Inter-Domain Dynamic Route Selection for Diversified IPv4 Networks

Jiva DeVoe (jiva@opnix.com) and Jay Jacobson (jay@opnix.com), Opnix, Inc.

Version 1.0, March 23, 2001.

Overview

The ARPANet was originally designed for high network reliability and resilience. Today's Internet, which evolved from ARPANet, is therefore is comprised of a multitude of diverse networks. As a result, information about inter-domain layer-3 routing on the Internet is decentralized. Individual networks are cognizant of their own and neighboring networks, but do not typically have detailed routing information for all networks comprising the Internet due to the volatility of routing information.

It is possible to make static routing decisions at any given point in time, but because of the dynamic nature of network performance characteristics, these routing decisions are inflexible and will likely be sub-optimal at various points in time.

This paper explores two possible methods of achieving dynamic inter-domain layer-3 route selection for diversified IPv4 networks. These methods are examined both in relation to static route implementation and to one another. The authors of this paper recognize there are many methods for determining network performance characteristics as well as methods for utilizing those characteristics to construct layer-3 routing information or to make layer-3 routing decisions. The authors welcome any comments about the items explored in this paper.

The Baseline – Static Layer-3 Routing

Static layer-3 routing implementations have no capability to dynamically address specific performance metrics of the Internet, such as:

- Packet Loss
- Latency Changes
- Throughput
- Layer-3 Hops
- Maximum Circuit Capacity
- Circuit Congestion
- Network Access Point (NAP) Congestion
- Historical Reliability
- Path Reachability
- Varying TCP/IP Characteristics

For the purpose of clarity, and standardization, these performance metrics are defined at the end of this paper.

Current inter-domain layer-3 routing protocols are incapable of automatically utilizing any of these metrics to decide on the best routes. Therefore, static routing fails to adequately address the inter-domain routing needs of diversified IPv4 networks.

The Standard Approach, BGP4

BGP4 attempts to address the inter-domain needs of diversified IPv4 networks via a cooperatively propagated decentralized route information base (RIB). This database consists of passively gathered information about connected networks. The primary purposes of the BGP4 protocol are network reachability, subnet aggregation, and elimination of inter-domain layer-3 routing loops, not performance optimization.

BGP4 – How It Works

BGP4 utilizes a set of Autonomous System (AS) numbers that are assigned to individual inter-domain networks or relatively large segments thereof. These AS numbers can be considered to be like political network boundaries within the Internet. An "AS Hop" is defined as a transition from one AS to another. The assumption made by BGP4 is that for any given path, the route with the least number of AS hops is preferable. In addition to this dynamic information, BGP4 allows administrators to define static path preferences using weights, local preferences, MEDs, etc. Utilizing this database of information, a layer-3 router is able to build a table of routes to describe how that router will make routing decisions. This table is populated with routes determined to be the "preferred routes" based first on the manually configured static preferences, and then on the dynamically propagated AS information.

(BGP4 – How It Works, Continued.)

These preferred routes from a BGP4 speaking layer-3 router's database are propagated through peering sessions with other BGP4 speaking layer-3 routers. For example, a router may receive dynamic BGP4 updates propagated from its BGP4 peers. It processes these updates, reevaluates its RIB, and then repropagates the updates to its other BGP4 peers.

BGP4 – Advantages

Compared to static routing, BGP4 addresses the need for dynamically updating route information for network reachability. As route information is updated from each AS, it propagates dynamically to all peers as configured. This allows each AS to be in control of its own routing policies, while also informing its AS peers about its preferred routes and peer network reachability. This improves the routing decision process by allowing a layer-3 router to choose routes based on BGP4's dynamically propagated network reachability information and/or manually configured static preferences. It is generally assumed by BGP4 that a path with fewer AS hops is the preferred route.

It is also important to recognize that varying vendor implementations of the BGP4 protocol are widely available for native operation on most modern layer-3 network devices. Thus, BGP4 is relatively easy to implement and is the default standard protocol for diverse inter-domain IPv4 routing.

BGP4 – Limitations

Unfortunately, BGP4 has no capability to discover any performance characteristics other than AS hops. BGP4 routing information, as related to network performance characteristics, is largely based on AS hops and manually configured static preferences. As a result, performance metrics such as packet loss, latency, throughput, layer-3 hops, circuit congestion, maximum circuit capacity, NAP congestion, historical reliability, and varying TCP/IP characteristics are not addressed by BGP4. BGP4 has no ability to actively discover any of these characteristics, and thus it has no ability to make routing decisions based on them. Therefore, a layer-3 router relying on BGP4 cannot make dynamic performance-optimized routing decisions.

An Alternate Approach, Opnix ORBIT

ORBIT attempts to address the needs of diversified IPv4 networks by building a centralized database of actively gathered performance metrics about Internet characteristics. This database is used by all client side ORBIT implementations to make dynamic performance-optimized routing decisions while still respecting network reachability. A table of preferred optimized routes is then transmitted to client routers.

ORBIT – How It Works

ORBIT consists of two main components: the ORBIT client premise equipment (CPE), which resides on a multi-homed IPv4 network (client network), and the Central Optimizing Route Engine (CORE) which resides at Opnix data centers. The two components work together to provide a system of Internet intelligence and dynamic Internet-wide routing knowledge.

(ORBIT – How It Works, Continued.)

The ORBIT CPE gathers Internet performance metrics from the perspective of the client network. ORBIT gathers these metrics through active probing of all available routes in IPv4 space, in parallel across all available peers, from as many ORBIT client network perspectives as possible. It sends this actively gathered data to the Opnix CORE for analysis, which computes a performance-optimized routing policy for the client network. The ORBIT CPE then announces the performance-optimized routes to the client network's layer-3 routers via a traditional BGP4 peering session.

The metrics currently gathered by the ORBIT system include packet loss, latency, NAP congestion, layer-3 hops, circuit congestion, and path reachability. Development is underway to add additional metrics such as throughput, historical reliability, maximum circuit capacity, and varying TCP/IP characteristics, among others.

The active probes generated by the ORBIT CPE consist of a small series of one-byte payload packets generated to the first available IP address on each known IPv4 route. The ORBIT CPE uses random high port numbers with increasing TTL values similar to the widely used network diagnostic tool *traceroute*. The result of these probes is ICMP_TTL_EXPIRED responses from each router in the path to the destination. When the destination is reached, the destination system responds with an ICMP_PORT_UNREACH error, indicating the host (destination) machine is not listening on the port to which the probe was sent.

The responses gathered thusly can be measured and the delta between the time sent and the time of the received response can be used to judge latency between each layer-3 hop in the path. Additionally, each responding router stamps the active probe packets with its own IP address, which can be used to determine the origin AS of the router by comparison against BGP4 propagated data. These IP addresses are also compared against a database of known NAPs, as determined by Internet research, to determine when the probe has traveled through a NAP. As the probes proceed, some packet loss may be measured as probes sent that do not have corresponding acknowledgements. This is also recorded for the purposes of determining reliability, and circuit congestion, as circuit congestion is often shown as packet loss.

When the current performance metrics for a configurable-sized portion of IPv4 routes are gathered, the ORBIT CPE compresses and encrypts the gathered data and sends it over a TCP connection through the Internet to the CORE. At this point, the ORBIT CPE proceeds with active probing on the next portion of IPv4 routes.

The CORE takes the data gathered by all remote ORBIT systems and combines that data to form a complete known view of the numerous Internet performance characteristics measured. It dissects the probe data from each ORBIT CPE into the smallest atomic components and merges it into the CORE's global collective.

For the purposes of analysis, the CORE considers each layer-3 hop in the gathered data to be an origin point, destination point, and transit point. The metrics reflect this by the difference in the data between these points. A complete path is represented as a series of these deltas resulting in a complete metric measurement. Using this methodology, metrics from numerous perspectives can be combined to create a cohesive whole. This also reduces the impact of abnormal performance characteristics specific to any one perspective since the measurements are based on deltas rather than any single source-to-destination relationship.

(ORBIT – How It Works, Continued.)

As the data is combined into the CORE's global collective, it is normalized using the following thresholds:

Metric	Lower	Upper
Layer-3 Hops	5	25
Latency	0.35ms	300ms
Packet Loss	0%	100%
AS Hops	1	10
NAP Hops	0	2

Applying the measured performance metrics with this table generates a normalized value for every known route. The normalizing process is done to be able to directly compare each metric. The normalized value for each metric is then multiplied by the following weighting values and added together:

- Packet Loss: 40%
- Latency: 30%
- Layer-3 Hops: 16%
- NAP Hops: 10%
- AS Hops: 4%

The resulting number is subtracted from 100 to yield an "OpScore" for any given IPv4 route. Each known route receives an OpScore, which is a weighted score for the normalized metric rankings for that particular IPv4 route.

After all the data is normalized and ranked, a complete "route view" can be built from any perspective or any point in the Internet, including from the perspective of the originating probes (client network). This route view can be used to create a performance-optimized route map of the Internet from a single ORBIT-connected network's perspective.

When the ORBIT CPE is ready to receive its optimized route map, it signals the CORE through a TCP connection. The CORE then builds a customized route map of optimized IPv4 routes specific to the perspective and configuration of the client network. This customized route map is encrypted and compressed by the CORE and sent over a TCP connection to the ORBIT CPE.

When the ORBIT CPE receives the optimized route map from the CORE, it creates new IPv4 routes in its internal BGP4 server. It sets the next-hop gateway appropriately according to the customized route map. Through the traditional BGP4 peering session with the client routers, this optimized route map is then propagated to the client routers. The client routers are configured to set a local preference for routes received from the ORBIT CPE, such that it will prefer to use the ORBIT system's routes over routes received from the routers' other peers.

Due to the highly volatile nature of Internet performance characteristics, collecting and computing metrics is a continuous and ongoing process.

ORBIT – Advantages

Compared to BGP4 for network performance, where BGP4 only takes into account AS hops and manually configured static preferences, the ORBIT system is more intelligent about network performance characteristics. The benefits of this system of route intelligence are that ORBIT routing decisions are based completely on actively collected performance metrics. This enables the ORBIT system to influence traffic from a client site through the best routes available to that client.

ORBIT – Limitations

It is difficult to establish a consistent and complete system of measuring performance metrics that can be agreed upon by a majority of the Internet community when discussing layer-3 routing characteristics. Therefore, the measurements of the ORBIT metric system may be subjective in its methodology. The client routers may experience some additional load as a result of the BGP4 interaction and updates with the ORBIT CPE, but the frequency of the BGP4 updates between the client routers and the ORBIT CPE is configurable.

Additionally, there are important measurement metrics that are still under development for the ORBIT system and may have a significant impact on layer-3 routing decisions. The active probes and updates done by the ORBIT system will generate a minimal amount of additional network traffic.

These factors will decrease on a per-site basis as ORBIT becomes more widely deployed. This is because a route through an identical next-hop gateway to a given destination only needs to be probed once for any given point in time. Thus, redundant probes from two independent ORBIT CPEs can be eliminated.

Interoperability Between BGP4 and Opnix ORBIT

ORBIT interfaces with BGP4-capable routers using an ordinary BGP4 peering session. In order to enable a router to utilize the routes sent to it by ORBIT, it is required that a local preference be configured on the router such that the routes from the ORBIT system are selected over routes received from the router's transit provider peers. In cases where the client network wishes to preserve a particular manually configured static routing policy over the routes that ORBIT is sending, the network administrator need only configure those policies to have a higher local preference than that of the ORBIT routes.

Once everything is configured, the ORBIT CPE will send a new set of performance-optimized routes to the client routers. These routes will look like any other routes from any other given peer, except that the next-hop gateway on these routes will be configured to be the preferred next-hop gateway according to the performance characteristic measurements and optimizations done by the ORBIT and CORE systems. Since the router has a local preference set for these routes, these routes will take precedence on the router and thus will influence the layer-3 routing of data from the client routers.

If the ORBIT CPE is shut down or fails, the client routers will fall back to the standard BGP4 routes received from the other peers.

Conclusion

This paper has examined two methods of achieving dynamic inter-domain routing for diversified IPv4 networks, with respect to reachability, performance optimization, and standards integration, in relation to a baseline of a static layer-3 routing implementation. Although the concepts of which network performance metrics are important, how to measure them at any given time, and how to apply that information to inter-domain layer-3 routing is not widely agreed upon by network operators, it is widely accepted that dynamic inter-domain route selection for diversified IPv4 networks is important for many factors, including network performance and reliability.

This paper has stated some of the reasons why static inter-domain layer-3 routing is neither optimal nor effective in today's network environments for attaining these requirements. This paper explored two possible methods of meeting these requirements: BGP4 and ORBIT. The positive and negative aspects of each were discussed. It is left to the reader to determine how this information and examination relates to his or her own network.

Definitions

Packet Loss

At any point in an end-to-end connection, as packets travel through networks, they will encounter varying network conditions including network failure, network congestion, or hardware instability. One result of these conditions is often packet loss, where either the data packet or its corresponding acknowledgement is dropped, discarded, or becomes corrupt in transit.

Latency

For the purposes of this paper, latency is the amount of time or delay between the sending of a data packet and the receipt of its corresponding acknowledgement packet at a layer-3 network device. The same varying network conditions that affect packet loss can also cause latency variations. Latency can also be introduced through a variety of other conditions including limitations on communication media, the speed of light, optical/electrical conversions, and protocol conversions.

Throughput

Throughput is the amount of the maximum usable bandwidth for any given end-to-end connectivity path at a point in time. Throughput is affected by a number of factors including roundtrip latency, packet loss, TCP/IP characteristics, and network/system utilization.

Layer-3 Hops

Each layer-3 device along an end-to-end connection that routes a packet at layer-3 (makes a routing decision; static, dynamic, or otherwise) introduces a layer-3 hop. Layer-3 hops introduce latency due to the decisions the layer-3 device has to make to move the packet to its destination. Generally, the more devices a packet passes through, there are more potential points of failure.

Maximum Circuit Capacity

Regardless of all other factors, each link in a connectivity path has a maximum circuit capacity. This is the maximum amount of data that can be continually sent across that link. When the maximum capacity is reached, negative performance characteristics can be introduced.

Circuit Congestion

Circuit congestion refers to an interval of time where data transiting a link, when combined with efficiency limitations and protocol/architecture overhead on that link, experiences negative

performance characteristics even if the theoretical maximum circuit capacity has not been reached.

NAP Congestion

For the purposes of this paper, a "Network Access Point" (NAP) is defined as a physical data exchange point where there is a large congregation of diverse networks, regardless of network architecture or protocols. NAPs could be problematic for data transit due to different possible circumstances such as legacy architecture, limited corporate/political cooperation, and overloaded capacity.

Historical Reliability

The statistical probability of a link failure at a point in time will vary based on the historical reliability of that link. Thus, observing the historical reliability of a link may help in determining the probability and the timing of a future failure or performance degradation.

Path Reachability

Similar to historical reliability, path reachability has to do with the state of a given link at an instantaneous point in time. If a link is shown to be down, then no data will be able to transit that link.

Varying TCP/IP Characteristics

Protocols using TCP/IP can be configured with varying parameters, such as segment size, window size, and the layer-2 maximum transfer unit (MTU) size. These parameters can have a significant impact on network performance characteristics.

Acknowledgements

The authors of this paper would like to thank the staff of Opnix, Inc., Darin Wayrynen, Mark Goldstein, Alan Gatlin, and Tom Coffeen for their invaluable assistance with the research and editing of this whitepaper.

References

RFC-1771 Y. Rekhter and T. Li, "A Border Gateway Protocol Version 4 (BGP-4)," March 1995.

RFC-1772 Y. Rekhter and P. Gross, "Application of the Border Gateway Protocol in the Internet," March 1995.

RFC-1773 P. Traina, "Experience with the BGP-4 protocol," March 1995.

Traceroute Thomas Kernen, http://www.traceroute.org

Oproute Jiva DeVoe and Opnix, Inc., http://oproute.net