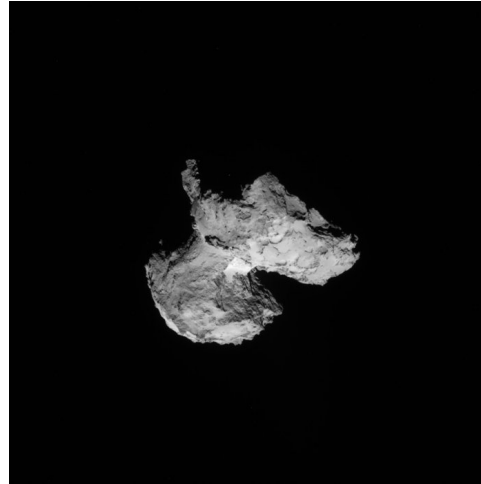# What's so special about the number 512?

Geoff Huston
Chief Scientist, APNIC Labs
1 September 2014

12 August 2014

World Elephant Day

Rosetta closes in on comet 67P/ Churyumov-Gerasimenko

Newborn Panda Triplets in China

Violence continues in the Gaza Strip

The internet apparently has a bad hair day

What happened?

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

12 August 2014

BGP update log for 12 August

12 August 2014

12 August 2014

BGP FIB Size

Cisco Cat 6500

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

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

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

Routing Behaviour

Was the AS701 Route Leak the problem?

Or was the FiB growth passing 512K entries the problem?

What does routing growth look like anyway?

20 years of routing the Internet

2011: Address Exhaustion

2009: The GFC hits the internet

2005: Broadband to the Masses

2001: The Great internet Boom and Bust

1994: introduction of CiDR
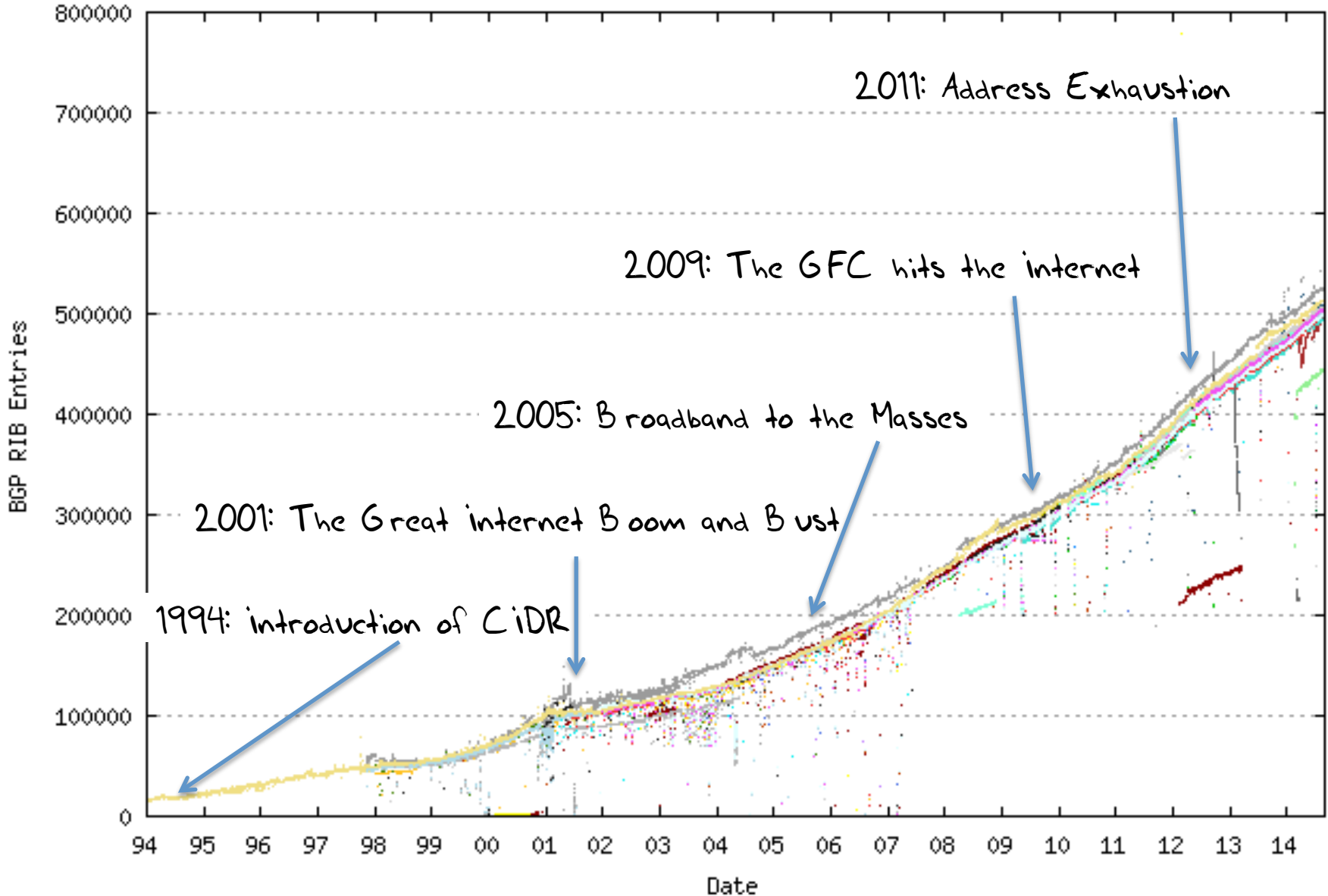
IPv4 BGP Prefix Count 2011 - 2014
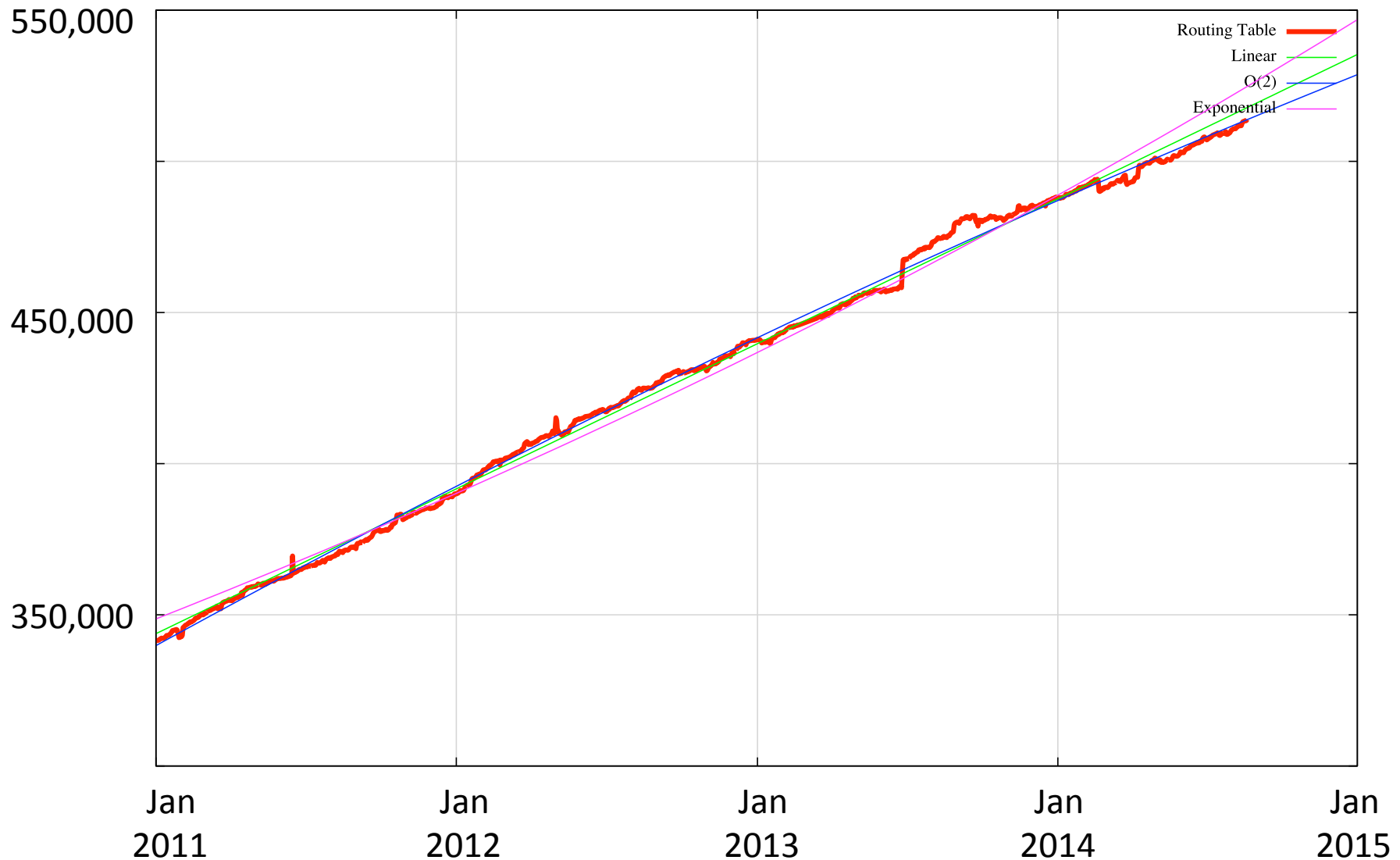
IPv4 2013- 2014 BGP Vital Statistics

	Jan-13	Aug-14	
Prefix Count	440,000	512,000	+ 11% p.a.
Roots	216,000	249,000	+ 9%
More Specifics	224,000	264,000	+ 11%
Address Span	156/8s	162/8s	+ **2%**
AS Count	43,000	48,000	+ 7%
Transit	6,100	7,000	+ 9%
Stub	36,900	41,000	+ 7%

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
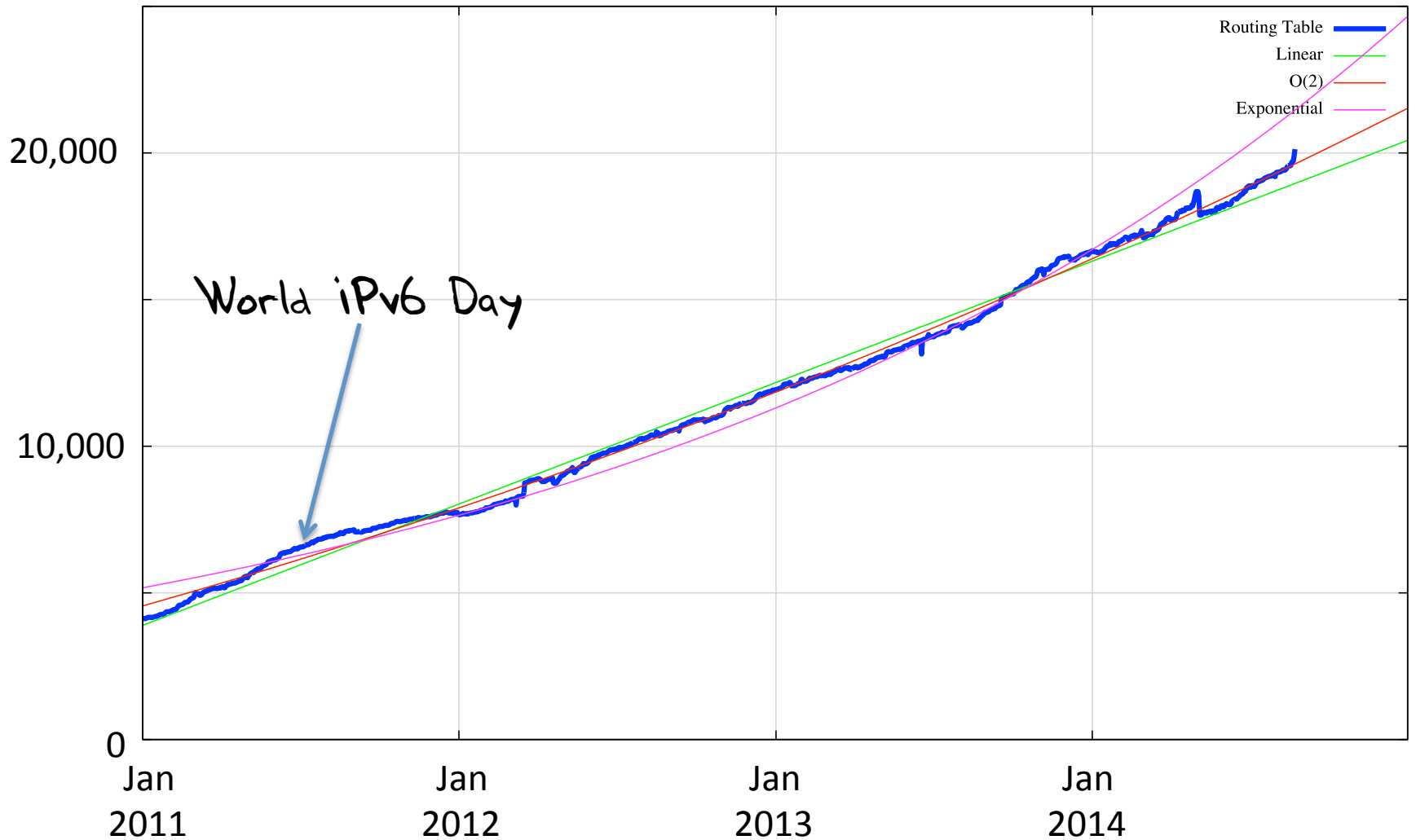
IPv6 2013-2014 BGP Vital Statistics

	Jan-13	Aug-14	p.a. rate
Prefix Count	11,500	19,036	+ 39%
Roots	8,451	12,998	+ 32%
More Specifics	3,049	6,038	+ 59%
Address Span (/32s)	65,127	73,153	+ 7%
AS Count	6,560	8,684	+ 19%
Transit	1,260	1,676	+ 20%
Stub	5,300	7,008	+ 19%

IPv6 in 2013

- Overall IPv6 Internet growth in terms of BGP is **20% - 40 % p.a.**

  - 2012 growth rate was ~ 90%.

(Looking at the AS count, if these relative growth rates persist then the IPv6 network would span the same network domain as IPv4 in ~16 years time -- 2030!)

IPv6 in 2013 - Growth is Slowing?

- Overall Internet growth in terms of BGP is at a rate of some **~20-40% p.a.**

- AS growth sub-linear

- The rate of growth of the IPv6 Internet is also slowing down

  – Lack of critical momentum behind IPv6?

  – Saturation of critical market sectors by IPv4?

  – *<some other factor>*?

What to expect

BGP Size Projections

- Generate a projection of the IPv4 routing table using a quadratic (O(2) polynomial) over the historic data
  - 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!

IPv4 Table Size

Linear: Y = -169638 + (t * 1517.584)

Quadratic: Y = -4973214 + (6484.294 * t) + (-1.837814* 10^{-12} * t^2)

Exponential: Y = 10**(3.540416 + (year * 1.547426e-09))

Legend: Routing Table, Linear, O(2), Exponential
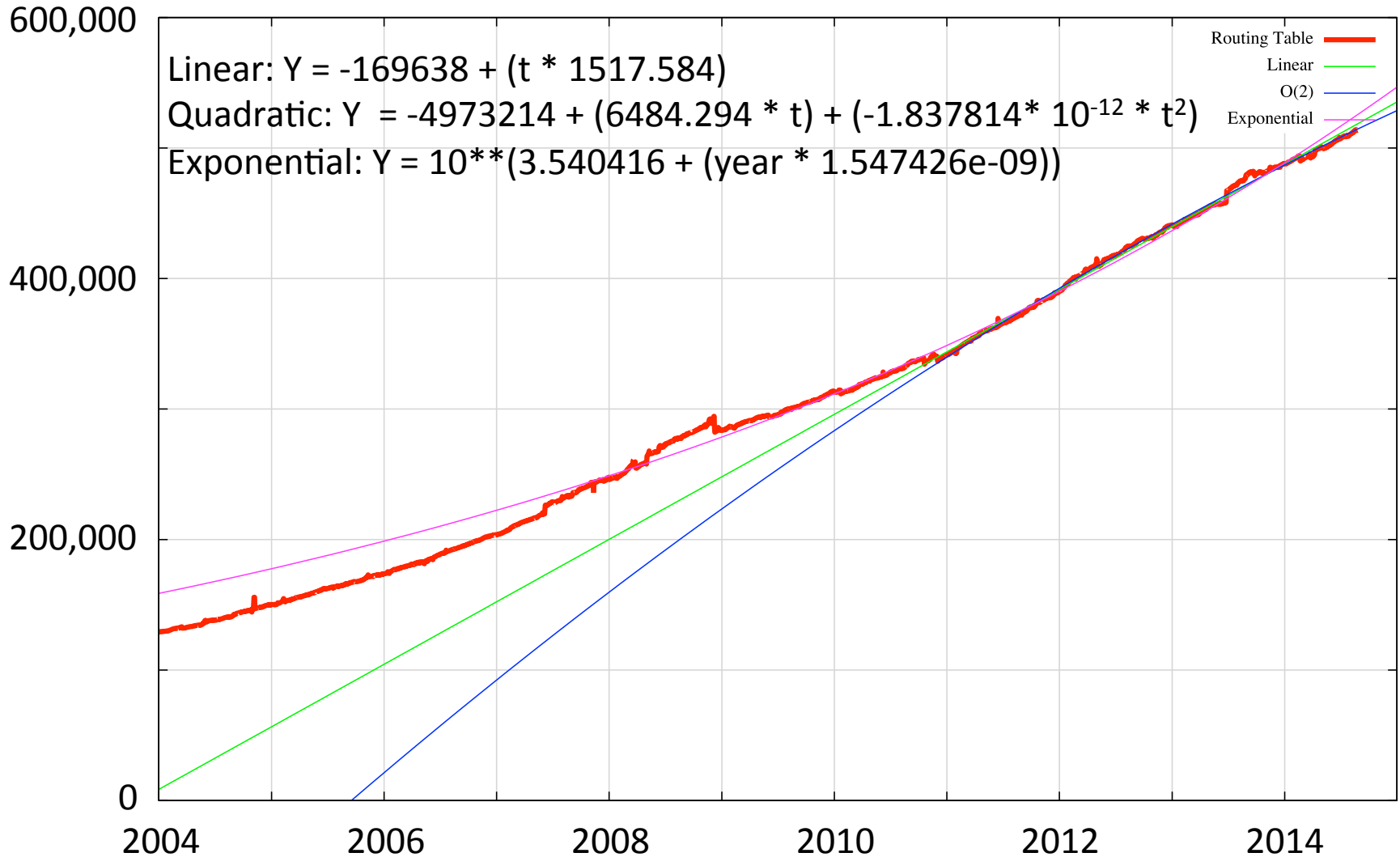
V4 - Daily Growth Rates
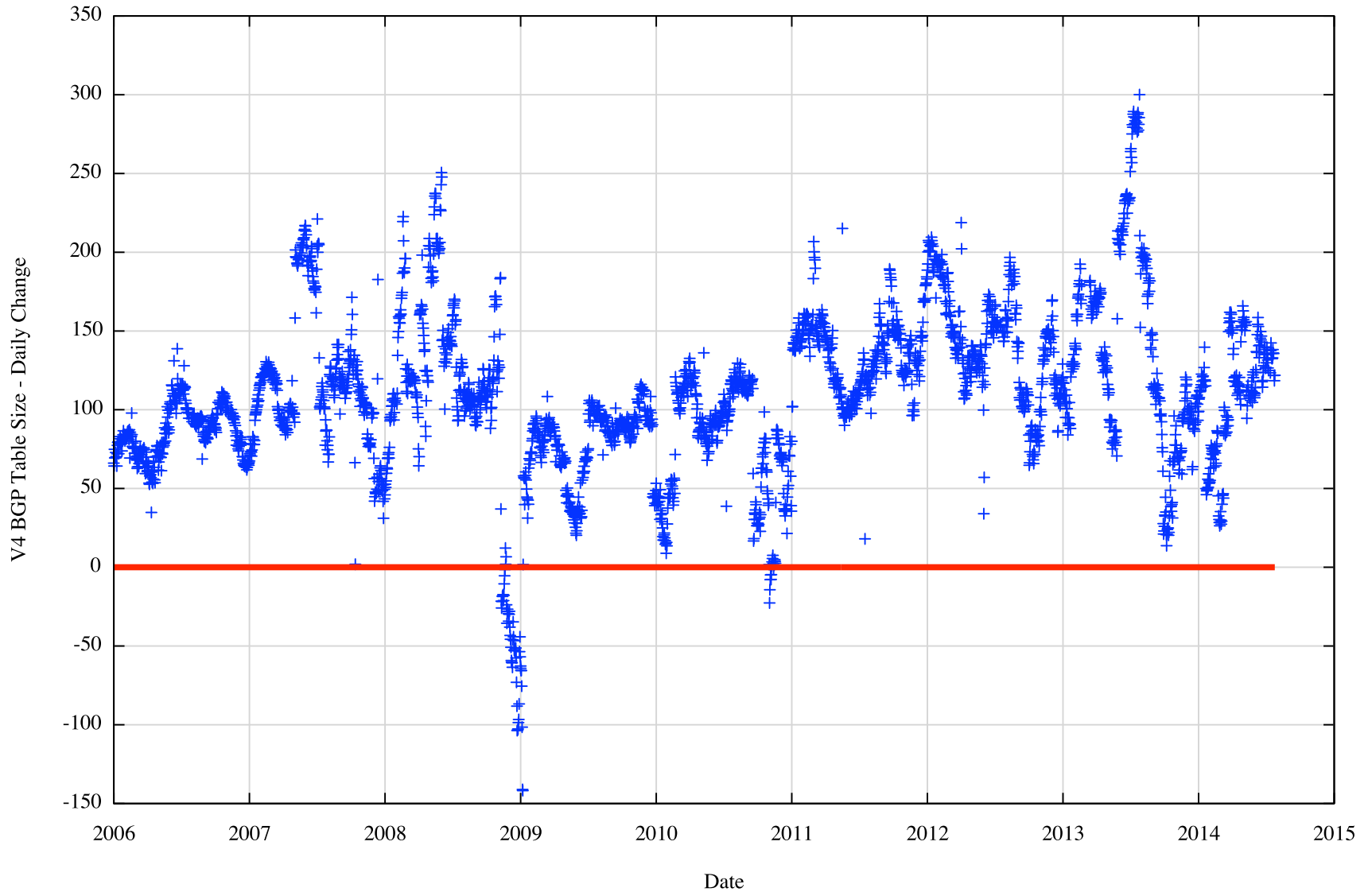
V4 - Relative Daily Growth Rates
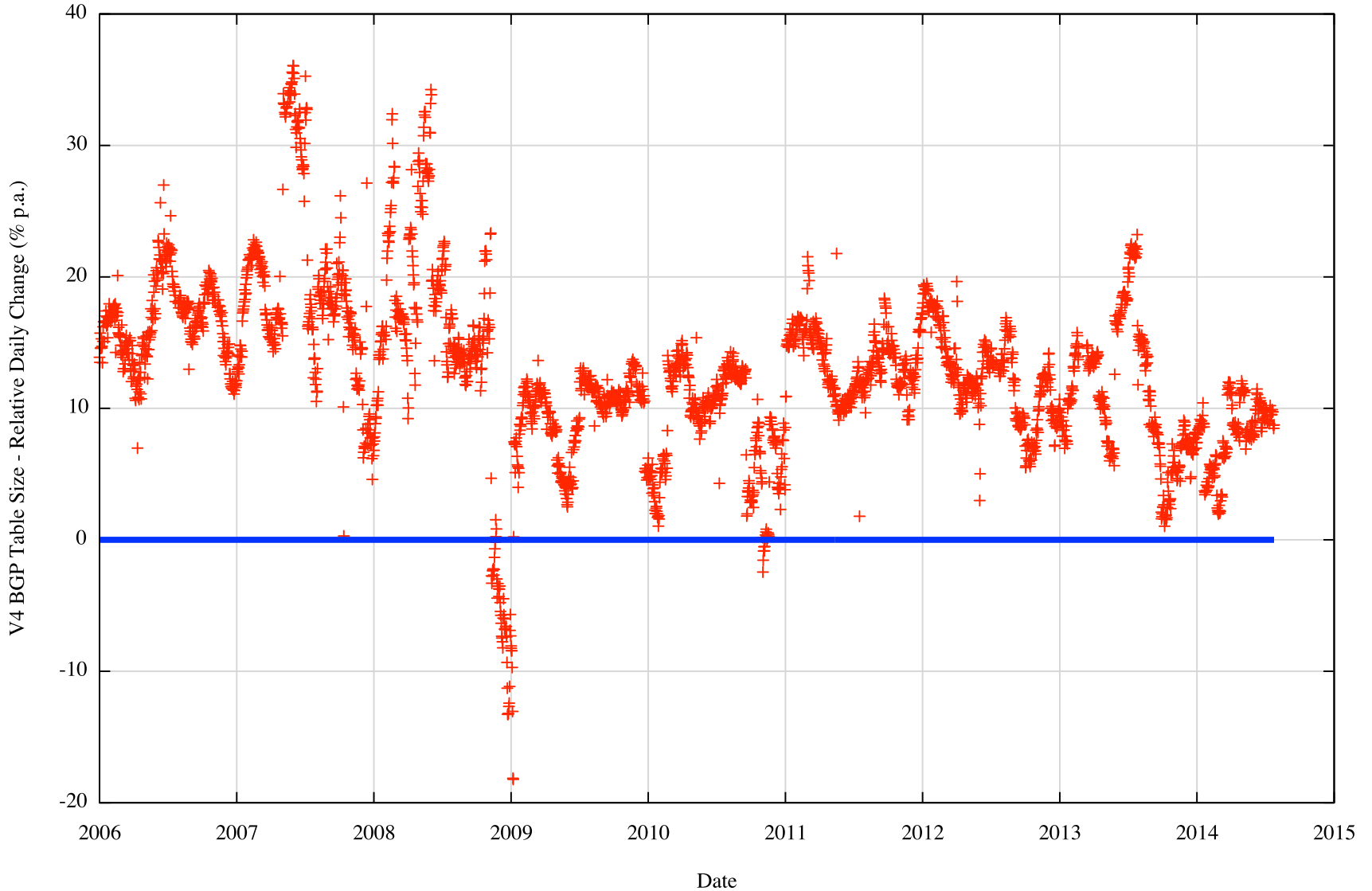
V4 - Relative Daily Growth Rates
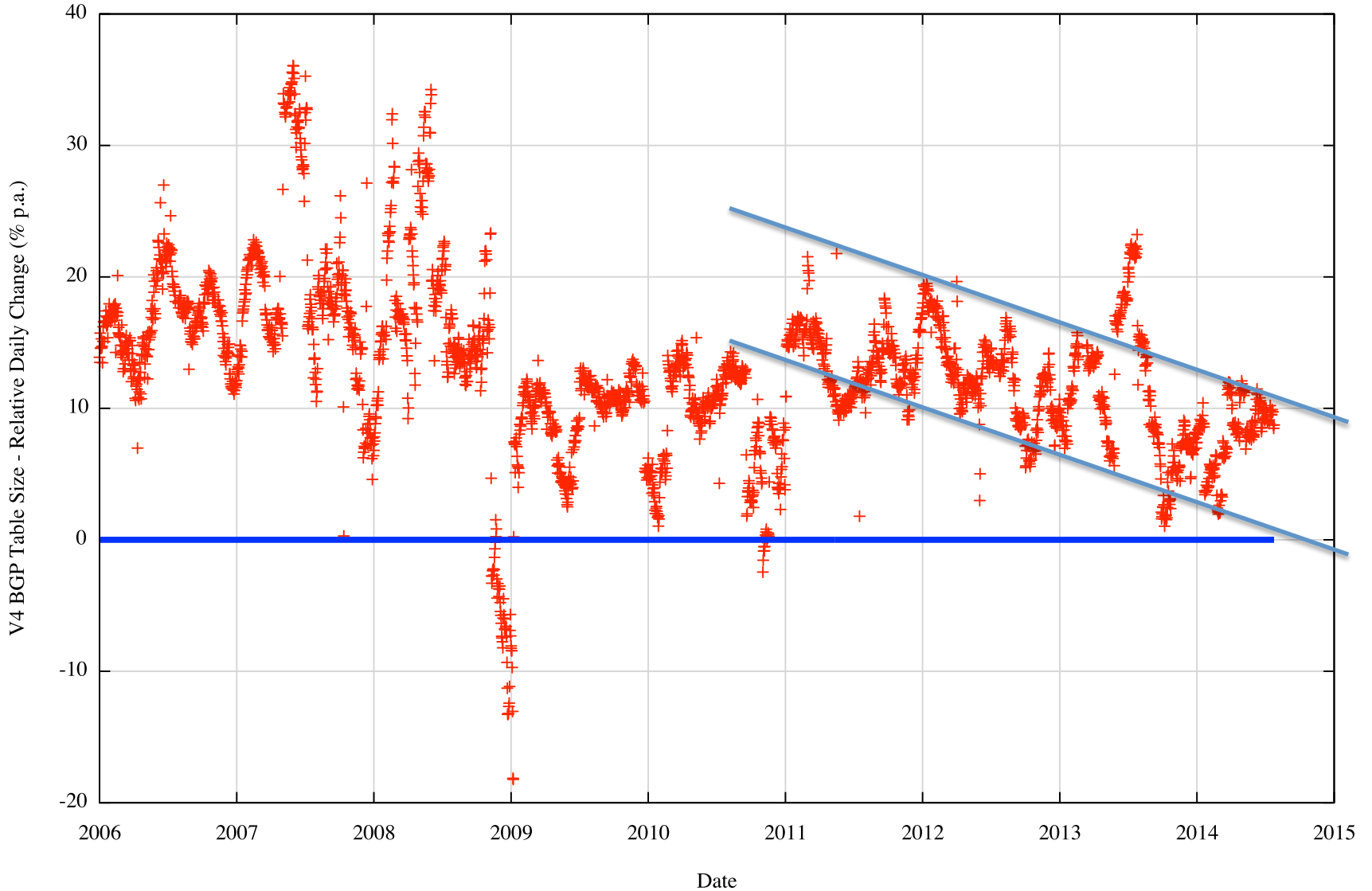
IPv4 BGP Table Size predictions

	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 BGP FIB Size

Legend:
- Routing Table
- Linear
- O(2)
- Exponential
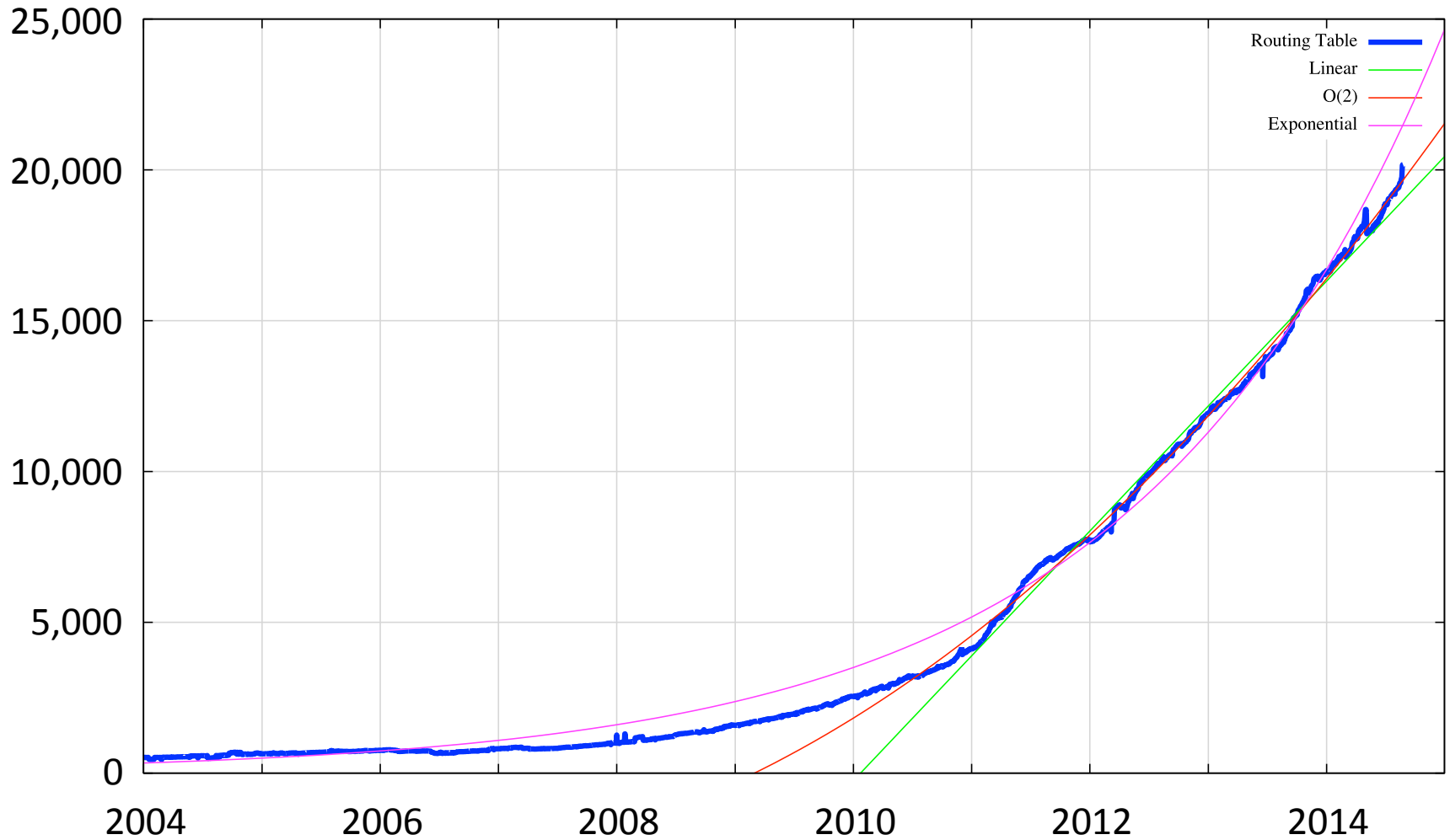
V6 - Daily Growth Rates
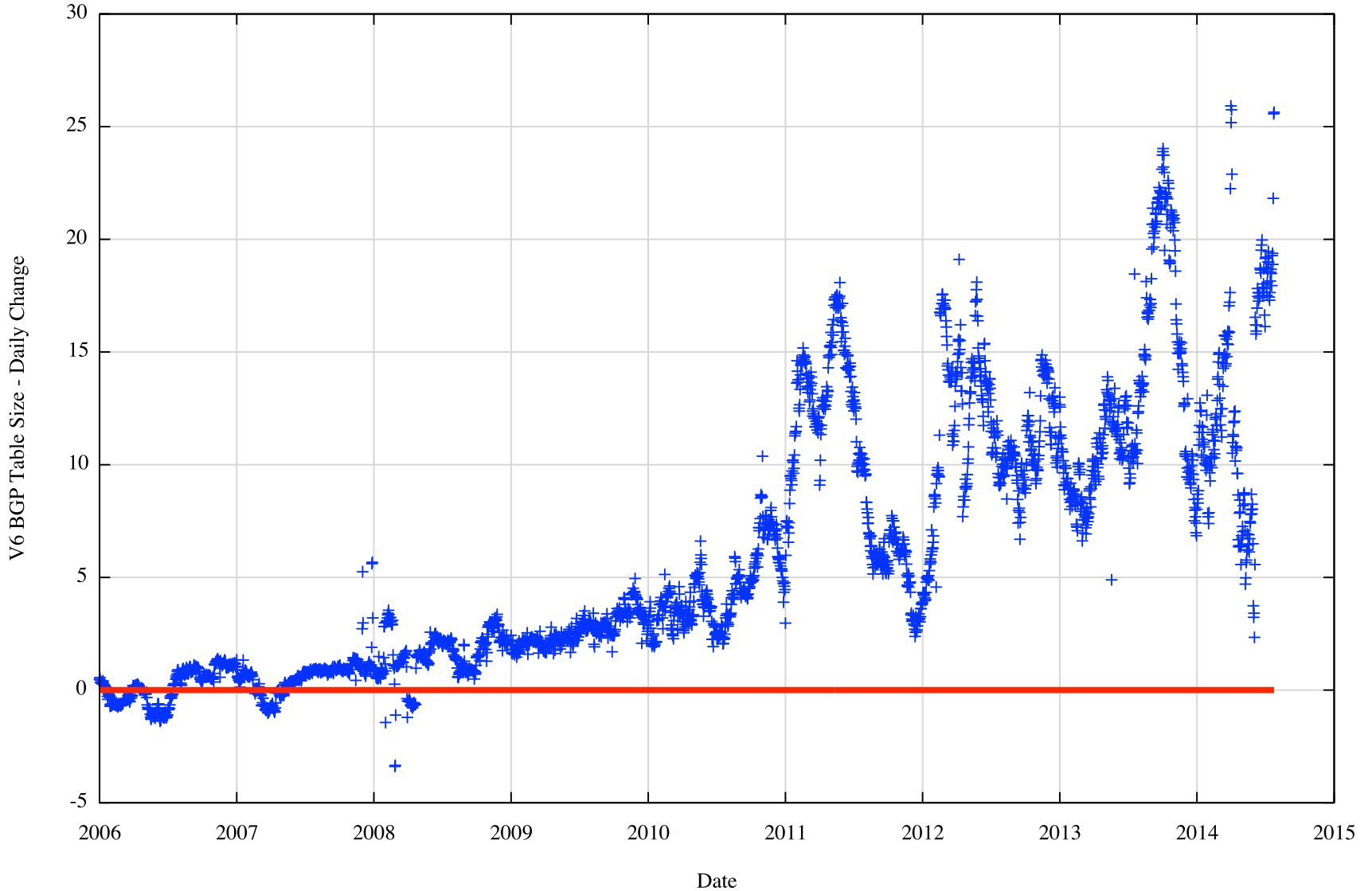
V6 - Relative Growth Rates
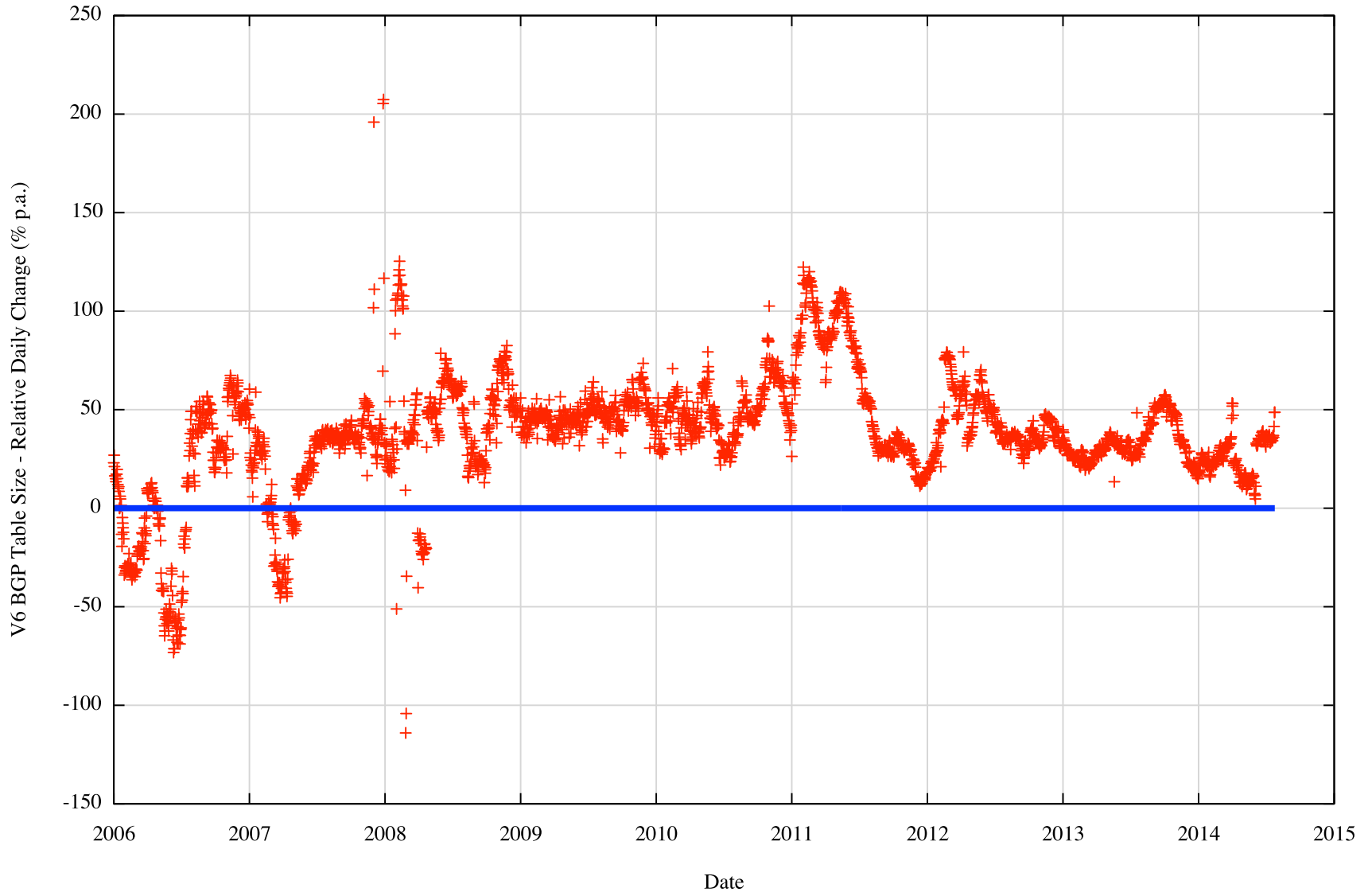
V6 - Relative Growth Rates
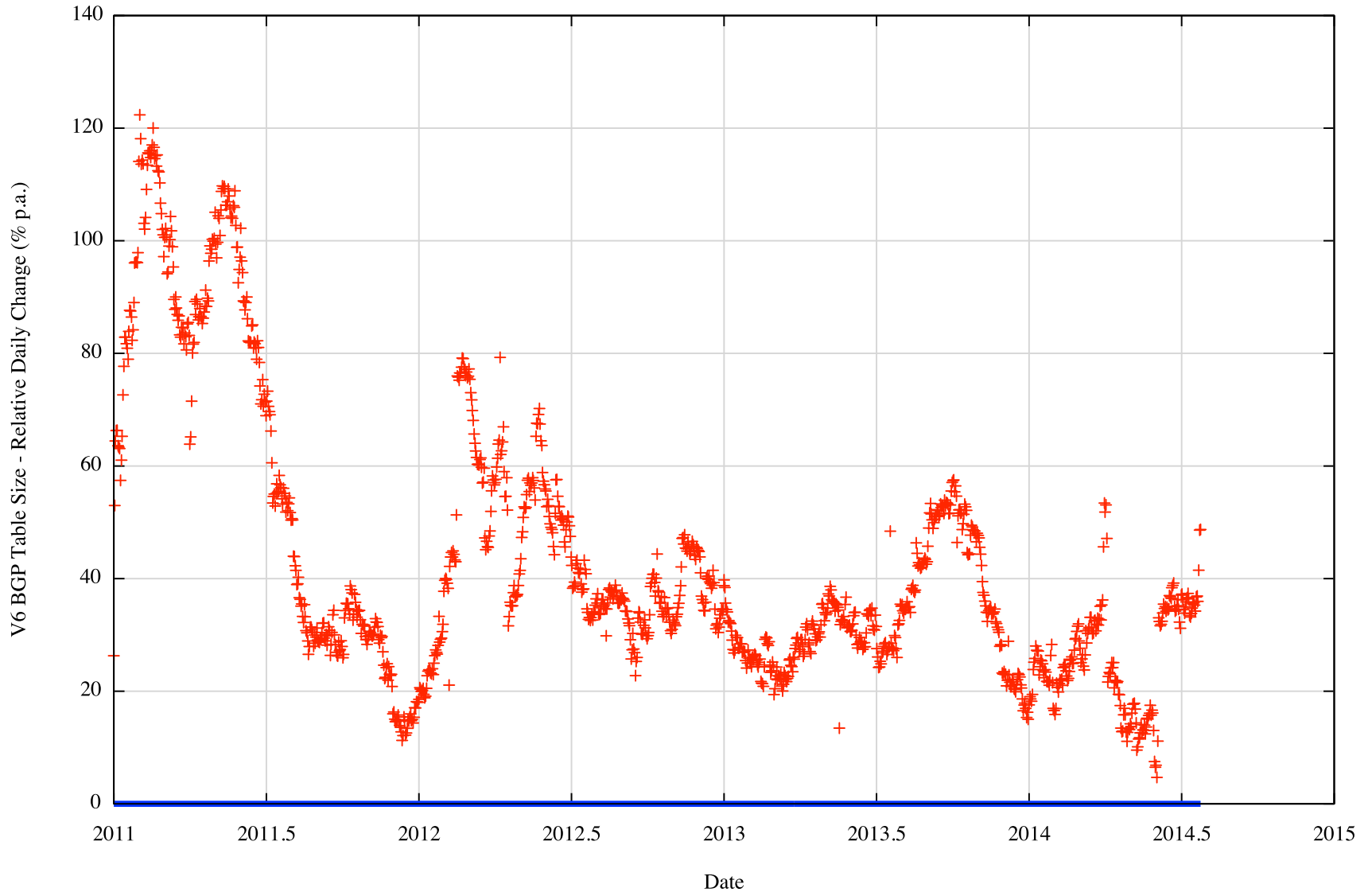
V6 - Relative Growth Rates
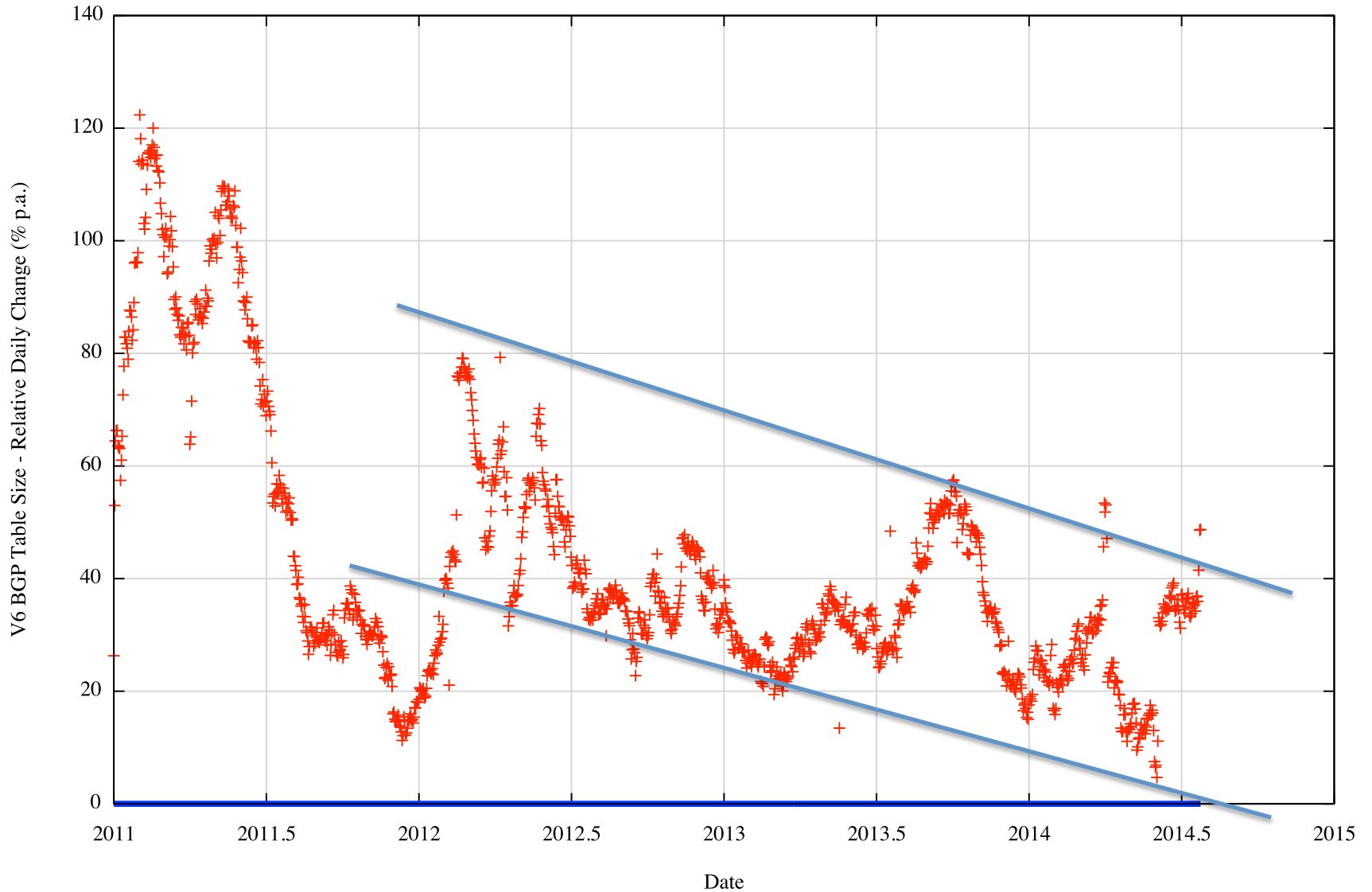
IPv6 BGP Table Size predictions

	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

Up and to the Right

- Most Internet curves are "up and to the right"
- But what makes this curve painful?
  - The pain threshold is approximated by Moore's Law

Microprocessor Transistor Counts 1971-2011 & Moore's Law
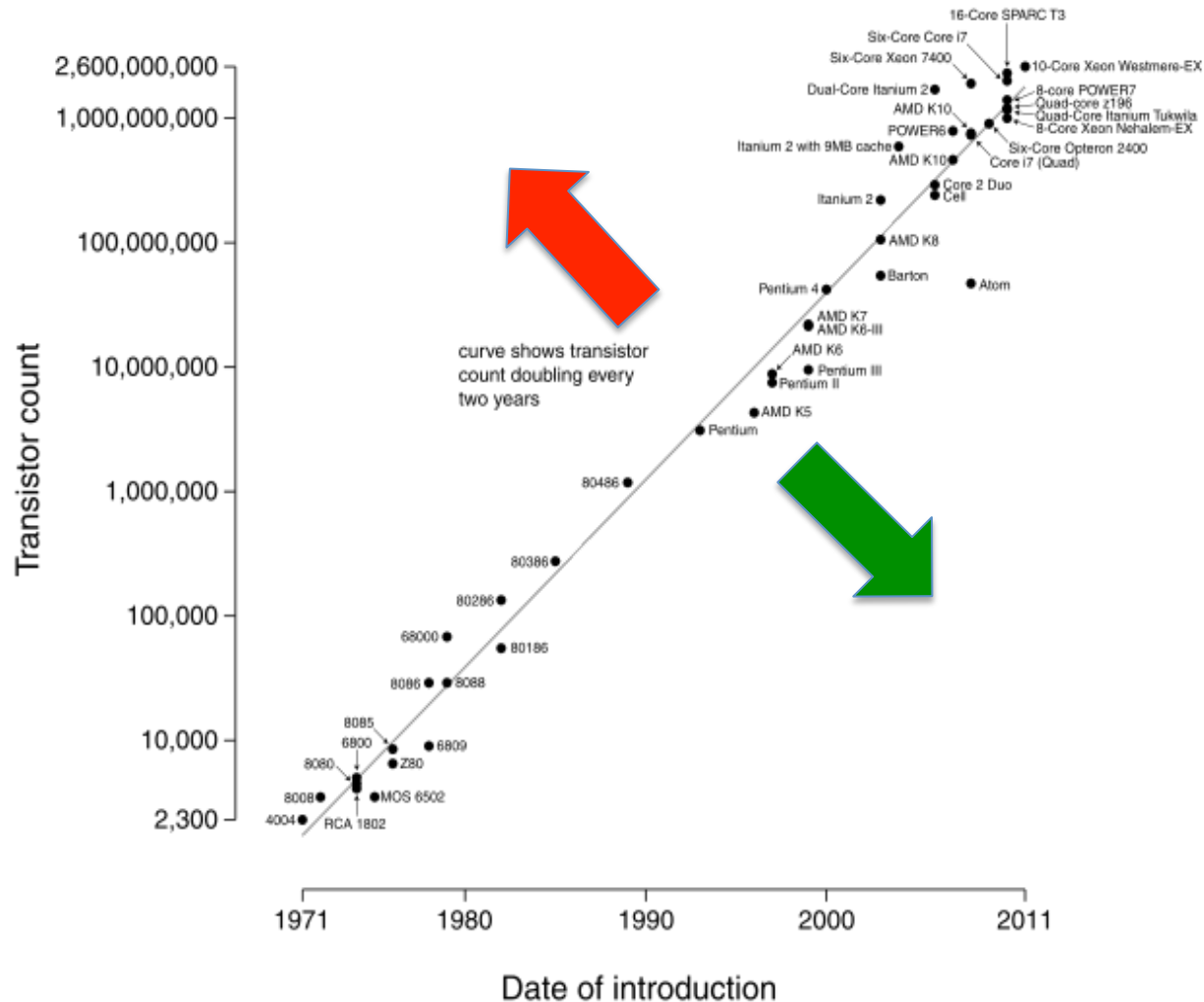
IPv4 BGP Table size and Moore's Law
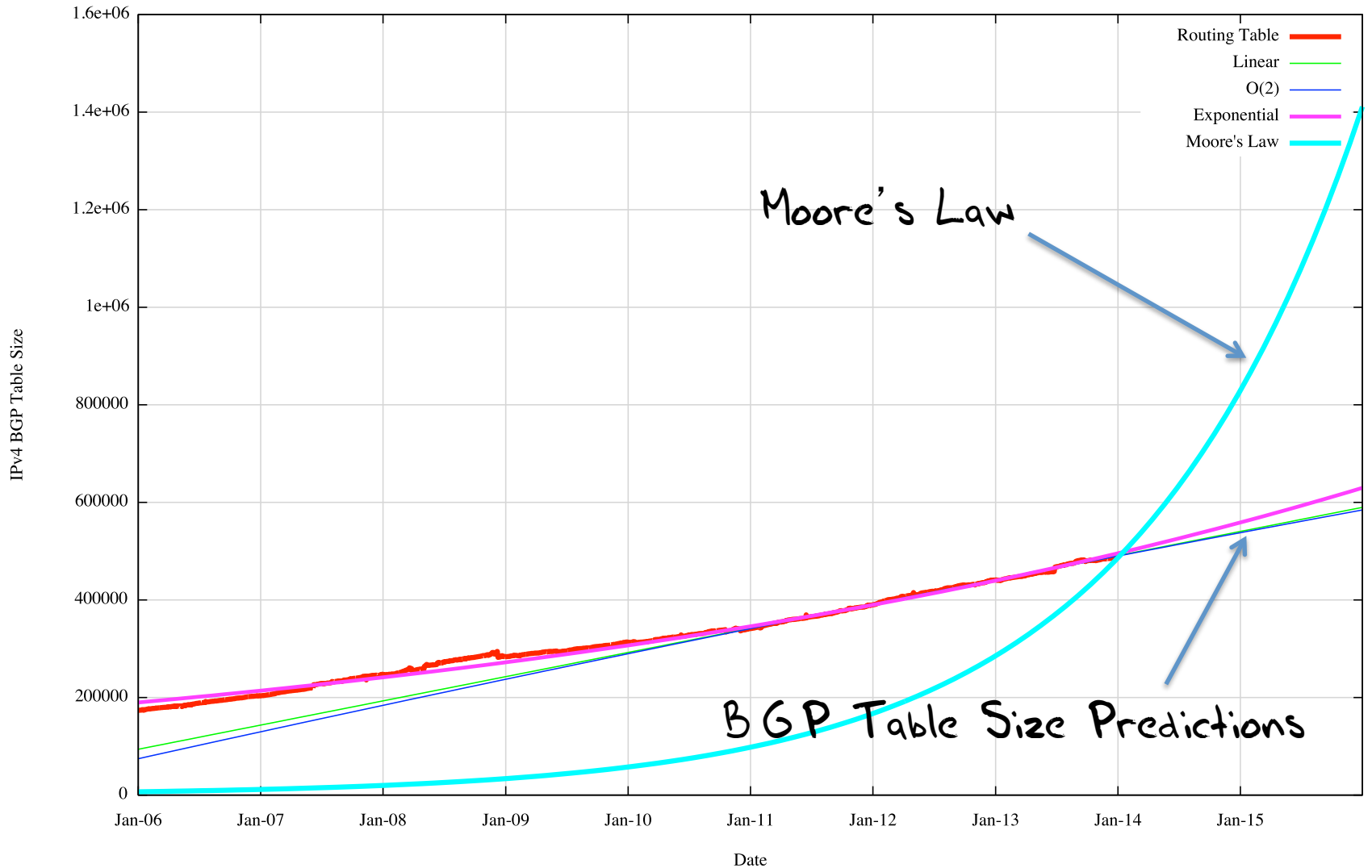
IPv6 Projections and Moore's Law
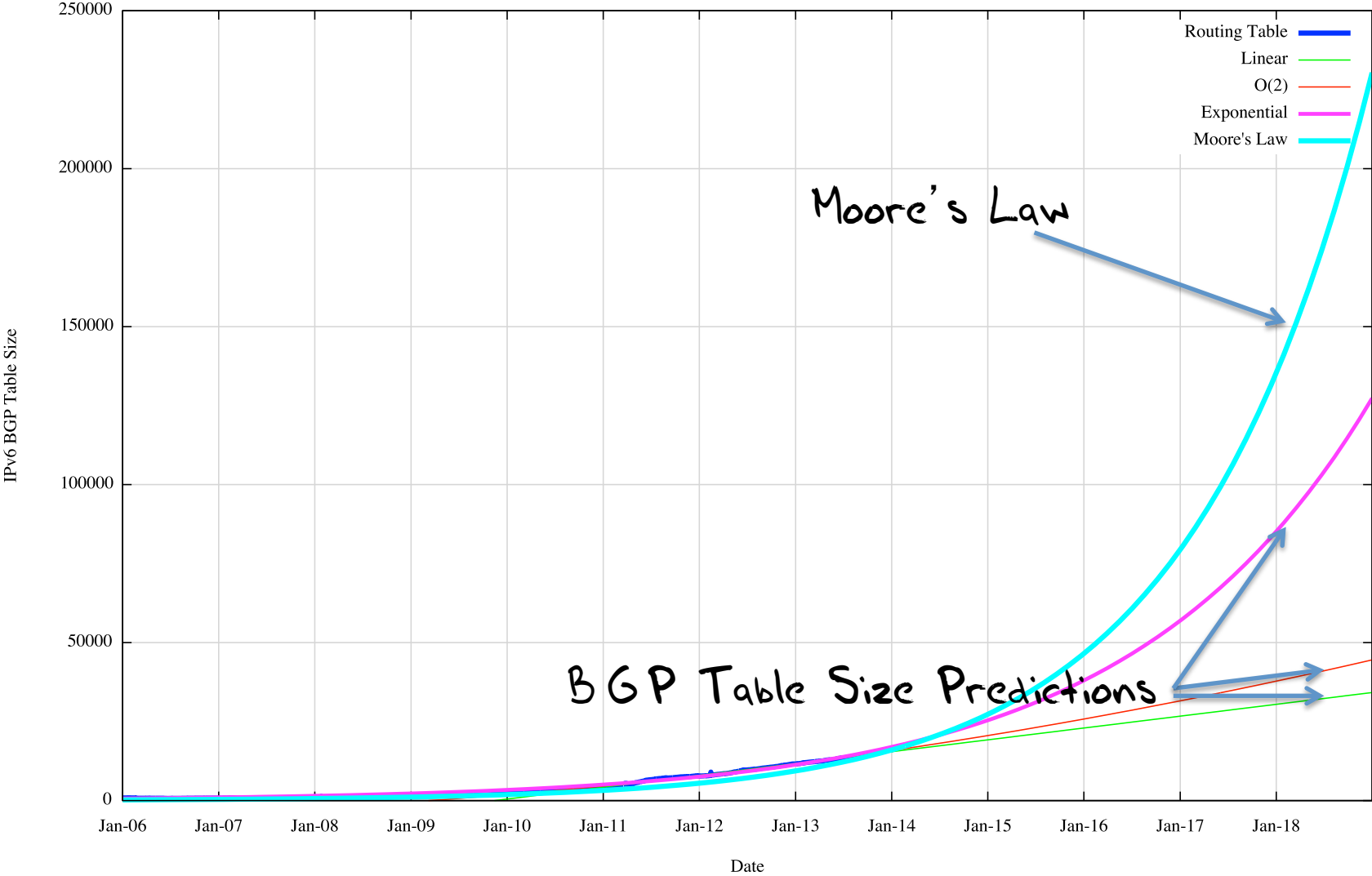
BGP Table Growth

- 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

Table Size vs Updates

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
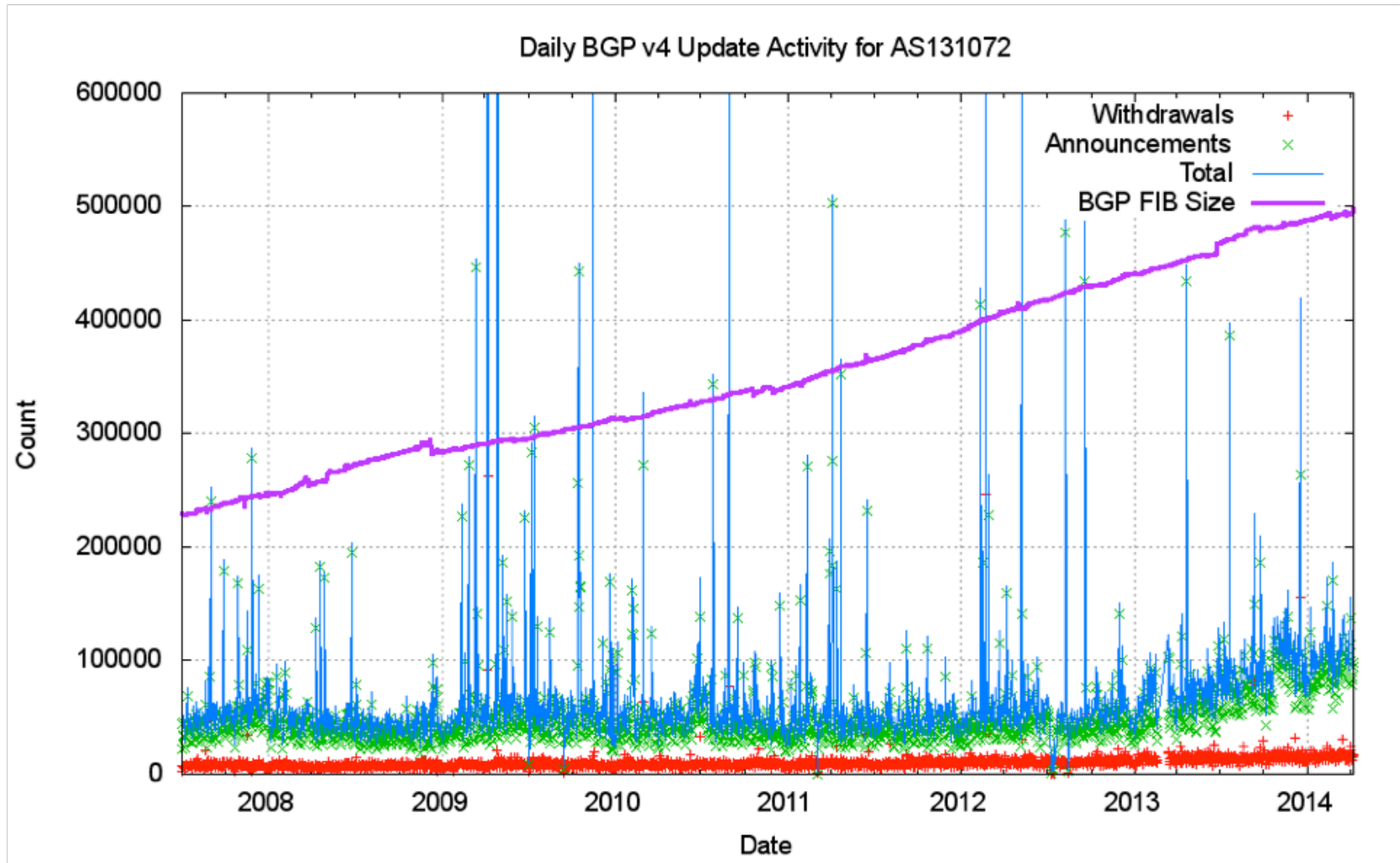
Convergence Performance

Average Convergence Time per day (AS 131072)
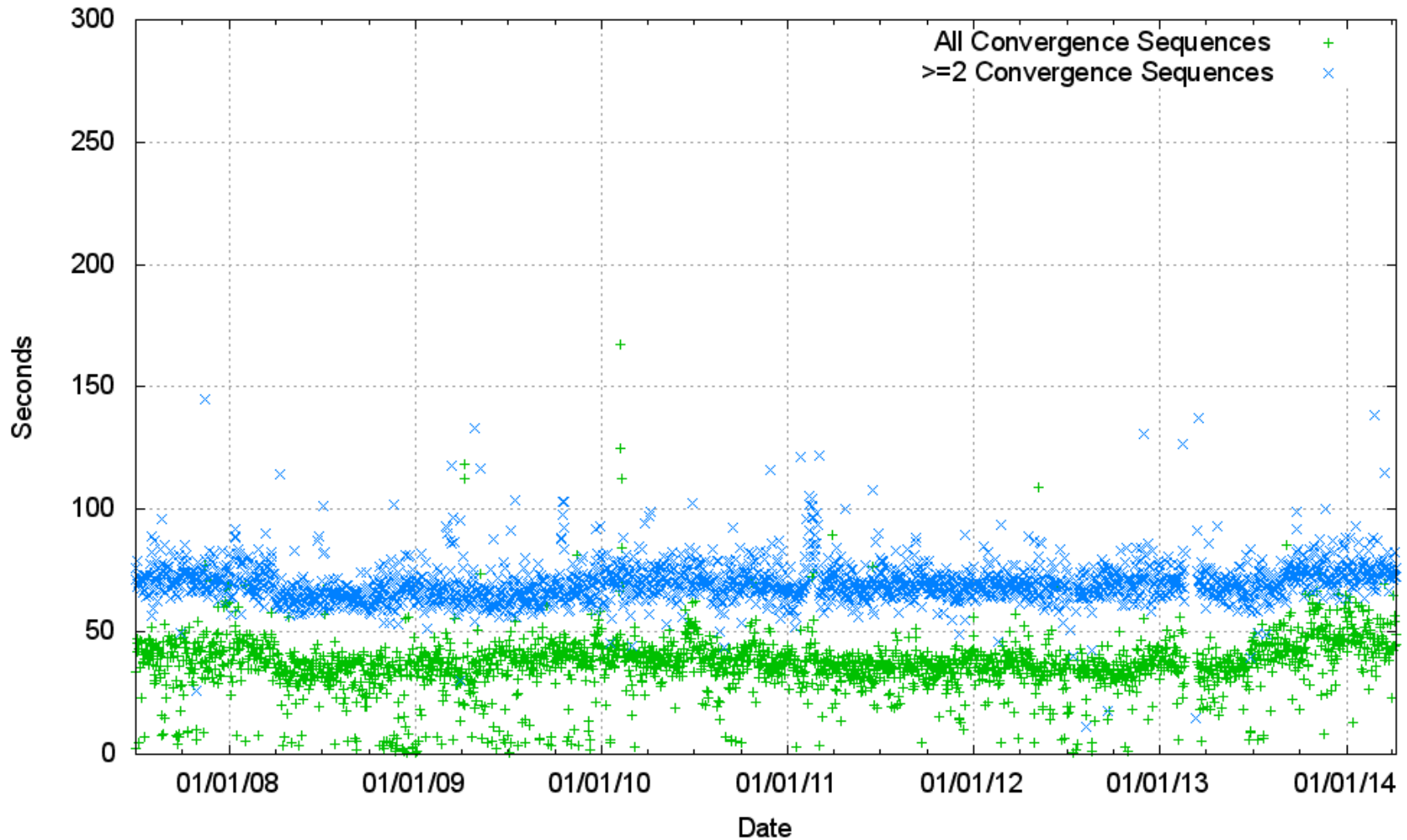
IPv4 Average AS Path Length
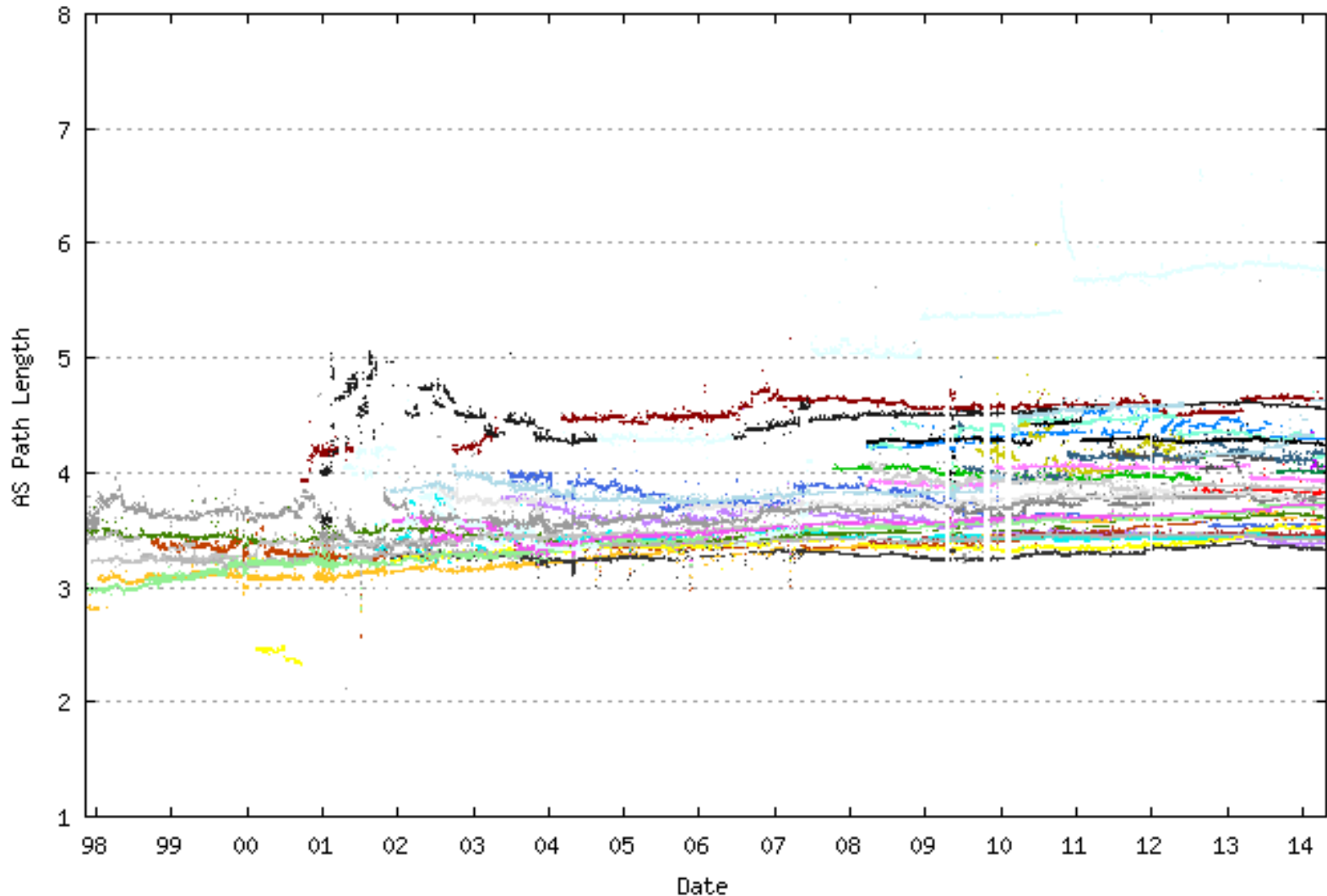
Data from Route Views

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
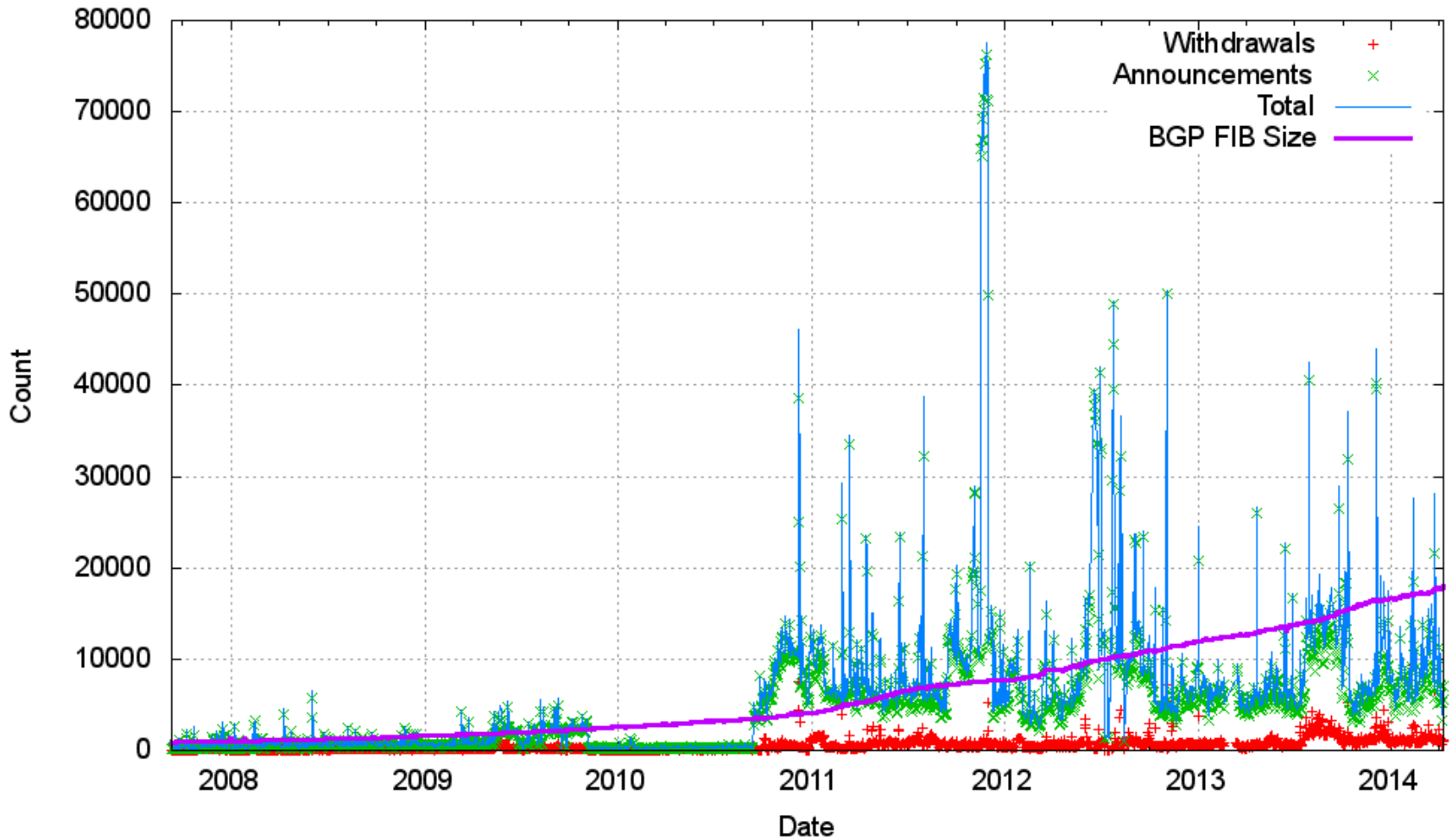
V6 Convergence Performance

Average Convergence Time per day (AS 131072)
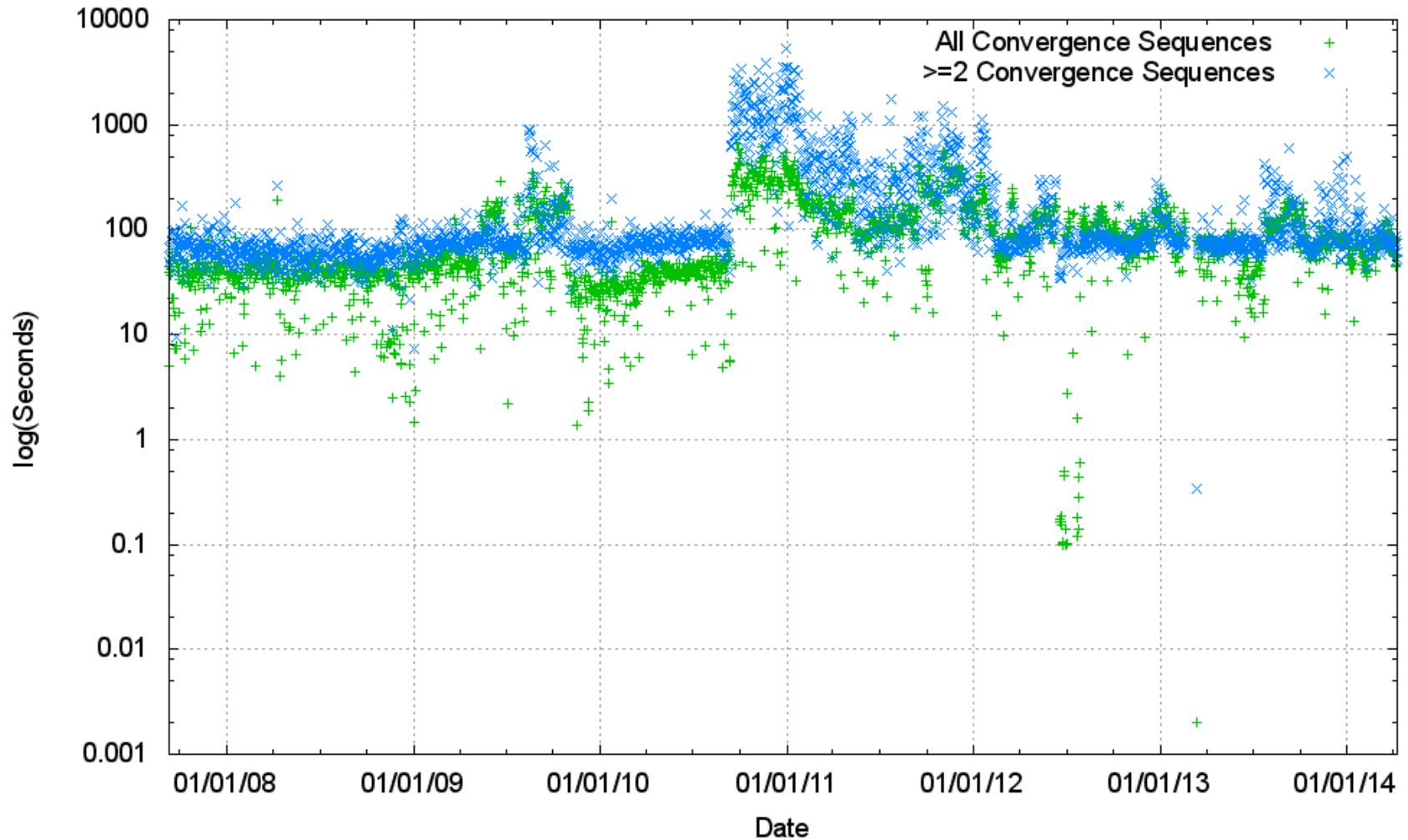
V6 Average AS Path Length
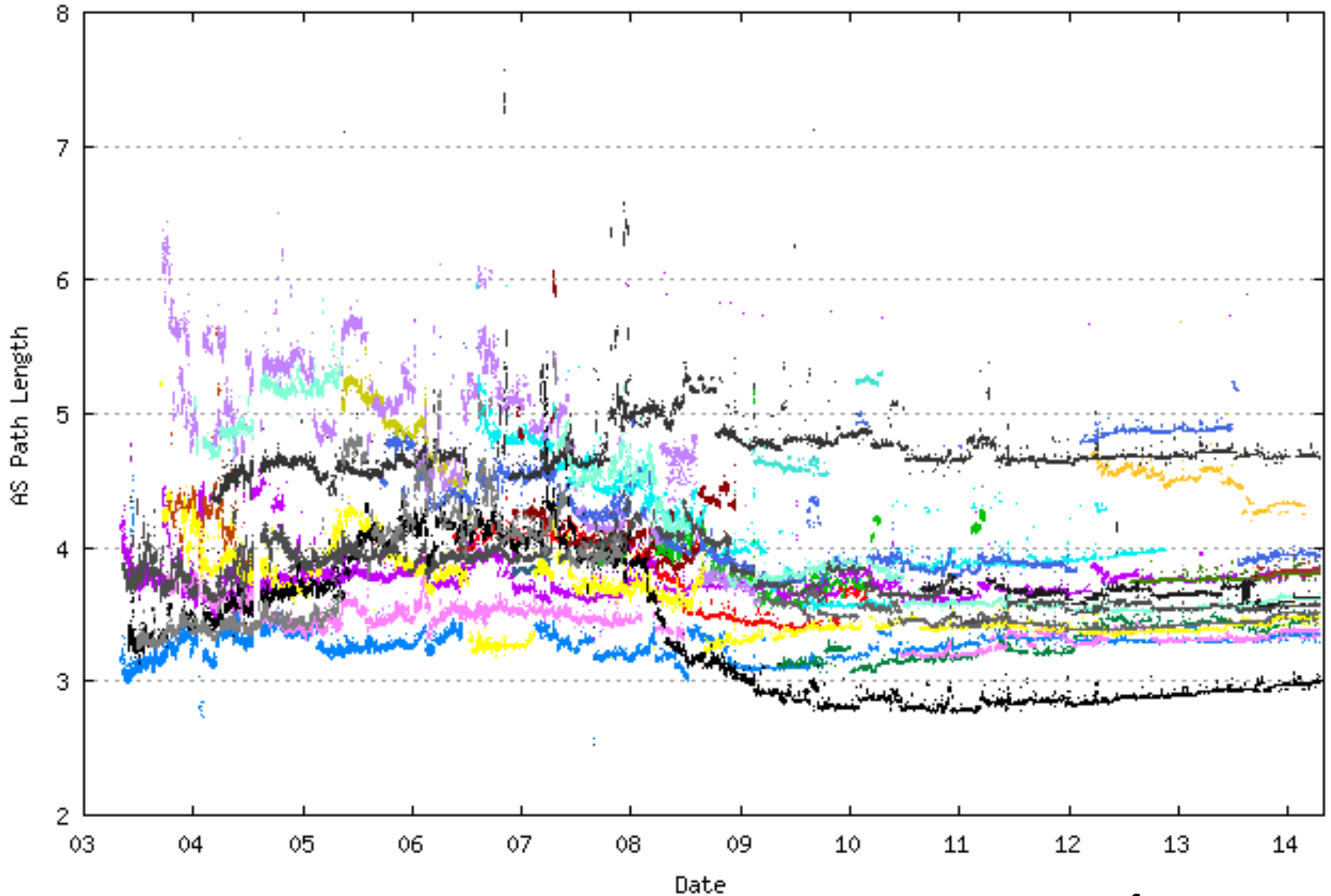
Data from Route Views

BGP Convergence

- The long term average convergence time for the IPv4 BGP network is some 70 seconds, or 2.3 updates given a 30 second MRAI timer

- The long term average convergence time for the IPv6 BGP network is some 80 seconds, or 2.6 updates

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

Backplane

Line interface Card

Switch Fabric Card

Management Card

Inside a line card

Backplane

Line Interface Cards

Switch Fabric Cards

Management Cards

Packet Buffer

FiB Lookup Bank

DRAM

TCAM

*DRAM

Backplane

Packet Manager

Network

PHY

Media

CPU

Thanks to Greg Hankins

Inside a line card

Backplane
Line Interface Cards
Switch Fabric Cards
Management Cards
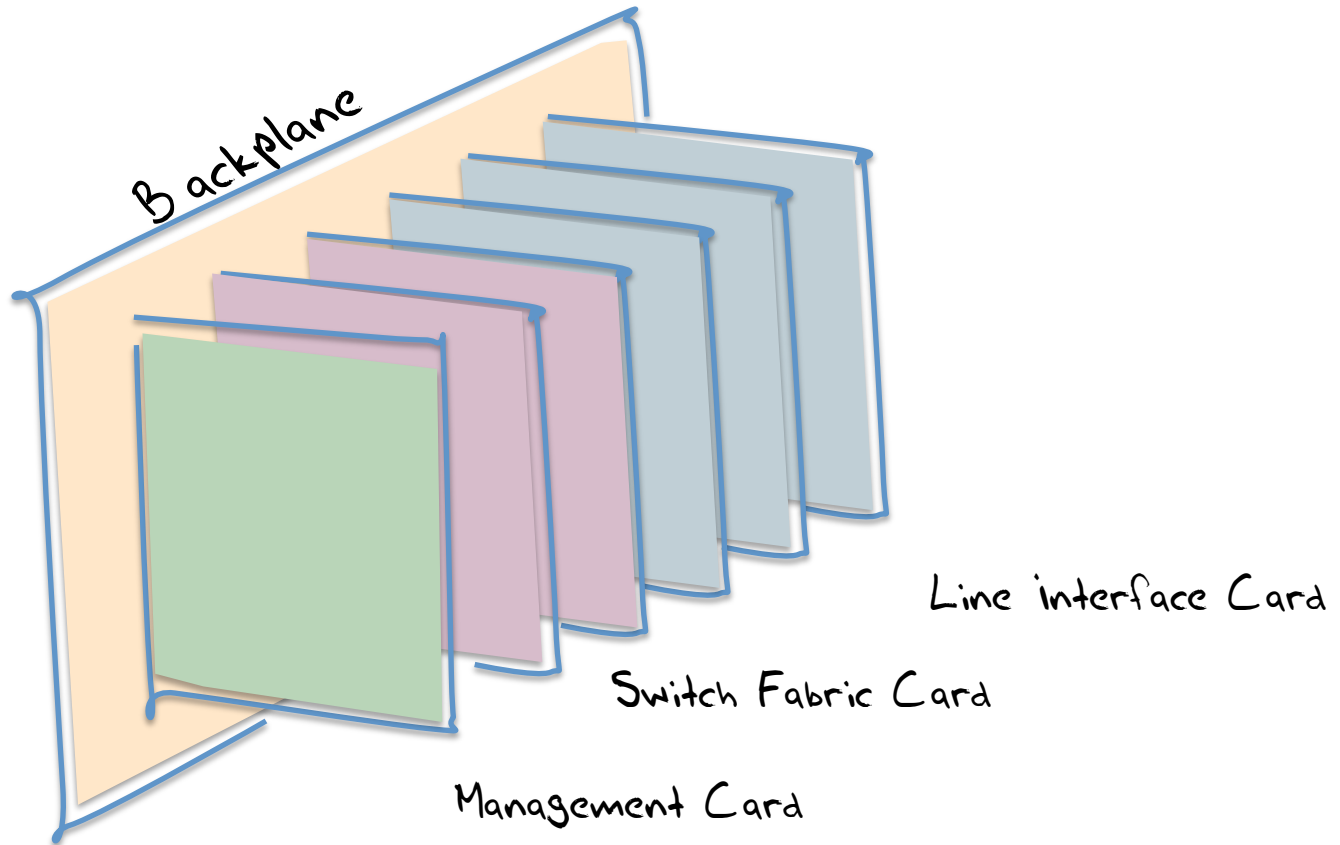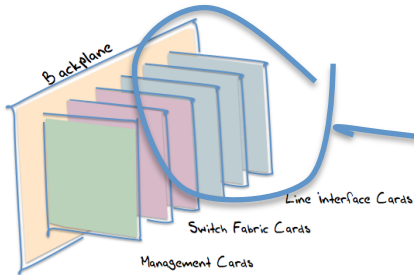
FiB Lookup Bank

Packet Buffer

Backplane

DRAM

TCAM   *DRAM

Packet Manager

Network

PHY

Media

CPU
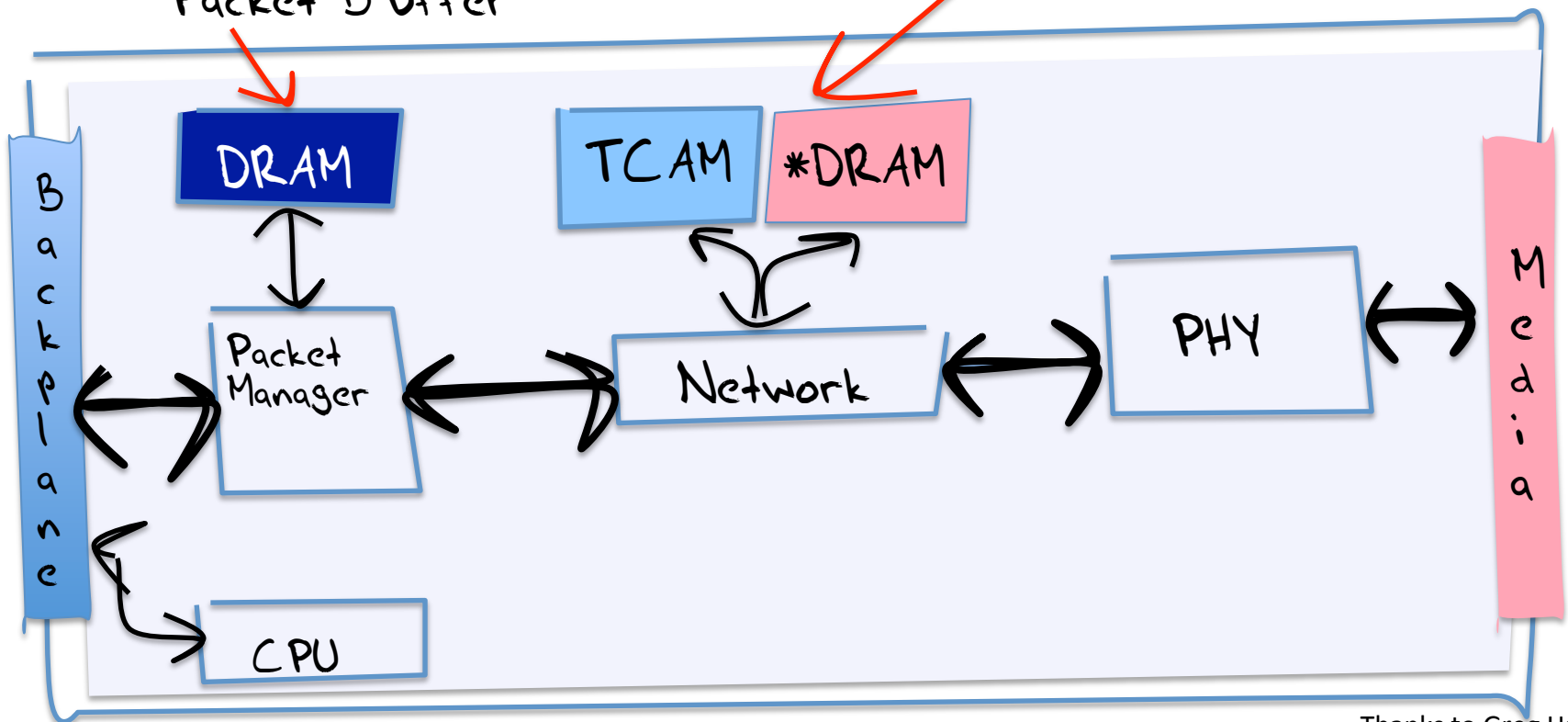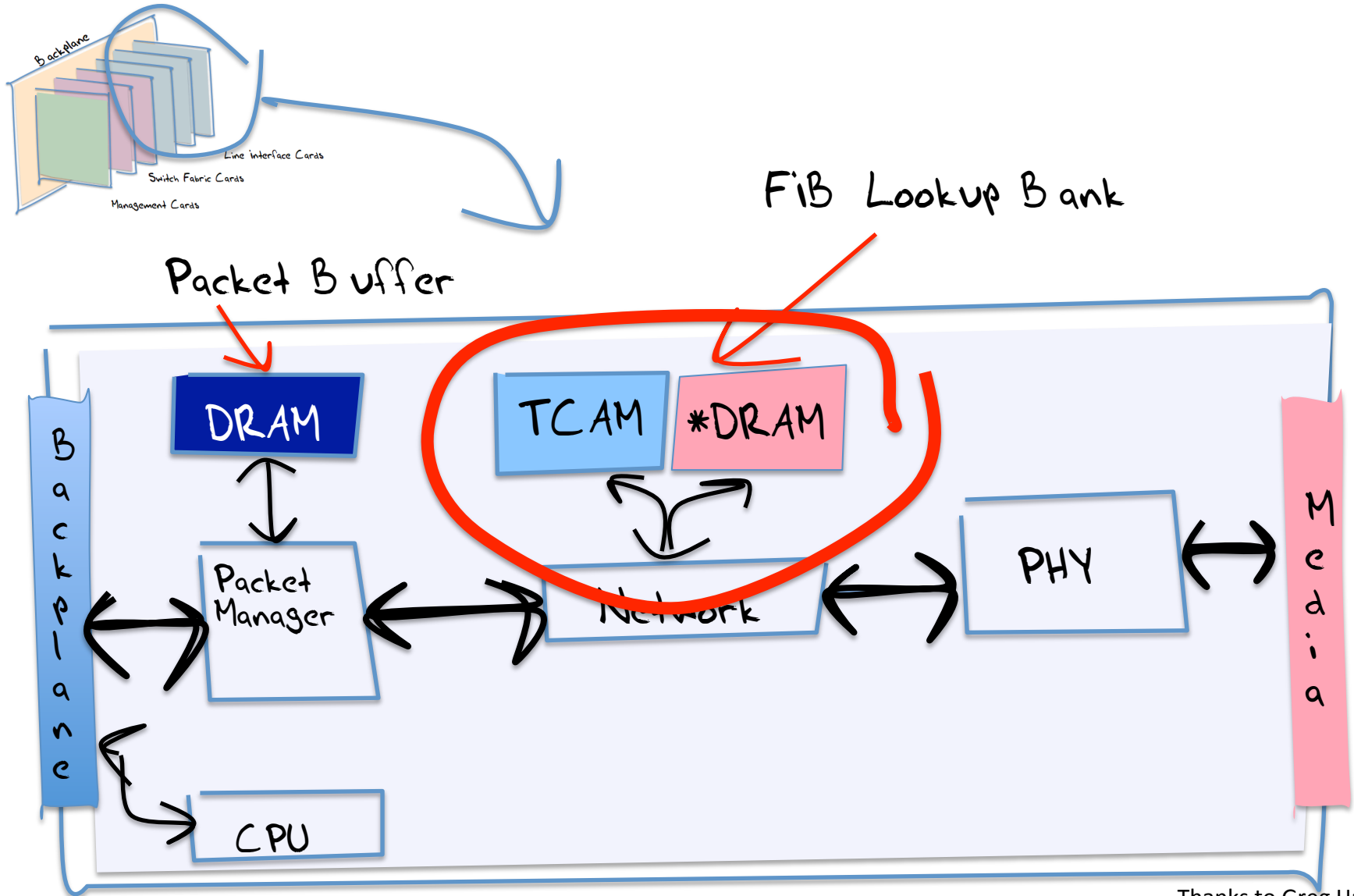
FIB Lookup Memory

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

192.0.2.1

11000000 00000000 00000010 00000001
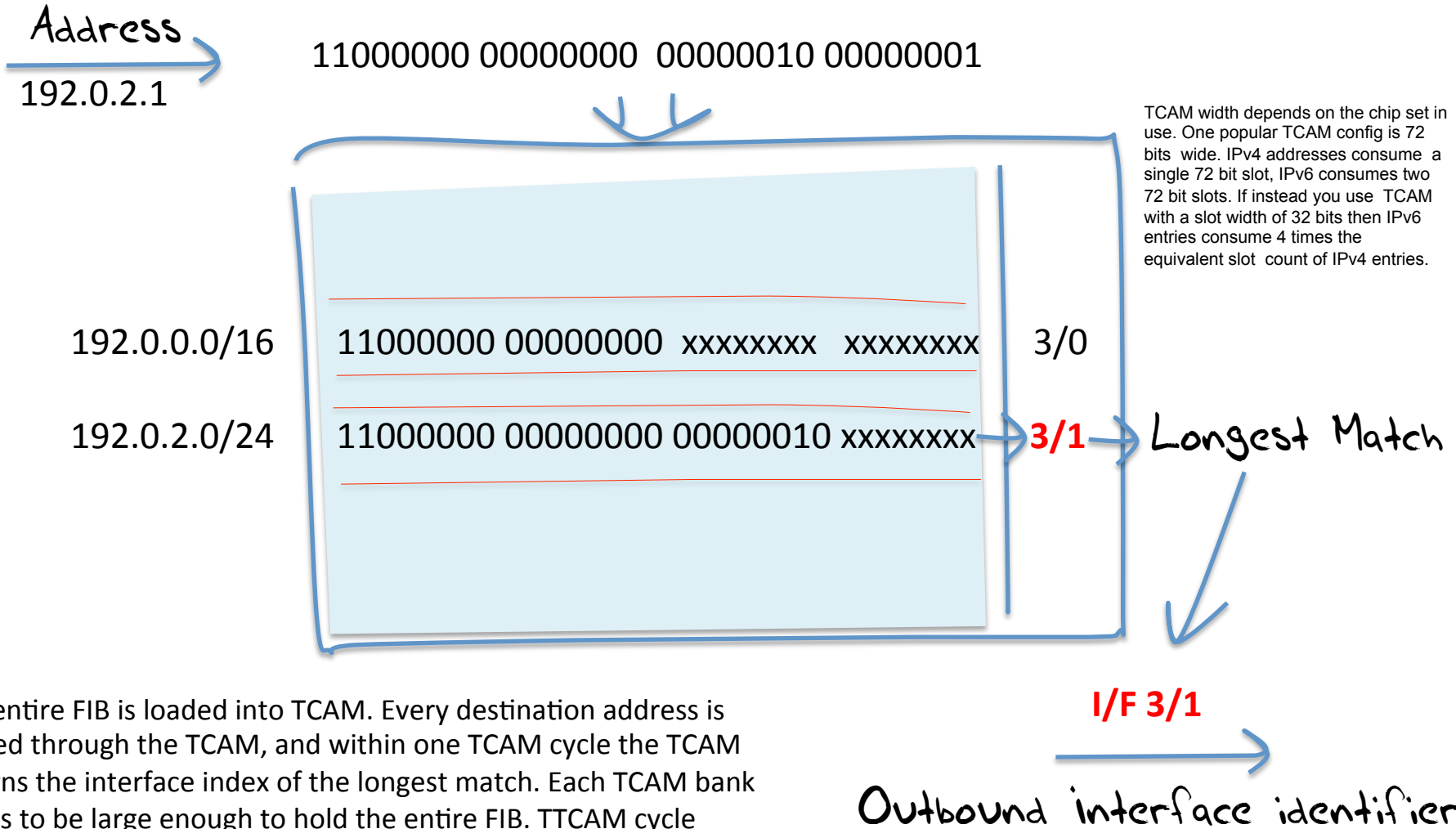
TCAM width depends on the chip set in use. One popular TCAM config is 72 bits wide. IPv4 addresses consume a single 72 bit slot, IPv6 consumes two 72 bit slots. If instead you use TCAM with a slot width of 32 bits then IPv6 entries consume 4 times the equivalent slot count of IPv4 entries.

192.0.0.0/16   11000000 00000000 xxxxxxxx xxxxxxxx   3/0

192.0.2.0/24   11000000 00000000 00000010 xxxxxxxx   **3/1**   Longest Match

I/F 3/1

Outbound interface identifier

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

ASIC

DRAM

1/0
?

1/0
?

1/0
?

1/0
?

1/0
?
x/0000

...
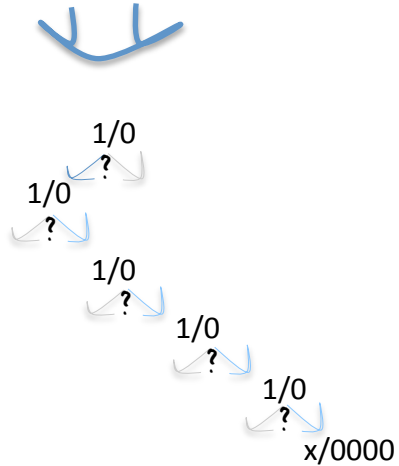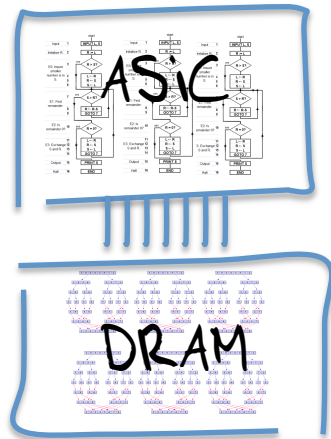
The entire FIB is converted into a serial decision tree. The size of decision tree depends on the distribution of prefix values in the FIB. The performance of the TRIE depends on the algorithm used in the ASIC and the number of serial decisions used to reach a decision

?

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	1Gbit

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 (16+)
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
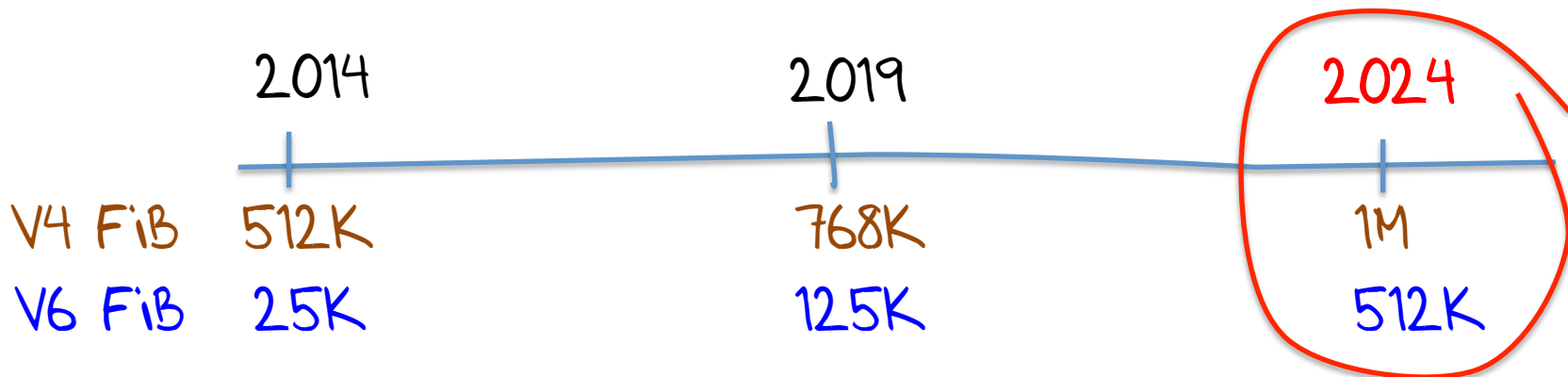
Scaling the FIB

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)

Speed

1Tb

400Gb/1Tb 2017?

1.5Gpps

100Gb

40Gb/100Gb 2010 / 150Mpps

10Gb

10Gb 2002 / 15Mpps

1Gb

1Gb 1999 / 1.5Mpps

100Mb

100Mb 1995 / 150Kpps

10Mb   10Mb 1982/15Kpps
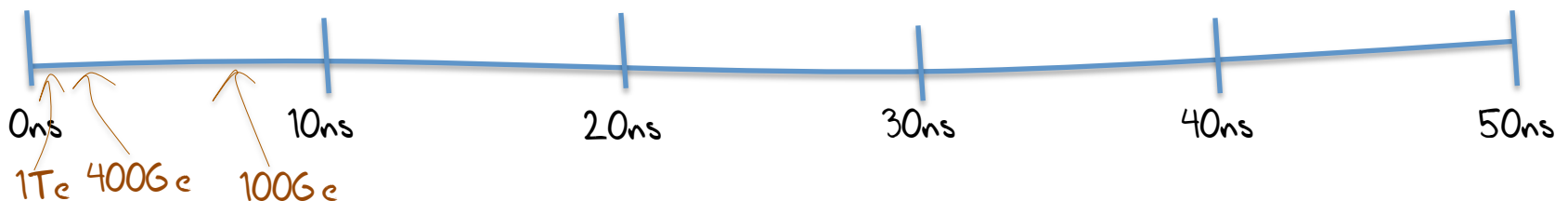
1980   1990   2000   2010   2020
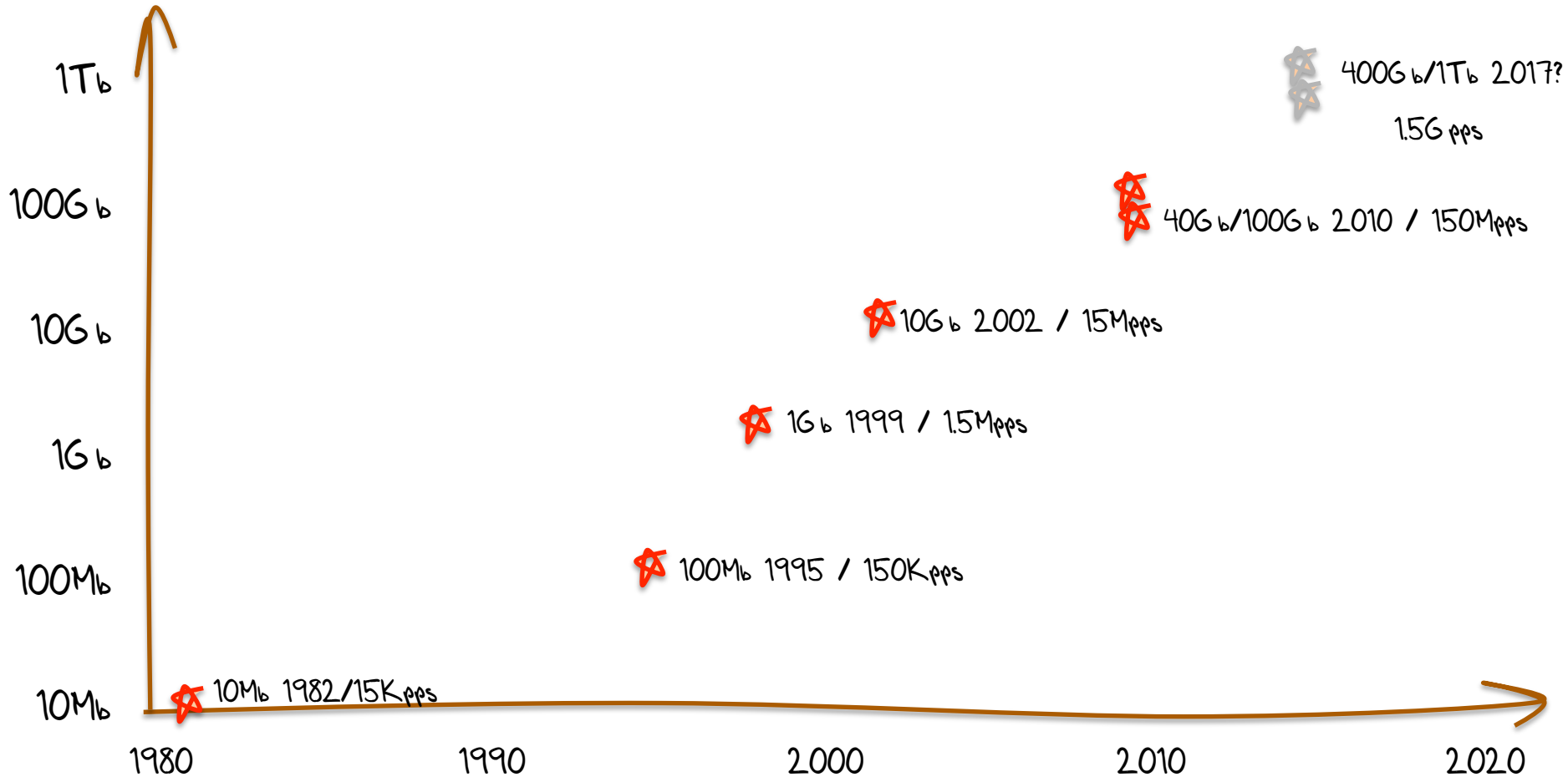
Speed, Speed, Speed

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

$100GE \approx 150Mpps \approx 6.7ns$ per packet

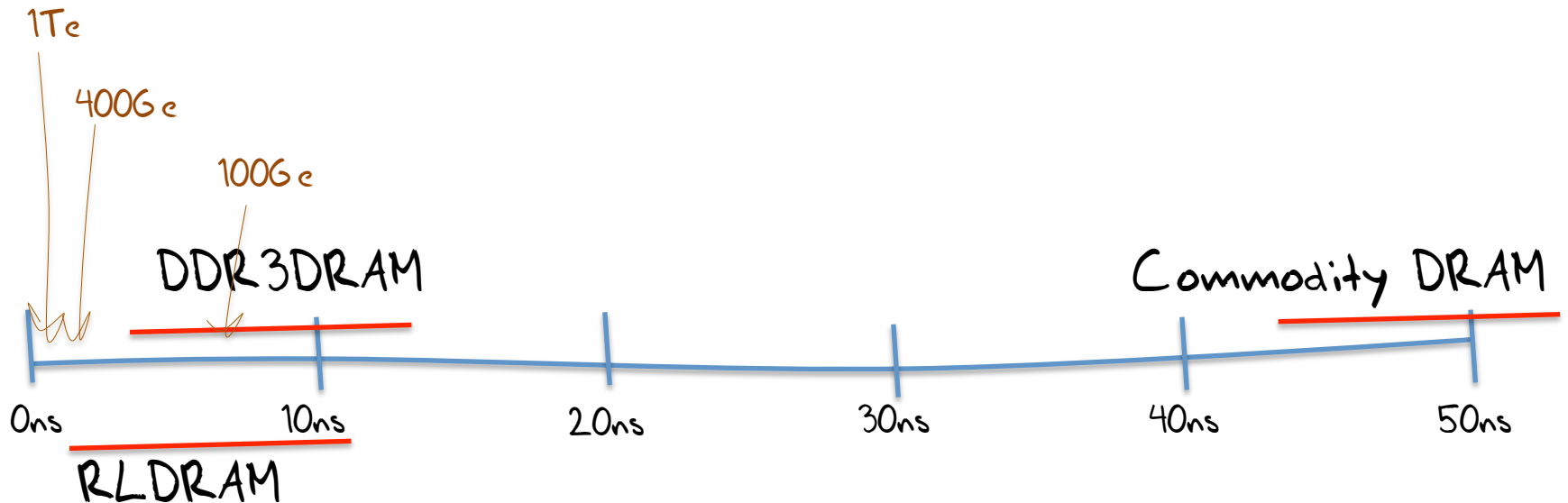$400Ge \approx 600Mpps \approx 1.6ns$ per packet

$1Te \approx 1.5Gpps \approx 0.67ns$ per packet

0ns — 10ns — 20ns — 30ns — 40ns — 50ns

1Te 400Ge 100Ge

Speed, Speed, Speed
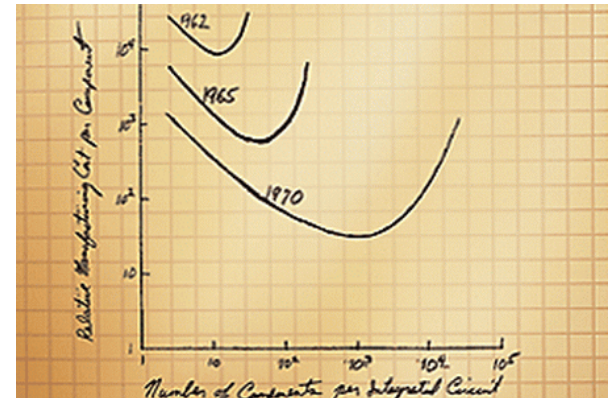
What memory speeds do we HAVE?

Scaling Speed

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 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?

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

Thank You

Questions?