30 Years of BGP

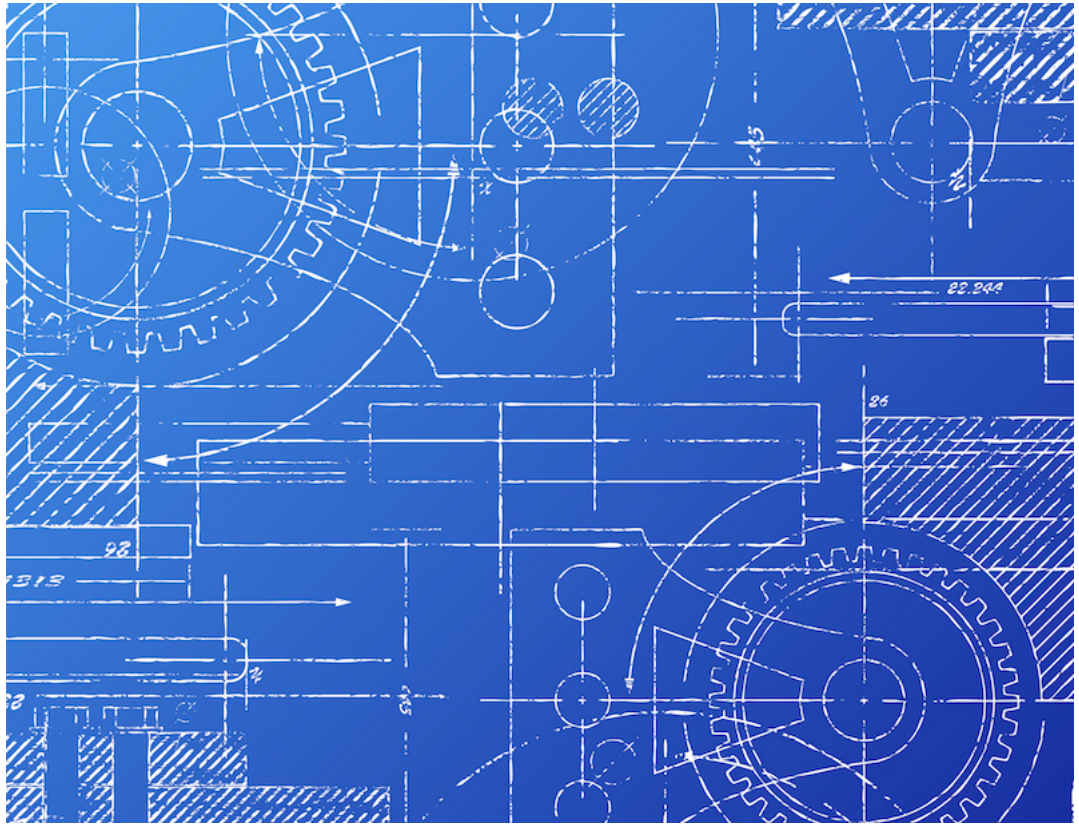
A Lesson in Protocol Evolution

Geoff Huston
APNIC

In the Beginning…

- BGP was an evolution of the earlier EGP protocol (developed in 1982 by Eric Rosen and Dave Mills)
- **BGP-1** – RFC 1105, June 1989, Kirk Lougheed, Yakov Rekhter
  - TCP-based message exchange protocol, based on distance vector routing algorithm with explicit path attributes
- **BGP-3** – RFC1267, October 1991, Kirk Lougheed, Yakov Rekhter
  - Essentially a clarification and minor tweaks to the basic concepts used in BGP
- **BGP-4** – RFC 1654, July 1994, Yakov Rekhter, Tony Li
  - Added CIDR (supporting explicit prefix lengths) and proxy aggregation

I - The Protocol Design of BGP



Routing Hierarchies

- Earlier protocols, notably DECnet Phase IV, supported scaling by hierarchies:
  - Within an “area” the routing protocol maintained a detailed topology that allowed all nodes within the area to reach any other node in the same area using links that were managed by the inter-area routing protocol
  - Area border routers maintained an inter-area topology
- BGP borrowed this concept, using the terminology of “Autonomous Systems” in a manner similar to the concept of “areas”
- Unlike DECnet, BGP did not define the ”interior” routing protocol, decoupling the concepts of internal and exterior routing in this two-level hierarchy

Routing Hierarchies

- Earlier protocols, notably DECnet Phase IV, su… scaling by hierarchies:
  - Within an "area" th… topology that allow… the same area u… …uting protocol
  - A … inter-area topology
- BGP … this concept, using the terminology of "Autonomous Systems" in a manner similar to the concept of "areas"
- Unlike DECnet, BGP did not define the "interior" routing protocol, decoupling the concepts of internal and exterior routing in this two-level hierarchy

Lesson:
Don't try to solve everything –
underachieving can be a virtue!

BGP Protocol

- BGP is a message passing protocol layered above TCP
- TCP manages:
  - Framing of individual elements of the protocol exchange
  - Reliability of the exchange
  - Flow control, including rate adaptation
- BGP assumes that as long as the TCP session remains up then everything that was passed to a peer is known by that peer for the duration of the session
  - BGP need only send changes, without periodic refresh for the lifetime of the session

BGP Protocol

- BGP is a message passing protocol layered above TCP
- TCP manages:
  - Framing of individual elements of the protocol exchange
  - Reliability of the exchange
  - Flow control
- BGP assumes that as long as the TCP session remains up then everything sent by each peer is known by that peer for the duration of the session
  - BGP need only send changes, without periodic refresh for the lifetime of the session

Lessons:
Reuse, don't re-invent!
Don't duplicate functionality

BGP and Packet Forwarding

- BGP does not alter IP packets
  - Its role is to inform routers on how to make forwarding decisions
- IP packets do not contain AS information
  - The association of IP addresses to an AS is a BGP concept. Within an AS, the interior routers and interior routing protocols and hosts have no knowledge of the local AS.
  - Which makes network rehoming in the AS space easy
  - Which prevents provider lock-in and aids in a competitive supply for transit

BGP and Packet Forwarding

- BGP does not alter IP packets
  - Its role is to inform routers on ha… …g decisions
- IP packets da…

Lessons:
Focus focus focus!
Limit side-effects as much as possible

…pt. Within an AS, the
…s have no knowledge of

…homing in the AS space easy
… prevents provider lock-in and aids in a competitive supply for transit

BGP Policy

- Each AS can determine its own traffic export policy autonomously
  - Within some constraints
- The AS Path concept was primarily there to prevent loops, nothing more
- BGP will by default prefer to use the shortest AS path
  - It's a crude LCD metric
  - But if the network admin wants to use some other route selection policy framework, then BGP won't stop you!
- Local BGP policy is opaque
  - Whatever your BGP policy settings may be, they are your policy settings, and no one else needs to know them!
  - What you accept from your peers and what you choose to re-advertise to your peers and why is your call and your business

BGP Policy

- Each AS can determine its own traffic export policy autonomously
  - Within some constraints
- The AS Path concept was primarily there to prevent loops, nothing more
- BGP will by default prefer to use the shortest AS path
  - It's a crude LCD metric
  - But if the network admin has a more elaborate policy framework, then BGP

Lesson:
Don't make the protocol force the business model

  - Your BGP policy settings may be, they are your policy settings, and no one else needs to know them!
  - What you accept from your peers and what you choose to re-advertise to your peers and why is your call and your business

BGP is Non-Deterministic

(Which is an odd property of a routing protocol!)

- BGP is best seen as a negotiation protocol, attempting to find a point of equilibrium between networks' export and import policies
- Subtle changes in timers and sequencing of BGP update processing means that the routing outcomes are not necessarily deterministic.

BGP is Non-Deterministic

(Which is an odd property of a routing protocol!)

- BGP is best seen as a negotiation protocol, attempting to find a point of equilibrium between networks' export and import polici
- Subtle changes in timers and sequencin ng means that the routing c... ...nistic.

Lesson:
Don't be OCD – any solution is still a solution!

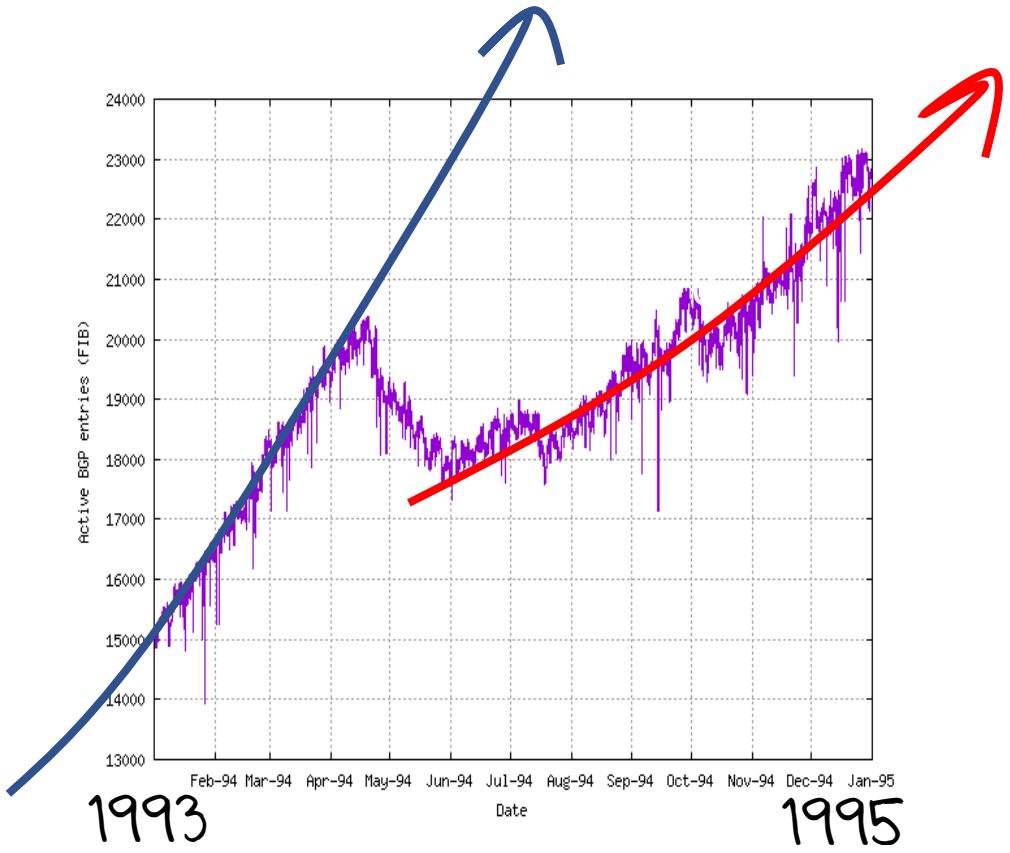
Why has BGP lasted?

- Don’t try to solve everything – underachieving can be a virtue!
- Reuse, don’t re-invent
- Don’t duplicate functionality
- Focus focus focus! Limit side-effects as much as possible
- Don’t make the protocol force the business model
- Don’t be OCD – any solution is still a solution!

II - BGP Deployment Experience

Containing the Routing "Explosion"

- IETF ROAD Efforts in 1992 (RFC1380)
  - Predicted exhaustion of IPv4 addresses and scaling explosion of inter-domain routing
- The chosen “solution” was to drop the concept of address classes from BGP
- It (sort of) worked for a while
  - Until it didn’t!



IPv6 and BGP

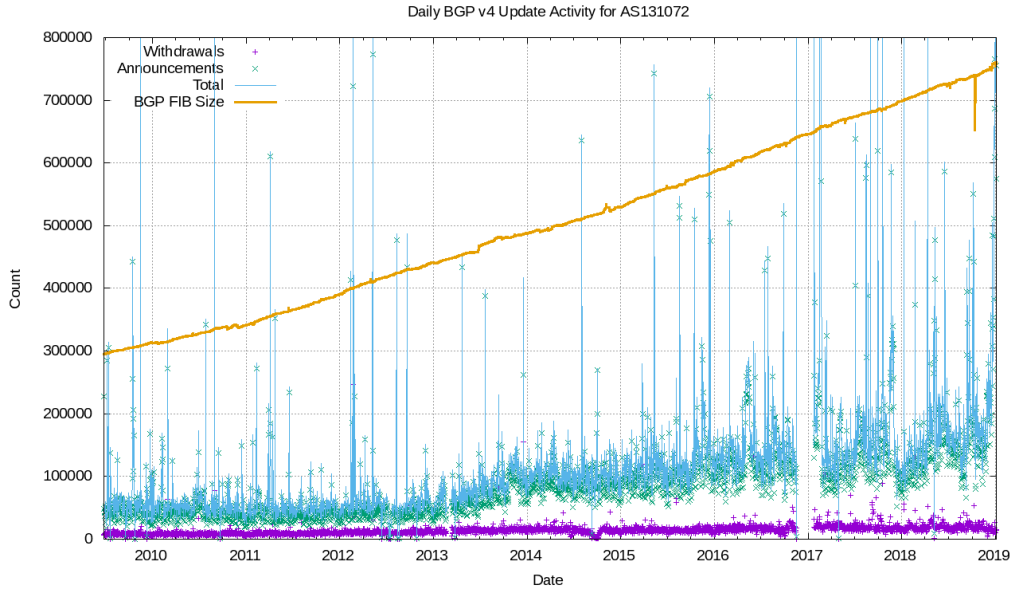
- While the IETF adopted the IPv6 address architecture for the address exhaustion issue, it was unable to find an IPv6 routing architecture that had similar scaling properties
  - IETF efforts to impose a routing hierarchy (TLAs and sub-TLAs – RFC 2928) got nowhere!
- So we just used BGP for IPv6 in the same way as we used BGP for IPv4
  - Address allocation policies that allocated ‘independent’ address blocks of /35 or larger
  - ISP traffic engineering and hijack “defence” by advertising more specifics

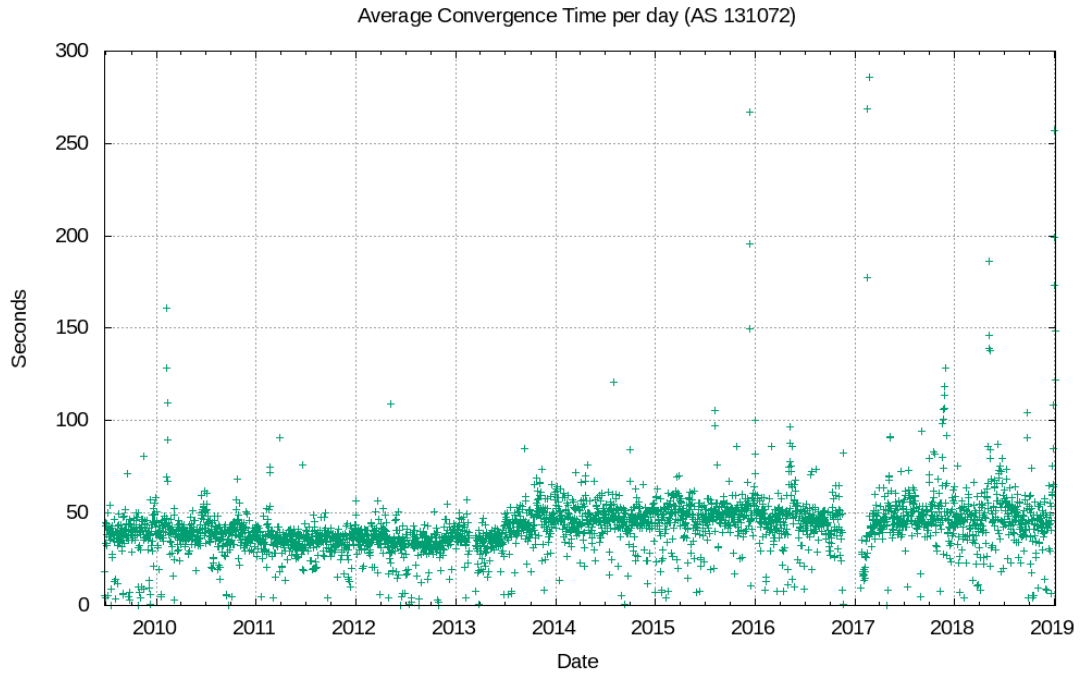
BGP and TE

- BGP cannot load-balance in the inter-AS space
  - It's a 'winner-take-all' best path selection protocol
  - It cannot load balance as it has no concept of feedback loops
- BGP cannot perform traffic engineering easily
  - Because routing policies are intrinsically non-transitive and AS prepending is completely unreliable, the only leverage left to engineer traffic is the selective advertisement of more specific routes
  - Which means that BGP carries large volumes of more specific routes whose primary purpose appears to relate to various efforts to perform traffic engineering of incoming traffic

BGP Scaling

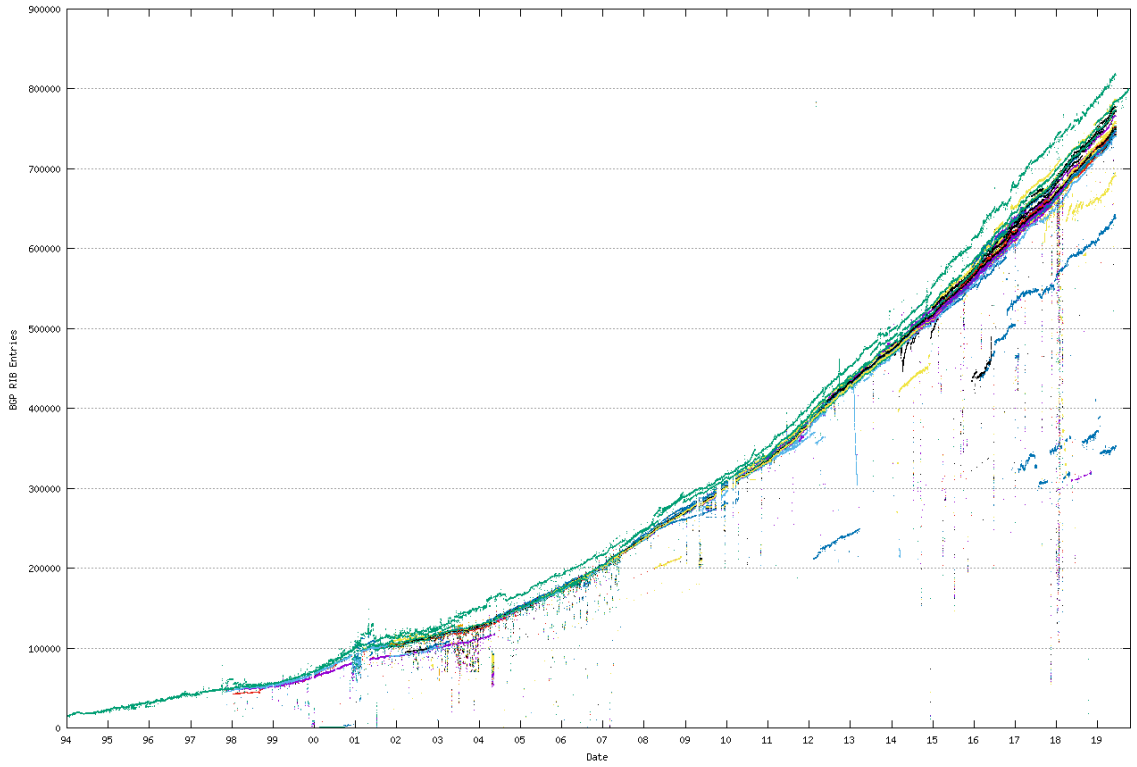
- BGP has scaled because the protocol only passes topology deltas - as long as the topology change rate is low, the BGP load is low
- The strongly clustered inter-AS topology of the Internet works in BGP's favour
- BGP has grown well beyond any original design expectations





But - Scale generates Inertia

- BGP-4 was introduced when the routing table contained ~ 10K entries – it is now ~800K entries and carries some 75K ASNs
- This has its own inertial mass that resists change
- Changing the routing environment to use a new IDR protocol would be incredibly challenging, even if we understood what we wanted from any candidate successor IDR protocol



Expectations vs Deployment

- Session lifetime
  - Expectations of short session lifetimes – experience of session longevity
- Session Security
  - Expectation of routing being a public function - experience of session attack
- Payload Integrity
  - Expectations of mutual trust – experience of malicious and negligent attack
- Protocol Performance
  - Expectations of slow performance – experience of more demanding environments
- Error Handling
  - Expectations of “clear session” as the universal solution – experience required better recovery without catastrophic session teardown
- Use
  - Expectations of simple topology maintenance – experience of complex traffic engineering

Deployment: BGP isn't perfect

- Session insecurity
- Payload insecurity
- Protocol instability
- Sparseness of signalling
- No ability to distinguish between topology maintenance, policy negotiation and traffic engineering

III - Where should we go with BGP?

Incremental tweaking?

Which as what we've been doing for 30 years:

- Capability negotiation
- Add Path
- Extended communities
- Fast BGP
- Graceful Restart
- 4-byte AS's
- …

Does tweaking "work"?

Not Really

- There are few BGP tweaks that provide substantial benefit to adopters in partial deployment scenarios in the Internet
  - Routing is a universal substrate and deviations from a common model are necessarily limited in scope and impact in order to interoperate with the common mass of behaviour
- As long as tweaks are localised in both impact and benefit they find it hard to gather sufficient impetus to impel common adoption
  - There are exceptions to this - like 4 byte ASN – but they are exceptions to the common behaviour model

Time for a "new" IDR?

What? Not again!

- We've been here before many times:

  "BGP is failing because <reasons> and we need to shift to a new IDR for the Internet"

- We have no new basic insights into routing in a diverse multi-provider space
  - Which means that we have no real assurance that we could improve on the basic BGP functions

"

Lessons from 30 years of BGP

- Enduring use is often an accidental and unintended outcome
- Simplicity is often undervalued
- Hop-by-Hop protocols are extremely flexible
- TCP is more powerful than anyone thought!
- Its by no means a perfect solution but it represents a set of compromises that we are willing to accept

What about the next 30 years?

I just don’t know!

What about the next 30 years?

I just don't know!

- There are major issues with content delivery systems and a major tension between carriage and content
  - In the multi-provider carriage environment BGP has a clear role to play for the near term future
  - In a future uni-provider content delivery system there are other approaches that can deliver better outcomes, incorporating feedback systems to support load balancing and adding fine-grained traffic steerage
  - So which way are we heading with the Internet?

What about the next 30 years?

I just don't know!

- There are major issues with content deli… tems and a major tension between carriage…
  - In the… clear role to play for the
  - …are other approaches …uting feedback systems to support …fine-grained traffic steerage
  - S… way are we heading with the Internet?

The entire internet may change and make BGP and IDR itself irrelevant! But that form of change is WAY more than just a discussion about BGP and routing!

Will it get better?

- Will we ever secure BGP?
- Will we clear out bogons?
- What about more specifics?
- Stop senseless prepending?
- See an end to massive route leaks?

Will it get better?

- Will we ever secure BGP?
- Will we clear o[illegible]gons?
- What ab[illegible]pecifics?
- Stop sens[illegible]ess prepending?
- See an end to massive route leaks?

My Opinions

We’re not going to change BGP anytime soon:

- It’s still functional
- We’ve grown used to working with its strengths and we’ve become accustomed to avoiding or tolerating its weaknesses
- Its adequately efficient
- The business model and the BGP model have managed to come to terms with each other
- The levels of abuse are tolerable (so far)
- And we’ve trained a large body of network operators who understand how to drive / abuse it for fun and profit!
- And we have no plan B!

Thanks!