# Scaling issues with routing+multihoming

Vince Fuller, Cisco Systems

http://www.vaf.net/~vaf/apricot-plenary.pdf

# Acknowledgements

This is not original work and credit is due:

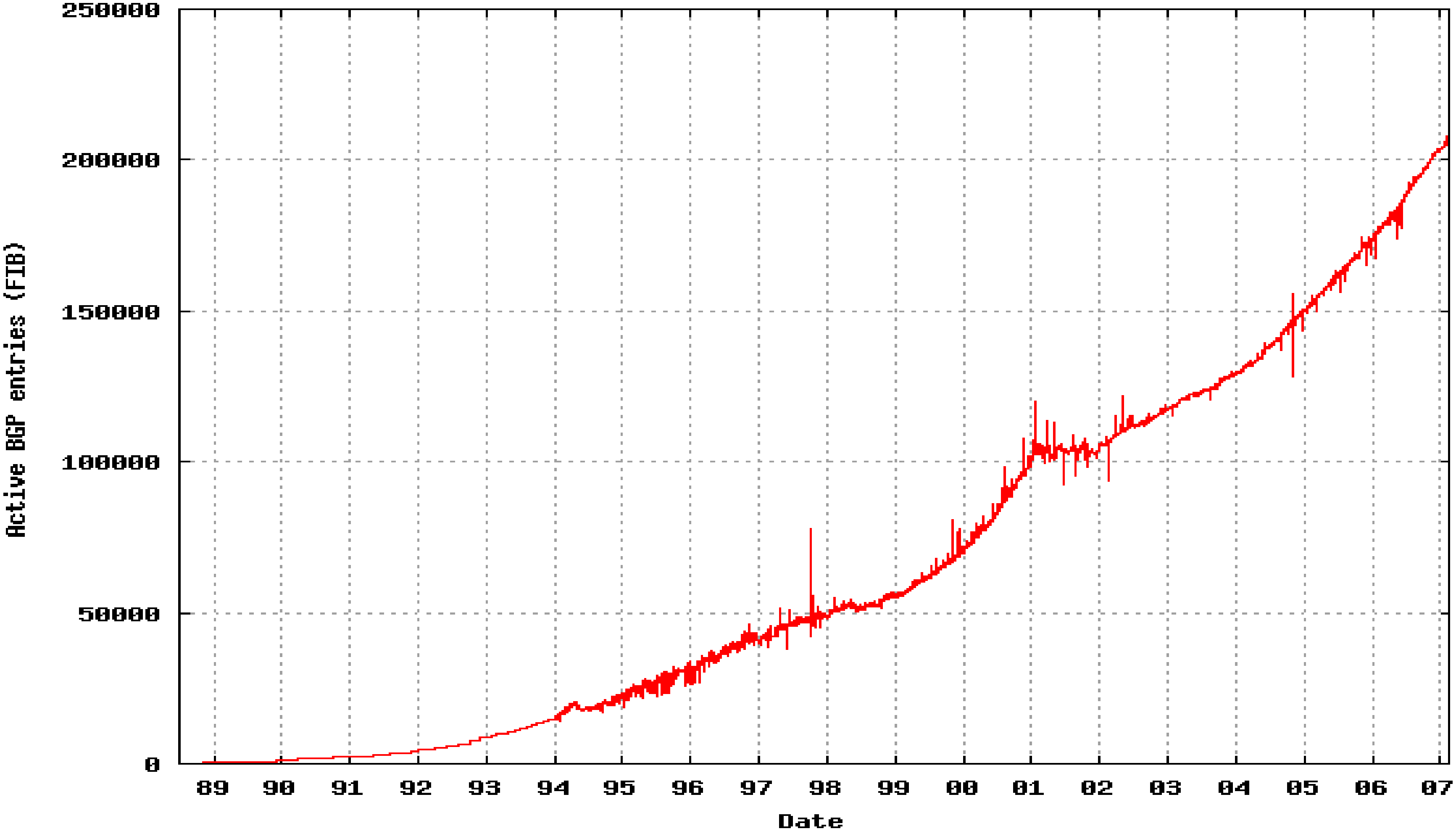
- Noel Chiappa for his extensive writings over the years on ID/Locator split
- Mike O'Dell for developing GSE/8+8
- Geoff Huston for his ongoing global routing system analysis work (CIDR report, BGP report, etc.)
- Jason Schiller and Sven Maduschke for the growth projection section (and Jason for tag-teaming to present this at NANOG)
- Tony Li for the information on hardware scaling
- Marshall Eubanks for finding and projecting the number of businesses (potential multi-homers) in the U.S. and the world

# Problem statement

- There are reasons to believe that current trends in the growth of routing and addressing state on the global Internet may cause difficulty in the long term
- The Internet needs an easier, more scalable mechanism for multi-homing with traffic engineering
- An Internet-wide replacement of IPv4 with ipv6 represents a one-in-a-generation opportunity to either continue current trends or to deploy something truly innovative and sustainable
- As currently specified, routing and addressing with ipv6 is not significantly different than with IPv4 – it shares many of the same properties and scaling characteristics

# A view of routing state growth: 1988 to now

***From bgp.potaroo.net/cidr/***



# IPv4 Current/near-term view - Geoff’s BGP report

- How bad are the growth trends? Geoff’s BGP reports show:
  - Prefixes: 130K to 170K (+30%) at end CY2005, 208K (+22%) on 2/15/07
    - ➢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)
- These kinda look exponential or quadratic; this is bad… and it’s not just about adding more cheap memory to systems

# Things are getting uglier… in many places

- Philip Smith's NANOG-39 "lightening talk":
  http://www.nanog.org/mtg-0702/presentations/smith-lightning.pdf
- Summary: de-aggregation is getting worse
  - De-aggregation factor: size of routing table/aggregated size
- For "original Internet", global de-agg factor is 1.85
  - North America: 1.69
  - EMEA: 1.53
- Faster-growing/developing regions are much higher:
  - Asia/Pacific: 2.48
  - Africa: 2.58
  - Latin/Caribbean: 3.40
- Trend may be additional pressure on table sizes, cause for concern

# What if we do nothing? Assume & project

- Assume ipv6 widely deployed in parallel with IPv4
  - Need to carry global state for both indefinitely
- Multihoming trends continue unchanged (valid?)
- ipv6 does IPv4-like mulithoming/traffic engineering
  - “PI” prefixes, no significant uptake of shim6
- Infer ipv6 table size from existing IPv4 deployment
  - One ipv6 prefix per ASN
  - One ipv6 more-specific per observed IPv4 more-specific
- Project historic growth trends forward
- Caveat: lots of scenarios for additional growth

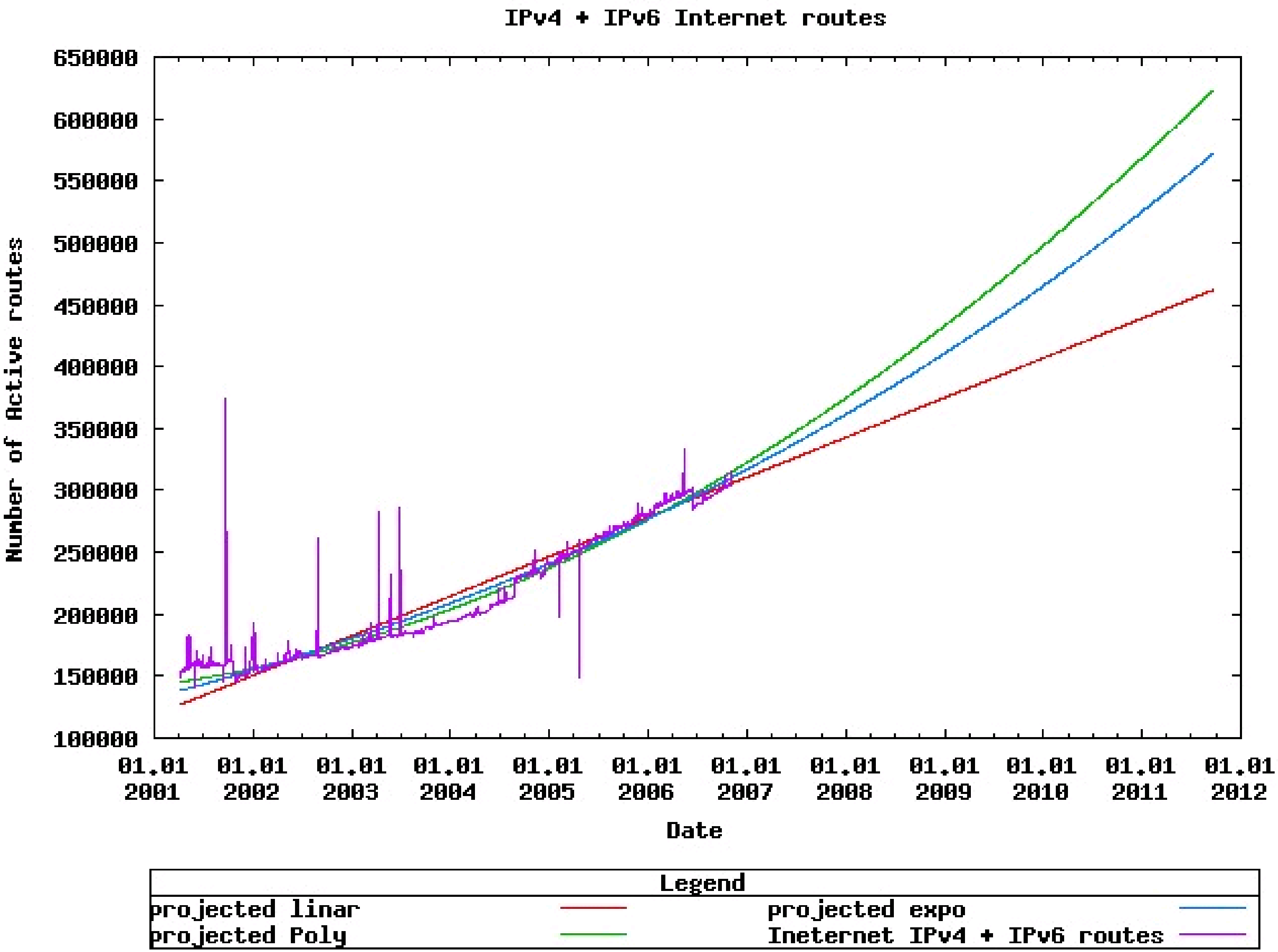
# Estimated IPv4+ipv6 Routing Table (Jason, 11/06)

## Assume that everyone does dual-stack tomorrow…

| | |
|---|---|
| Current IPv4 Internet routing table: | 199K routes |
| New ipv6 routes (based on 1 prefix per AS): | + 23K routes |
| Intentional ipv6 de-aggregates: | + 69K routes |
| Combined global IP-routing table | 291K routes |

- These numbers exceed the FIB size of some deployed equipment
- Of course, ipv6 will not be ubiquitous overnight
  - but if/when it is, state growth will approach projections
- This is only looking at the global table
- We'll consider the reality of "tier-1" routers next

# Plot: projection of combined IPv4 + ipv6 global routing state

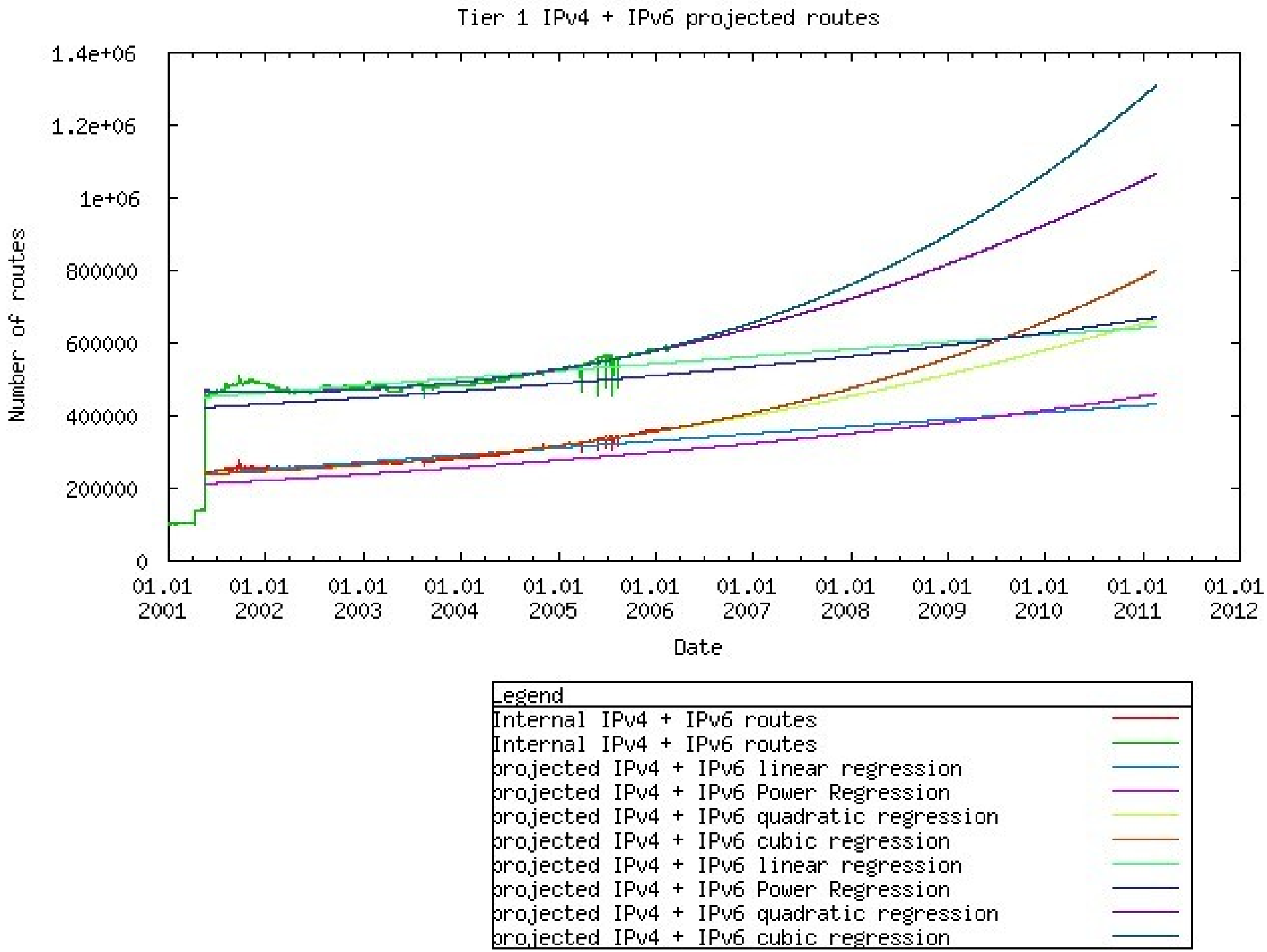


# “tier-1” internal routing table is bigger

| | |
|---|---|
| Current IPv4 Internet routing table: | 199K routes |
| New ipv6 routes (based on 1 prefix per AS): | + 23K routes |
| Intentional de-aggregates for IPv4-style TE: | + 69K routes |
| Internal IPv4 customer de-aggregates | + 50K to 150K routes |
| Internal ipv6 customer de-aggregates | + 40K to 120K routes |
| (projected from number of IPv4 customers) | |
| Total size of tier-1 ISP routing table | 381K to 561K routes |

**These numbers exceed the FIB limits of a lot of currently-deployed equipment… and this *doesn’t* include routes used for VPNs/VRFs (estimated at 200K to 500K for a large ISP today)**

# Plot: global routing state + "tier-1" internals



# Summary of big numbers

| Route type | 11/01/06 | 5 years | 7 years | 10 Years | 14 years |
|---|---|---|---|---|---|
| IPv4 Internet routes | 199,107 | 285,064 | 338,567 | 427,300 | 492,269 |
| IPv4 CIDR Aggregates | 129,664 | | | | |
| IPv4 intentional de-aggregates | 69,443 | 144,253 | 195,176 | 288,554 | 362,304 |
| Active Ases | 23,439 | 31,752 | 36,161 | 42,766 | 47,176 |
| Projected ipv6 Internet routes | 92,882 | 179,481 | 237,195 | 341,852 | 423,871 |
| Total IPv4/ipv6 Internet routes | 291,989 | 464,545 | 575,762 | 769,152 | 916,140 |
| | | | | | |
| Internal IPv4 (low est) | 48,845 | 101,390 | 131,532 | 190,245 | 238,494 |
| Internal IPv4 (high est) | 150,109 | 311,588 | 404,221 | 584,655 | 732,933 |
| | | | | | |
| Projected internal ipv6 (low est) | 39,076 | 88,853 | 117,296 | 173,422 | 219,916 |
| Projected internal ipv6 (high est) | 120,087 | 273,061 | 360,471 | 532,955 | 675,840 |
| | | | | | |
| Total IPv4/ipv6 routes (low est) | 381,989 | 654,788 | 824,590 | 1,132,819 | 1,374,550 |
| Total IPv4/ipv6 routes (high est) | 561,989 | 1,049,194 | 1,340,453 | 1,886,762 | 2,324,913 |

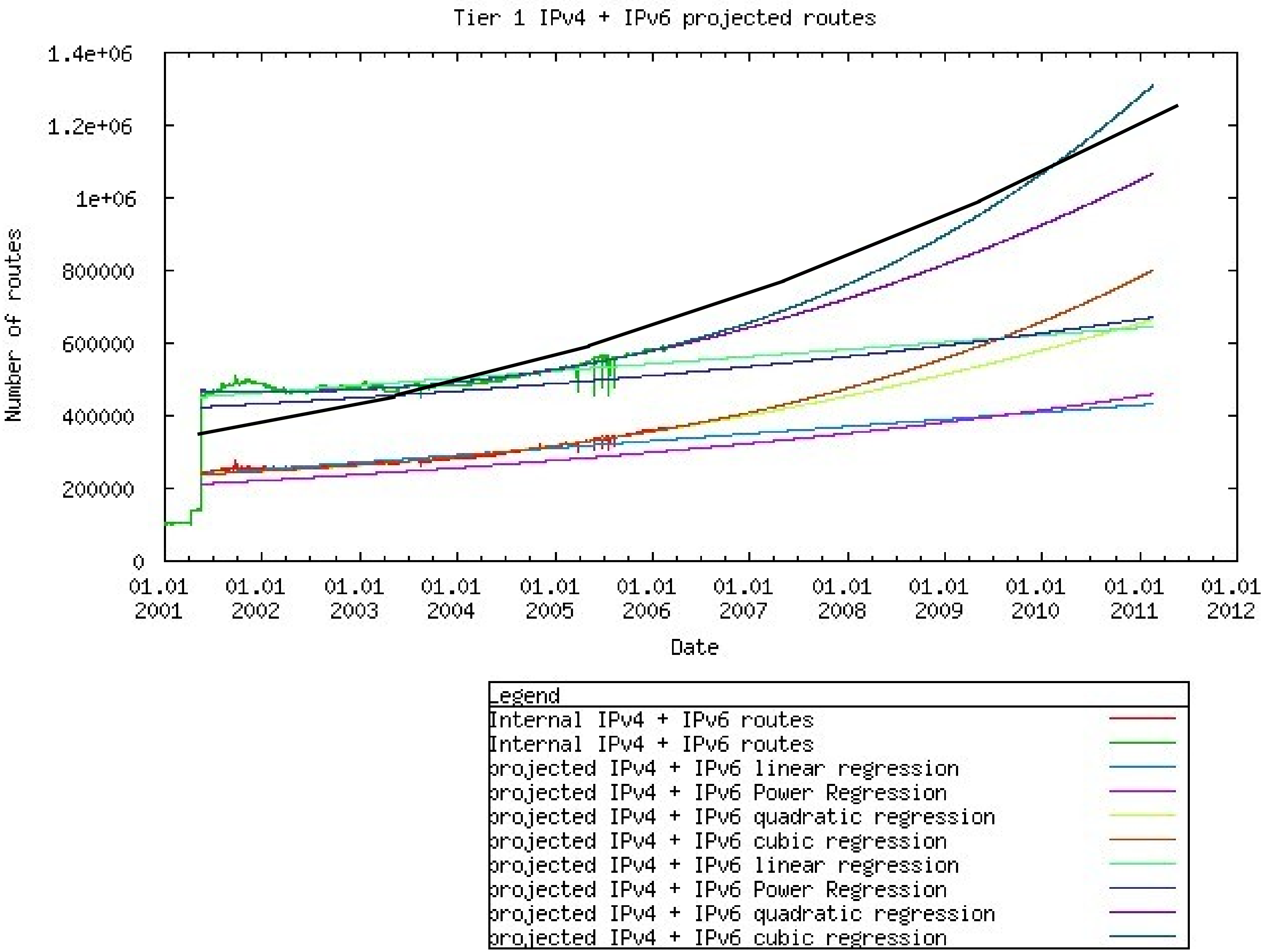
# Are these numbers insane?

- Marshall Eubanks did some analysis during discussion on the ARIN policy mailing list (PPML):
- 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, …?)

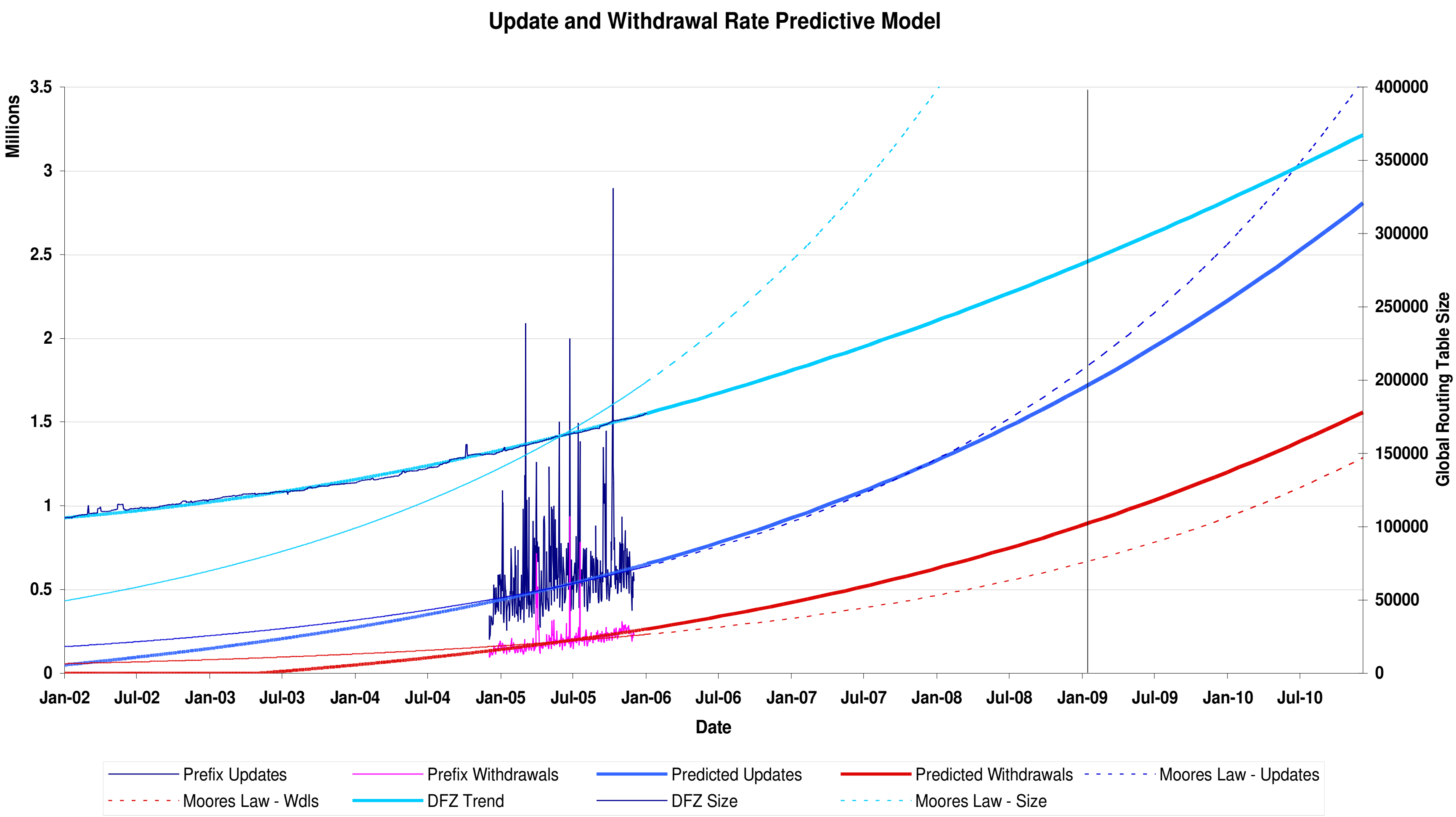
# Won’t “Moore’s Law” save us? Maybe

- DRAM-based RIB/FIB should be able to ride growth curve, so raw size may not be a problem
  - Designers says no problem building 10M-entry RIB/FIB)
  - But with what tradeoffs? Power/chip space are real issues
- TCAM/SRAM are low-volume and have much lower growth rates; platforms that using those will have issues
- Forwarding ASICs already push limits of tech.
- “Moore’s Law” tracks component density, not speed
  - Memory speeds improve at only about 10% per year
- BGP and RIB/FIB update rates are bounded by memory/CPU speeds and seem to be growing non-linearly; “meshiness” of topology is an issue

# Hardware growth vs. routing state growth



# Plot of growth trends vs. “Moore’s Law”




**Source: Huston/Armitage - http://www.potaroo.net/papers/phd/atnac-2006/bgp-atnac2006.pdf**

# Current direction doesn't seem to be helping

- **Original ipv6 strict hierarchical assignments**
  - **Fails in the face of large numbers of multi-homed sites**
  - **RIRs already moving away**
- **"PI for all" – see the earlier growth projections**
- **"geographic/metro/exchange" – constrains topology, requires new regulatory regime**
  - **"*Addressing can follow topology or topology can follow addressing; choose one" – Y. Rekhter***
- **Shim6 – maybe workable for SOHO but nobody (SPs, hosting providers, end-sites) wanting it**

# So, why doesn't IP routing scale?

- **It's all about the schizophrenic nature of addresses**
  - **they need to provide location information for routing**
  - **but also identify the endpoints for sessions**
- **For routing to scale, locators need to be assigned according to topology and change as topology changes ("*Addressing can follow topology or topology can follow addressing; choose one" – Y. Rekhter)***
- **But as identifiers, assignment is along organizational hierarchy and stability is needed – users and applications don't want renumbering when network attachment points change**
- **A single numbering space cannot serve both of these needs in a scalable way (see "further reading" section for a more in depth discussion of this)**
- **The really scary thing is that the scaling problem won't become obvious until (and if) ipv6 becomes widely-deployed**

# Maybe we something other than “addresses”?

- **What if instead of addresses there were “endpoint identifiers” associated with sites and “locators” used by the routing system?**
  - **Identifiers are hierarchically assigned to sites along administrative lines (like DNS hostnames) and do not change on devices that remain associated with the site; think “provider-independent” numbering but not routable**
  - **Locators are assigned according to the network topology; think “provider-based” CIDR block address assignments**
  - **Locators are aggregated/abstracted at topological boundaries to keep routing state scalable**
  - **When site’s connection to network topology changes, so do the locators – aggregation is preserved**

# A new approach - continued

- **This is not a new idea – see the "additional reading" section for more discussion about the concepts of endpoint naming and topological locators**
- **October IAB-sponsored workshop found fairly good consensus among a group of ISPs, vendors, IESG, and IAB that the problem exists and needs to be solved… ID/LOC separation seems likely part of the solution**
- **More recent email list discussions suggest that we are far from good consensus (and ugly politics/egos in the IETF may be muddling things a bit)**

# ID/LOC separation – a little bit of why and how

- **Common concepts:**
  - **Topologically-assigned locators (think “PA”)**
  - **Organizationally-assigned identifiers (think “PI”)**
- **Two different dimensions of approaches/trade-offs:**
  - **Host-based vs. network/router-based (which devices change?)**
  - **New name space vs. re-use/re-purpose of existing name space**
- **Several past and present approaches:**
  - **8+8/GSE – ipv6 address format (split into two parts), router changes, limited host changes**
  - **shim6/HIP/SCTP – new name space, major host changes**
  - **LISP – IPv4/ipv6 address format (different roles for prefixes), no host changes, some router changes**
  - **NIMROD – new name space, new routing architecture, no host changes (maybe)**

# Conclusions and recommendation

- **Currently specified IPv4 and ipv6 do not offer a scalable routing and addressing plans**
- **None of the options proposed in recent Internet drafts on address assignment policies offer a viable solution; in fact, they generally make the problem worse by codifying the construction of a brand-new "routing swamp"**
- **Work on a scalable solution is needed. That work will probably involve separation of the endpoint-id and locator functions of addresses used today**
- **The problem may become urgent; given vendor development and SP testing/deployment schedules, a solution needs to be designed within the next year or so if it is to be deployed in time to avoid problems with routing state projections in the 5-to-7 year timeframe.**
- **Next step: working group/design team? Vendors/providers already discussing this (a la CIDR deployment). Does IETF want to be part of the solution or part of the problem?**

# Recommended Reading - historic

"The Long and Winding ROAD", a brief history of Internet routing and address evolution, http://rms46.vlsm.org/1/42.html

"Endpoints and Endpoint names: A Proposed Enhancement to the Internet Architecture", J. Noel Chiappa, 1999, http://ana.lcs.mit.edu/~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

"GSE - An Alternative Addressing Architecture for IPv6", M. O'Dell, http://ietfreport.isoc.org/idref/draft-ietf-ipngwg-gseaddr

# Recommended Reading - recent work

"2005 – A BGP Year in Review", G. Huston, APRICOT 2006, http://www.apnic.net/meetings/21/docs/sigs/routing/routing-pres-husto

"Projecting Future IPv4 Router Requirementas from Trends in Dynamic BGP Behavior", G. Huston and G. Armitage, http://www.potaroo.net/papers/phd/atnac-2006/bgp-atnac2006.pdf

"Report from the IAB Workshop on Routing and Addressing", Meyer, D., Zhang, L., and Fall, K. (editors), http://www.ietf.org/internet-drafts/draft-iab-raws-report-00.txt

"Locator/ID Separation Protocol", Farainacci, D., Fuller, V., and D. Oran, http://www.ietf.org/internet-drafts/draft-farinacci-lisp-00.txt