</td>
<td>Higher</td>
<td>Lower</td>
</tr>
<tr>
<td>Density</td>
<td>Higher</td>
<td>Lower</td>
</tr>
<tr>
<td>Physical Size</td>
<td>Larger</td>
<td>Smaller</td>
</tr>
<tr>
<td>Capacity</td>
<td>80Mbit</td>
<td>16 bit</td>
</tr>
</tbody>
</table>

Thanks to Greg Hankins
Memory Tradeoffs

TCAMs are higher cost, but operate with a fixed search latency and a fixed add/delete time. TCAMs scale linearly with the size of the FIB.

ASICs implement a TRIE in memory. The cost is lower, but the search and add/delete times are variable. The performance of the lookup depends on the chosen algorithm. The memory efficiency of the TRIE depends on the prefix distribution and the particular algorithm used to manage the data structure.
What memory size do we need for 10 years of FIB growth from today?

**TCAM**
- V4: 2M entries (1Gt)
- plus
- V6: 1M entries (2Gt)

**Trie**
- V4: 100Mbit memory (500Mt)
- plus
- V6: 200Mbit memory (1Gt)

“*The Impact of Address Allocation and Routing on the Structure and Implementation of Routing Tables*, Narayn, Govindan & Varghese, SIGCOMM ’03
Scaling the FIB

BGP table growth is slow enough that we can continue to use simple FIB lookup in linecards without straining the state of the art in memory capacity.

However, if it all turns horrible, there are alternatives to using a complete FIB in memory, which are at the moment variously robust and variously viable:

- FIB compression
- MPLS
- Locator/ID Separation (LISP)
- OpenFlow/Software Defined Networking (SDN)
But it's not just size

It’s speed as well.  
10Mb Ethernet had a 64 byte min packet size, plus preamble plus inter-packet spacing  
  =14,880 pps  
  =1 packet every 67usec

We’ve increased speed of circuits, but left the Ethernet framing and packet size limits largely unaltered. What does this imply for router memory?
Wireline Speed – Ethernet

- 10Mb 1982 / 15Kpps
- 100Mb 1995 / 150Kpps
- 1Gb 1999 / 1.5Mpps
- 10Gb 2002 / 15Mpps
- 40Gb/100Gb 2010 / 150Mpps
- 400Gb/1Tb 2017? 15G pps
Clock Speed - Processors

Intel CPU Speeds Over Time

- 486DX 66MHz
- Pentium 100MHz
- Pentium III 600MHz
- Pentium 4 EE 3.2GHz
- Core 2 QX6700 2.66GHz
- Core i7 920 2.66/2.93GHz
- Core i7 3960X 3.33/3.9GHz
- Core i7 2600K 3.4/3.8GHz

MIPS (logarithmic scale)

Year

1 core 4 cores
CPU vs Memory Speed

The graph shows the performance of processors and memories over the years from 1980 to 2010. The performance is measured on the vertical axis, and the year is marked on the horizontal axis. The performance gap between processors and memories is highlighted with an arrow, indicating a significant difference in their growth rates over the years.
Speed, Speed, Speed

What memory speeds are necessary to sustain a maximal packet rate?

100G E ≈ 150Mpps ≈ 6.7ns per packet

400G e ≈ 600Mpps ≈ 1.6ns per packet

1Te ≈ 1.5Gpps ≈ 0.67ns per packet
Speed, Speed, Speed

What memory speeds do we have today?

- DDR3DRAM: 9ns - 15ns
- RLDRAM: 1.9ns - 12ns
- Commodity DRAM

Thanks to Greg Hankins
Scaling Speed

Scaling size is not a dramatic problem today
Scaling speed is going to be tougher over time

Moore’s Law talks about the number of gates per circuit, but
not circuit clocking speeds
Speed and capacity could be the major design challenge for
network equipment in the coming years
If we can’t route the max packet rate for a terrabit wire
then:

• If we want to exploit parallelism as an alternative to
wireline speed for terrabit networks, then is the use of
best path routing protocols, coupled with destination-
based hop-based forwarding going to scale?
• Or are we going to need to look at path-pinned routing
architectures to provide stable flow-level parallelism
within the network to limit aggregate flow volumes?
• Or should we reduce the max packet rate by moving
away from a 64byte min packet size?

http://www.startupinnovation.org/research/moores-law/
Thank You

Questions?