Scaling issues with ipv6 routing & multihoming

Vince Fuller, Cisco Systems

RIPE-53, Amsterdam, NL

Session Objectives

- A brief look at how we got where we are today
- Define "locator", "endpoint-id", and their functions
- Explain why these concepts matter and why this separation is a good thing
- Understand that IPv4 and ipv6 co-mingle these functions and why that is problematic
- Examine current ipv6 multi-homing direction and project how that will scale into the future
- Determine if this community is interested in looking at a solution to the scaling problem

Acknowledgements

This is not original work, so credit must be given to:

- **Noel Chiappa for his extensive writings over the years on ID/Locator split**
- **Mike O’Dell for developing GSE/8+8**
- **Geoff Huston for his ongoing global routing system analysis work (CIDR report, BGP report, etc.)**
- **Jason Schiller for the growth projection section (and for tag-teaming to present this at NANOG)**
- **Marshall Eubanks for sanity-checking the growth projections against economic reality**

A brief history of Internet time

- Recognition of exponential growth – late 1980s
- CLNS as IP replacement – December, 1990 IETF
- ROAD group and the “three trucks” – 1991-1992
  - Running out of “class-B” network numbers
  - Explosive growth of the “default-free” routing table
  - Eventual exhaustion of 32-bit address space
  - Two efforts – short-term vs. long-term
  - More at “The Long and Winding ROAD” http://rms46.vlsm.org/1/42.html
- Supernetting and CIDR – 1992-1993

A brief history of Internet time (cont'd)

- IETF "ipng" solicitation – RFC1550, Dec 1993
- Direction and technical criteria for ipng choice – RFC1719 and RFC1726, Dec 1994
- Proliferation of proposals:
  - TUBA – RFC1347, June 1992
  - PIP – RFC1621, RFC1622, May 1994
  - CATNIP – RFC1707, October 1994
  - SIP – RFC1710, October 1994
  - NIMROD – RFC1753, December 1994
  - ENCAPS – RFC1955, June 1996

A brief history of Internet time (cont'd)

- Choice came down to politics, not technical merit
  - Hard issues deferred in favor of packet header design
- Things lost in shuffle…err compromise included:
  - Variable-length addresses
  - De-coupling of transport and network-layer addresses
  - Clear separation of endpoint-id/locator (more later)
  - Routing aggregation/abstraction
- In fairness, these were (and still are) hard problems… but without solving them, long-term scalability is problematic

Identity - “what’s in a name”?

- **Think of an “endpoint-id” as the “name” of a device or protocol stack instance that is communicating over a network**
- **In the real world, this is something like “Dave Meyer” - “who” you are**
- **A “domain name” can be used as a human-readable way of referring to an endpoint-id**

Desirable properties of endpoint-IDs

- Persistence: long-term binding to the thing that they name
  - These do not change during long-lived network sessions
- Ease of administrative assignment
  - Assigned to and by organizations
  - Hierarchy is along these lines (like DNS)
- Portability
  - IDs remain the same when an organization changes provider or otherwise moves to a different point in the network topology
- Globally unique

Locators – “where” you are in the network

- Think of the source and destination “addresses” used in routing and forwarding
- Real-world analogy is street address (i.e. 3700 Cisco Way, San Jose, CA, US) or phone number (408-526-7128)
- Typically there is some hierarchical structure (analogous to number, street, city, state, country or NPA/NXX)

Desirable properties of locators

- Hierarchical assignment according to network topology (“isomorphic”)
- Dynamic, transparent renumbering without disrupting network sessions
- Unique when fully-specified, but may be abstracted to reduce unwanted state
  - Variable-length addresses or less-specific prefixes can abstract/group together sets of related locators
  - Real-world analogy: don’t need to know exact street address in Australia to travel toward it from San Jose
- Possibly applied to traffic without end-system knowledge (effectively, like NAT but without breaking the sacred End-to-End principle)

Why should I care about this?

- In IPv4 and ipv6, there are only “addresses” which serve as both endpoint-ids and locators
- This means they don’t have the desirable properties of either:
    - Assignment to organizations is painful because use as locator constrains it to be topological (“provider-based”)
    - Exceptions to topology create additional, global routing state - multihoming is painful and expensive
    - Renumbering is hard – DHCP isn’t enough, changing address disrupts sessions, weak authentication used, source-based filtering, etc.
- Doesn’t scale for large numbers of “provider-independent” or multi-homed sites

Why should I care (continued)?

- The really scary thing is that the scaling problem won't become obvious until (and if) ipv6 becomes widely-deployed
  - Larger ipv6 address space could result in orders of magnitude more prefixes (depending on allocation policy, provider behavior, etc.)
  - NAT is effectively implementing id/locator split – what happens if the ipv6 proponents' dream of a "NAT-free" Internet is realized?
  - Scale of IP network is still relatively small
  - Re-creating the "routing swamp" with ipv6 would be… ugly/bad/disastrous; it isn't clear what anyone could do to save the Internet if that happens
- Sadly, this has been mostly ignored in the IETF for 10+ years
- …and the concepts have been known for far longer… see "additional reading" section

Can ipv6 be fixed? (and what is GSE, anyway?)

- **Can we keep ipv6 packet formats but implement the identifier/locator split?**
- **Mike O'Dell proposed this in 1997 with 8+8/GSE**

  http://ietfreport.isoc.org/idref/draft-ietf-ipngwg-gseaddr

- **Basic idea: separate 16-byte address into 8-byte EID and 8-byte "routing goop" (locator)**
  - **Change TCP/UDP to only care about EID (requires incompatible change to tcp6/udp6)**
  - **Allow routing system to modify RG as needed, including on packets "in flight", to keep locators isomorphic to network topology**

GSE benefits

- Achieves goal of EID/locator split while keeping most of ipv6 and (hopefully) without requiring a new database for EID-to-locator mapping
- Allows for scalable multi-homing by allowing separate RG for each path to an end-system; unlike shim6, does not require transport-layer complexity to deal with multiple addresses
- Renumbering can be fast and transparent to hosts (including for long-lived sessions) with no need to detect failure of usable addresses

GSE issues

- Incompatible change needed to tcp6/udp6 (specifically, to only use 64 bits of address for TCP connections)
  - in 1997, no installed base and plenty of time for transition
  - may be more difficult today (but it will only get a lot worse…)
- Purists argue violation of end-to-end principle
- Perceived security weakness of trusting “naked” EID (Steve Bellovin says this is a non-issue)
- Mapping of EID to EID+RG may add complexity to DNS, depending on how it is implemented
- Scalable TE not in original design; will differ from IPv4 TE, may involve “NAT-like” RG re-write
- Currently not being pursued (expired draft)

GSE is only one approach

- GSE isn’t the only (or perhaps easiest) way to do this but it is a straightforward retro-fit to the existing protocols
- Other approaches include:
  - Full separation of EID/locator (NIMROD…see additional reading section)
  - Tunnelling (such as IP mobility and/or MPLS)
  - Associating multiple addresses with connections (SCTP)
  - Adding hash-based identifiers (HIP)
- Each has pluses and minuses and would require major changes to protocol and application implementations and/or to operational practices
- More importantly, each of these is either not well enough developed (GSE, NIMROD) or positioned as a general-purpose, application-transparent retrofit to existing ipv6 (tunelling, SCTP, HIP, NIMROD); more work is needed

What about shim6/multi6?

- Approx 3-year-old IETF effort to retro-fit an endpoint-id/locator split into the existing ipv6 spec
- Summary: end-systems are assigned an address (locator) for each connection they have to the network topology (each provider); one address is used as the id and isn't expected to change during session lifetimes
- A "shim" layer hides locator/id split from transport (somewhat problematic as ipv6 embeds addresses in the transport headers)
- Complexity around locator pair selection, addition, removal, testing of liveness, etc… to avoid address changes being visible to TCP…all of this in hosts rather than routers

What about shim6/multi6? (continued)

- Some perceive as an optional, “bag on the side” rather than a part of the core architecture…
- Will shim6 solve your problems and help make ipv6 both scalable and deployable in your network?
- Feedback thus far: probably not (to be polite…)
    - SP objection: doesn’t allow site-level traffic-engineering in manner of IPv4; TE may be doable but will be very different and will add greater dependency on host implementations and administration
    - Hosting provider objection: requires too many addresses and too much state in web servers
    - End-users: still don’t get “provider-independent addresses” so still face renumbering pain
- Dependencies on end-hosts (vs. border routers with NAT or GSE) have implications for deployment, management, etc.

What if nothing is changed?

- How about a “thought experiment”?
- Make assumptions about ipv6 and Internet growth
- Take a guess at growth trends
- Pose some questions about what might happen
- What is the “worst-case” scenario that providers, vendors, and users might face?

My cloudy crystal ball: a few assumptions

- **ipv6 will be deployed in parallel to IPv4 and will be widely adopted**
- **IPv4 will be predominant protocol for near-to-mid term and will continue to be used indefinitely**
- **IPv4 routing state growth, in particular that for multi-homed sites, will continue to grow at a greater than linear rate up to or beyond address space exhaustion; ipv6 routing state growth curve will be similar - driven by multihoming**
- **As consequence of above, routers in the “DFZ” will need to maintain full routing/forwarding tables for both IPv4 and ipv6; tables will continue to grow and will need to respond rapidly in the face of significant churn**

A few more assumptions

- ipv6 prefix assignments will be large enough to allow virtually all organizations to aggregate addresses into a single prefix; in only relatively few cases (consider acquisitions, mergers, etc.) will multiple prefixes need to be advertised for an organization into the “DFZ”
- shim6 will not see significant adoption beyond possible edge use for multi-homing of residences and very small organizations
- IPv4-style multi-homing will be the norm for ipv6, implying that all multi-homed sites and all sites which change providers without renumbering will need to be explicitly advertised into the “DFZ”

Even more assumptions

- as the Internet becomes more mission-critical a greater fraction of organizations will choose to multi-home
- IPv4-style traffic engineering, using more-specific prefix advertisements, will be performed with ipv6; this practice will likely increase as the Internet grows
- Efforts to reduce the scope of prefix advertisements, such as AS_HOPCOUNT, will not be adopted on a large enough scale to reduce the impact of more-specifics in the "DFZ"

Questions to ask or worry about

- How much routing state growth is due to organizations needing multiple IPv4 prefixes? Some/most of these may be avoided with ipv6.
- As a result of available larger prefixes, will the number of prefixes per ASN decrease toward one? What is the likelihood that ASN usage growth will remain linear? (probably low)
  - Today, approximately 30,000 ASNs in use, so IPv4 prefixes-per ASN averages around 6-to-1 or so… how much better will this be with ipv6? 1-to-1? 2-to-1? More?
- How much growth is due to unintentional more-specifics? These may be avoided with ipv6.

More questions, more worries

- **How much growth is due to TE or other intentional use of more-specifics? These will happen with ipv6 unless draconian address allocation rules are kept (which is unlikely)**
  - **This appears to be an increasing fraction of the more-specifics**
- **What's the routing state "churn rate" and is it growing, shrinking, or remaining steady? (growing dramatically)**
- **What happens if we add more overhead to the routing protocols/system (think: SBGP/SoBGP)?**
  - **If the routing table is allowed to grow arbitrarily large, does validation become infeasible?**

Geoff Huston’s IPv4 BGP growth report

- How bad are the growth trends?
  - Prefixes: 130K to 170K in 2005 (196K as of 10/2006)
    - projected increase to ~370K within 5 years
    - global routes only – each SP has additional internal routes
  - Churn: 0.7M/0.4M updates/withdrawals per day
    - projected increase to 2.8M/1.6M within 5 years
  - CPU use: 30% at 1.5Ghz (average) today
    - projected increase to 120% within 5 years
- These are guesses based on a limited view of the routing system and on low-confidence projections (cloudy crystal ball); the truth could be worse, especially for peak demands
- No attempt to consider higher overhead (i.e. SBGP/SoBGP)
- Trend lines look exponential or quadratic; this is bad…
  - 200K (4Q06/1Q07) is an interesting number for some hardware…

Jason Schiller’s analysis: future routing state size

- Assume that wide spread ipv6 adoption will occur at some point
  - Put aside when - just assume it will happen
- What is the projection of the of the current IPv4 growth
  - Internet routing table
  - International de-aggregates for TE in the Internet routing table
  - Number of Active ASes
- What is the ipv6 routing table size interpolated from the IPv4 growth projections assuming everyone is doing dual stack and ipv6 TE in the “traditional” IPv4 style?
- Add to this internal IPv4 de-aggregates and ipv6 internal de-aggregates
- Ask vendors and operators to plan to be at least five years ahead of the curve for the foreseeable future

Current IPv4 Route Classification

- Three basic types of IPv4 routes
  - Aggregates
  - De-aggregates from growth and assignment of a non-contiguous block
  - De-aggregates to perform traffic engineering
- March 2006 Tony Bates CIDR report showed:

DatePrefixes	Prefixes	CIDR Agg
14-03-06	180,219	119,114

- Can assume that 61K intentional de-aggregates

Estimated IPv4+ipv6 Routing Table Size

Assume that tomorrow everyone does dual stack...

Current IPv4 Internet routing table:	180K routes
New ipv6 routes (based on 1 prefix per AS):	+ 21K routes
Intentional de-aggregates for IPv4-style TE:	+ 61K routes
Internal routes for tier-1 ISP	+ 50K to 150K routes
Internal customer de-aggregates	+ 40K to 120K routes
(projected from number of customers)	
Total size of tier-1 ISP routing table	352K to 532K routes

Given that tier-1 ISPs require IP forwarding in hardware (6Mpps), these numbers easily exceed the current FIB limitations of some deployed routers

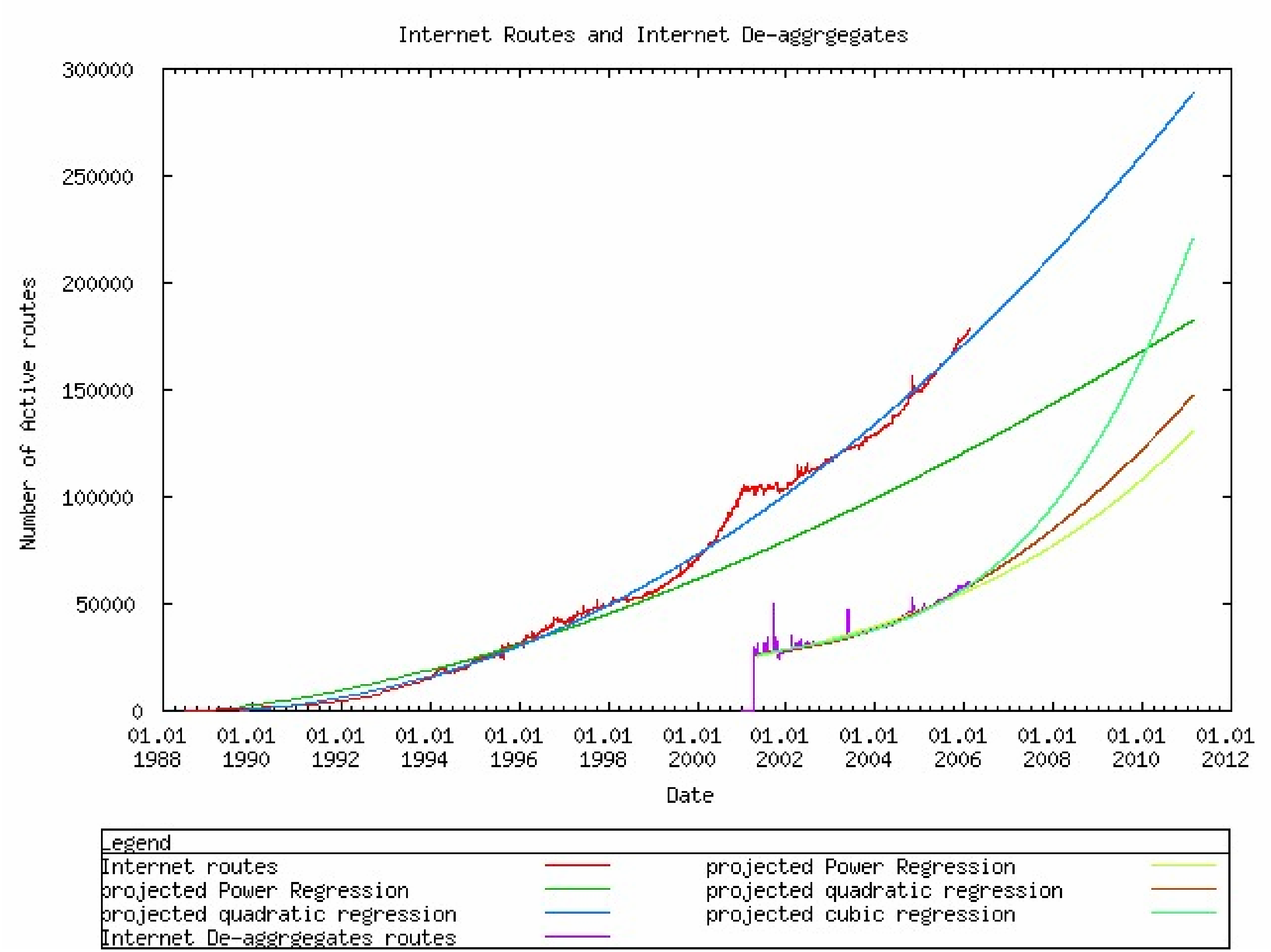
What this interpolation doesn't include

- A single AS that currently has multiple non-contiguous assignments that would still advertise the same number of prefixes to the Internet routing table if it had a single contiguous assignment
- All of the ASes that announce only a single /24 to the Internet routing table, but would announce more specifics if they were generally accepted (assume these customers get a /48 and up to /64 is generally accepted)
- All of the networks that hide behind multiple NAT addresses from multiple providers who change the NAT address for TE. With ipv6 and the removal of NAT, they may need a different TE mechanism.
- All of the new ipv6 only networks that may pop up: China, Cell phones, coffee makers, toasters, RFIDs, etc.

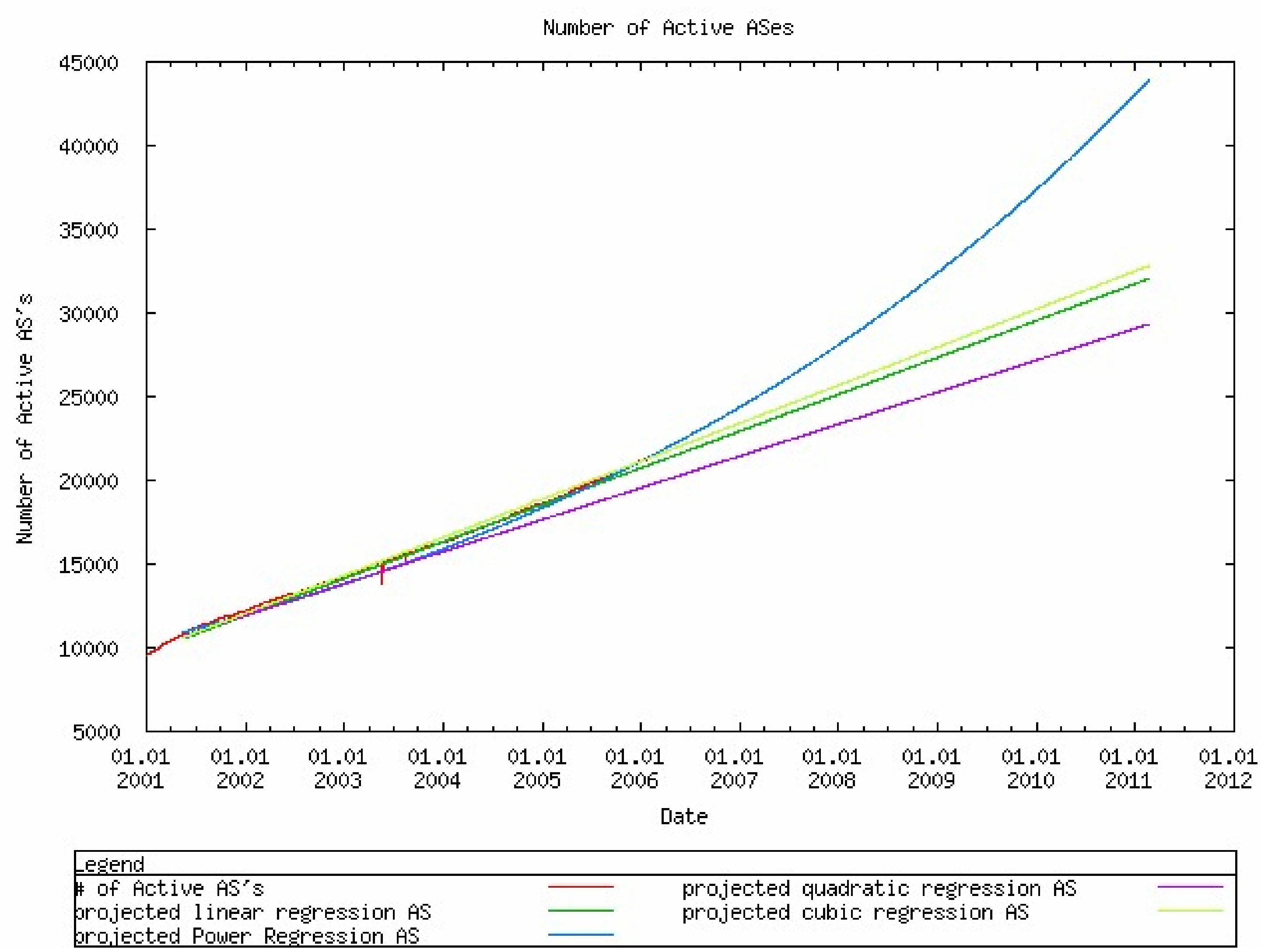
Projecting IPv6 Routing Table Growth

- Let's put aside the date when wide spread IPv6 adoption will occur
- Let's assume that wide spread IPv6 adoption will occur at some point
- What is the projection of the of the current IPv4 growth
    - Internet routing table
    - International de-aggregates for TE in the Internet routing table
    - Number of Active ASes
- What is the IPv6 routing table size interpolated from the IPv4 growth projections assuming everyone is doing dual stack and IPv6 TE in the "traditional" IPv4 style?
- Add to this internal IPv4 de-aggregates and IPv6 internal de-aggregates
- Ask vendors and operators to plan to be at least five years ahead of the curve for the forseeable future

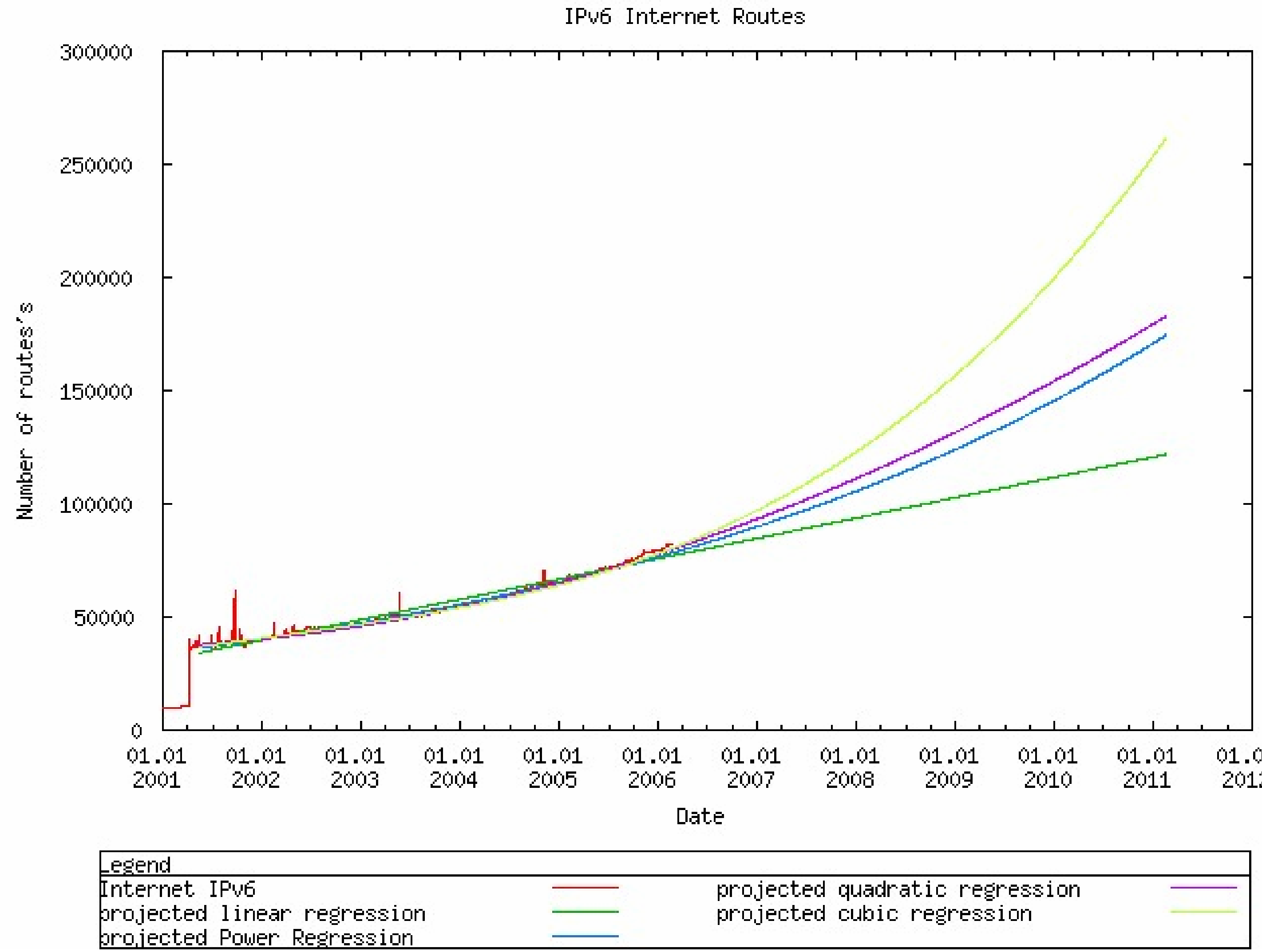
Trend: Internet CIDR Information
Total Routes and Intentional de-aggregates



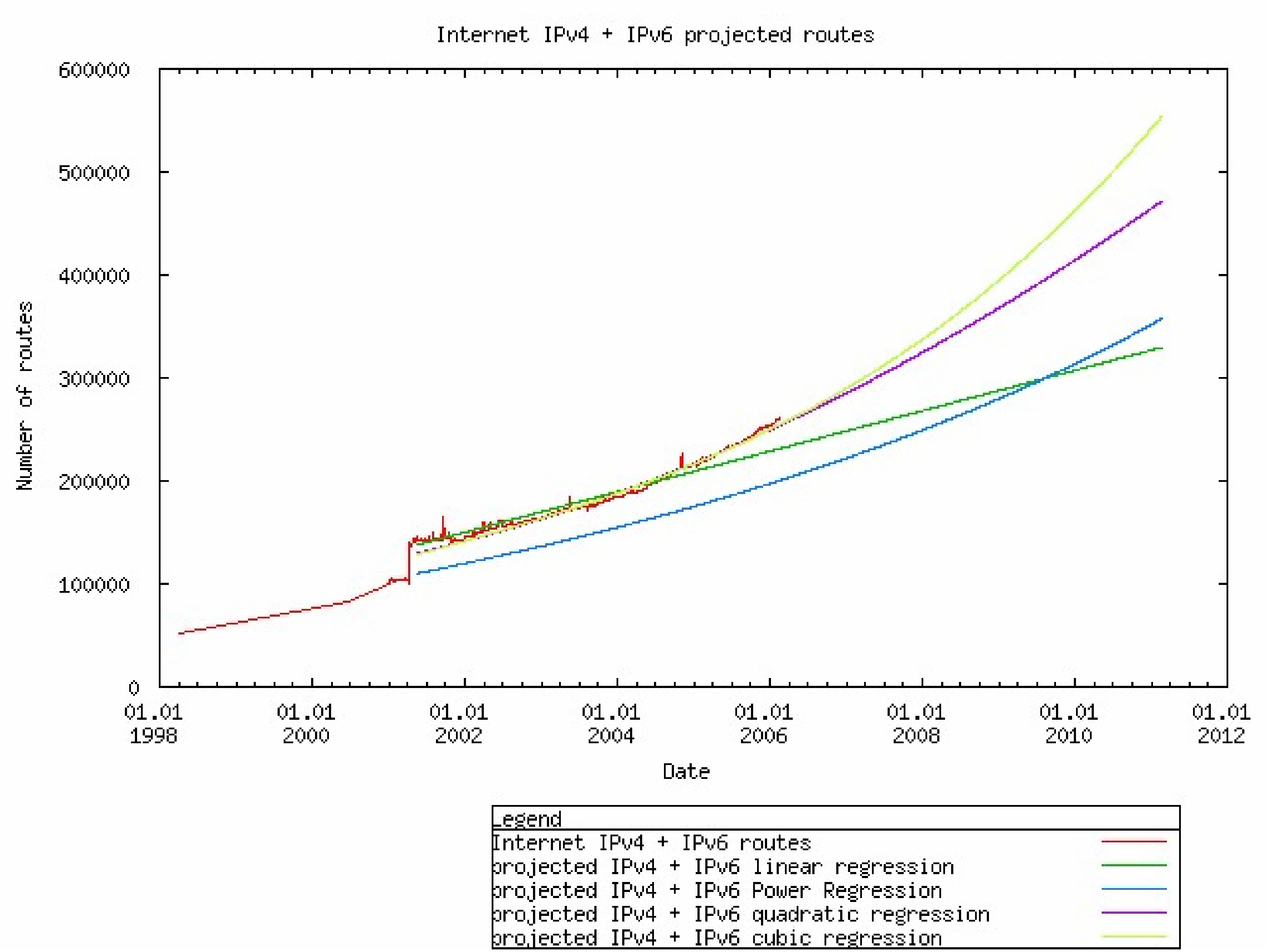
Trend: Internet CIDR Information Active ASes



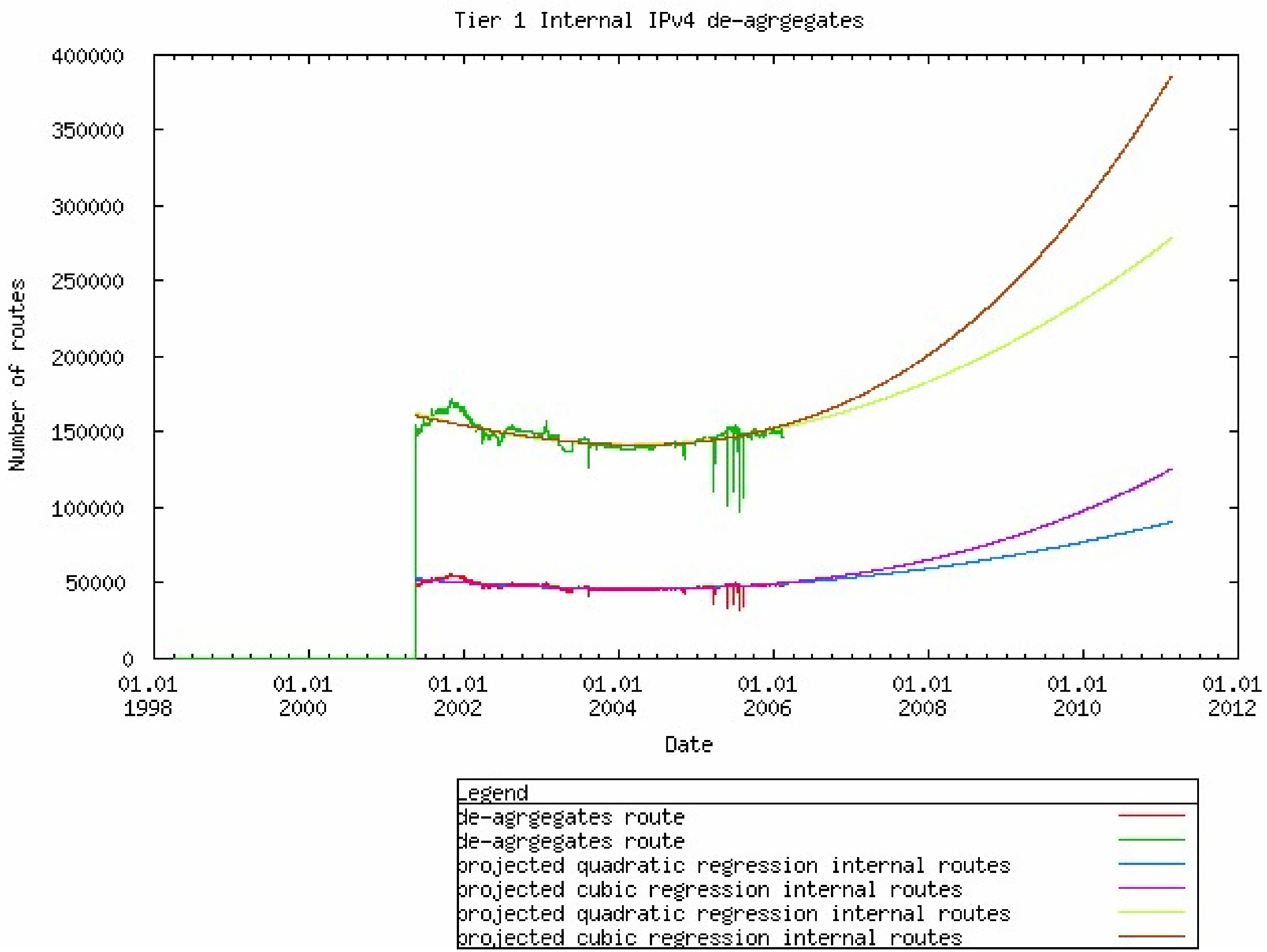
Future Projection of IPv6 Internet Growth
(IPv4 Intentional De-aggregates + Active ASes)



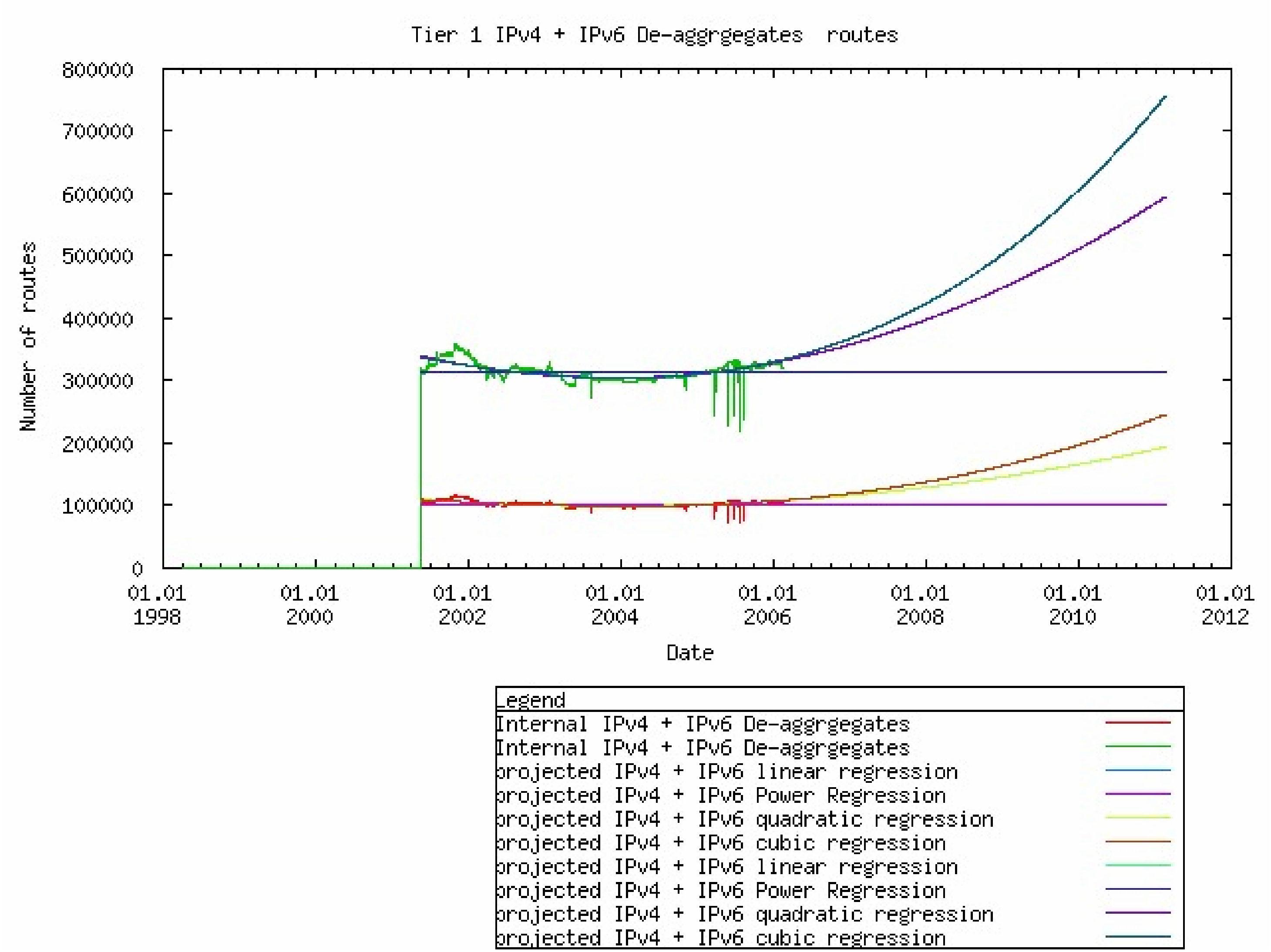
Future Projection of Combined IPv4 and IPv6 Internet Growth



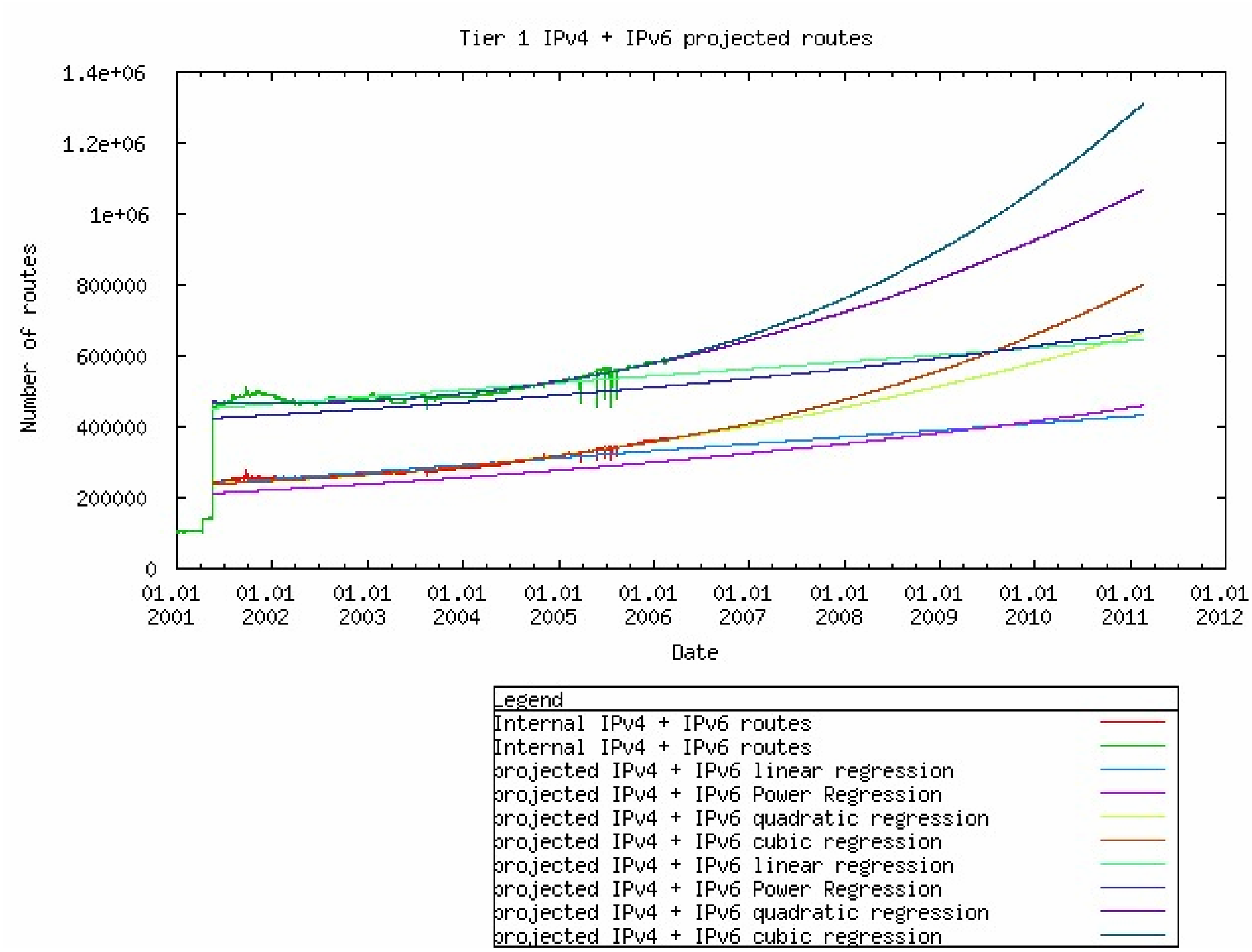
Tier 1 Service Provider IPv4 Internal de-aggregates



Future Projection Of Tier 1 Service Provider IPv4 and IPv6 Internal de-aggregates



Future Projection Of Tier 1 Service Provider IPv4 and IPv6 Routing Table



Summary of scary numbers

Route type	2006.03	5 years	7 years	10 Years	14 years
IPv4 Internet routes	180,219	285,064	338,567	427,300	492,269
IPv4 CIDR Aggregates	119,114				
IPv4 intentional de-aggregates	61,105	144,253	195,176	288,554	362,304
Active Ases	21,646	31,752	36,161	42,766	47,176
Projected ipv6 Internet routes	82,751	179,481	237,195	341,852	423,871
Total IPv4/ipv6 Internet routes	262,970	464,545	575,762	769,152	916,140
Internal IPv4 low number	48,845	88,853	117,296	173,422	219,916
Internal IPv4 high number	150,109	273,061	360,471	532,955	675,840
Projected internal ipv6 (low)	39,076	101,390	131,532	190,245	238,494
Projected internal ipv6 (high)	120,087	311,588	404,221	584,655	732,933
Total IPv4/ipv6 routes (low)	350,891	654,788	824,590	1,132,819	1,374,550
Total IPv4/ipv6 routes (high)	533,166	1,049,194	1,340,453	1,886,762	2,324,913

An upper bound? (Marshall Eubanks on PPML)

- **Are these numbers ridiculous?**
- **How many multi-homed sites could there really be? Consider as an upper-bound the number of small-to-medium businesses worldwide**
- **1,237,198 U.S. companies with >= 10 employees**
  - **(from http://www.sba.gov/advo/research/us_03ss.pdf)**
- **U.S. is approximately 1/5 of global economy**
- **Suggests up to 6 million businesses that might want to multi-home someday… would be 6 million routes if multi-homing is done with “provider independent” address space**
- **Of course, this is just a WAG… and doesn’t consider other factors that may or may not increase/decrease a demand for multi-homing (mobility? individuals’ personal networks, …?)**

Big Concerns

Current equipment purchases

- **Assuming wide spread IPv6 adoption by 2011**
- **Assuming equipment purchased today should last in the network for 5 years**
- **All equipment purchased today should support 1M routes**

Next generation equipment purchases

- **Assuming wide spread IPv6 adoption by 2016**
- **Assuming equipment purchased in 2012 should last in the network for 5 years**
- **Vendors should be prepared to provide equipment that scales to 1.8M routes**

Concerns and questions

- Can vendors plan to be at least five years ahead of the curve for the foreseeable future?
- How do operator certification and deployment plans lengthen the amount of time required to be ahead of the curve?
- Do we really want to embark on a routing table growth / hardware size escalation race for the foreseeable future? Will it be cost effective?
- Is it possible that routing table growth could be so rapid that operators will be required to start a new round of upgrades prior to finishing the current round? (remember the 1990s?)

What’s next?

- **Is there a real problem here? Or just “chicken little”?**
- **Should we socialize this anywhere else?**
- **Is the Internet operations community interested in looking at this problem and working on a solution? Where could/should the work be done?**
  - **IETF? Been there – IAB/IESG not very receptive**
    - **but soon an IAB workshop (good news?)**
  - **NANOG/RIPE/APRICOT?**
  - **ITU? YFRV? Research community? Other suggestions?**
- **Some discussion earlier this year at:**

  architecture-discuss@ietf.org

  ppml@arin.net
- **Sign up to help at: ipmh-interest@external.cisco.com**

Recommended Reading

“Endpoints and Endpoint names: A Proposed Enhancement to the Internet Architecture”, J. Noel Chiappa, 1999, http://users.exis.net/~jnc/tech/endpoints.txt

“On the Naming and Binding of Network Destinations”, J. Saltzer, August, 1993, published as RFC1498, http://www.ietf.org/rfc/rfc1498.txt?number=1498

“The NIMROD Routing Architecture”, I. Castineyra, N. Chiappa, M. Steenstrup. February 2006, published as RFC1992, http://www.ietf.org/rfc/rfc1992.txt?number=1992

“2005 – A BGP Year in Review”, G. Huston, APRICOT 2006, http://www.apnic.net/meetings/21/docs/sigs/routing/routing-pres-huston-routing-update.pdf