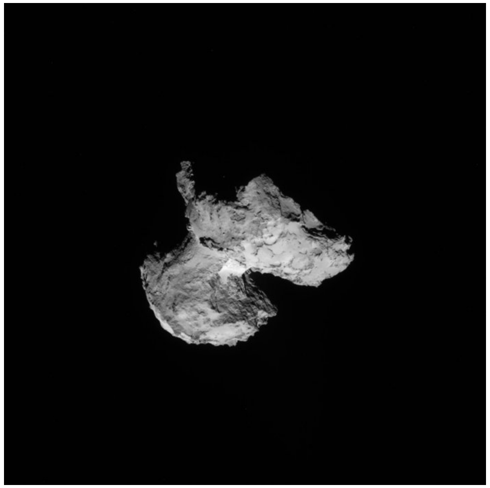
The 512K route thing

12 August 2014

World Elephant Day



Rosetta closes in on comet 67P/ Churyumov-Gerasimenko

Newborn Panda Triplets in China

The internet apparently has a bad hair day

The Telegraph

Home News World Sport Finance Comment Culture

Technology News | Technology Companies | Technology Reviews

HOME » TECHNOLOGY » INTERNET

Is the Internet full? Major sites brought problems

Likely repeat of this week's technical problems affecting eBay would cost the economy millions as the Internet runs out of space, experts fear

The Washington Post

Sections Search Sign In Subscribe

The Switch

Here's why your Internet might have been slow on Tuesday

12 More

By Andrea Peterson August 13 Follow @kansasalps

Some users were frustrated to find some of their favorite Web sites were unresponsive or otherwise inaccessible Tuesday. But it wasn't a data center outage or a squirrel chewing through a cable line causing the disruption. Instead, structural problems with one of the core technologies that keeps the Internet working were to blame, researchers say.

THE WALL STREET JOURNAL. TECH

LOG IN | SUBSCRIBE

TOP STORIES IN TECH

1 of 12 Data Breach Puts Focus on Beefed-Up Car...

2 of 12 What Happens When Police Wear Cameras

3 of 12 Vintage Videogame Venues

TWC Deal Complicates Comcast Merger Plan

TECHNOLOGY

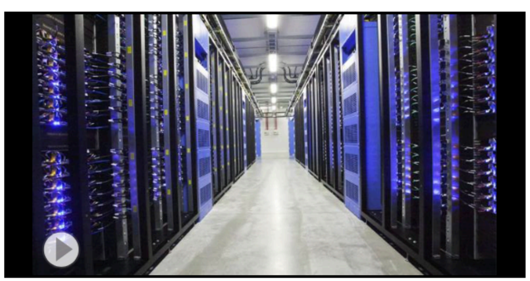
Echoes of Y2K: Engineers Buzz That Internet Is Outgrowing Its Gear

Routers That Send Data Online Could Become Overloaded as Number of Internet Routes Hits '512K'

Email Print 53 Comments

ARTICLE FREE PASS
Enjoy your free sample of exclusive subscriber content.

By DREW FITZGERALD CONNECT
Updated Aug. 13, 2014 7:38 p.m. ET



Network engineers are buzzing that that the internet is outgrowing some of its gear. WSJ's Drew FitzGerald discusses what that means on Lunch Break with Sara Murray. Photo: Getty

Network engineers are buzzing this week as the Internet outgrows some of its gear.

MarketWatch MARKETS NEWS

10:26 pm / Nov 6, 2013

Popular Now What's This?

ARTICLES

1 Things That Jeff Bridges Can't Abide

2 Autopsy Finds 6 Shots Killed Teen

news.com.au

National World Finance Sport Entertainment Lifestyle Travel Technology Video

You are here: news.com.au Technology Online Social

online

LATEST IN ONLINE

Internet service dead as a dodo

Social media reacts to vigilante shop owners

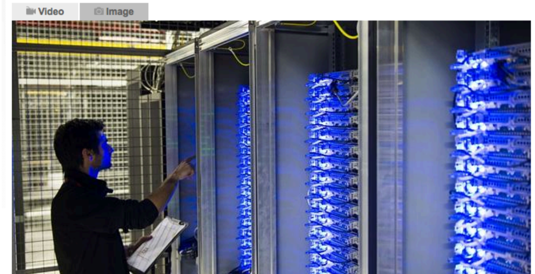
Piracy coming apart at the streams

These people will help you change your life

The internet broke yesterday and it was all because of the number 512

This story was published: 4 DAYS AGO | AUGUST 14, 2014 2:15PM

Video Image

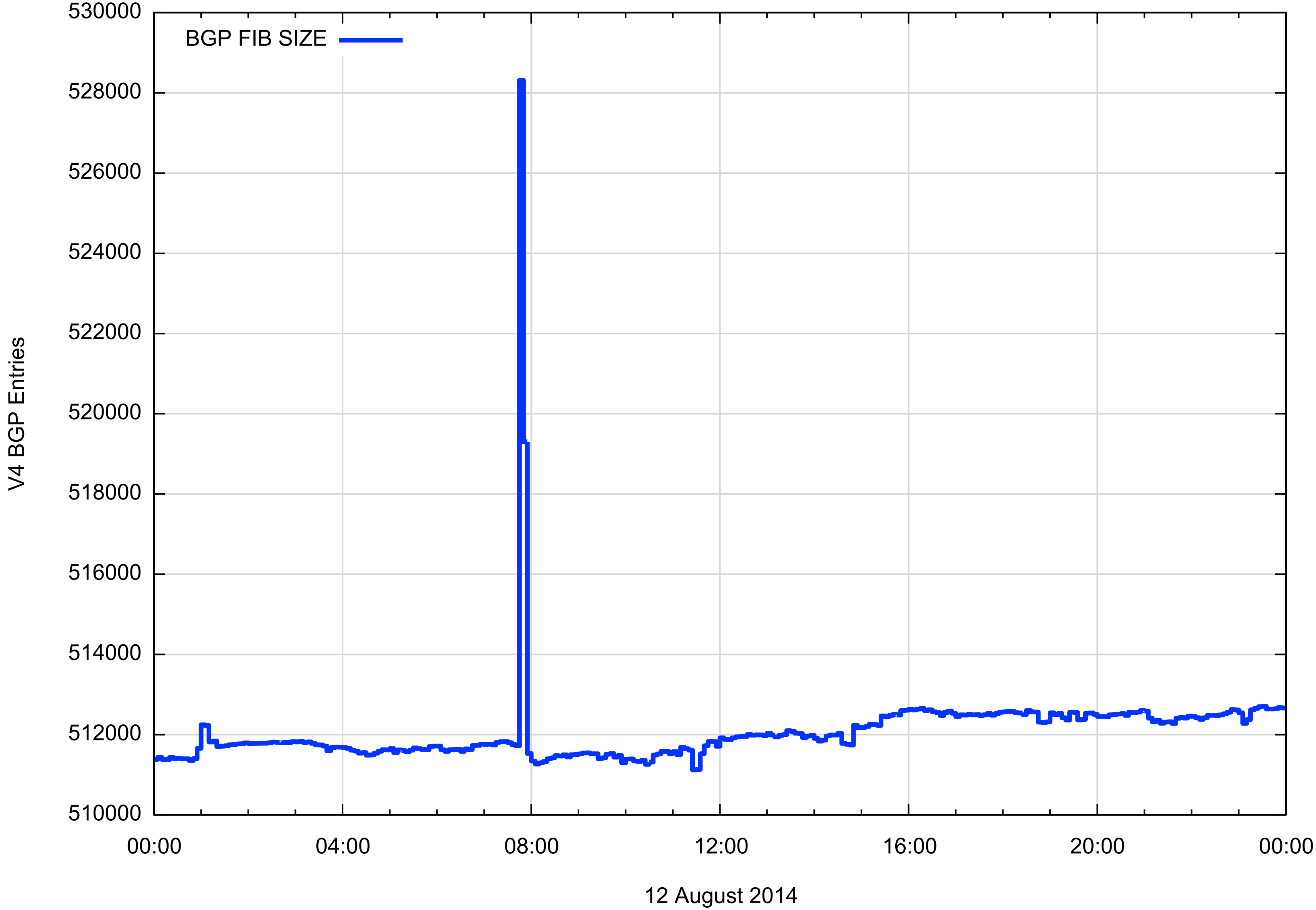


STORY BY
Network Writer
News Corp Australia

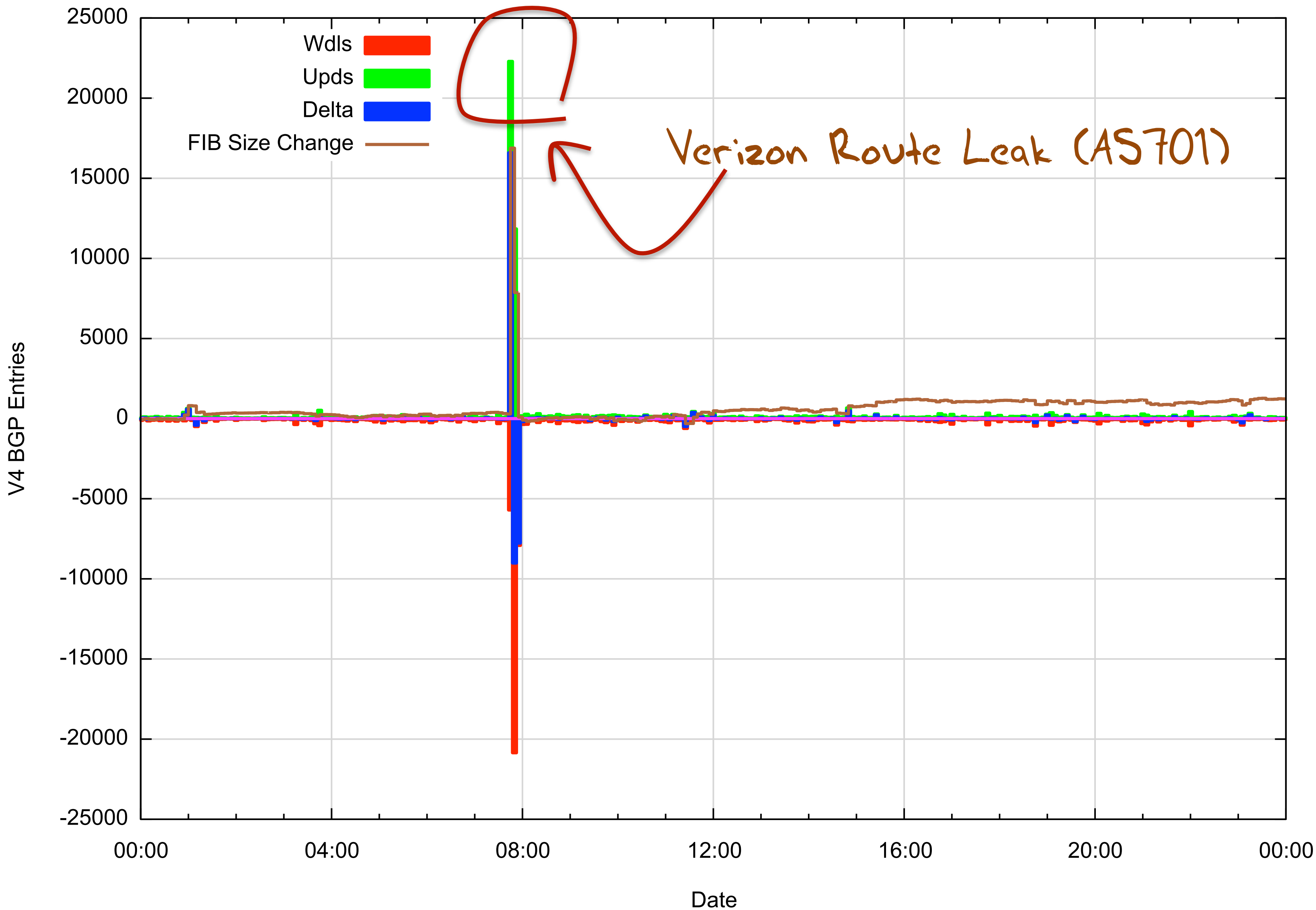
What happened?

Did we all sneeze at once and cause the routing system to fail?

Well someone sneezed!



12 August 2014



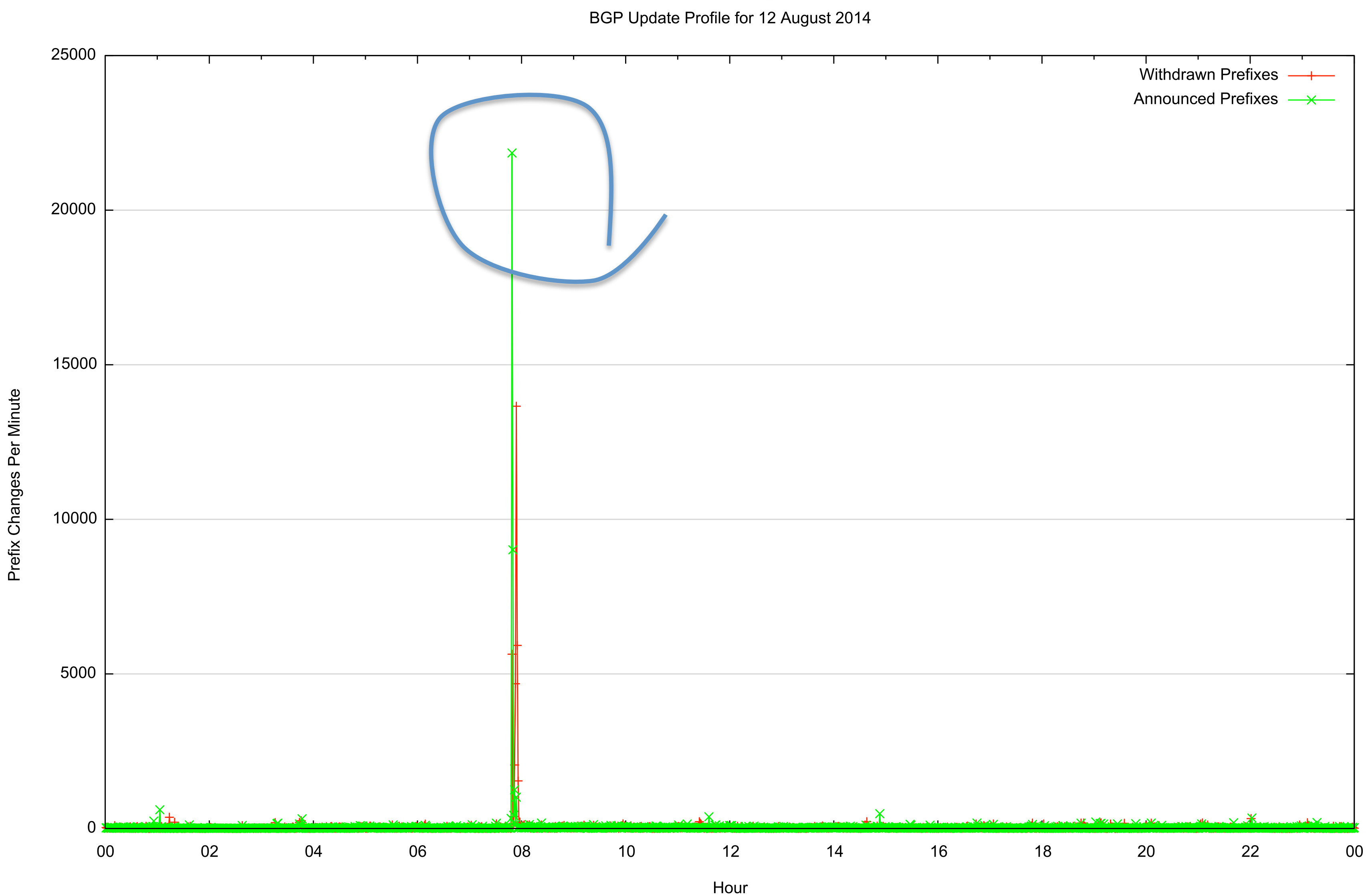
But route leaks happen all the time

But not from AS701!

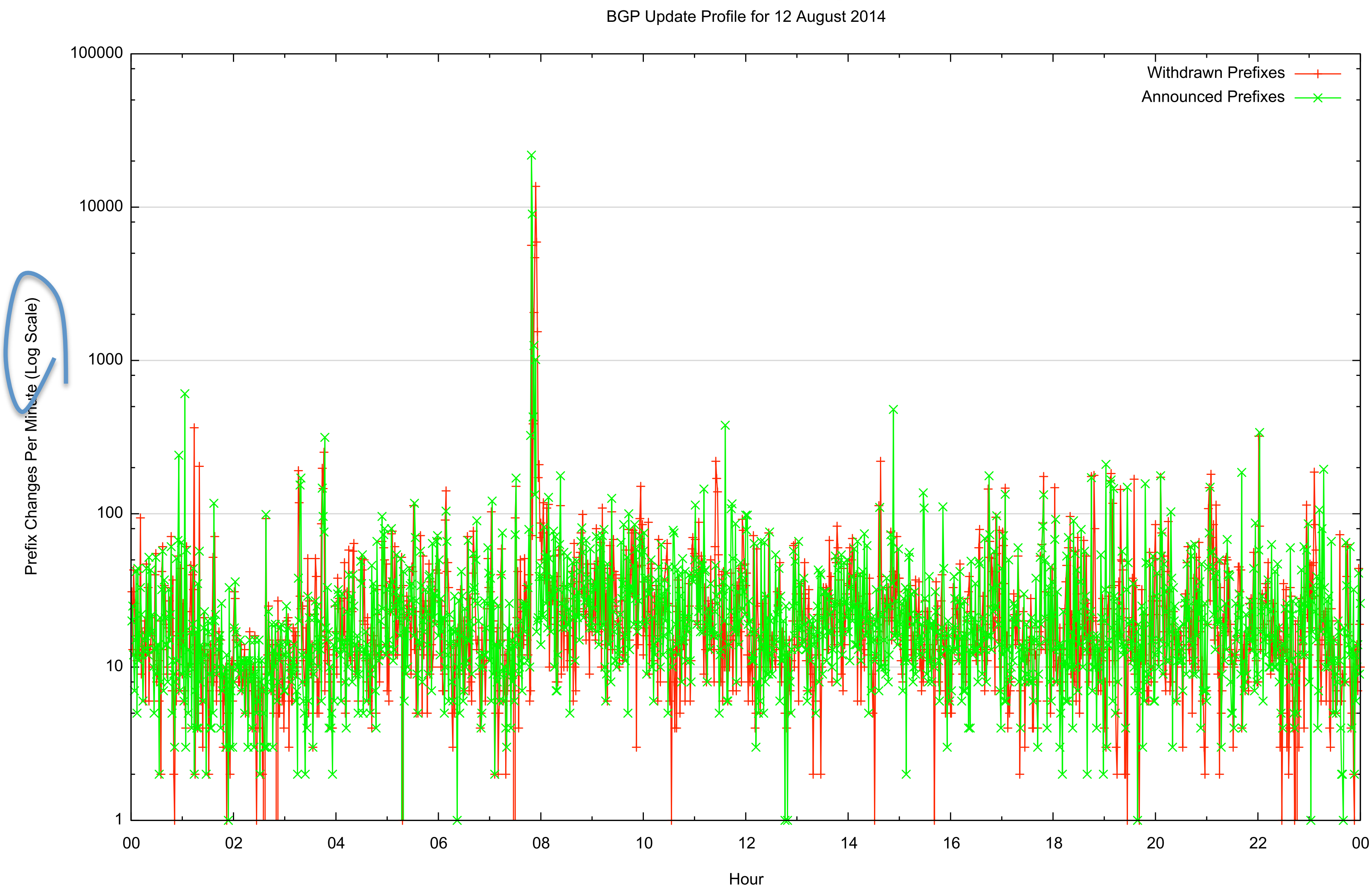
- AS701 is a tier 1 ISP
- So very few (noone?) filters what they hear from AS701
- Which means that when AS701 leaks all non-default AS's (and a few more besides) are likely to hear the route leak

So everybody saw a bunch of routes for a small amount of time…

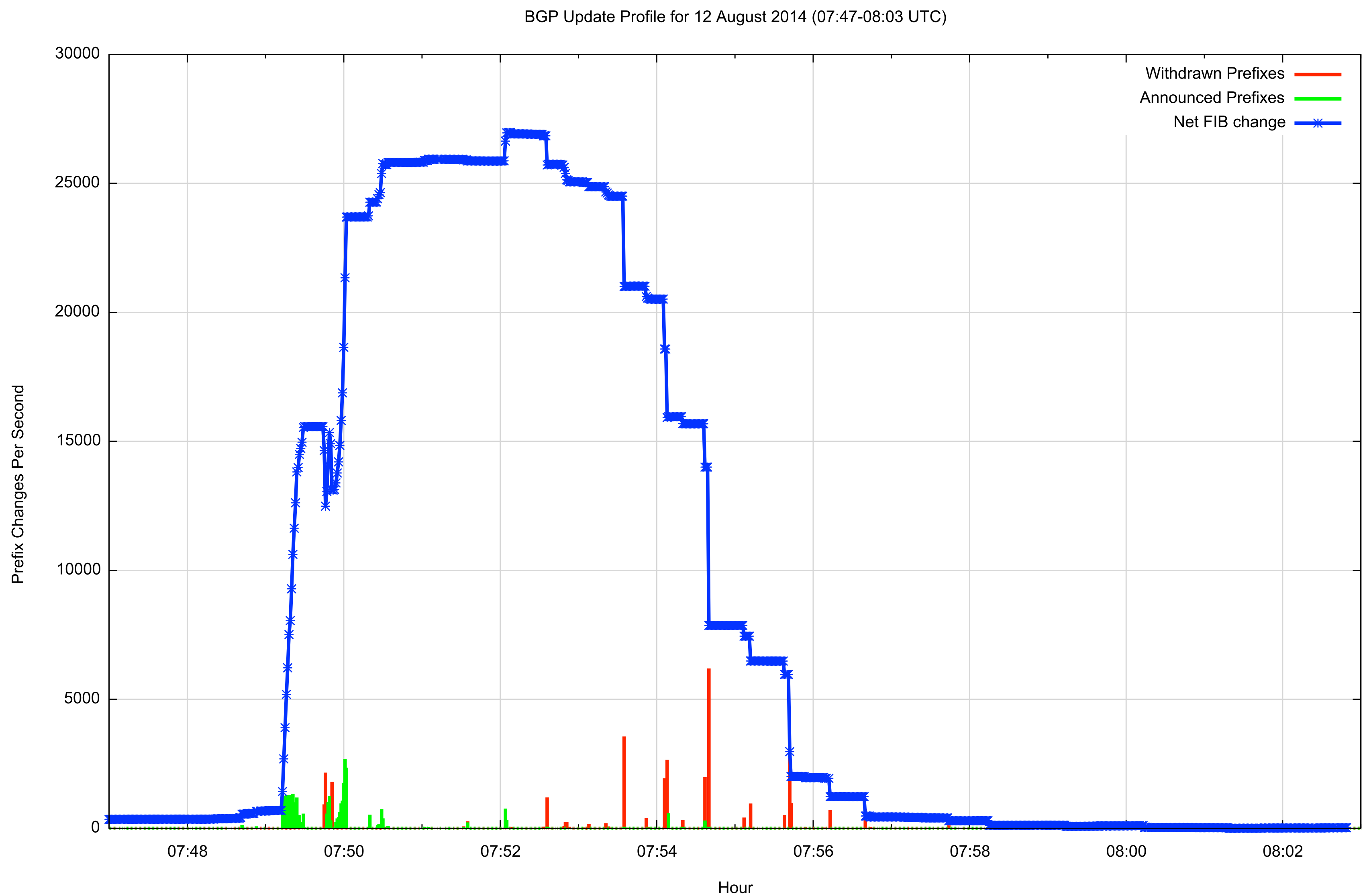
Minute by Minute ANNces & WDLs



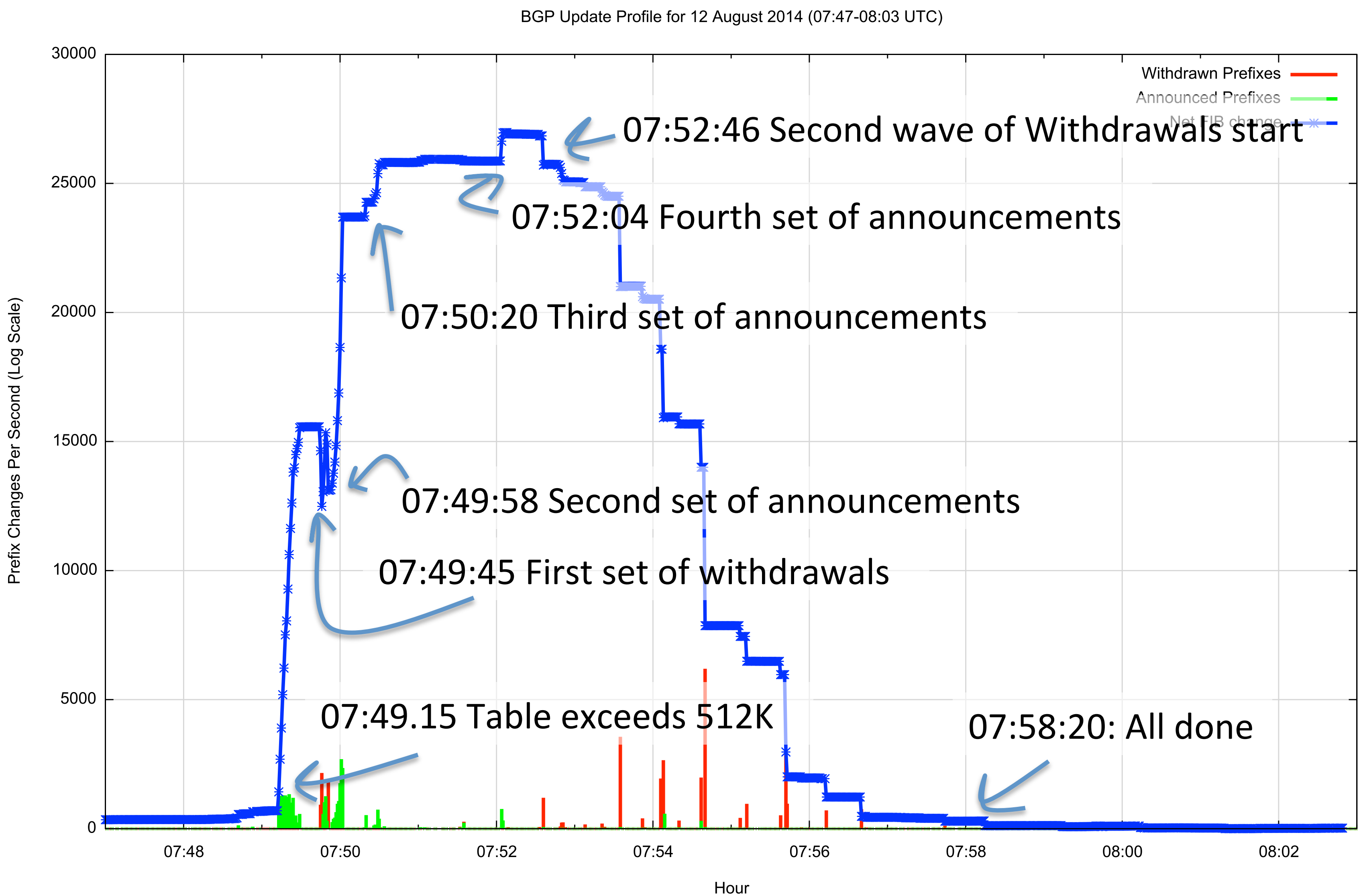
Minute by Minute ANNces & WDLs



Second by Second



Second by Second



http://www.cisco.com/c/en/us/support/docs/switches/catalyst-6500-series-switches/116132-problem-catalyst6500-00.html

512K is a default constant in some of the older Cisco and Brocade products

Worldwide [change] | Log In | Account | Register | My Cisco

cisco | Products & Services | Support | How to Buy | Training & Events | Partners

Cisco Catalyst 6500 Series Switches

Catalyst 6500 Switches Ternary Content Addressable Memory Customization

HOME
SUPPORT
PRODUCT SUPPORT
SWITCHES
CISCO CATALYST 6500 SERIES SWITCHES
TROUBLESHOOT AND ALERTS
TROUBLESHOOTING TECHNOTES
Catalyst 6500 Switches Ternary Content Addressable Memory Customization

Contents

Introduction

This document describes how to customize the forwarding information base (FIB) ternary content addressable memory (TCAM) on Catalyst 6500 switches that run the Supervisor Engine 720.

Prerequisites

Requirements

There are no specific requirements for this document.

Components Used

The information in this document is based on a Cisco Catalyst 6500 switch that runs on a Supervisor Engine 720 with PFC3BXL/PFC3CXL.

The information in this document was created from the devices in a specific lab environment. All of the devices used in this document started with a cleared (default) configuration. If your network is live, make sure that you understand the potential impact of any command.

Problem

As outlined in the datasheet, PFC3BXL and PFC3CXL support one millon (1M) IPv4 routes and 512,000 (512k) IPv6 routes. However, default outputs look different:

```
6500#show mls cef maximum-routes
FIB TCAM maximum routes :
=======================
Current :
---------
 IPv4 + MPLS         - 512k (default)
 IPv6 + IP Multicast - 256k (default)
```

Solution

The default numbers for PFC3BXL/PFC3CXL are 512k IPv4 routes and 256k IPv6 routes. These numbers can be increased to 1M IPv4 OR 512k IPv6 routes if you enter **mls cef maximum-routes ip|ipv6 []** and reload. But, you cannot achieve both 1M IPv4 AND 512k IPv6 routes at the same time. If you increase the IPv4 TCAM size above the default value, it automatically takes up the IPv6 space and vice versa.

TAC

Document ID: 116132
Updated: Jun 04, 2013

Contributed by Shashank Singh, Cisco TAC Engineer.

Download PDF Print

[+] Feedback

Related Documents

Catalyst 6500 Switches Ternary Content Addressable Memory Customization - Cisco
Cisco Catalyst 6500 Borderless Services Node - Cisco
Cisco Catalyst 6509 Switch - Cisco
Cisco Catalyst 6503-E Switch - Cisco
Cisco Catalyst 6504-E Switch - Cisco.
More...

Related Products

Cisco Express Forwarding (CEF)
Cisco Catalyst 6500 Series Switches

The default numbers for PFC3BXL/PFC3CXL are 512k IPv4 routes and 256k IPv6 routes. These numbers can be increased to 1M IPv4 OR 512k IPv6 routes if you enter **mls cef maximum-routes ip|ipv6 []** and reload. But, you cannot achieve both 1M IPv4 AND 512k IPv6 routes at the same time. If you increase the IPv4 TCAM size above the default value, it automatically takes up the IPv6 space and vice versa.

Cisco Cat 6500

Brocade NetIron XMR

http://www.brocade.com/downloads/documents/html_product_manuals/NI_05600_ADMIN/wwhelp/wwhimpl/common/html/wwhelp.htm#context=Admin_Guide&file=CAM_part.11.2.html

Multi-Service Ironware Administration Guide
R05.6.00
Part Number: 53-1003028-02
documentation@brocade.com

Foundry Direct Routing and CAM Partition Profiles for the NetIron XMR and the Brocade MLX Series : CAM partition profiles

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.

```
Brocade(config) cam-partition profile ipv4
```

Syntax: cam-partition profile [ipv4 | ipv4-ipv6| ipv4-ipv6-2 | ipv4-vpls | ipv4-vpn | ipv6 | l2-metro | l2-metro-2 | mpls-l3vpn | mpls-l3vpn-2 | mpls-vpls | mpls-vpls-2 | mpls-vpn-vpls | multi-service | multi-service-2 |multi-service-3 | multi-service-4]

The **ipv4** parameter adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series, to optimize the device for IPv4 applications.

The **ipv4-ipv6** parameter adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series, to optimize the device for IPv4 and IPv6 dual stack applications

The **ipv4-ipv6-2** parameter that was introduced in version 03.7.00, adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series, to optimize the device for increased IPv4 routes with room or IPv6.

The **ipv4-vpls** parameter adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series, to optimize the device for IPv4 and MPLS VPLS applications

The **ipv4-vpn** parameter adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series, to optimize the device for IPv4 and MPLS Layer-3 VPN applications

The **ipv6** parameter adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series, to optimize the device for IPv6 applications.

The **l2-metro** parameter adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series routers, to optimize the device for Layer 2 Metro applications.

The **l2-metro-2** parameter provides another alternative to **l2-metro** to optimize the device for Layer 2 Metro applications. It adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series routers.

The **mpls-l3vpn** parameter adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series routers, to optimize the device for Layer 3, BGP or MPLS VPN applications.

The **mpls-l3vpn-2** parameter provides another alternative to **mpls-l3vpn** to optimize the device for Layer 3, BGP or MPLS VPN applications. It adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series routers.

The **mpls-vpls** parameter adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series routers, to optimize the device for MPLS VPLS applications.

The **mpls-vpls-2** parameter provides another alternative to **mpls-vpls** to optimize the device for MPLS VPLS applications. It adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series routers.

The **mpls-vpn-vpls** parameter adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series routers, to optimize the device for MPLS Layer-3 and Layer-2 VPN applications.

The **multi-service** parameter adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series routers, to optimize the device for Multi-Service applications.

The **multi-service-2** parameter provides another alternative to **multi-service** to optimize the device for Multi-Service applications. It adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series routers.

NOTE: You must reload your device for this command to take effect.

The **multi-service-3** parameter provides another alternative to multi-service to optimize the device for Multi-Service applications to support IPv6 VRF. It adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series routers.

The **multi-service-4** parameter provides another alternative to multi-service to optimize the device for Multi-Service applications to support IPv6 VRF. It adjusts the CAM partitions, as described in Table 47 for Brocade NetIron XMR and Table 48 for Brocade MLX series routers.

There are fourteen CAM partitioning profiles for Brocade NetIron XMR and Table 48 for Brocade MLX series routers.The profiles for Brocade XMR routers are described in Table 47 and the profiles for Brocade MLX routers are described in Table 48.

TABLE 47 CAM partitioning profiles available for Brocade NetIron XMR routers

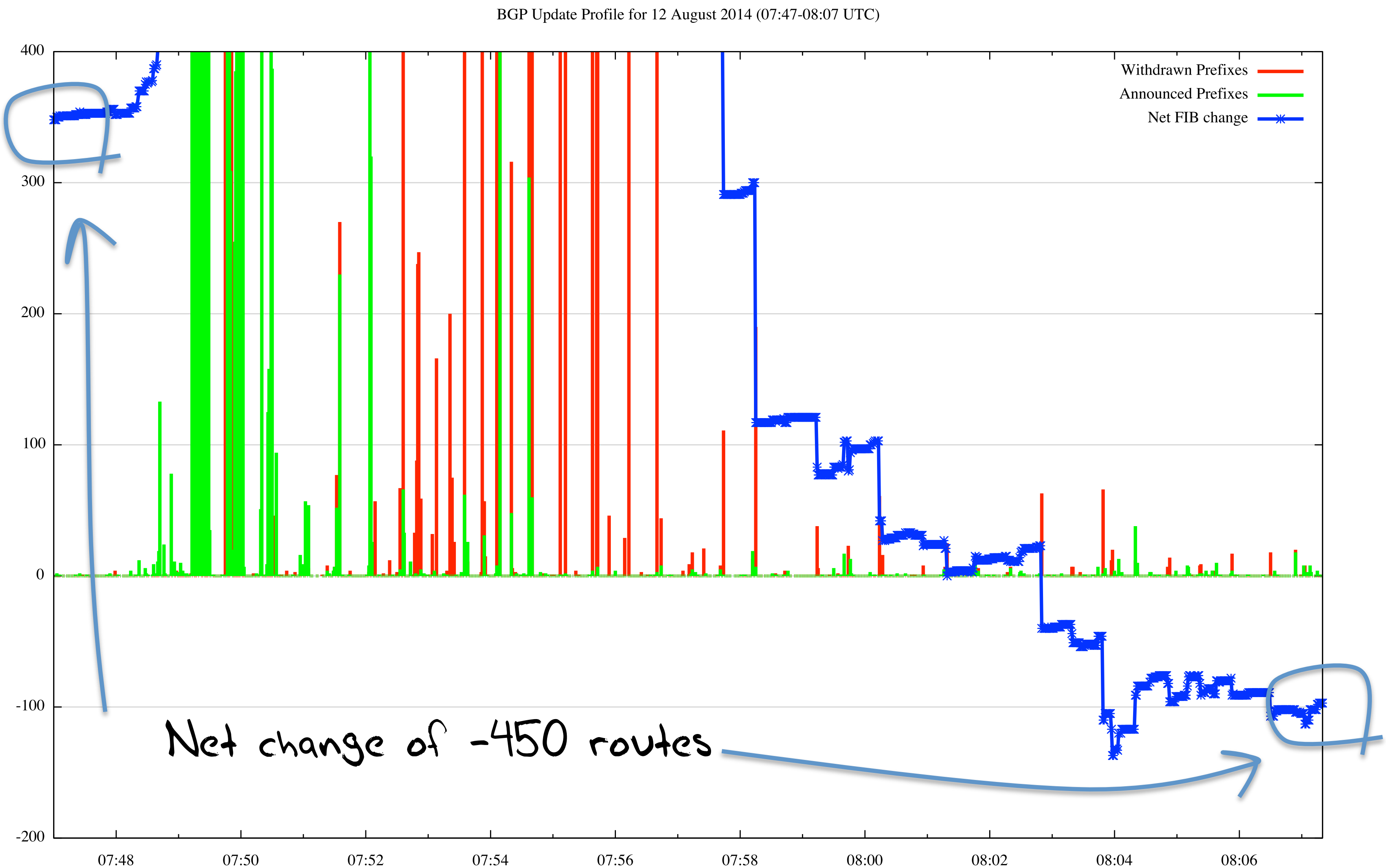
Profile	IPv4	IPv6	MAC or VPLS MAC	IPv4 VPN	IPv6 VPN	IPv4 or L2 Inbound ACL	IPv6 Inbound ACL	IPv4 or L2 Outbound ACL	IPv6 Outbound ACL
Default Profile	Logical size: 512K	Logical size: 64K	Logical size: 128K	Logical size: 128K	0	Logical size: 48K	Logical size: 4K	Logical size: 48K	Logical size: 4K
ipv4 Profile	Logical size: 1M	0	Logical size: 32K	0	0	Logical size: 112K	0	Logical size: 64K	0

What happens then?

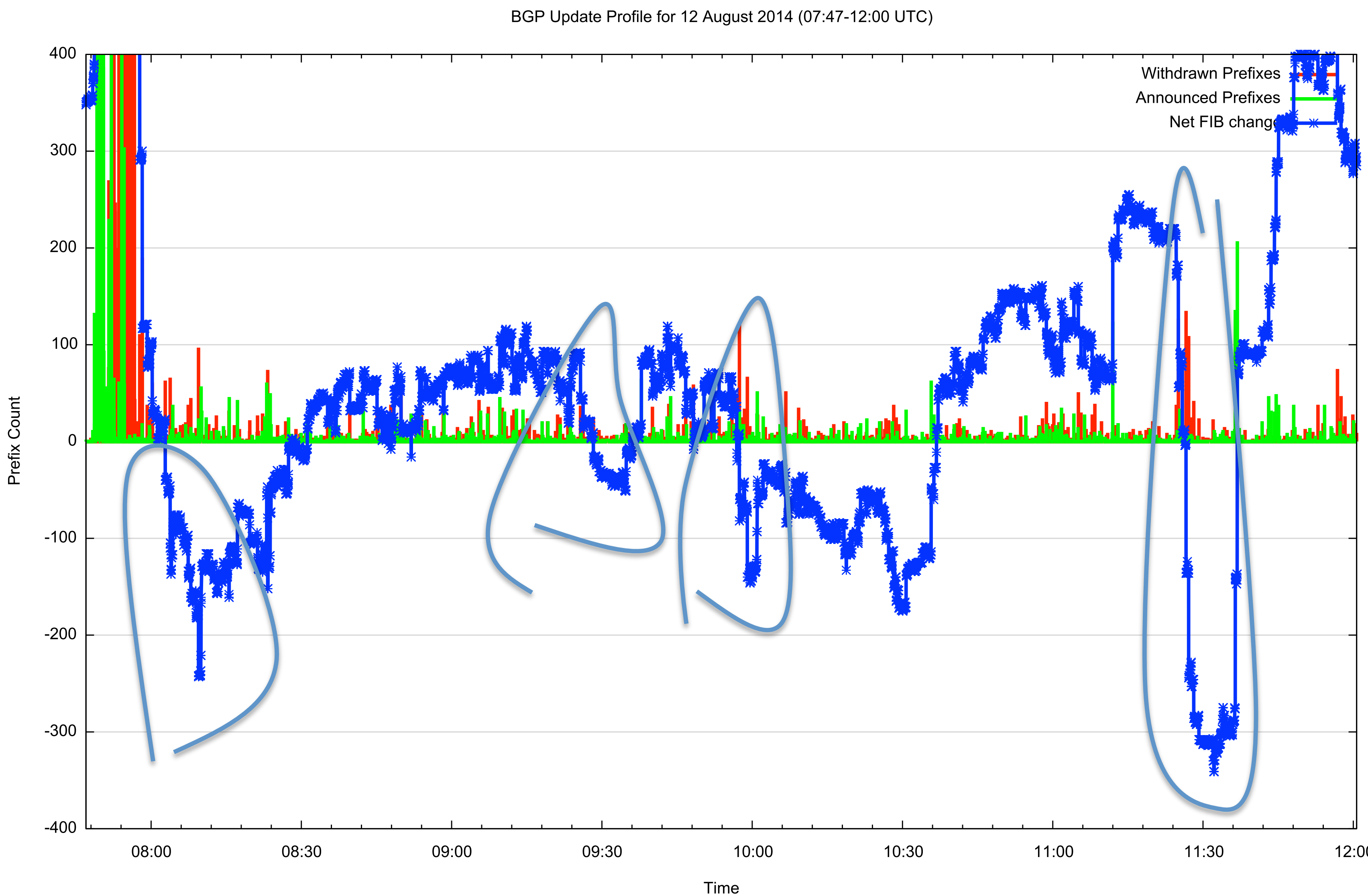
- ❑Crash and reboot?
- ❑Crash and die?
- ❑Push excess routes to slow path?
- ❑Discard excess routes

Was there any evidence of dropped routes?

Dropped Routes?



Maybe there's more…



Collateral Damage

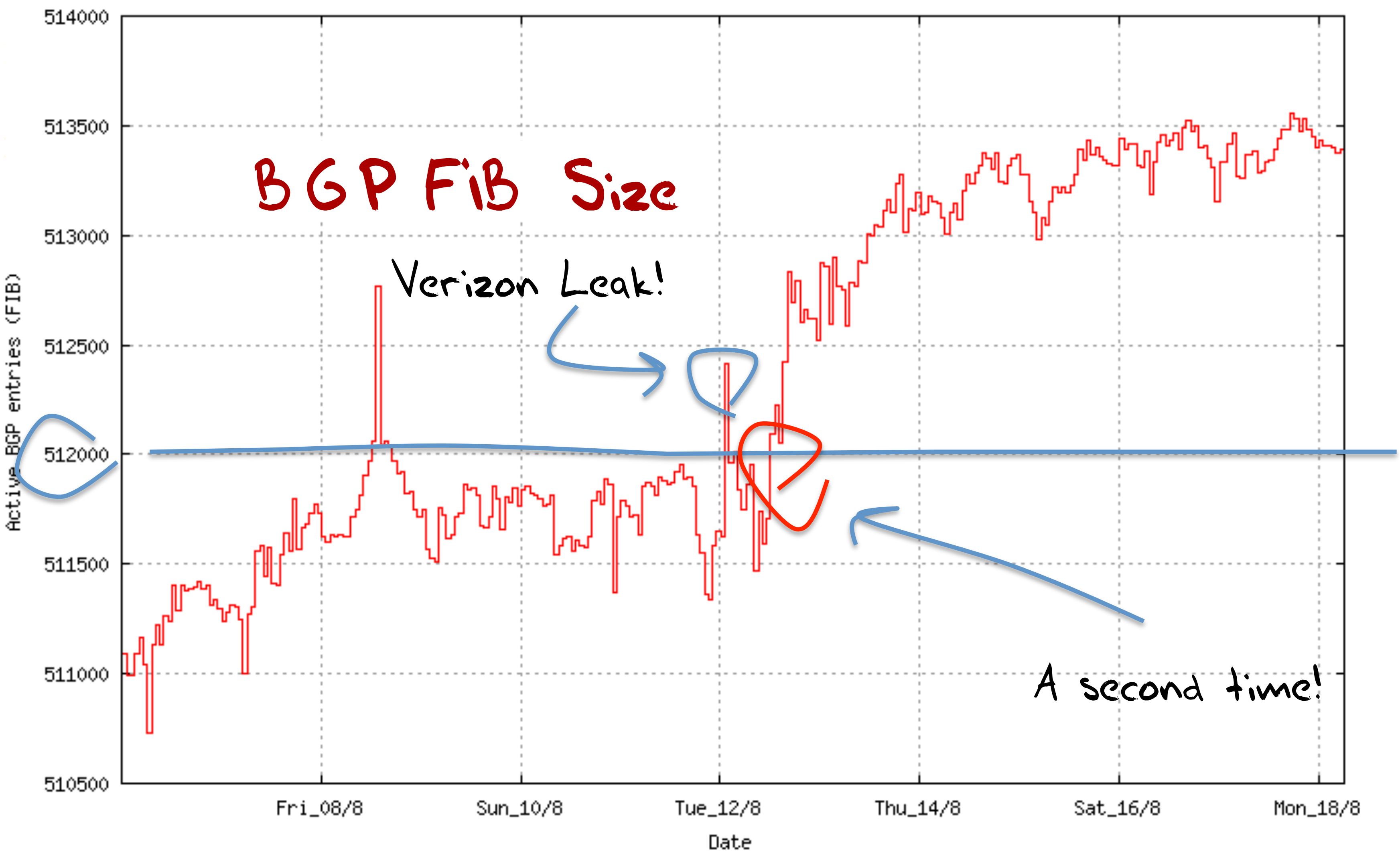
Outside of AS701, a further ~**2,000** routes were withdrawn between 07:47 and 12:00, but some of these were probably part of the route leak as they appeared to be part of the Verizon enterprise structure. But there were others who were clearly unrelated to Verizon…

Collateral Damage

763 Origin ASes were probably affected

AS	Pfxs	AS Name
9658	391	ETPI-IDS-AS-AP Eastern Telecoms Phils., Inc.,PH
6648	127	BAYAN-TELECOMMUNICATIONS Bayan Telecommunications, Inc.,PH
23498	75	CDSI - COGECODATA,CA
21332	60	NTC-AS OJSC "Vimpelcom",RU
27882	59	Telefonica Celular de Bolivia S.A.,BO
30036	56	MEDIACOM-ENTERPRISE-BUSINESS - Mediacom Communications Corp,US
131222	51	MTS-INDIA-IN 334,Udyog Vihar,IN
6459	45	TRANSBEAM - I-2000, Inc.,US
46805	42	CACHED - CachedNet LLC,US
45664	42	LBNI Liberty Broadcasting Network Inc,PH
8402	40	CORBINA-AS OJSC "Vimpelcom",RU
55465	38	TTT-AS-AP TT&T Co,TH
18025	38	ACE-1-WIFI-AS-AP Ace-1 Wifi Network,PH
22363	36	PHMGMT-AS1 - Powerhouse Management, Inc.,US
15085	33	IMMEDION - Immedion, LLC,US
50710	32	EARTHLINK-AS EarthLink Ltd. Communications&Internet Services,IQ
27229	32	WEBHOST-ASN1 - Webhosting.Net, Inc.,US
21284	32	VIVODI-AS ON S.A.,GR
23606	31	BELLTELECOM-PH-AS-PH Bell Telecommunication Philippines,PH
7018	30	ATT-INTERNET4 - AT&T Services, Inc.,US
50576	30	KRASNET-UA-AS Krasnet ltd.,UA
16058	30	Gabon-Telecom,GA
13188	30	BANKINFORM-AS TOV "Bank-Inform",UA

But then it happened again!



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!

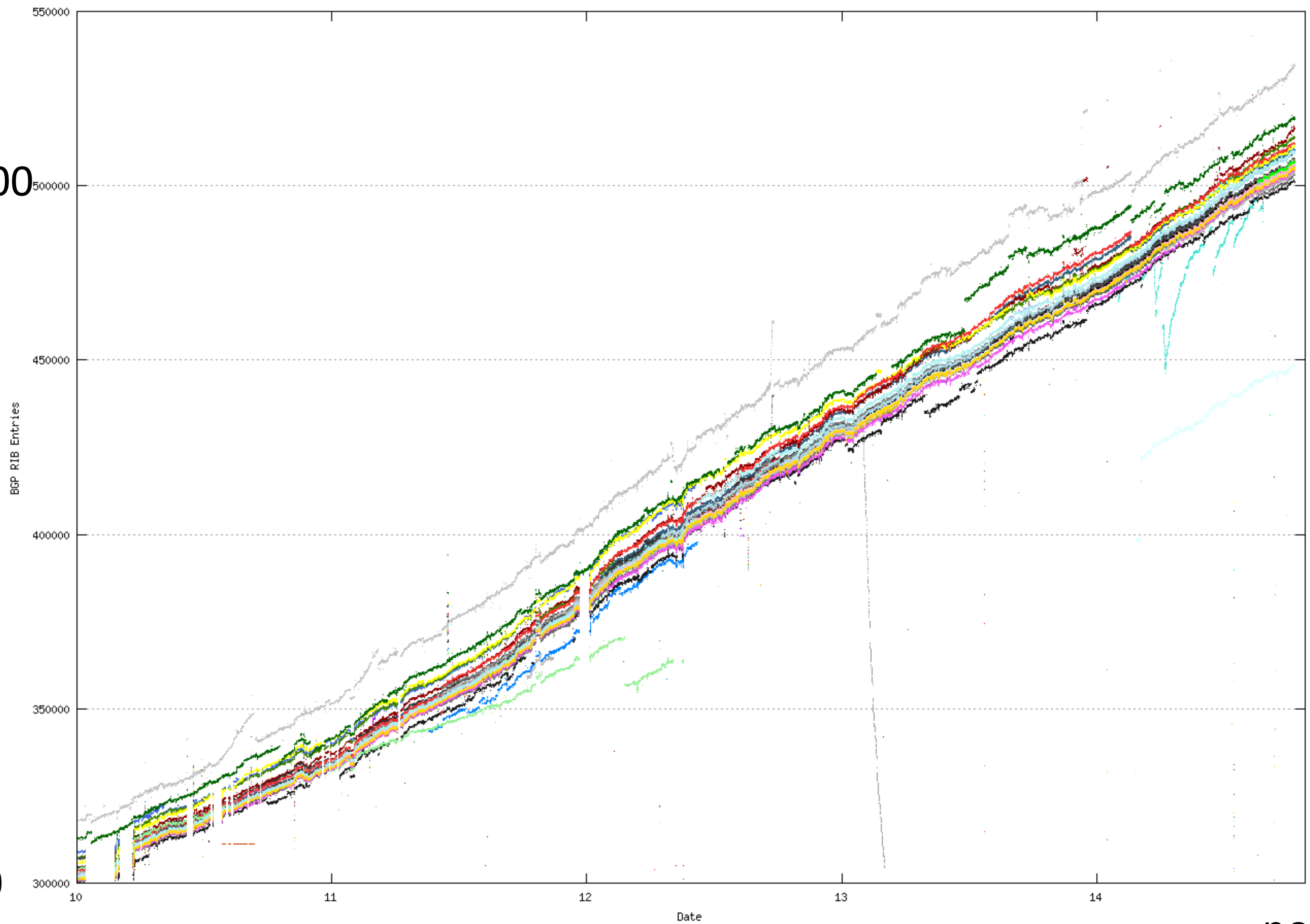
There is no single objective "out of the system" view of the Internet's Routing environment.

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

500,000

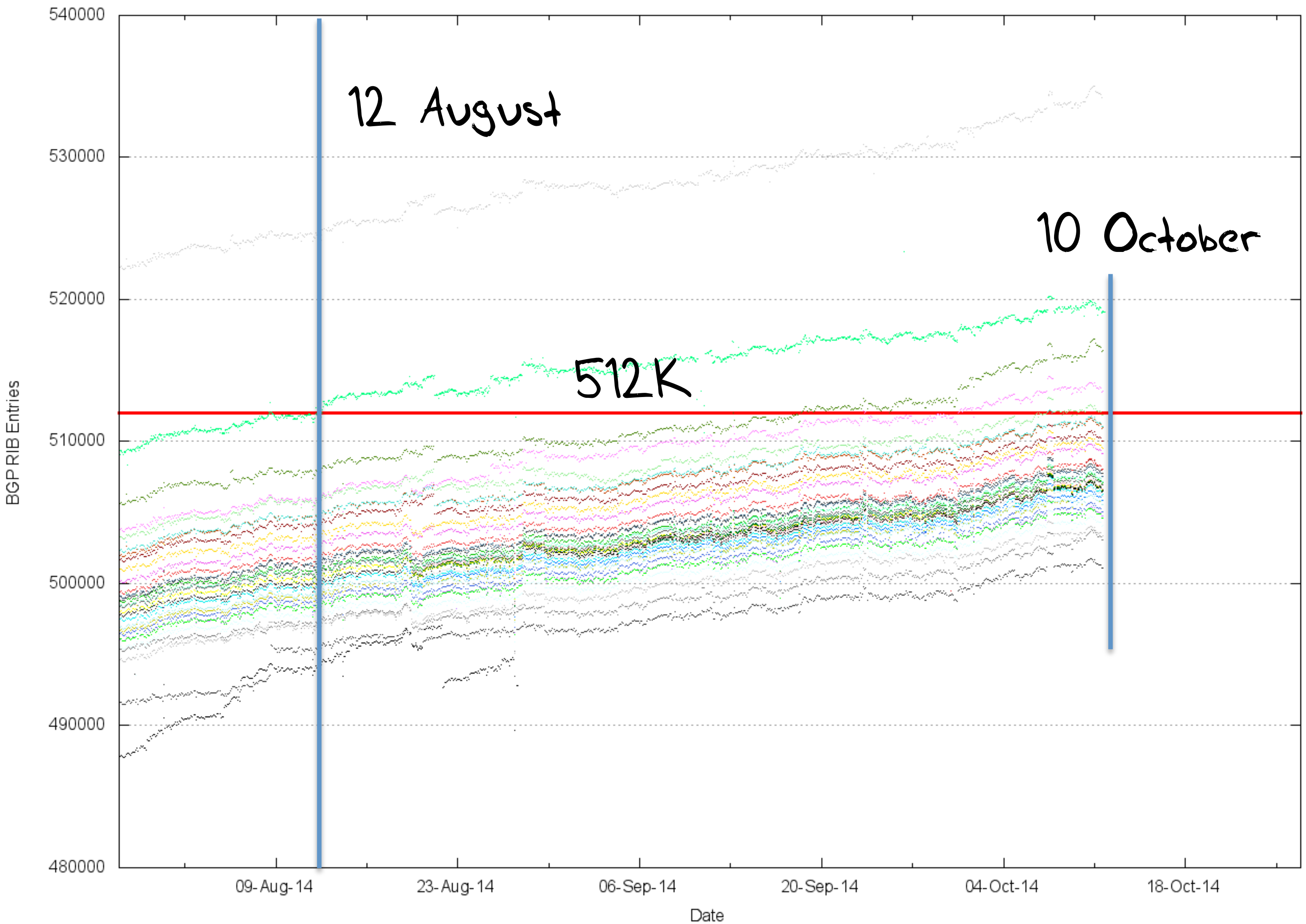


300,000

2010

now

Zooming in



12 August
10 October
512K
540000
530000
520000
510000
500000
490000
480000
BGP RIB Entries
09-Aug-14
23-Aug-14
06-Sep-14
20-Sep-14
04-Oct-14
18-Oct-14
Date

For most networks…

(probably including yours)

Its 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 networks…

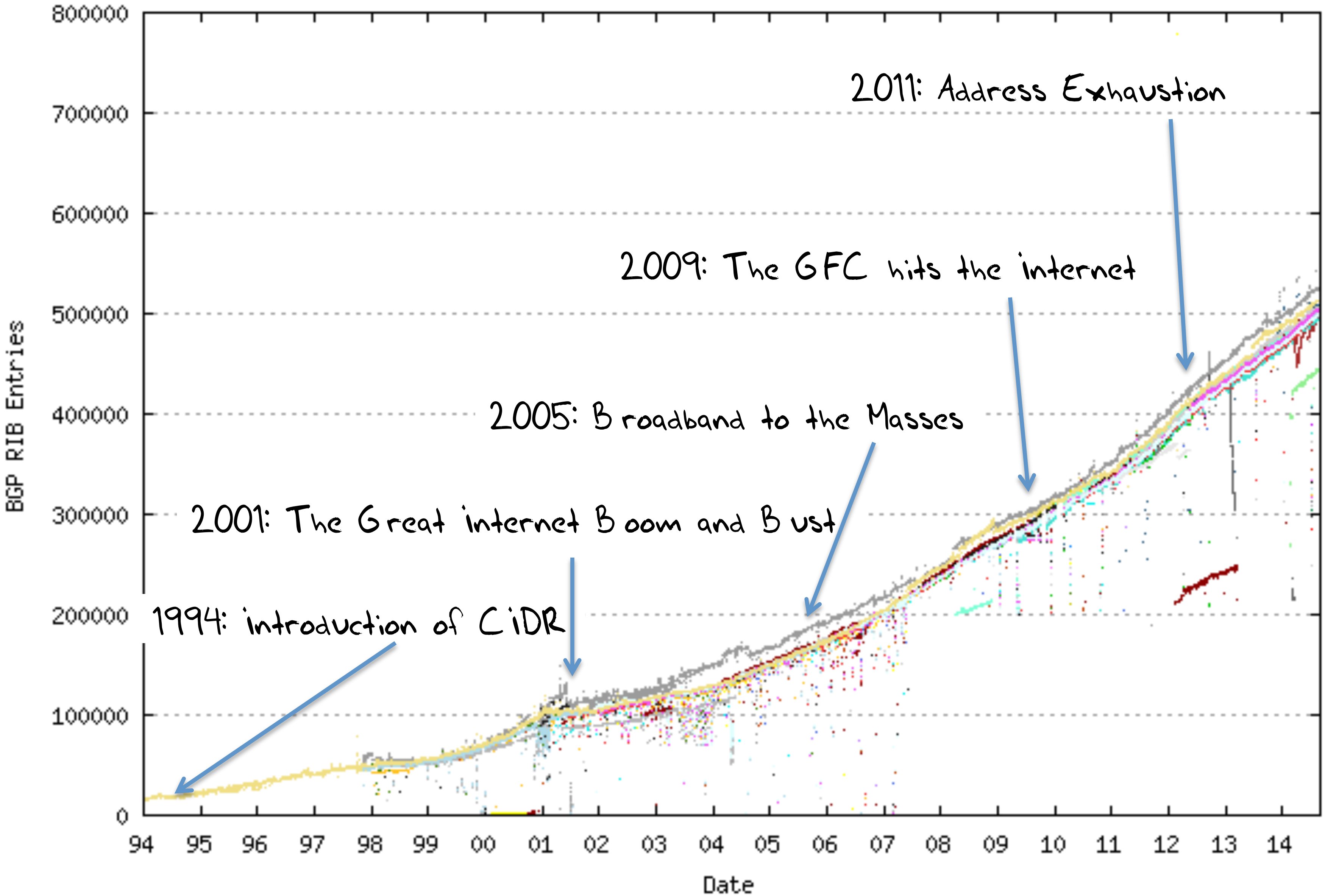
For most affected networks, it was the AS701 route leak that tipped them over the edge on the day

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

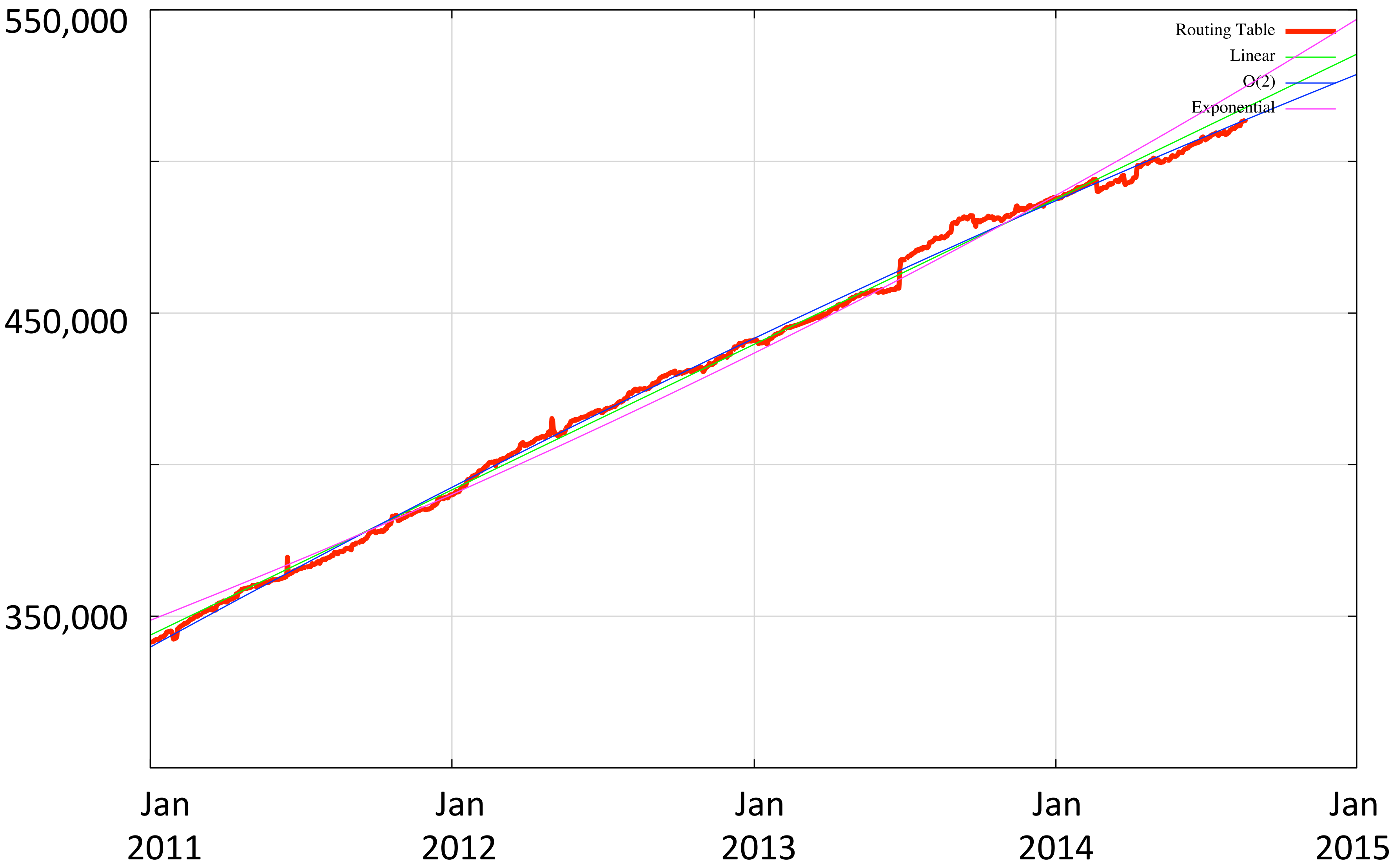
When? And what's next?

How quickly is the routing table growing?

20 years of routing the Internet



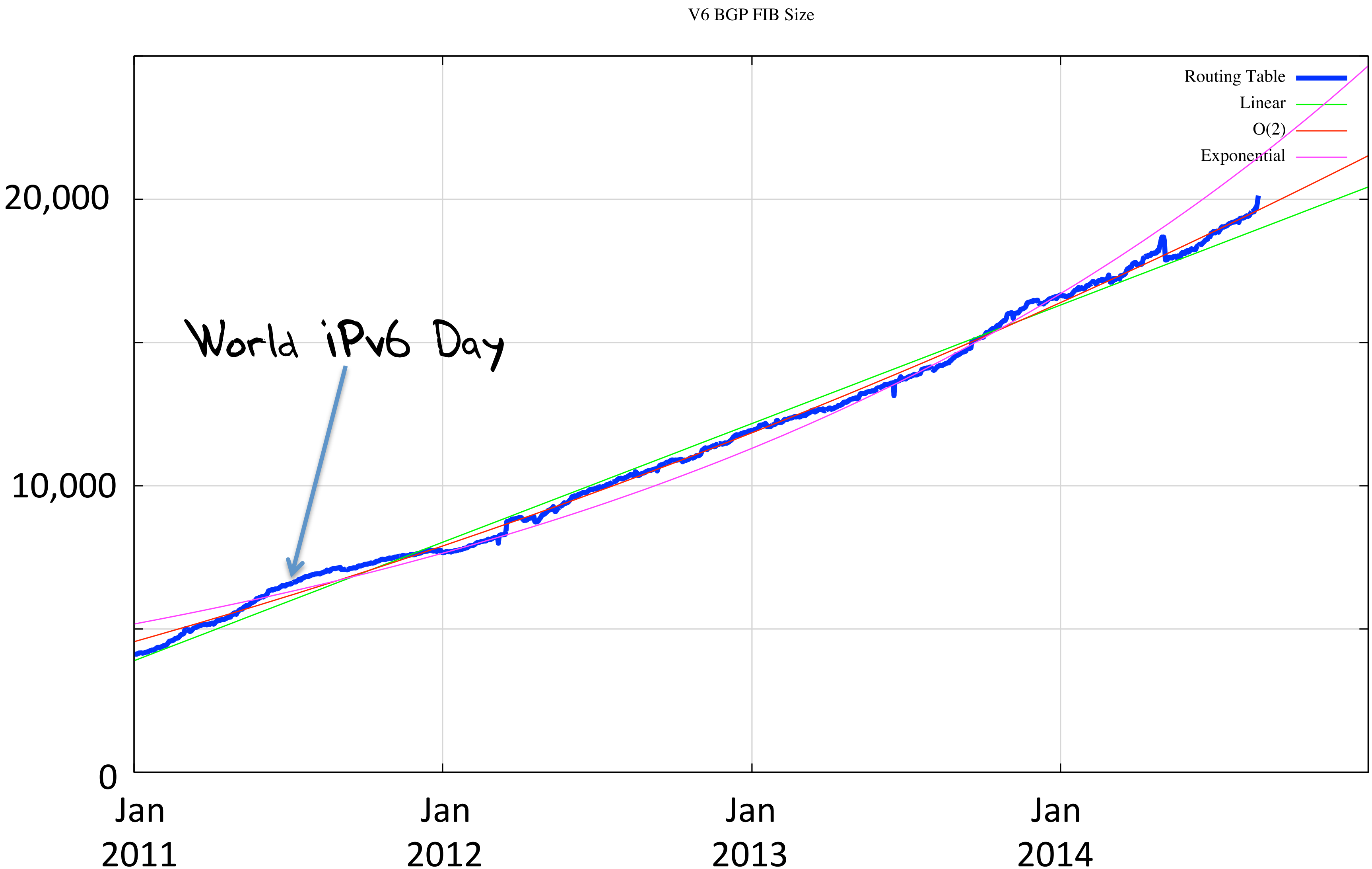
IPv4 BGP Prefix Count 2011 - 2014



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



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

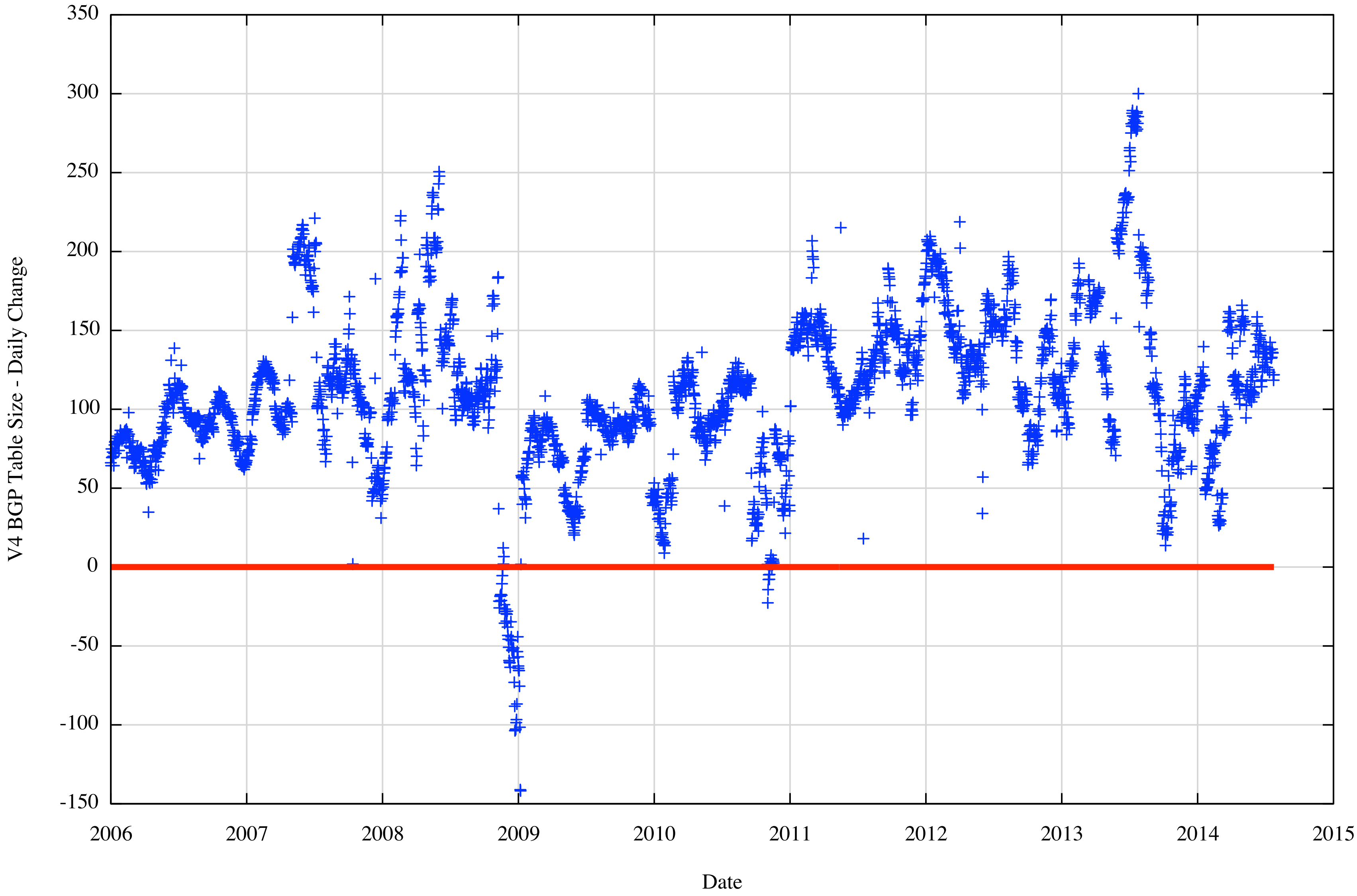
BGP Size Projections

- 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

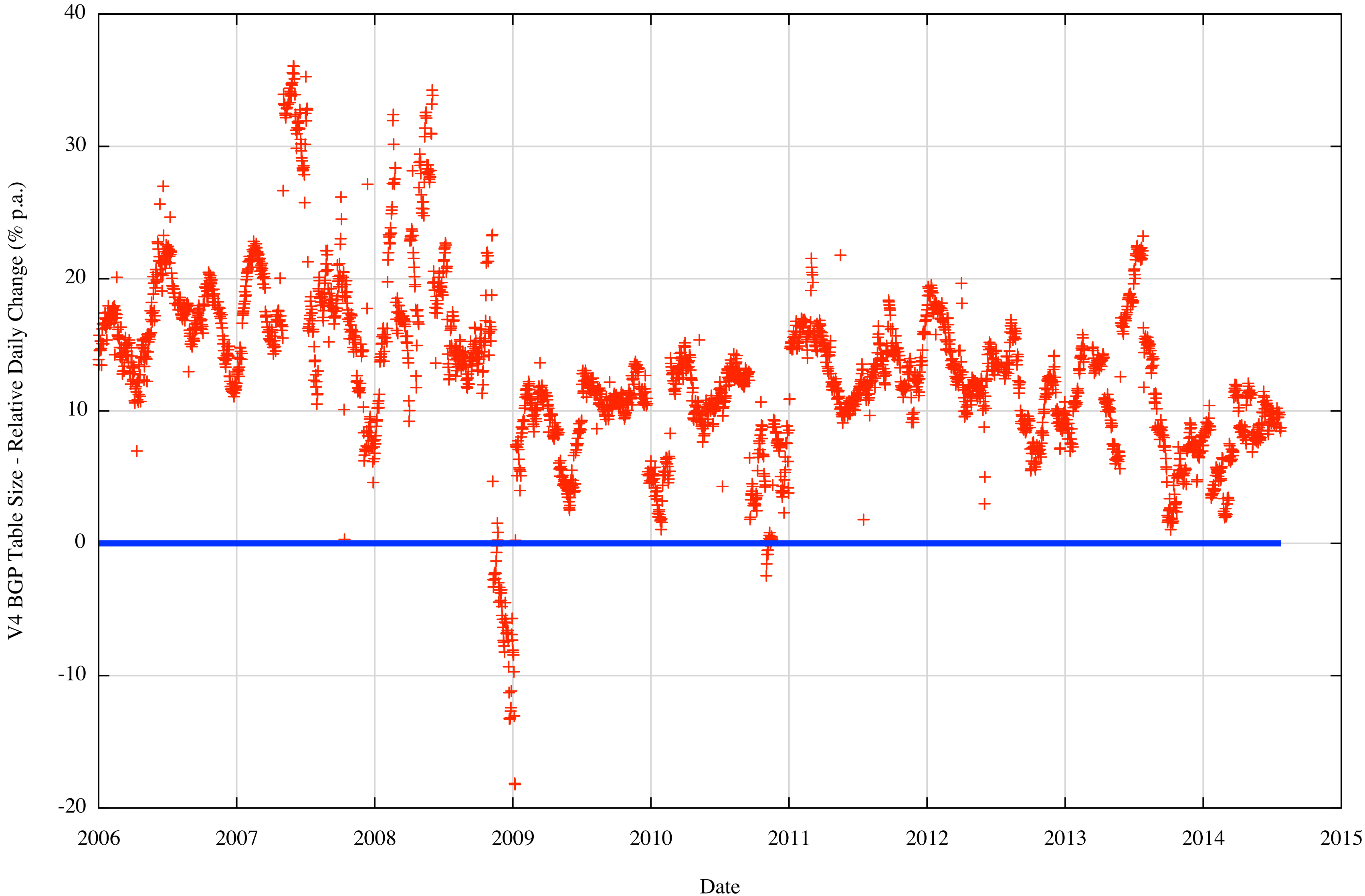
  which makes this projection even more

  speculative than normal!

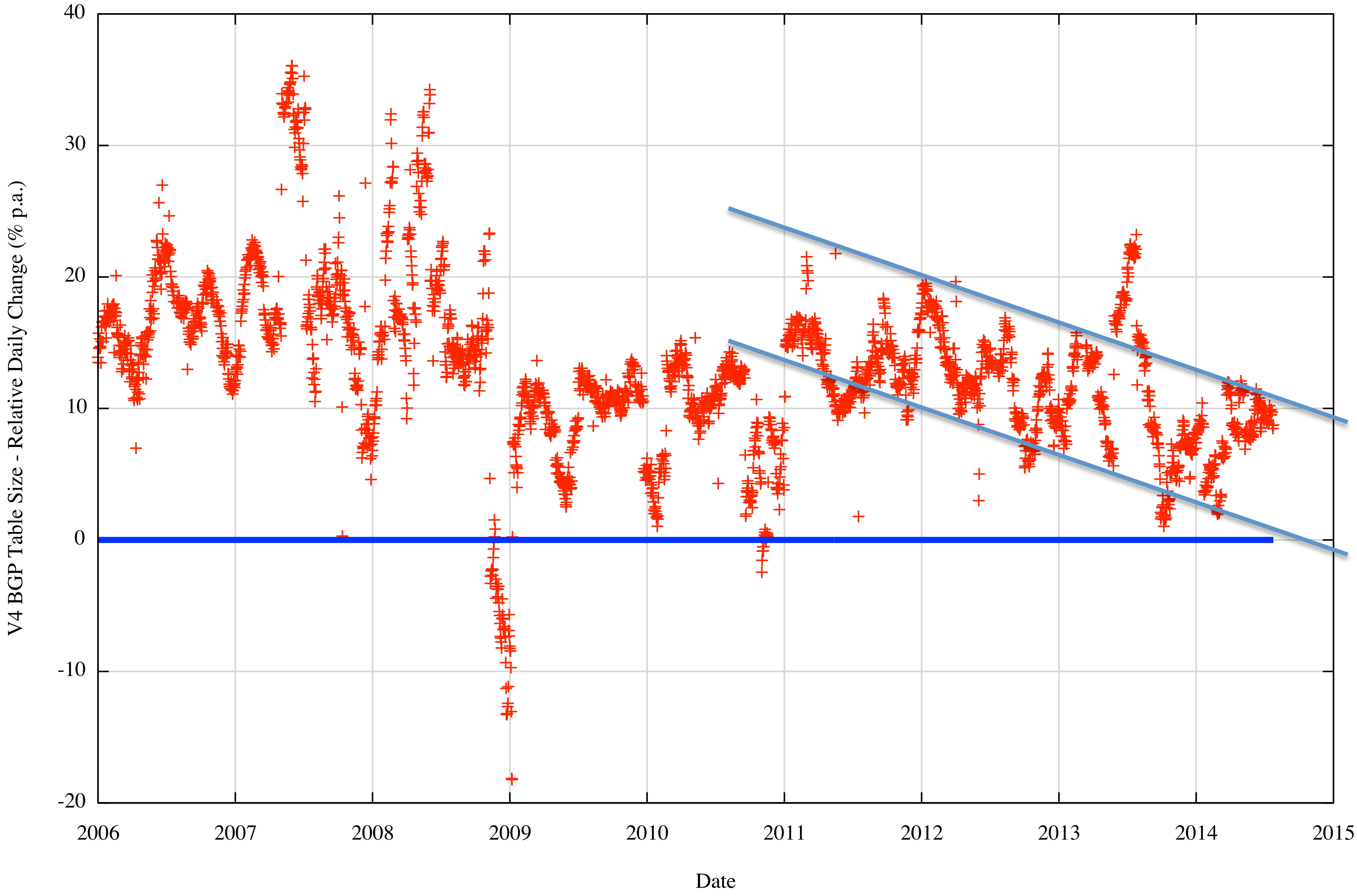
V4 - Daily Growth Rates



V4 - Relative Daily Growth Rates



V4 - Relative Daily Growth Rates

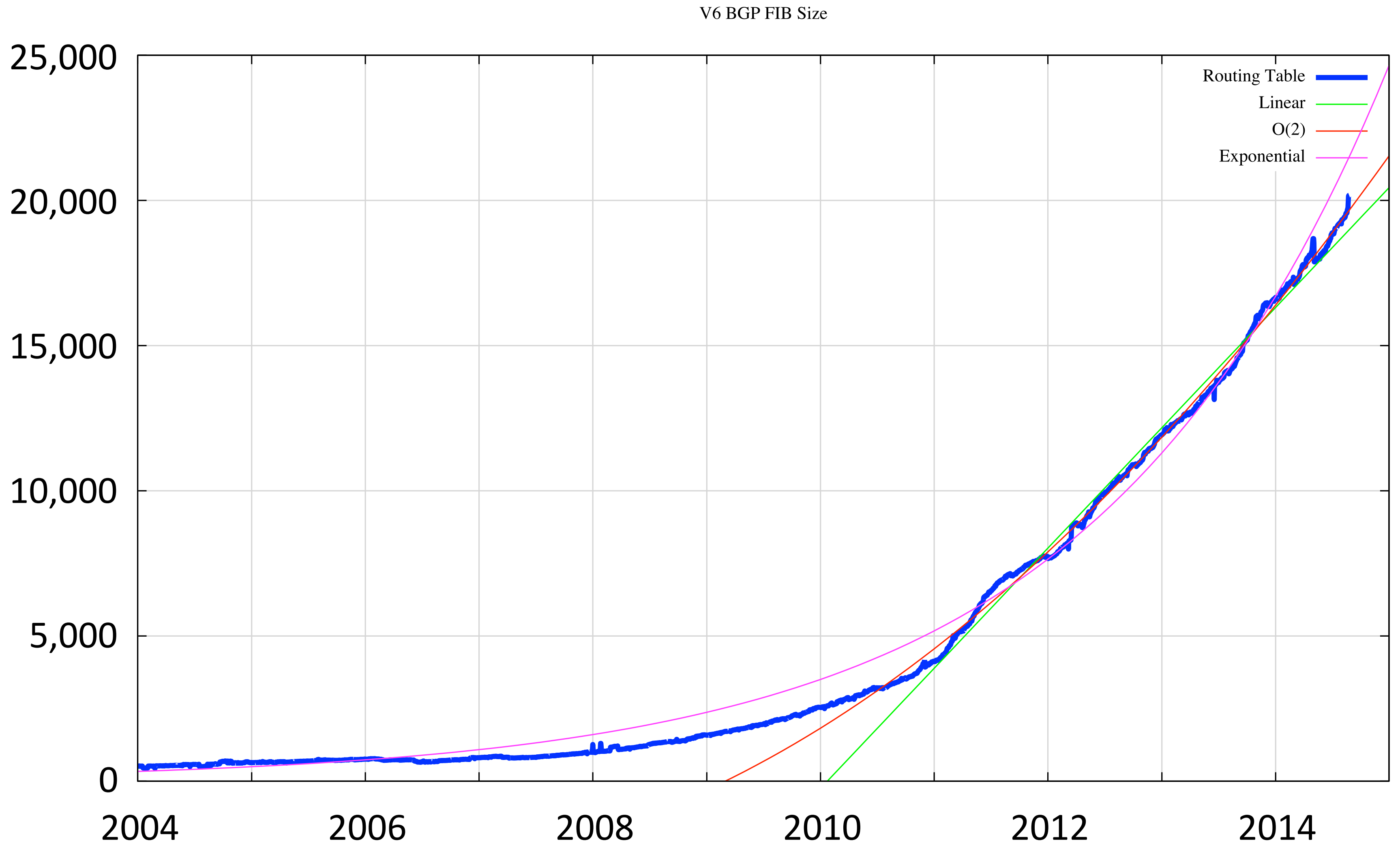


IPv4 BGP Table Size predictions

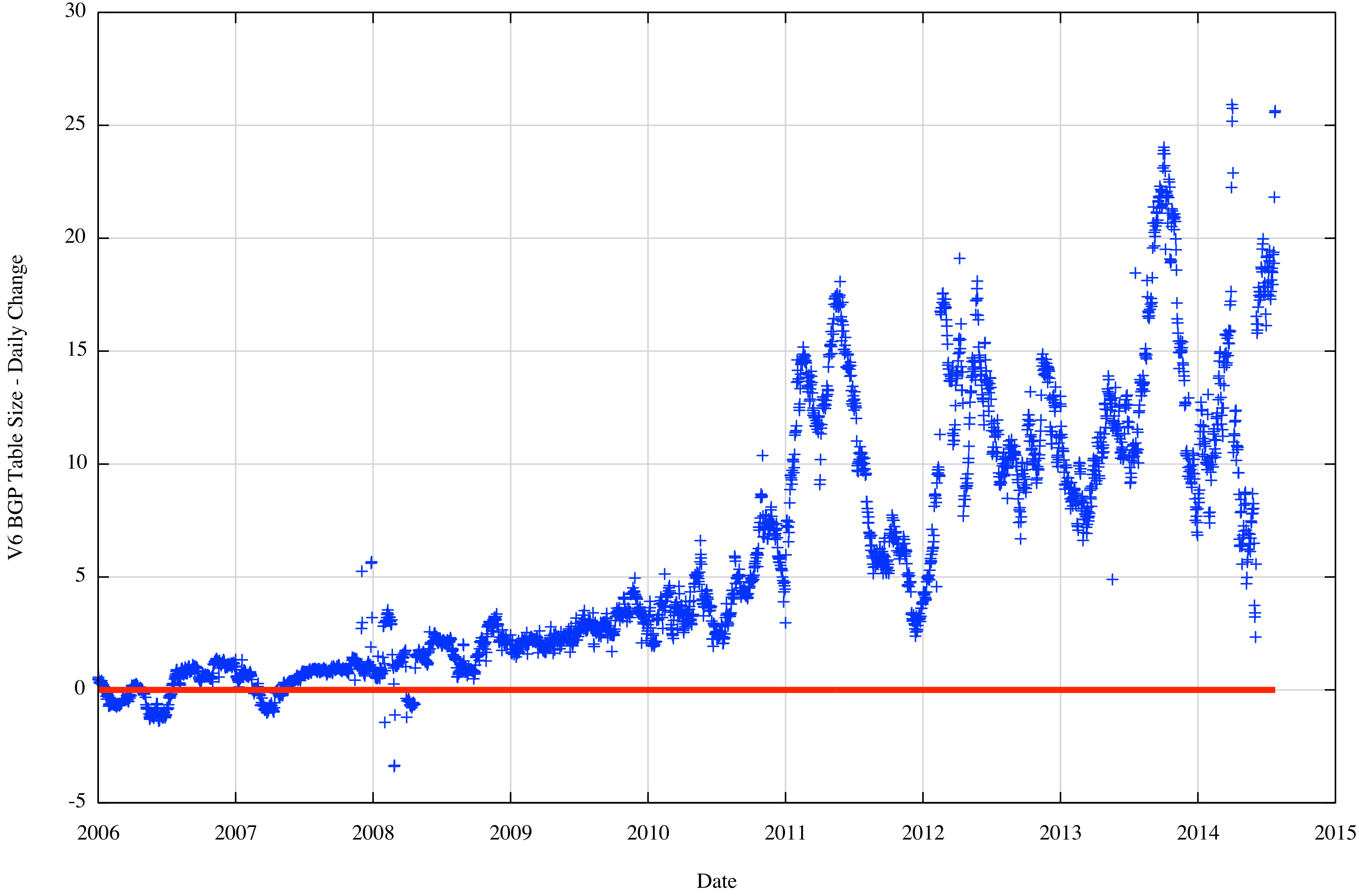
	Linear Model	Exponential Model
Jan 2013	441,172 entries	
2014	488,011 entries	
2015	*540,000 entries*	*559,000*
2016	*590,000 entries*	*630,000*
2017	*640,000 entries*	*710,000*
2018	*690,000 entries*	*801,000*
2019	*740,000 entries*	*902,000*

These numbers are dubious due to uncertainties introduced by IPv4 address exhaustion pressures.

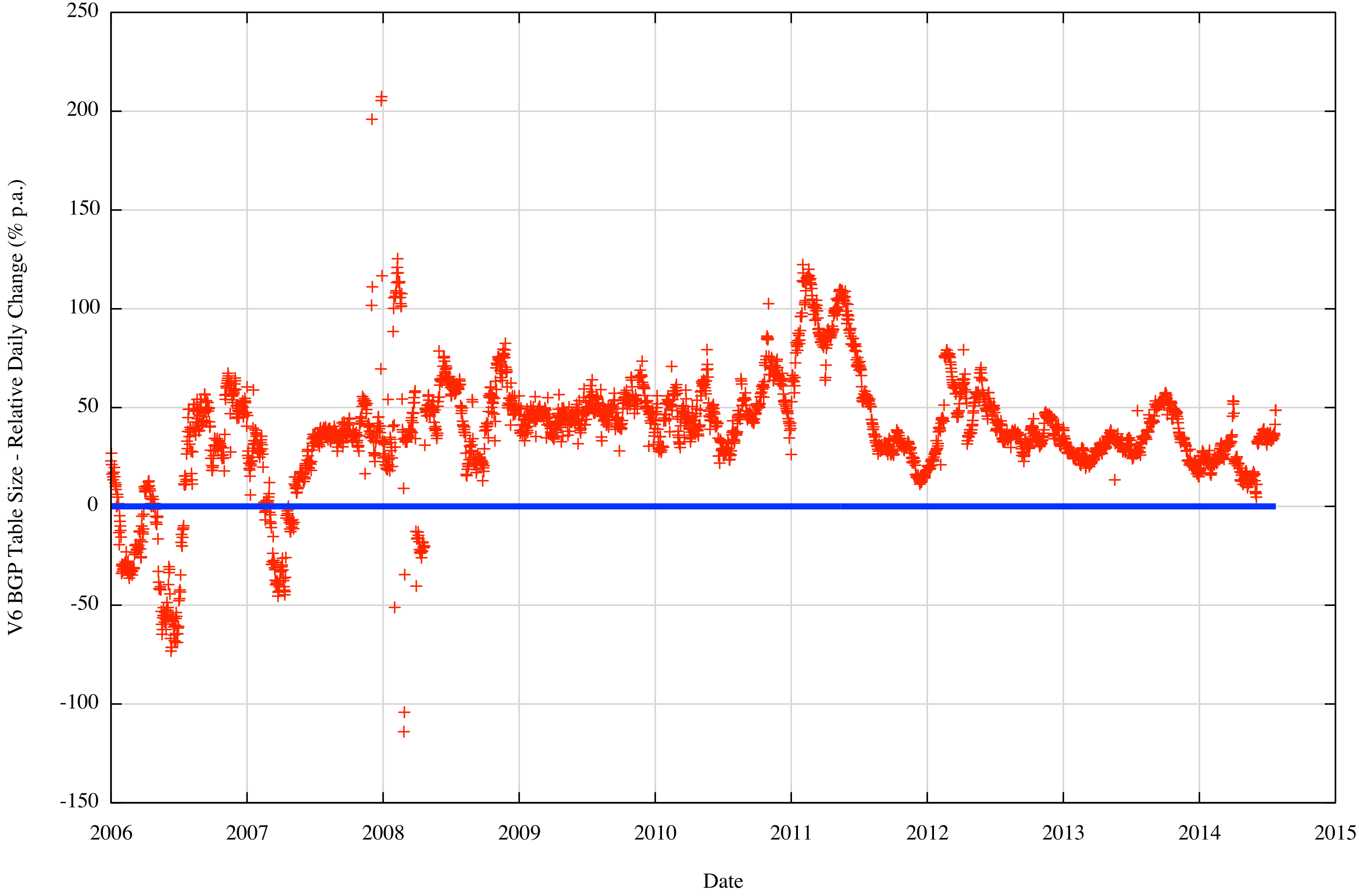
IPv6 Table Size



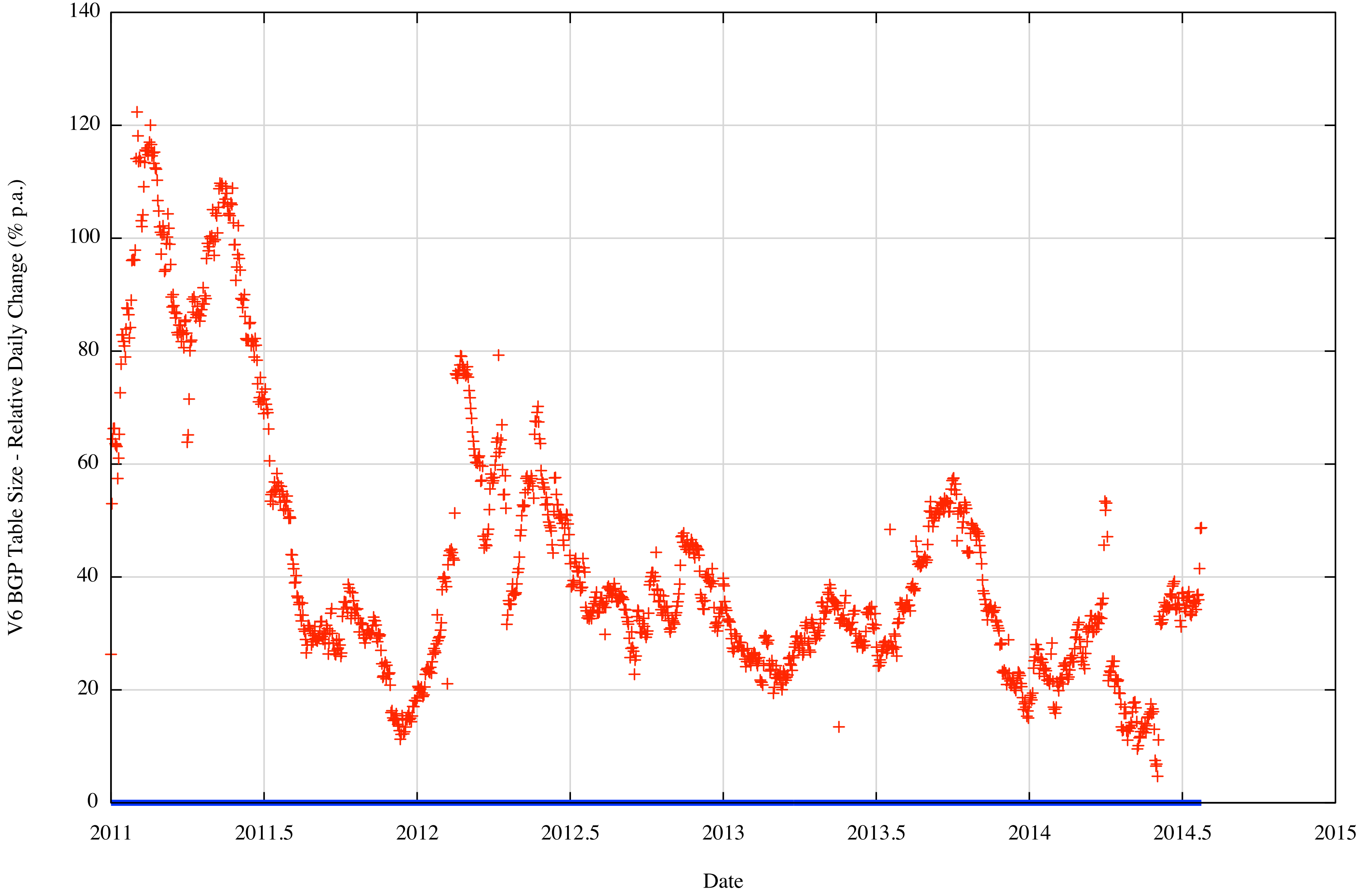
V6 - Daily Growth Rates



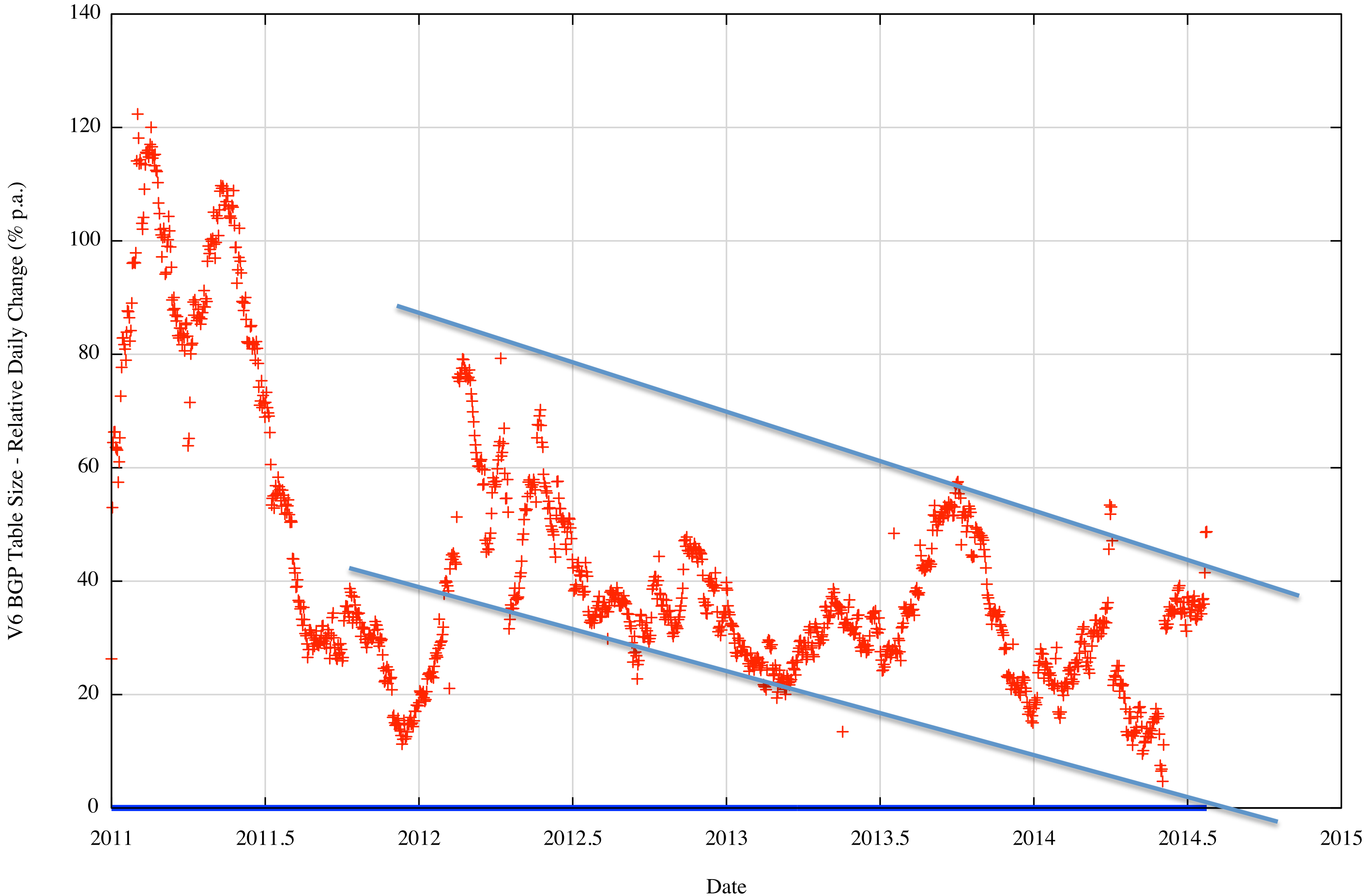
V6 - Relative Growth Rates



V6 - Relative Growth Rates



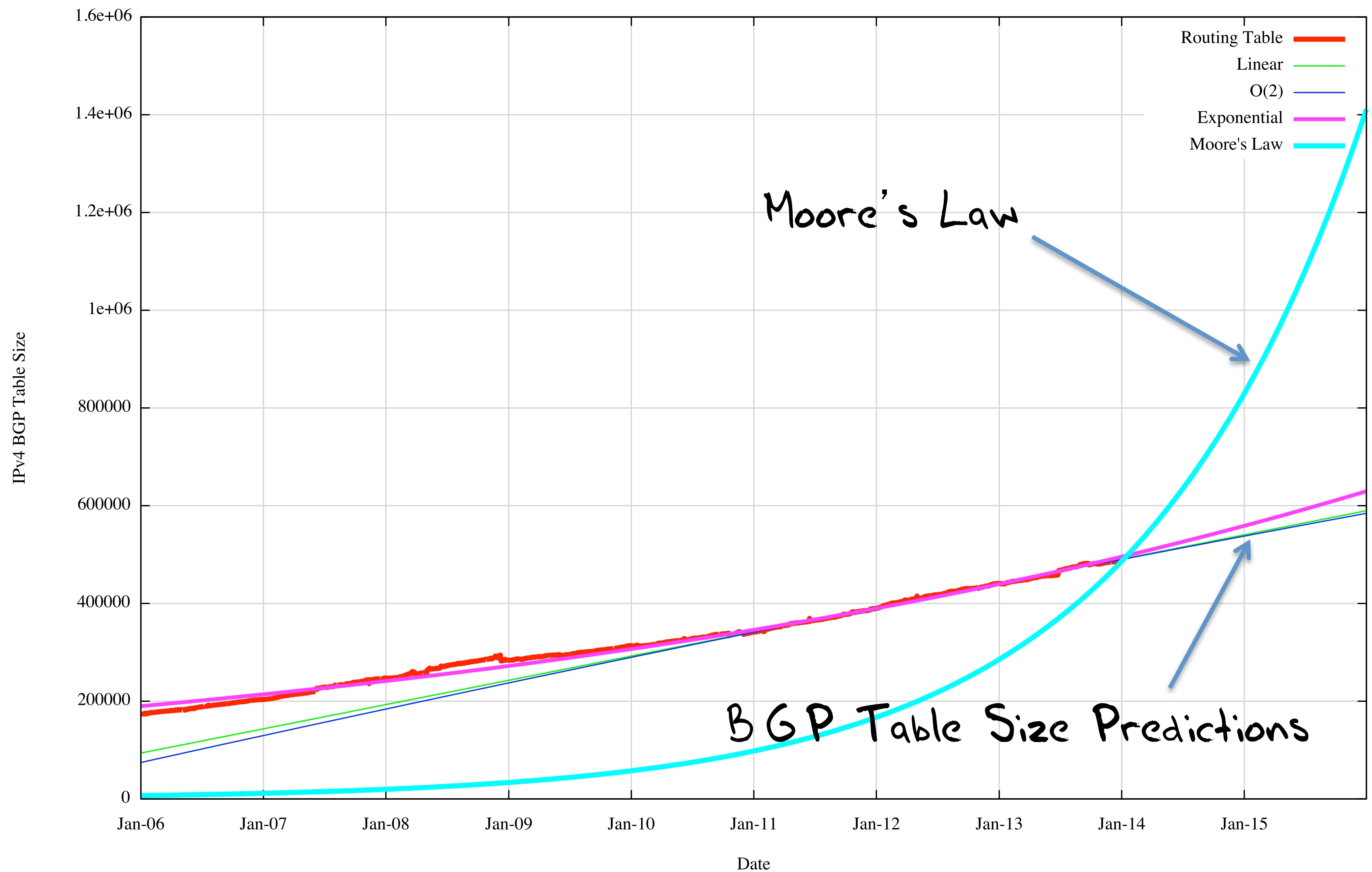
V6 - Relative Growth Rates



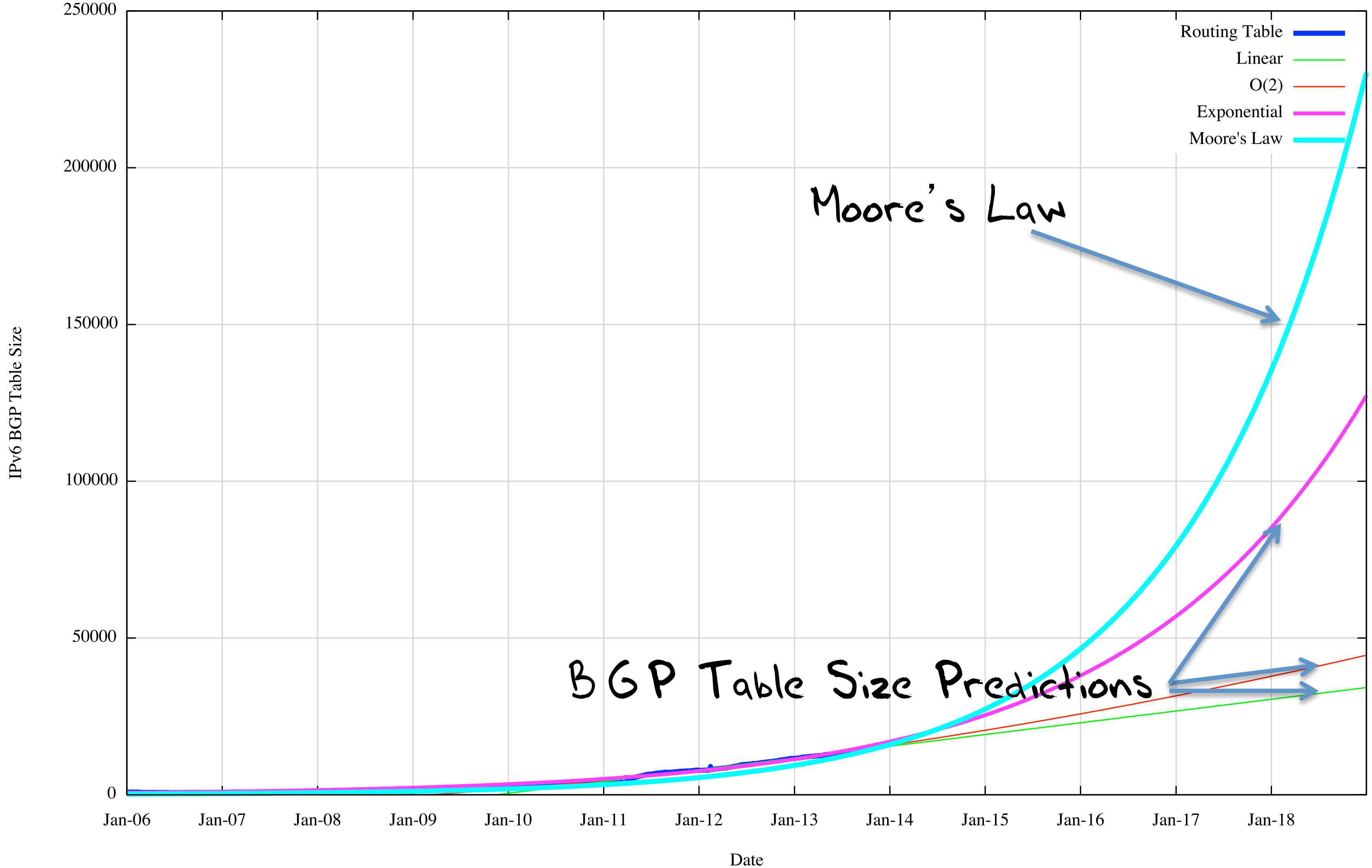
IPv6 BGP Table Size predictions

	Exponential Model	LinearModel
Jan 2013	11,600 entries	
2014	16,200 entries	
2015	*24,600 entries*	*19,000*
2016	*36,400 entries*	*23,000*
2017	*54,000 entries*	*27,000*
2018	*80,000 entries*	*30,000*
2019	*119,000 entries*	*35,000*

IPv4 BGP Table size and Moore's Law



IPv6 Projections and Moore's Law



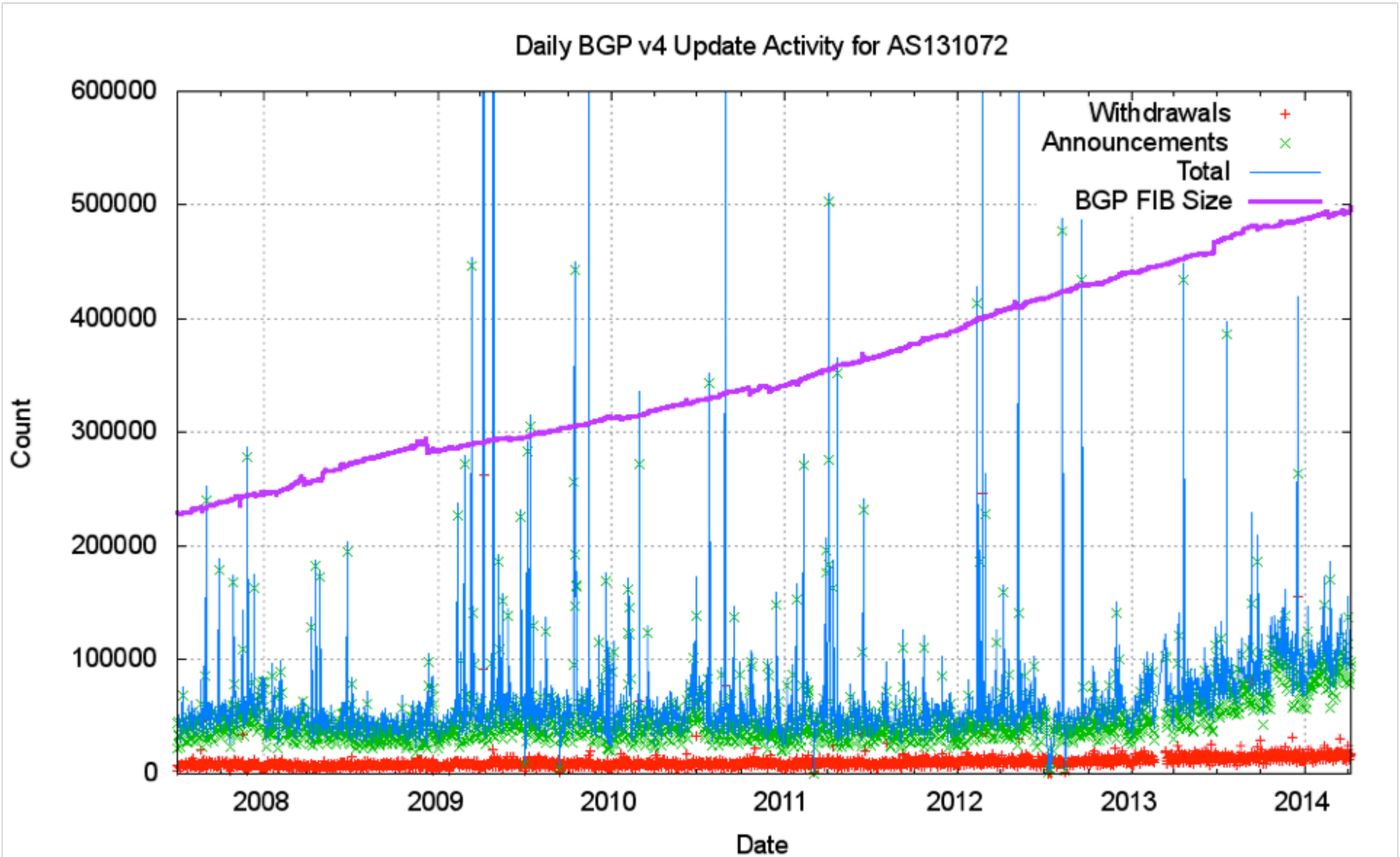
BGP Table Growth

- 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

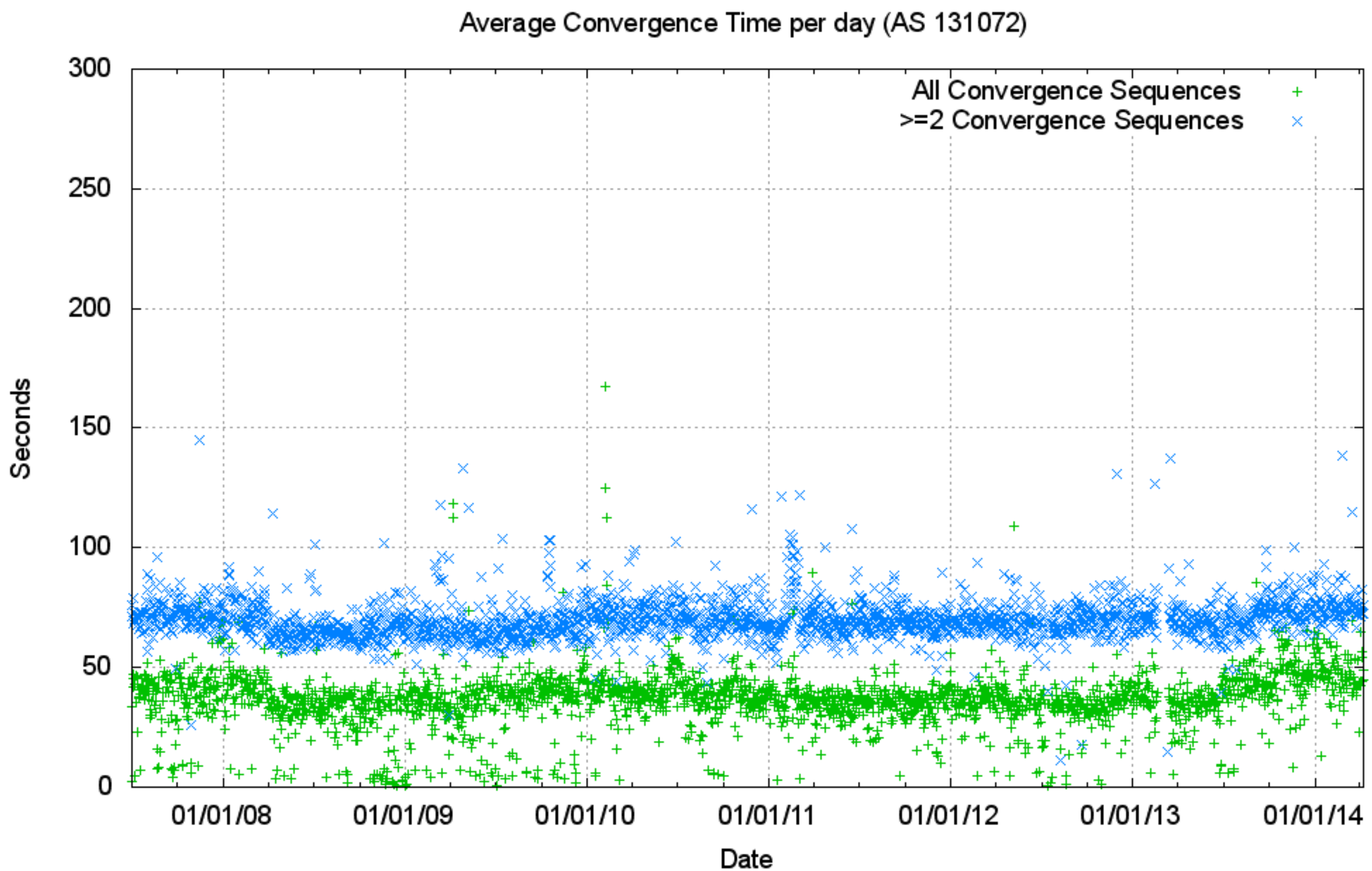
BGP Updates

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

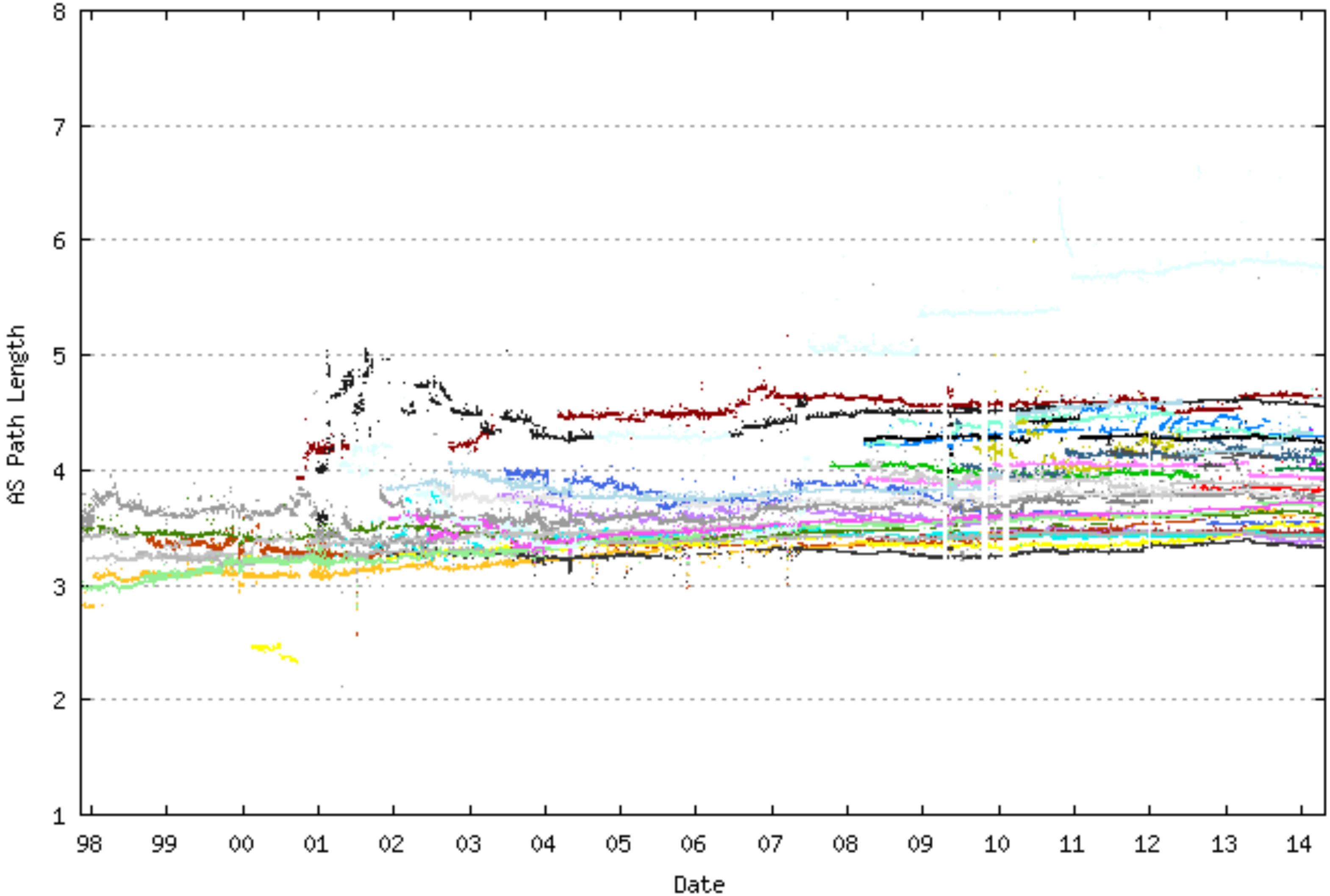
Announcements and Withdrawals



Convergence Performance



IPv4 Average AS Path Length



Data from Route Views

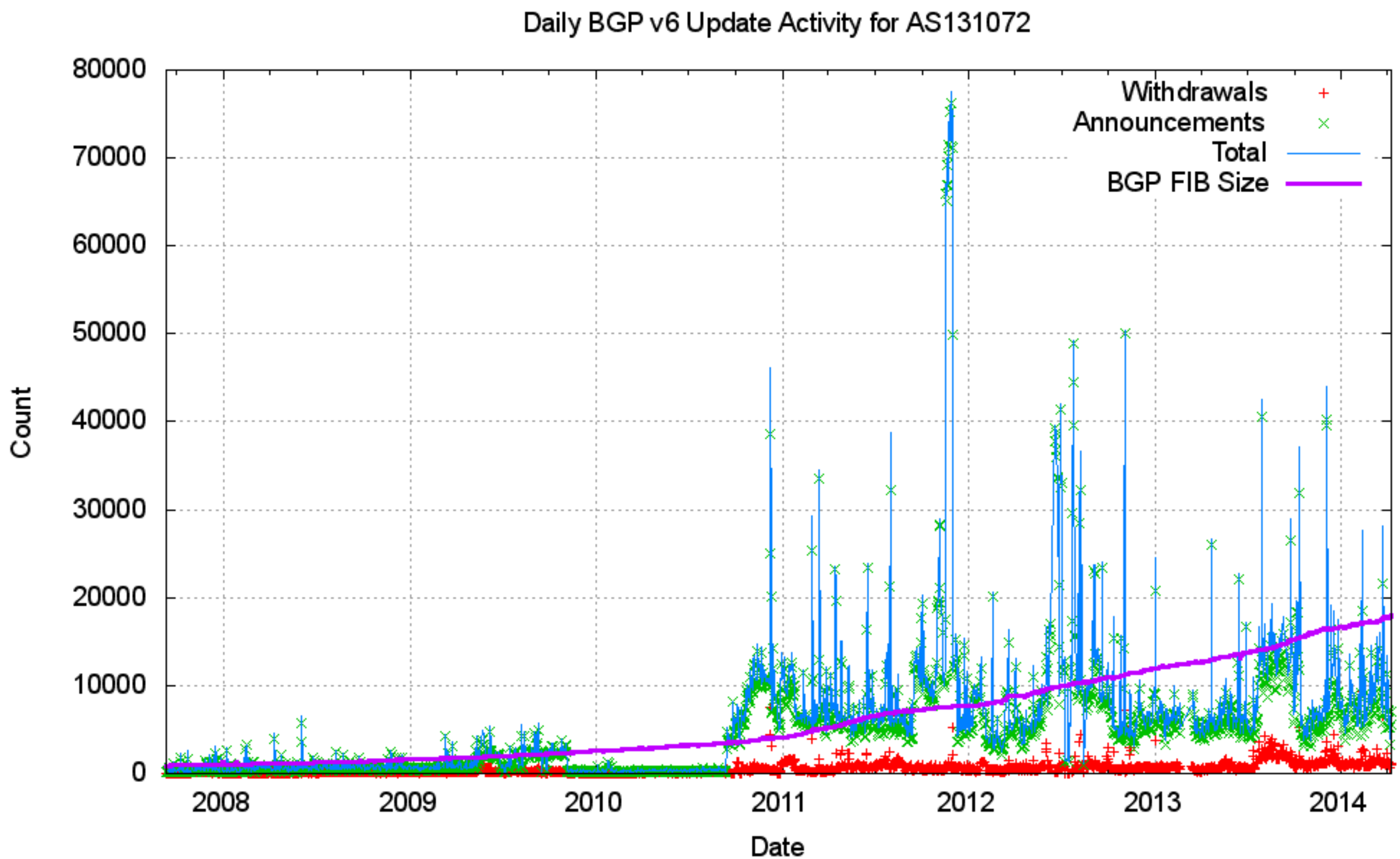
Updates in IPv4 BGP

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

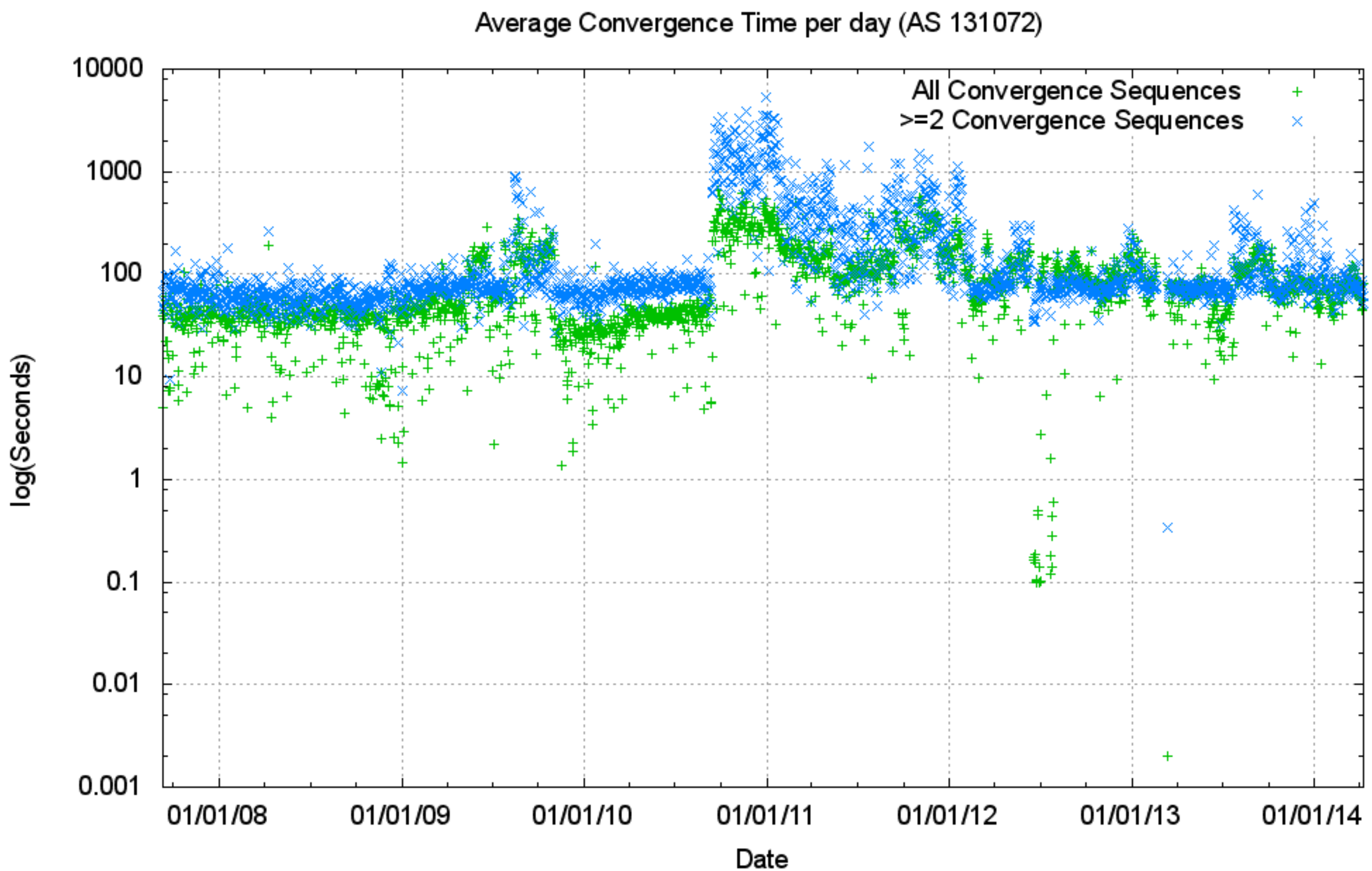
- The number of updates per instability event has been constant, due to the damping effect of the MRAI interval, and the relatively constant AS Path length over this interval

What about IPv6?

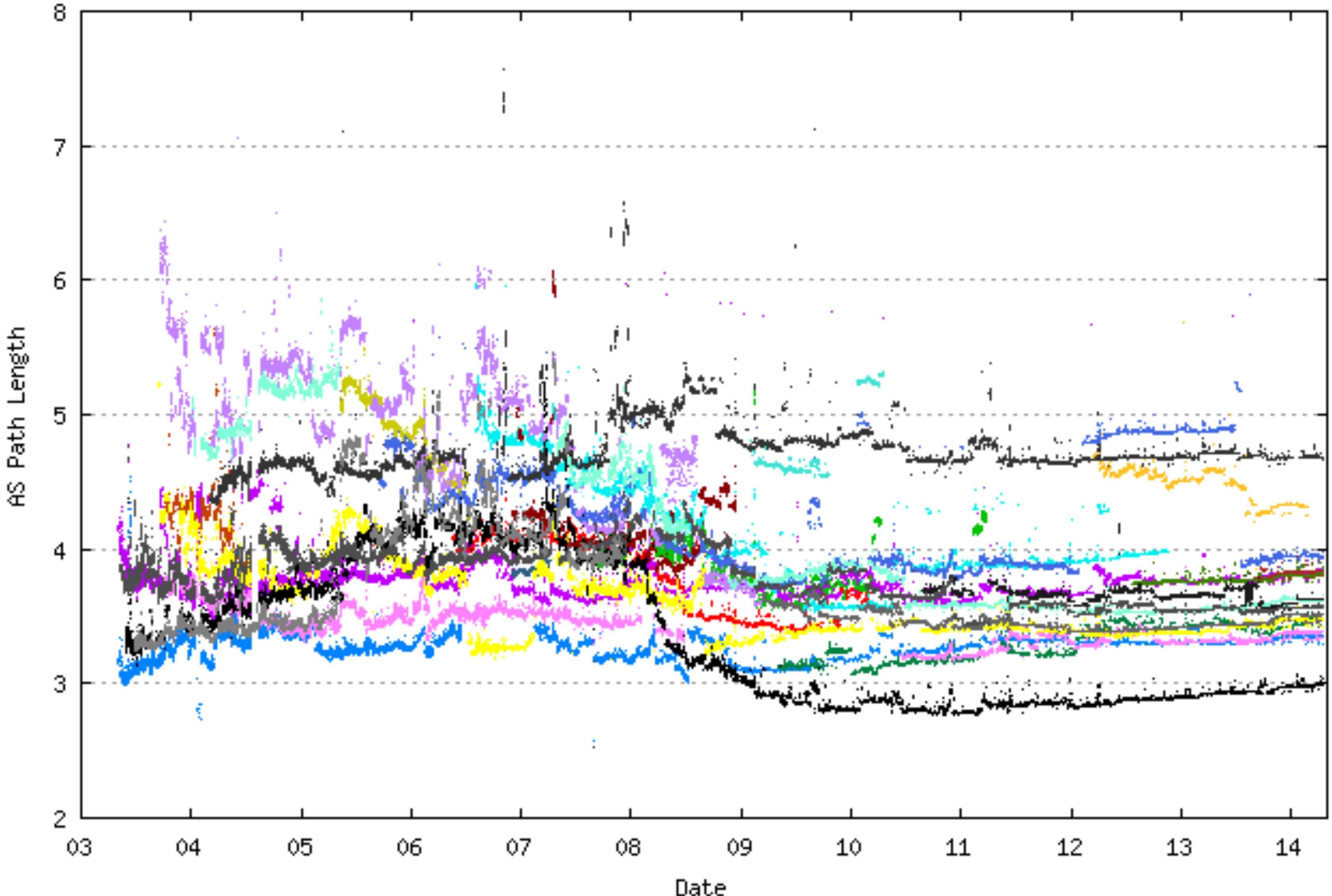
V6 Announcements and Withdrawals



V6 Convergence Performance



V6 Average AS Path Length



Data from Route Views

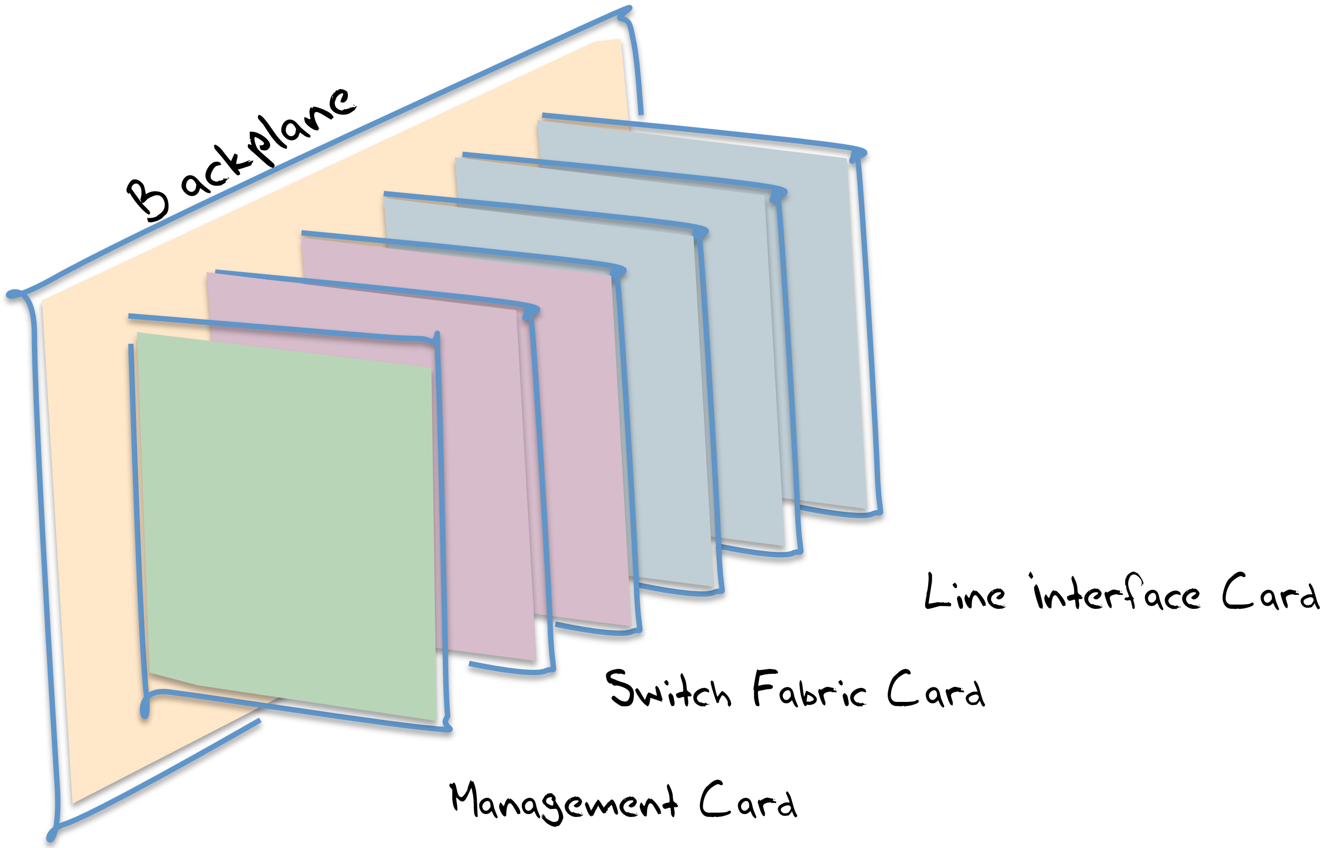
Problem? Not a Problem?

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.

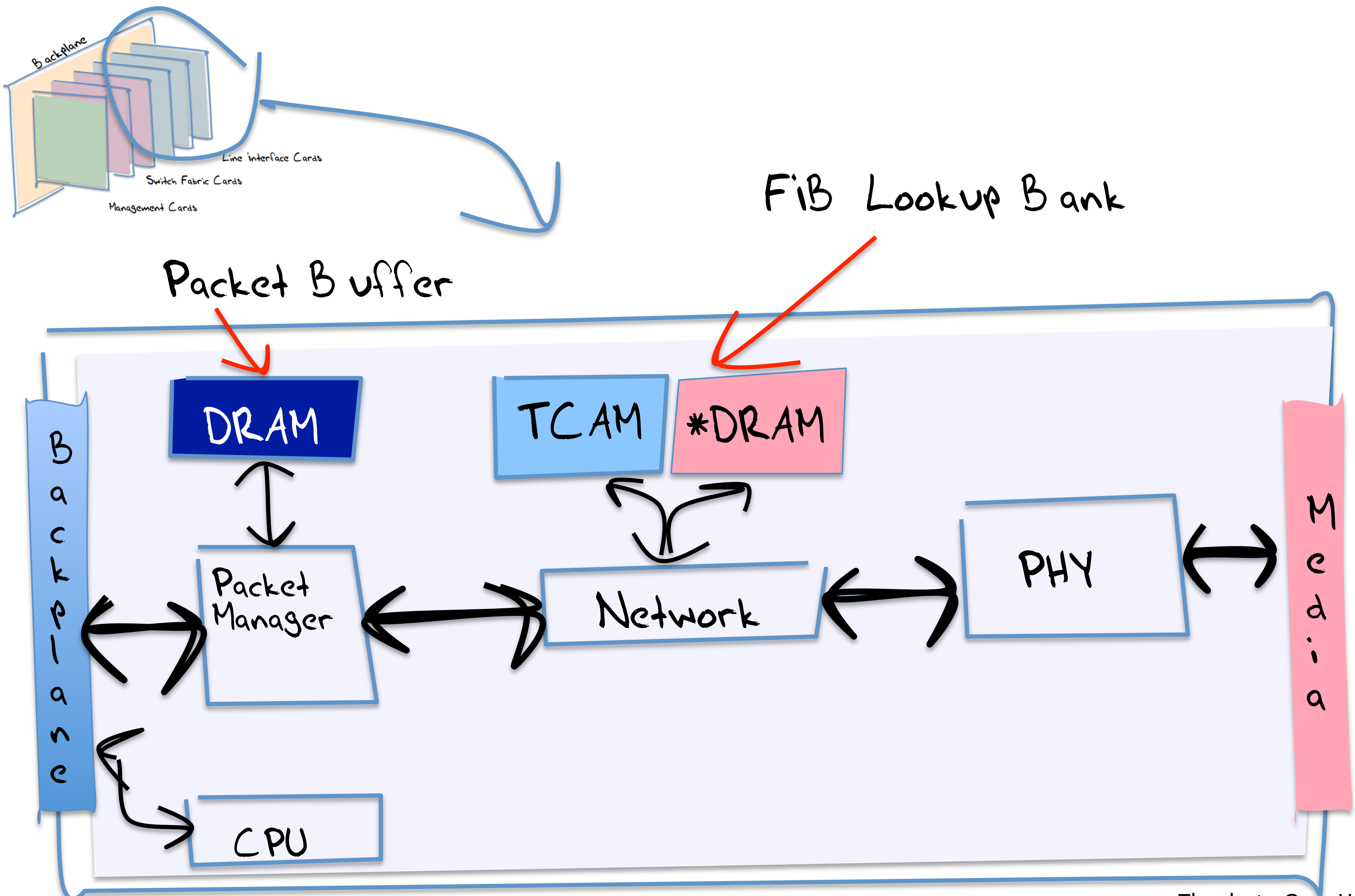
So lets take a quick look at routers…

Inside a router



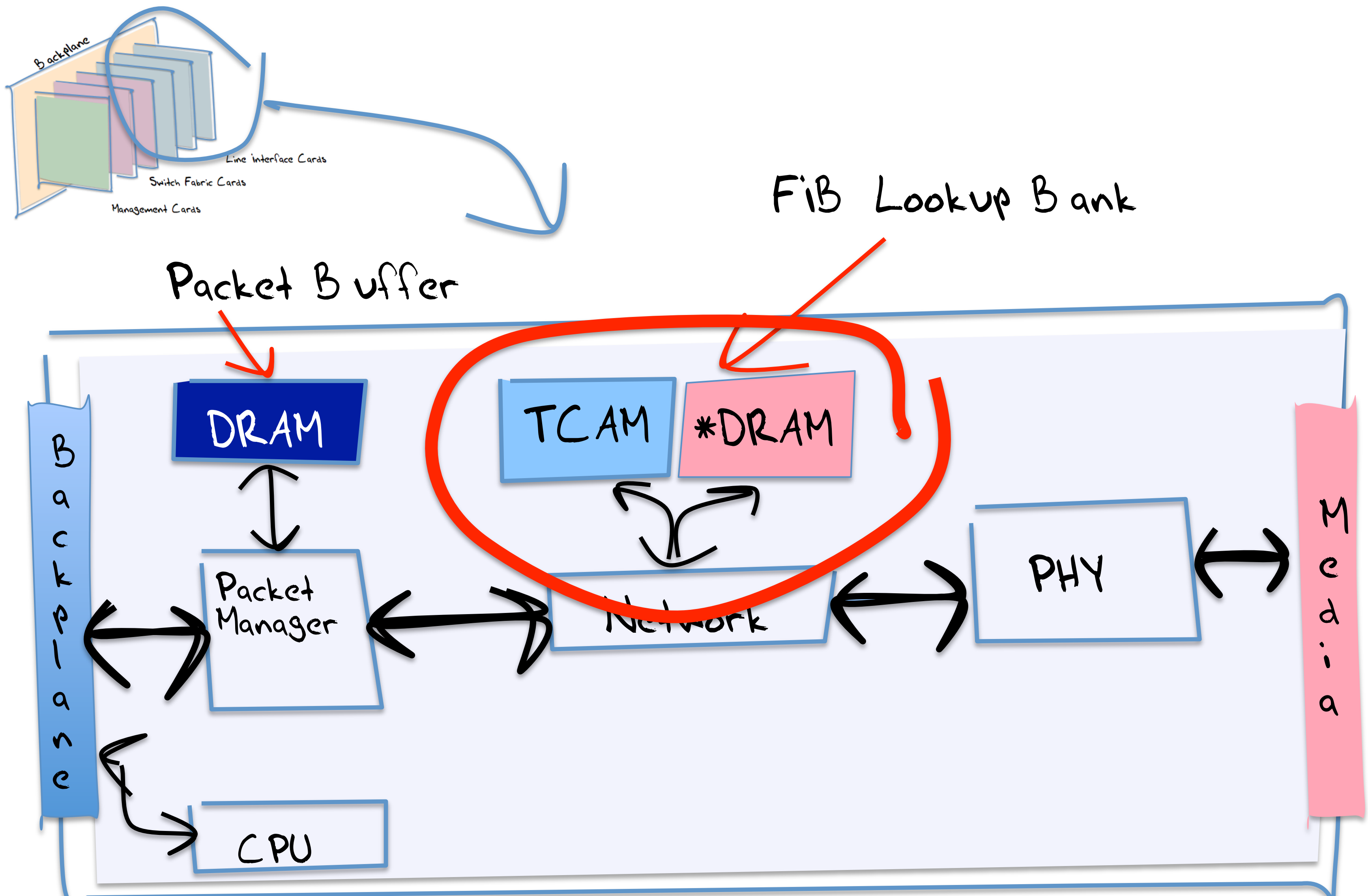
Thanks to Greg Hankins

Inside a line card



Thanks to Greg Hankins

Inside a line card



Thanks to Greg Hankins

FIB Lookup Memory

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:

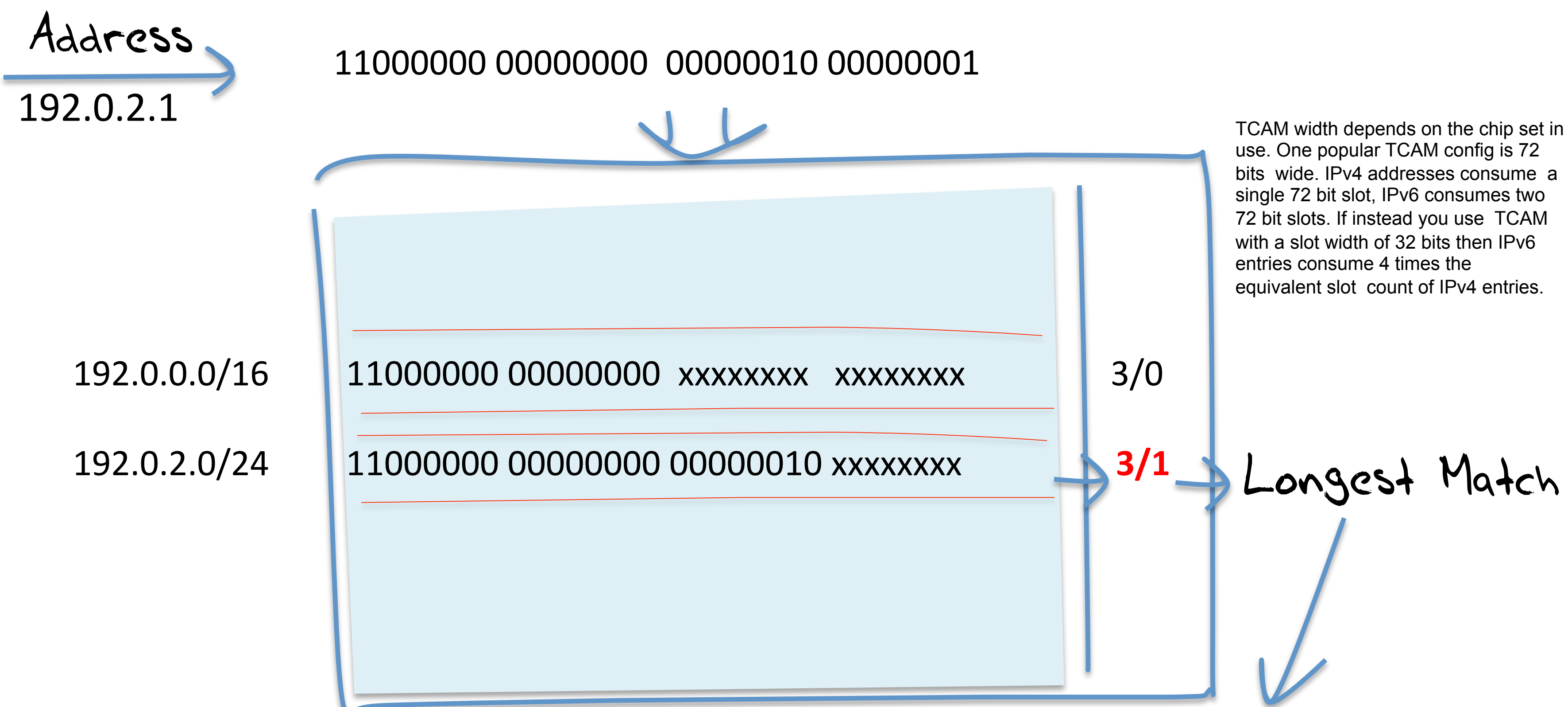
- Loading the entire routing table into a Ternary Content Addressable Memory bank (TCAM)

or

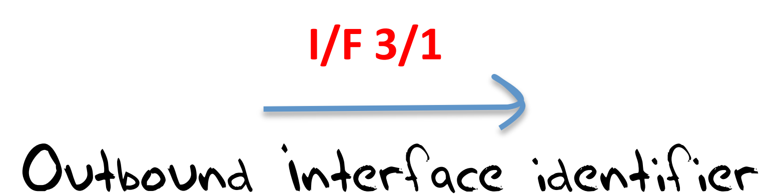
- Using an ASIC implementation of a TRIE representation of the routing table with DRAM memory to hold the routing table

Either way, this needs **fast** memory

TCAM 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.

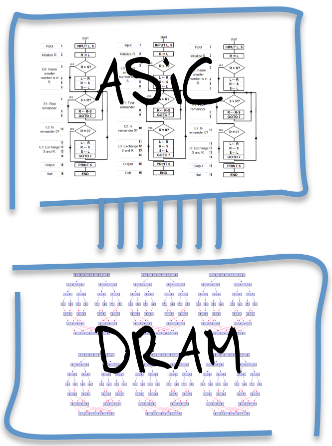


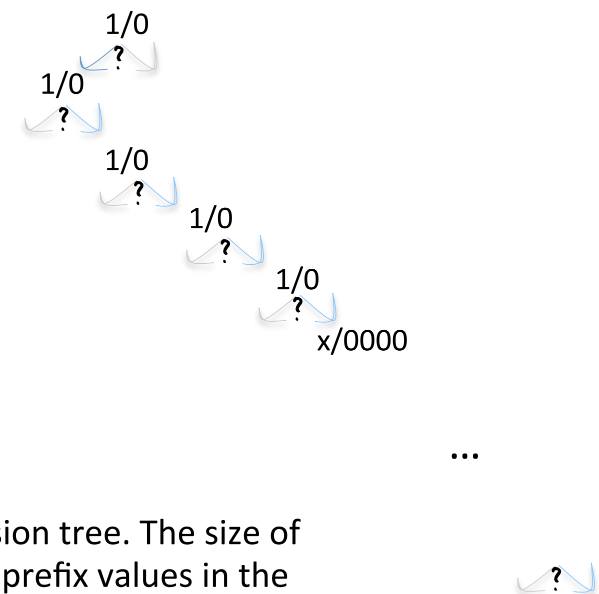
TRIE Lookup

Address

192.0.2.1

11000000 00000000 00000010 00000001





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

I/F 3/1

Outbound interface identifier

Memory Tradeoffs

	TCAM	ASIC + RLDRAM 3
Access Speed	Lower	Higher
$ per bit	Higher	Lower
Power	Higher	Lower
Density	Higher	Lower
Physical Size	Larger	Smaller
Capacity	80Mbit	1G bit

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

Size

What memory size do we need for **10 years** of FIB growth from today?

TCAM

V4: 2M entries (1G$)
plus
V6: 1M entries (2G$)

Trie

V4: 100Mbit memory (500M$)
plus
V6: 200Mbit memory (1G$)

	2014	2019	2024
V4 FIB	512K	768K	1M
V6 FIB	25K	125K	512K

"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.

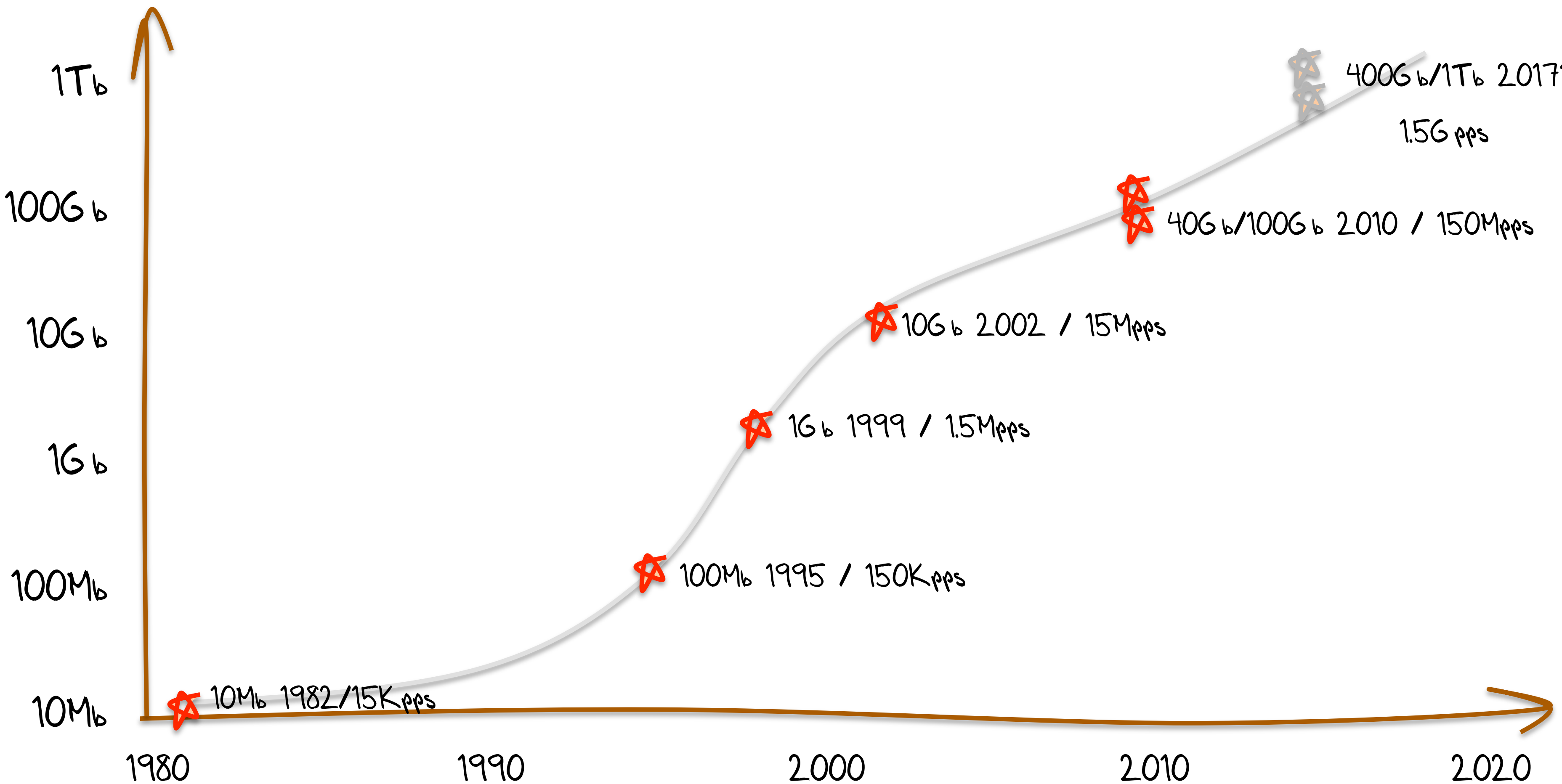
10Mb Ethernet had a 64 byte min packet size, plus preamble plus inter-packet spacing

=14,880 pps

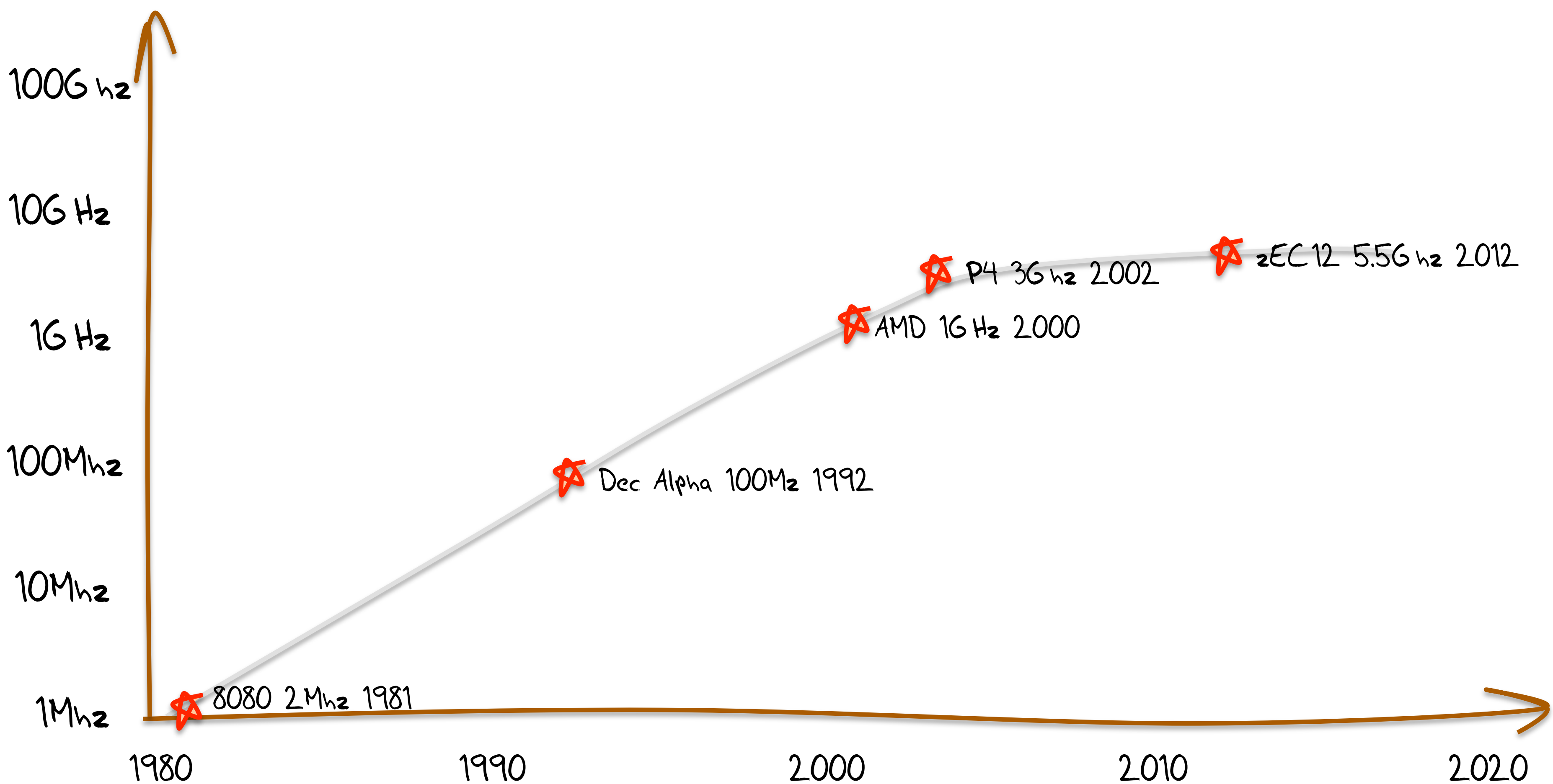
=1 packet every 67usec

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

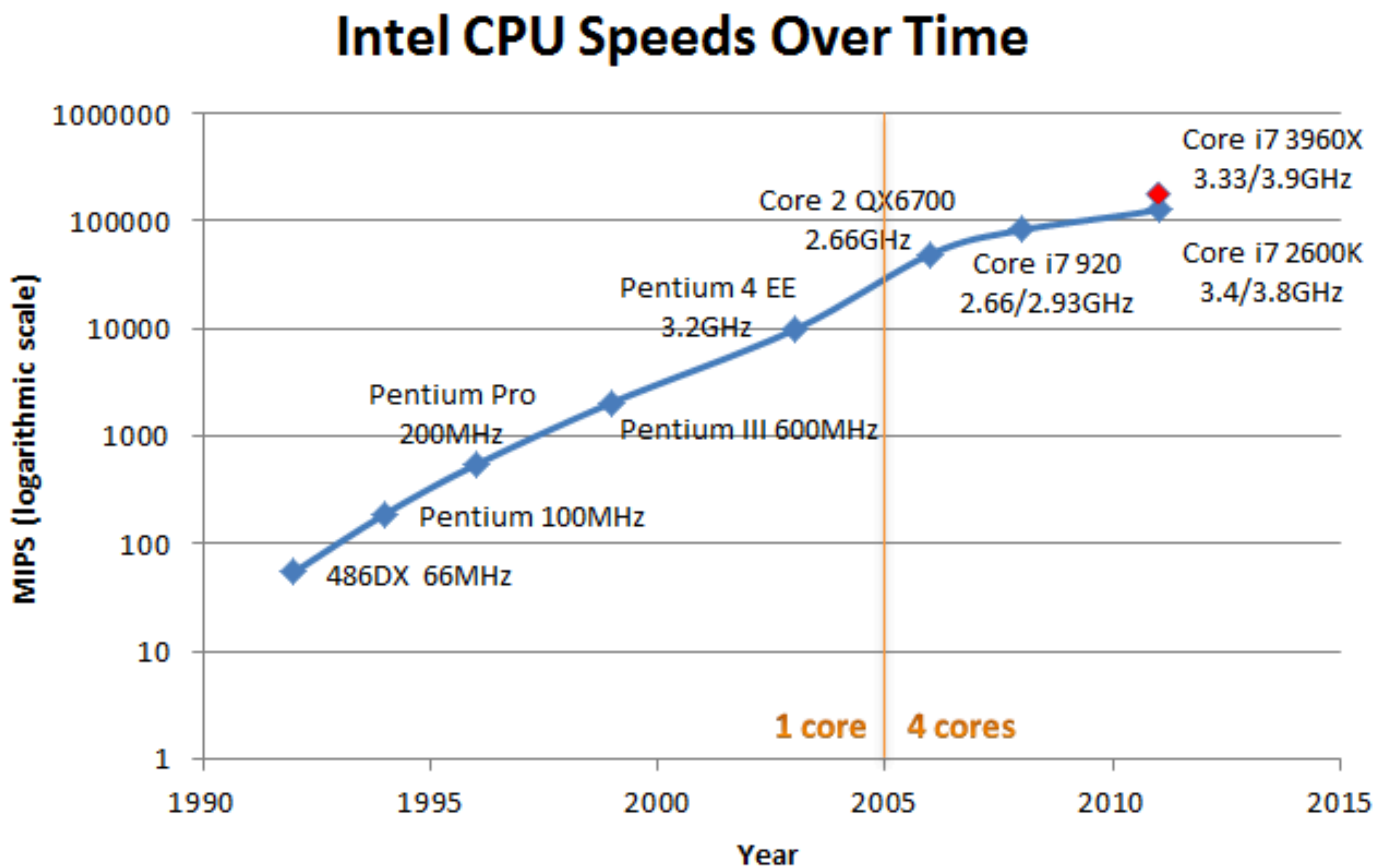
Wireline Speed - Ethernet



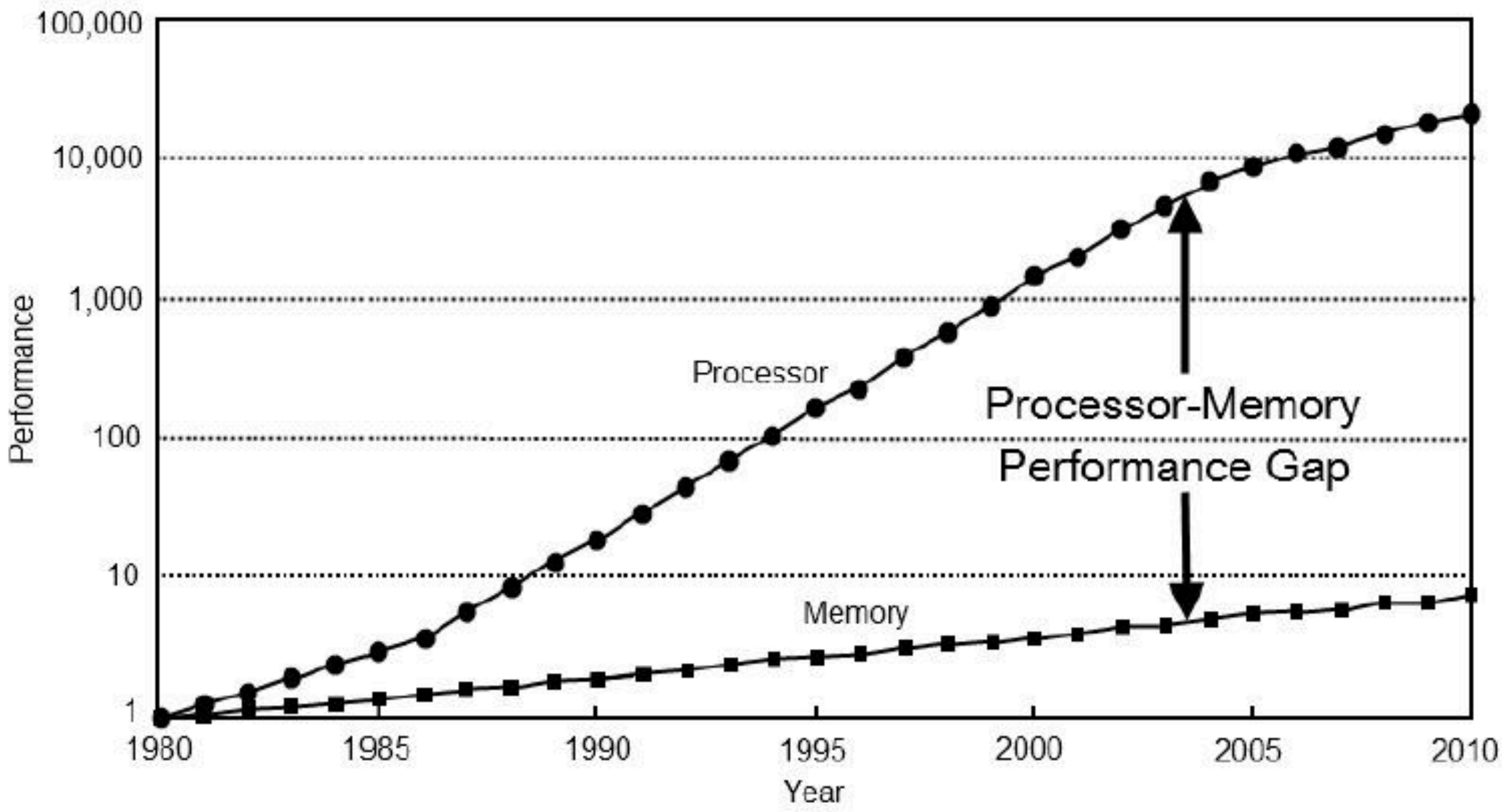
Clock Speed - Processors



Clock Speed - Processors



CPU vs Memory Speed



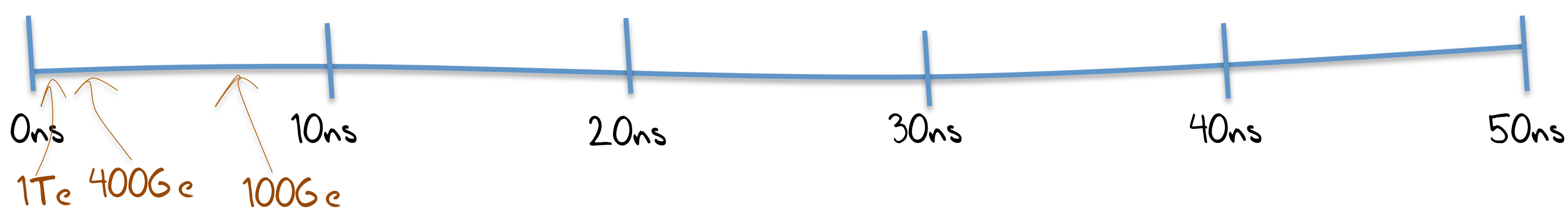
Speed, Speed, Speed

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

100GE ≈ 150Mpps ≈ 6.7ns per packet

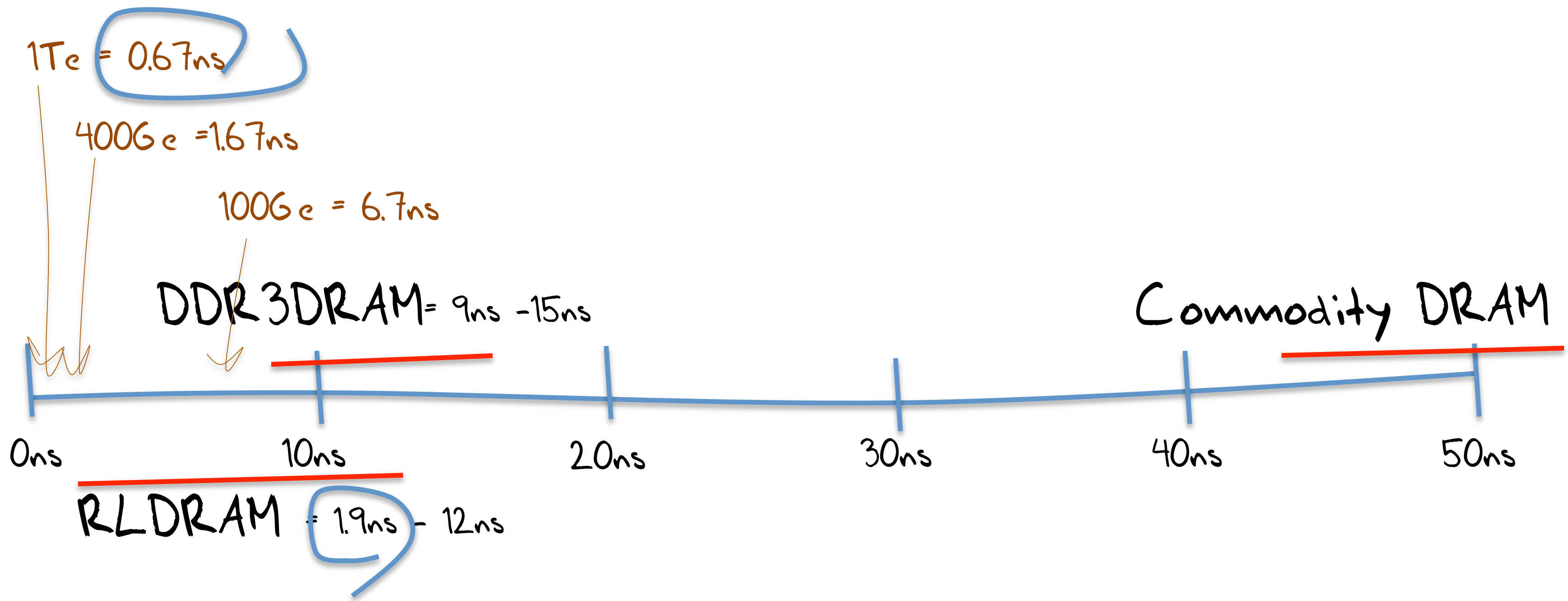
400Ge ≈ 600Mpps ≈ 1.6ns per packet

1Te ≈ 1.5Gpps ≈ 0.67ns per packet



Speed, Speed, Speed

What memory speeds do we have today?



Thanks to Greg Hankins

Scaling Speed

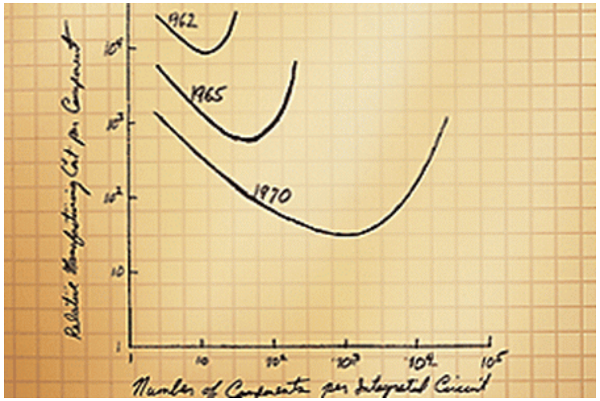
Scaling size is not a dramatic problem today
Scaling speed is going to be tougher over time

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?