BGP scalability

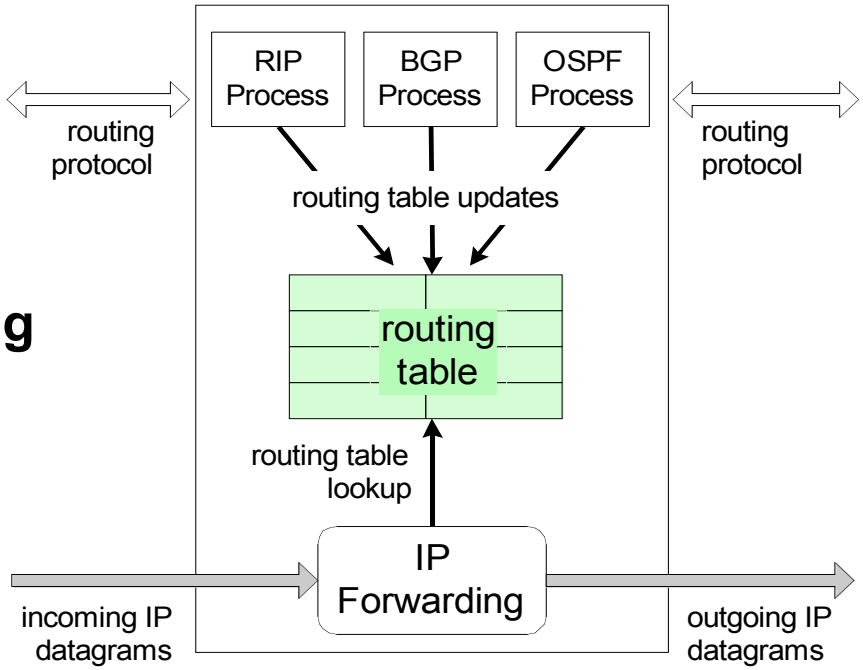
Eduardo Grampín
Universidad Carlos III de Madrid

IAB Workshop on Inter-Domain routing in October 2006 – RFC 4984:

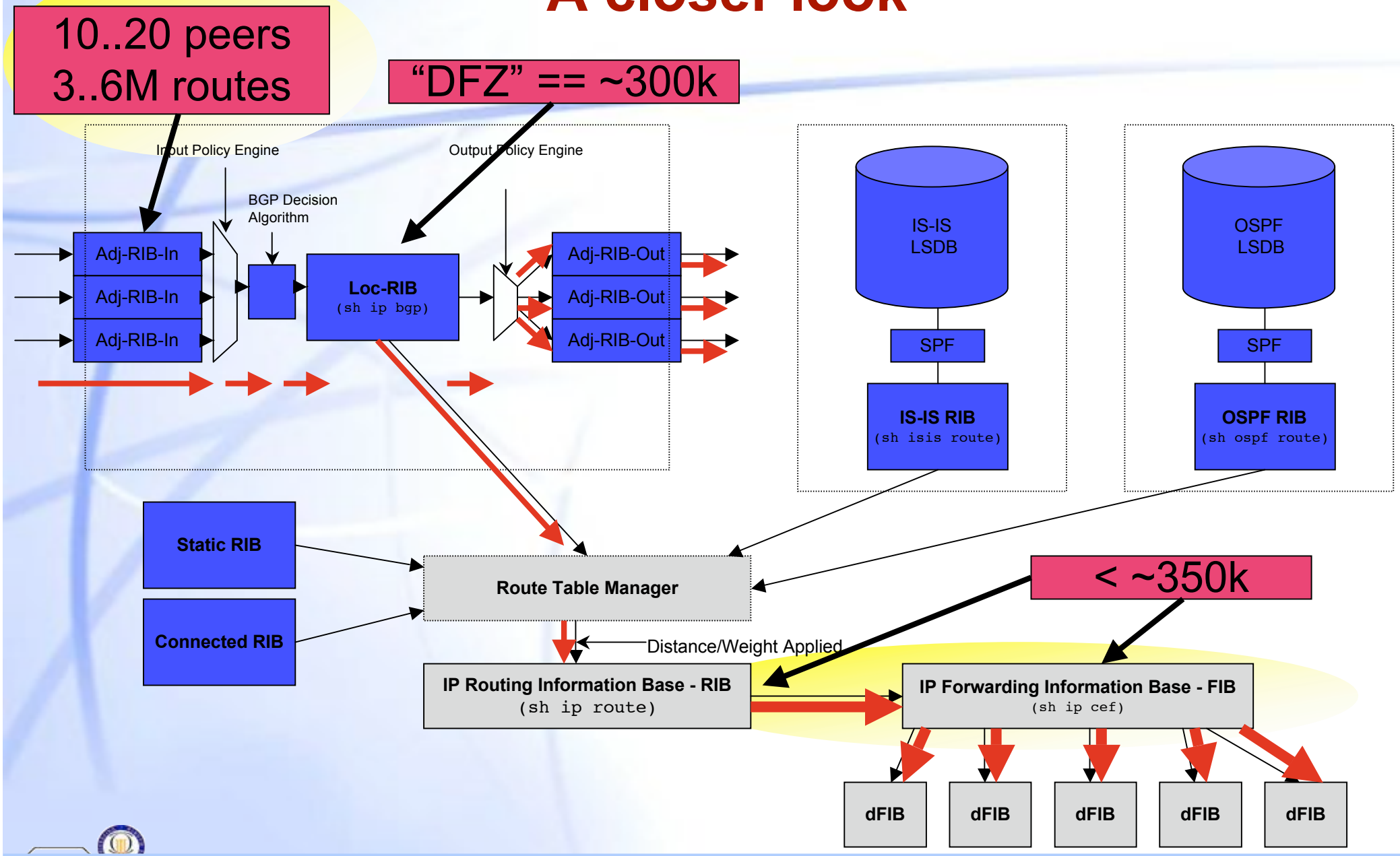
"routing scalability is the most important problem facing the Internet today and must be solved"

Scalability metrics

- **Data Plane**
  - **Internet Traffic**
    - ✓ **Fast forwarding engines & big pipes**
- **Control Plane**
  - **Size and dynamics of the routing table**
    - ✓ **# of routing entries**
    - ✓ **# and frequency of updates (a.k.a. "BGP churn")**
- **Remember…**
  - **Multiple routing protocols can run on the same router**
  - **Each routing protocol updates the routing table**



A closer look



show ip bgp

```
route-views>sh ip bgp
BGP table version is 4228790247, local router ID is 128.223.51.103
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
              r RIB-failure, S Stale
Origin codes: i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
*  1.9.0.0/16       144.228.241.130                        0 1239 3320 4788 i
*                   194.85.40.15                           0 3267 9002 4788 i
*                   194.85.102.33                          0 3277 3216 1273 4788 i
*                   64.71.255.61                           0 812 1273 4788 i
*                   154.11.98.225            0             0 852 3320 4788 i
*                   154.11.11.113            0             0 852 3320 4788 i
*                   209.124.176.223                        0 101 101 11164 4788 i
*                   69.31.111.244          183             0 4436 1273 4788 i
*                   66.185.128.48            7             0 1668 7018 4788 i
*                   129.250.0.11           302             0 2914 4788 i
*                   4.69.184.193             0             0 3356 1273 4788 i
*                   207.172.6.1              0             0 6079 4788 4788 i
*                   193.0.0.56                             0 3333 8218 4788 i
*                   65.106.7.139             3             0 2828 3549 4788 i
*                   208.74.64.40                           0 19214 26769 1273 4788 i
*                   207.46.32.34                           0 8075 4788 i
*>                  12.0.1.63                              0 7018 4788 i
*                   216.218.252.164                        0 6939 4788 i
*                   207.172.6.20             0             0 6079 2914 4788 i
*
*
```

Source: telnet route-views.routeviews.org

Full table: http://bgp.potaroo.net/as2.0/bgptable.txt

show ip route

```
route-views>show ip route 1.9.0.0 255.255.0.0
Routing entry for 1.9.0.0/16
  Known via "bgp 6447", distance 20, metric 0
  Tag 7018, type external
  Last update from 12.0.1.63 2w4d ago
  Routing Descriptor Blocks:
  * 12.0.1.63, from 12.0.1.63, 2w4d ago
      Route metric is 0, traffic share count is 1
      AS Hops 2
      Route tag 7018
```

Source: telnet route-views.routeviews.org

Large BGP Tables Considered Harmful

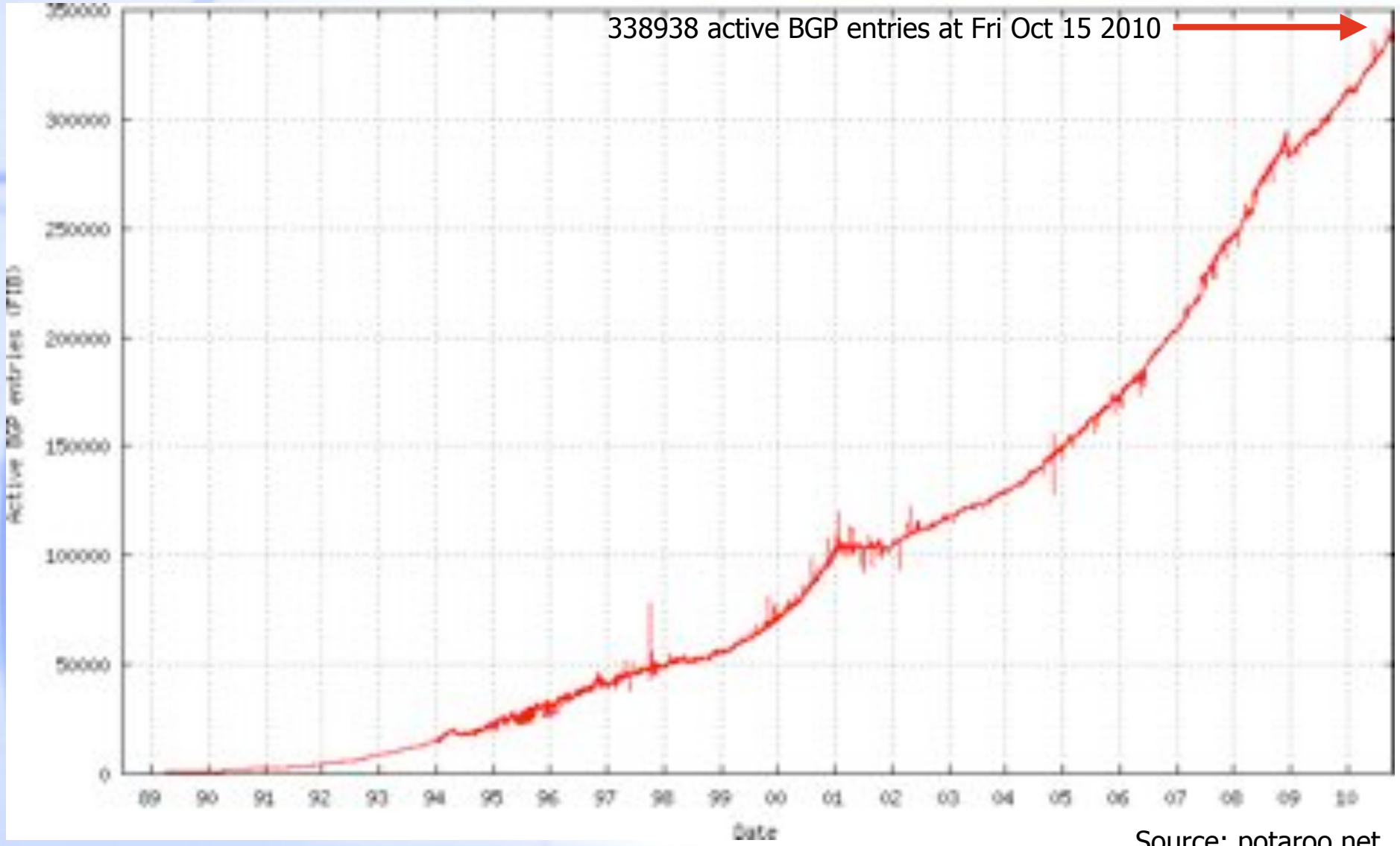
- Routing tables must store best routes and alternate routes
  - Burden can be large for routers with many alternate routes (route reflectors for example)
  - Memory requirements
- Increases CPU load, especially during session reset
  - Routers have been known to die
- Moore's Law may save us in theory. But in practice it means spending money to upgrade equipment …

BGP measurements

There are a number of ways to "measure" BGP:

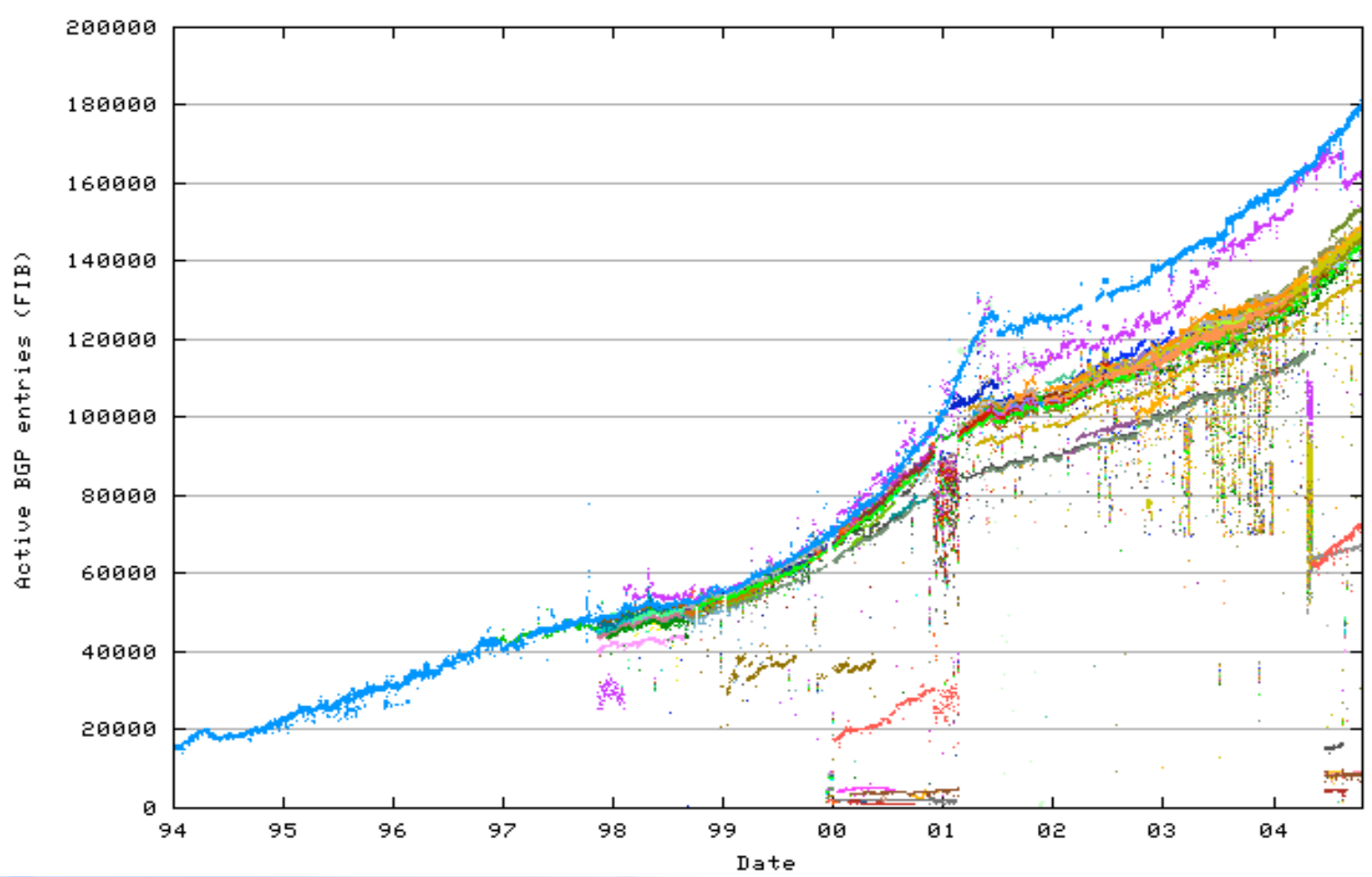
1. **Assemble a large set of BGP peering sessions and record everything**
   - ✓ RIPE NCC's RIS service
   - ✓ Route Views
2. **Perform carefully controlled injections of route information and observe the propagation of information**
   - ✓ Beacons
   - ✓ AS Set manipulation
   - ✓ Bogon Detection and Triangulation
3. **Take a single BGP perspective and perform continuous recording of a number of BGP metrics over a long baseline -> potaroo.net**

BGP Routing Table Size



Source: potaroo.net

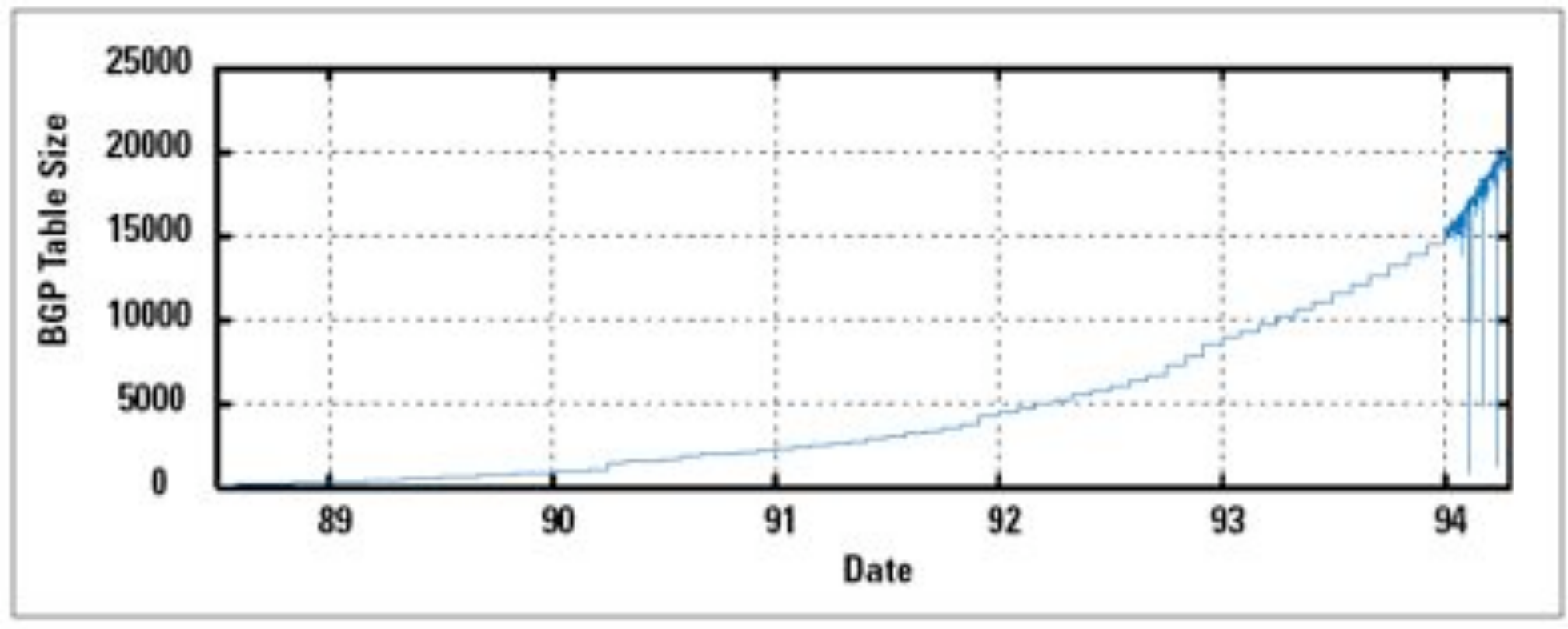
BGP Routing Table Size: multiple views



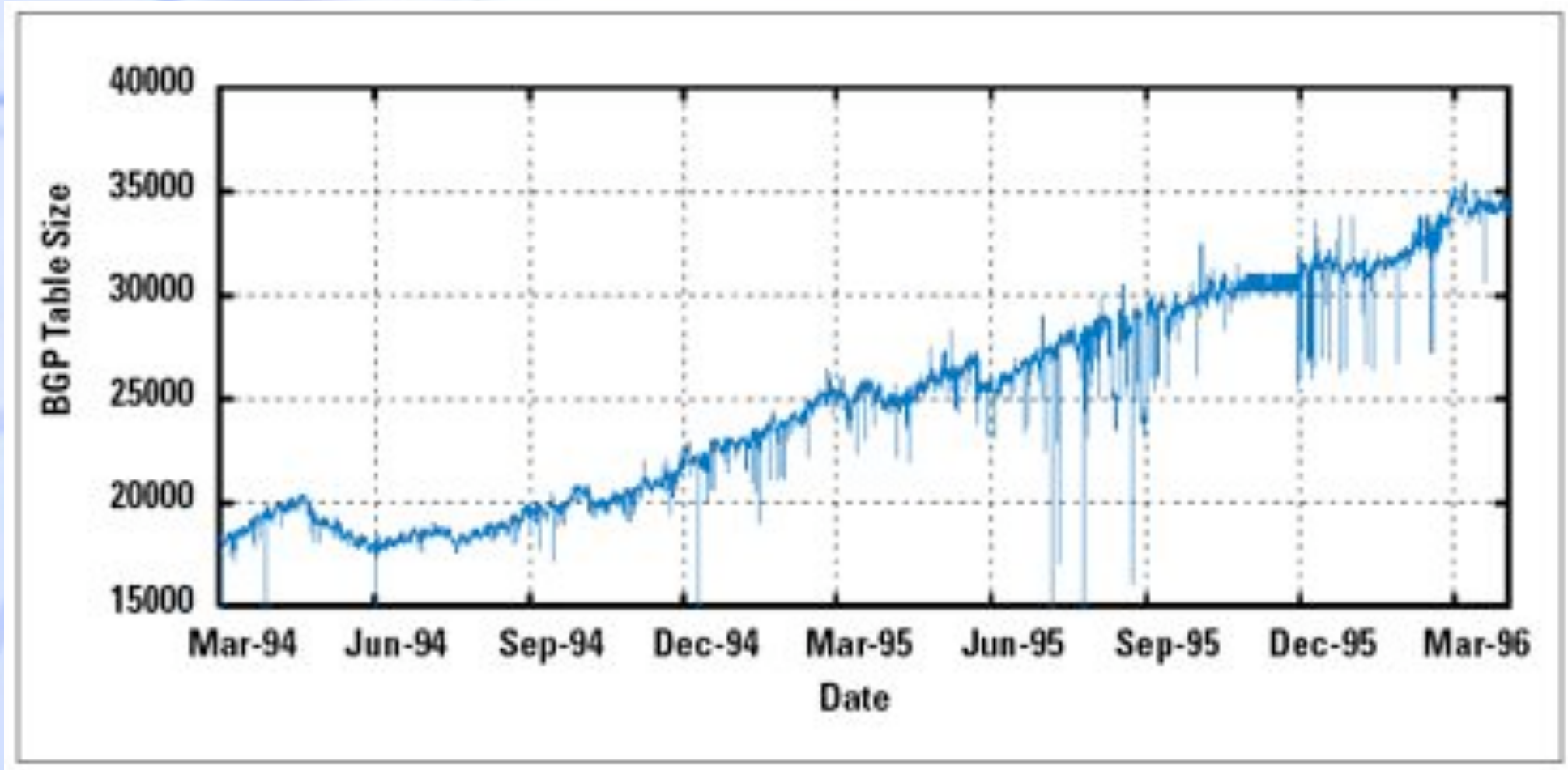
Data assembled from a variety of sources, Including Surfnet, Telstra, KPN and Route Views. Each colour represents a time series for a single AS.

The major point here is that there is no single view of routing. Each AS view is based on local conditions, which include some local information and also local filtering policies about external views.

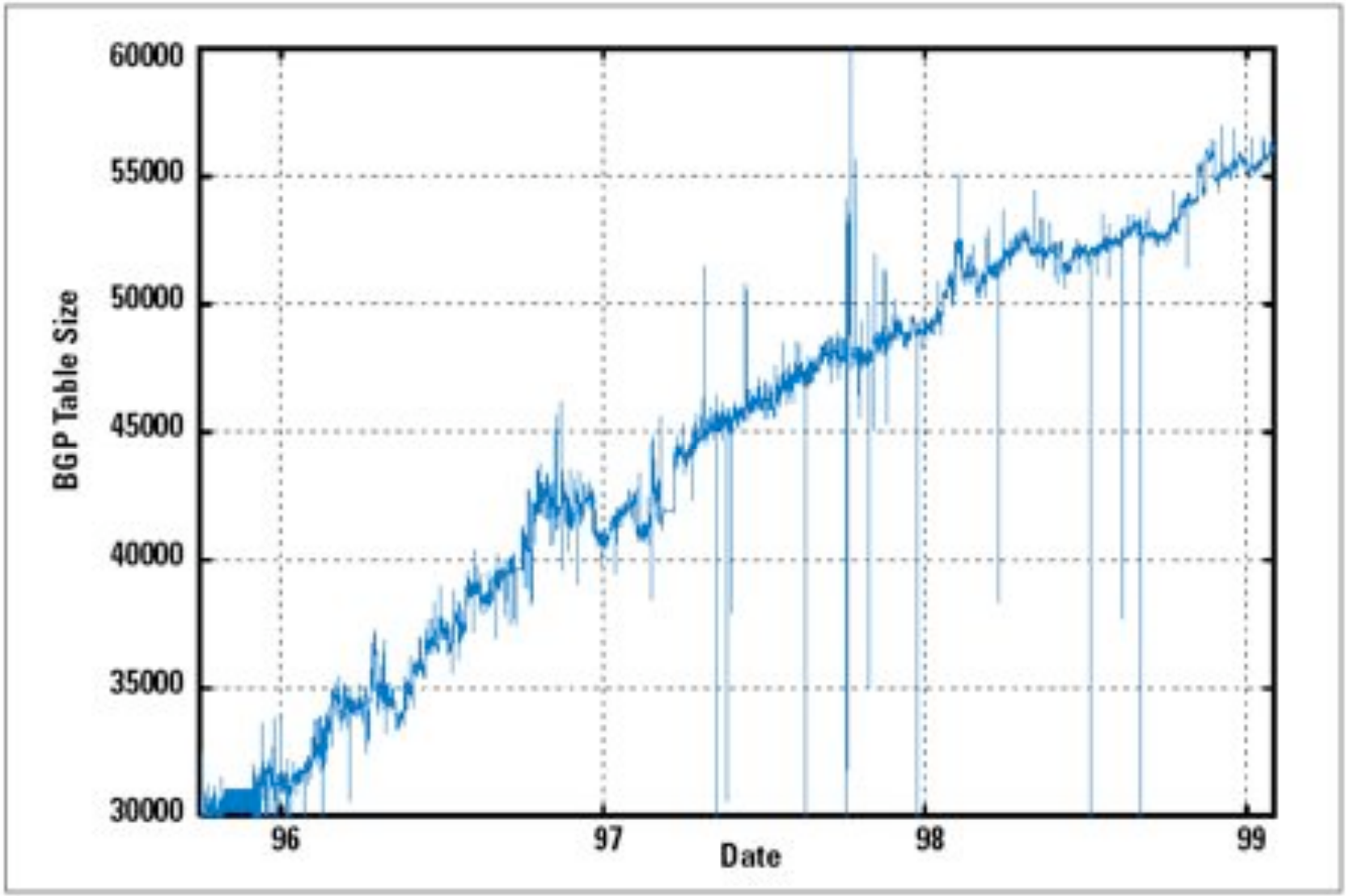
Pre-CIDR growth (1989-1994)



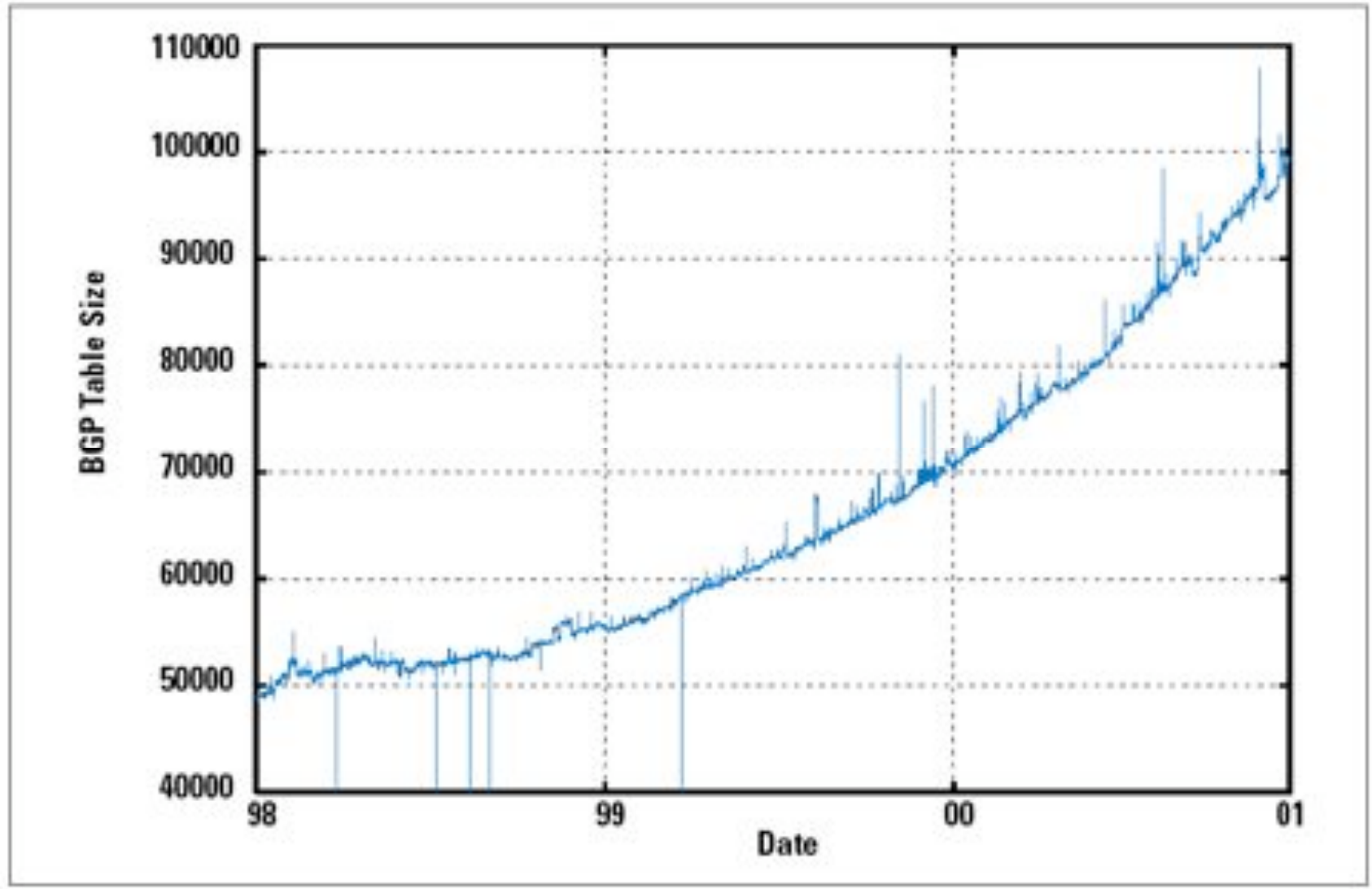
CIDR deployment (1994-1995)



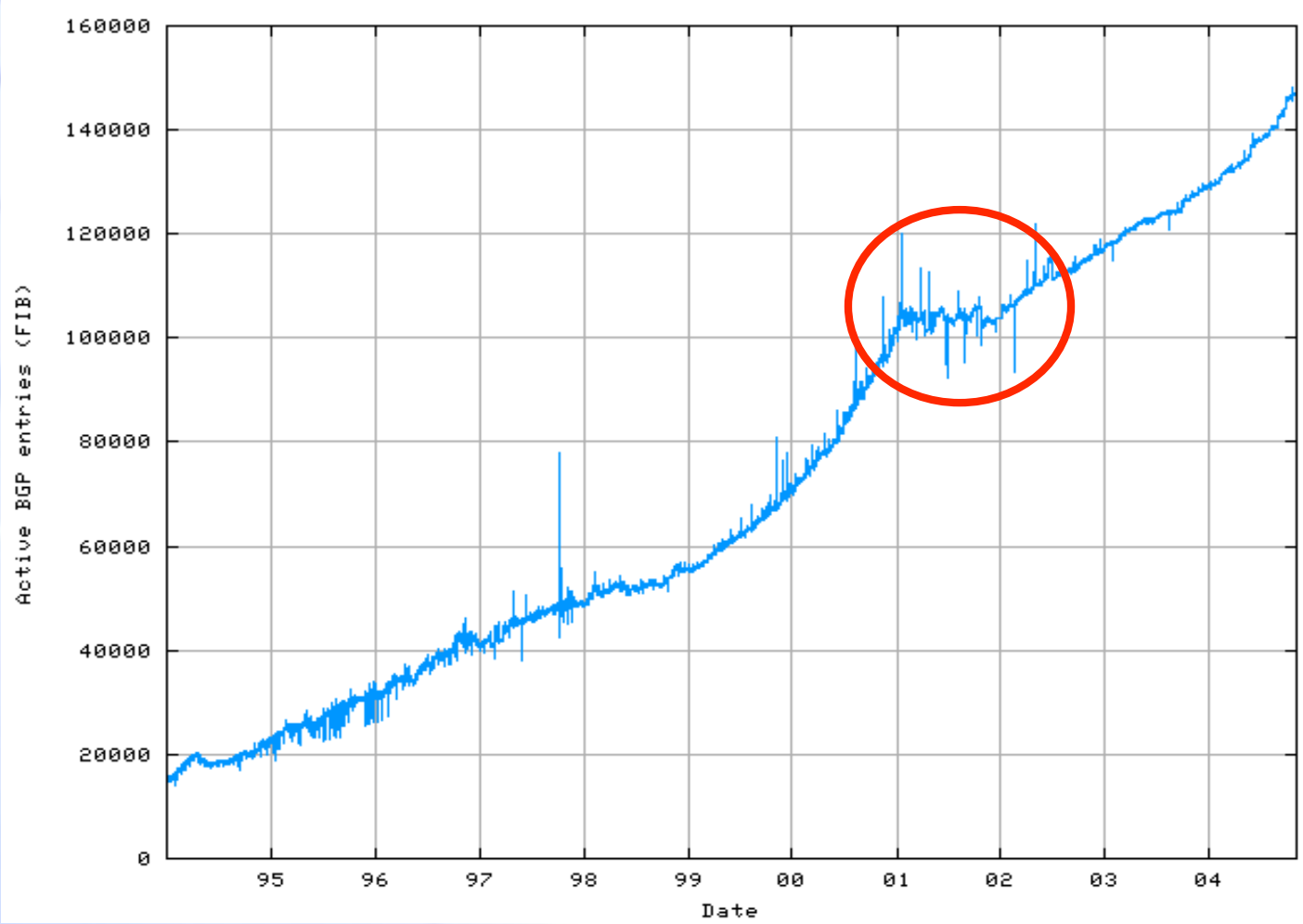
CIDR Growth (1995-1998)



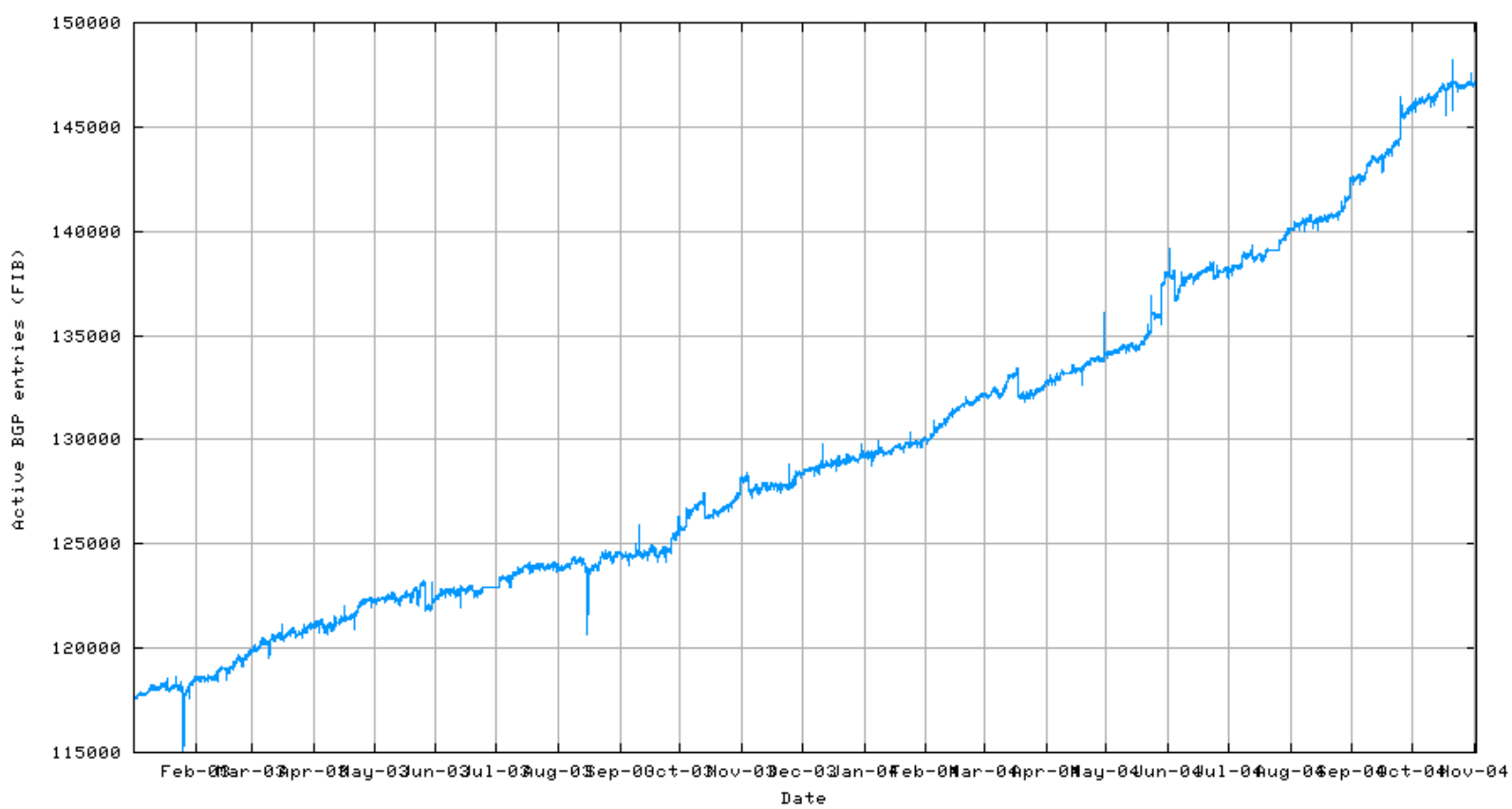
Back to exponential (1998-2000)



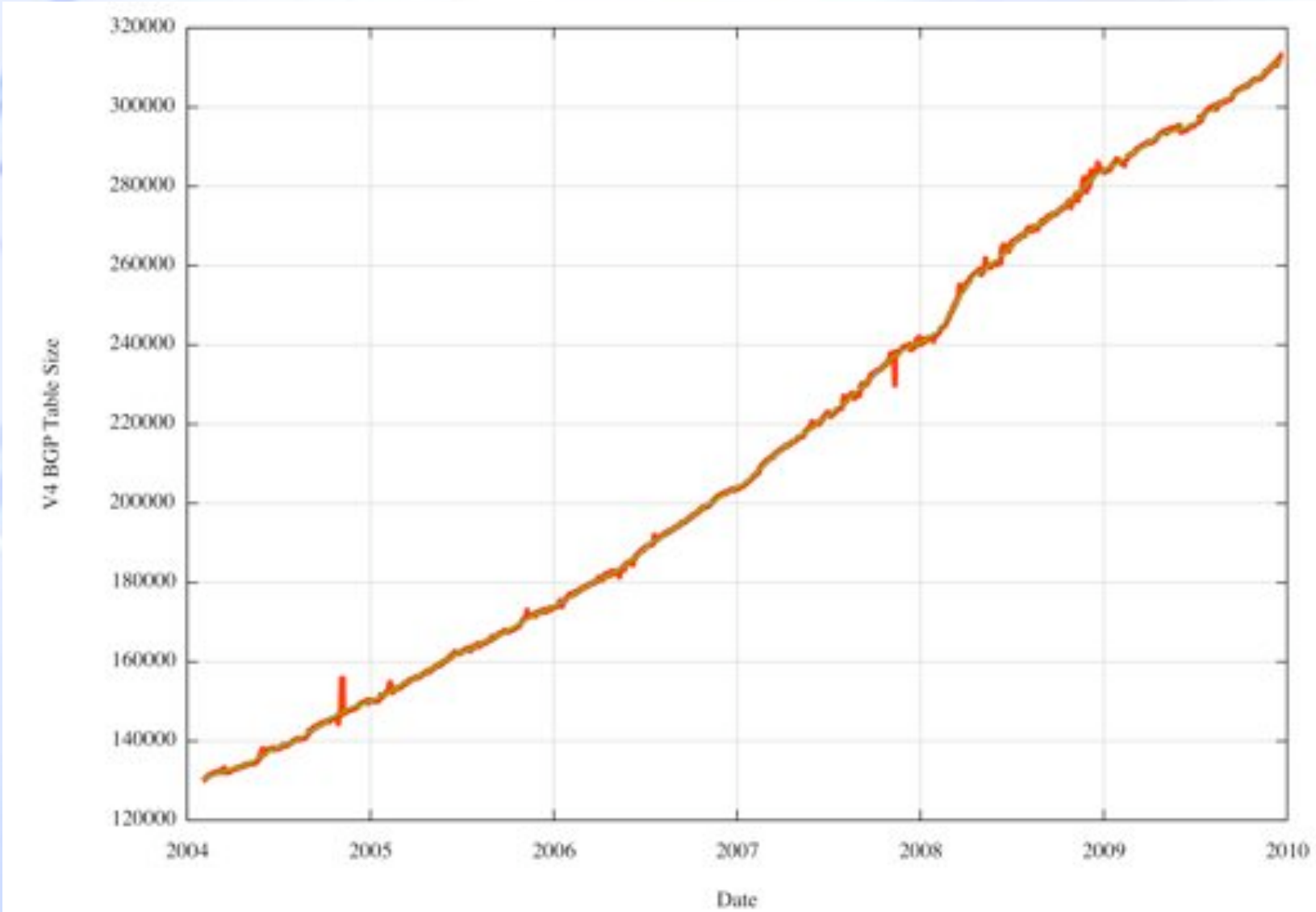
dot.com bubble crash and beyond (2000+)



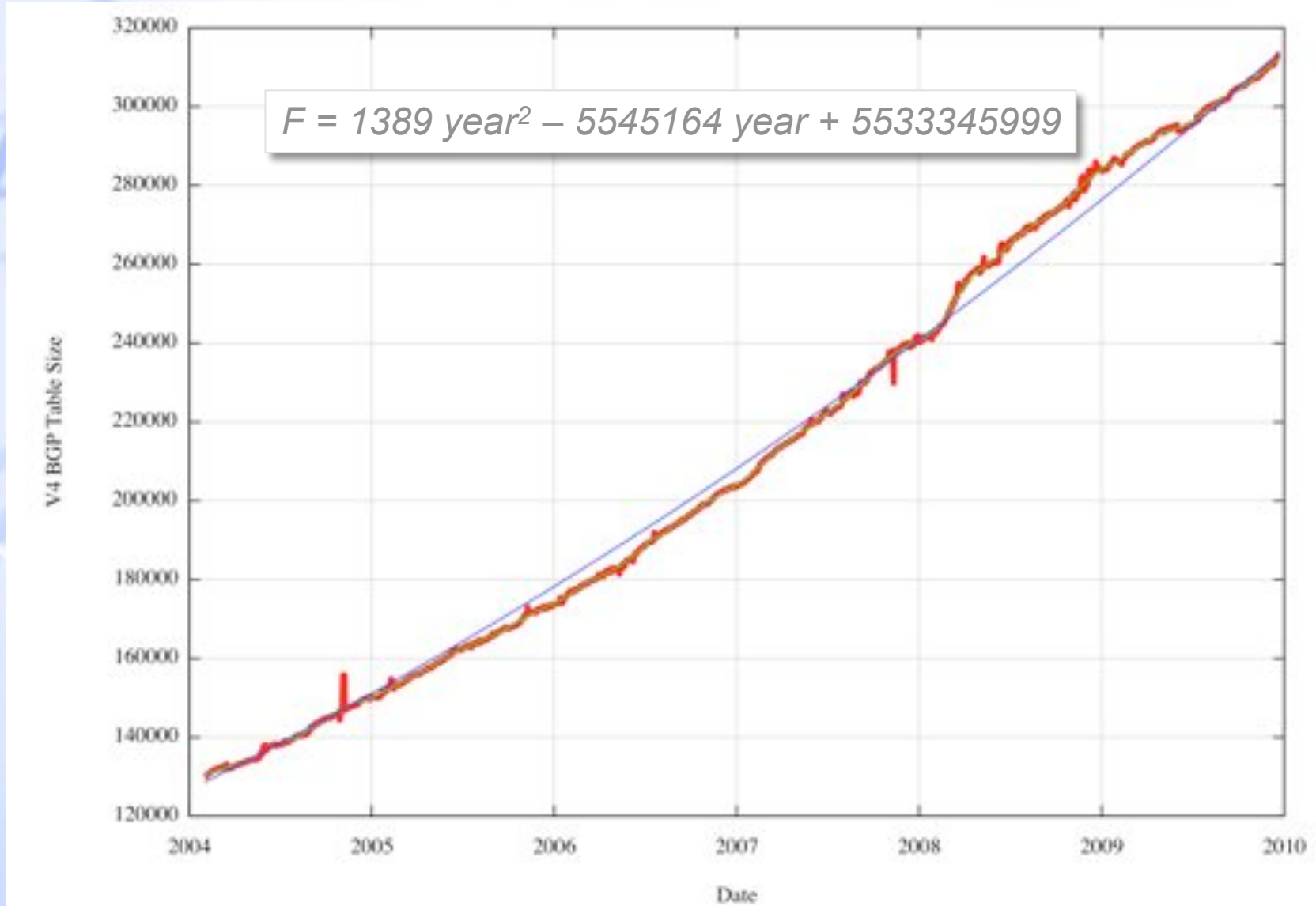
Linear growth (2003-2004)



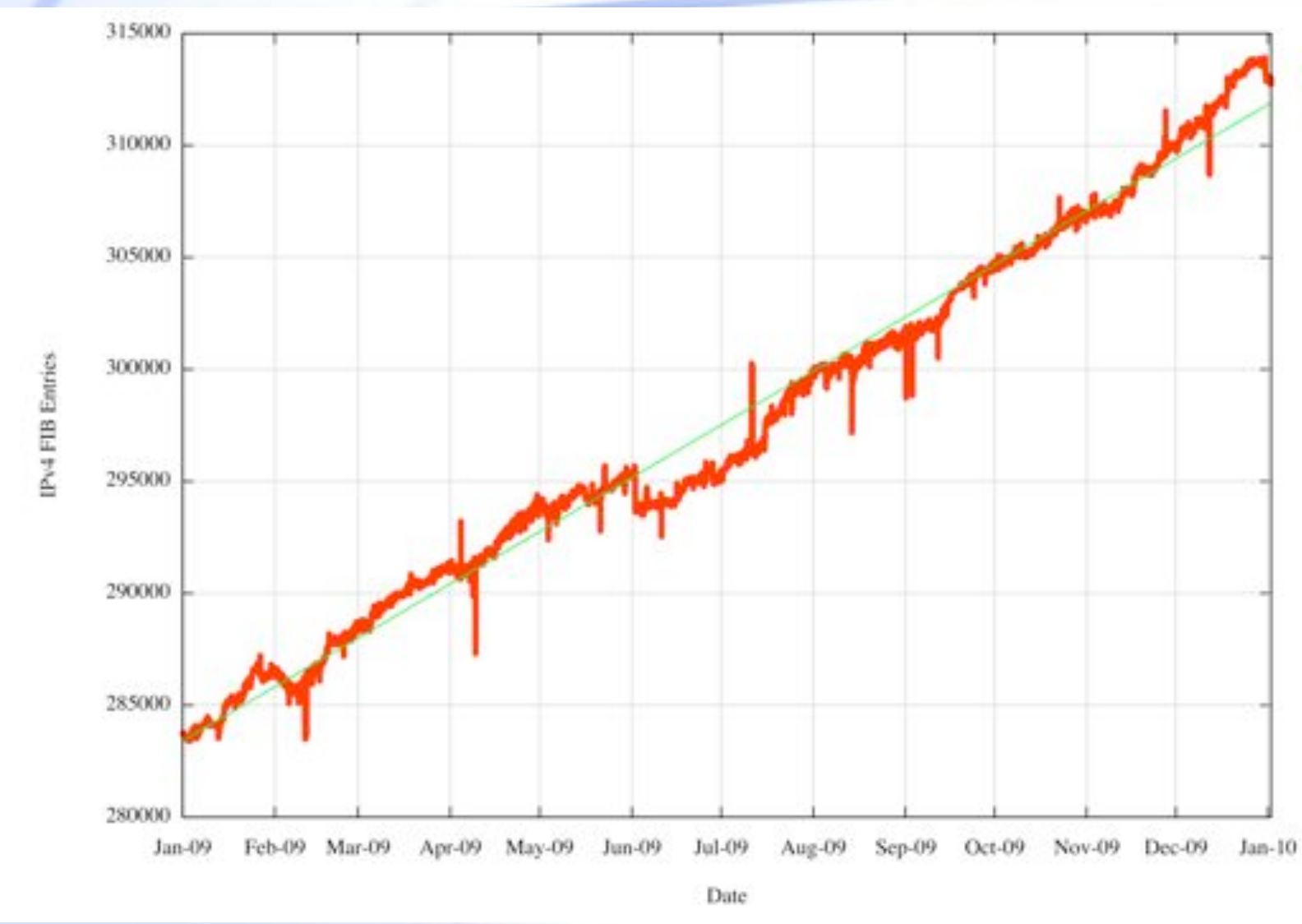
Last years (2004-2010)



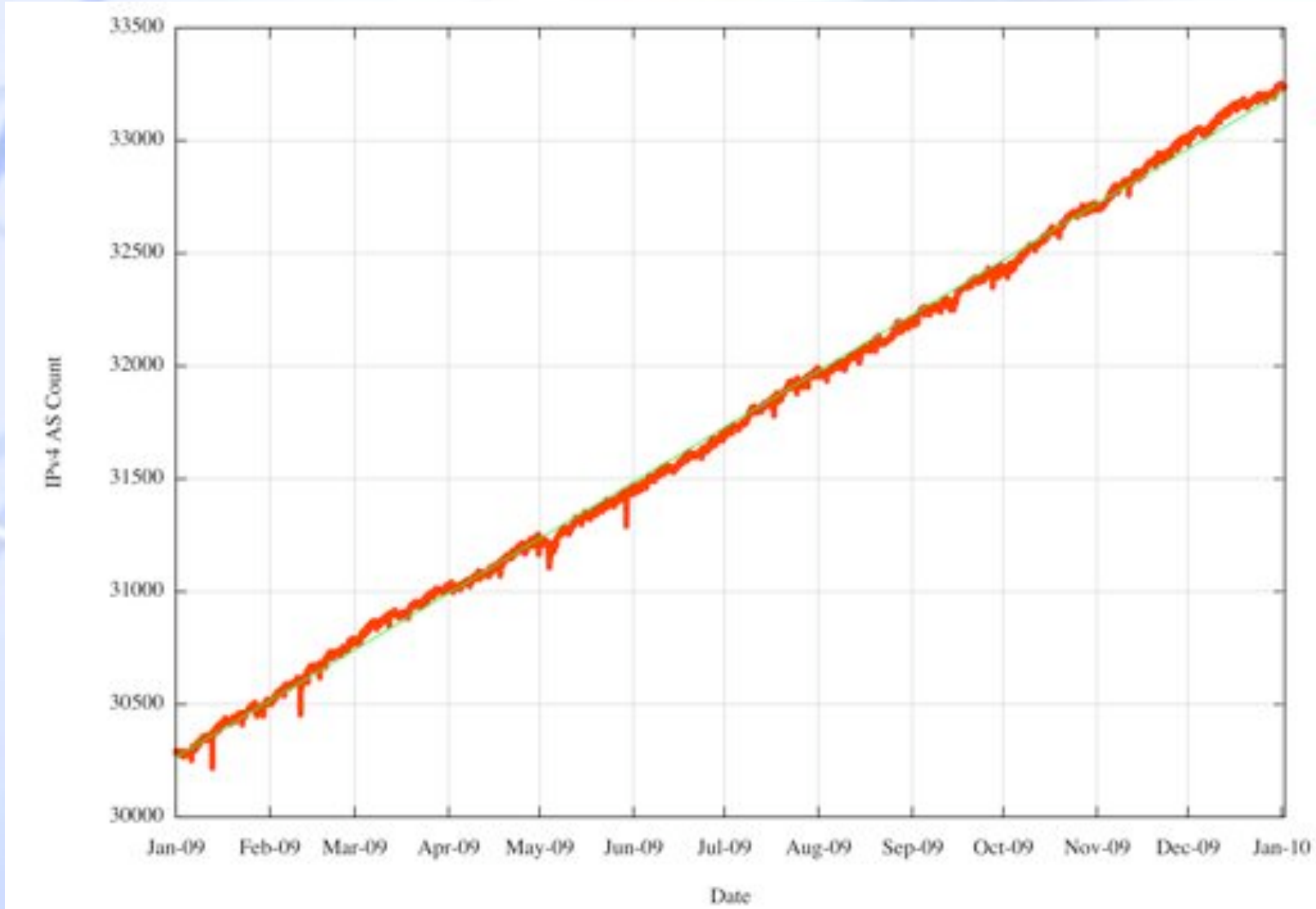
Quadratic Growth Model (2004-2010)



A closer look: 2009



IPv4 AS Count in 2009



The Internet in 2009

	Jan-09	Dec-09	
Prefix Count	283,000	312,000	**+10%**
AS Count	30,200	33,200	**+10%**

- **The IPv4 Routing table grew by 10% over 2009**
  - **compared with 12% - 15% growth in 2008**
- **Is this an indicator of reduced growth overall in the Internet?**
- **Or an indicator of reducing diversity in the supply side, and increasing market dominance by the larger providers?**

BGP Table Size Predictions

Jan 2010	$313{,}000_4 + 2{,}400_6$ entries
2011	$356{,}000_4 + 3{,}600_6$ entries
2012*	$400{,}000_4 + 5{,}000_6$ entries
2013*	$447{,}000_4 + 6{,}700_6$ entries
2014*	$496{,}000_4 + 8{,}600_6$ entries

** These numbers are dubious due to IPv4 address exhaustion pressures. It is possible that the number will be larger than the values predicted by this model.*

BGP Scaling and Table Size

- As we get further into the IPv6 transition we may see:
  - accelerated IPv4 routing fragmentation as an outcome from the operation of a V4 address trading market that starts to slice up the V4 space into smaller routed units
  - parallel V6 deployment that picks up pace
- These projections of FIB size are going to be *low*.
- Just how low it will be is far harder to estimate.

Is this a Problem?

- What is the anticipated end of service life of the core routers?
- What's the price/performance curve for forwarding engine ASICS?
- What's a sustainable growth factor in FIB size that will allow for continued improvement in unit costs of routing?
- What is a reasonable margin of uncertainty in these projections?

BGP Scaling and Stability

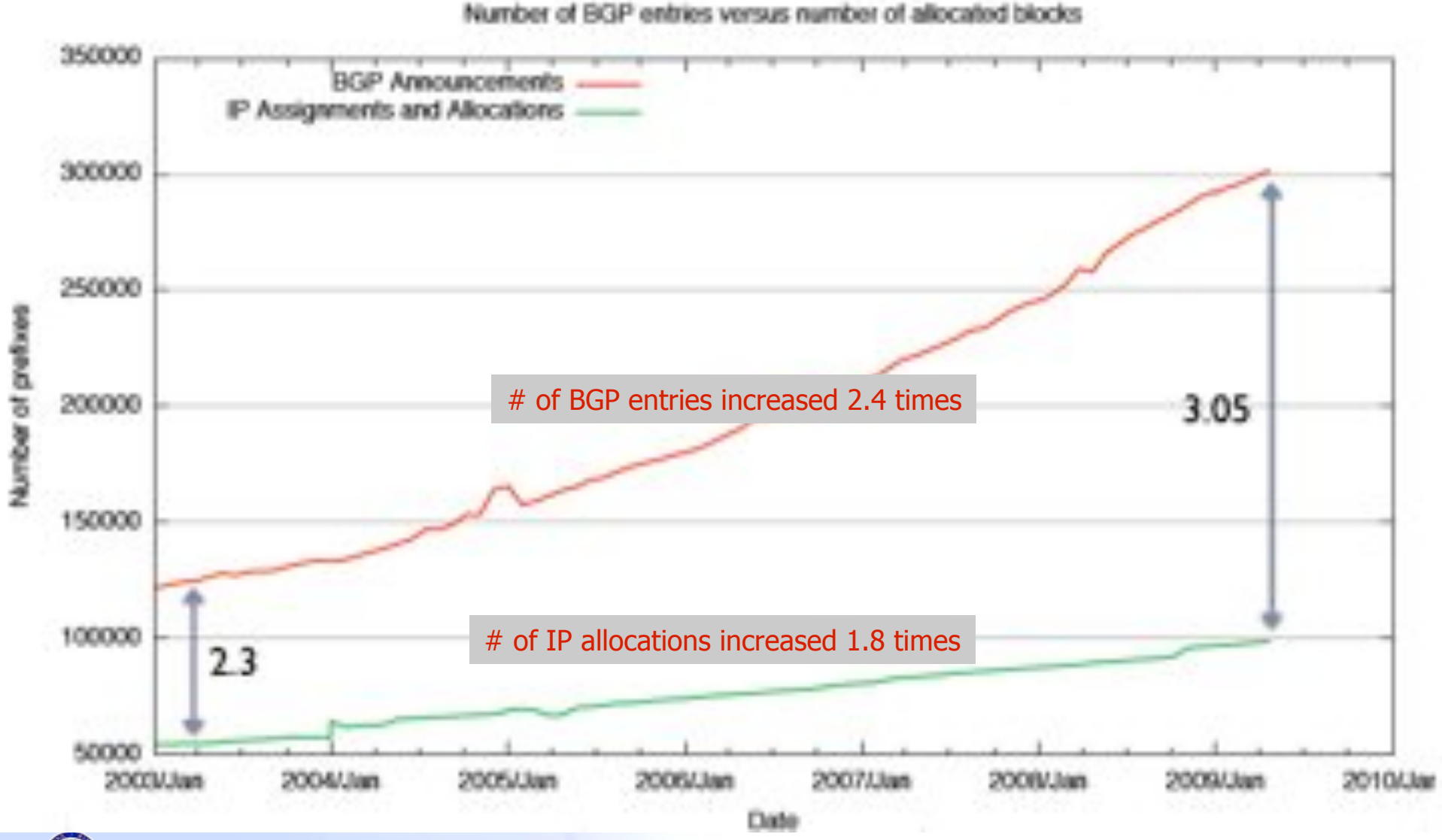
Is it the size of the RIB **or the** level of dynamic update and routing stability **that is the concern here?**

Later we'll take a look at update trends in BGP…

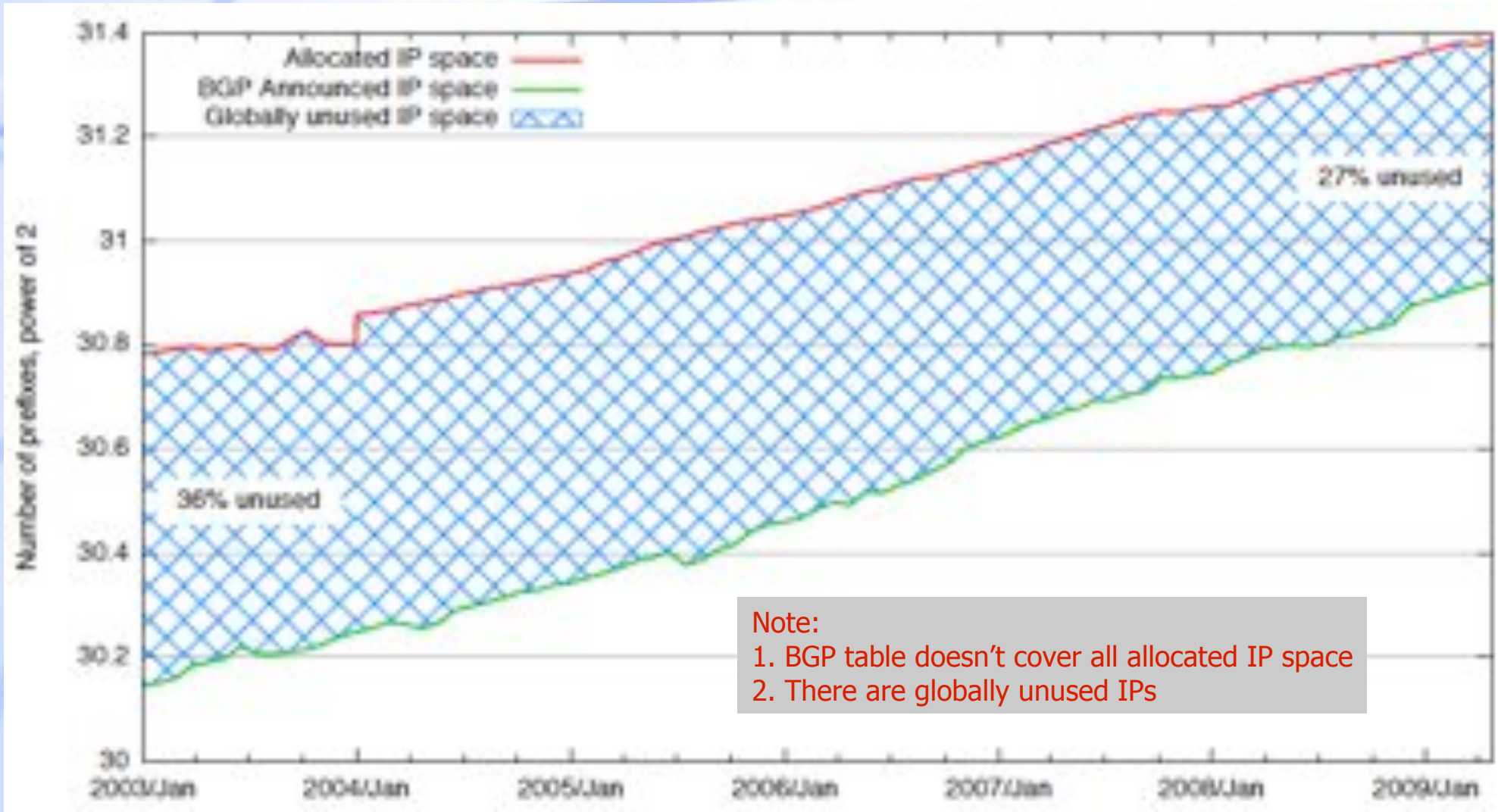
Allocation of IP addresses by RIRs

- ISPs obtain addresses blocks and further delegate to customers
  - Provider Assigned (PA) IP blocks
- Customers obtain addresses blocks directly from RIRs
  - Provider Independent (PI) IP blocks

BGP announced prefixes versus IP allocated prefixes (BGP vs RIRs)

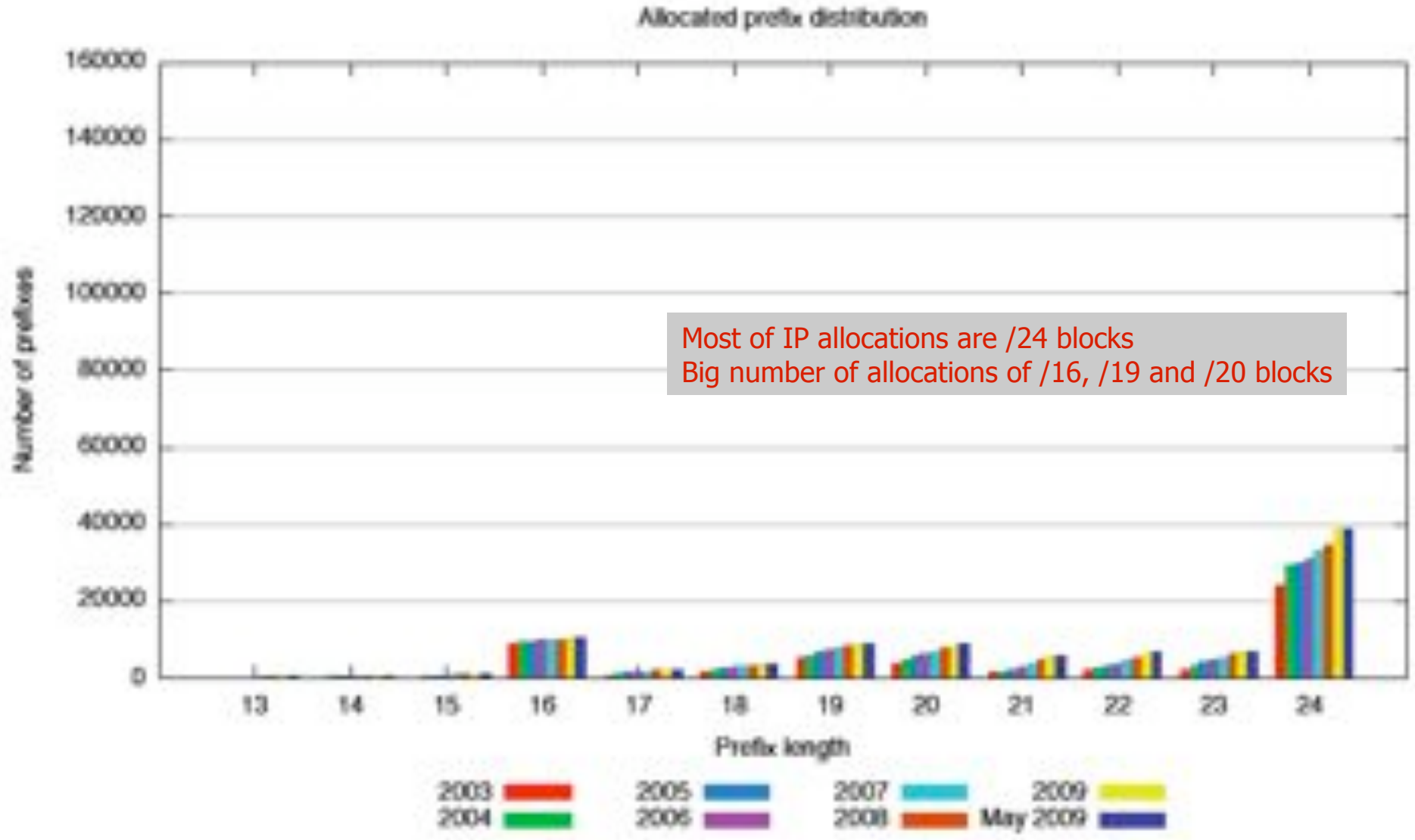


Allocated vs. announced IP space

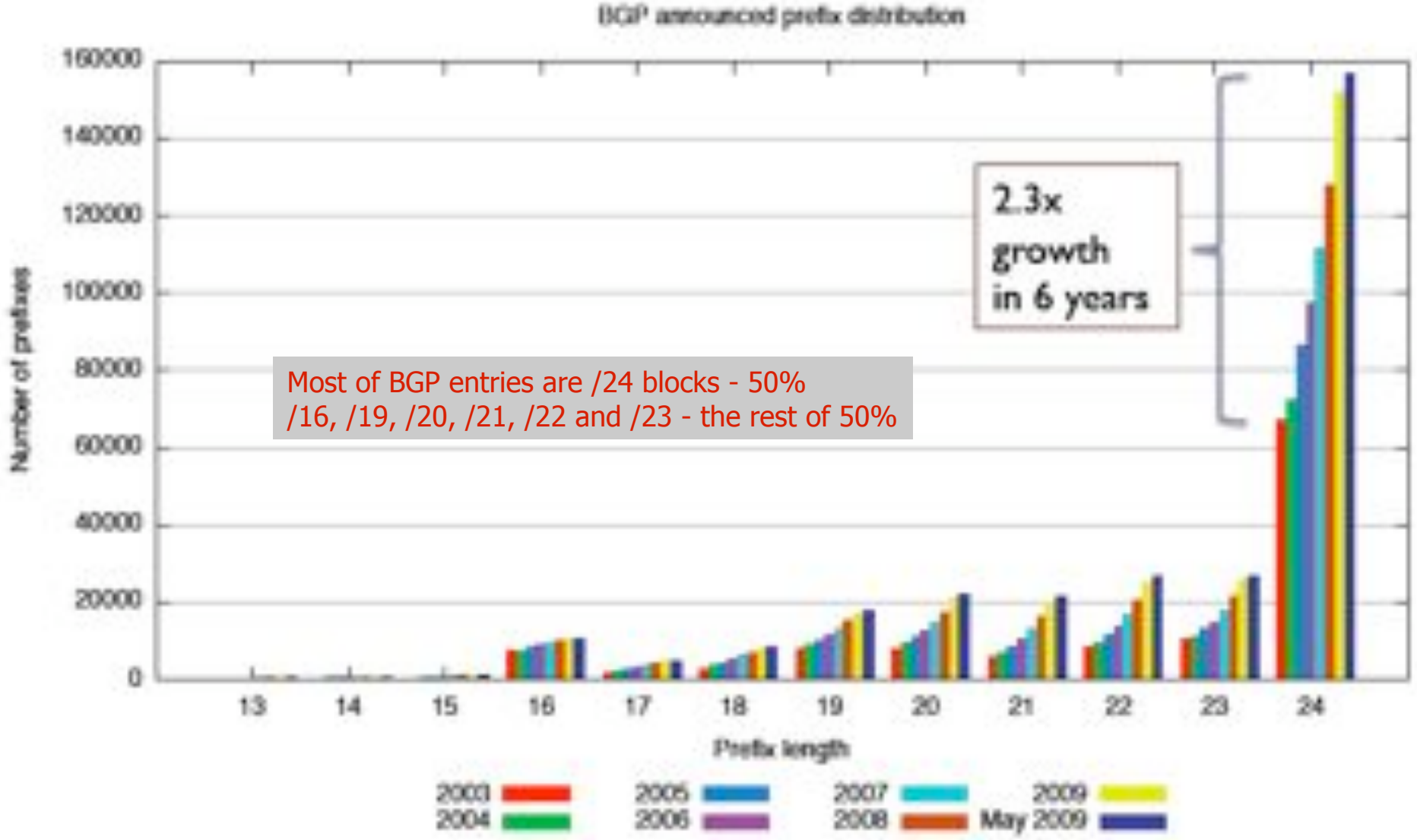


Note:
1. BGP table doesn't cover all allocated IP space
2. There are globally unused IPs

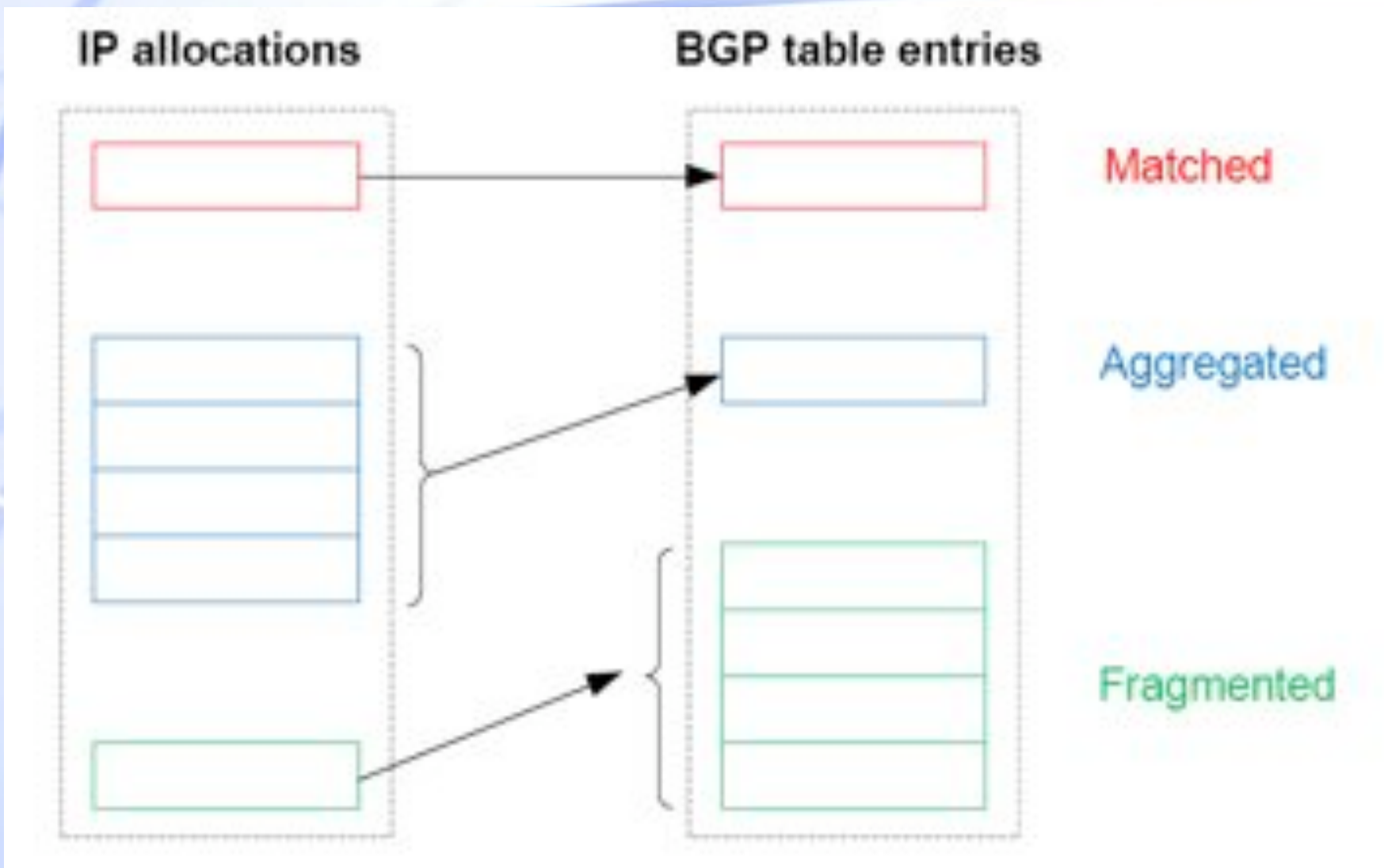
Distribution of prefix lengths: IP allocations



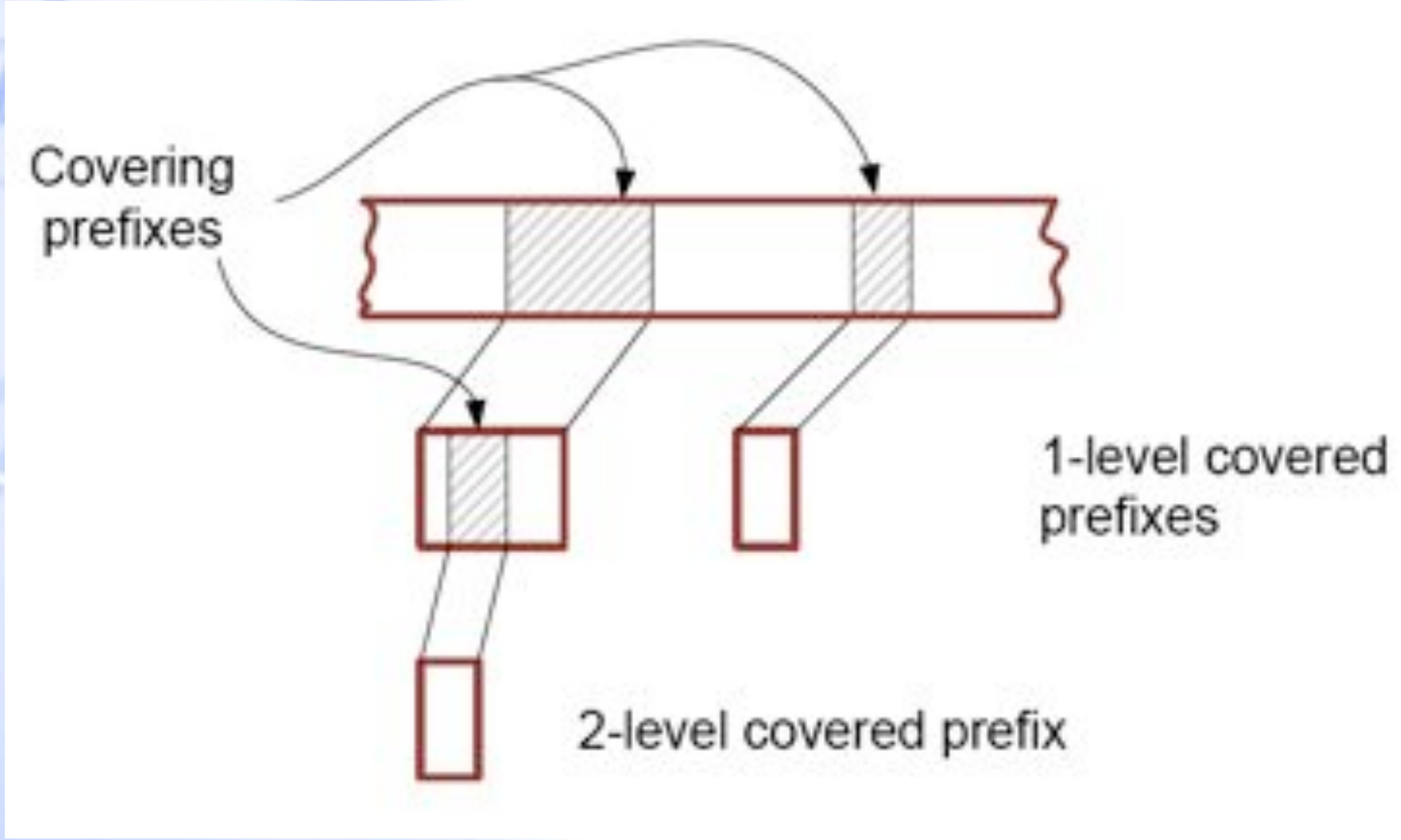
Distribution of prefix lengths BGP announcements



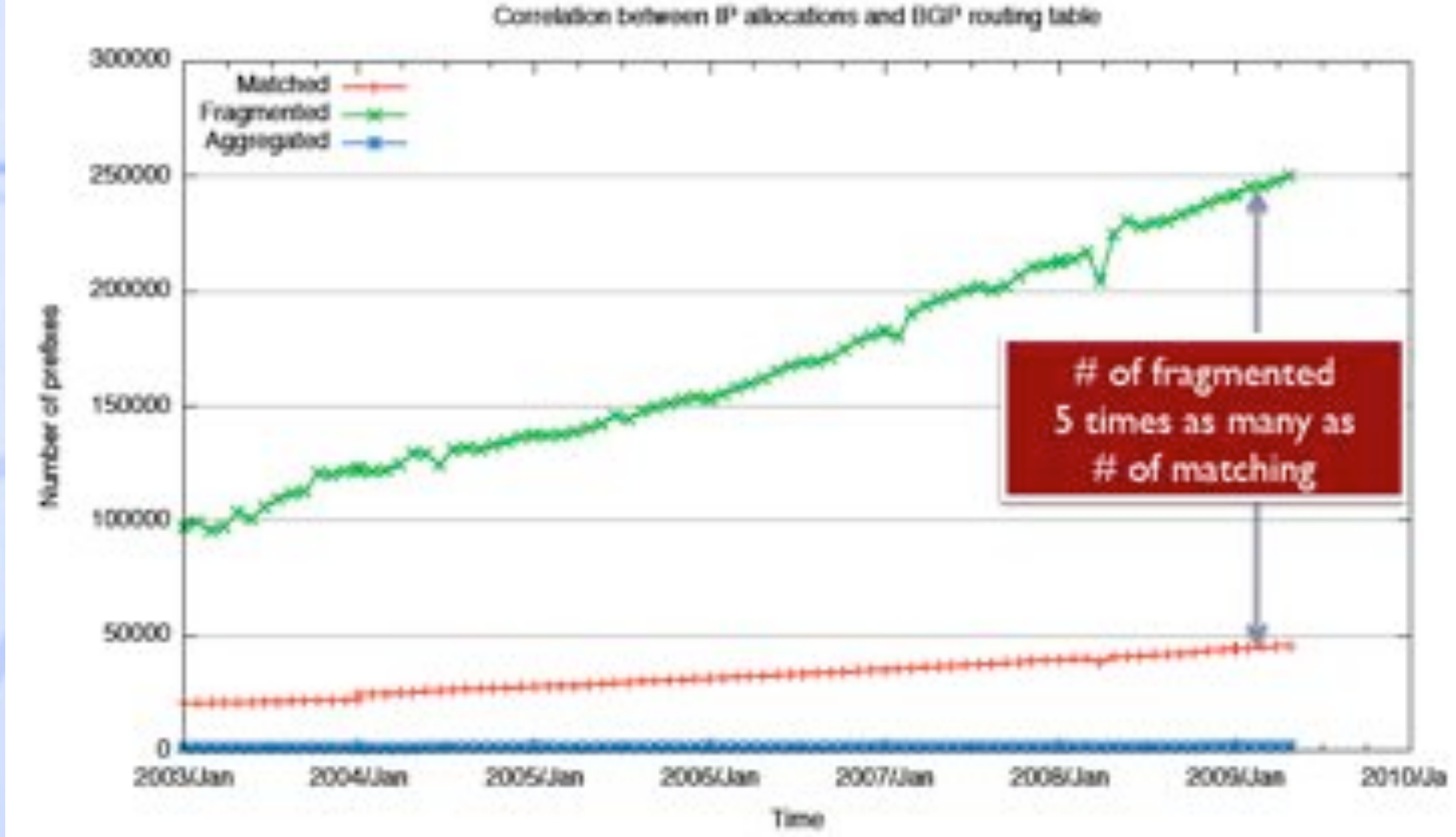
Some definitions



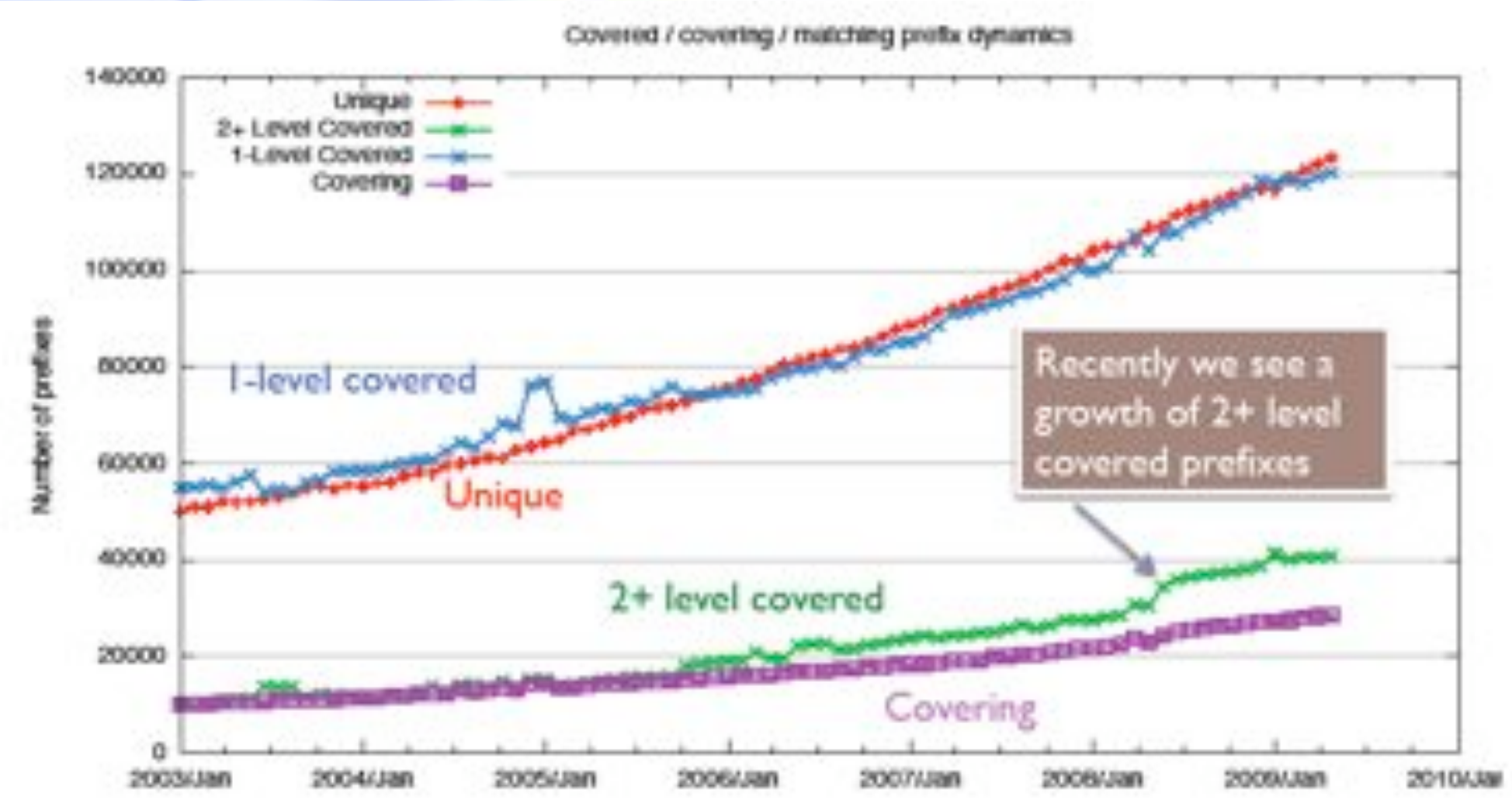
Some definitions



Impact of the fragmentation on the BGP routing table size



Duplication in BGP routing table
Unique, Covered and Covering prefixes



BGP growth and IP allocation

- BGP table has more than doubled in 6 years
  - Even though the growth rate is not exponential
- The BGP table growth outstrips IP allocation rate
- Multihoming and traffic engineering techniques introduce redundancy in BGP table (58% in 2009)

Deaggregation, why?

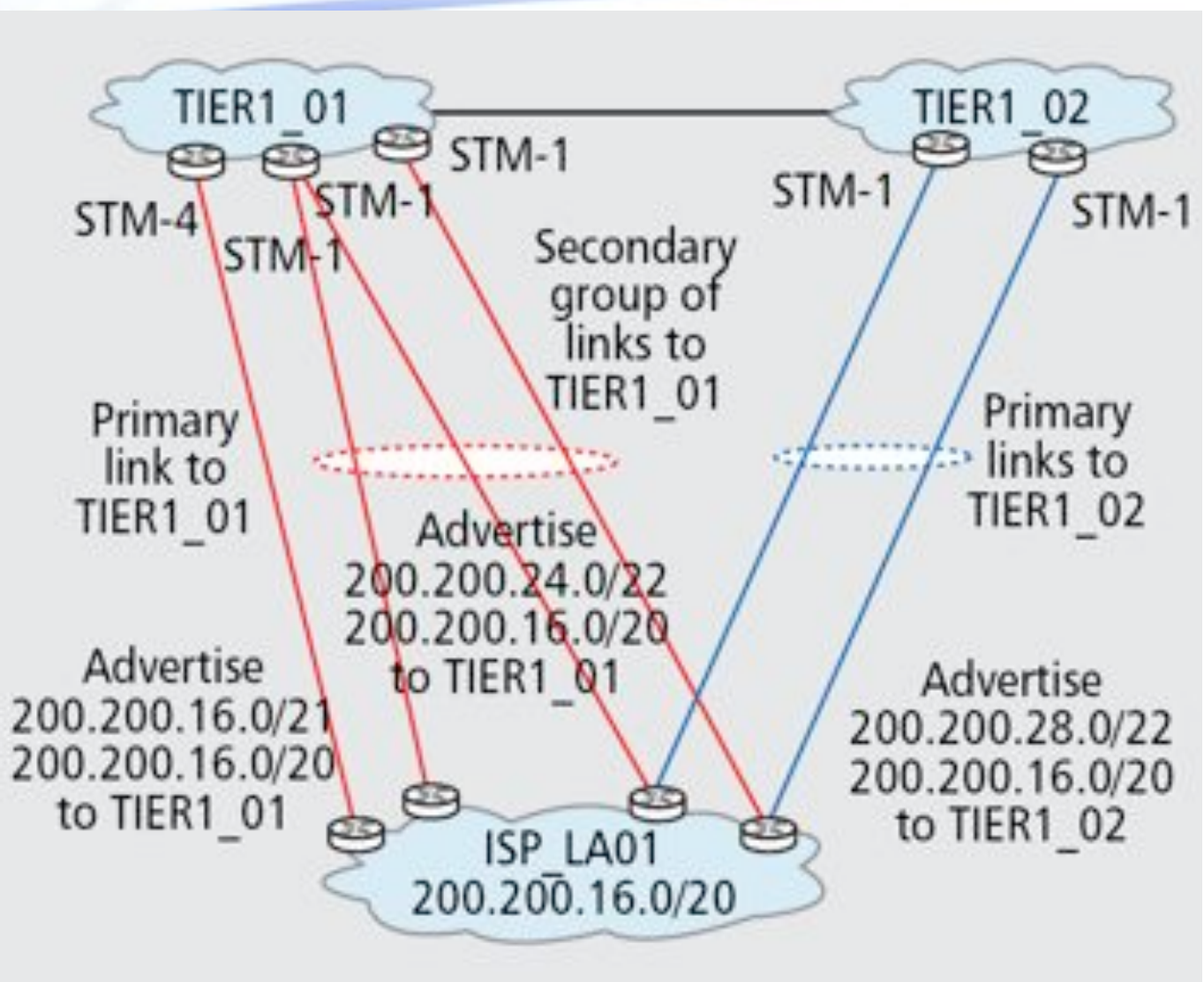
- Traffic Engineering for Multihoming
  - Spraying out /24s hoping it will work…
  - …rather than do any real engineering (work with BGP attributes!)
- Security
  - Announcing /24s to prevent DoS attacks
  - Announcing only address space in use
  - The rest attracts ‘noise’”
- Leakage of iBGP outside of local AS
  - eBGP is NOT iBGP…
- Legacy Assignments
  - Both RIR and legacy assignments in place
- Commercial reasons

Deaggregation factor

- *Prefixes in Global Routing Table / Aggregatable Prefixes*
  - **Measure of Routing Table size/Aggregated Size**
- **Global value has been increasing slowly and steadily since records began in 1999**

Deaggregation example

- Multihomed ISP with links of different capacity, load sharing, and backup routing policy



Deaggregation impacts

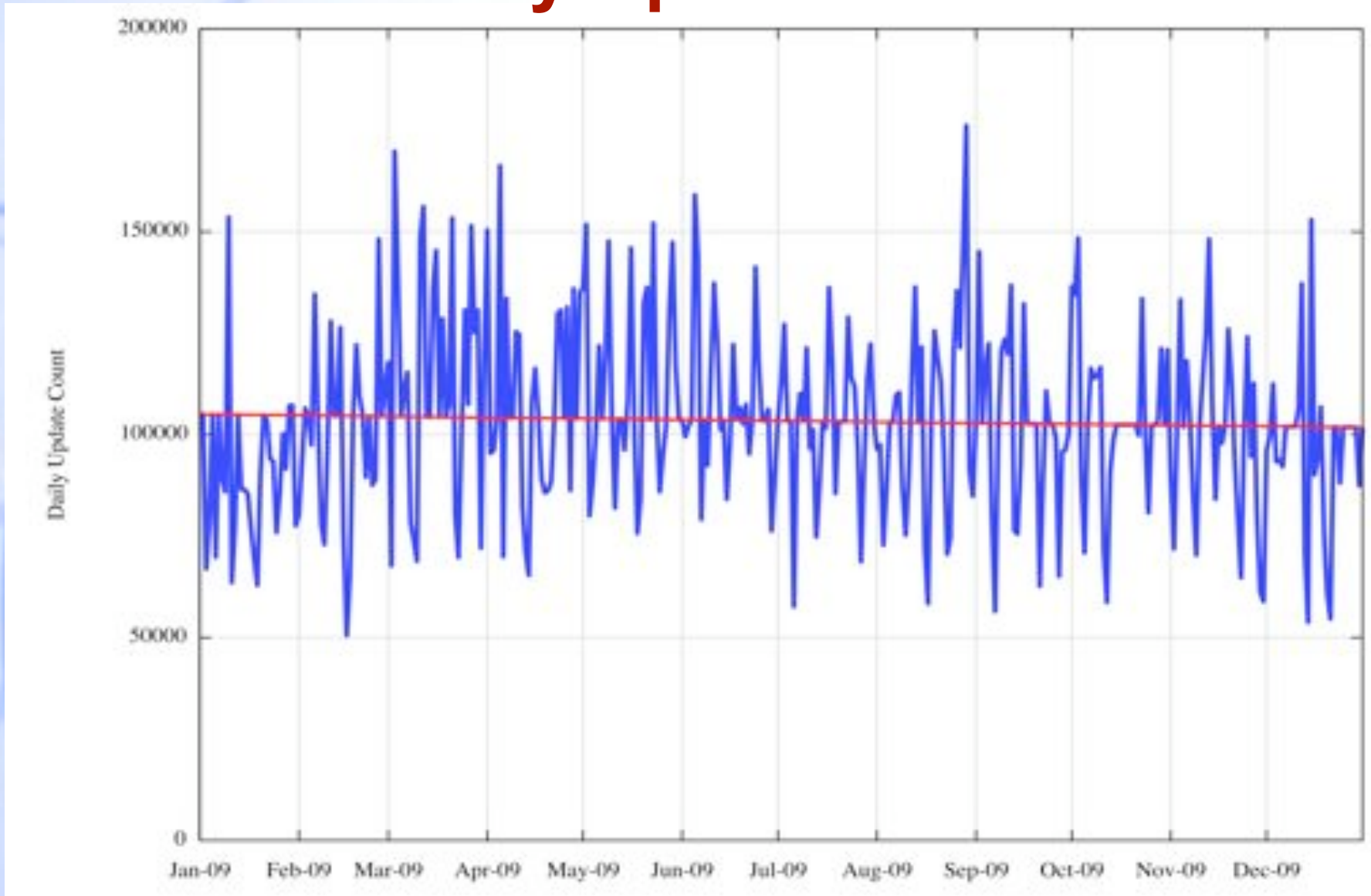
- **Router memory & processing power**
  - Shortens router life time when growth requirements are underestimated
- **Routing System convergence**
  - Larger routing table → slowed convergence
- **Network Performance & Stability**
  - Slowed convergence → slowed recovery from failure
  - Slowed recovery → longer downtime

- **Is BGP churn growing as much as BGP table size?**

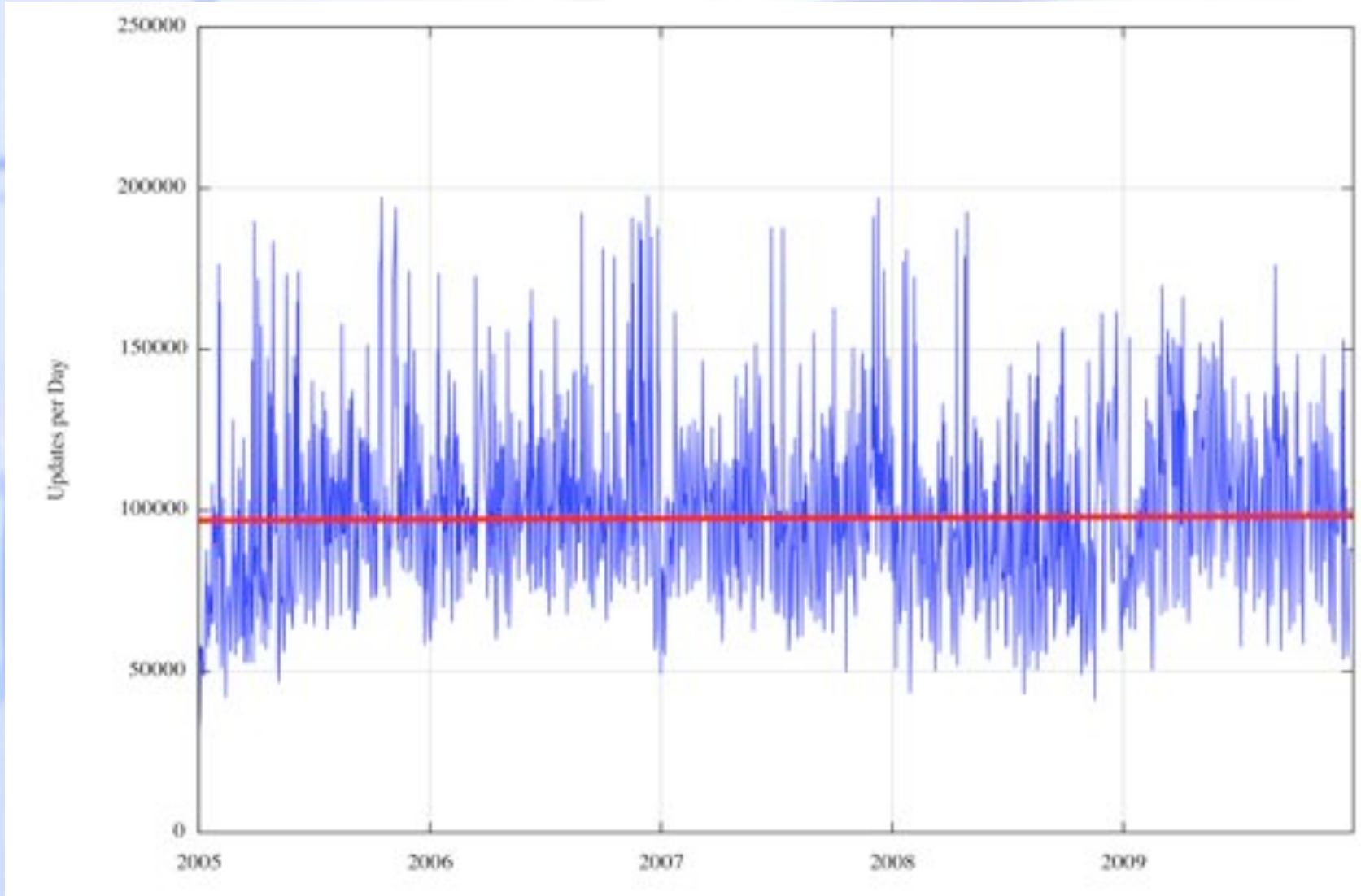
Causes for the observed BGP churn rate

- The size of the network
  - More elements that can fail/change/act
- The structure of the network topology
  - Who peers with who?
  - How many and which providers does an AS have?
  - Depth of Internet hierarchy/path lengths
- Policies and protocol configuration
  - MRAI timer
  - Route Flap Dampening
  - Route filtering and aggregation
- Event types and frequencies
  - Prefix withdrawals, link failures, TE operation…

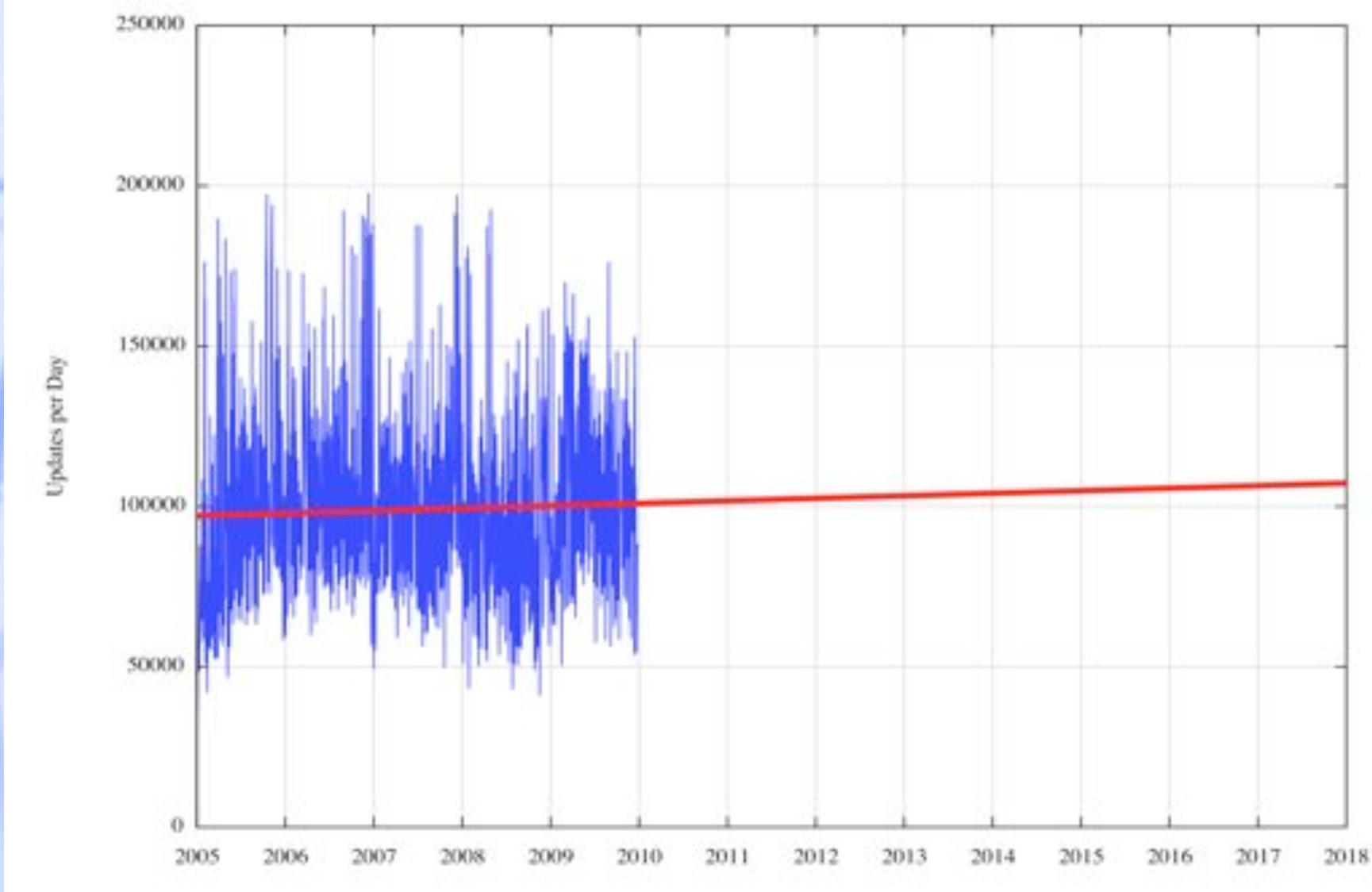
Daily Updates - 2009



BGP Updates – 2005 - 2009
Extended Data Set

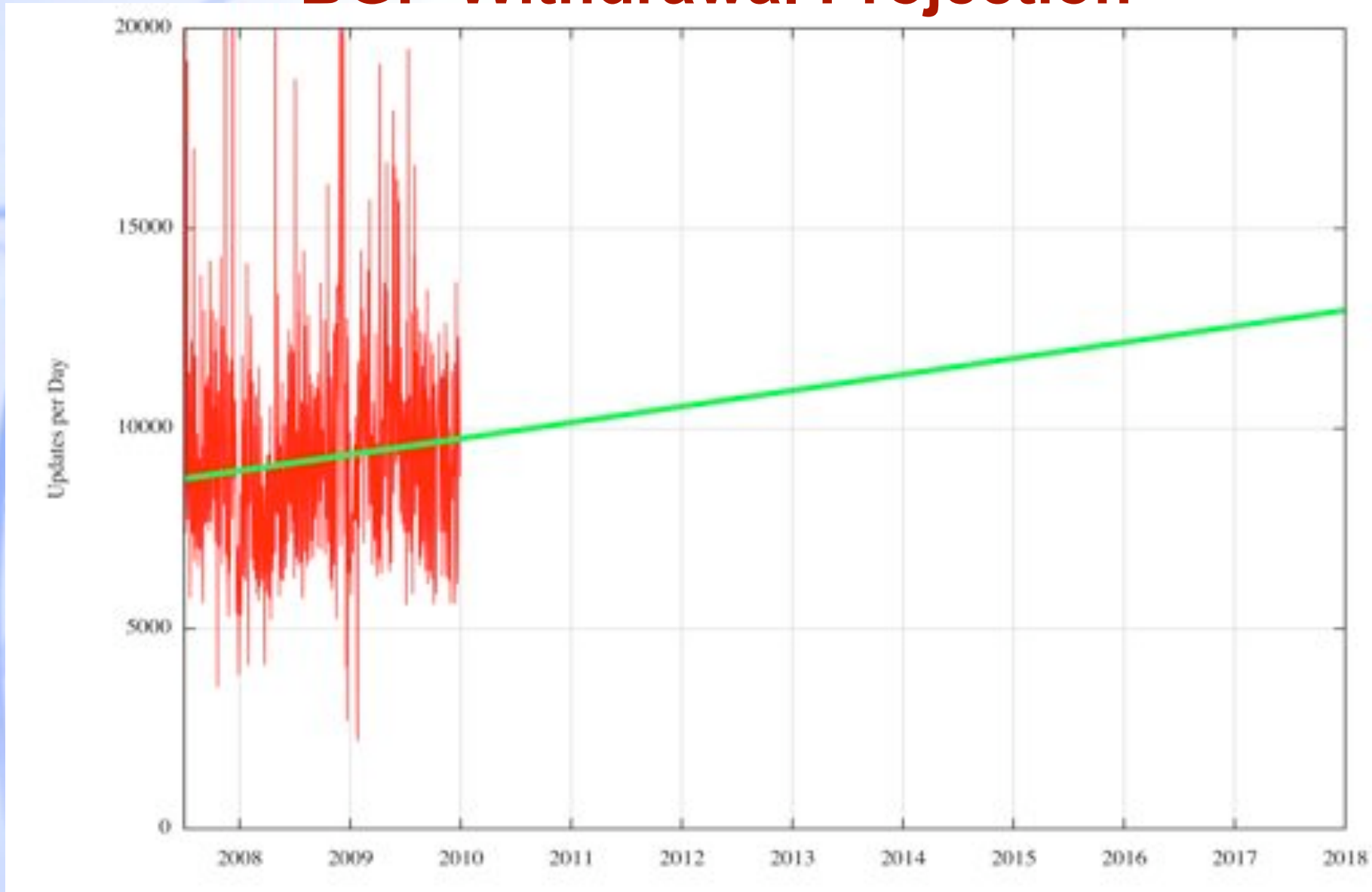


BGP Update Projection



Daily Withdrawals - 2009

BGP Withdrawal Projection

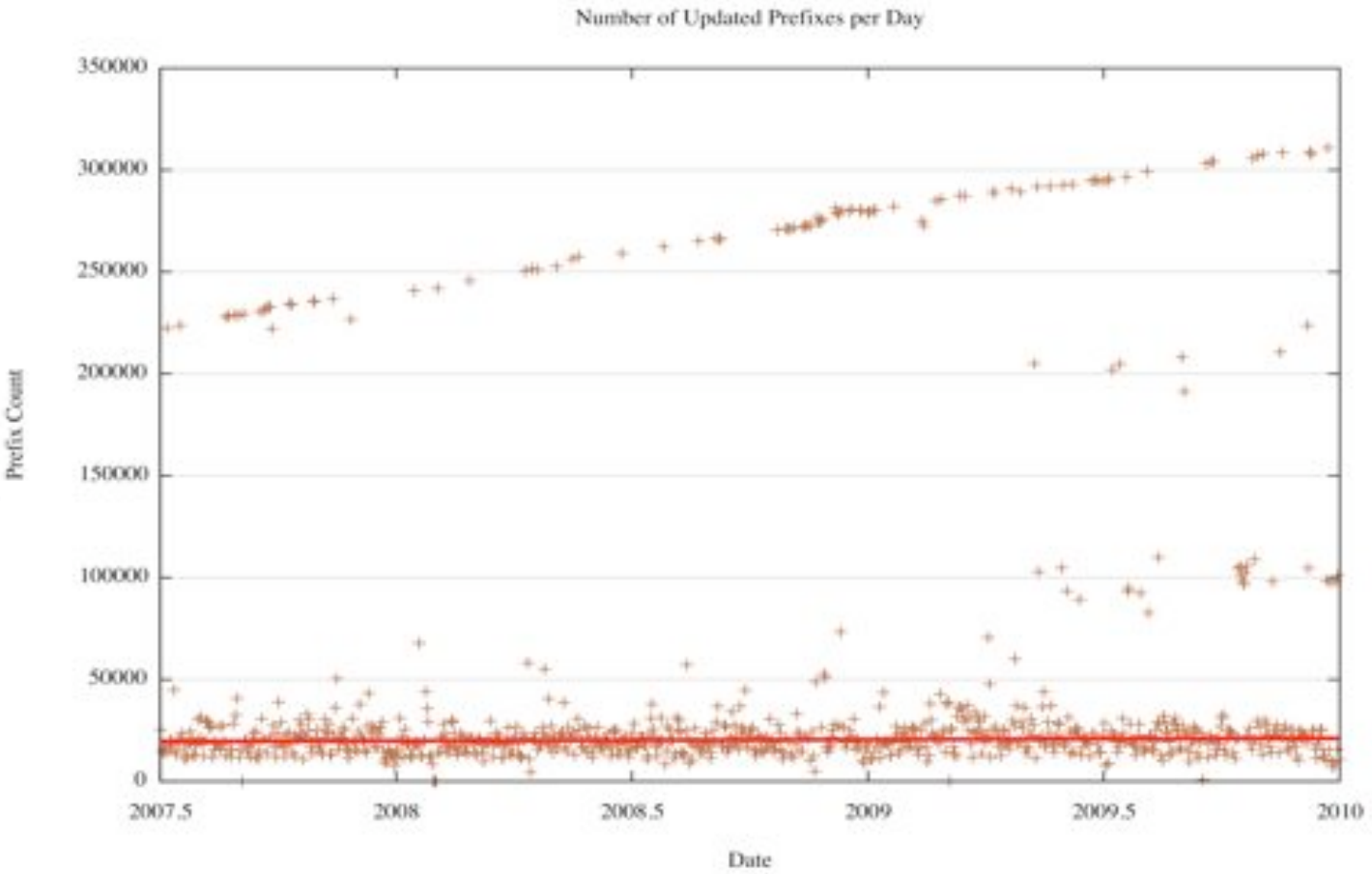


Why is this so flat?

- Growth rates of BGP update activity appear to be far smaller than the growth rate of the routing space itself
  - Remember we're not considering duplicates
- Another study from a core perspective leads to the same conclusion when duplicates are taken aside
- Why are the levels of growth in BGP updates not proportional to the size of the routing table?

(In)Stability

- Over the past 1,000 days the number of announced prefixes increased by 40% (225,000 to 313,000)
- But the average number of unstable prefixes on any day increased by only 7% in 1,000 days(19,600 to 21,000)
- Routing instability is not directly related to the number of advertised objects
- What is routing instability related to?



Convergence in BGP

- BGP is a path vector protocol
- This implies that BGP may send a number of updates in a tight "cluster" before converging to the "best" path
- This is clearly evident in withdrawals and convergence to (longer) secondary paths

Withdrawal at source at 08:00:00 03-Apr of 84.205.77.0/24 at MSK-IX, as observed at AS 2.0
Announced AS Path: <4777 2497 9002 12654>
Received update sequence:

08:02:22 03-Apr	+ <4777 2516 3549 3327 12976 20483 31323 12654>
08:02:51 03-Apr	+ <4777 2497 3549 3327 12976 20483 39792 8359 12654>
08:03:52 03-Apr	+ <4777 2516 3549 3327 12976 20483 39792 6939 16150 8359 12654>
08:04:28 03-Apr	+ <4777 2516 1239 3549 3327 12976 20483 39792 6939 16150 8359 12654>
08:04:52 03-Apr	- <4777 2516 1239 3549 3327 12976 20483 39792 6939 16150 8359 12654>

1 withdrawal at source generated a convergence sequence of 5 events, spanning 150 seconds

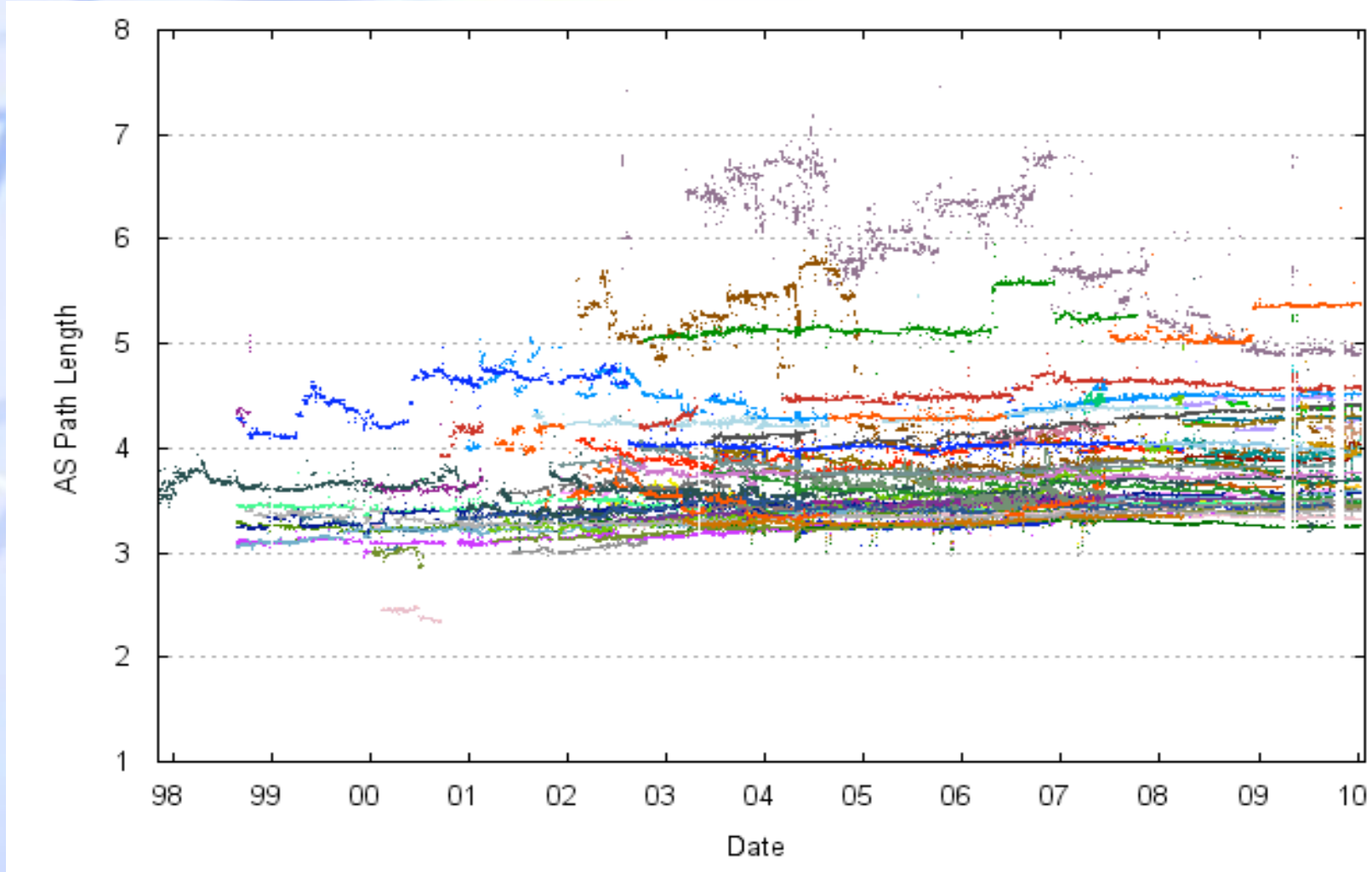
Observations

- There are two types of updates:
  - updates that are part of a convergence sequence
  - updates that are single isolated events
- Average Convergence Time
  - An unstable prefix takes, on average around 70 seconds to reach a stable state (between 2 and 3 MRAI intervals)
  - This has remained constant for almost two years
  - As the network expands, the distance vector operation to achieve convergence is taking the same elapsed time
- Average Convergence Updates
  - The average number of updates to reach a converged state has remained constant for the past 2 ½ years at 2.7 updates

Observations

- The growth of the network appears to have been achieved by increasing the density of connectivity, rather than increasing the network's diameter
- There is a reasonable correlation between AS Path Length Distribution and Convergence Update Distribution
- The number of updates to reach convergence and the time to reach convergence is related to AS Path Length for most of all instability events
- Persistent instability events (around 1.3% of all such events) are probably related to longer term instability that may have causes beyond conventional protocol convergence behaviour of BGP

Average AS Path Length is long term stable



What is going on?

- The convergence instability factor for a path vector protocol like BGP is related to the AS path length, and average AS Path length has remained steady in the Internet for some years
- Taking MRAI factors into account, the number of received Path Exploration Updates in advance of a withdrawal is related to the propagation time of the withdrawal message. This is approximately related to the average AS path length
- Today's Internet of 30,000 ASes is more densely interconnected, but not more "stringier" than the internet of 5,000 ASes of 2,000
- This is consistent with the observation that the number of protocol path exploration transitions leading to convergence to a new stable state is relatively stable over time

- **Beware!**
  - ***maybe densification limits the visibility of routing changes (?)***

Conclusions

- BGP table size predictions, considering both IPv4 and IPv6 are fairly low
  - Scalability doesn't seem to be compormised by this factor
  - Even though de-aggregation is persistently growing
- BGP churn behaviour is independent of network growth
  - Beacuse the Internet is getting more dense, but the average AS Path lenght is stable
- Beware!
  - We're not considering duplicates, which accounts for 40% of BGP updates
  - We need to consider the effect of BGP churn intra-AS, where hardware is more stressed (route reflectors)

References

1. *Analyzing the Internet BGP Routing Table*, Geoff Huston, The Internet Protocol Journal - Volume 4, Number 1, March 2001.

2. *BGP in 2009*, Geoff Huston, RIPE 60, Prague, May 2010.

3. *BGP Routing Table: Trends and Challenges*, Alexander Afanasyev, Neil Tilley, Brent Longsta, and Lixia Zhang, 2010.

4. *BGP Churn Evolution: A perspective from the core*, Ahmed Elmokashfi, Amund Kvalbein, Constantine Dovrolis, INFOCOM 2010.