Scaling of Internet Routing and Addressing:

past view, present reality, and possible futures

Vince Fuller, Cisco Systems

http://www.vaf.net/~vaf/iepg.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

Agenda

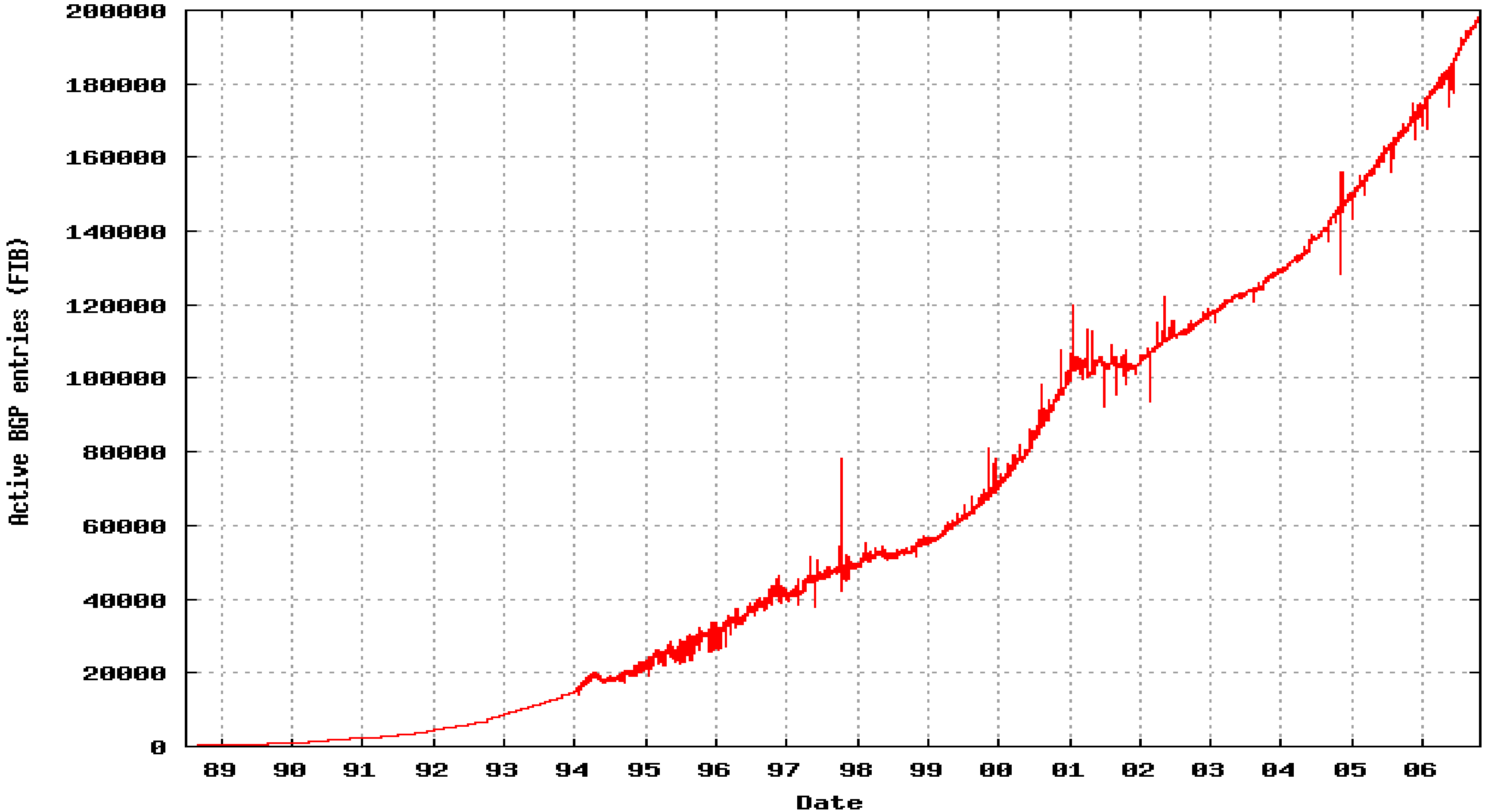
- Look at the growth of routing and addressing on the Internet
- Review the history of attempts to accommodate growth
- Examine current trends, scaling constraints imposed by hardware/cost limitations, and how the future might look if nothing changes
- Explore an alternative approach that might better serve the Internet community

Problem statement

- There are reasons to believe that current trends in the growth of routing and addressing state on the global Internet may not be scalable in the long term
- 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 doesn’t really differ from 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/***

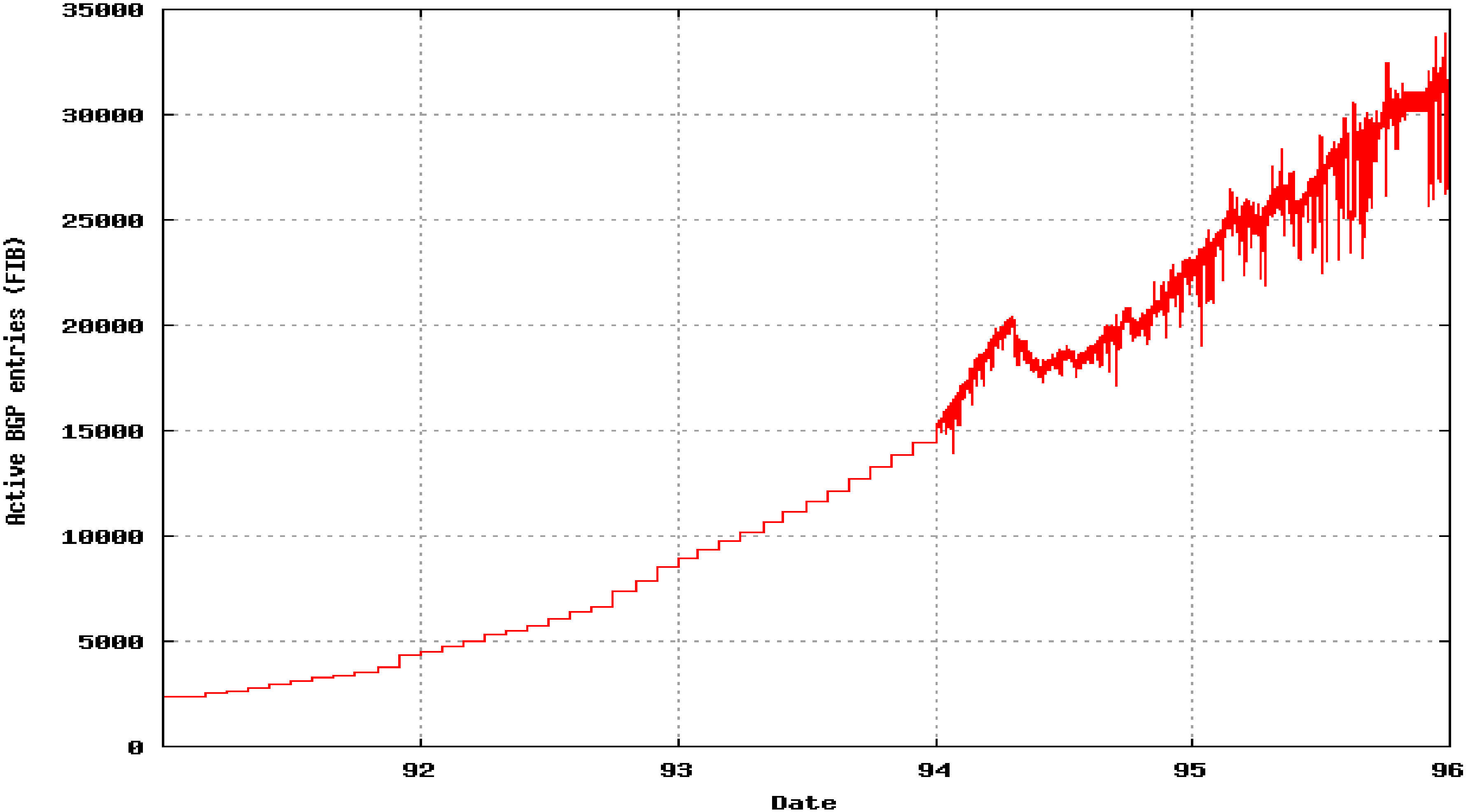


A brief history of Internet time

- 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 – described and proposed in 1992-1993, deployed starting in 1994

Pre- and early post-CIDR: 1991 - 1996

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



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)

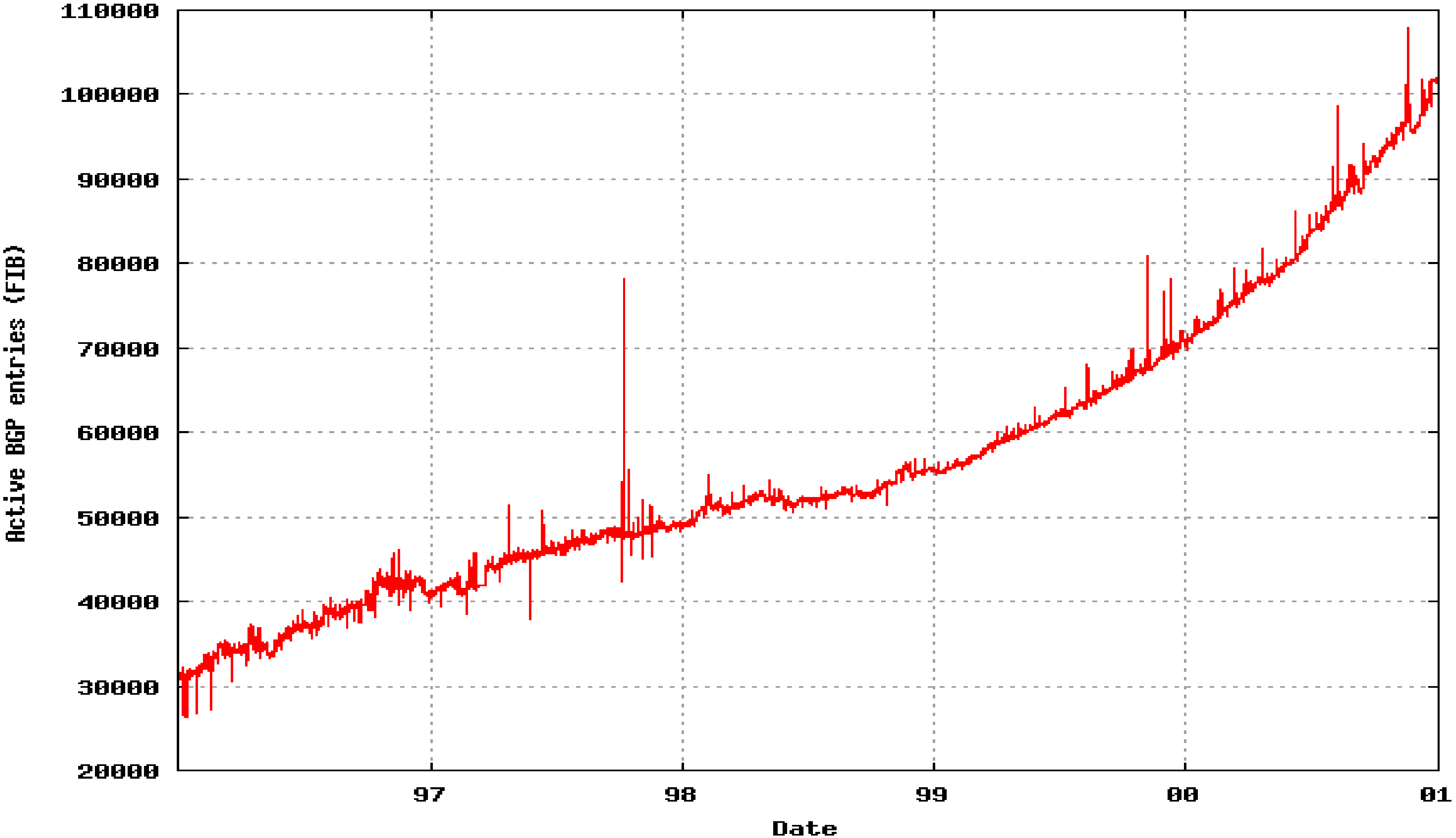
- Choice came down to politics, not technical merit
  - Hard issues deferred in favor of packet header design
- Things lost in shuffle…err compromise included:
  - Variable-length addresses
  - De-coupling of transport and network-layer addresses and clear separation of endpoint-id/locator (more later)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

Why doesn't ipv6 (or IPv4) routing scale?

- It's all about the schizophrenic nature of addresses
  - they need to be "locators" for routing information
  - but also serve as "endpoint id's" for the transport layer
- 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 (more on how to change this later)
- The really scary thing is that the scaling problem won't become obvious until (and if) ipv6 becomes widely-deployed

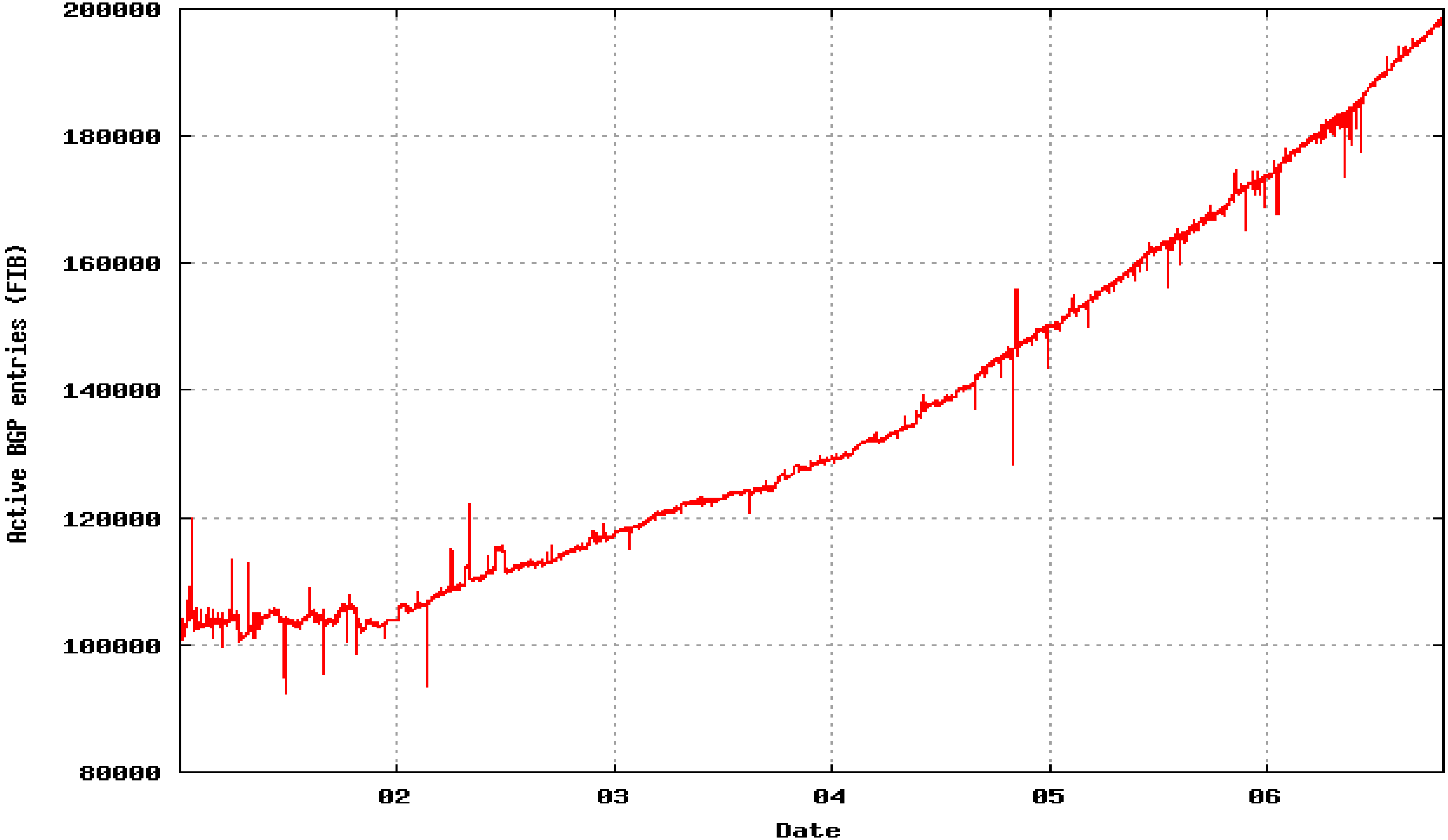
Internet boom: 1996 - 2001

From bgp.potaroo.net/cidr/



Post-boom to present: 2001 - 2006

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



View of the present: Geoff's IPv4 BGP report

- How bad are the growth trends? Geoff's BGP reports show:
  - Prefixes: 130K to 170K (+30%) in 2005 (200K/+17% thru 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)
- These kinda look exponential or quadratic; this is bad… and it's not just about adding more cheap memory to systems

What if we do nothing? Assume & project

- 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

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**

- Tony Bates CIDR report shows:

DatePrefixes	Prefixes	CIDR Agg
01-11-06	199,107	129,664

- Can assume that 69K intentional de-aggregates

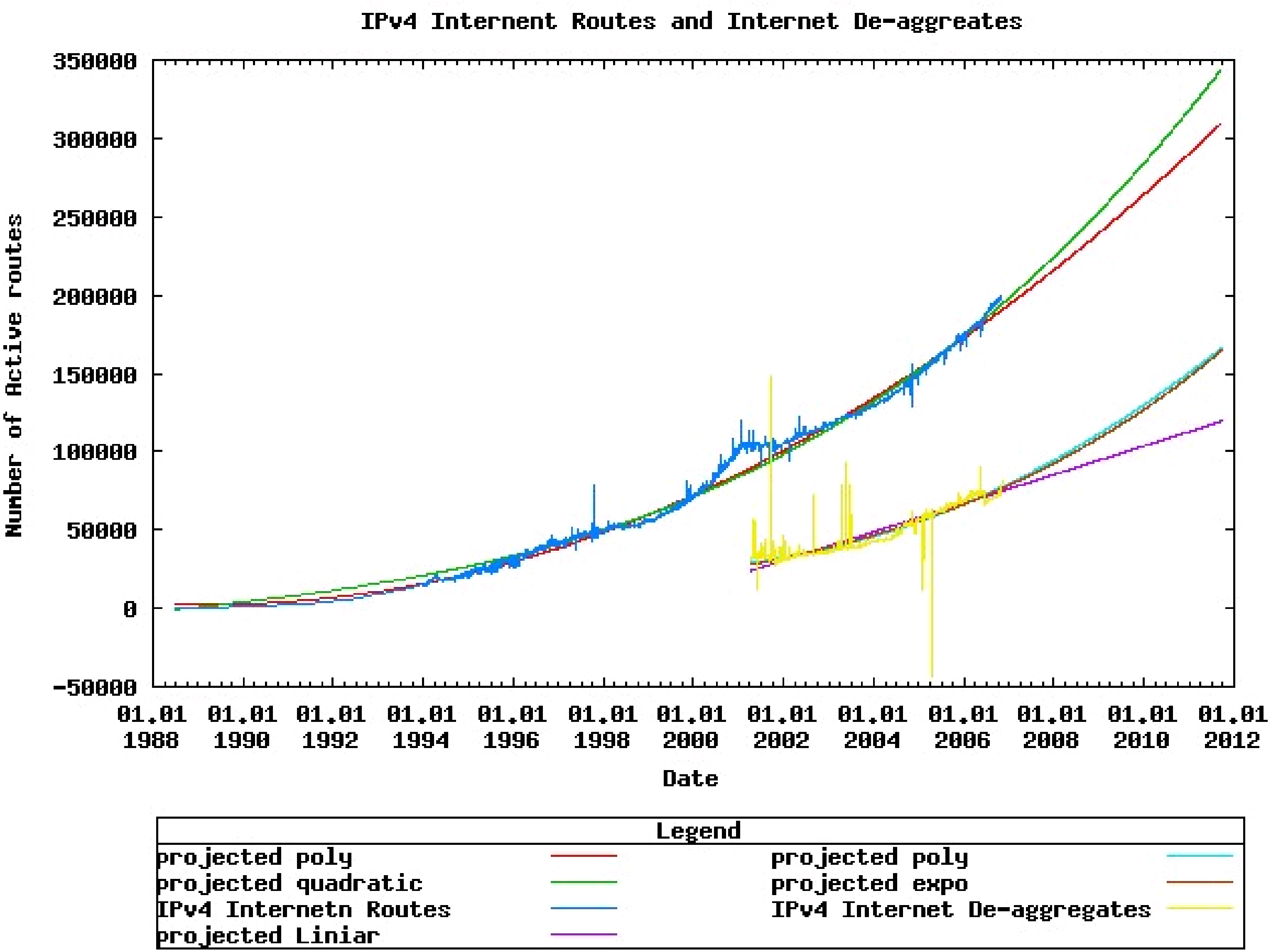
Estimated IPv4+ipv6 Routing Table (Jason, 11/06)

Assume that tomorrow everyone does dual stack...

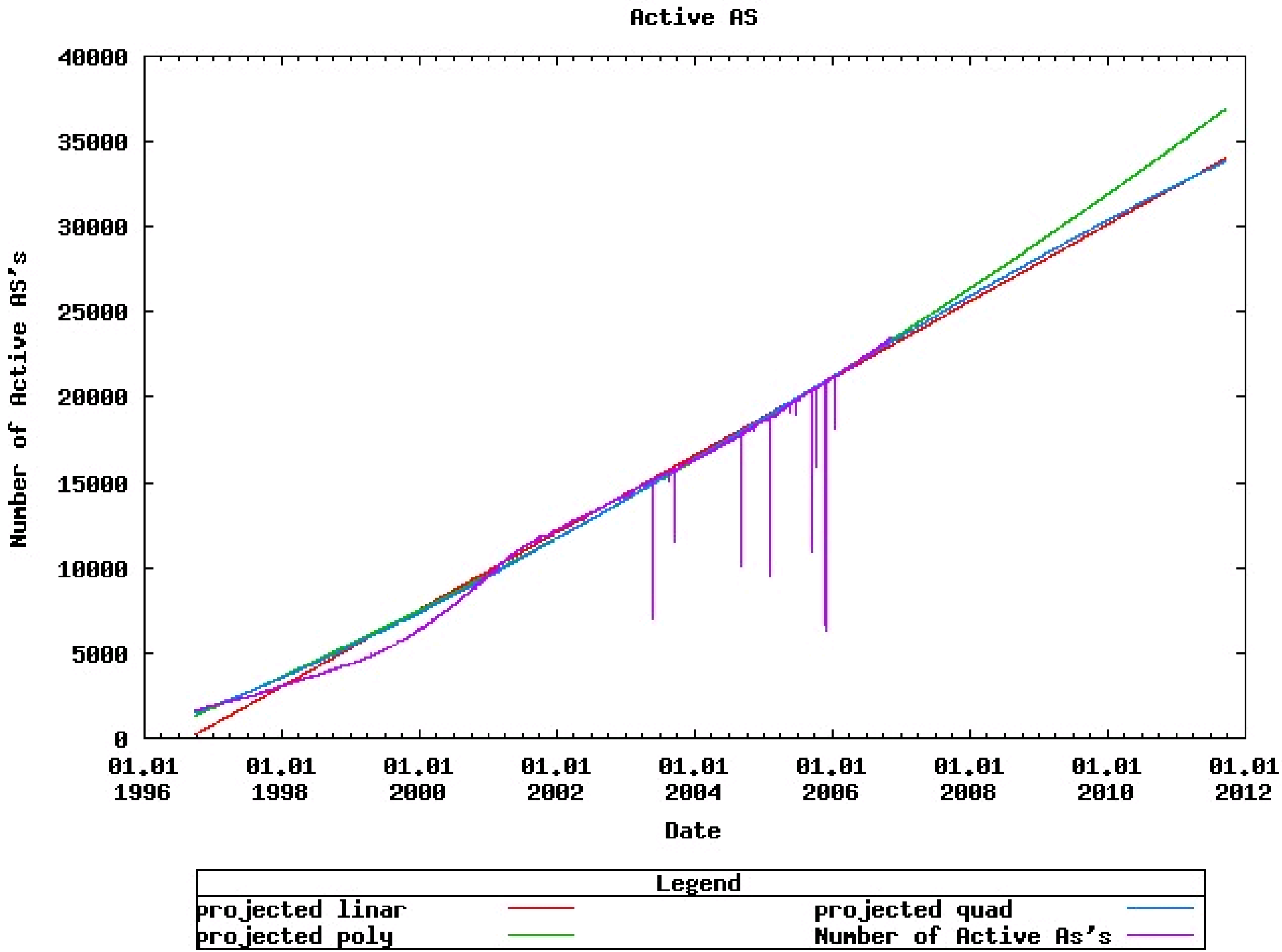
Current IPv4 Internet routing table:	**199K routes**
New ipv6 routes (based on 1 prefix per AS):	**+ 23K routes**
Intentional de-aggregates for IPv4-style TE:	**+ 69K routes**
Internal IPv4 customer de-aggregates	**+ 50K to 150K routes**
Internal ipv6 customer de-aggregates	**+ 40K to 120K routes**
(projected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

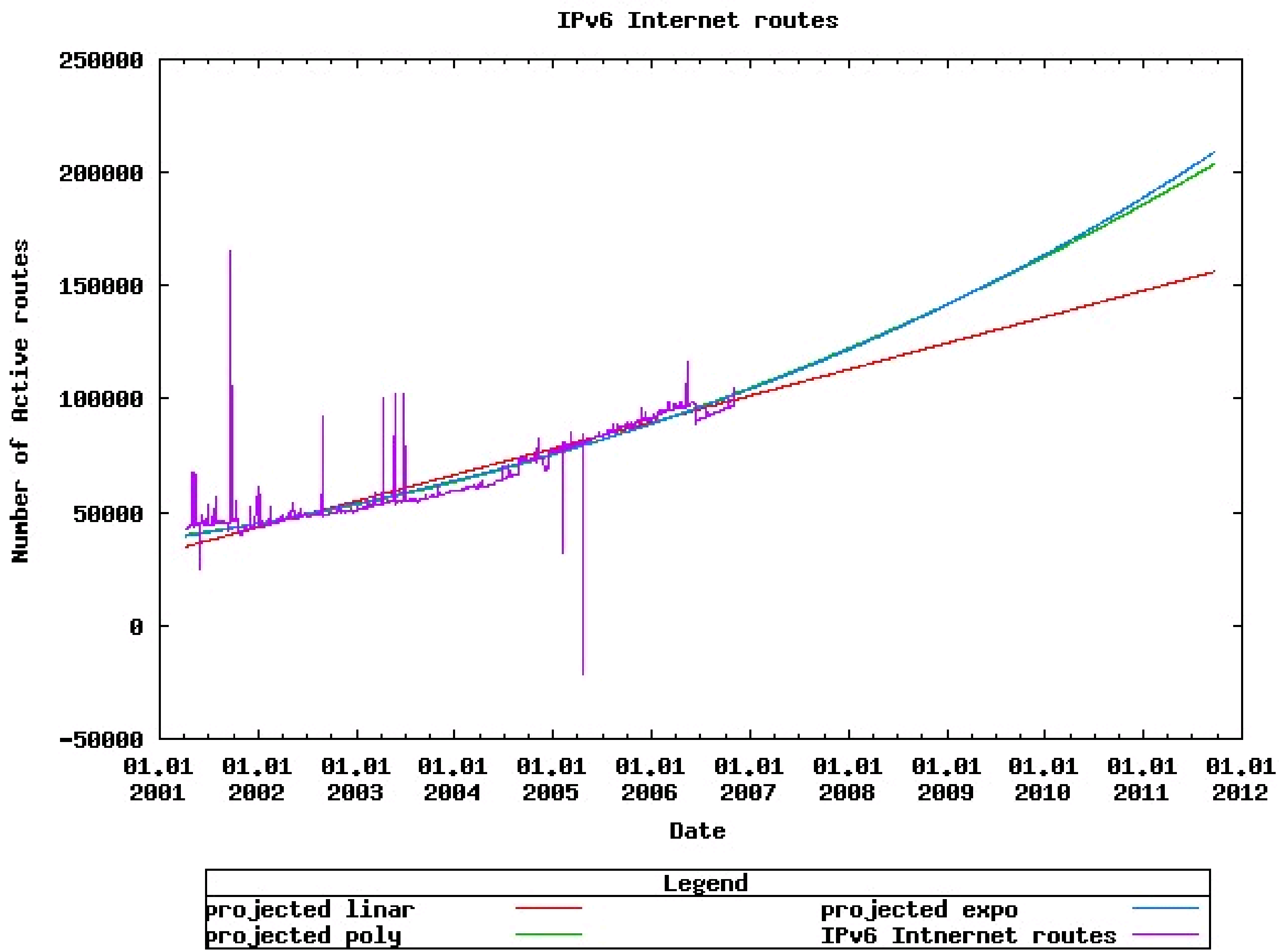
Trend: Internet CIDR Information
Total Routes and Intentional de-aggregates



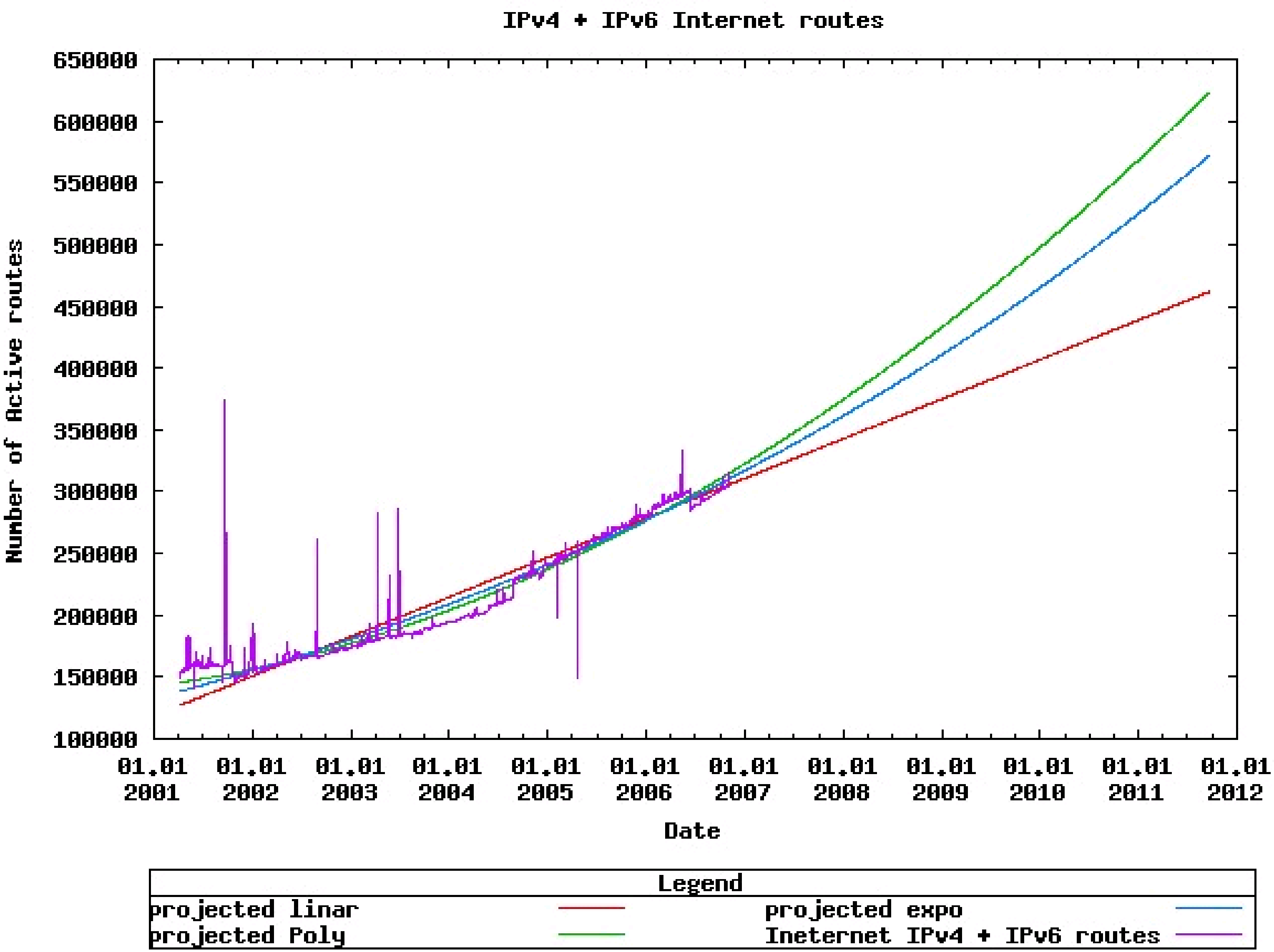
Trend: Internet CIDR Information Active ASes



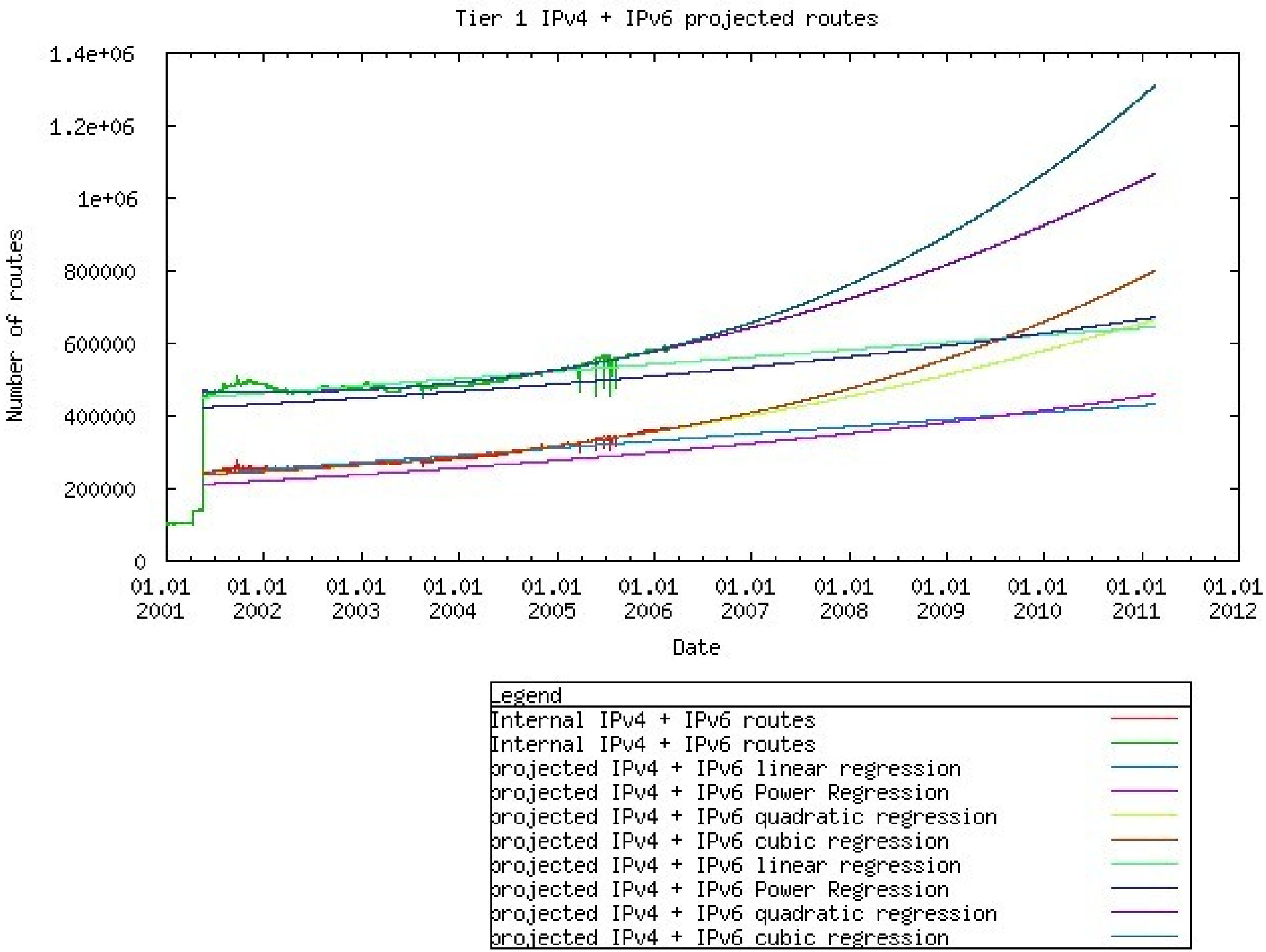
Future Projection of IPv6 Internet Growth
(IPv4 Intentional De-aggregates + Active ASes)



Future Projection of Combined IPv4 and IPv6 Internet Growth



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



Summary of scary 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

“it could be worse” - what this interpolation doesn’t try to consider

- **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.**

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, …?)

Router Scalability & Moore's "Law"

So, how do these growth trends compare to those for hardware size and speed? Won't "Moore's Law" just take care of that for us?

Definition:

> Moore's Law **is the empirical *observation* that the transistor density of integrated circuits, with respect to minimum component cost, doubles every 24 months. (Wikipedia)**

It isn't a *law* it's an *observation* that has nicely fit semiconductor growth trends since the 1960s

Moore's "Law" – assumptions and constraints

- Applicable to high volume components - think PC's, main (DRAM) memories, and disk drives
- Low volume applications can ride technology curve, not cost curve
- Critical router components don't fit this model
- Yes, DRAM capacity grows 4x/3.3yrs (2.4x/2yrs)
- …speed increases only about 10%/yr (1.2x/2yrs)
- …and BGP convergence degrades at table growth rate/speed improvement

Off-chip SRAM

- Requires high-speed, high-capacity parts
- Driver was PC cache, now on-chip
- Most of market is cell-phones, for low-power small-capacity parts
- Big fast SRAMs, which are critical in forwarding path of certain big routers, are not volume parts; off the cost curve
- Similar story for TCAMs – specialized, low-volume components in some routing/switching systems

Forwarding engines

- Forwarding engines most sophisticated ASICs being built, second only to CPUs
- Currently one generation behind CPUs
- Already past knee on price/performance
- High performance requires bandwidth; favors on-chip SRAM
- Gains so far have leveraged technology; little gain to be had
- Technological leadership will be expensive

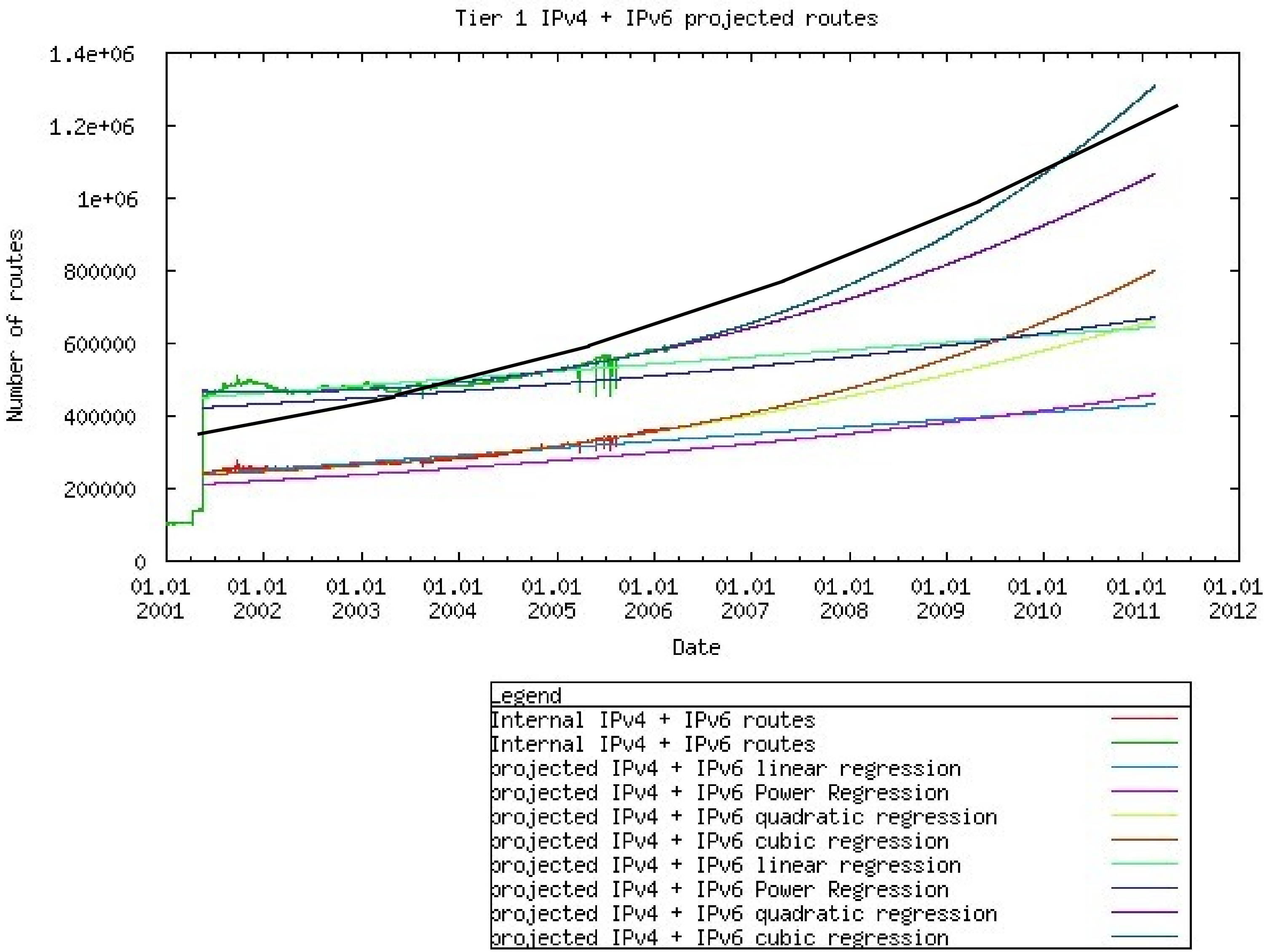
Chip costs

- Tapeout costs rising about 1.5x/2yrs (Wikipedia)
- Chip development costs rising similarly
- Net per-chip costs rising about 1.5x/2yrs
- Progress faster than 1.3x/2yrs will require non-linear cost
- Does not include CapEx, OpEx from continual upgrades

Moore’s “Law” - Summary

- Constant convergence growth rate is about 1.2x/2yrs
- Constant cost growth rate is about 1.3x/2yrs
- Current growth is from 1.3x/2yrs – >2x/2yrs
- Without architectural or policy constraints, costs are potentially unbounded
- Even with constraints, SPs are doomed to continual upgrades, passed along to consumers

Hardware growth vs. routing state growth



So, what's driving this problematic growth?

- In IPv4 and ipv6 use *addresses* both as session-layer *identifiers* and as routing *locators*
- This dual usage is problematic because:
    - Assignment to organizations is painful because use as *locator* constrains it to be topological ("provider-based") for routing to scale
    - Organizations would rather have *identifiers* so that they don't have to renumber if they change providers or become multi-homed within the network topology
- This dual-use doesn't scale for large numbers of "provider-independent" or multi-homed sites
- Perhaps a change to explicit use of *identifiers* and *locators* would offer scaling benefits… this general concept is termed the ID/LOC split

Digression: identifiers and locators

- **Think of an endpoint *identifier* 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 *identifier***

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 like 3700 Cisco Way, San Jose, CA, US or phone number (prior to mandated number portability) such as +1 408 526 7000
- 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 stuff?

- The scaling problem isn’t obvious now and won’t be 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 today; 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…bad; 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
  - ipv6 designers punted this problem to the RIRs by mandating that all ipv6 address-assignments would be “PA”; reality is that all RIRs are revising assignment policies to allow “PI” for all
- …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" (LOC)**
  - **Change TCP/UDP to only care about ID (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 ID/LOC split while keeping most of ipv6 and (hopefully) without requiring a new database for id-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 id/loc (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.**

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?)**

Conclusions and recommendations

- Projected growth trends of routing state will exceed the cost-effectiveness of hardware improvements
- Big implications for service provider expense, not only in $$ but also in space, power, cooling, and equipment refresh cycles
- More profit for vendors in short-term (remember the 1990s?) but more pain for all in the long-term
- An Internet-wide replacement of IPv4 with ipv6 represents a unique opportunity to either continue current trends or to deploy something truly innovative and sustainable
- As currently specified, routing and addressing with ipv6 is much the same as IPv4, with similar properties and scaling characteristics; perhaps a new approach, based on identifier/locator split, would be a better path forward

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?
    - Recent IAB workshop was good – problem recognized
    - NANOG/RIPE/APRICOT?
    - ITU? YFRV? Research community? Other suggestions?
- Some discussion earlier this year at:

  architecture-discuss@ietf.org

  ppml@arin.net
- More discussion at: ipmh-interest@external.cisco.com
- Stay tuned… more to come

Recommended Reading

"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://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-pr**