Routing 2014

Geoff Huston
APNIC

Looking through the Routing Lens

Looking through the Routing Lens

There are very few ways to assemble a single view of the entire Internet

The lens of routing is one of the ways in which information relating to the entire reachable Internet is bought together

Even so, its not a perfect lens…

There is no Routing God!

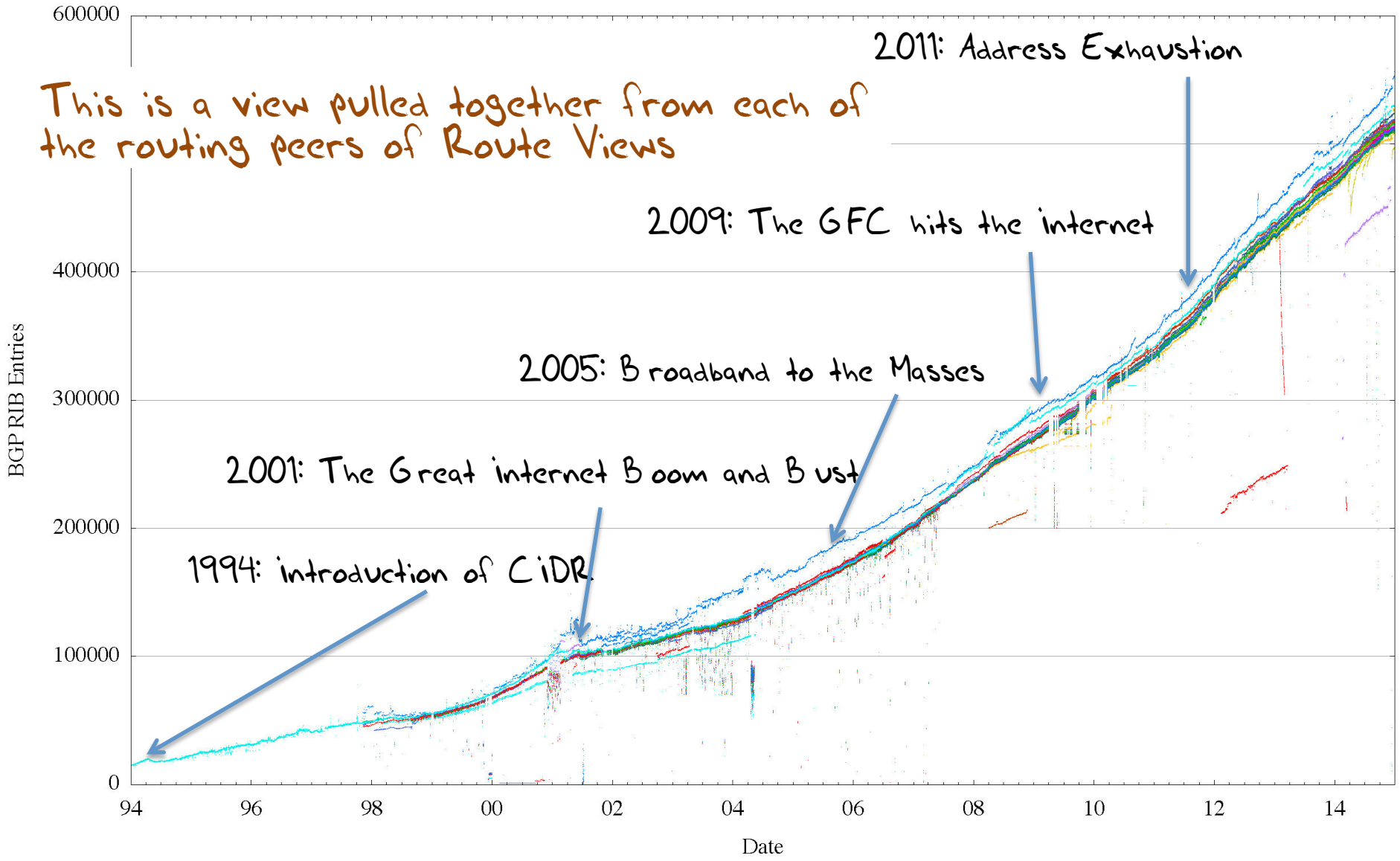
There is no single objective “out of the system” view of the Internet’s Routing environment.

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

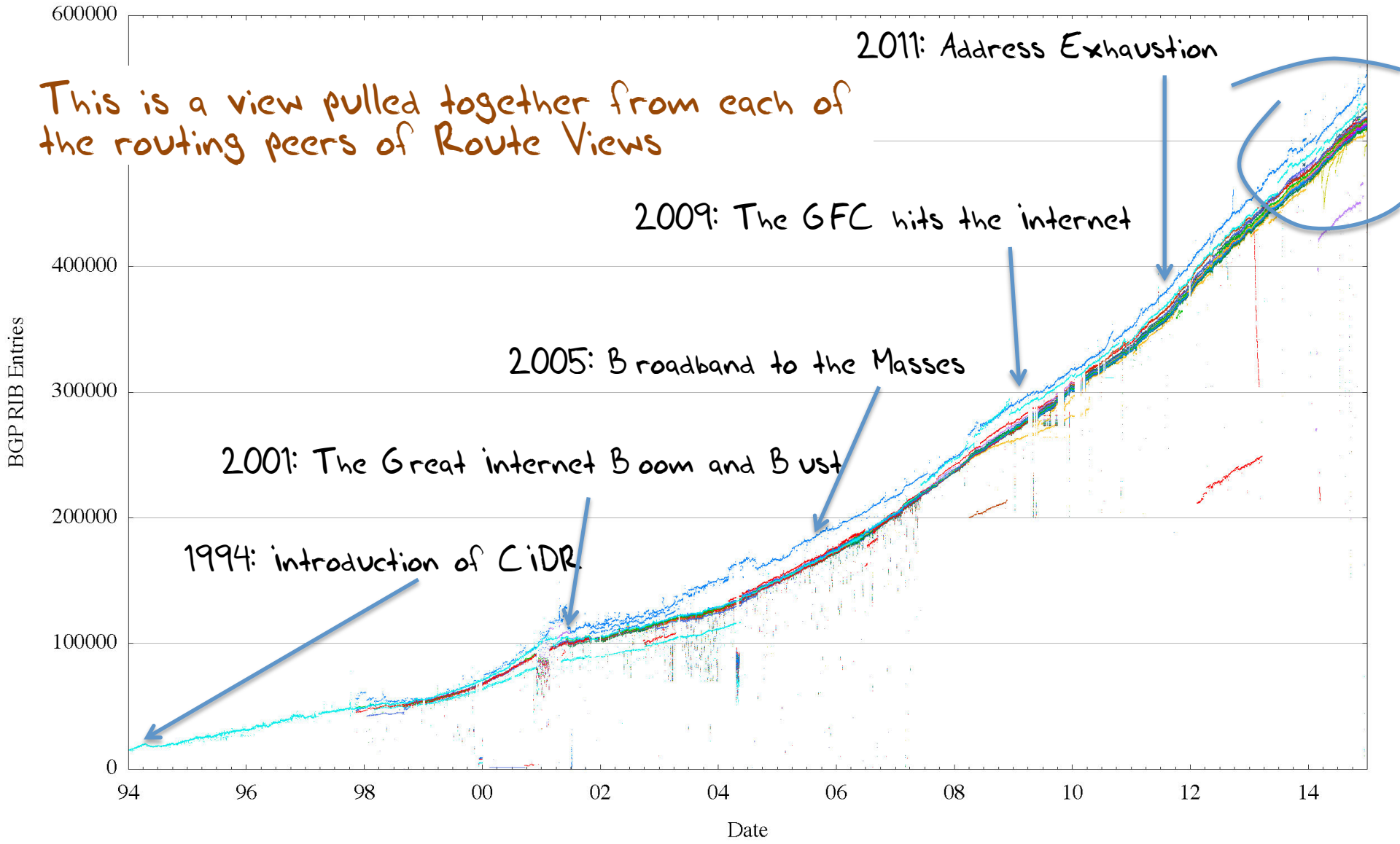
So the picture I will be painting here is one that is drawn from the perspective of AS131072. This is a stub AS at the edge of the Internet.

You may or may not have a similar view from your network.

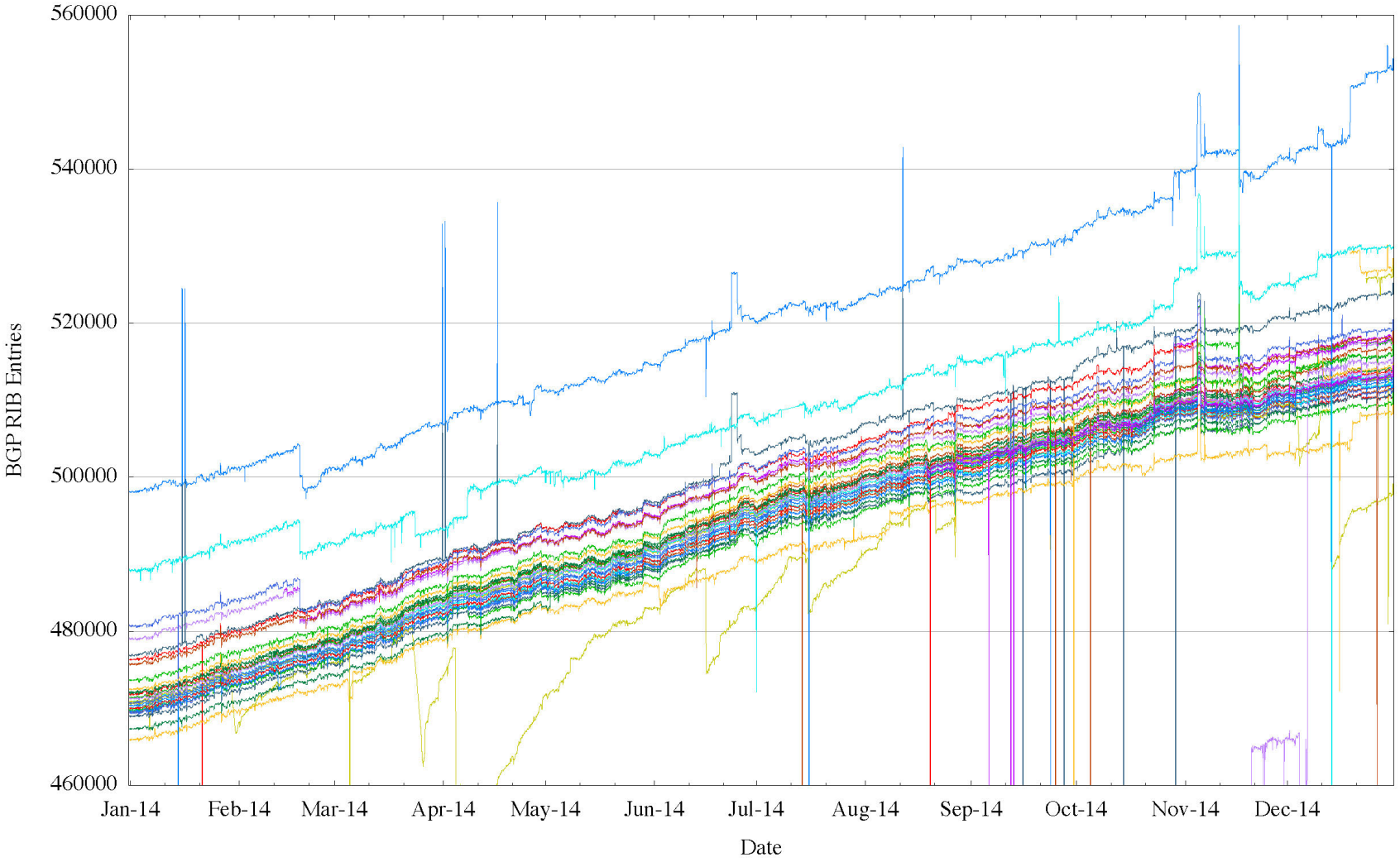
20 Years of Routing the Internet



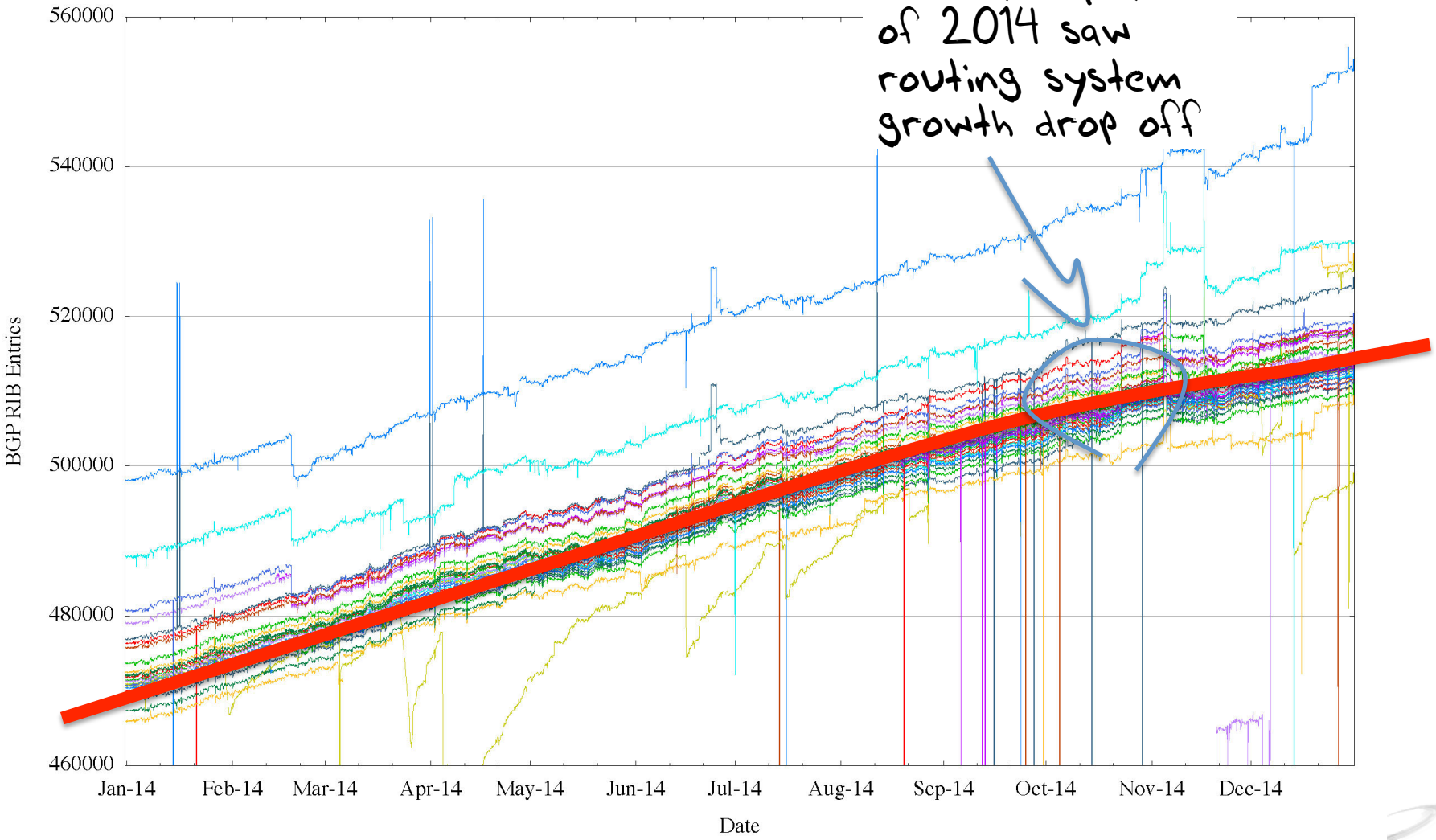
20 Years of Routing the Internet



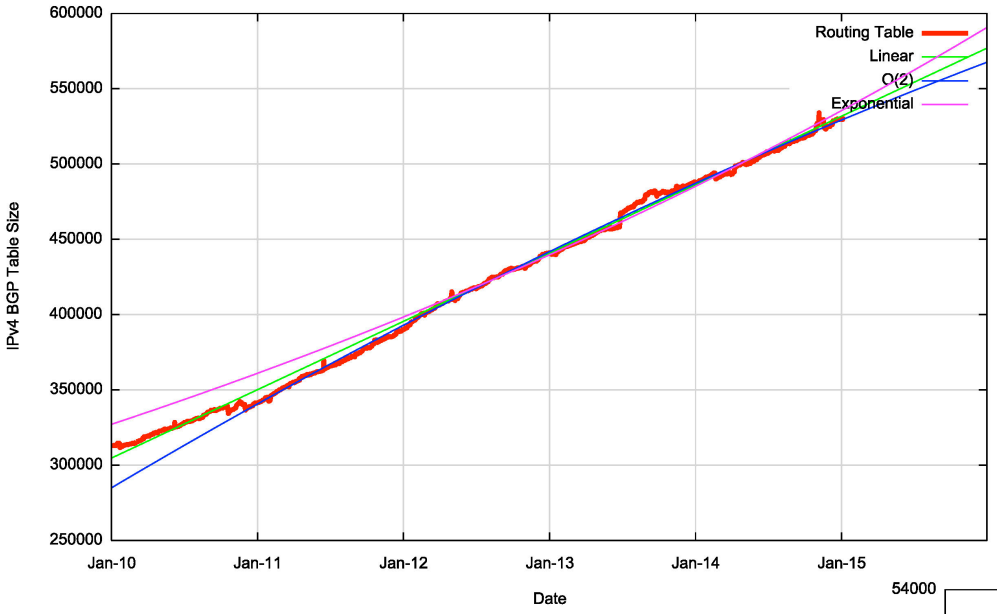
2014, as seen at Route Views



2014, as seen at Route Views

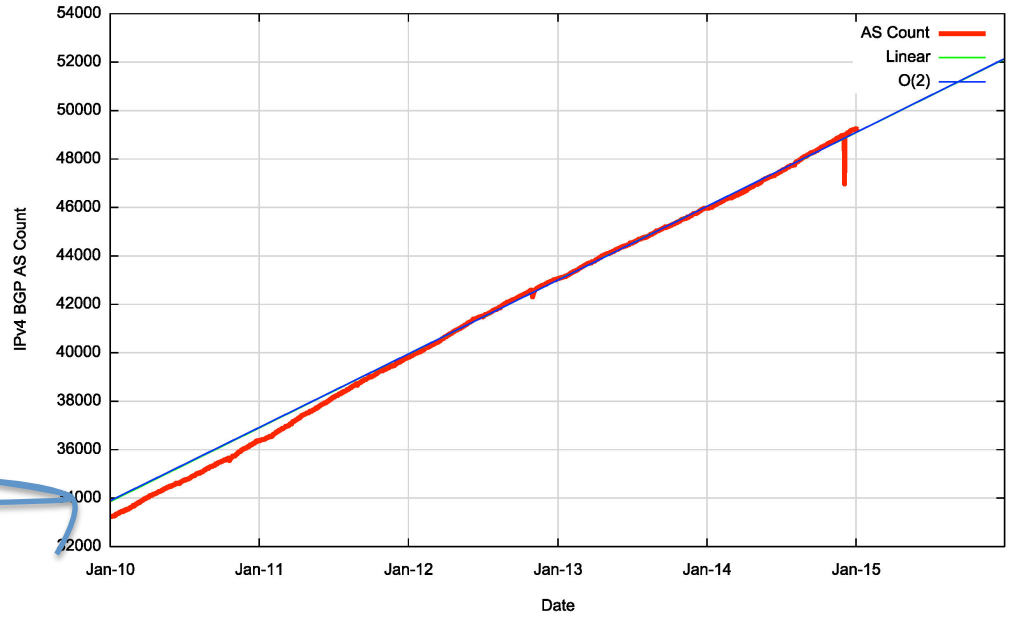


Routing Indicators for IPv4

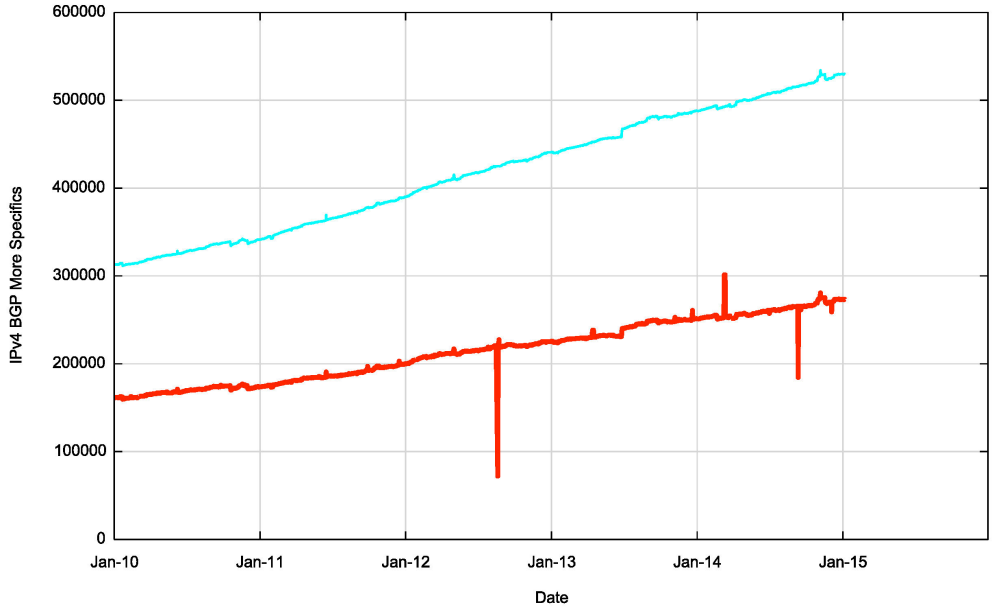


Routing prefixes - growing by some 45,000 prefixes per year

AS Numbers- growing by some 3,000 prefixes per year

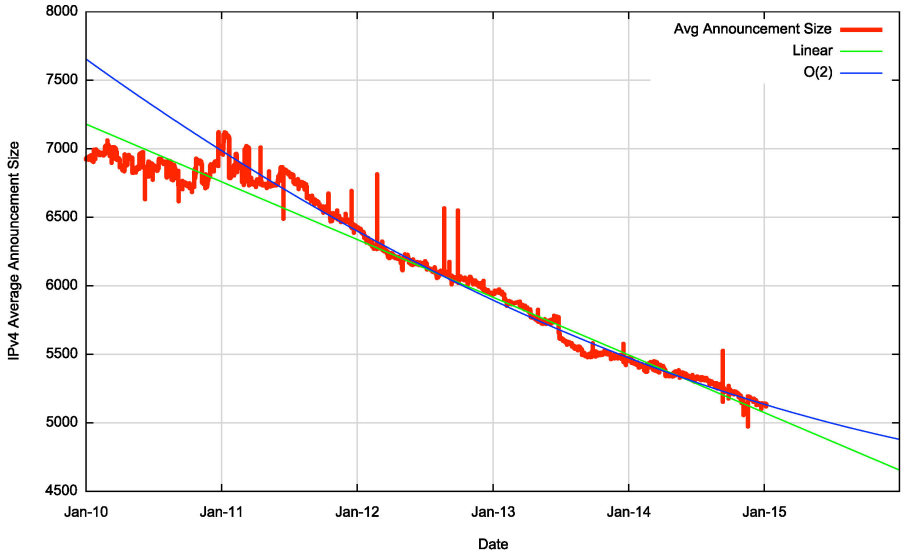


Routing Indicators for IPv4

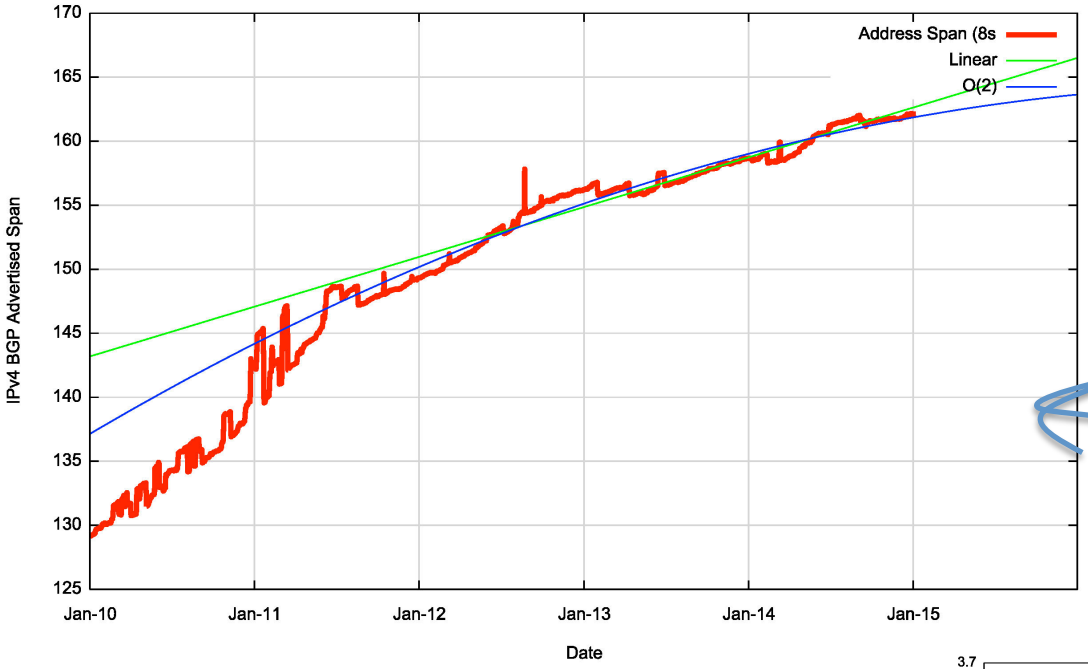


More Specifics are still taking up one half of the routing table

But the average size of a routing advertisement is getting smaller

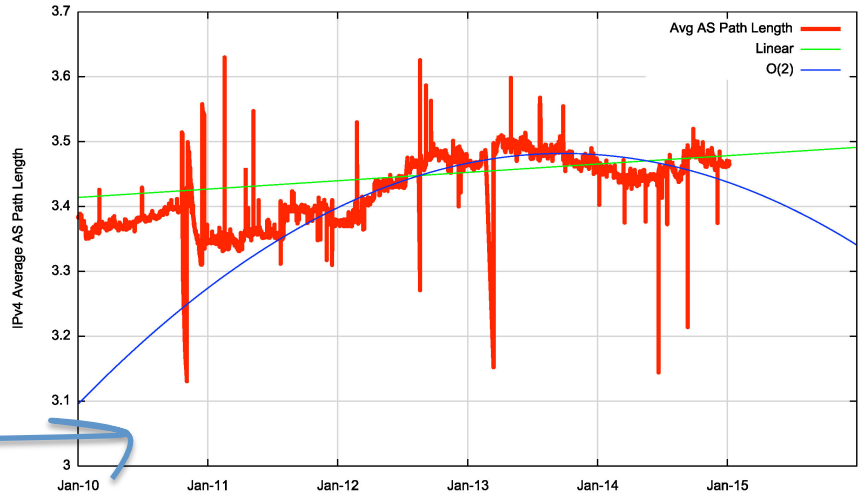


Routing Indicators for IPv4



Address Exhaustion is now visible in the extent of advertised address space

The "shape" of inter-AS interconnection appears to be steady, as the Average AS Path length has been held steady through the year



What happened in 2014 in V4?

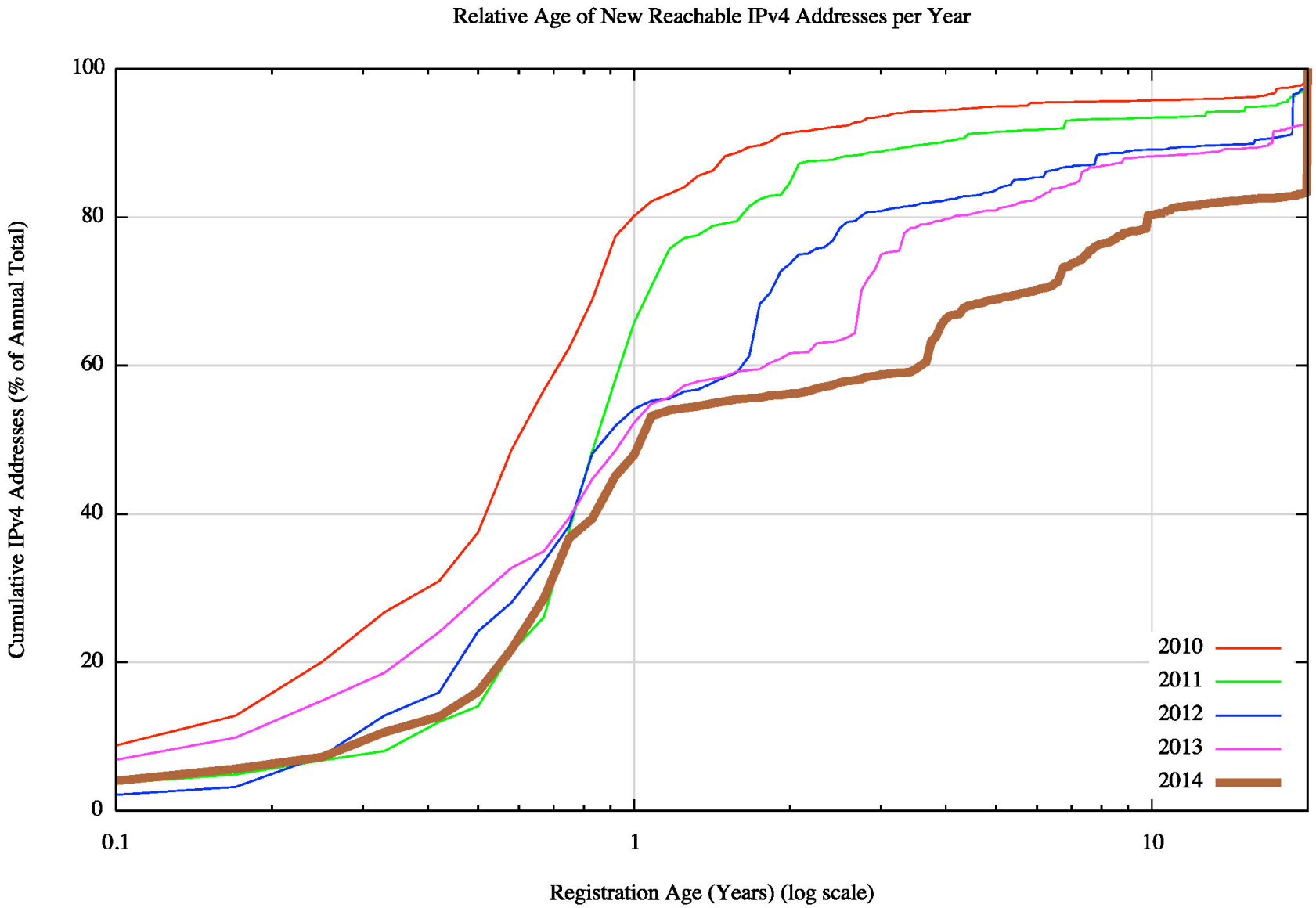
- From the look of the growth plots, its business as usual, despite the increasing pressure on IPv4 address availability
- You may have noticed that the number of IPv4 routes cross across the threshold value of 512,000 routes in the last quarter of 2014
  - And for some routers this would've caused a hiccup or two
- You can also see that the pace of growth of the routing table is dropping off towards the end of the year
  - IPv4 address exhaustion is probably to blame here!

How can the IPv4 network continue to grow when we are running out of IPv4 addresses?

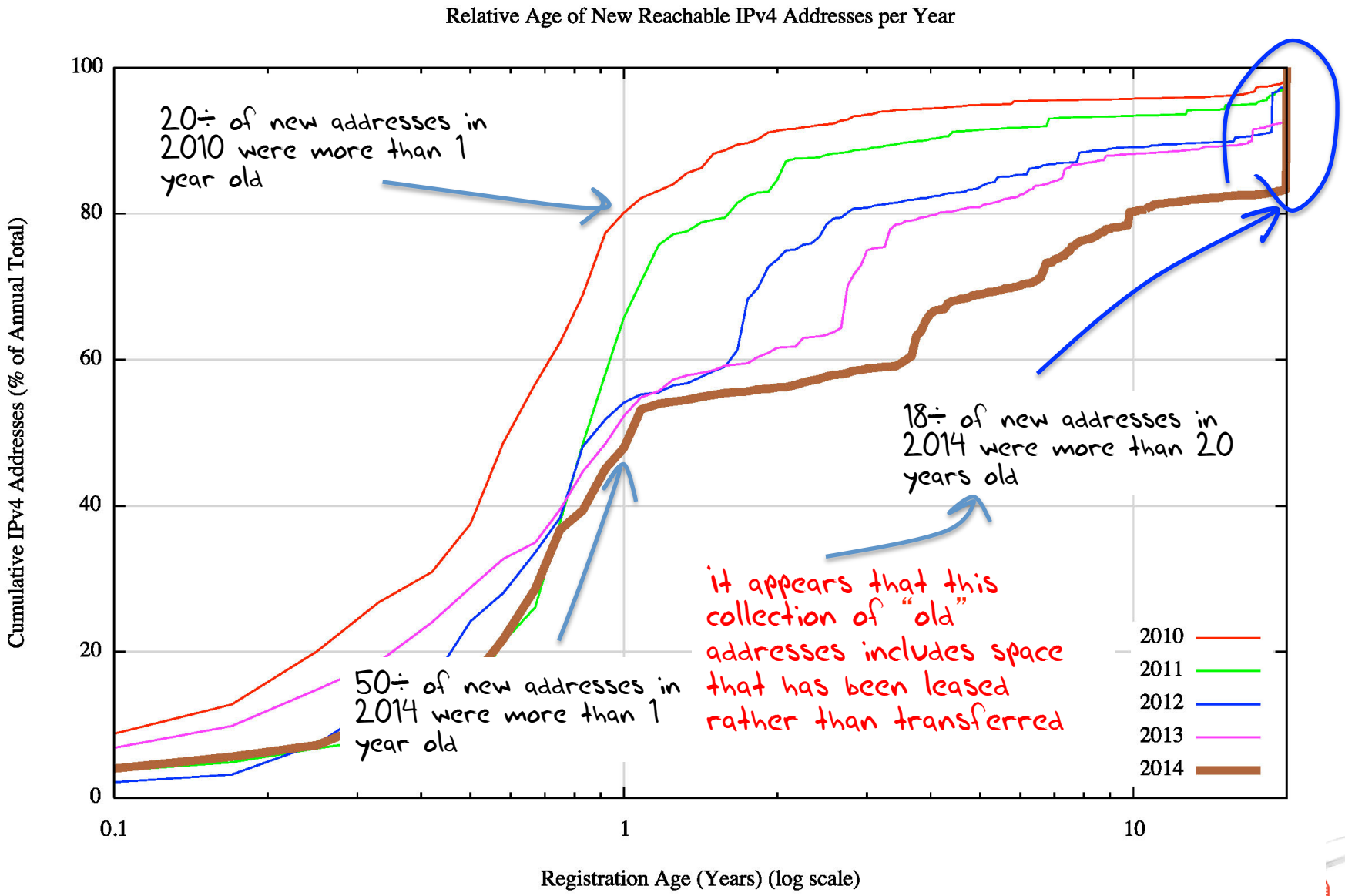
We are now recycling old addresses back into the routing system

Some of these addresses are transferred in ways that are recorded in the registry system, while others are being “leased” without any clear registration entry that describes the lessee

IPv4 Address Reuse



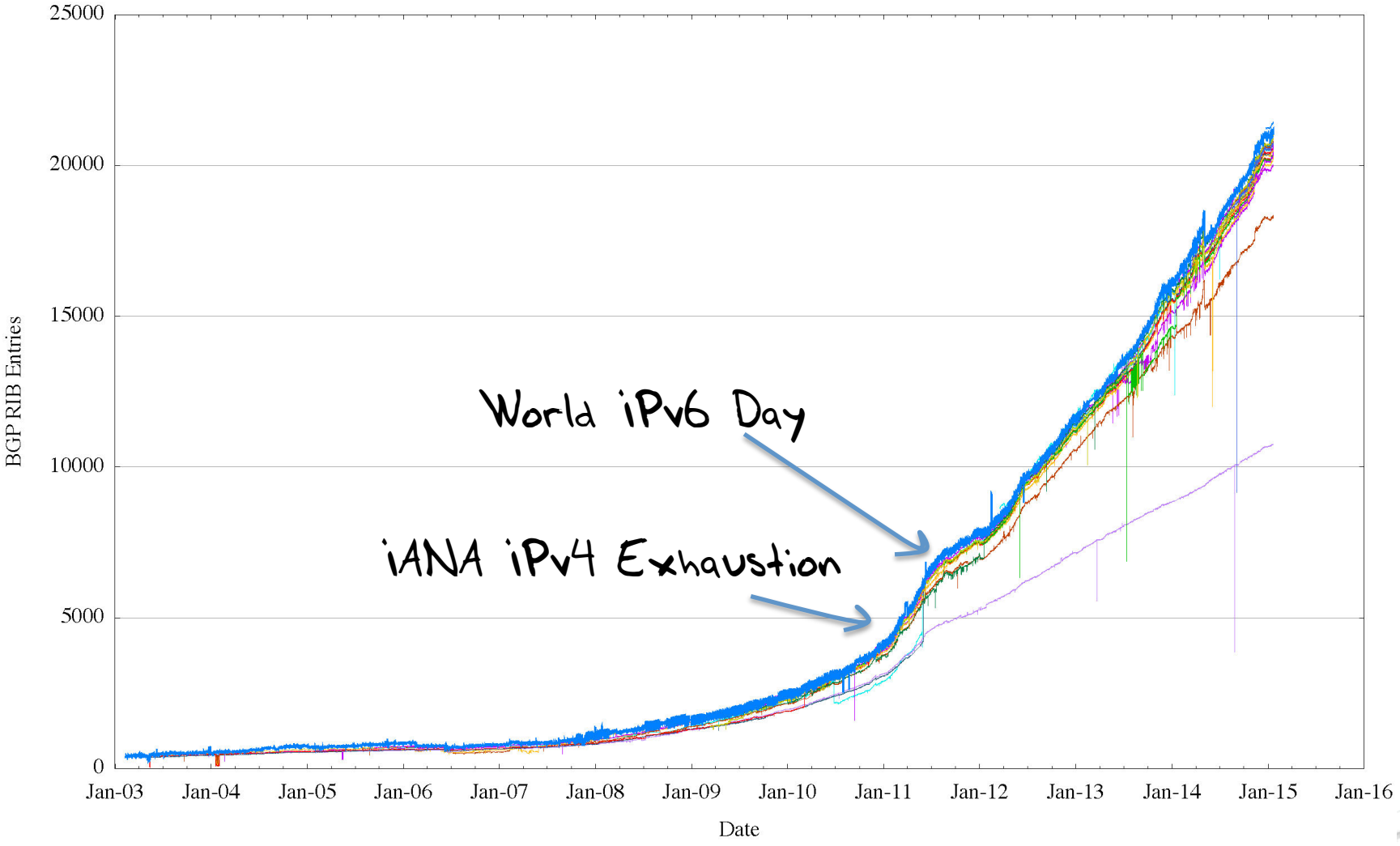
IPv4 Address Reuse



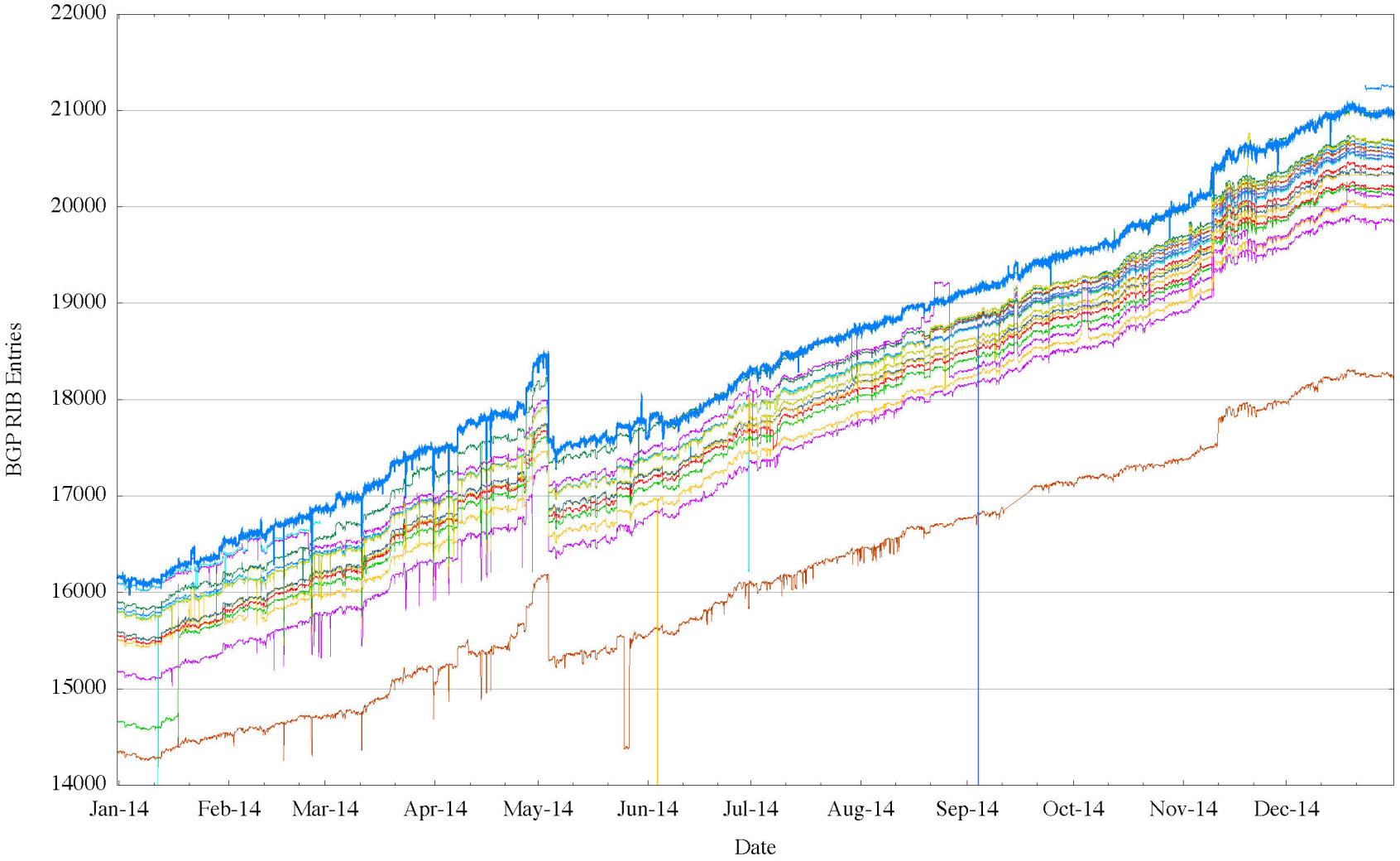
IPv4 in 2014 - Growth is Slowing

- 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 ’14 caused a visible blip in the address consumption rate)
- The rate of growth of the IPv4 Internet is slowing down, due to:
  - Address shortages
  - Masking by NAT deployments
  - Saturation of critical market sectors
  - Transition uncertainty

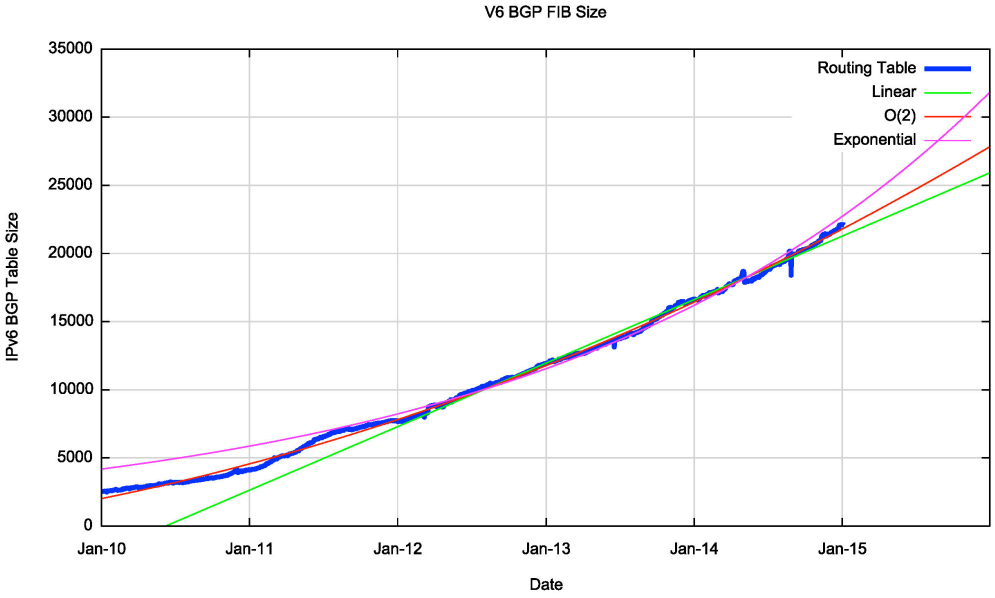
The Route Views view of IPv6



2014 for IPv6, as seen at Route Views

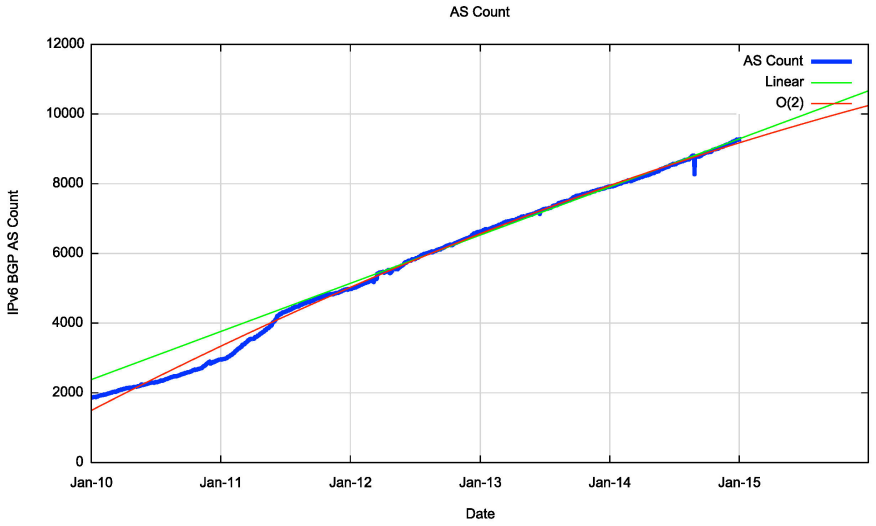


Routing Indicators for IPv6

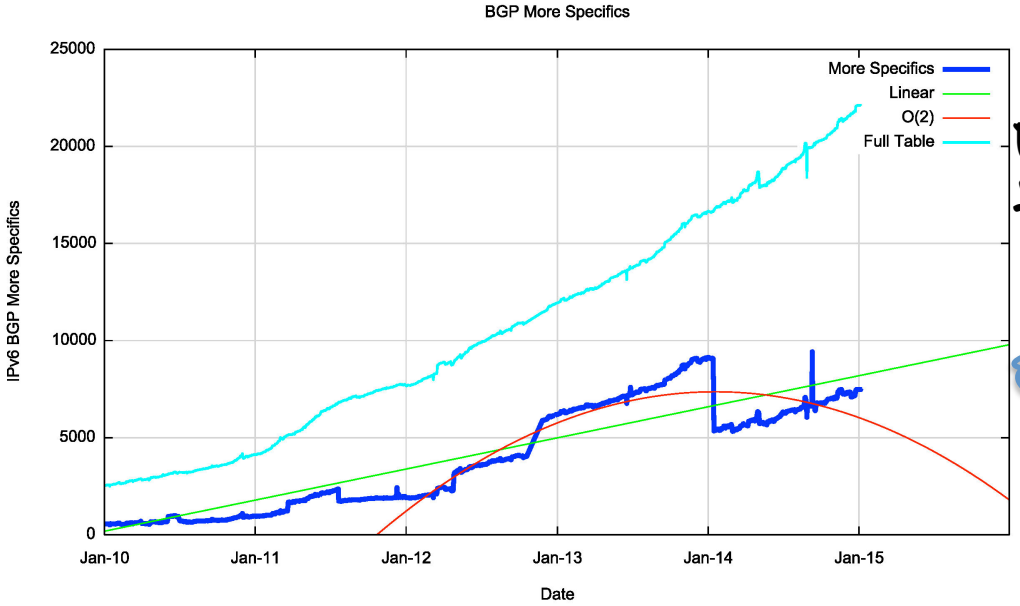


Routing prefixes - growing by some 6,000 prefixes per year

AS Numbers- growing by some 1,600 prefixes per year (which is half the V4 growth)

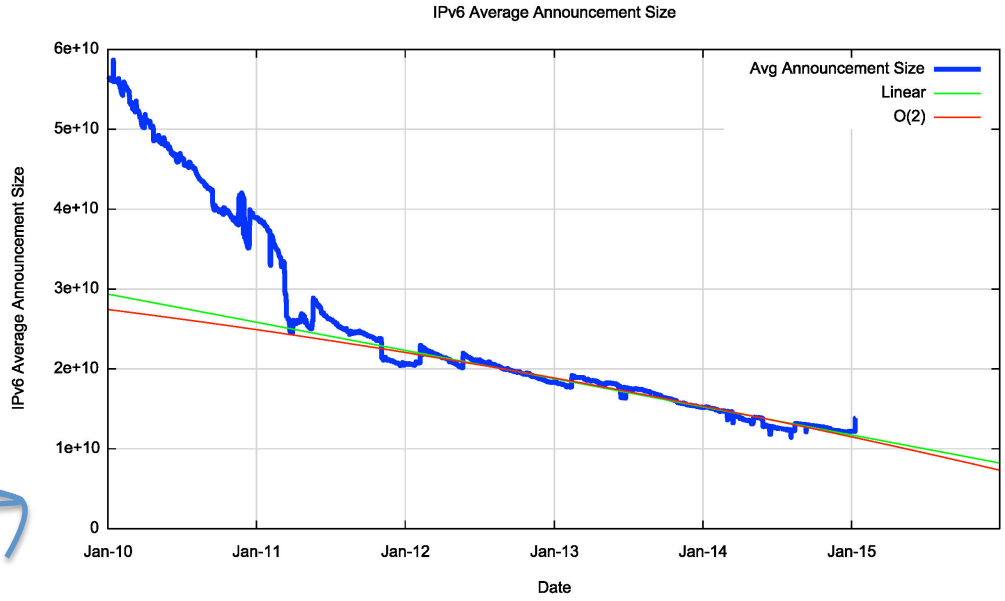


Routing Indicators for IPv6

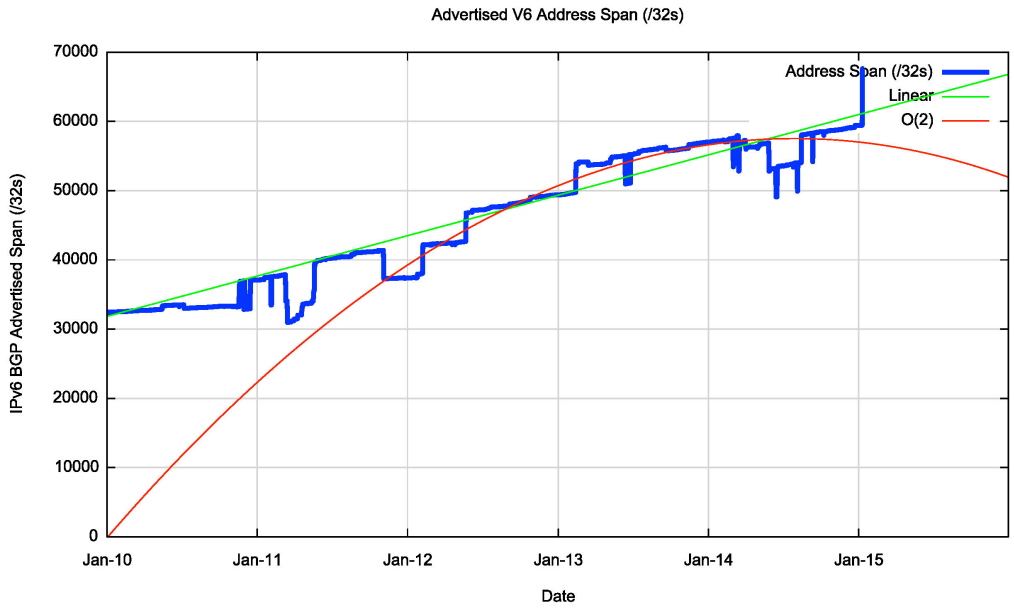


More Specifics now take up one third of the routing table

The average size of a routing advertisement is getting smaller

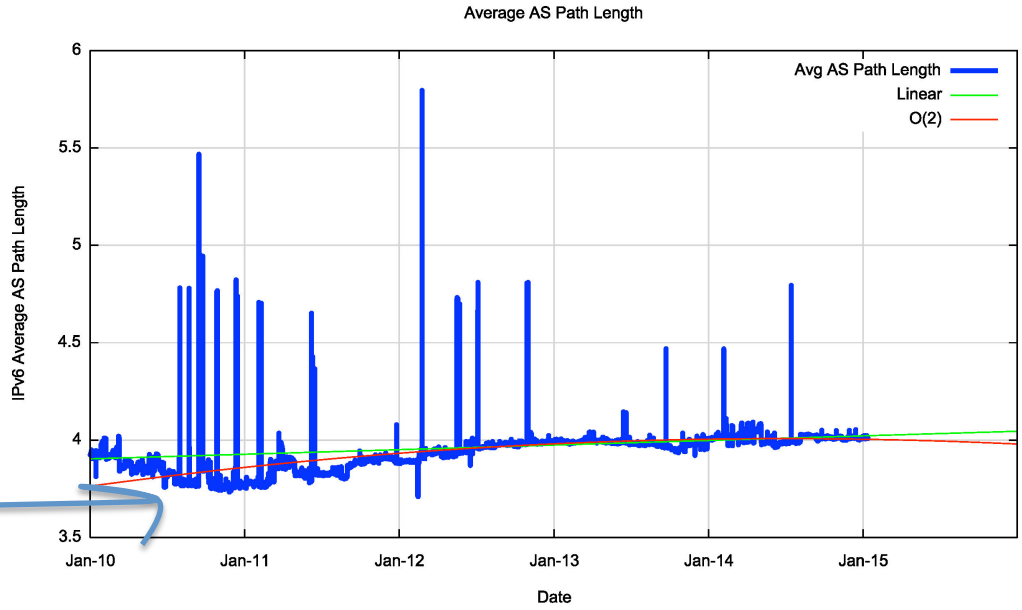


Routing Indicators for IPv6



Address consumption is happening at a constant rate, and not growing year by year

The "shape" of inter-AS interconnection appears to be steady, as the Average AS Path length has been held steady through the year



IPv6 in 2014

- 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

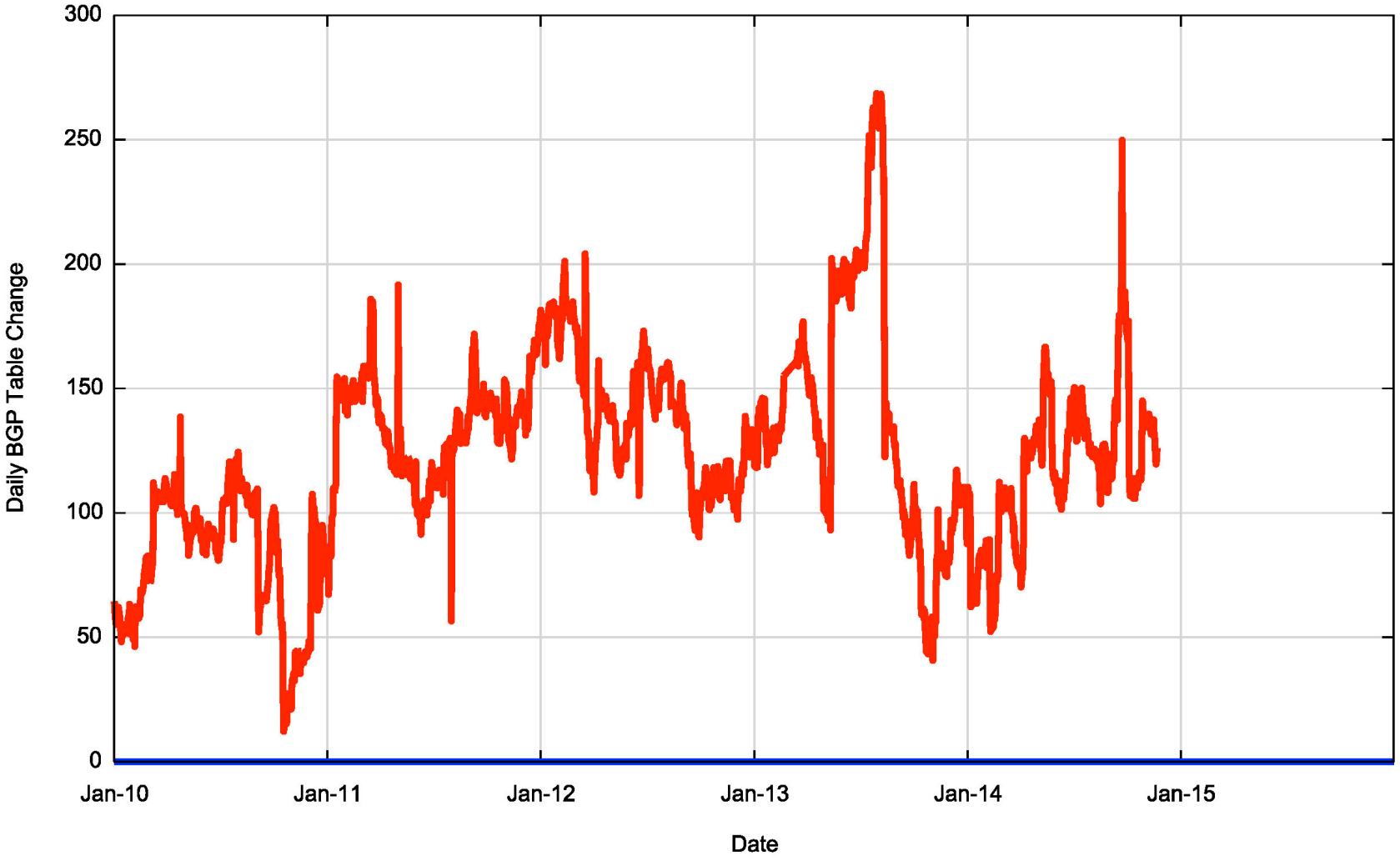
BGP Size Projections

For the Internet this is a time of **extreme uncertainty**

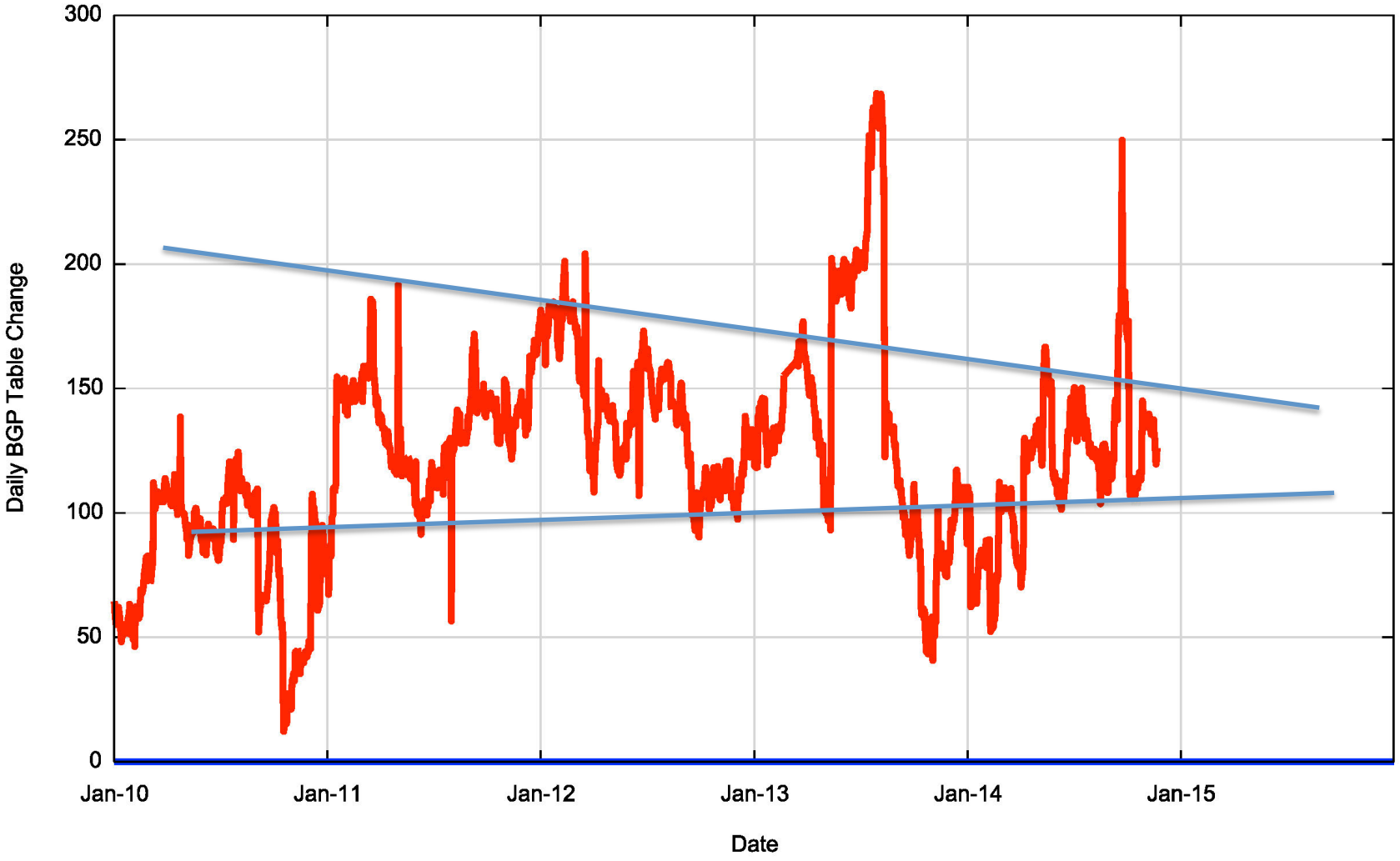
- Registry IPv4 address run out
- Uncertainty over the impacts of any after-market in IPv4 on the routing table
- Uncertainty over IPv6 takeup leads to a mixed response to IPv6 so far, and no clear indicator of trigger points for change

all of which which make this year's projection even more speculative than normal!

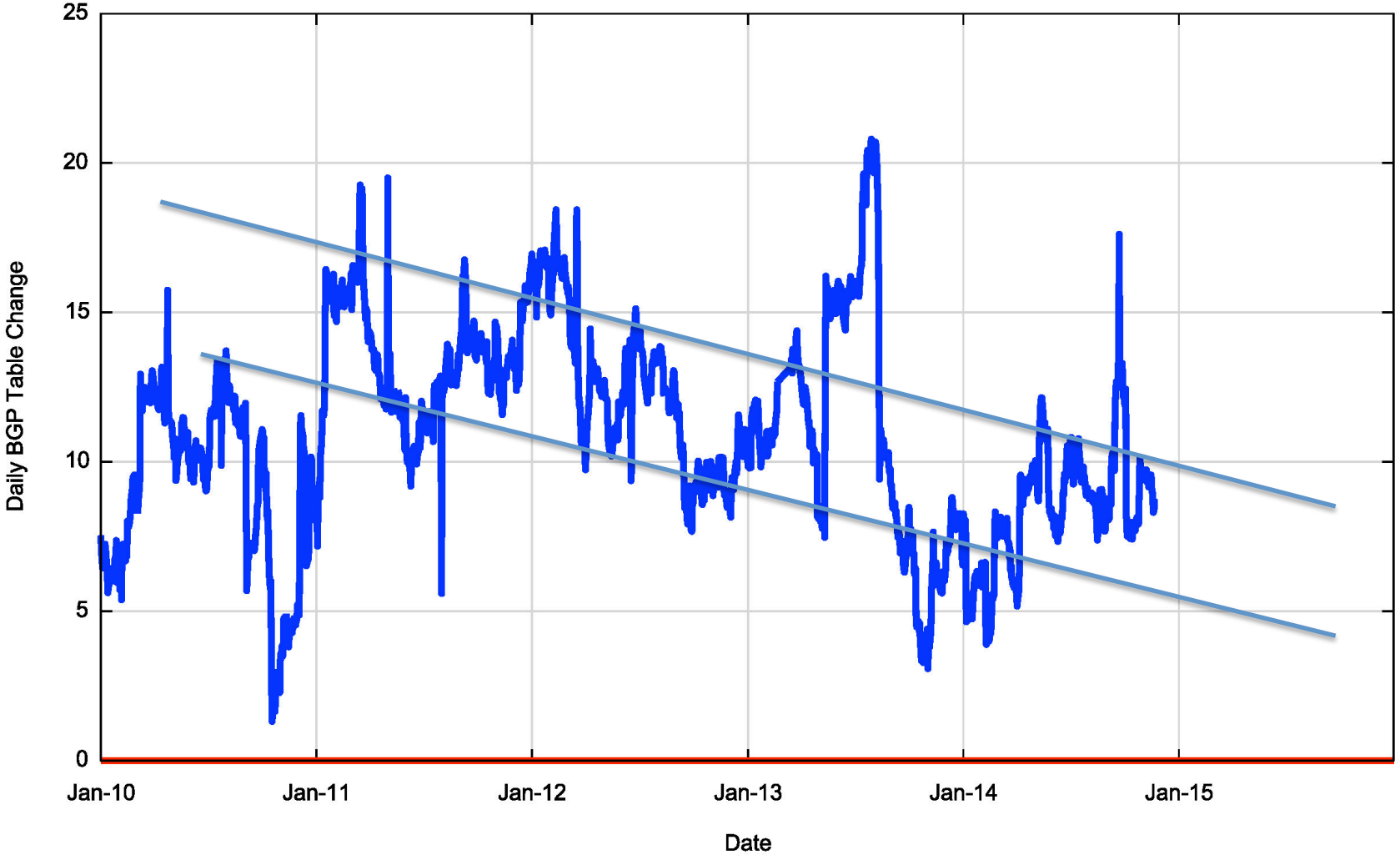
V4 - Daily Growth Rates



V4 - Daily Growth Rates



V4 - Relative Daily Growth Rates

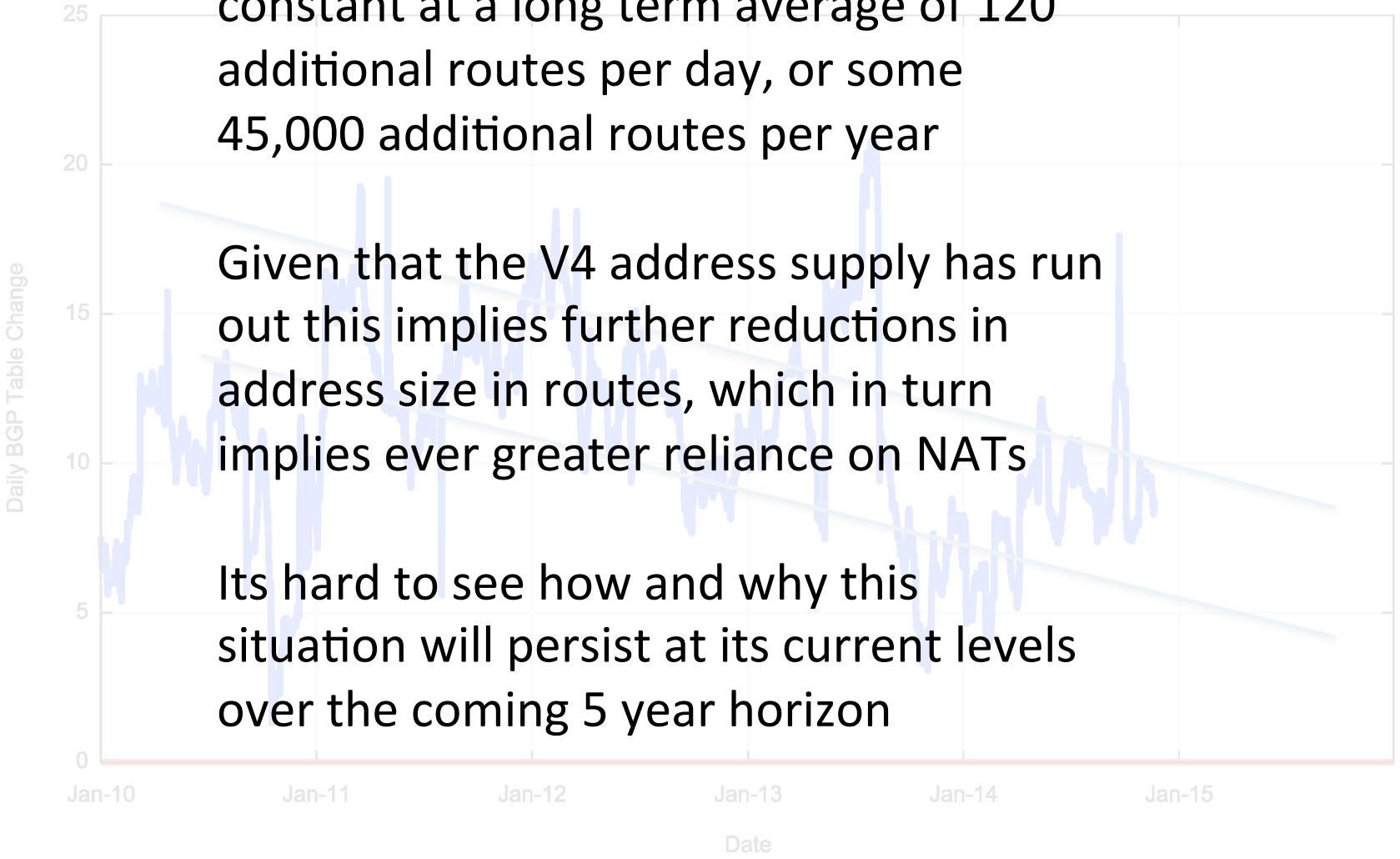


V4 - Relative Daily Growth Rates

Growth in the V4 network appears to be constant at a long term average of 120 additional routes per day, or some 45,000 additional routes per year

Given that the V4 address supply has run out this implies further reductions in address size in routes, which in turn implies ever greater reliance on NATs

Its hard to see how and why this situation will persist at its current levels over the coming 5 year horizon

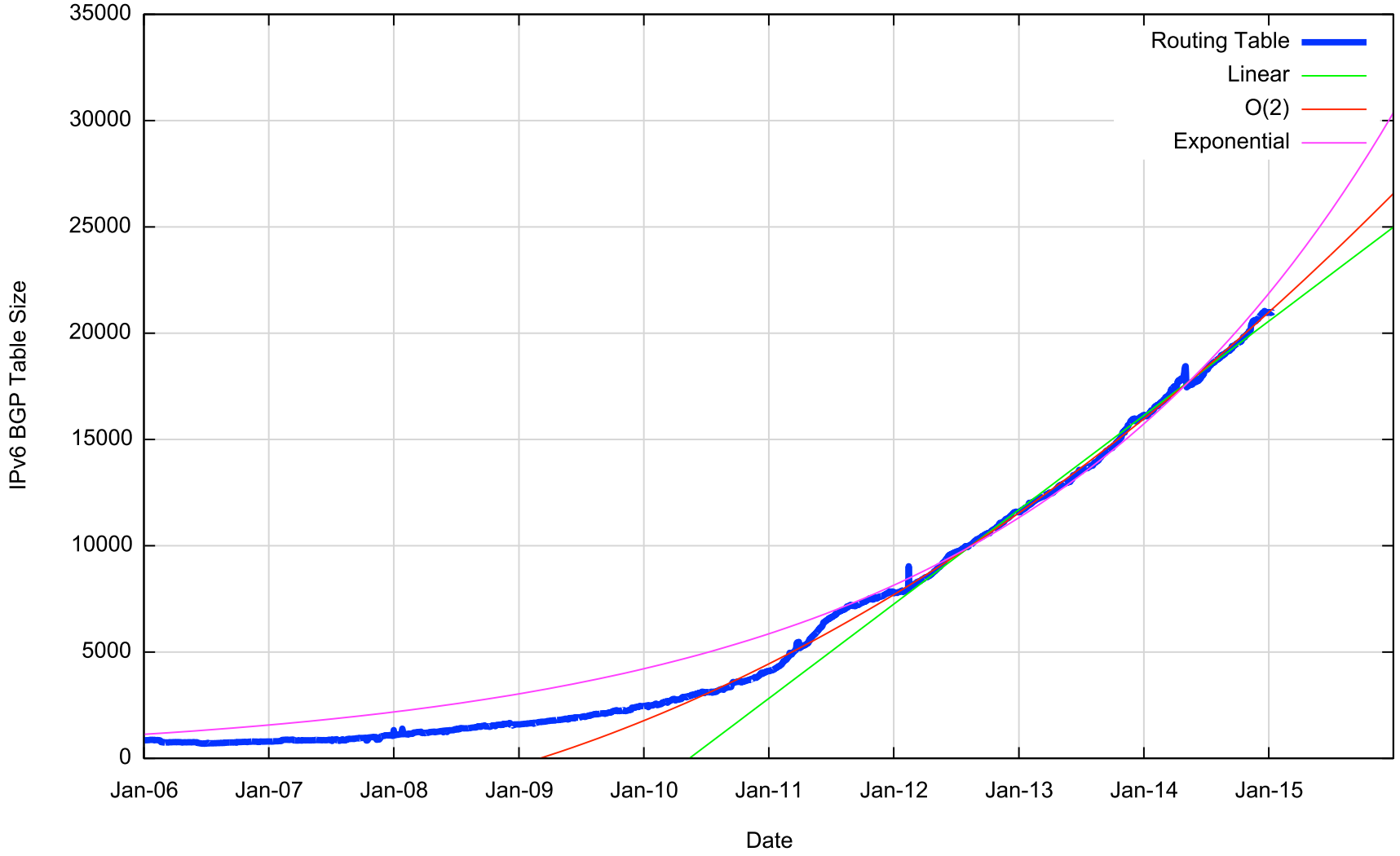


IPv4 BGP Table Size predictions

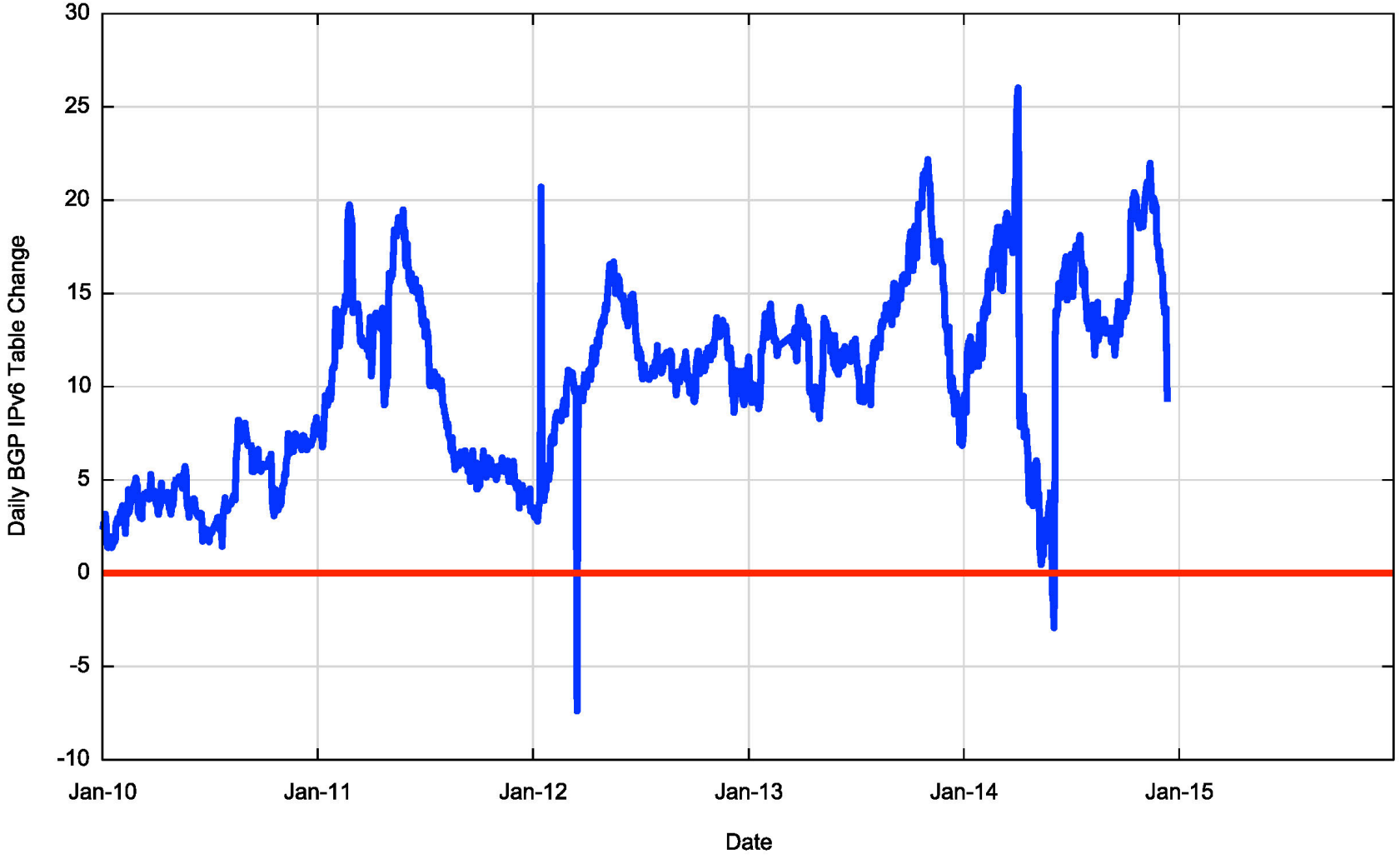
Jan 2013	441,000 entries
2014	488,000
2015	*530,000*
2016	*580,000*
2017	*620,000*
2018	*670,000*
2019	*710,000*
2020	*760,000*

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

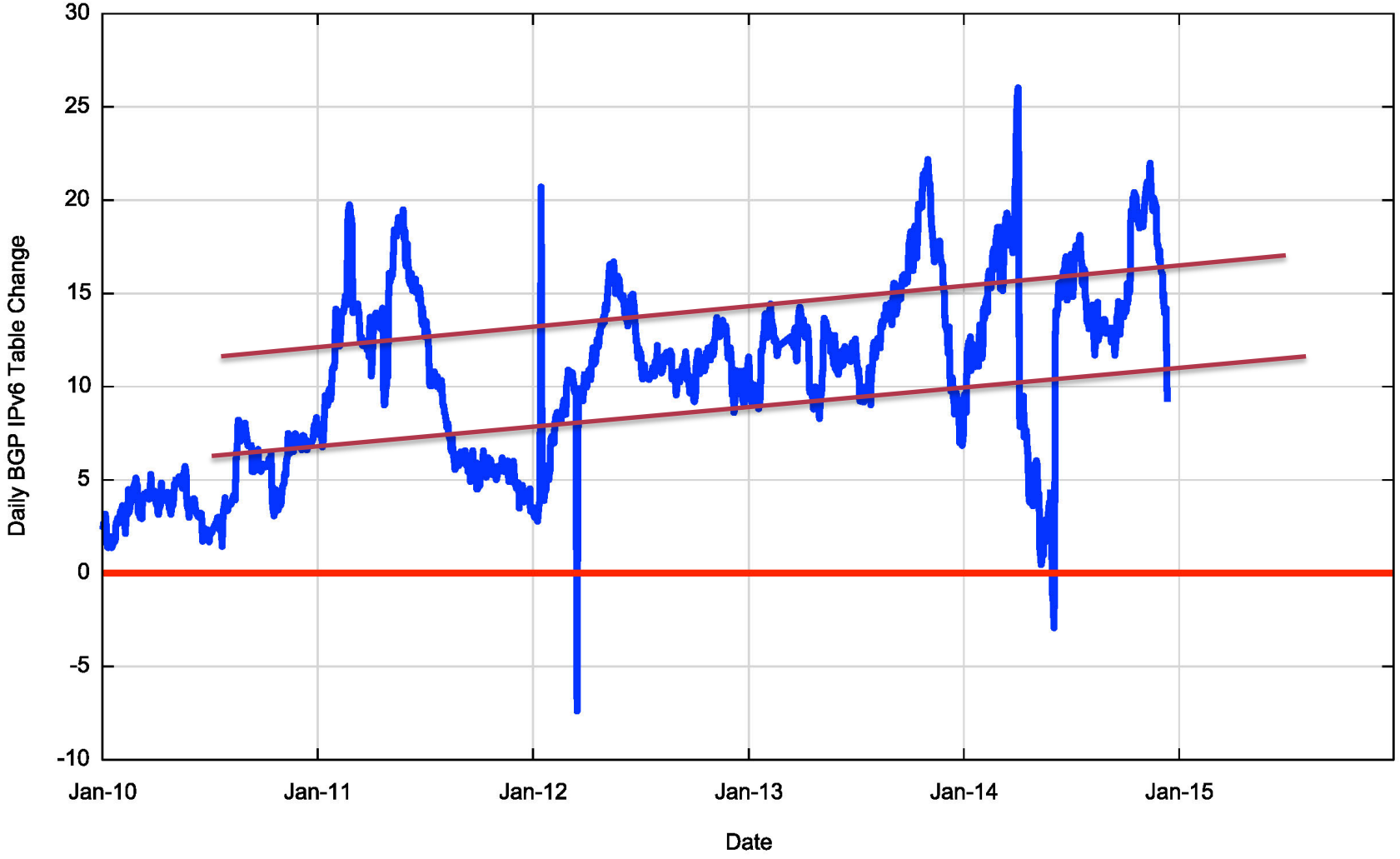
IPv6 Table Size



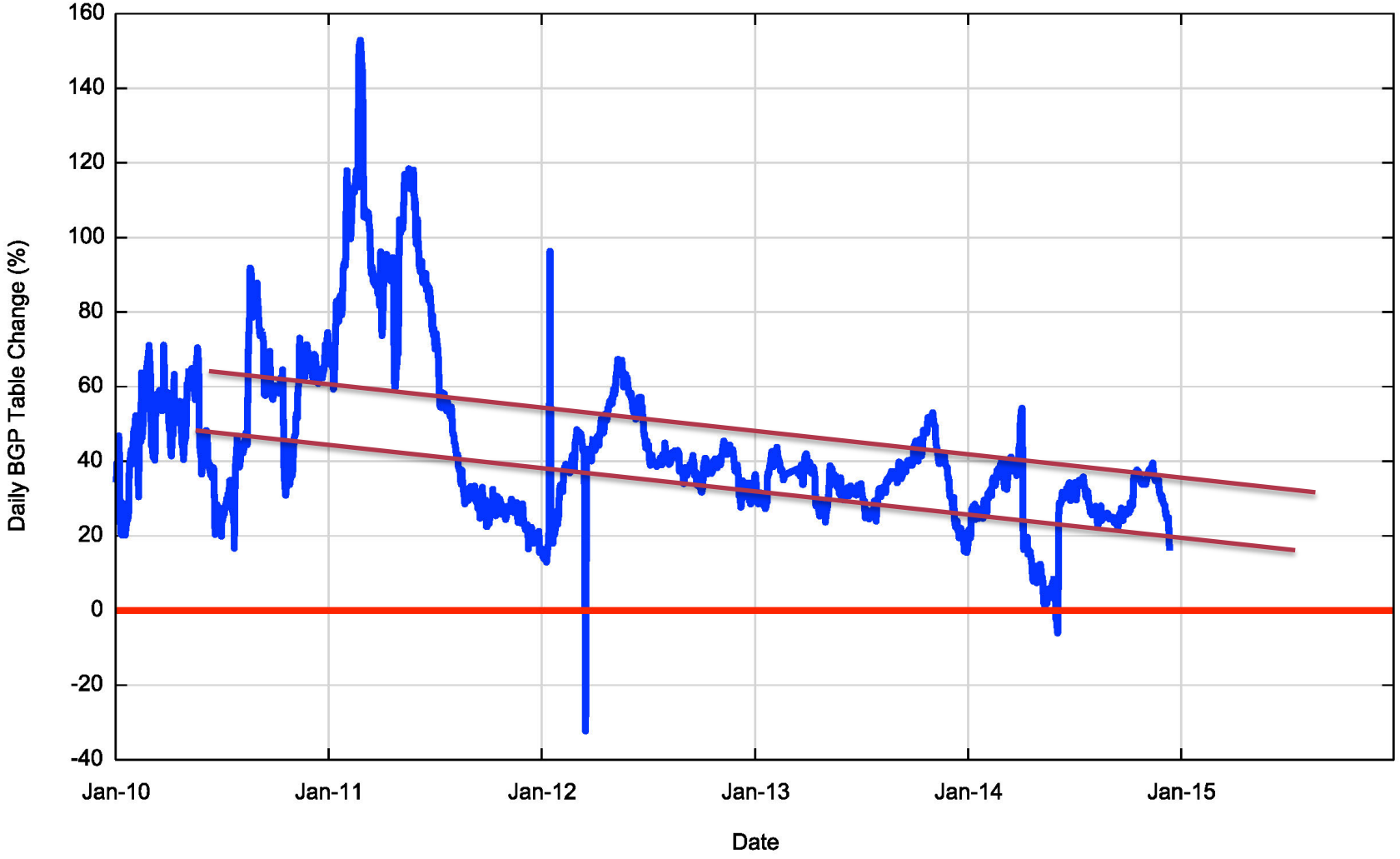
V6 - Daily Growth Rates



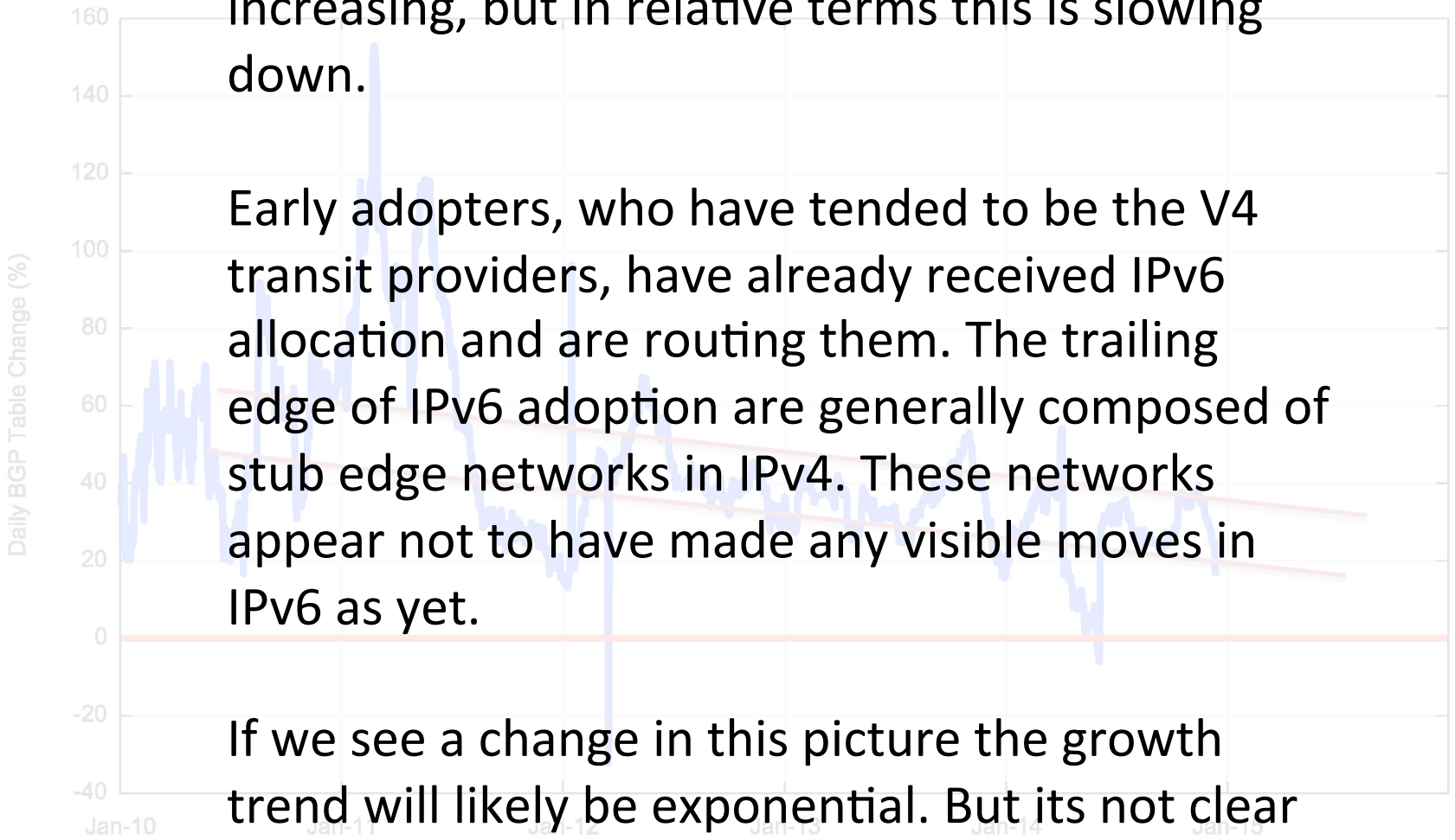
V6 - Daily Growth Rates



V6 - Relative Growth Rates



V6 - Relative Growth Rates



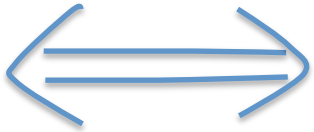
Growth in the V6 network appears to be increasing, but in relative terms this is slowing down.

Early adopters, who have tended to be the V4 transit providers, have already received IPv6 allocation and are routing them. The trailing edge of IPv6 adoption are generally composed of stub edge networks in IPv4. These networks appear not to have made any visible moves in IPv6 as yet.

If we see a change in this picture the growth trend will likely be exponential. But its not clear when such a tipping point will occur

IPv6 BGP Table Size predictions

	Exponential Model	Linear Model
Jan 2013	11,600 entries	
2014	16,200	
2015	*21,000*	
2016	*30,000*	*25,000*
2017	*42,000*	*29,000*
2018	*58,000*	*34,000*
2019	*82,000*	*38,000*
2019	*113,000*	*43,000*



Range of potential outcomes

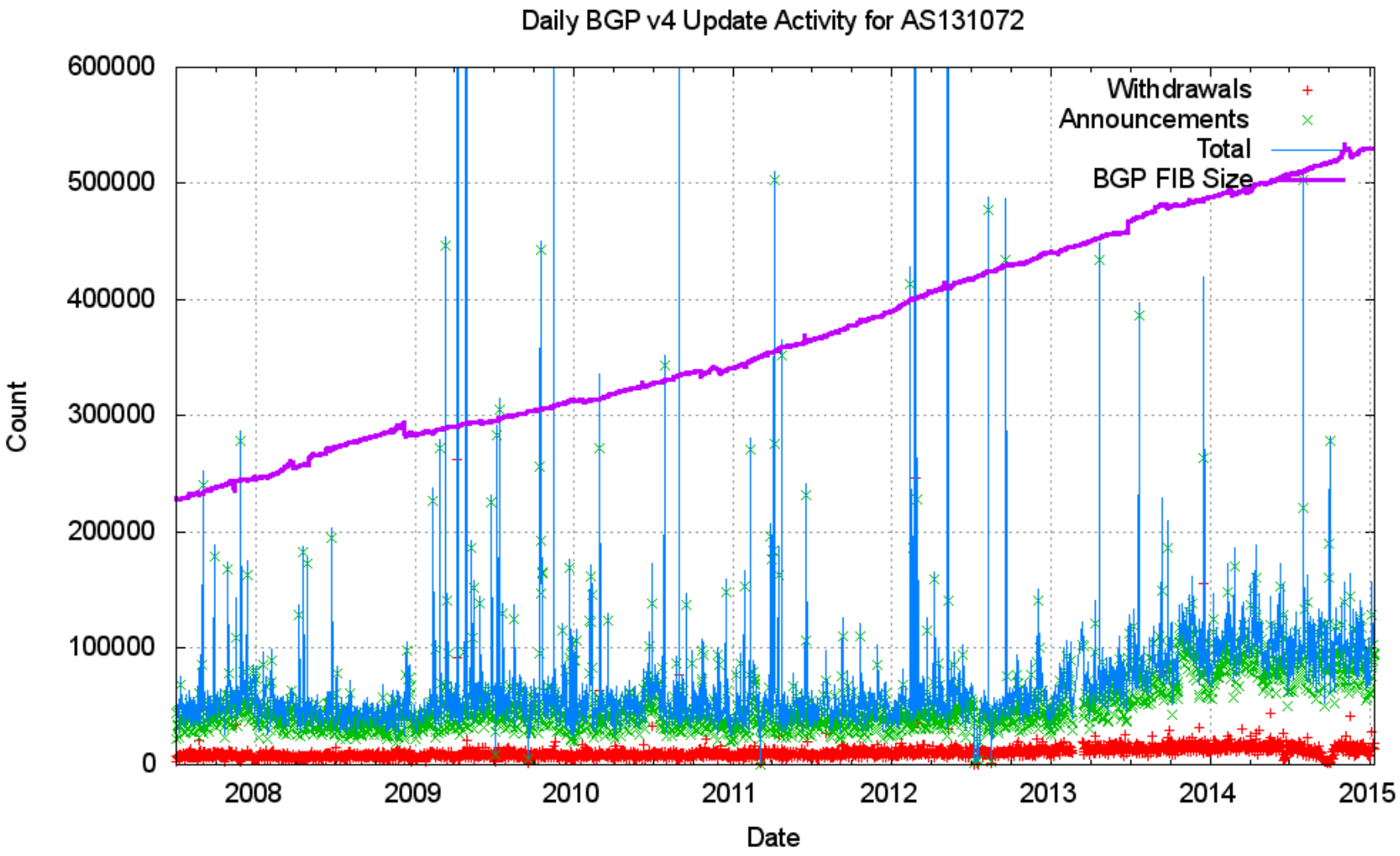
BGP Table Growth

- 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 at a faster relative rate, is still small in size in absolute terms
- As long as we are prepared to live within the technical constraints of the current routing paradigm, the Internet's use of BGP will continue to be viable for some time yet
- Nothing is melting in terms of the size of the routing table as yet

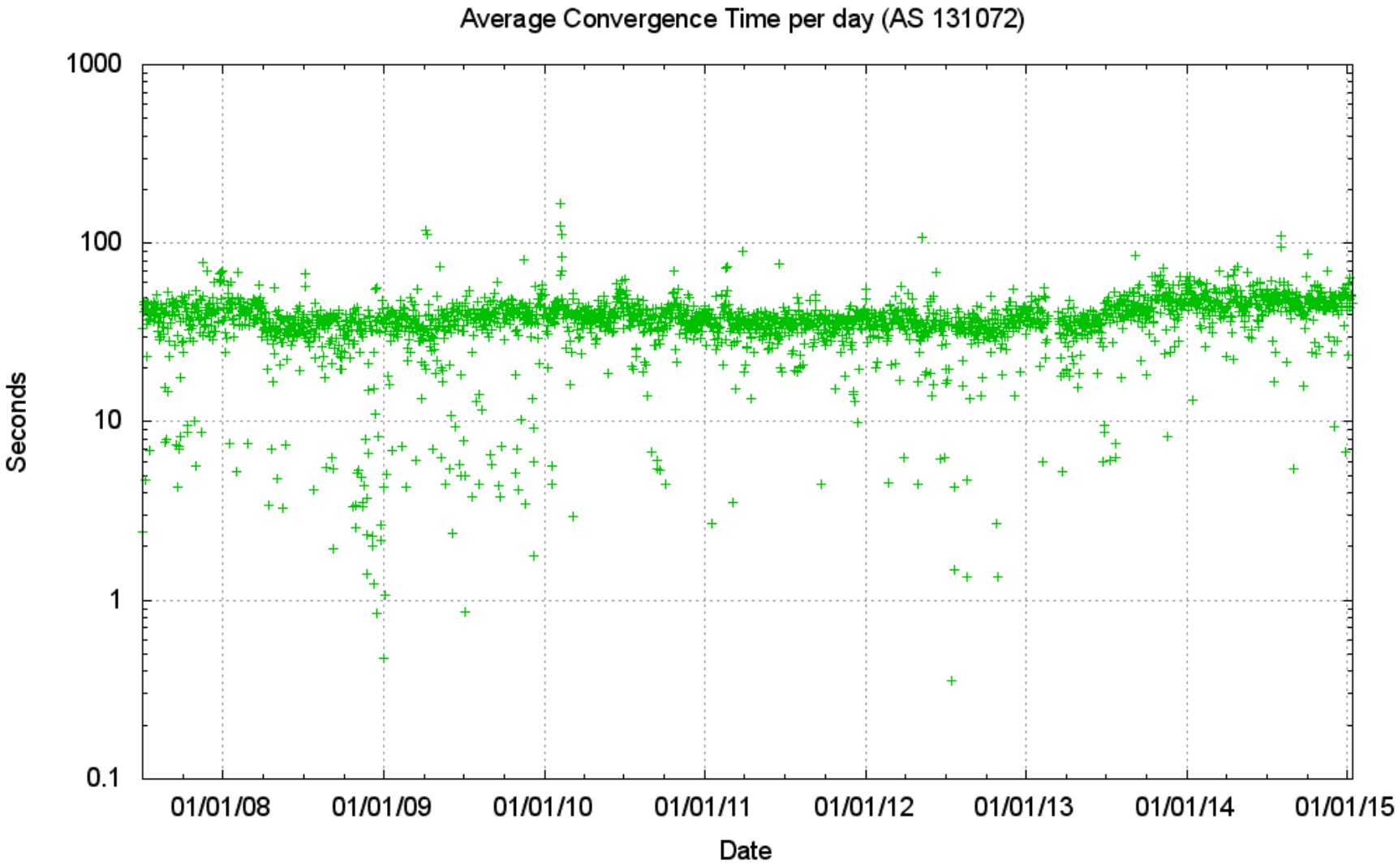
BGP Updates

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

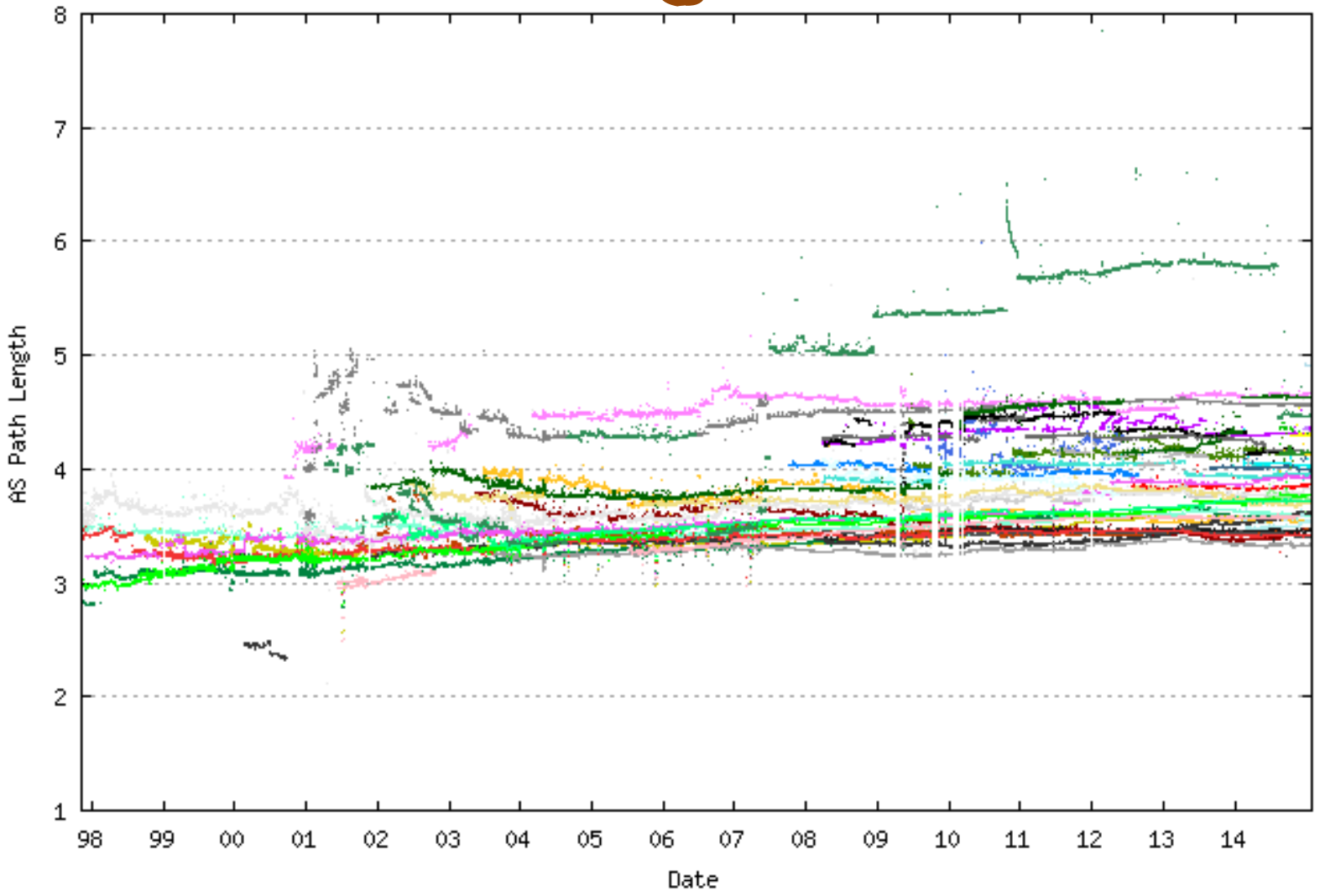
Announcements and Withdrawals



Convergence Performance



IPv4 Average AS Path Length



Data from Route Views

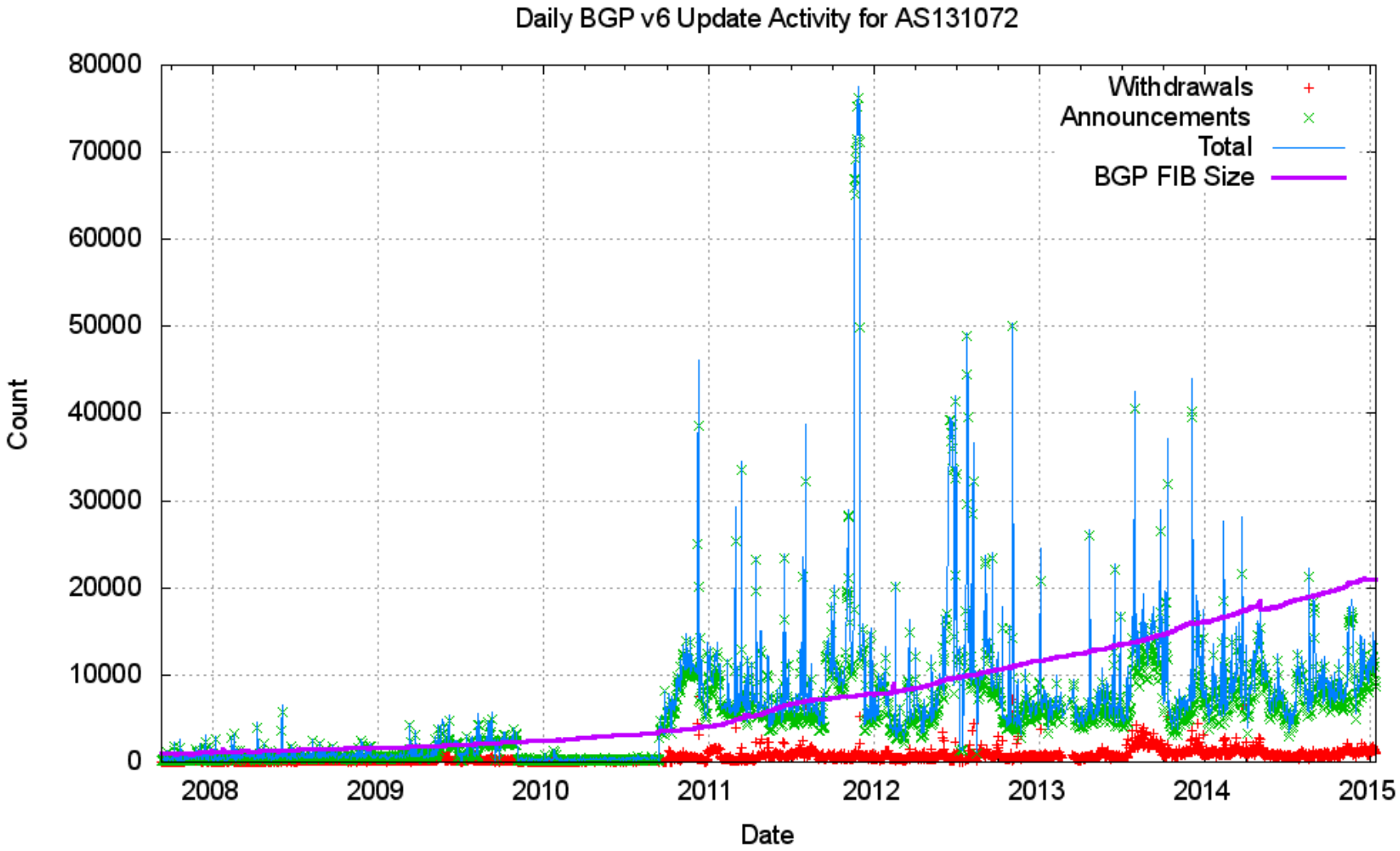
Updates in IPv4 BGP

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

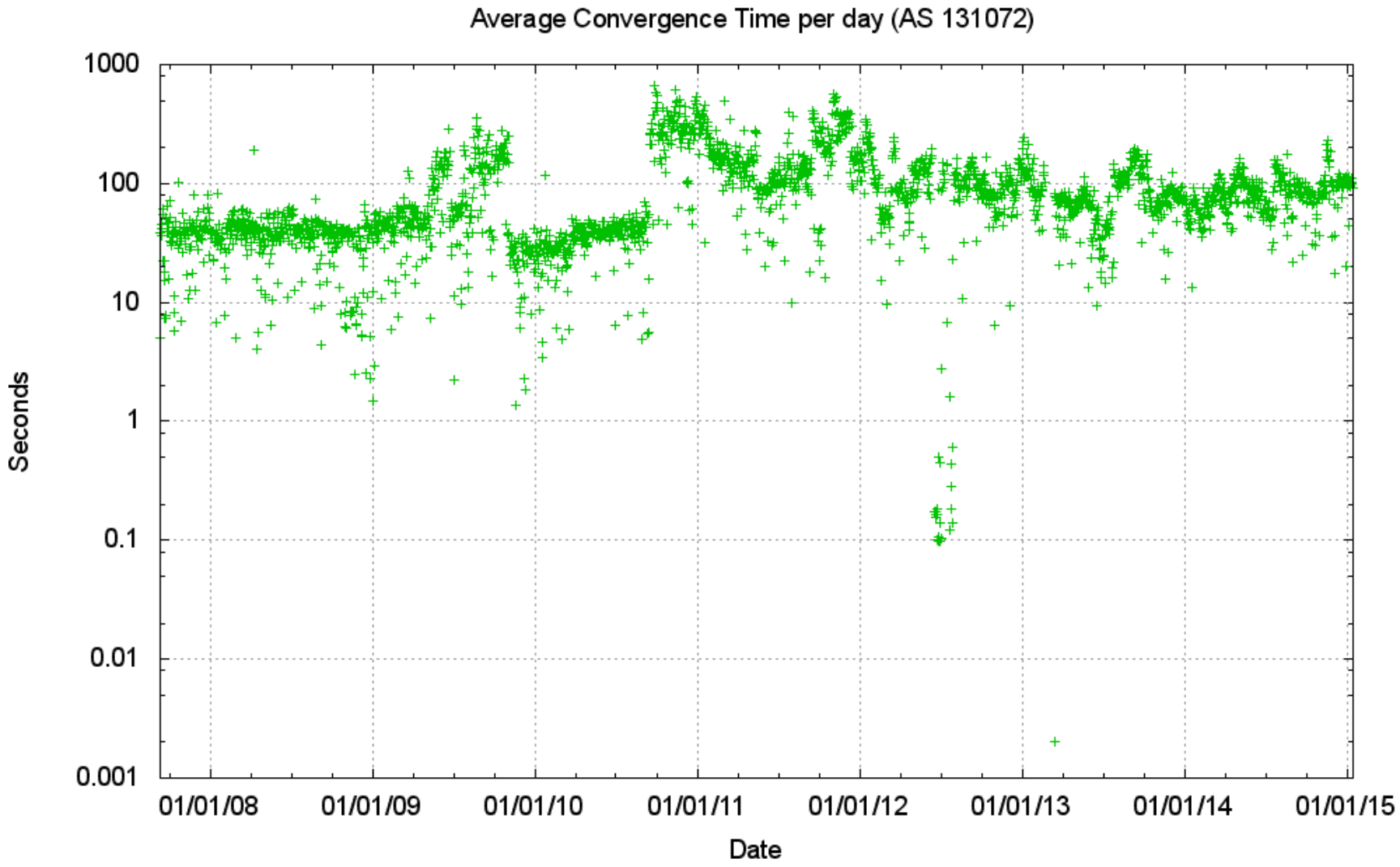
- The number of updates per instability event has been constant, which for a distance vector routing protocol is weird, and completely unanticipated. Distance Vector routing protocols should get noisier as the population of protocol speakers increases, and the increase should be multiplicative.
- But this is not happening in the Internet
- Which is good, but why is this not happening?

Likely contributors to this +ve outcome are the damping effect of widespread use of the MRAI interval, and the topology factor, as seen in the relatively constant AS Path length over this interval

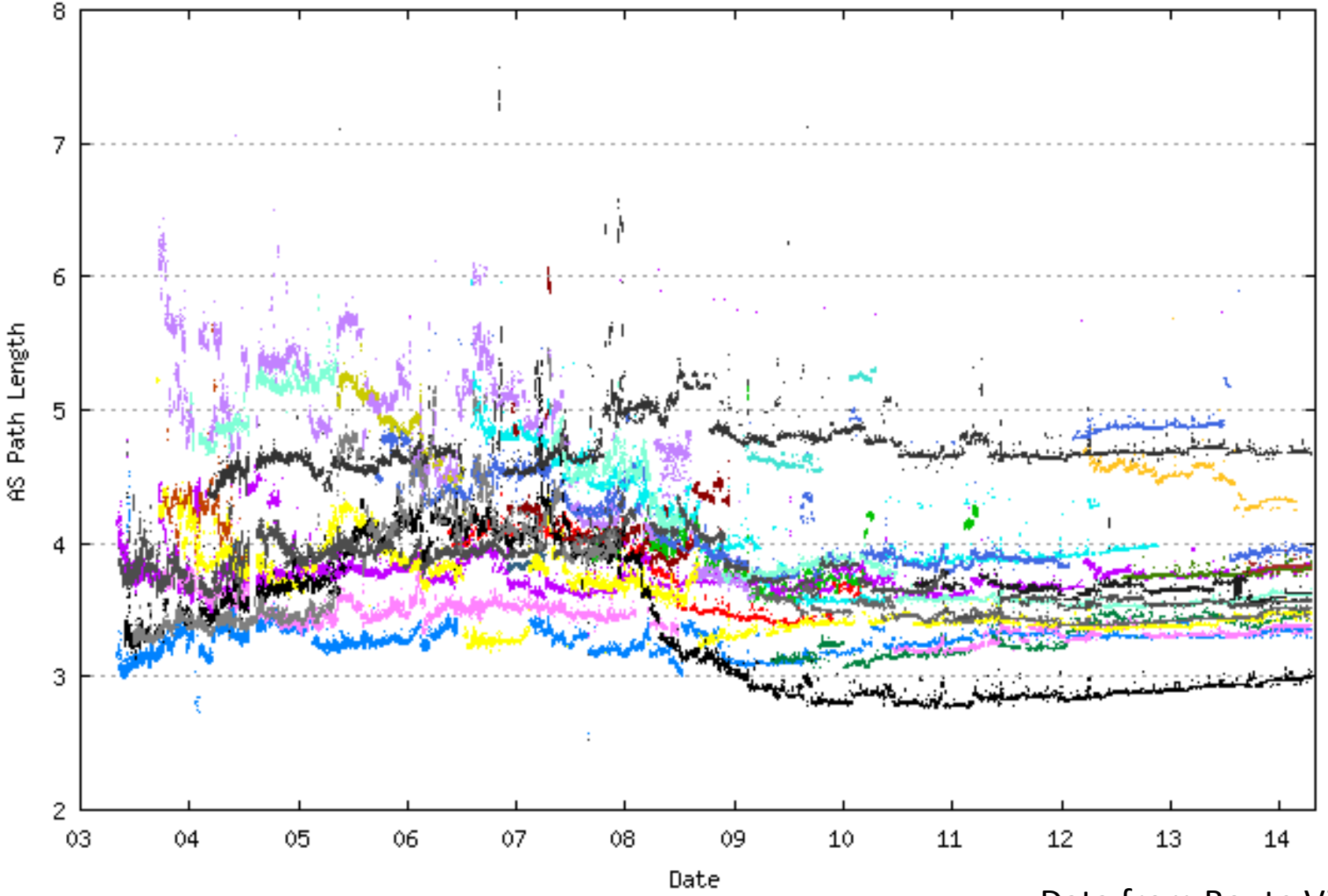
V6 Announcements and Withdrawals



V6 Convergence Performance



V6 Average AS Path Length



Data from Route Views

Updates in IPv6 BGP

IPv6 updates look a lot like IPv4 updates.

Which should not come as a surprise

It's the same routing protocol, and the same underlying inter-AS topology, and the observation is that the convergence times and instability rate appear to be unrelated to the population of the routing space.

So we see similar protocol convergence metrics in a network that is 1/20 of the size of the IPv4 network

It tends to underline the importance of dense connectivity and extensive use of local exchanges to minimize AS path lengths as a means of containing scaling of the routing protocol

Problem? Not a Problem?

There is nothing is this data to suggest that we will need a new inter-domain routing protocol in the next 5 years

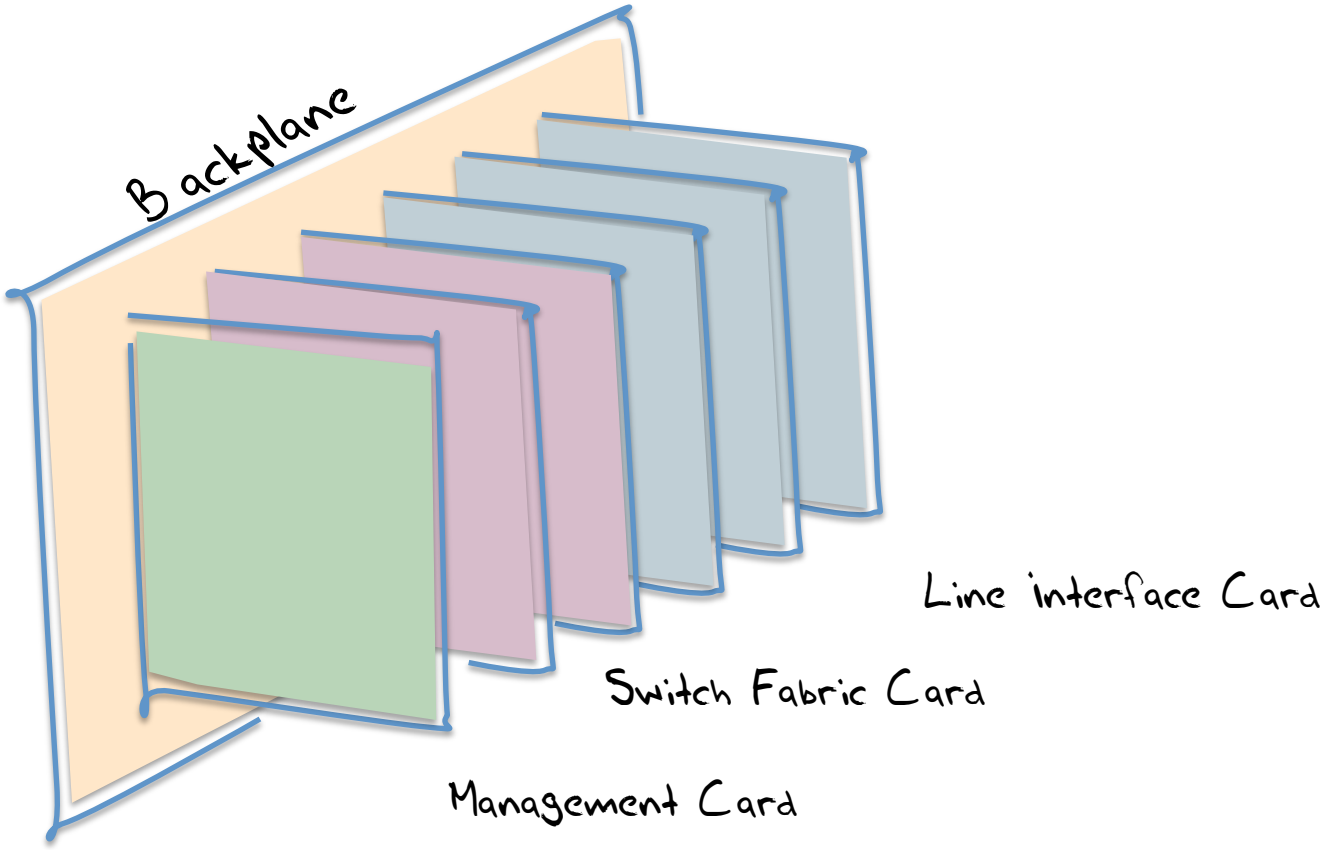
Or even in the next 10 to 15 years

But this is not the only scaling aspect of the Internet

Remember that BGP is a Best Path selection protocol. i.e. a single path selection protocol.

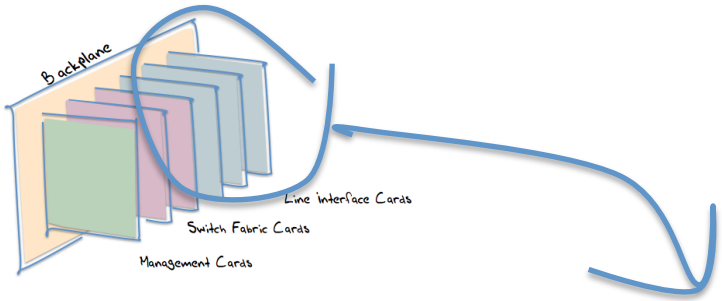
And that might contribute to the next scaling issue…

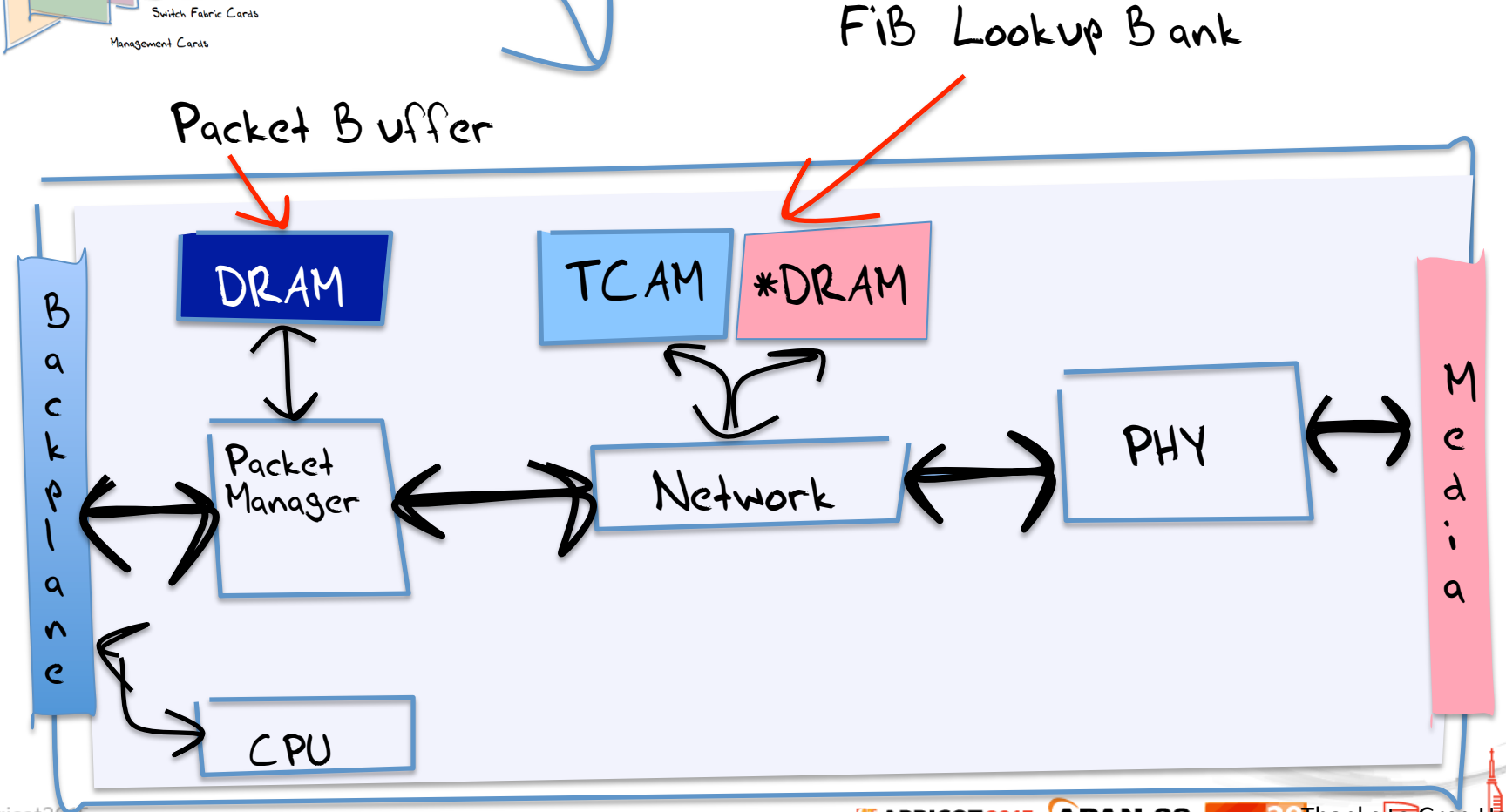
Inside a router



Thanks to Greg Hankins

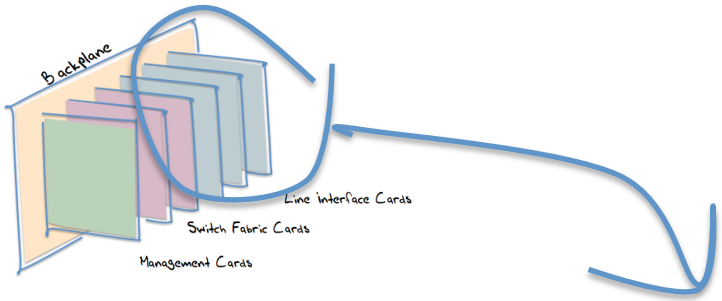
Inside a line card

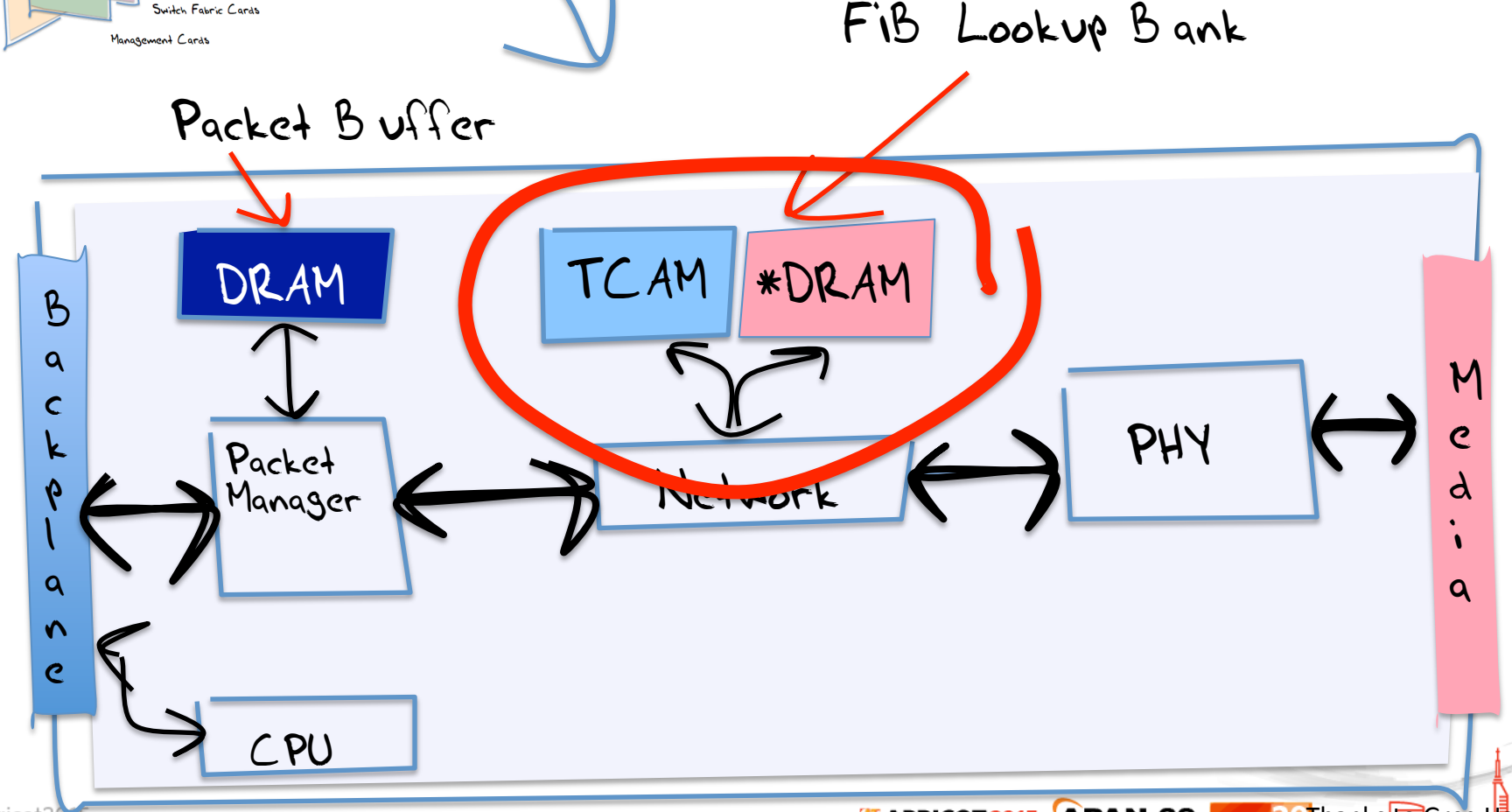




Thanks to Greg Hankins

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:

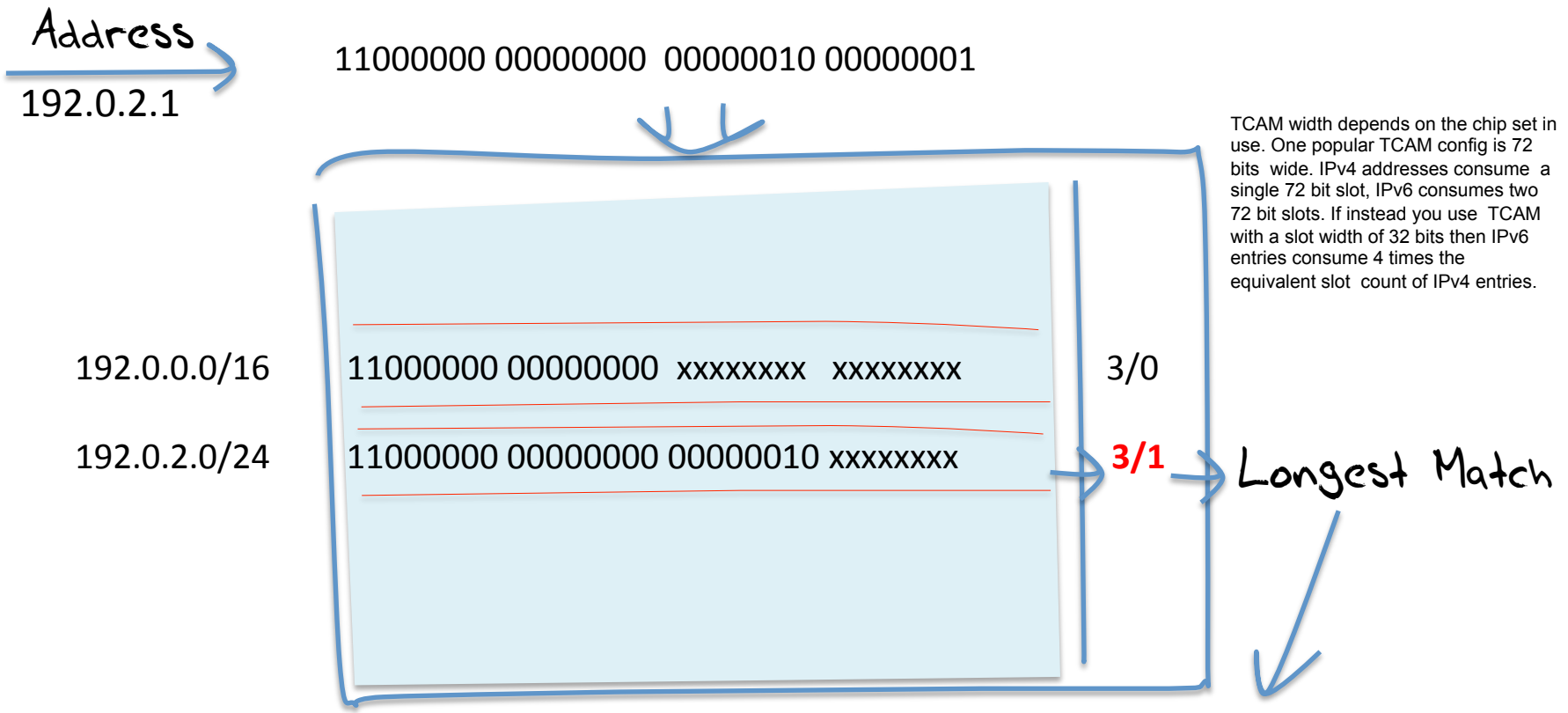
- Loading the entire routing table into a Ternary Content Addressable Memory bank (TCAM)

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



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.

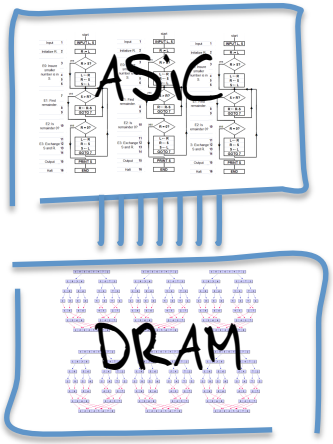
The entire FIB is loaded into TCAM. Every destination address is passed through the TCAM, and within one TCAM cycle the TCAM returns the interface index of the longest match. Each TCAM bank needs to be large enough to hold the entire FIB. TTCAM cycle time needs to be fast enough to support the max packet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

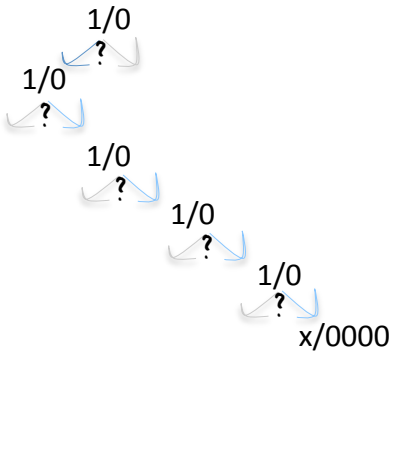
TRIE Lookup

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	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

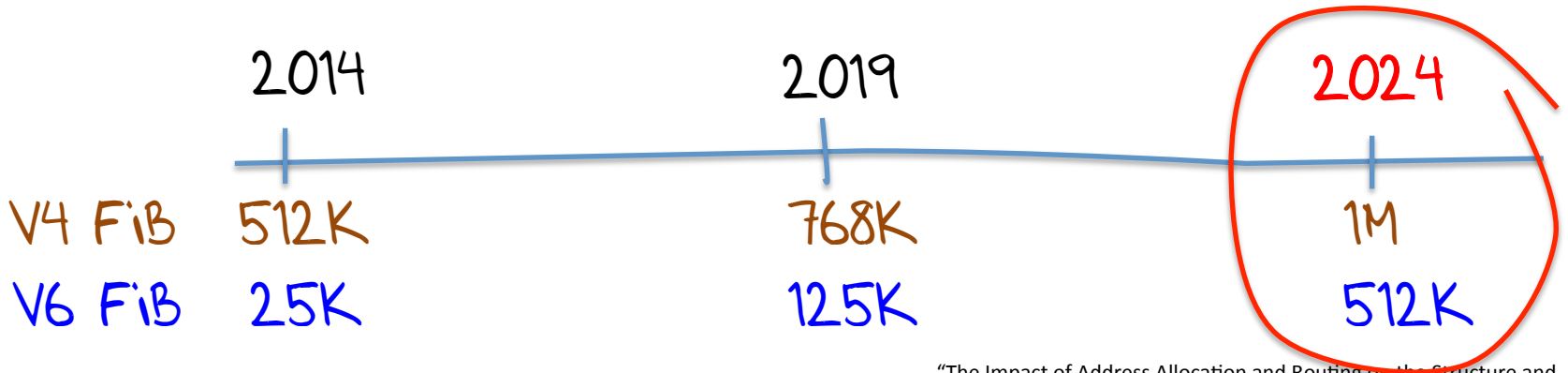
What memory size do we need for **10 years** of FIB growth from today?

TCAM

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

Trie

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



"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.

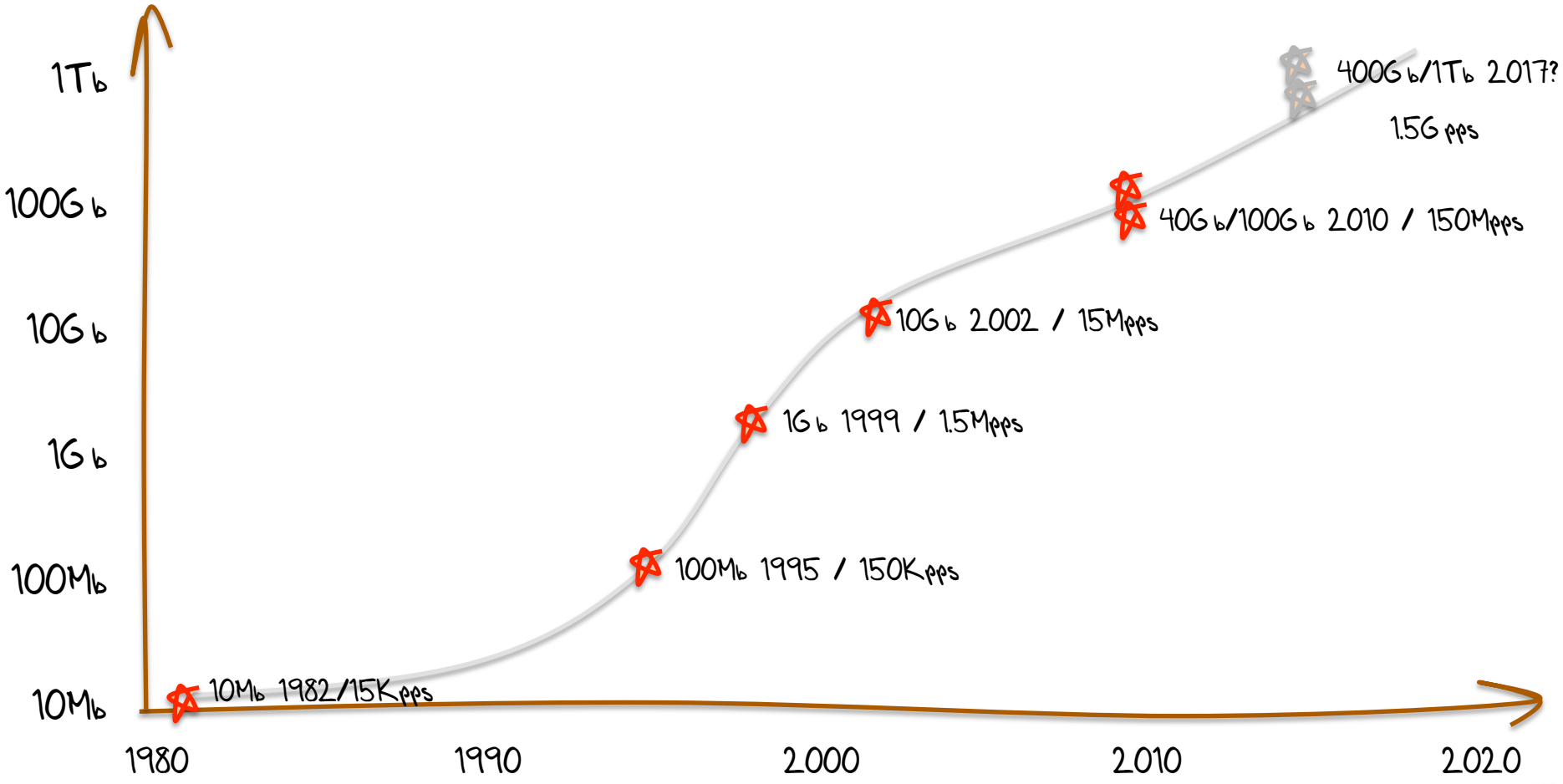
10Mb Ethernet had a 64 byte min packet size, plus preamble plus inter-packet spacing

=14,880 pps

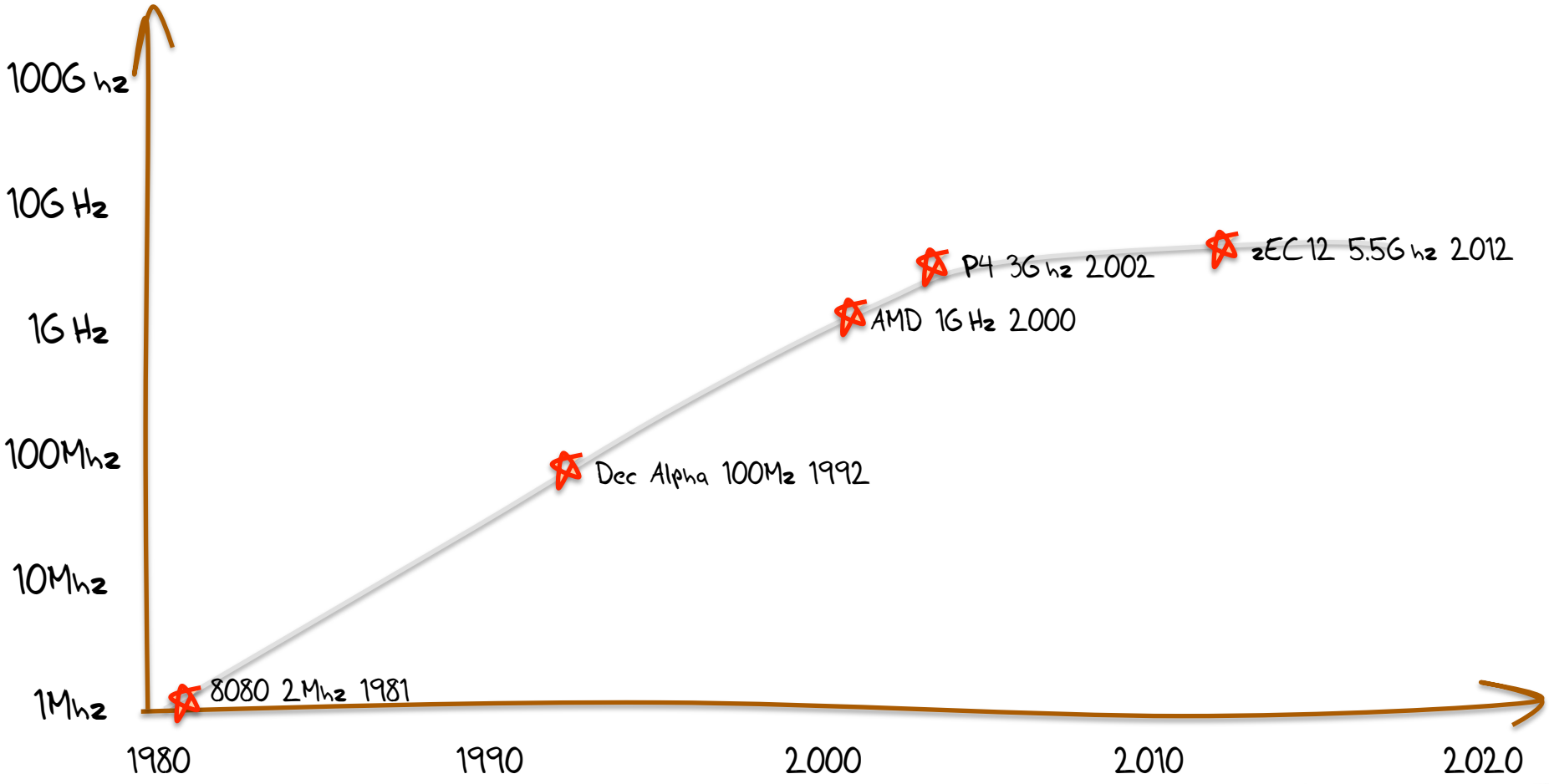
=1 packet every 67usec

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

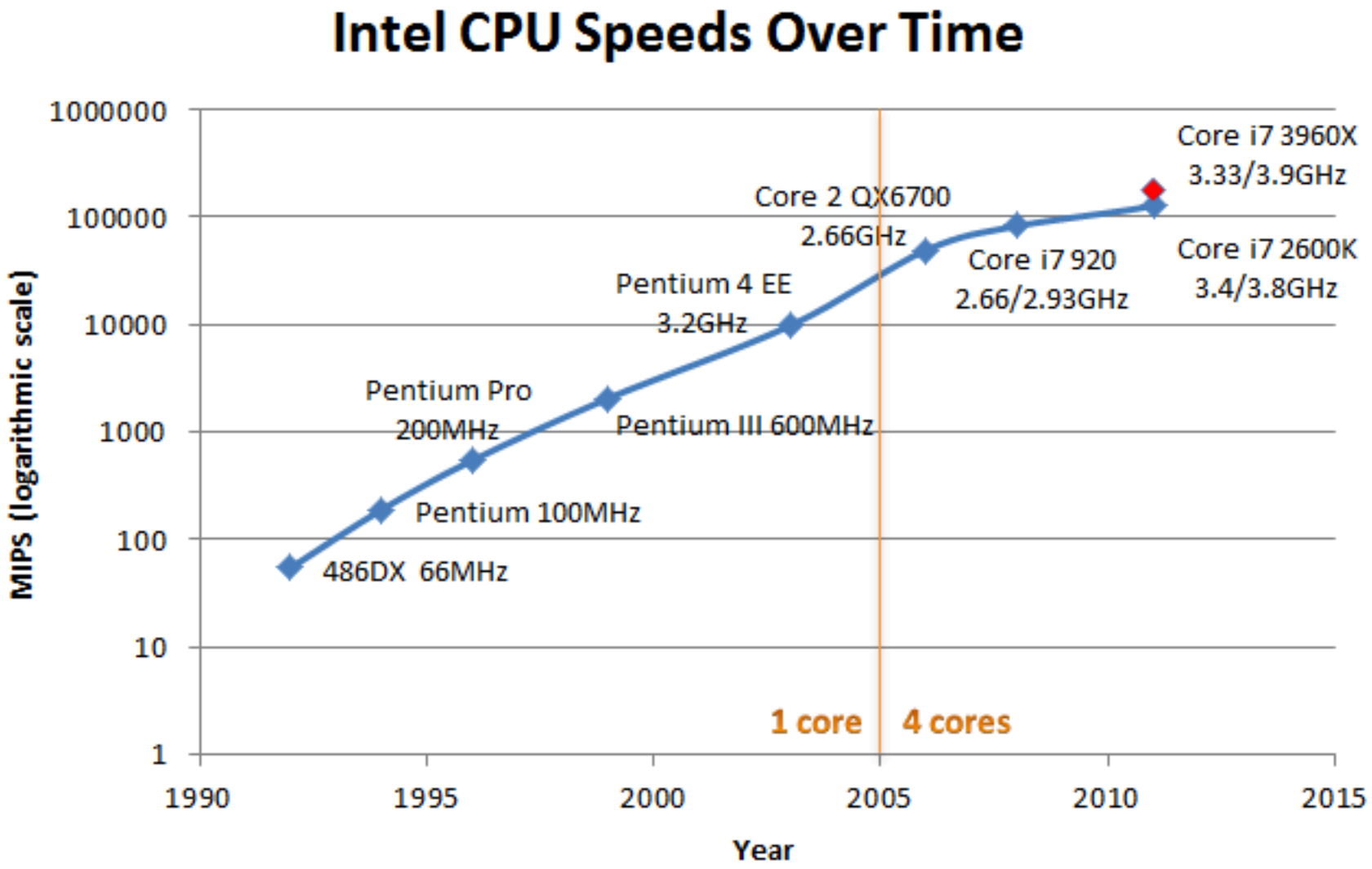
Wireline Speed - Ethernet



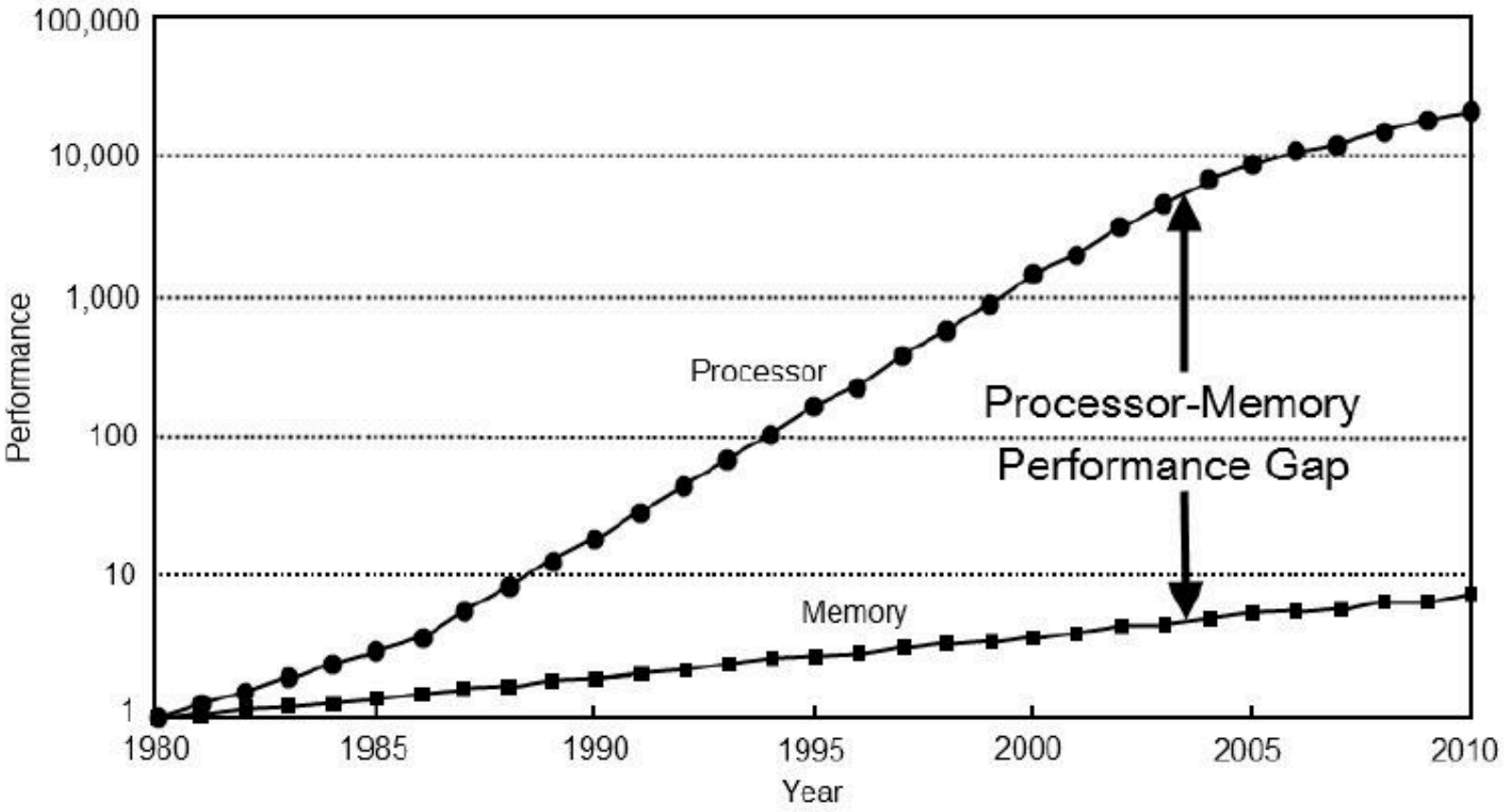
Clock Speed - Processors



Clock Speed - Processors



CPU vs Memory Speed



Speed, Speed, Speed

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

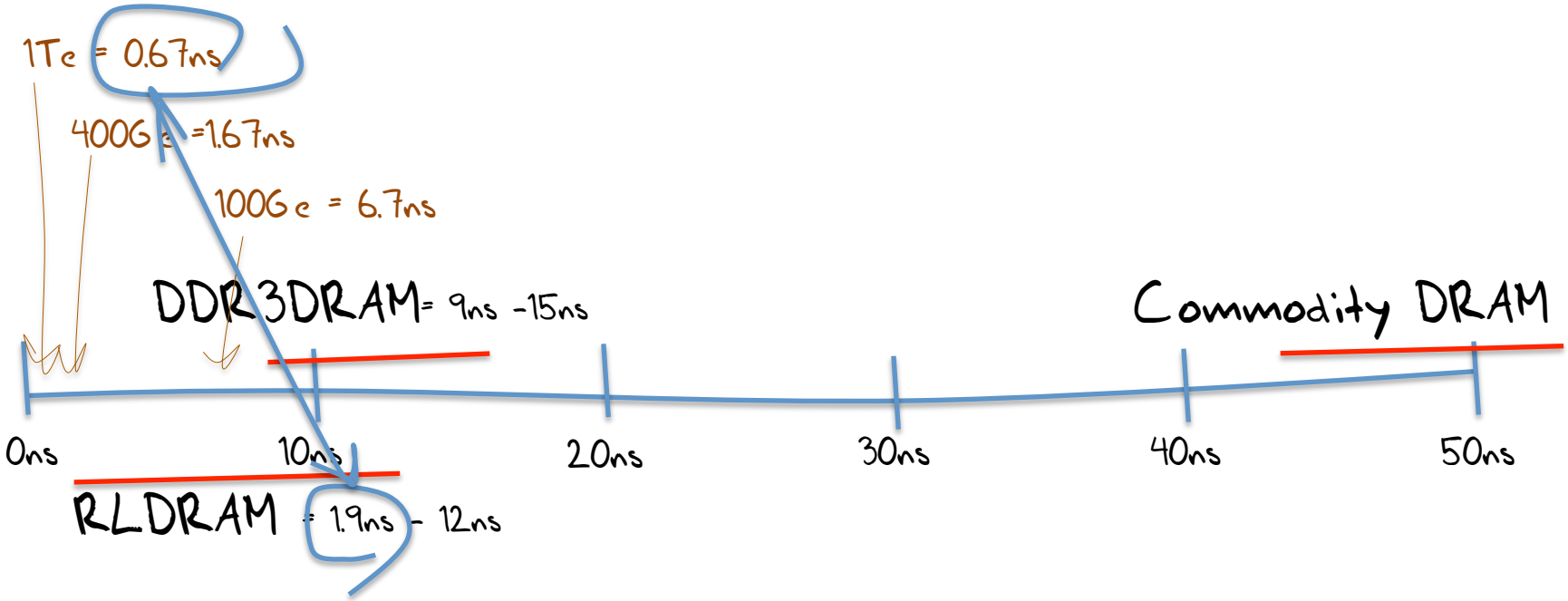
100GE ≈ 150Mpps ≈ 6.7ns per packet

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

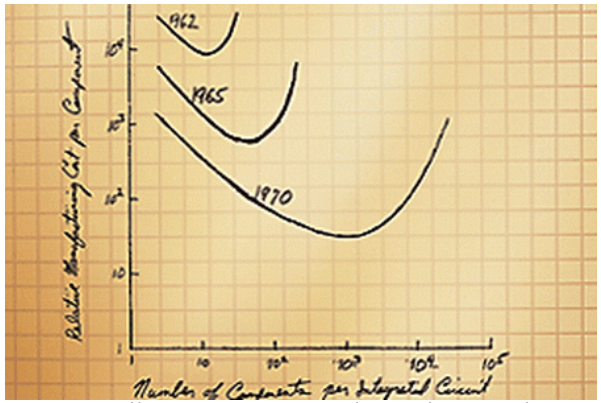
Scaling Speed

Scaling **size** is not a dramatic problem for the Internet of today or even tomorrow

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

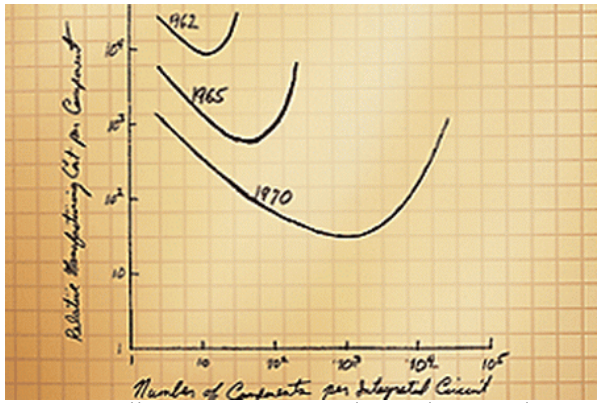


http://www.startupinnovation.org/research/moores-law/

Scaling Speed

If we can't route the max packet rate for a Terrabit wire then:

- Should the IEEE standards group recognise this and support a case to reduce the max packet rate by moving away from a 64byte min packet size for the 1Tb 802.3 standard?
- Can we push this into Layer 3? 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 do we start to tinker with the IP architecture itself? 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?



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

Thank You

Questions?