Routing Arbiter in the post-NSFNET Service World

Bill Manning <bmanning@isi.edu>

Abstract

The United States National Science Foundation (NSF) has funded the ROUTING ARBITER[1] (RA) to provide stable, coherent routing in the Internet. With the Internet doubling every 13 months (according to some measurements) this is not as easy as it might be. The problem is compounded by the withdrawal of the NSFNET Service and the poliferation of Internet Service Providers and exchange points. A brief view from the RA perspective is given with some attention to tools and techniques that will facilitate the continued growth of the Internet in size, features, and function.

1 Introduction

On April 30th, 1995, the NSFNET Service was terminated, ending a nine year era of explosive growth for the Internet. By some measures, the Internet has doubled every 13 months and is showing no signs of slowing down. The latest figures from ISOC, TIC and NW, along with figures from the InterNIC back up this premise. Perhaps the single most stabilizing influence has been the NSFNET, with its Policy Routing database and "default-free" transit Service. This stability has not been without cost. The increasingly commercial Internet community has been concerned with the enforcement of the NSFNET Acceptable Use Policy and the resultant breaks in reachability.

These facilities have been replaced with commercial services for transit, exchange points or Network Access Points (NAPS) for peering, and the Routing Arbiter. The Routing Arbiter has as its charter the continued maintance of stable, unbiased global routing. To meet these tasks in the short term, the RA team has focused on a replacement for the PRDB, which is now known as the RADB. As an adjunct to the database, we have written tools that allow any Network or Internet Service Provider (ISP) to define and register their own routing policies. With a large number of providers and several exchange points the Internet community finds itself in a policy rich environment, with levels of complexity that did not exist in the NSFNET Service era. Use of these tools allows ISP policy expression to be codified as router configurations. In addition, the RA has deployed Route Servers at the NSF identified exchange points and other directed locations.

[1] The work described in this paper is supported by the NSF under the following agreements: NCR-9321043 and NCR-9321060. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not reflect the views of the National Science Foundation.

2 NSF-9352

The U.S. National Science Foundation, in winding down its support of what has been one of the better examples of technology transfer, recognized that it needed to focus on support of High Performance Computing and Communications. To this end, it released a soliciation [1] for a number of interlocking elements: a very high speed backbone to link its supercomputer centers, places where this backbone would be able to communicate with the rest of the Internet and its service providers, and an entity to facilitate stable, scaleable, global routing so the Internet can continue to grow.

2.1 The vBNS

The vBNS is a private backbone, originally specified to run at 155Mbps (OC3c) and connecting up the NSF supercomputer centers (Cornell, NCSA, SDSC, PSC, and UCAR). The NSF is retaining its acceptable use policy on this infrastructure, in that it is to be utilized for research and education use only. The supplier of the vBNS service is required to connect to the Internet at all of the exchange points specified by the NSF.

2.2 The Network Access Points

The NAPS are level 2 interconnect or exchange points. NSF awarded three priority NAPs and one non-priority NAP. The NAPs are located in New Jersey (Sprint), Washington DC (MFS), Chicago (Bellcore and AADS), and the San Francisco area (Bellcore and PACBELL). The NAP architectures are currently either a bridged FDDI/ethernet hybrid or an ATM(OC3/DS3)/FDDI hybrid. An additional exchange point is being constructed to support the other U.S. Federal internets access to the Internet at the NASA Ames facilities. The architecture of exchange points are being replicated around the global Internet, with exchange points in Europe and Japan. At each of these exchange points in the U.S., commercial and private use internets and ISPs touch down to exchange routing information and to transit traffic. A recent review of the exchange points has shown that the single 45Mbps NSFNET Service backbone has been replaced with as many as nine U.S. wide ISPs running 45Mbps backbones. Some studies [2] have indicated that with the increased load, the NAP fabrics as currently designed will not support the load offered by these ISPs.

2.3 The Routing Arbiter

The Routing Arbiter component has the charter to establish and maintain stable, unbiased, global routing and to advance the art and state of routing technology. Our initial efforts have gone into the NSFNET transition support and postioning to have the capabilities to support an increasingly rich environment for policy expression and interconnection. The architecture describing this phase of the RA activities is found in [3] and will be explored in the next section.

3 Routing Arbiter Elements

The RA, in its efforts to meet the requirements for stable, unbiased, and global routing have laid out the following architectural elements, which we believe will meet ISP needs and will support the growth in Internet services. The RADB, which is part of the total Internet Routing Registry, the Configuration and Analysis suite of tools, and the Route Servers form the implementation today. Other, less tangible activities are education, engineering, and research, so we can stay ahead or right on the growth curve.

3.1 The IRR and RADB

Internet Operations have become dependent on two types of registries, a delegation registry such as the InterNIC, RIPE/NCC, or APNIC, and a routing registry such as the PRDB or RIPE-81.

Delegation registries are tasked with the transfer of authority and responsibility to manage Internet Assets for the public good. They assign blocks of address space, AS numbers and DNS names. In doing so they track the points of delegation. Over the years they have discovered that it is no longer feasible to maintain a monolithic registration service and expect it to scale. A couple of examples will illustrate this;

* The migration from a flat host file to DNS.

* The use of distributed NICs by region [4].

* The deployment of the Rwhois Service [5].

For the last few years, the NSF policy routing database was authoritative for general Internet traffic transit. However, with the increase in the number of public exchange points, it is no longer feasible to presume that this registry would scale. The RIPE staff recognized this problem as the infrastructure grew richer in Europe and they created the first public routing registry description and software[6]. Experience with this initial release led to refinement. Refinement brought it to the point that it was deemed appropriate to try and reconcile discrepancies between the PRDB and the RIPE registry for production use. The results of these efforts have resulted in the RIPE-181 database and policy description. This release was stable enough that it was also released for general use in the Internet [7] and the code was widely distributed. ISPs with active registries based on the RIPE-181 code are the RIPE, the RA, MCI, CA*net, and others. At a meeting in the San Jose IETF, representatives from these groups met and agreed that the collective information represented in these databases would be referred to as the *Internet Routing Registry* (IRR) and to ensure that the information was replicated, they would exchange the information on a periodic basis.

It was from this unified base that the RA team selected its initial database. There were and are a series of problems related to the widespread use of the RIPE-181 registry. The problems we know of today are related to data duplication and directory synchronization. These are being addressed within the Routing Policy System working group [8] in the IETF. Until there is a resolution of these concerns, the RA team, in an effort to support unbiased access has adopted the view that the RADB is and can be considered a route repository of last resort. Anyone is free to register attributes within the RADB.

Once the base was selected, a thorough review was done of the database and the policy language to ensure that it could accurately and unambiguously represent the routing information and policies that were requested by ISPs. ISI was able to identify several inconsistencies with the RIPE-181 policy language and database [9] and has provided feedback to the community on changes that have been made in the RADB to support accurate representations of desired policies.

While this analysis was being undertaken, parallel efforts were proceeding to migrate the data from the old PRDB to the RADB. Perhaps the most difficult part of this effort was and is the on-going need to retrain people to use the new registration procedures and tools. Although the tools have a common heritage. [10] they must be tuned to a specific registry. ISPs must be aware of the subtle differences in tools between the PRIDE tools and the RPS updates to the PRIDE tools. It is important to note that for the RADB and the IRR in general, the intent is to place control of routing announcement in the direct hands of the Internet Service Providers and their clients. From a scaling perspective, a routing registry can no longer be run in a monolithic fashion with human intervention at every step. An added benefit is that with Internet users creating their own routing policies in the IRR there is less chance of bias or preferential treatment being injected by the RA or any operator of a component of the IRR.

Current directions on how to register in the RADB can be found in http://www.merit.edu/routing.arbiter/RA/RADB.tools.docs.html.

3.2 Configuration & Policy Analysis

Since registration in a routing registry is usually an extra step, ISI has provided ISPs with tools to provide them with direct operational advantage for the effort of registration. This advantage is in the form of auto-configuration tools which build router configurations. These tools are able to evaluate policy expressions based on the RIPE-181 or the proposed RPS formats and generate router configuration files based on the outcome of the evaluations. The RA team utilizes these tools to generate configurations for the fielded route servers. CA*net has built a port to generate cisco router configurations that they use internally. Merit used these tools to maintain the NSFnet

Service router configurations in the last few weeks of its life. We would like to thank ANS for the use of their network to test out yet another configuration file format. The RA team realizes that as needs change, this toolkit will need to be upgraded.

Current directions on how to get the RTconfig Toolkit can be found in http://info.ra.net/div7/ra/

The current interface to the RADB is through email. This constraint effectively limits the tools available today to essentially batch processing. The RA team recognizes that there are problems with this approach and realizes the need to have more interactive tools, such as a telnet interface to the RADB, as well as some what-if tools to allow an ISP the ability to explore reachability options before committing changes to the RADB. Such tools are being developed now.

In addition to the development of these new tools, the RA team has picked up the PRIDE tools and is porting them to support RIPE-181 and RPS formats.

3.3 RSd

The target for all these efforts is to support configuration of routers. To show proof of concept and to add value to the NAPs, the RA has deployed route servers at each NAP. Given that the traffic load on a NAP is expected to be high, and that ISP routers would best be able to use their memory and cycles forwarding packets, the route server code was designed to compute a unique, composite view of the Internet on a per-peer basis. This is a novel change in router design and use.

The end result is that the NAP fabric can be viewed as a router system bus, with the ISP routers as the interfaces and the route server and the forwarding table computation engine. This design could be modified [11] to incorporate the separation of the control channel (routing updates) from the data channel (packet forwarding). To do this would allow better tuning of required bandwidths needed by the exchange point parties. To achieve this design, the RS software was adapted from the GateD Consortium's Gated version 3.5; we made extensive modifications to Gated to support per-ISP routing tables. A number of releases have been made, with each successive release incorporating either functionality requested by service providers (e.g., correct handling of the Border Gateway Protocol (BGP) multi-exit-discriminator (MED) attribute, knobs to configure the insertion of the Route Server's Autonomous System (AS) number in advertised AS paths) or those requested by RA team members (e.g. binary dumps of the routing tables).

Assuming the Internet will continue to double every 13 months encourages us to ensure that the choices we make will be viable at least for the short term. ISI has rigorously derived Route Server behavior from a formal characterization of the behavior of BGP speaking routers. This work [12] analyzes the storage requirements of Route Servers and suggests ways in which these storage requirements may be reduced.

This work has also led us to a complete redesign of our Route Server software. The new design reduces Route Server storage requirements (by more than an order of magnitude in some cases) by trading off some processing for lesser storage. Since the resulting implementation is significantly different from Gated, and is designed and optimized for Route Servers specifically, we have labeled this software RSd (for Route Server daemon).

Work is also currently underway to design more efficient policy filtering in Route Servers. This is driven by the emergence of the need for more fine-grain policy; this need implies that policy filtering could become a dominant component of routing update processing in Route Servers. This improved design will be implemented in a future release of RSd.

Current directions on how to get the RSd software can be found in http://info.ra.net/div7/ra/

3.4 Operations & Management[2]

Placement of the route servers presents a number of interesting challenges. Since they are effectively stand-alone devices that may be unreachable, they have acquired many of the characteristics of intermittently visible devices.

The RA team has deployed custom software into each route server that collects performance statistics, delay matrix measurements, and throughput measurements. This software discovers the state and topology of the NAPs once a minute, and automatically configures itself and its ping daemon to monitor all peers and peering sessions. Alerts such as "peer not reachable" or "peering session down" are generated and stored in a Problem Table. Both the Network State Table and the Problem Table are externalized via a bilingual SNMPv1/v2 agent.

As far as the RA team is aware, this is the first deployment of SNMPv2 technology in an operational environment [13], and as such, a number of problems have been found. These problems have been conveyed to the appropriate IETF working group for discussion. Primarily, configuration of the security features of SNMPv2 have proven to be difficult.

The Route Servers have also been configured to collect and store NAP performance statistics as seen from the Route Server. These statistics include:

* Interface statistics.
  (in/out packets, in/out bytes, in/out errors)
* IP layer statistics.
* BGP layer statistics.

Other statistics collected include measurements of delay and packet loss characteristics between the RS

[2] Special thanks to Bill Norton and Jessica Yu for this information.

and its peers, and throughput performance between each RS and an RA data collection machine.

The RA team has had a number of requests for additional reports that the ISP community is interested in for routing analysis. The RA has begun collecting and formatting data to present these types of information to the Internet community:

* Frequency of route flaps.

* Total number of routes.

* Aggregation statistics.

* Stability of Route Server BGP sessions.

* Volume of BGP updates.

Other reports will note who is peering with the RS at various NAPs, how frequently the RADB is updated and configurations are run, and the stability of routing at the NAPs. The RA team will carefully consider privacy issues when making these reports available to the Internet community

3.5 Education & Engineering

The RA team is active in a number of forums where operational, research and administrative issues are discussed. There has been active participation in the routing designs at the hybrid NAPs and in outreach to the Internet Community. We encourage the formation and ISP participation in Operations forums like the North American Network Operations Group (NANOG). RA information is available from two servers:

http://www.merit.edu/routing.arbiter

http://info.ra.net/div7/ra

4.0 Futures & Research

The RA team is committed [14] to ensuring that the Internet continues to have stable, consistent, global routing with a goal for end to end reachability. In the current environment, we believe that the best way to reach this goal is through the use of the elements set forth above. Short term requirements that need work are:

* Improving the user interface for the RADB.

* Correcting distributed database problems.

* Releasing better what-if analysis tools.

* Speeding up configuration generation time.

* Adding support for increased NAP speeds.

* Improving Route Server asset utilization.

* Better tools for fine-grained policy filters.

* Protecting the Route Servers from attack.

Longer term, ISI and IBM are investigating the requirements for routing in and with IPv6. The RA team[3] has identified the following issues for concentration of our research focus in this area:

1) Detailed analysis of IDRP dynamics with diverse topologies and routing policies. As IDRP is deployed more widely any constraints on route selection and policy expression must be articulated. More over the routing registry provides a unique opportunity to detect configuration problems. Among other issues we will investigate a) the introduction of new route selection methods such as cisco's communities and symmetric-bilateral routing agreements, and b) the effect of addressing assignment practices. Particular attention will be paid to IDRP running in a Route Server (RS) supported context.

2) In conjunction with our IDRP analysis and development of the RS system, we will be pursuing work in the area of routing protocol testing and emulation. The target environment for IDRP/BGP is too large to begin to fabricate in a laboratory setting. Emulation is one technique for realizing reasonable scale experiments. Moreover, it allows more realistic investigation of protocol interaction than is usually achieved in simulations. Such emulation methods could become a critical part of protocol design, in addition to testing.

3) IDRP is a very flexible and extensible path vector routing protocol. Two extensibility issues in particular will require attention in the coming year. The first is how to practically deploy IDRP for IPv6. The second issue is how to use IDRP across ATM clouds. cisco has proposed introduction of a new NHRP routing algorithm. We will investigate the trade-offs associated with incorporating the desired functionality into IDRP, versus interoperating IDRP with NHRP and based on our conclusion focus the necessary design and implementation activities.

Amidst all the disagreements regarding routing architectures, most researchers agree that some form of explicit routing will be needed to accommodate heterogeneous routing demands, driven by both policy and quality of service. During the coming year we will be testing and refining our design of explicit route construction techniques. Three route construction mechanisms are under investigation. The first is RIFs, mechanism that uses IDRP queries with specified filters to obtain information from RIBs. The second is a path-explore mechanism to invoke constrained IDRP route computations. The final technique will be based on link-state style computations using the Routing Registry database. These explicit routes will be usable in conjunction with Source Demand Routing Protocol (SDRP), Explicit Routing Protocol (ERP), and PIM, however only the SDRP design is complete. Therefore, we will also complete our ongoing analysis of ERP and PIM-SM based on explicit routes.

3. Deborah Estrin and Yakov Rekhter, along with the other members of the RA team have provided valuable insights and direction on routing system protocol design and operations.

5.0 Author Information

Bill Manning
USC/ISI
4676 Admiralty Way
Marina del Rey, CA. 90292
USA
01.310.822.1511
bmanning@isi.edu

Bill Manning is currently with USC/ISI, working on the Routing Arbiter project.

References

[1] CISE/NCR, "NSF 93-52 - Network Access Point Manager, Routing Arbiter, Regional Network Providers, and Very High Speed Backbone Network Services Provider for the NSFnet and the NREN(tm) Program," Program Guideline nsf9352, May 1993.

[2] J. Scudder, and S. Hares, "NSFnet Traffic Projections for NAP Transition," NANOG Presentation, URL http://www.merit.edu/routing.arbiter/NANOG/Scudder-Hares.html, Oct. 1994

[3] D. Estrin, J. Postel, and Y. Rekhter, "Routing Arbiter Architecture," *ConneXions*, vol. 8, no. 8, Aug. 1994.

[4] RIPE NCC, "Delegated Internet Registry in Europe," RIPE-NCC F-2 Version 0.3, Mar. 1994.

[5] S. Williamson and M. Kosters. "Referral Whois Protocol," RFC 1714, URL http://www.isi.edu/div7/, Nov. 1994.

[6] J-M. Jouanigot et.al., "Policy based routing within RIPE," RIPE-060, URL ftp://ftp.ripe.net/ripe/docs/ripe-060.txt, May 1992.

[7] T. Bates, E. Gerich, L. Joncheray, J-M. Jouanigot, D.Karrenberg, M. Terpstra, & J. Yu, "Representation of IP Routing Policies in a Routing Registry (ripe-81++)," RFC 1786 URL http://www.isi.edu/div7/, Mar. 1995.

[8] rps-request@isi.edu

[9] C. Alaettinoglu, and J. Yu, "Autonmous System Path Expression Extension to RIPE-181," Technical Report, USC/ISI, Mar. 1995.

[10]T. Bates, M. Terpstra, D. Karrenberg, "The PRIDE project directory," URL ftp://ftp.ripe.net/pride/, Mar. 1994.

[11]P.Lothberg, "D-GIX design," Private communication, 1993

[12]Govindan, R., Alaettinoglu, C., Varadhan, K., and Estrin, D., "A Route Server Architecture for Inter-Domain Routing," USC Technical Report # 95-603, Jan. 1995.

[13]MERIT Network Inc., "Routing Arbiter for the NSFNET and the NREN, First Annual Report," Apr. 1995.

[14]USC/ISI, and IBM, "Routing Arbiter for the NSFNET and the NREN - Annual Report 1994," Apr. 1995.