The ISP Column
An occasional column on things Internet

April 2007

Geoff Huston

More ROAP – Routing and Addressing at IETF68

Over the past year or so we've seen a heightened level of interest in the topic of Internet routing and addressing. The continued intense examination of the IPv6 protocol and the associated speculation regarding the future role of the Internet raises the possibility of the Internet supporting a world of tens or hundreds of billions of chattering devices. What does such a future imply in terms of the core technologies of the Internet? Does what we use right now scale into such a possible tomorrow? Consideration of this topic has prompted a critical examination of aspects of the architecture of the Internet, including the scaling properties of routing systems the forms of interdependence between addressing plans and routing and the roles of addresses within the architecture. The IAB has been active in facilitating discussion of this topic, both in the IETF and in various Internet operational gatherings around the world. This IAB effort culminated in a 2 day workshop on routing and addressing in October 2006 to examine the characteristics of this space and start to identify some of the interdependencies that appear to exist here (the workshop report is close to completion, and there is also the author's informal report of impressions gained at the workshop).

IETF68 saw some further steps in analysing these issues, and during the week there was a plenary session on routing and addressing and meetings of the Internet and Routing Areas devoted to aspects of routing and addressing. This is a report of these sessions, and some conjecture as to what lies ahead along this path.

Plenary ROAP - The Plenary session on Routing and Addressing

The plenary session at IETF68 presented an overview of the topic, looking at the previous initiatives in routing and addressing as well as providing some perspectives on the current status of work in this area. Routing and addressing, in the context of the Internet has been visited on a number of occasions over the years, starting with the shift from the original 8/24 network and host part addressing to the Class A, B and C addressing structures, and the subsequent shift to the prefix-plus-length concepts of classless addressing. In the routing area there was the adoption of a peer model of routing with the introduction of BGP and the shift in BGP to support classless addressing in the form of CIDR. And, of course, there has been the design of IPv6. However, there still remains the concern that this is not completed work, and that the technology is not in an ideal state to scale by further orders of magnitude without further refinement. There are concerns over the scalability of routing, the 'transparency" of the network, renumbering issues, provider-based addressing and provider lock-in, service and traffic engineering and routing capabilities, to name but a few relevant issues that are relevant and challenging today, and appear to be even more so for the Internet of tomorrow.

Are there architectural principles that are relevant here? In the large, diverse but coupled set of networks that collectively define the Internet it appears that each component network should operate within a general principle of containment or insulation of impact. The principle is that each network should be able to implement reasonable choices in their local configuration without undue impact on the operation, or range of choices available to all other networks. In other words each network should be able to make such local configuration choices relatively independently of the choices made by any other network. The relevant issue here is balancing this principle against the operation of the network as a whole, which can be seen as a binding of networks together as a coherent entity, supporting consistent and robust communications paths through this collection of networks.

We do not use a routing technology that effectively isolates individual network elements from each other, or even manages to localize the external impacts of local choices. On the contrary, far from being a protocol that damps instability, BGP manages to be a highly effective amplifier of noise components of routing events. So while it is a remarkably useful information dissemination protocol with considerable flexibility, the properties of BGP in an ever-more connected world with ever-finer granularity of information raises some questions about its scaling properties. Will the imposed 'noise' of the protocol's behaviour completely swamp the underlying information content? Will we need to deploy significantly larger routers to support a much larger routing protocol load, but route across a network of much the same size as today's network? The prospect here that routing may become far less efficient because as we increase the degree of interconnection and the information load simultaneously the inability to insulate network elements from each other and inability to effectively localize information creates a disproportionately higher load in network routing.

As well as these observations about routing, there is the continuing suspicion that the semantic load of addresses in the Internet architecture, where an address conveys simultaneously the concepts of "who", "where" and "how" has some side-effects that cause complexity other aspects of the network, including

routing complexity. To what extent can the semantic intent of endpoint identity (or "id") be pulled apart from the semantic intent of network location and forwarding lookup token (or "loc") is a question of considerable interest. While the current IP address semantics removes the need to support an explicit mapping operation between identity and location, the cost lies in the inability to support an address plan that is cleanly aligned to network topology, and the inability to cleanly support functionality associated with device or network mobility. In the end it's the routing system that carries the consequent load here. The questions in this area include an evaluation of the extent to which identity can be separated from location, and the impact of such a measure on the operation of applications. How much of today's Internet architecture would be impacted by such a change, and what would be the resultant benefits if this were to be deployed? Would the benefits of such a deployment be realized directly by those actors who would be carrying the costs? Is deployment a complete and disruptive phase shift in the Internet, or are there mechanisms that support incremental deployment? Are we looking at one single model of such an id/loc split, or should we think about this in a more general manner with a number of potential id/loc splits?

As well as consideration of these general architectural principles and their application in routing and addressing there are also more specific sets of objectives that relate to Internet actors. For users there are objectives here about maximising the user's service and provider choices without cost escalation, and for service provides there are the objectives of using cost-effective technologies that can accommodate a broad diversity of both current and projected business needs, and well as the very real need to maximise the value of existing investments in plant and operational capability.

Behind this is the observation that the routing and addressing space is not infinitely flexible, and, on the contrary, form a highly constrained space. Part of the motivation behind the id/loc splits is to take some of the inflexibility of the id part of an address, where persistence is a key attribute and remove that from the locator part of an address. In split id/loc terms a mobile device is one that maintains a constant identity but changes locators. Multi-homing can be expressed in id/loc terms as a single identity simultaneously associated with 2 or more locators. Traffic engineering can be expressed in terms of locator attributes without reference to identifiers, and so on.

Obviously the study of this topic of routing and addressing, and the related aspects of name space attributes and mapping and binding properties is one with a very broad scope. The larger question posed here is whether this an issue where resolution can be deferred to a comfortably distant future, or whether we are seeing some of these issues impact on the network of the here and now. Are we accelerating towards some form of near term technical limit that will cause a significant disruptive event within the deployed Internet, and will volume-based networks economics hold or will bigger networks start to experience disproportionate cost bloat or worse? Is it time to become alarmed? Well there is the certainty of exhaustion of the unallocated IPv4 address pool in the coming years, but this sense of alarm in routing and addressing is more about whether there are real limits in the near future over the capability to continue to route the Internet within the deployed platform, using the current technologies, and working within current cost performance relationships irrespective of whether the addresses in the packet headers are 32 or 128 bits in size. There was a strong sense of "Don't Panic!" in the plenary presentation, with the relatively confident expectation that BGP will be able to carry the Internet's routing load over the next 3 -5 years without the need for major protocol surgery and that Moore's Law would continue to ensure that the capacity and speed of hardware would track the anticipated growth rates. There was the expectation that the current technologies and cost performance parameters would continue to prevail in this time frame. However, the subsequent plenary discussion exposed the viewpoint that such a prediction does not imply cause for complacency, and some sense of urgency is warranted given the criticality of this topic, the high level of uncertainty when looking at even near term growth prospects, and the ease with which this industry adopts a comprehensive state of denial over pending events, irrespective of their potential severity.

What we are up against as we consider these objectives as they relate to a future Internet is the relentless expansion of the network. Today the Internet sits in an order of size of dimension or around 10^9. There are some 1.6 x 109 routed addresses in the Internet and an estimate of between 10^8 and 10^9 attached devices. If we look out as far as four decades to around 2050 we may be looking at between 10^{11} to 10^{14} connected devices. (Yes, there's a large uncertainty factor in such projections!) Can we take the Internet along such a trajectory from where we are today? And if that's the objective, then how can we phrase our objectives over the next 5 years that are steps along this longer term path?

The immediate steps at the IESG level have been to take the IAB's initiative and work with a focus group, the Routing and Addressing Problem Directorate (ROAP), to refine the broad space into a number of more specific work areas, or "problem statements", and undertake a role of coordination and communication across the related IETF activities. In addition, as there is a relatively significant research agenda posed by such long term questions, the Routing Research Group of the IRTF has been rechartered and, judging by the participation at its most recent meeting just prior to IETF68, effectively reinvigorated to investigate various approaches to routing that take us well beyond tweaking the existing routing toolset.

Internet ROAP – The Internet Area meeting

The Internet Area meeting concentrated on aspects of this approach of supporting an identifier / locator split within the architecture of the Internet, and, specifically, at the internetworking layer of the protocol stack, and gathering some understanding as to whether this approach would assist with routing scaling. One of the key considerations in this area is working through what could be called boundary conditions of the study. For example is this purely a matter for protocol stacks within an endpoint, or are distributed approaches that have active elements within the network also part of the consideration? To what extent should a study consider mobility, traffic engineering, NATs and MTU behaviour? What appears to be clear at the outset is that this is not a 'clean slate' network, and any approach should be deployable on the existing infrastructure, use capability negotiation to trigger behaviours so that deployment can be incremental and piecemeal, allows existing applications and their identity referential models to operate with no changes, and, hopefully, have a direct benefit to those parties who decide to deploy the technology.

From the routing perspective the overall desire is to reduce the growth rates of the inter-domain routing space. The desired intent is to reduce the amount of information associated with locators so that locators reflect primarily network topology in such a way that the locators can be efficiently aggregated within the routing system that attempts to maintain a highly stable view of the network's topology.

The resultant system must be able to express, in routing terms, most of the flexibility we see in today's system, perhaps on a more ubiquitous scale. This includes site multi-homing across multiple providers, ease of provider switching and locator renumbering (assuming that locators may include some provider-based hierarchy), support for mobility, roaming and Traffic Engineering, and allowing for session resilience across various locator switch events. In and of itself these objectives form a challenging set, but it's not the complete set of objectives. In addition, it is necessary that these outcomes are achieved within tight cost constraints and volume economics that allow for scaling without disproportionate cost escalation, and, of course, such systems should be resilient to various known (and currently unknown) forms of hostile attack).

Today's system uses two critical mapping databases to support the discovery of the binding between identifiers and addresses. The Domain Name System (DNS) is used to map between a human-oriented name space used at the application level (domain names) and IP addresses, and the routing database in each router is used to map from addresses to particular local forwarding decisions (the forwarding mapping from the RIB to the FIB data structures). The current mapping system assumes stable endpoints with simple resource requirements and rudimentary security.

When we consider in further detail the implications of disambiguating aspects of identity from those of network location there are a number of dimensions to such a study, including the structure of the spaces, the mapping functions and the practicalities of any form of deployment of such a technology.

The first of these topics is the desired properties and structure of these distinct identification and locator spaces. Should the identity space be a 'flat' space of token values, or use some internal structure within the token that matches some distribution hierarchy? Is "identity" something that is embedded into a device at the point of manufacture (such as IEEE-48 MAC addresses), or at the point of deployment (such as Domain Names)? Is uniqueness a statistically likely outcome or one that is assured though the structure of the token space? Are there properties of the identity space that aid or hinder the security properties of the use functions in terms of mapping and referral operations? Is there necessarily one identifier space or potentially many such spaces? There are similar questions about a dedicated locator space, particularly relating to the time and space properties of locator tokens.

The next critical topic appears to be how an identity mapping function relates to the forwarding mapping function. Assuming that the existing name spaces remain unaltered, then the resultant framework appear to require distinct 'name' to 'identifier' mappings, 'identifier' to locator' mappings and a 'locator to forwarding' mappings. Where these mapping functions should be performed, who should perform these functions, when they should be performed and the duration of the validity of the outcomes, whether the mapping function outcomes are relative or universal, the scope and level of granularity in time and space of the map elements, the security of these mapping functions, and whether there is a simple operation in each mapping function or multiple operations all remain undefined at this point. There is also the issue of whether the mapping is explicit or implicit, and what evidence of a previous mapping operation is held in a packet in a visible manner, and what is occluded from further inspection once the mapping operation has been performed. What level of state is required in each host, and is there true end-to-end transparency and at what level? To illustrate some of the dimensions here, a particular approach to an identifier / locator split could see identifiers in the role of the end-to-end-tokens that are used by upper levels of the protocol stack, where identifiers are preserved in such a manner that both parties to an packet exchange use the same identifier pair for each transmitted packet, while locators would have a more elastic in intent and various identifier-to-locator and even locator-to-locator mappings could be performed while the packet is in transit. Another approach would take a more constrained view of locators and attempt to protect the initial locator value in such a way that any attempts to alter this value during transit would be detected and discarded by the receiver.

The other aspect to consider here is what one presentation termed the "Incentive structure", where it was advocated that the most effective incentives are those where local change is performed as a means of alleviating local 'pain'. This would indicate that routing scalability is predominately concern of service providers, whereas host mobility and service multi-homing and session resilience are matters of concern to the host and service provider and consumer. Its also useful in an incentive structure that benefit is realized unilaterally, in that one party's efforts at deployment provide local benefit to that party without regard to the actions of others, so that the problems of initial deployer penalties and lock-stepping are avoided.

It is likely, at least at this stage of the study, that there are a diversity of approaches to such a split, both in the intended roles of identifier and location tokens, and in their means of binding. Already in the HIP and SHIM6 approaches we've seen a difference of approach, where the SHIM6 approaches coopts locators as identifiers on a per-host-pair basis, while the HIP approach uses a persistent identity value that cannot assume the role of a locator. The expectation at this stage of the study is that further ideas will surface here and such ideas are helpful rather than distracting. It is unclear of a single solution can emerge from this activity, or whether different actors have a sufficiently different set of relative priorities that multiple approaches each of which express different prioritization of functionality are viable longer term outcomes.

The critical consideration here is that it is unlikely that scaling routing over the longer term to very much larger network is simply a matter of just changing the operation of the routing system itself. Real leverage in this area appears to also require an understanding of the meaning of the objects, or 'addresses' that are being passed within the routing system. The motivation for opening up the identifier / locator space within the Internet Area appear to be strongly tied to the notion that if you can unburden some of the roles of the addresses used in routing, and treat these routed tokens as unadorned network locality tokens, then you gain some additional capability in routing. The intended outcomes include being able to group 'equivalent' locators together and thereby reduce the number of elements being passed within the routing system, ensure that the locator set readily maps into local forwarding actions and also, hopefully, reduce the amount of dynamic change that is propagated in routing. It would also be useful if such an approach facilitates traffic engineering, site multi-homing, various forms of mobility and roaming. It might also be possible to remove from the application's end-to-end model the consideration of not just endpoint locality but also the tokens used in the transport protocol, proving a different approach to IPv4 and IPv6 interoperability.

At this juncture there is no unity or even clarity of the exact requirements, system design let alone solutions for this work. The exploration of the inter-dependencies of mapping functions, the properties of identity and locator spaces and the ways in which mapping functions can be supported in this environment is still at an early stage.

Routing ROAP – The Routing Area meeting

The last of these ROAP sessions in IETF68 was that of the Routing Area.

The first part of the Routing ROAP session looked at the trends in the routing system over 2005 and 2006. The overall trend appears to be a system that is increasingly densely interconnected carrying more information elements each of which expresses finer levels of granularity in reachability. As an example of some of the relativities here, it was reported that the amount of address space advertised in 2006 increased by 12% from January 2006 to December 2006, while the number of advertised AS's increased by 13% and the number of advertised prefixes increased by 17% over the same period. The report also looked at the dynamic behaviour of the routing space, looking at various distributions of the 90 million prefix updates that were recorded for the year. One of the major aspects of BGP updates in both 2005 and 2006 is the skewed distribution of updates, where, in 2006, 10% of the announced prefixes are the subject of 60% of the BGP updates and 60% of the announced prefixes generate just 10% of all updates. Looking at some known control prefixes it appears that BGP appears to be an effective noise amplifier, where a single origin event can generate up to 11 updates at the measurement point.

There appears to be two forms of dynamic BGP load: the BGP "supernova" that burst with an intense BGP update load over some weeks and then disappear, and "background radiation" generators that appear to be unstable at a steady update rate for months or even the entire year.

In looking at scaling the BGP routing environment it appears that one form of approach is to look in further detail at this subset of prefixes and AS's that are associated with the overall majority of BGP updates. One approach is to investigate whether damping of unstable prefixes in some fashion, or detecting routing instability that is an artefact of origination withdrawal, or deployment of propagation controls on advertisements would be effective in reducing the overall dynamic load of BGP updates. This approach represents a behavioural change in local instances of BGP that reduce the potential for unnecessary updates to be propagated beyond a "need-to-know-now" radius. Another approach is to consider changes to BGP in terms of additional attributes to BGP updates, such as "withdrawal-at-origin" flag, or selective advertisement of "next best path", both of which are intended to limit the span of advertised intermediate transitions while the BGP distance vector algorithm converges to a stable state.

Again, the considerations of deployment were noted, where the Internet's routing system is now a large system with considerable inertia. The implication is that any change to the routing system needs to use mechanisms that allow for piecemeal incremental deployment, and where incremental benefit is realized by those who deploy. One potential case study of such a change is the 4-Byte AS Number deployment.

It appears that we could improve our understanding of the operational profile of the routing space, particularly looking at the various forms of pathological routing behaviours and comparing these against the observations of known control points. Such a study may also lead to some more effective models of projections of the size of the routing space in the near and medium term future, and allow some level of quantification as to what "scaling of the routing space" actually implies.

The second part of the Routing ROAP session took a look at the current status of the routing world, updating some of the observations made at the IAB Routing Workshop and outlining some further perspectives on this space.

One critical perspective on BGP is the behaviour of BGP under load. BGP uses TCP as its transport protocol and this is a flow-controlled protocol, where the sender must await an advertisement of reception capability from the receiver (an advertised "window") before being able to send data. When this session is uncongested then a BGP speaker will send updates as fast as they are locally generated (depending on the Minimum Route Advertisement Interval (MRAI) timer). When the transmission is congested a local send buffer will form. Unlike conventional applications who treat TCP as a simple black box most deployed BGP implementations use state compression on the advertisement queues (as a simple example, the queuing of a withdrawal should remove any already queued but as yet unsent updates for this prefix). This state compression of the advertisement queue should be on a peer-by-peer basis, so that a congested BGP peer does not slow down an uncongested peer. The implication is that the load characteristics of BGP alter as the load level increases, and BGP attempts to ensure that its peer only receives the latest state information when the peer signals (via TCP flow control) that it is not keeping pace with the update rate.

Another critical factor is the nature of "convergence" in BGP. Convergence is at least an O(n) sized issue, where n is the number of discrete routing entries. This may appear daunting, but the real question is how important is convergence? The presentation included the claim that this was BGP's biggest, yet least important, problem. Convergence delays can be mitigated by graceful restart, non-stop routing, and fast re-route. One of the measures that exacerbates convergence is the use of Route Reflectors. Their model of information hiding is intended to reduce the number of BGP peer sessions and the update load, but what benefits they achieve they do so at the cost of slower convergence with a higher message rate during the intermediate state transitions. Perhaps it is appropriate to consider small scale changes to BGP behaviour to mitigate the transient BGP update bursts caused by path hunting, including those already mentioned of "withdrawal-at-origin" notification and propagation of backup paths.

One approach is to take the current set of potential tools that are proposed to addresses or mitigate various BGP pathologies and prune this set by looking at those that align cost and benefit in deployment, allow piecemeal incremental deployment, and have beneficial changes on the load properties of BGP.

The approach advocate here is based on the perspective that BGP is not in danger of imminent collapse, and there is still considerable "headroom" for BGP operation in today's Internet. This allows the IDR Working Group of the IETF to focus on measures that include tools and behaviours that tweak the current behaviour of BGP in ways that could mitigate some of the more excessive behaviours of BGP, and allows the Routing Research Group the latitude to study the broader topis of fundamental changes that may be associated with novel routing and addressing architectures.

More ROAP?

So is there some urgency here in looking at this problem? It's not clear that the problem is pressing, in that it is likely that the Internet will still be around tomorrow and probably the day after tomorrow as well. However, like many other issues where there are complex feedback loops here with internal amplification factors, so it may not be apparent that there is a near term problem with the health of the routing system until such time as the problems have already surfaced, and by then dire warnings of impending trouble are just too late! Also by that stage there is not the time to think about the various approaches to the space and the relative drawbacks and merits of each, as the pressure to simply deploy any measure to mitigate the issue is overwhelming.

The routing space is a classic example of the commons, where each party is at liberty to generate as many or as few routing entries as they see fit, and also free to adjust these entries as often as they see fit. This allows each party to use routing to solve a multitude of business issues, including, for example, using routing to perform load balancing of traffic over a set of transit providers, using a 'spot market' in Internet transit services, creating differentiated transit offerings using more specific routes and selective advertisements. The ultimate cost of these local efforts in optimising business outcomes through loading of the routing system is not necessarily a cost that is imposed back on the originating party. The ultimate cost lies in the increasing bloat in the routing system and the consequent escalation in costs across the entire network in supporting the routing system. There are no "routing police" nor is there a "routing market". There is no way to impose administrative controls on the global routing system, nor have we been able to devise a economic model of routing where the incremental costs of local routing decisions are visible to the originator as true economic costs for the business, and the benefit of a conservative and prudent use of the routing system reaps economic dividends in terms of relatively lower costs for the business. Like the commons there are no effective feedback mechanisms to impose constraint on actors in the routing space, and also like the commons there is the distinct risk that the cumulative effect of local actions in routing creates a situation that pushes the routing system, either as a whole or in various locales, into a non-functioning state.

It appears that there are a number of avenues of approach here in attempting to place some constraints on the potential expansion of the routing system. What is less than clear is the ultimate value of such approaches in the context of the future Internet. Is making a functionally richer endpoint protocol stack a course of action that sits comfortably within a world of communicating RFID labels? Is the lack of a routing market and an associated routing economy such a fundamental weakness that no technical efforts to alleviate the situation can gain traction in a world dominated by the desire to perform local optimizations in the cheapest possible

manner? Have we already constructed a massive multi-trillion dollar industry that now uses business models that assume particular routing behaviours, and would efforts to alter those behaviours simply founder because of trenchant resistance to change in the business models within the communications industry?

Whether it needs a sense of urgency to motivate the work, or a sense that there can and should be a better way to plan a future than crude crisis management, the underlying observation is that the routing and address world is fundamental to tomorrow's Internet. Unless we make a concerted effort to understand the various inter-dependencies and feedback systems that exist in the current environment, and understand the interdependences that exist between network behaviours and routing and addressing models, then I'm afraid that the true potential of the Internet will always lie within our vision, but frustratingly just beyond our grasp.

Yes, more ROAP please!

Further Reading

This is the set of references to further material on this topic, as presented in the plenary session.

- http://www.ietf.org/internet-drafts/draft-iab-raws-report-01.txt
- http://submission.apricot.net/chatter07/slides/future_of_routing/apia-future-routing-john-scudder.pdf
- http://submission.apricot.net/chatter07/slides/future_of_routing/apia-future-routing-jari-arkko.pdf
- http://www3.ietf.org/proceedings/07mar/slides/plenaryw-3.pdf
- http://www3.ietf.org/proceedings/07mar/agenda/intarea.txt
- http://www3.ietf.org/proceedings/07mar/agenda/rtgarea.txt
- http://www1.tools.ietf.org/group/irtf/trac/wiki/RRG
- http://www.ietf.org/IESG/content/radir.html

Disclaimer

The above views do not necessarily represent the views or positions of the Asia Pacific Network Information Centre, nor those of the Internet Society.

About the Author

GEOFF HUSTON holds a B.Sc. and a M.Sc. from the Australian National University. He has been closely involved with the development of the Internet for many years, particularly within Australia, where he was responsible for the initial build of the Internet within the Australian academic and research sector. He is author of a number of Internet-related books, and is currently the Chief Scientist at APNIC, the Regional Internet Registry serving the Asia Pacific region. He was a member of the Internet Architecture Board from 1999 until 2005, and served on the Board of the Internet Society from 1992 until 2001.

http://www.potaroo.net