BGP Issues

Geoff Huston

May 2001

What is BGP?

- The Internet is composed of a collection of *networks*
  - Each *network* is autonomously managed
- The Internet uses a two layer routing hierarchy:
  - Within a network the *interior gateway protocol* manages the internal topology of the network
  - Summaries of reachable address prefixes are passed between networks using an *exterior gateway protocol*
- BGP is today's *exterior gateway protocol* for the internet
- A BGP routing table contains a set of address prefixes and the associated path of autonomous networks to transit to reach each address prefix

A sample of the BGP Table

```
show ip bgp
BGP table version is 80367535, local router ID is 203.62.248.4
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal
Origin codes: i - IGP, e - EGP, ? - incomplete

   Network          Next Hop              Metric LocPrf Weight Path
*>i3.0.0.0          134.159.0.3               55         0 16779 1 701 80 i
*>i4.0.0.0          134.159.0.3               55         0 16779 1 i
*>i6.1.0.0/16       134.159.0.3               55         0 16779 1 7170 1455 i
*>i6.2.0.0/22       134.159.0.3               55         0 16779 1 7170 1455 i
*>i6.3.0.0/18       134.159.0.3               55         0 16779 1 7170 1455 i
*>i6.4.0.0/16       134.159.0.3               55         0 16779 1 7170 1455 i
*>i6.5.0.0/19       134.159.0.3               55         0 16779 1 7170 1455 i
```

Flags   Address Prefix   Transit path to reach the address

Why measure BGP?

- BGP describes the structure of the Internet, and an analysis of the BGP routing table can provide information to help answer the following questions:

  - What is changing in the deployment environment?
  - Are these changes sustainable?
  - How do address allocation policies, BGP and the Internet inter-relate?
  - Are current address allocation policies still relevant?
 - What are sensible objectives for address allocation policies?

Techniques

- ## Passive Measurement
  - Takes measurements from a default-free router at the edge of the local network
  - Easily configured
  - Single (Filtered) view of the larger Internet
    - What you see is a collection of best paths from your immediate neighbours

eBGP

Measurement Point

Local AS
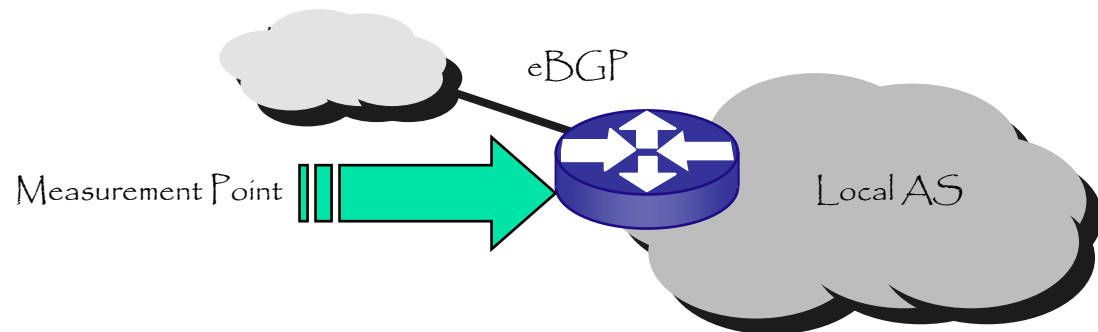
Techniques

- ## Multiple Passive measurement points
 - Measure a number of locations simultaneously
  - Can be used to infer policy

AS2

AS1

AS3

Measurement Points
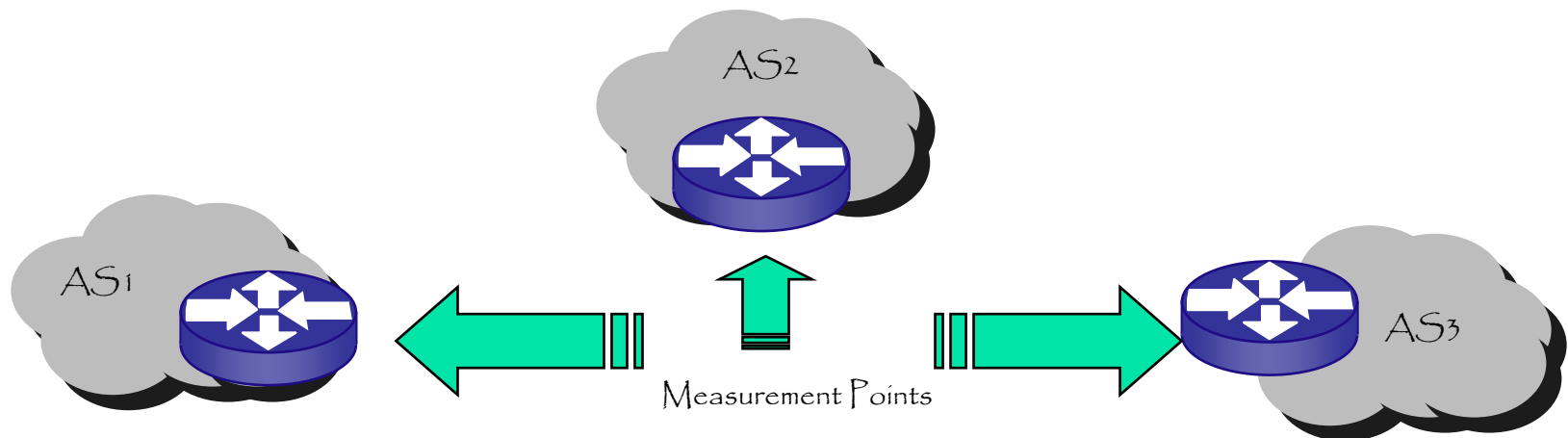
Techniques

- Single passive measurement point with multiple route feeds
  - Best example:
    - Route-views.oregon-ix.net
    - Operating since 1995
    - 42 peers
    - Uses eBGP multihop to pull in route views

Techniques

- ## Active Measurement Tests
 - ### Convergence
    - #### Announcement and withdrawal
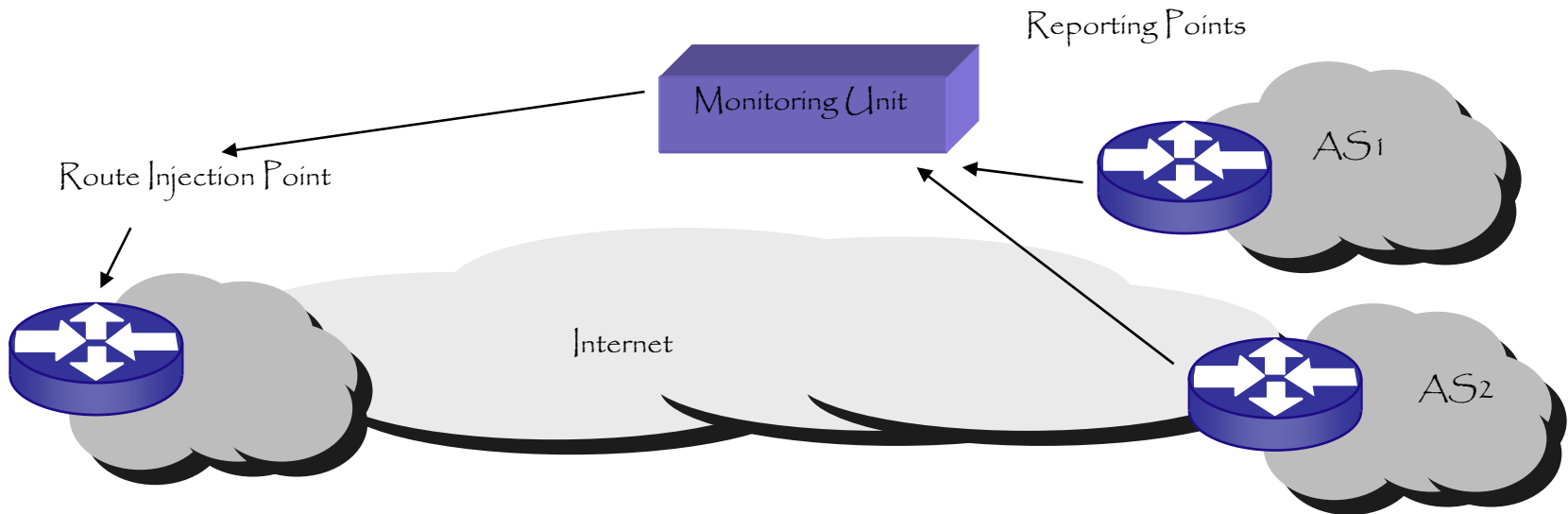
Interpretation

- BGP is not a link state protocol
- There is no synchronized overview of the entire connectivity and policy state
- Every BGP viewing point contains a filtered view of the network
  - Just because you can't see it does not mean that it does not exist
- BGP metrics are sample metrics
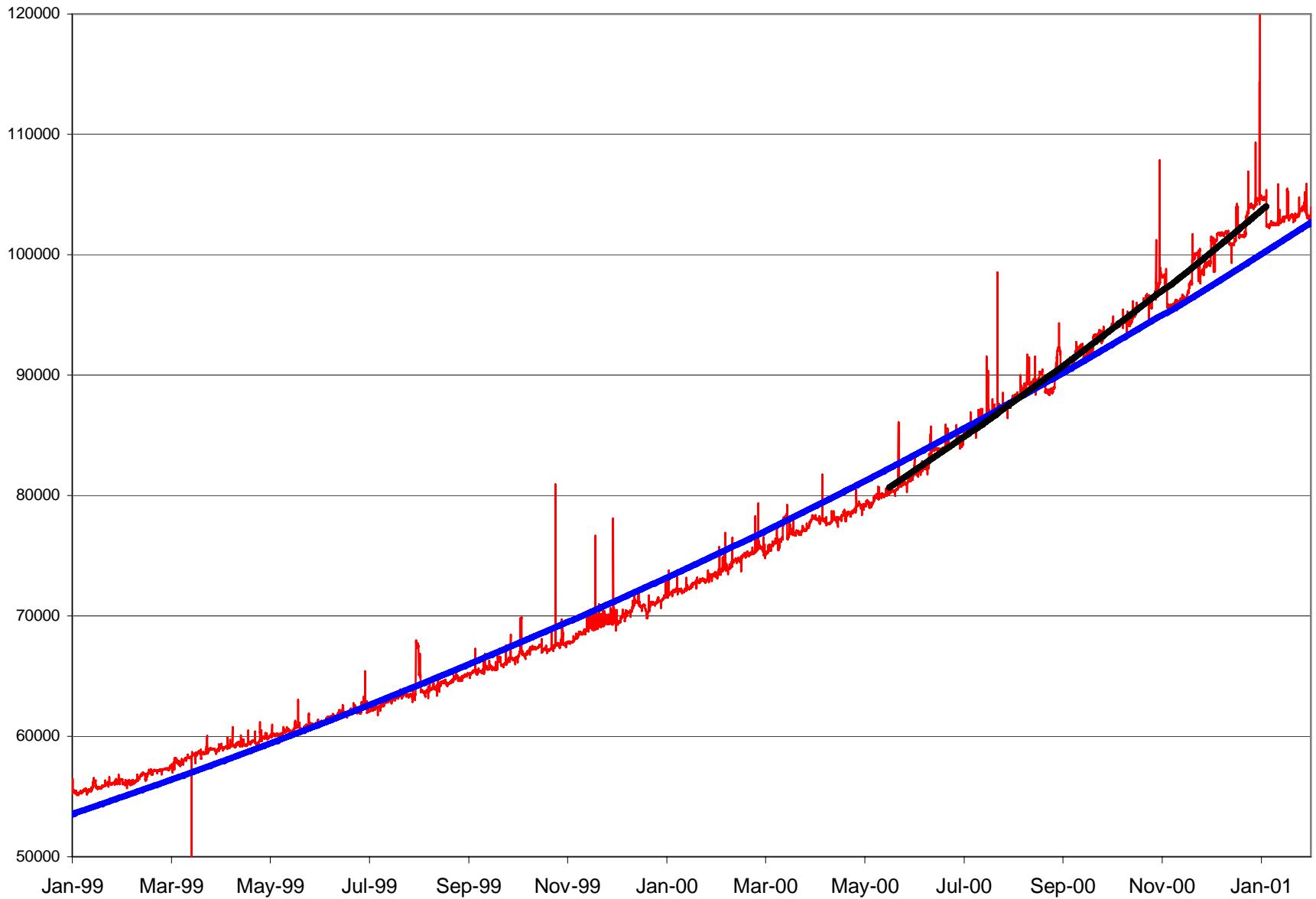
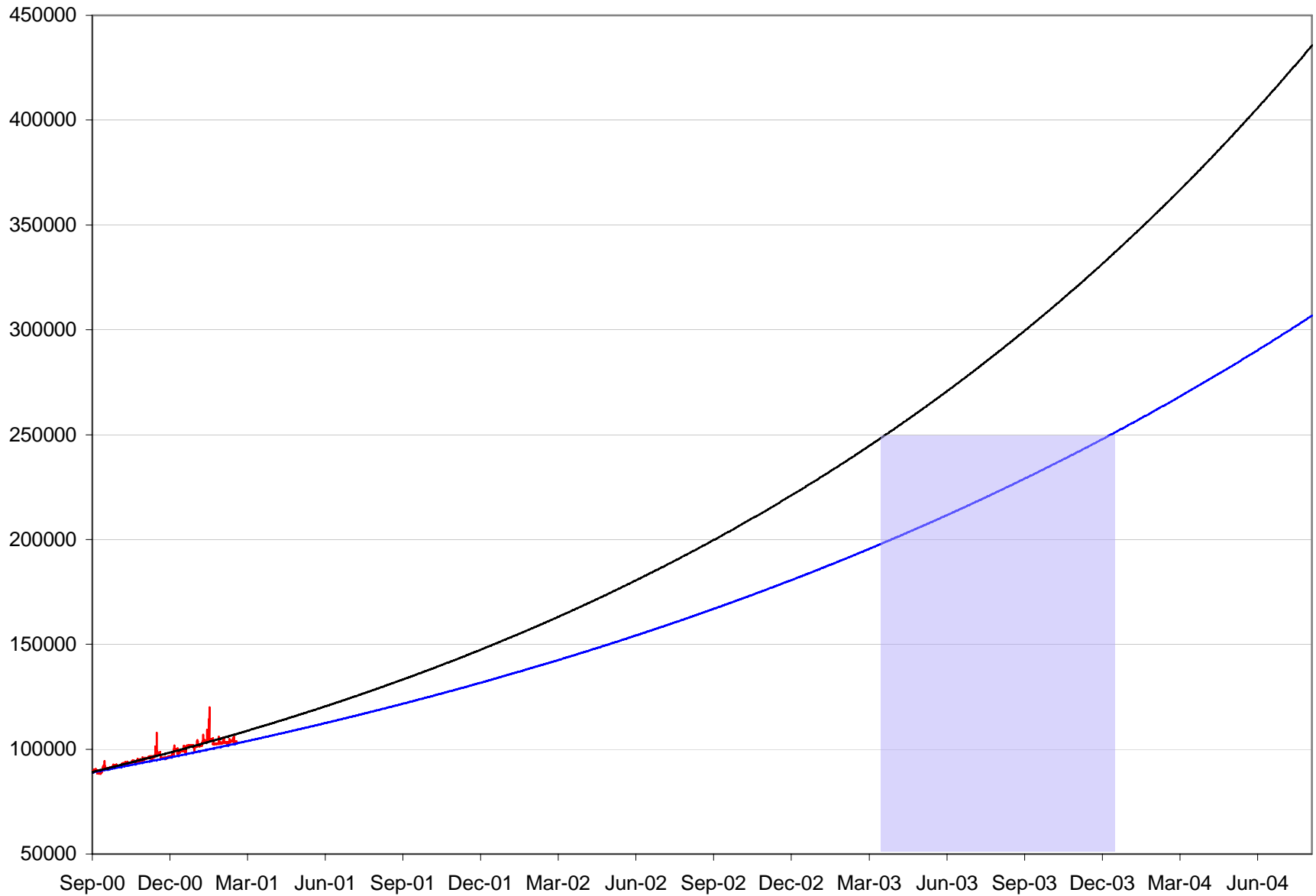# BGP Table Growth – 12 year history

BGP Table Growth
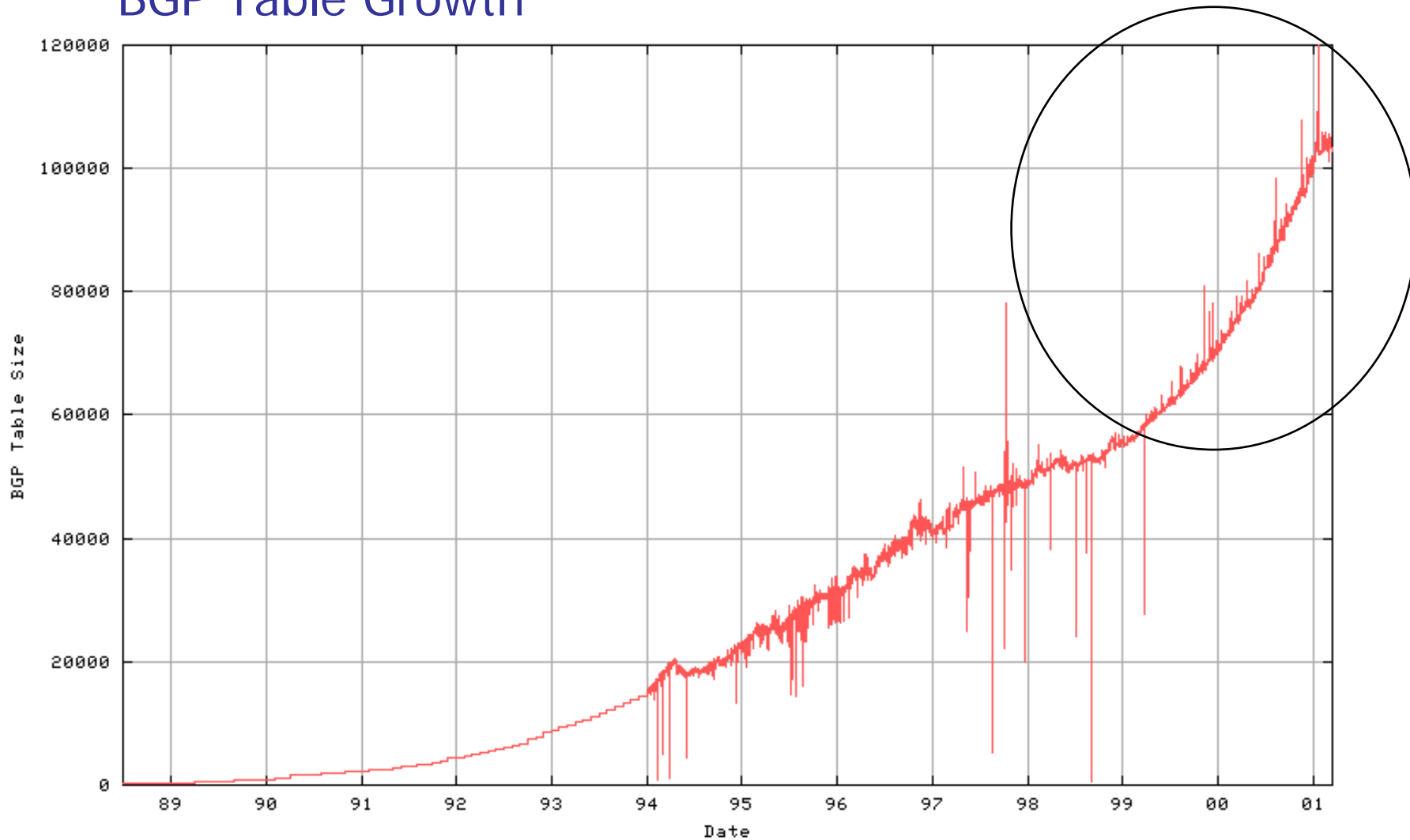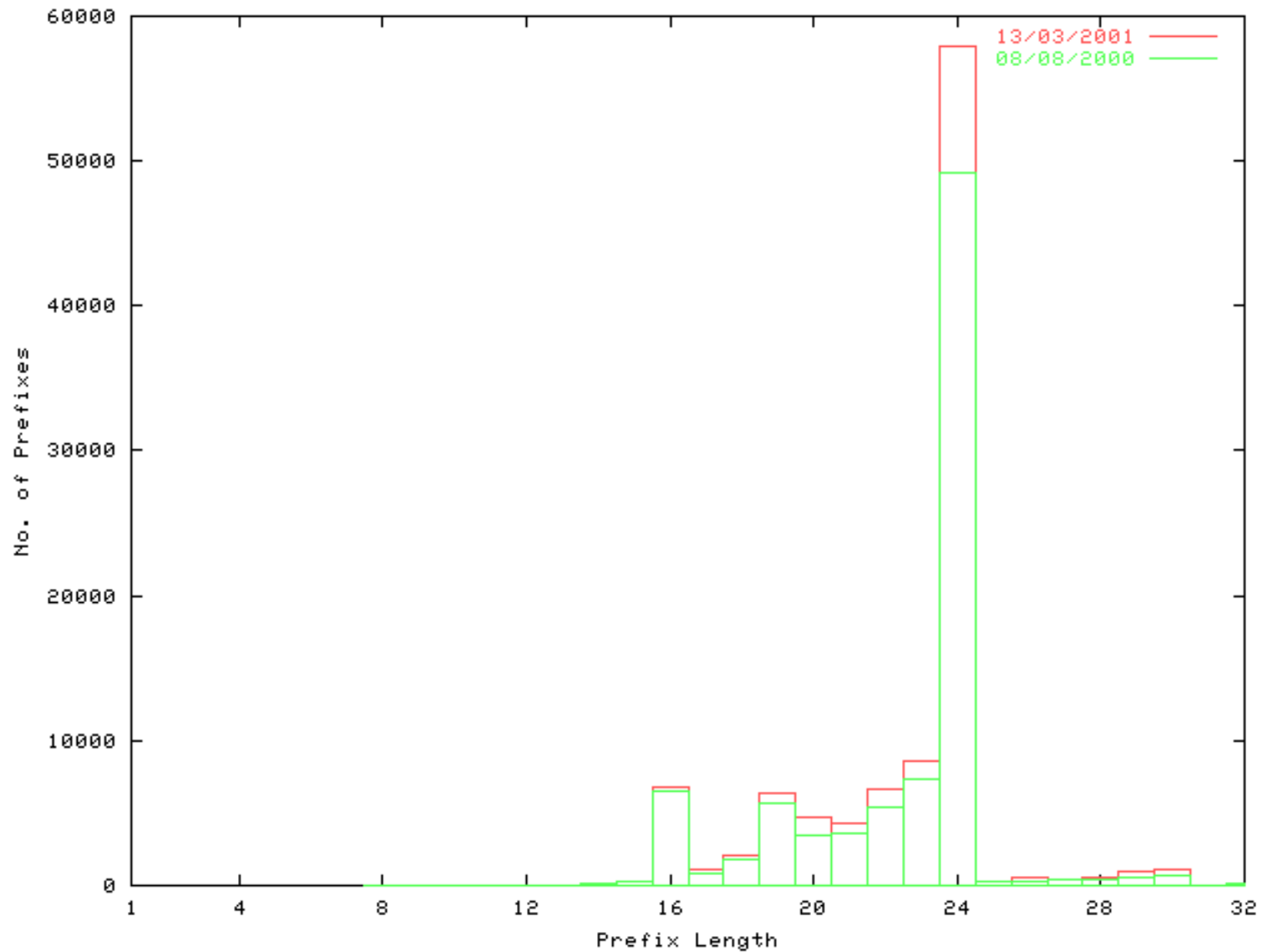
BGP Table Growth – 2 year history
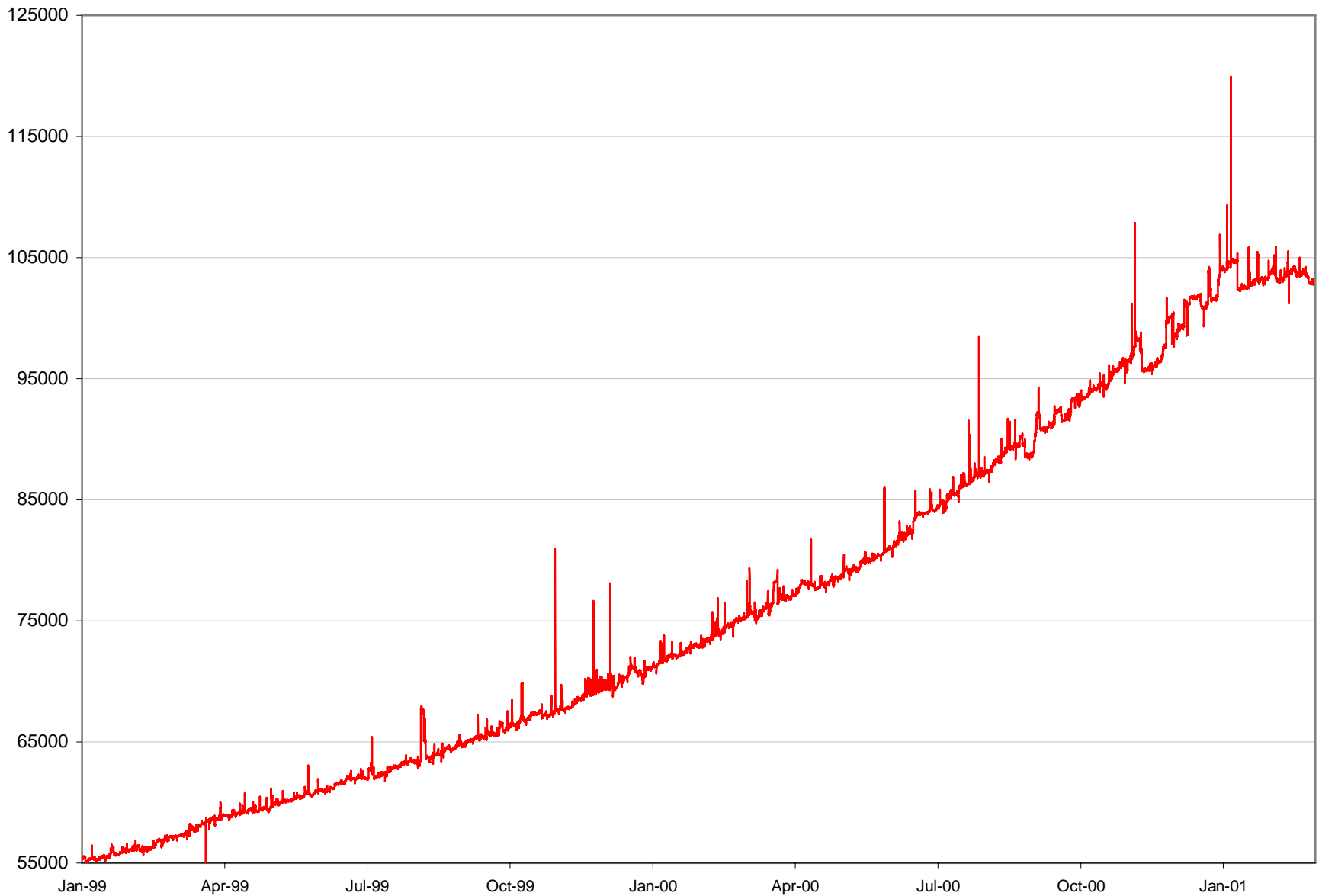
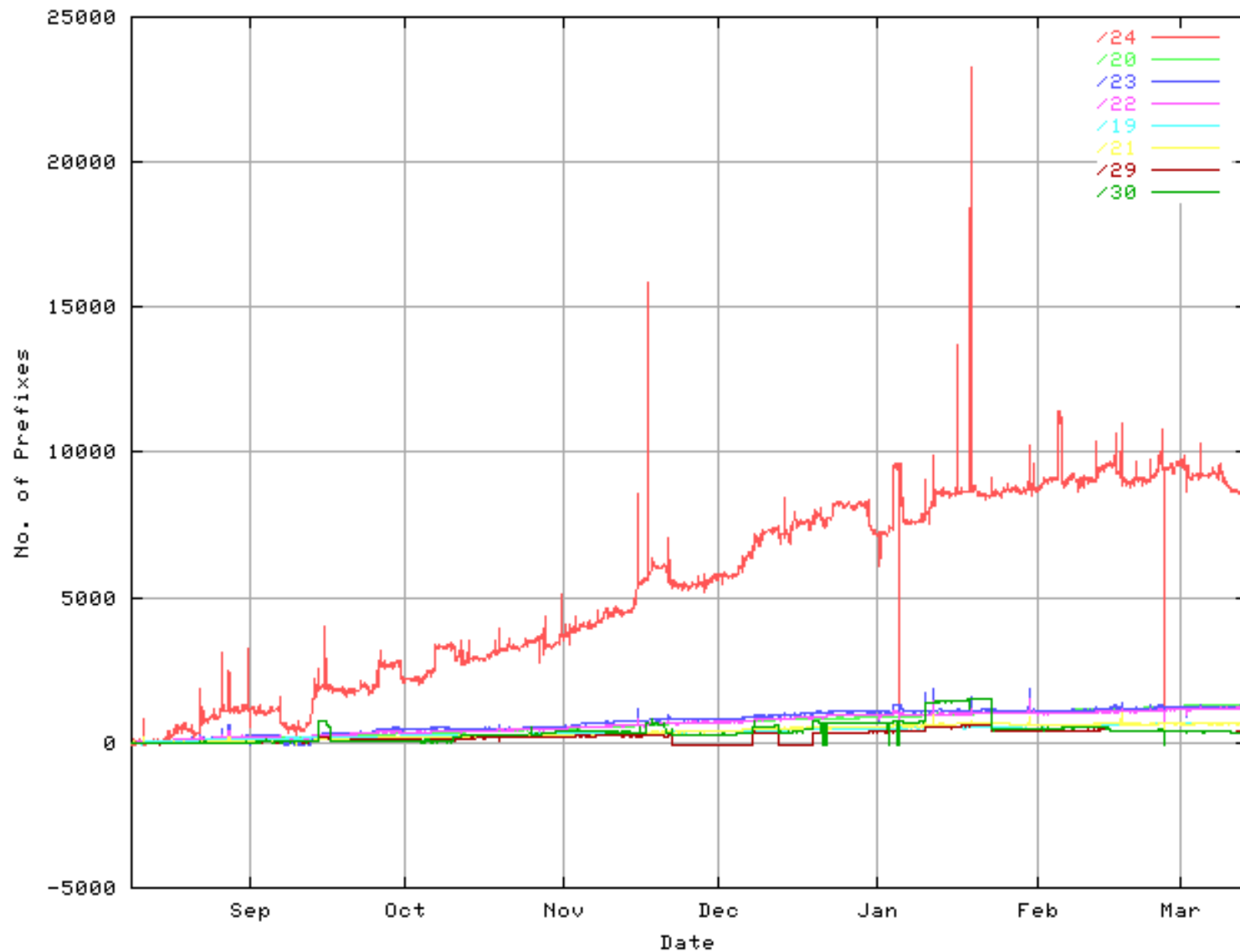# BGP Table Growth – 2 year & 6 month trends

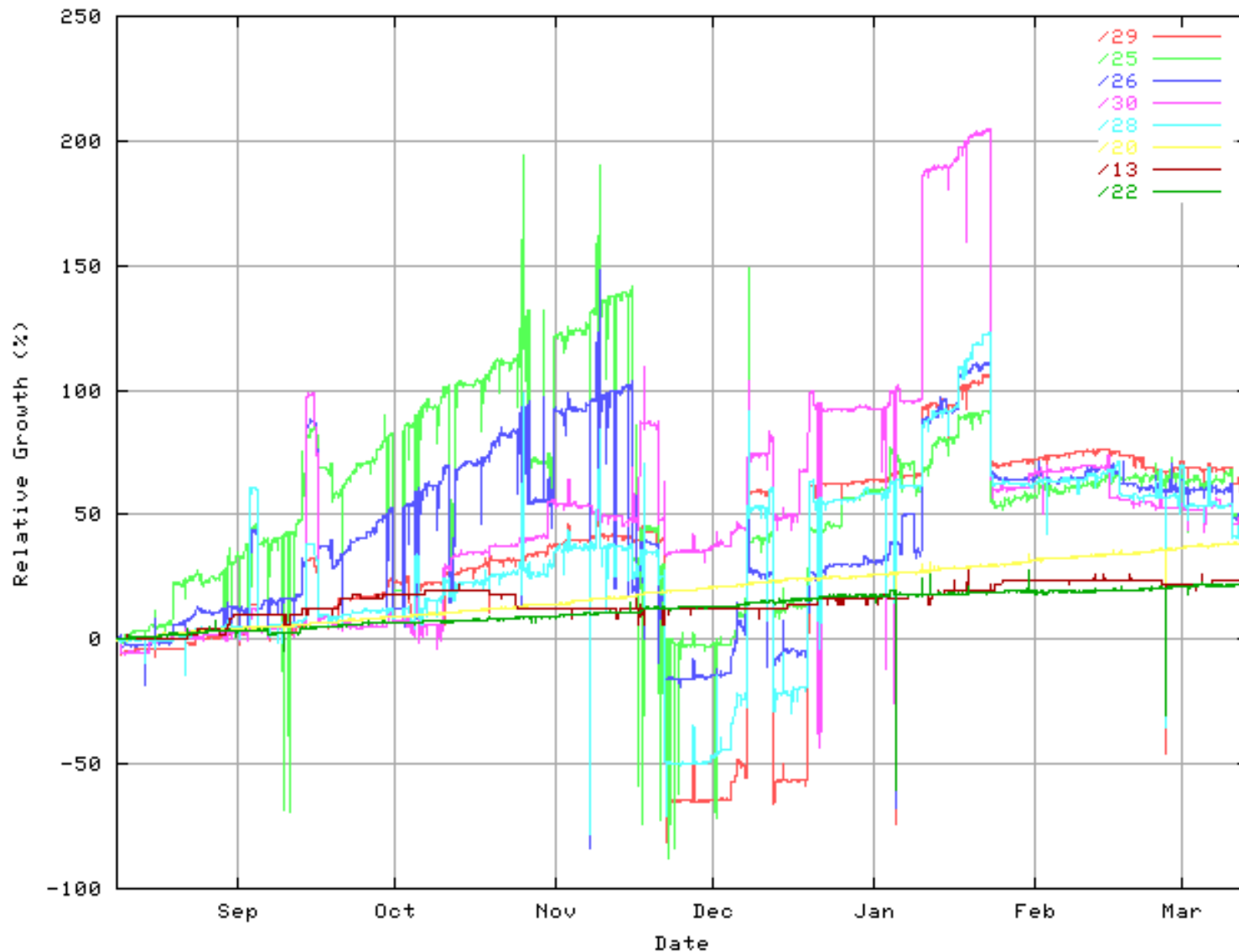# BGP Table Growth – Projections

Prefix size distribution in the BGP table
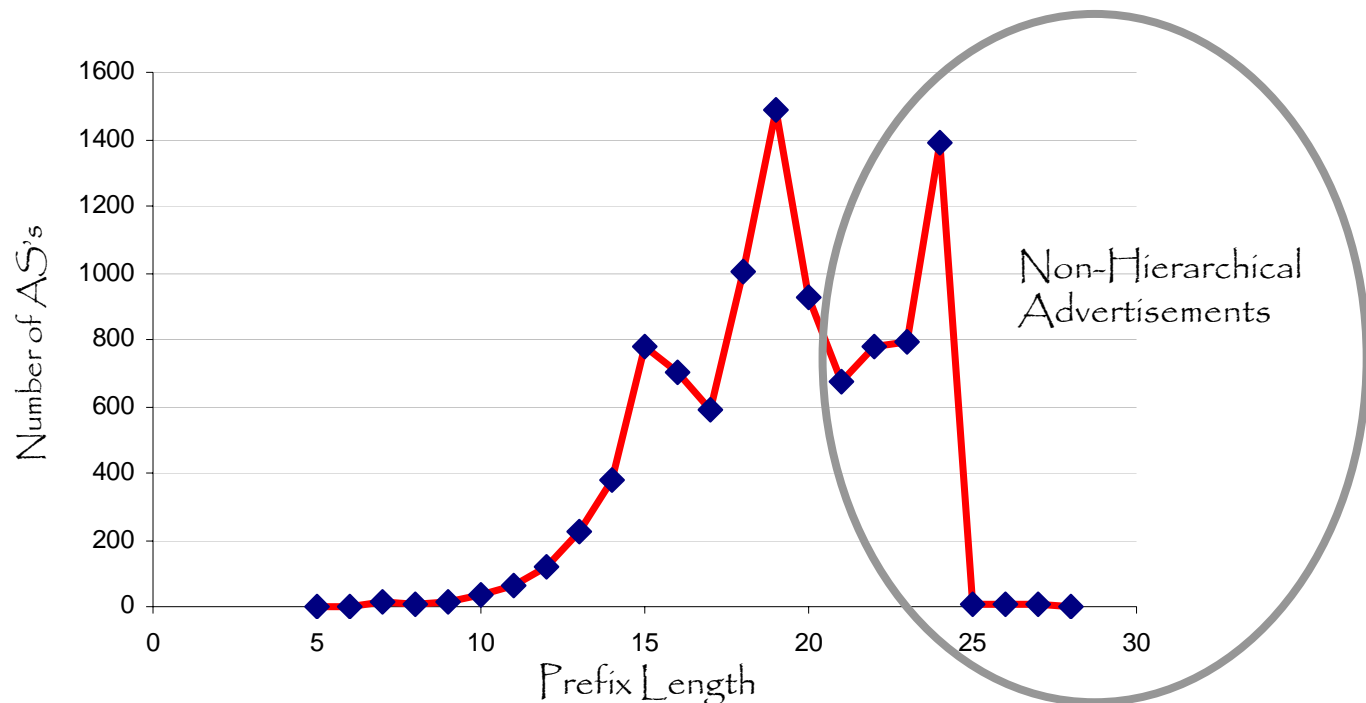
/24 is the fastest growing prefix length

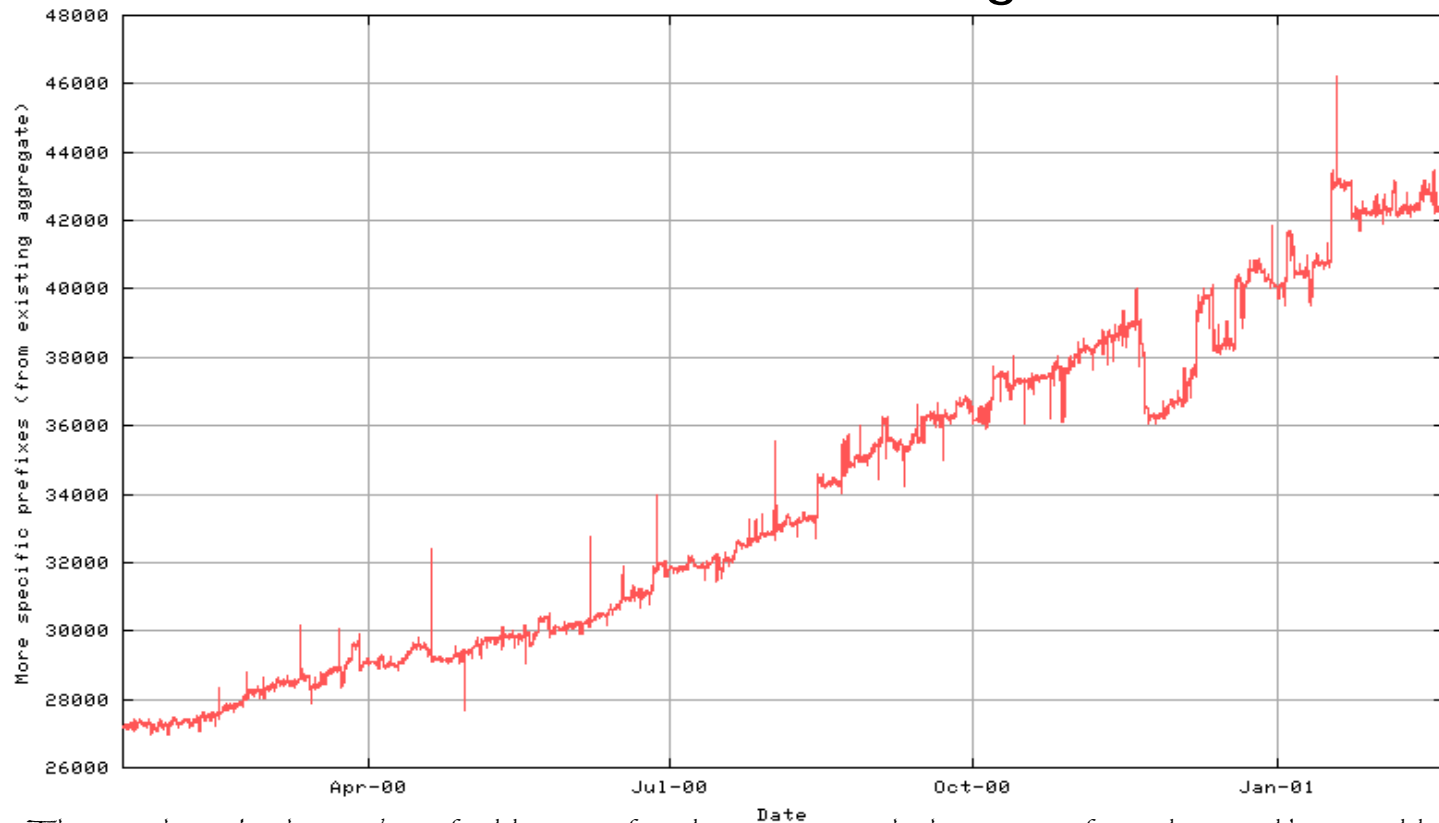# /25 and smaller are the fastest growing prefixes in relative terms

Prefixes by AS

- Distribution of originating address sizes per AS
- Address advertisements are getting smaller

Multi-homing on the rise?

- Track rate of CIDR "holes" – currently 41% of all route advertisements are routing 'holes'

This graph tracks the number of address prefix advertisements which are part of an advertised larger address prefix

Proportion of BGP advertisements which are more specific advertisements of existing aggregates
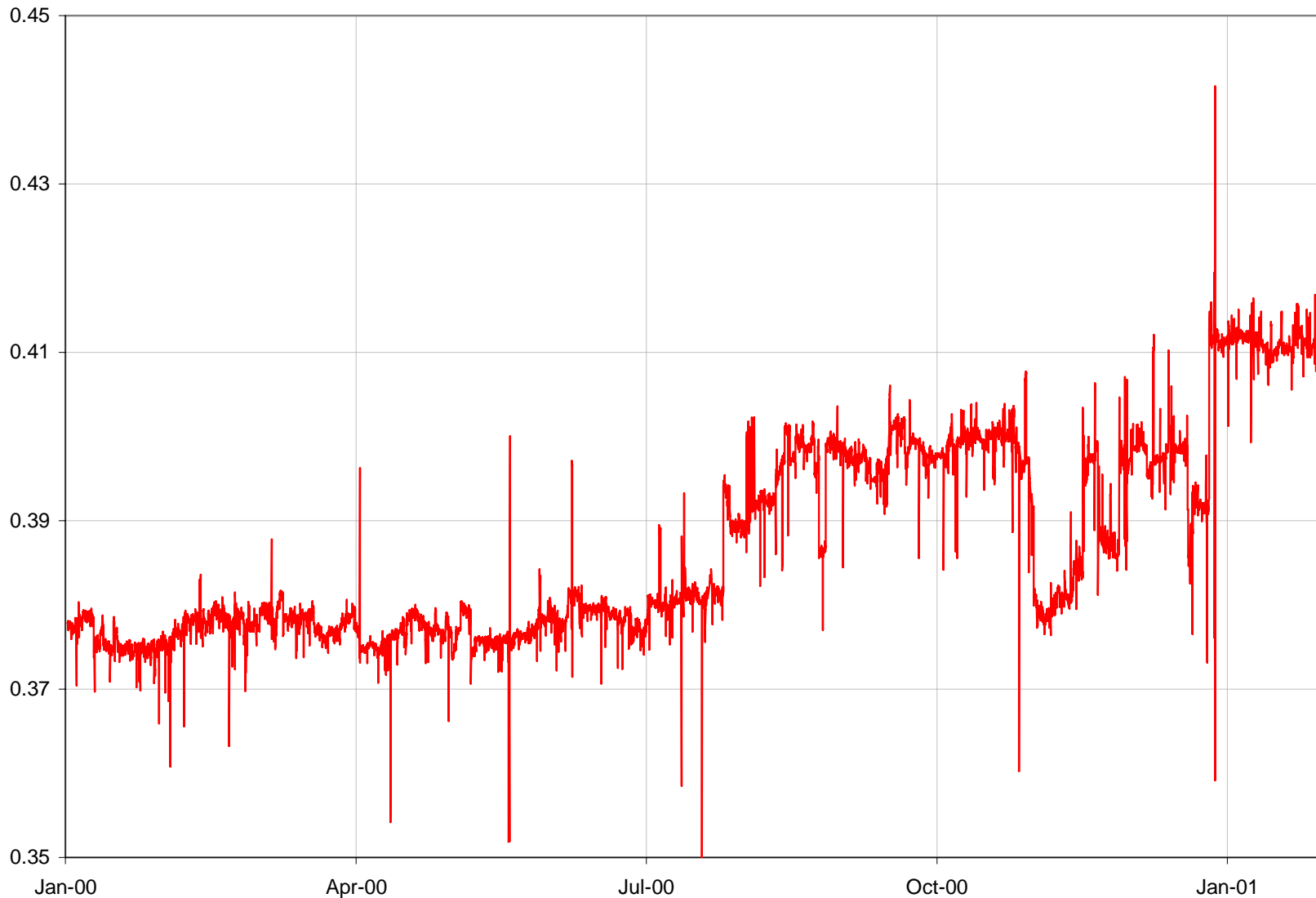
OOPS

- Program bug! The number is larger than that.
- More specific advertisement of existing aggregates account for 54% of the BGP selected route table from the perspective of AS1221
  - 56,799 entries from a total of 103,561
- Older (mid Jan) data from AS286 has the number at 53,644 from a total of 95,036 (56%)

Routed Address Space
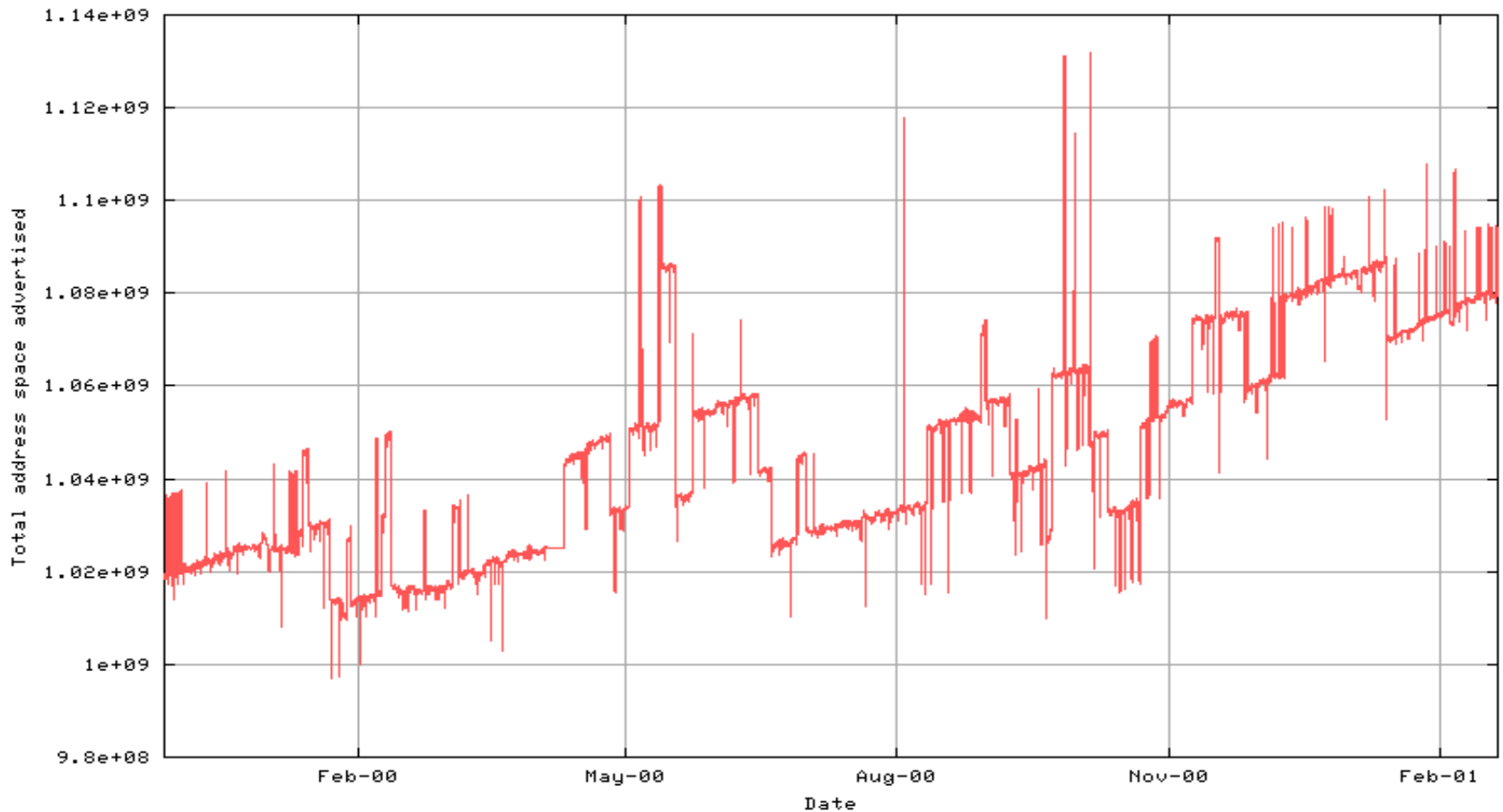
Large fluctuation is due to announcement / withdrawals of /8 prefixes
12 months of data does not provide clear longer growth characteristic

Routed Address Space (/8 Corrected)

Annual compound growth rate is 7% p.a.
Most address consumption today appears to be
ocurring behind NATs
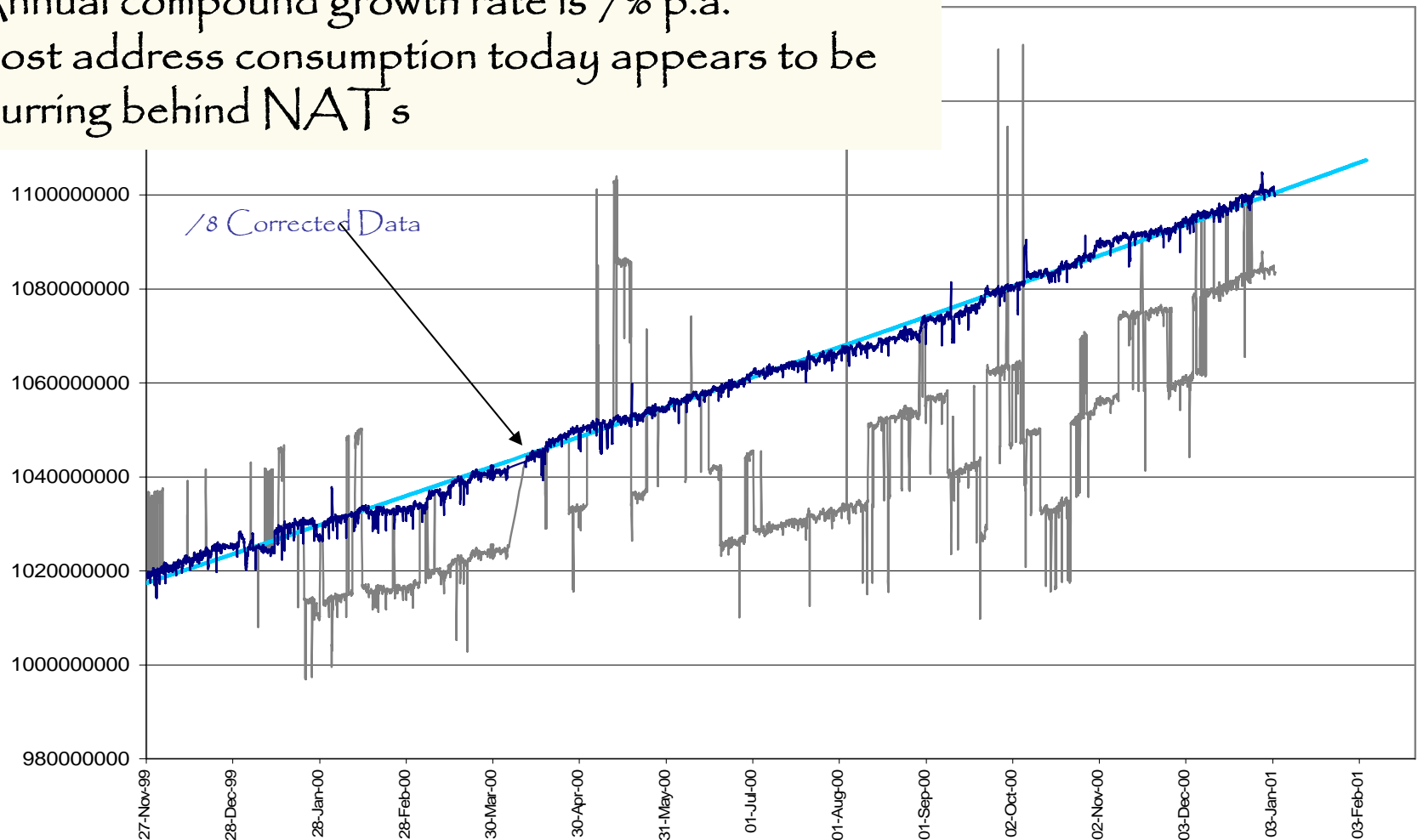
AS Number Growth
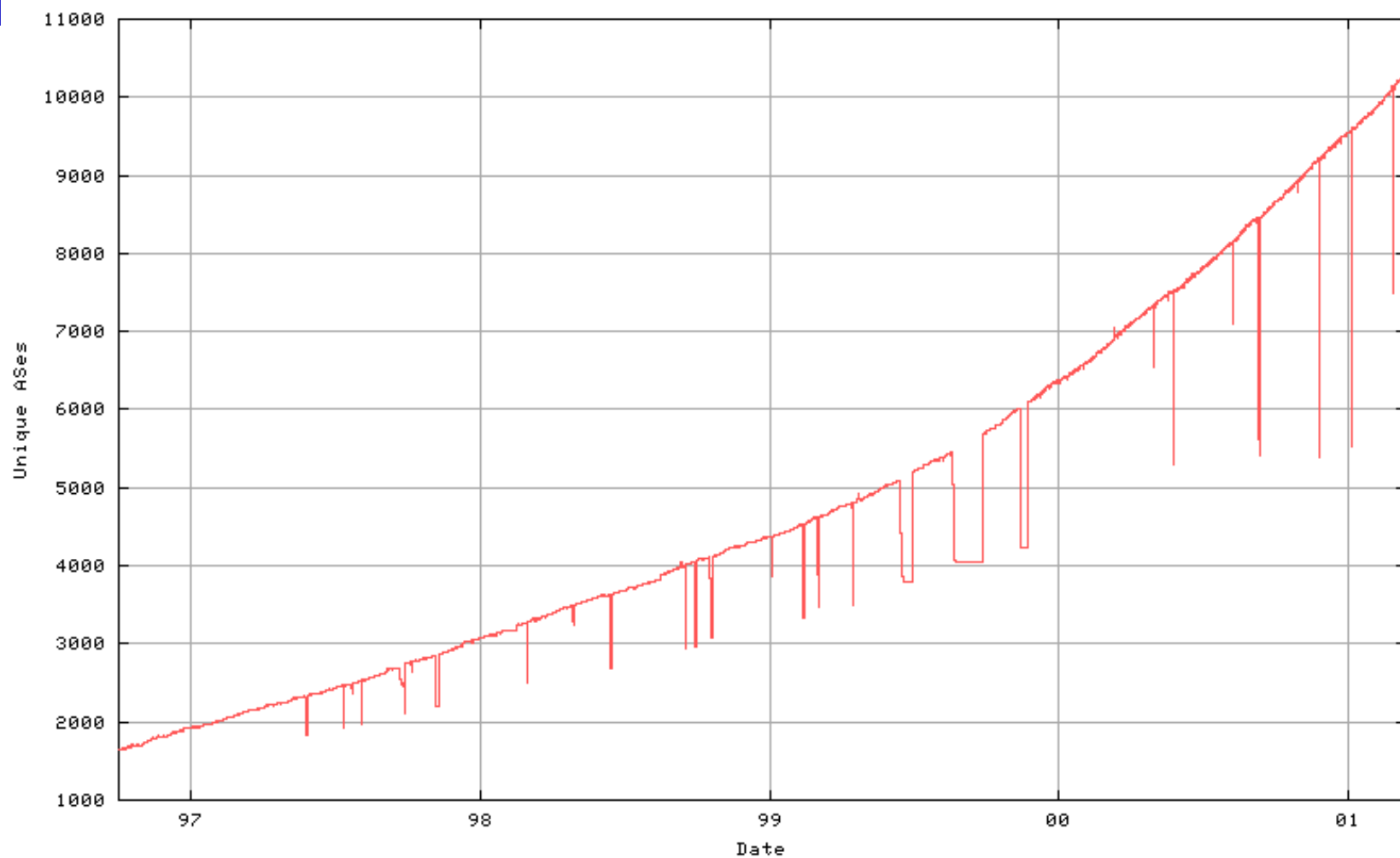
AS Number Use - Extrapolation

Continued exponential growth implies AS number exhaustion in 2005
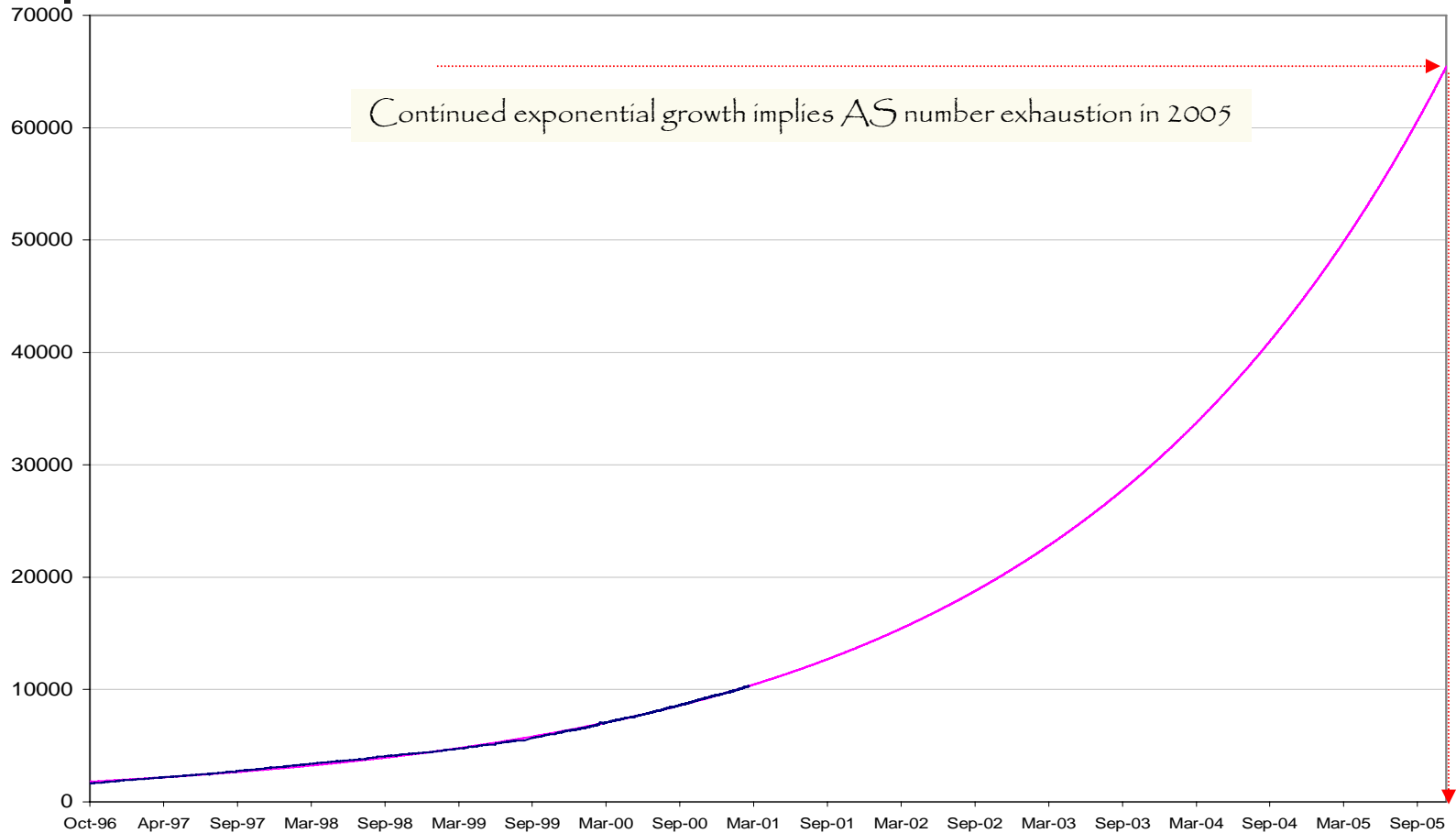
Average size of a routing table entry
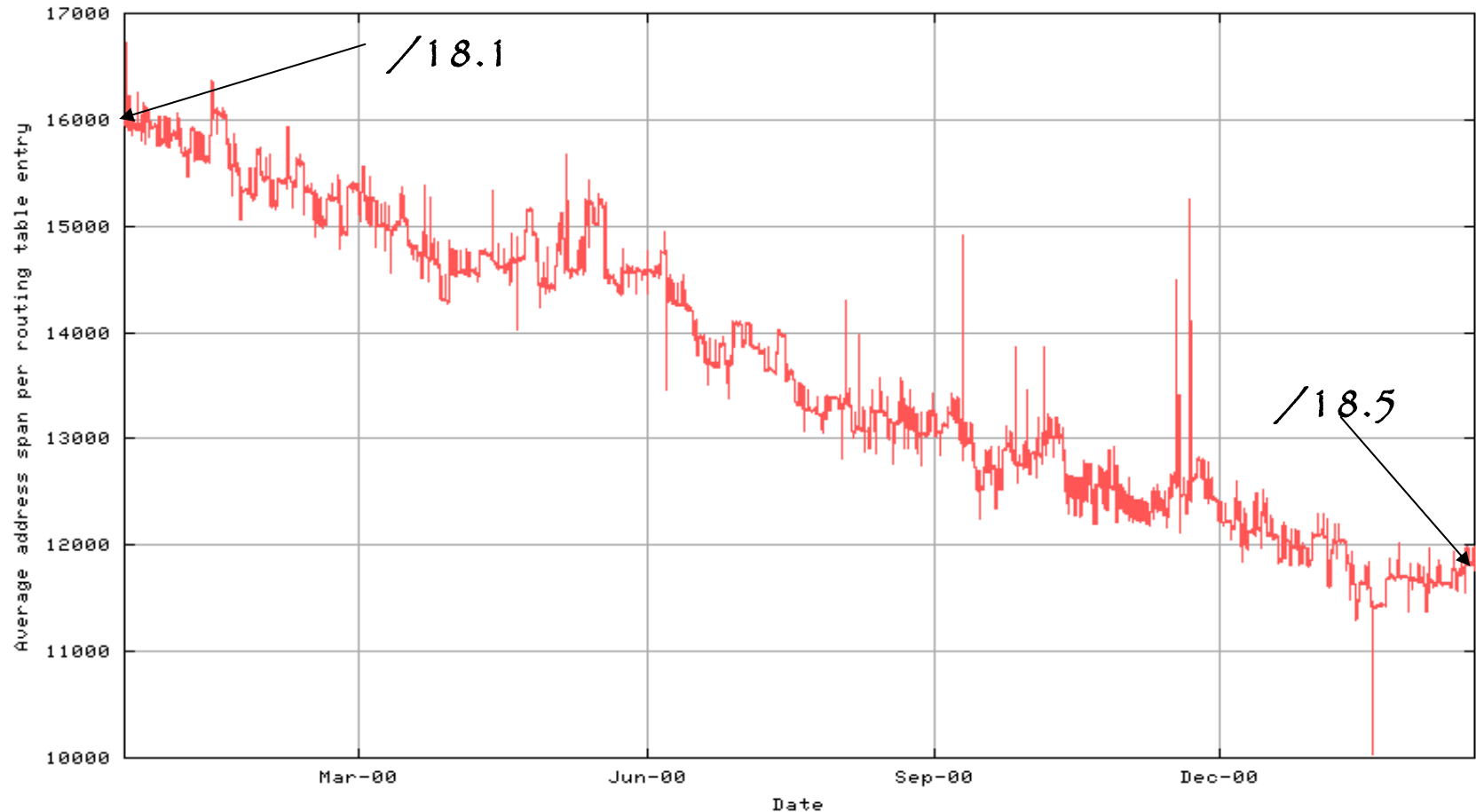
The BGP routing tale is growing at a faster rate than the rate of growth of announced address space

Denser Internet Structure
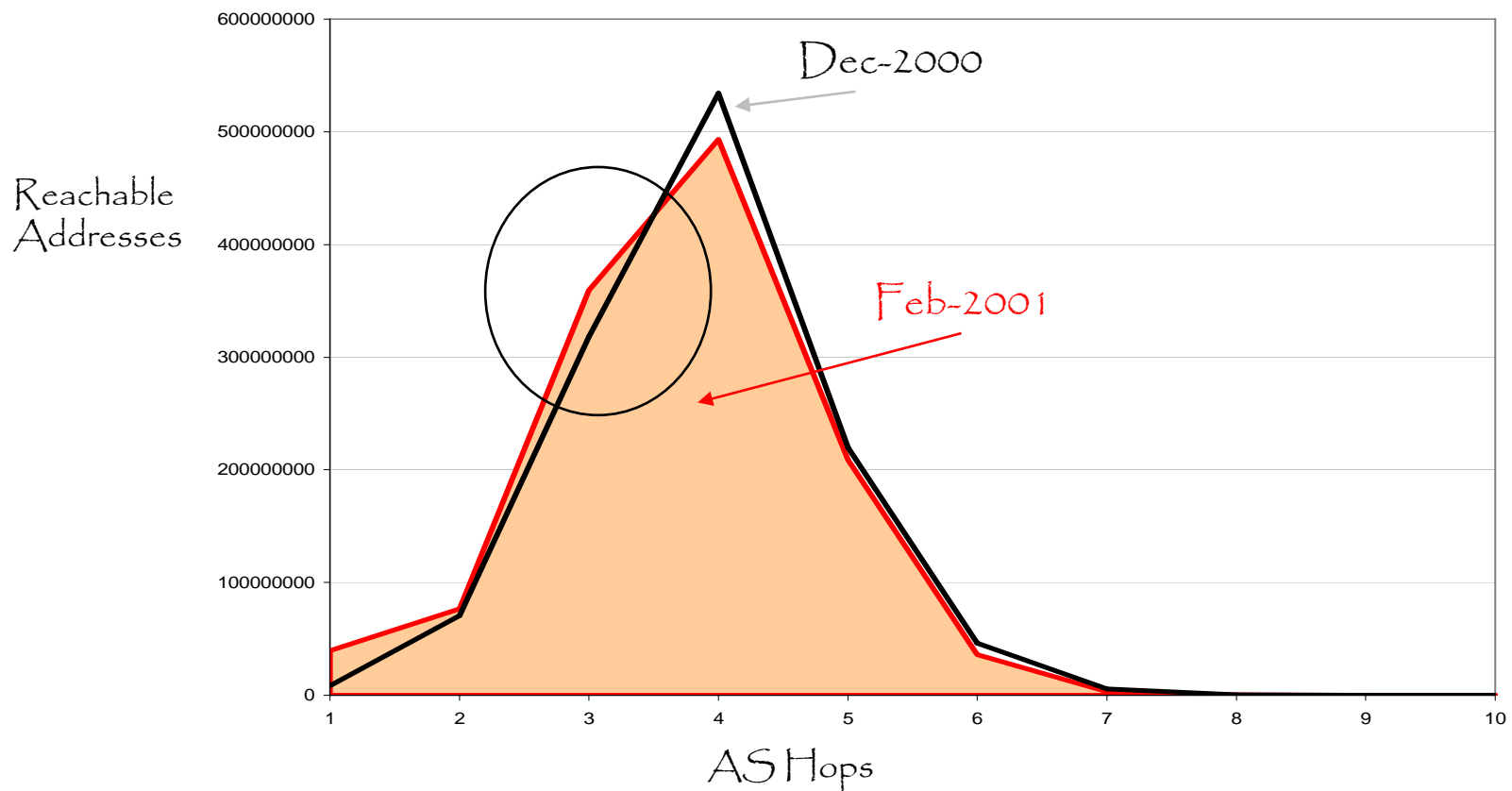
Denser Internet Structure
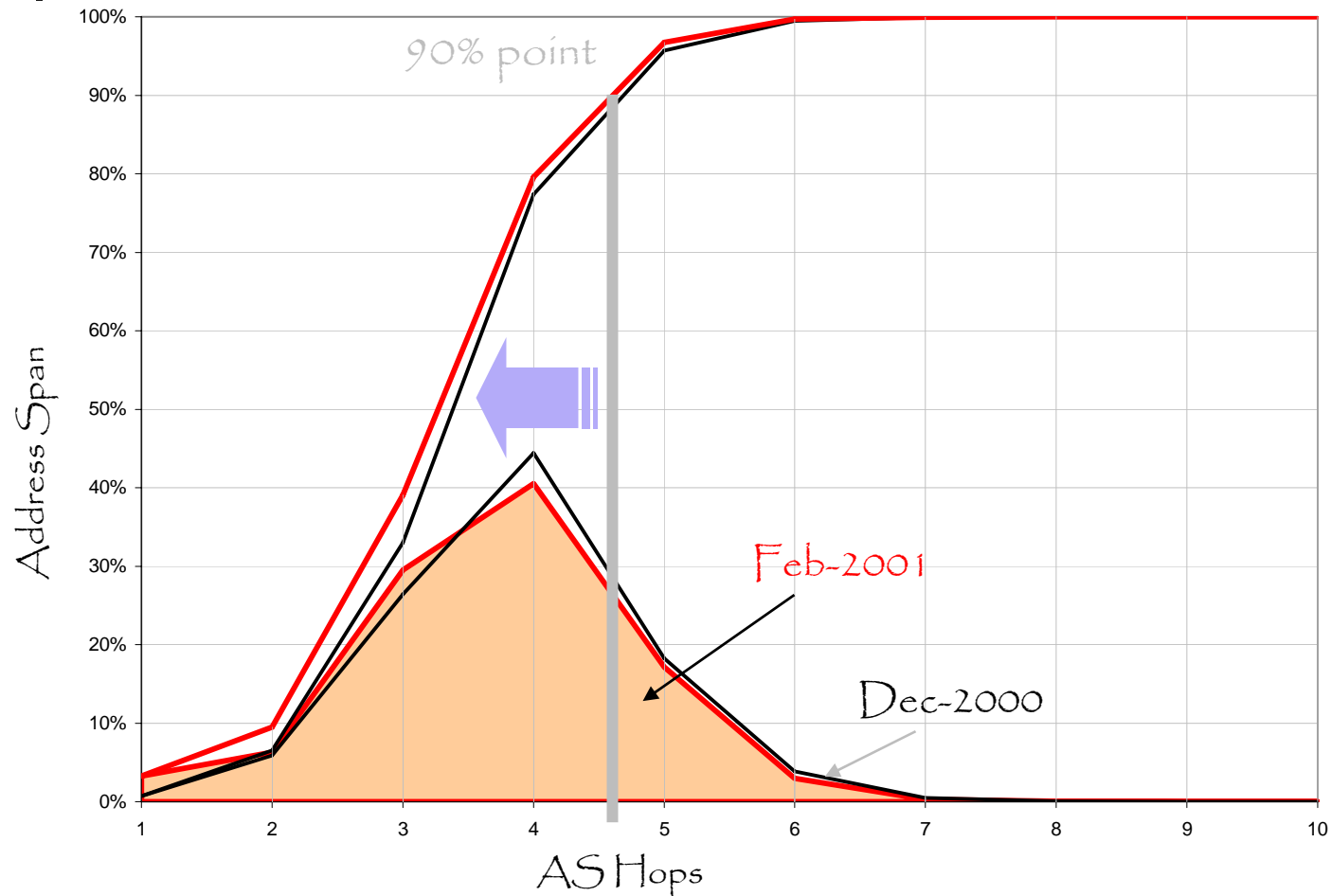
Internet 'Shape'

- The network is becoming less 'stringy' and more densely interconnected
  - i.e. Transit depth is getting smaller
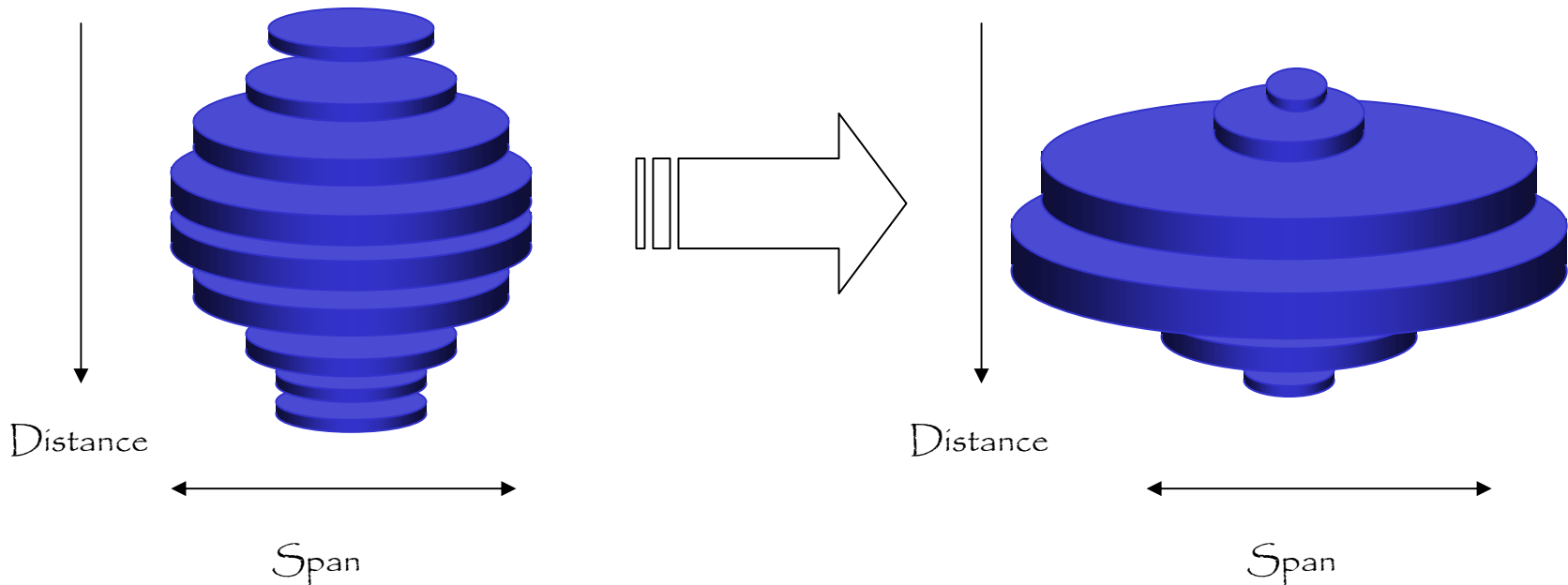
Aggregation and Specifics

- Is the prevalence of fine-grained advertisements the result of deliberate configuration or inadvertent leakage of advertisements?

Different Views
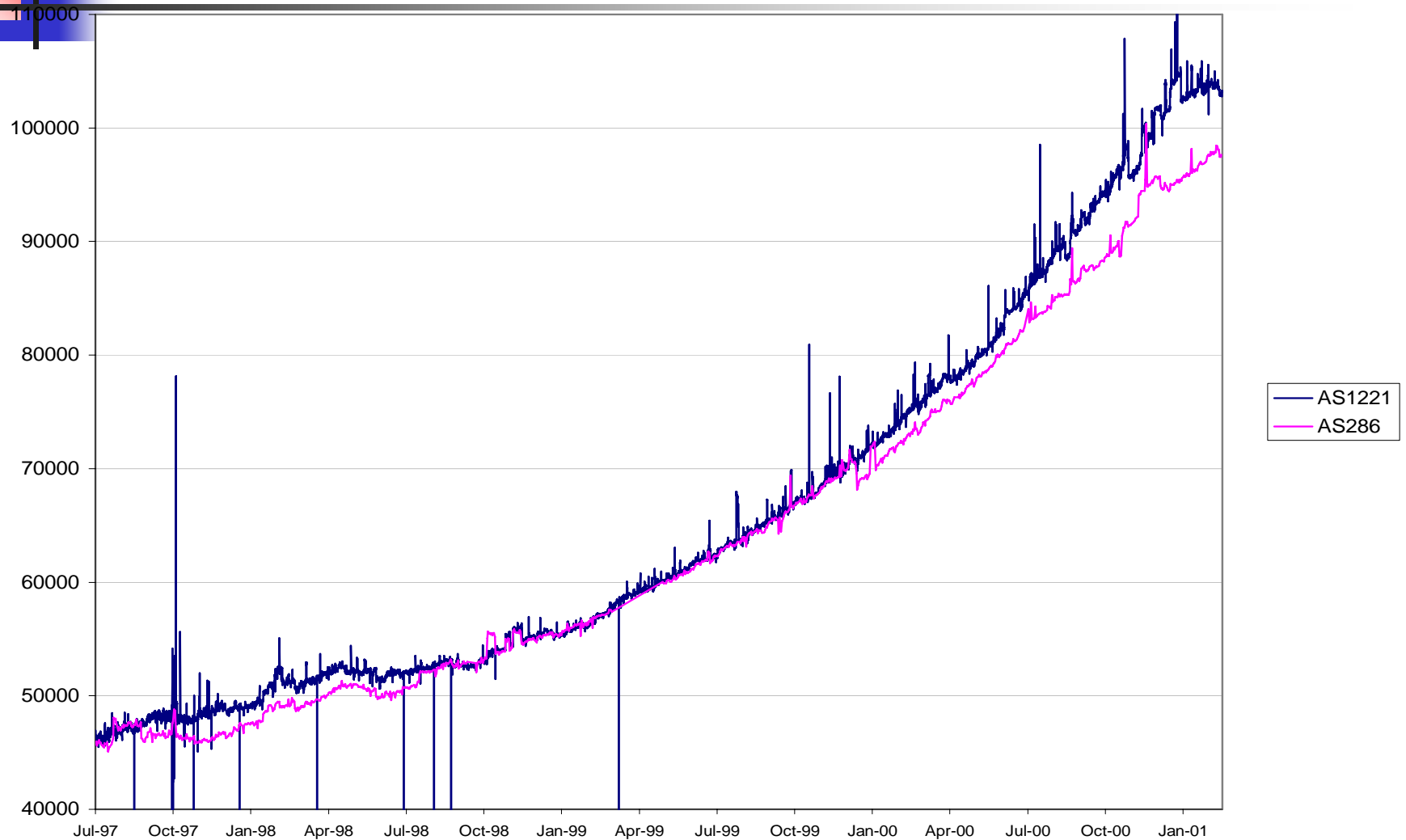
Different Views

- Route views in prefix-length-filtered parts of the net do not show the same recent reduction in the size of the routing table.
- It is likely that the reduction in routes seen by AS1221 appears to be in the prefix-length filtered ranges
    - Either more transit networks are prefix length filtering or origin AS's are filtering at the edge, or both
- The underlying growth trend in BGP table size remains strong

Aggregation possibilities

- ## What if all advertisements were maximally aggregated* ?
  - ### 27% reduction (103126 -> 74427) using AS Path aggregation
  - ### 33% reduction (103126 -> 68504) using AS Origin aggregation

* This assumes that the specific advertisements are not matched by other specific advertisements which have been masked out closer to the origin AS – this is not a terribly good assumption, so these numbers are optimistic to some extent

Aggregation Potential from AS1221
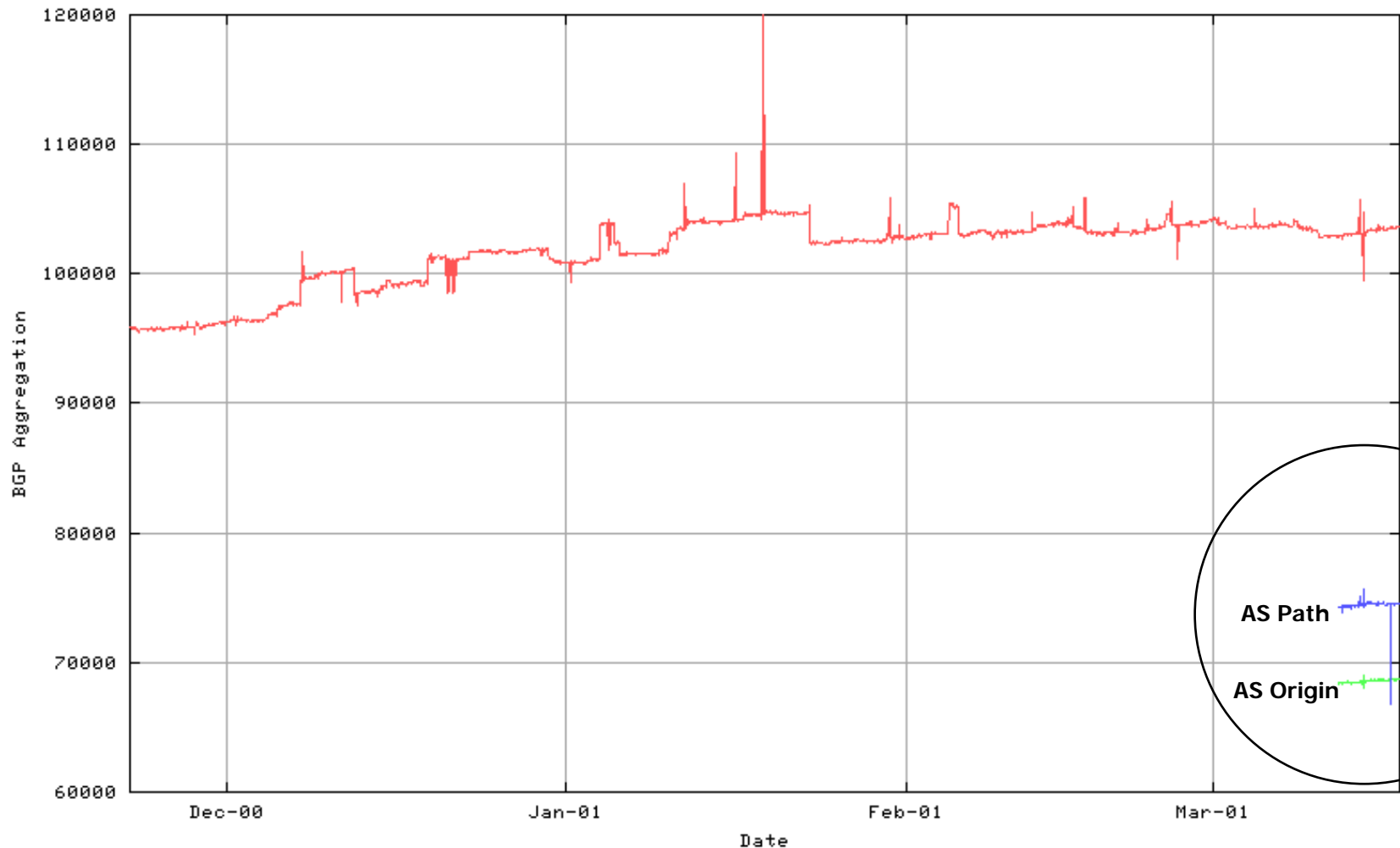
The aggregation potential view from KPNQwest

Data from James Aldridge, KPNQwest - http://www.mcvax.org/~jhma/routing/
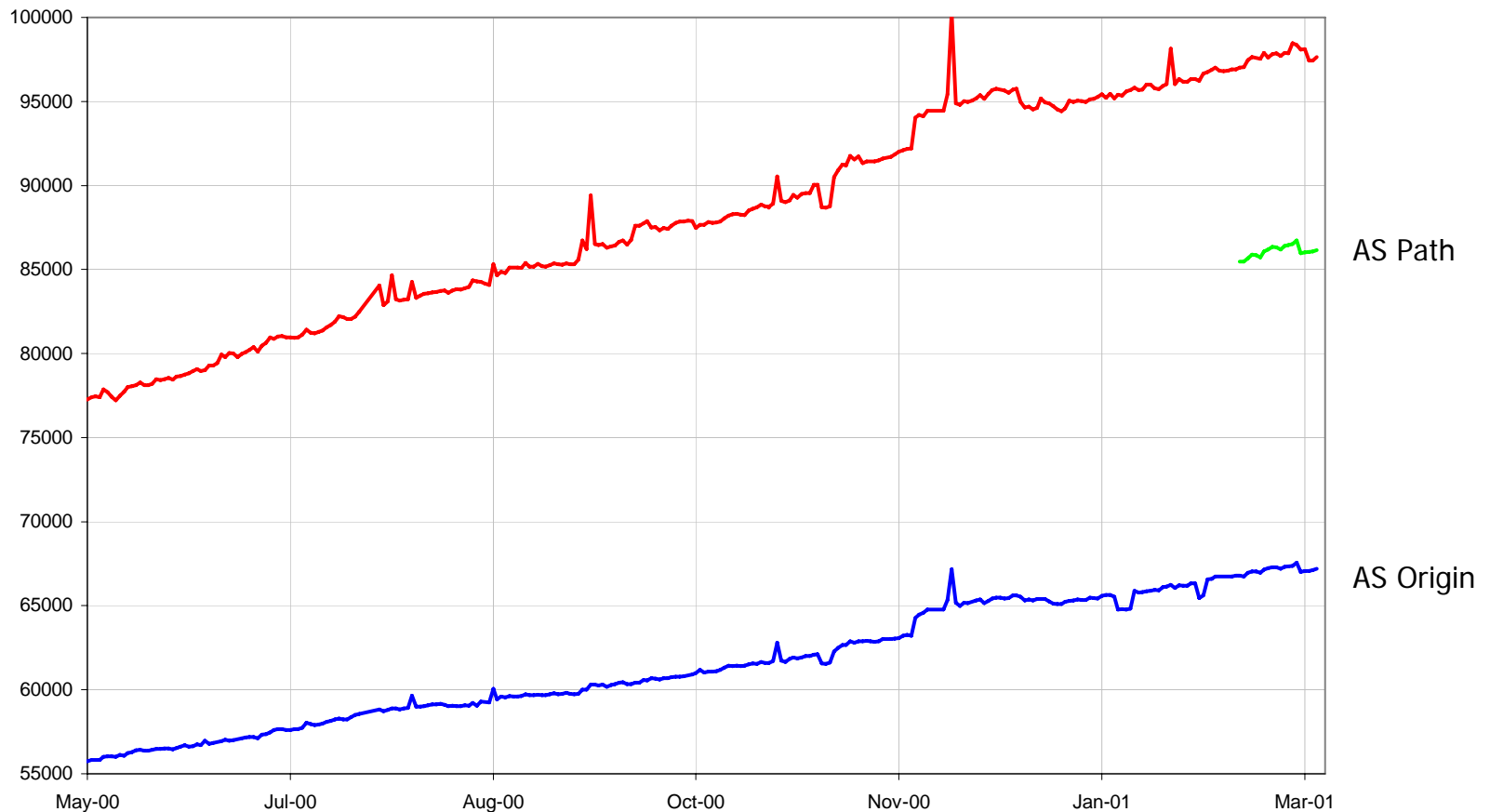
A Longer Term View from AS286

Routing Table Growth (2001-03-17)
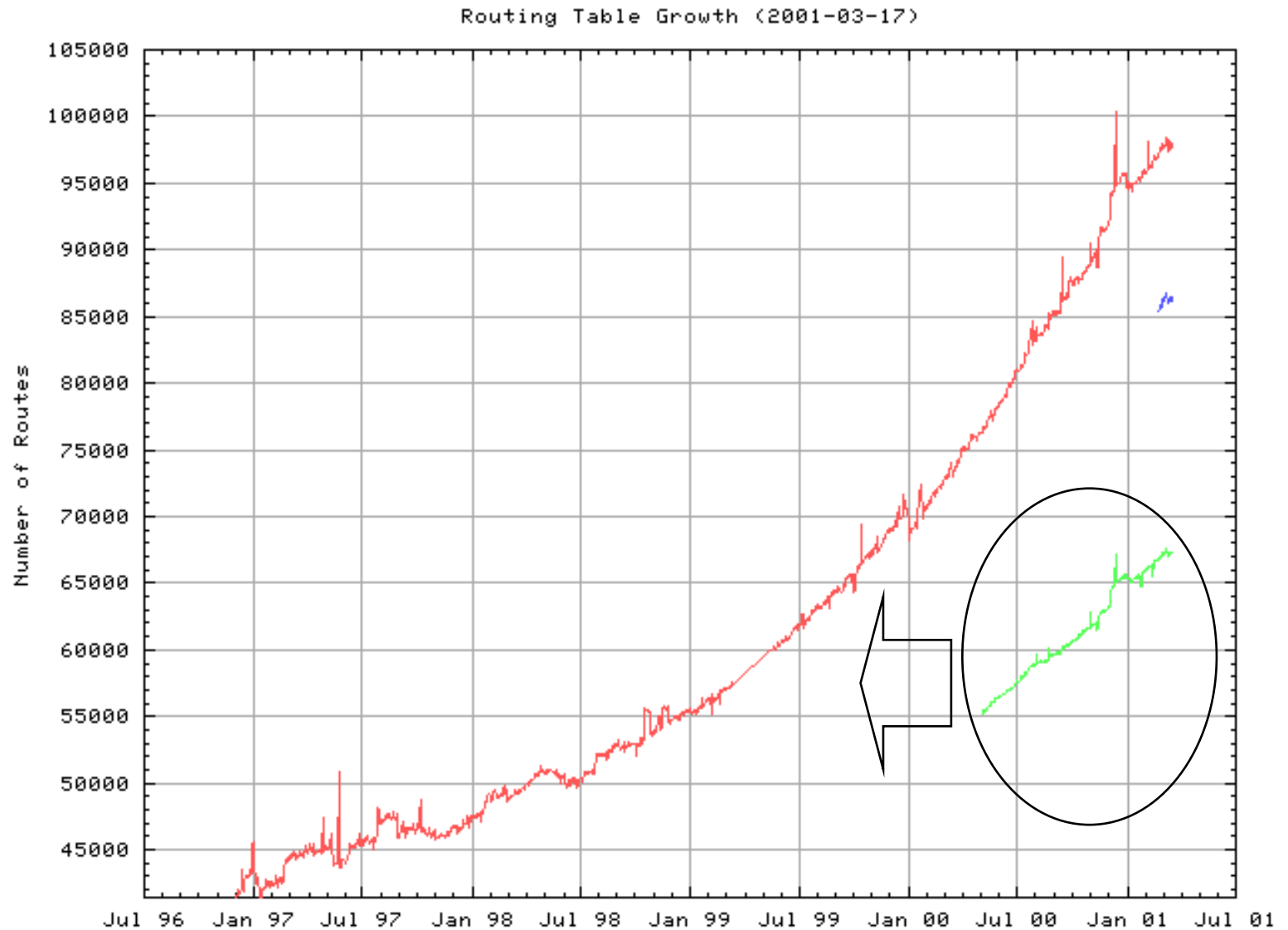
Aggregatability?

- A remote view of aggregation has two potential interpretations:
    - Propose aggregation to the origin AS
    - Propose a self-imposed proxy aggregation ruleset
- Any aggregation reduces the information content in the routing table. Any such reduction implies a potential change in inter-domain traffic patterns.
- Aggregation with preserved integrity of traffic flows is different from aggregation with potential changes in traffic flow patters

Aggregatability

- Origin AS aggregation is easier to perform at the origin, but harder to determine remotely IF traffic flows are to be preserved

- Proxy Aggregation is only possible IF you know what your neighbors know
  Yes this is a recursive statement

    - If an AS proxy aggregates will it learn new specifics in response?

BGP as a Routing Protocol

- How quickly can the routing system converge to a consistent state following dynamic change?
- Is this time interval changing over time?

Increased convergence time intervals for BGP

- Measured time to withdraw route:
    - Up to 2 minutes
- Measured time to advertise new route:
    - Up to 30 minutes

Withdraw Convergence

- Probability distribution
- Providers exhibit different, but related convergence behaviors
- 80% of withdraws from all ISPs take more than a minute
- For ISP4, 20% withdraws took more than three minutes to converge
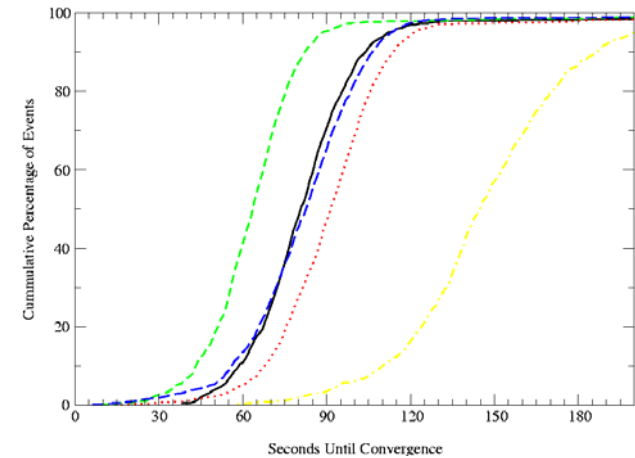
Failures, Fail-overs and Repairs

- Bad news does not travel fast...
- Failures and short-long fail-overs (e.g. primary to secondary path) also similar
  - 60% take longer than two minutes
- Fail-over times degrade the greater the degree of multi-homing!
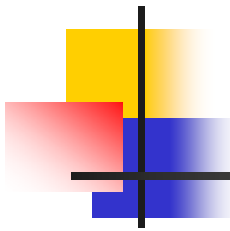
Conjectures....

BGP table size will continue to rise exponentially

- Multi-homing at the edge of the Internet is on the increase
- The interconnectivity mesh is getting denser
  - The number of AS paths is increasing faster than the number of AS's
  - Average AS path length remains constant
- AS number deployment growth will exhaust 64K AS number space in August 2005 if current growth trends continue

More conjecturing....

- Inter-AS Traffic Engineering is being undertaken through routing discrete prefixes along different paths -- globally (the routing mallet!)
  - AS Origin aggregation < AS Path aggregation
- RIR allocation policy (/19, /20) is driving one area of per-prefix length growth in the aggregated prefix area of the table
- BUT - NAT is a **very** common deployment tool
  - NAT, multihoming and Traffic Engineering is driving even larger growth in the /24 prefix area

And while we are having such a good time conjecturing…

- Over 12 months average prefix length in the table has shifted from /18.1 to /18.5
- More noise (/25 and greater) in the table, but the absolute level of noise is low (so far)
- Most routing table flux is in the /24 to /32 prefix space – as this space gets relatively larger so will total routing table flux levels
  - "Flux" here is used to describe the cumulative result of the withdrawals and announcements
  - This space appears to be susceptible to social pressure – at present

This is fun – lets have even more conjectures...

- CIDR worked effectively for four years, but its effective leverage to support long term dampened route table growth and improved table stability has now finished

- Provider-based service aggregation hierarchies as a model of Internet deployment structure is more theoretic than real these days

i.e. provider based route aggregation is leaking like a sieve!

Commentary

- draft-iab-bgparch-00.txt
  - Exponential growth of BGP tables has resumed
  - AS number space exhaustion
  - Convergence issues
  - Traffic Engineering in a denser mesh

  - What are the inter-domain routing protocol evolutionary requirements?

Objectives and Requirements

- Supporting a larger and denser interconnection topology
- Scale by x100 over current levels in number of discrete policy entities
- Fast Convergence
- Security
- Integration of Policy and Traffic Engineering as an overlay on basic connectivity
- Control entropy / noise inputs

Available Options

- Social Pressure on aggregation
- Economic Pressure on route advertisements
- Tweak BGP4 behavior
- Revise BGP4 community attributes
- BGPng
- New IDR protocol(s)
- New IP routing architecture

Social Pressure

- Social pressure can reduce BGP noise
- Social pressure cannot reduce pressures caused by
  - Denser interconnection meshing
  - Increased use of multi-homing
  - Traffic engineering of multiple connections
- Limited utility and does not address longer term routing scaling

Economic Pressure on Routing

- Charge for route advertisements
  - Upstream charges a downstream per route advertisements
  - Peers charge each other
- This topic is outside an agenda based on technology scope
- Raises a whole set of thorny secondary issues:
  - Commercial
  - National Regulatory
  - International
- Such measures would attempt to make multi-homing less attractive economically. It does not address why multi-homing is attractive from a perspective of enhanced service resilience.

Tweaking BGP4

- Potential tweak to BGP-4
  - Auto-Proxy-Aggregation
    - Automatically proxy aggregate bitwise aligned route advertisements
    - Cleans up noise – but reduces information
    - Cannot merge multi-homed environments unless the proxy aggregation process makes sweeping assumptions, or unless there is an overlay aggregation protocol to control proxy aggregation (this is then no longer a tweak)

Extend BGP4 Communities

- We already need to extend community attributes to take on the 2 / 4 octet AS number transition.
- Can we add further community attribute semantics to allow proxy aggregation and proxy sublimation under specified conditions?
- Extend commonly defined transitive community attributes to allow further information to be attached to a routing advertisement
  - Limit of 'locality' of propagation
  - Aggregation conditions or constraints
- If we could do this, will this be enough? Can this improve
  - Scaling properties
  - convergence properties

BGPng

- Preserve: AS concept, prefix + AS advertisements, distance vector operation, AS policy "opaqueness"

- Alter: convergence algorithm (DUAL?), advertisement syntax (AS + prefix set + specifics + constraints), BGP processing algorithm

- Issues:
  - Development time
  - Potential to reach closure on specification
  - Testing of critical properties
  - Deployment considerations
  - Transition mechanisms

IDR

- ## A different IDR protocol?
  - ### Can we separate connectivity maintenance, application of policy constraints and sender- and/or receiver- managed traffic engineering?
    - SPF topology maintenance
    - Inter-Domain Policy Protocol to communicate policy preferences between policy "islands"
    - Multi-domain path maintenance to support traffic engineering requirements
  - ### Eliminate the need to advertise specifics to undertake traffic engineering
  - ### Multi-homing may still be an issue – is multi-homing a policy issue within an aggregate or a new distinct routing "entity"?
  - ### Can SPF scale? Will SPF routing hierarchies impose policy on the hierarchy elements?

New IP Routing Architecture

- Separate Identity, Location and Path at an architectural level?
- Identity
  - How do you structure an entirely new unique identity label space? How do you construct the "identity lookup" mechanism?
- Location
  - How can location be specified independent of network topology?
- Path:
  - Is multi-homing an internal attribute within the network driven by inter-domain policies, or is multi-homing an end-host switching function

New IP Routing Architecture

- Other approaches?
  - Realms and RSIP
  - Inter-Domain CRLDP approaches where policy is the constraint