Network Working Group                                          J. Yu
Request for Comments: 2791                      CoSine Communications
Category: Informational                                    July 2000


                   Scalable Routing Design Principles

Abstract

   Routing is essential to a network. Routing scalability is essential
   to a large network. When routing does not scale, there is a direct
   impact on the stability and performance of a network. Therefore,
   routing scalability is an important issue, especially for a large
   network. This document identifies major factors affecting routing
   scalability as well as basic principles of designing scalable routing
   for large networks.

Table of Contents

1. Introduction

   Routing is essential to a network. Without routing, packets cannot be
   delivered to desired destinations and the network would be non-
   functional. The challenge of designing the routing for a large
   network, such as a large ISP backbone network, is not only to make it
   work, but also to make it scale. Without a scalable routing system, a
   network may suffer from severe performance penalties, as
   unfortunately proven by disastrous events in large networks. This
   document attempts to analyze routing scalability issues and define a
   set of principles for designing scalable routing system for large
   networks.

   The organization of this document is as follows: Section 2 describes
   routing functions and design goals. Sections 3 and 4 discuss the

characteristics of today's large networks and the associated routing
scaling issues. Section 5 explores routing protocol scalability, and
Section 6 presents scalable routing design principles. Section 7
provides a conclusion to the document.

2. Common Routing Design Goals

   The basic goals a routing system should achieve are as follows:

      o Stability
      o Redundancy and robustness
      o Reasonable convergency time
      o Routing information integrity
      o Sensible and manageable routing policy

   The challenge of designing routing in a large network is not only to
   achieve these basic goals but also to make the routing system scale.

3. Characteristics of Today's Large Networks

   Today's large networks typically possess the following features:

      o They are composed of a large number of nodes (routers and/or
        switches), typically in the hundreds. Some provider networks
        include customer CPE routers within their administrative domain,
        which increases the number of nodes to thousands.

      o They have rich connectivity to meet redundancy and robustness
        requirements, and they consequently have complex topologies.

      o They are default-free; that is, they carry all the routes known
        to the entire Internet. Currently, the total number is
        approximately 70,000.

      o The customer aggregation routers inside the large networks
        connect sometimes hundreds of customer routers.

   These characteristics impose a direct challenge to the routing
   scalability of the network.

4. Routing Scaling Issues

   Today, the main issues surrounding routing scaling are: i) excessive
   router resource consumption, which can potentially increase routing
   convergency difficulties thus destabilize a network; and ii) routing
   complexity, resulting in poor management of network, producing low
   service quality.

4.1. Router Resource Consumption

   The routing process puts bursty loads on routers, especially under
   unstable network conditions. In the extreme case, the routing process
   takes all available resources from the routers, which results in slow
   routing convergence or no convergence. A network is paralyzed when it
   cannot converge internal routing information.

   It's worthy noting that routers with internal architectures that
   tightly couple forwarding and routing processes tend to handle the
   excessive routing load poorly. The emerging new generation of routers
   with the architecture of separating resource used for forwarding and
   routing could provide better routing scalability.

   Today, a large network typically employs IS-IS [1,2] or OSPF [3] as
   an Interior Routing Protocol(IGP) and BGP [4] as an Exterior Routing
   Protocol(EGP), respectively. The IGP calculates paths across the
   interior of the network. BGP facilitates routing exchange between
   routing domains, or Autonomous Systems (AS). BGP also processes and
   propagates external routing information within the network. The
   presence of a large number of routers and adjacencies in a network,
   coupled with frequent topology changes due to link instability, will
   contribute to excessive resource consumption by the interior routing.
   In the case of exterior routing, a large quantity of routers in a BGP
   system plus frequent routing updates (route flapping) would put a
   heavy burden on the routers. Section 5 describes scaling issues with
   IS-IS, OSPF and BGP in detail.

   In addition, having many destinations in a routing system, combined
   with multiple paths associated with these routes, impose the
   following scaling issues on BGP:

      o A large number of routes combined with multiple paths for each
        increases the cost of routing processing for route selection,
        routing policy application and filtering.

      o Too many routes combined with multiple paths requires large
        amounts of memory on routers for storage. The demand is even
        higher at InterExchange Points such as NAPs.

      o The larger the number of routes, the greater the chance route
        flapping will occur and the more BGP routing updates will happen
        as a result. Based on statistics collected by [5], thousands of
        BGP updates in a measured 15 minute interval can occur on a
        typical default-free router at a NAP.

Route flapping refers to frequent routing updates occurring due
to network instability, for example, when the state of a
physical link in the network is fluctuating, or when a BGP
session is torn down and re-established numerous time within a
short period of time.

To facilitate fast convergence, topology change information must
be propagated in a timely fashion. When a route becomes
unavailable and is withdrawn, the information is typically sent
immediately. If the affected routes have been announced to the
global Internet, the update information is likely to be
propagated to the entire Internet.

Route flapping has a profound impact on routers running BGP. The
routers have to process routing information frequently and this
consumes a tremendous amounts of the available resources. When a
local route or link is oscillating, interior routing is affected
as well by excessive topology information flooding and
subsequent shortest path calculations. However, OSPF (or IS-IS)
imposes rate limits on such activity to reduce the burden on the
routers. For example, OSPF specifies that an individual SLA can
be updated at most once every 5 seconds. This essentially
dampens the flapping.

Moreover, large numbers of E-BGP sessions processed by a single
router create another potential scaling issue. Large networks usually
have huge customer subscriptions and connections. To scale the
hardware and the number of nodes in the network, providers tend to
dedicate a group of customer aggregation routers, each connecting as
many customer CPE routers as possible. As a result, it's not uncommon
for a customer aggregation router to handle hundreds of E-BGP
sessions, which imposes potential problems, such as BGP session
processing and maintenance, route processing, filtering and route
storage.

4.2. Routing Complexity

Routing complexity can lead to network management difficulties, which
will have an impact on trouble shooting and quick problem resolution.
It can result in a less than desirable service quality across the
network. Complicated routing policies and special cases or exceptions
in a routing design can contribute to routing complexity in a large
system.

Routing Policy refers to the administrative criteria for handling
routing information, commonly in the form of routing path selection
and route filtering. The way routing information is handled has a
direct impact on traffic flow within a network and across domains. As

a result, it affects business agreements among different networks.
Therefore, the determination of routing policy is largely dominated
by non-technical concerns, such as business considerations. Routing
policy can be very complex, which would make management and
configuration an unscalable task.

The keys to reducing routing complexity are systematic as well as
consistent routing scheme and a routing policy that is simple but
meets the requirement of administrative polices.

Another factor contributing to the complexity of routing management
is prefix-based route filtering. As is well known, prefix-based
filtering is necessary in order to protect the integrity of the
routing system. This becomes a challenge when the number of routes
known to the Internet is as large as it is today.

5. Routing Protocol Scalability

   Today's commonly deployed routing protocols are IS-IS or OSPF for
   Interior routing (aka IGP) and BGP for exterior routing (aka EGP). In
   terms of scaling and other aspects, these protocols are already an
   improvement over the previous generation of protocols, such as RIP
   and EGP. However, scalability is still a major issue when a network
   is large, when a routing design is insensitive to scaling issues, or
   the protocol implementation is inefficient.

5.1. IS-IS and OSPF

   As described earlier in the document, IS-IS and OSPF are Link State
   routing protocols. The basic components of a link state routing
   protocol are i) generation and maintenance of a Link-State-DataBase
   (LSDB) that describes the routing topology of a given routing area;
   and ii) route calculation based on the topology information in the
   database. Each node in a routing area is responsible for describing
   its local routing topology in a Link State Advertisement or LSA (LSP
   in the case of IS-IS.) Each individually generated LSA will be
   distributed or flooded to all the routers in the area. Each router
   receives LSAs from all the other routers, forming a link-state-
   database that reflects the routing topology of the entire routing
   area.

   The main associated scaling issues are the complexity of the link
   state flooding and routing calculation, plus the size of the LSDB
   which contributes to the cost of routing calculation and router
   memory consumption.

Flooding is the process by which a router distributes its self-
originated LSA to the rest of the routers in the area in case of any
link state change. A router will send the LSA via all its interfaces.
When receiving an LSA update, a router validates the information and
updates its local LSDB before sending it out via all its own
interfaces, except the one from which it received the original LSA
update. Given the nature of IS-IS or OSPF flooding, a full-mesh
network with N routers would have $O(N^2)$ of LSAs flooded in the
network when a single link failure occurs. A single router outage
would cause LSA in the order of $O(N^3)$ to be flooded in the system.

In the case of OSPF, the protocol will refresh or flood every 30
minutes even under stable network conditions, which could increase
the problem for an already highly loaded router.

From the above discussion, one can easily observe that the more
routers and adjacencies in a Link State IGP routing area, the more
CPU burden there are for each router to bear. When a network is
unstable, the load will be amplified.

A link-state protocol typically uses Dijkstra's Shortest Path First
(SPF) algorithm for route calculation. The Dijkstra algorithm scales
to the order of $O(N^2)$, where N is the number of nodes. The algorithm
could be improved to the order of $O(l*logN)$ where l is the number of
links in the network and N is the number of destinations or routers
[6].

Consequently, link state routing protocols do not scale to a network
topology with many routers and excessive adjacencies in an area. When
the network topology is unstable, the computation, processing and
bandwidth costs are magnified, which causes excessive consumption of
router resources. When the instability prevents IS-IS or OSPF from
maintaining adjacencies, a network routing meltdown occurs.

Node adjacencies are discovered and maintained through the exchange
of HELLO messages sent periodically from each node. When a node fails
to receive HELLO messages from its neighbor within a certain period
of time (40 seconds for OSPF and less for IS-IS), it considers the
neighbor down. When heavy flooding, re-calculation and other
activities happen that make router CPU a scarce resource, a router
may not be able to allocate CPU time to send or process HELLO
packets. Routers in the network then lose adjacency, which magnifies
the instability. As a result, an isolated instability can escalate to
a routing failure across the entire network.

Link-state IGPs also do not scale well to carry a large number of
routes such as the 70,000 routes known to the Internet today. Since
external routes are included in the link-state-database and in LSA

(LSP for IS-IS) updates, the link bandwidth and router memory
consumption will be tremendous. Moreover, due to the large size of
LSA updates, it would aggravate router resource consumption in the
process of LSA flooding, especially under unstable network condition.

To summarize, a scalable design should avoid inclusion of too many
routers in an IGP routing area, a large external routes carried by
IGP and, more important, excessive adjacencies in the area.

5.2. BGP

BGP is an inter-domain routing protocol allowing the exchange of
routing or reachability information between different Autonomous-
System networks. Functionally, BGP is composed of External BGP(E-BGP)
and Internal BGP(I-BGP). E-BGP is used for exchanging external routes
while I-BGP is typically used for distributing externally learned
routes within an AS.

The general costs of BGP are as follows:

      o CPU consumption in BGP session establishment, route selection,
        routing information processing, and handling of routing updates

      o Router memory to install routes and multiple paths associated
        with the routes.

The major scaling issue associated with BGP lie in the full mesh I-
BGP connections. Since it does not scale for an IGP to carry
externally learned prefixes, as mentioned in the previous section,
I-BGP assumes this duty. In order to prevent routing loops, prefixes
learned via I-BGP are prohibited from being advertised to another I-
BGP speaker. As a result, a full mesh of I-BGP sessions among the
routers within an AS is required. In an AS with N routers, each
router will have to establish I-BGP sessions with N-1 routers, and
the system complexity is in the order of $O(N^2)$. Therefore, BGP
scales poorly when the number of routers involved in I-BGP mesh is
large.

A large network normally learns all the routes known to the Internet,
which is approximately 70,000. I-BGP will need to carry all these
routes.

The large number of I-BGP sessions and routes consumes tremendous
resources from each router, especially during BGP session
establishment and during periods of heavy route flapping.

Frequent routing updates are another potential scaling problem in
large networks. BGP uses incremental updates and sends out routing
information about unreachable routes quickly for fast convergence.
This is a great improvement from EGP, in which the whole routing
table is updated at a fixed time interval. However, when a network is
unstable the updates, especially those containing route withdrawals,
are sent immediately, causing global BGP updates. As a result,
network instability initiated anywhere in a network triggers updates
all over the Internet. This effect is magnified when large amounts of
routes are visible to the Internet, putting a heavy load on routers
that participate in BGP.

The introduction of a routing hierarchy in BGP, through I-BGP Route
Reflectors [7] and BGP Confederations [8], for example, will help
alleviate the scaling problem caused by the requirement of full mesh
I-BGP establishment.

Another potential solution is to avoid the requirement of full mesh
pairwise I-BGP connections. This will change the way that BGP
distributes routing information among the I-BGP peers. Mechanisms
worth considering are using multicast to distribute information or
adopting flooding mechanisms similar to those used in IS-IS or OSPF.
Further investigation of the implication of using such mechanism for
BGP route distribution is needed.

Route dampening [9] is one way to reduce excessive updates triggered
by route flapping. The trade-off between fast convergence and
stability of the network should be considered, as discussed in
section 6.3.

6. Scalable Routing Design Principles

The routing design for a large-scale network should achieve the basic
goals of accuracy, stability, redundancy and convergence as described
in Section 2 and moreover should achieve it in a scalable fashion.

How routing scales is influenced by protocol design decisions,
protocol implementation decisions, and network design decisions. A
network engineer has direct control over network design decisions and
can have substantial influence over protocol design and
implementation. The focus of this document is network design
decisions.

Following is a set of design principles for making a large network
routing system more scalable:

    o Building hierarchy
    o Compartmentalization
    o Making proper trade-offs
    o Reducing route processing burdens
    o Defining scalable routing policies and implementation
    o Utilizing out-of-band routing assistance

6.1. Building Hierarchy

As discussed in Section 5.1, OSPF and IS-IS scale poorly when a
network has a large number of routers and in particular, a large
quantity of adjacencies. This has unfortunately been proven by
networks that deploy IP over ATM with full mesh adjacencies among the
routers. The full mesh overlay design combined with the inefficient
protocol implementation led to disastrous network outages. A lesson
learned from this is to avoid full mesh overlay topology in a large
network with a large, flat network routing structure.

Building hierarchical routing structures in the network is the key to
achieving routing scalability in a large network. As discussed
earlier in this document, large networks are usually composed of many
routers with a complex topology, which results in a large number of
adjacencies. As also discussed earlier, currently available routing
protocols scale poorly for handling a large number of routers in a
routing domain or many adjacencies among the routers. Therefore, it
is sensible to build a routing hierarchy to reduce the number of
routers as well as the number of adjacencies in a routing domain.

The current common practice is to build a two-tiered hierarchy in a
network with a center component (or transit core network) to which a
number of outskirt components (or access networks) attach. The
transit core network covers the entire geographical area the network
serves; each access network (aka regional network) covers one region.
There are usually no direct link connections among the regional
components. Traffic from one regional network to another traverses
the transit core. Customer networks connect only to access or
regional networks. There are a number of ways to build a routing
hierarchy in the above described hierarchical network topology.

    1) Completely Separate Routing Domains

    This design treats the transit core network and each regional
    network as completely independent ASs with respect to routing, and
    each AS runs an independent IGP. Each regional network E-BGP with
    the transit core for exchanging routing knowledge. Full I-BGP

connections need to be established only within each component
network. With this design, the maximum number of routers in an IGP
domain is the total number of routers in each component. As a
result, the IGP processing load is reduced, and the number of
routers in an I-BGP mesh in the network routing system is
decreased dramatically.

Another advantage of this design is that it compartmentalizes the
routing system so that instability in one such component has less
impact on the entire system. See the discussion in section 6.2.

The main disadvantage of this scheme is that it inserts one extra
AS in the routing path when routes are advertised to the Internet
via BGP. This extra AS in the path may cause route selection
difficulties for other providers.

2) One Domain with IGP and BGP Hierarchy

This method includes the transit core and each regional network
into one AS domain. The routing hierarchy is realized by utilizing
multi-level IS-IS or OSPF areas and either BGP Confederation or
I-BGP Reflector or a combination of the two.

This mechanism avoids the introduction of an extra AS in the
routing path, which is an advantage over the method described in
Point 1).  However, multi-area hierarchical IGP is rarely used
now-a-days in large networks since most of them are using IS-IS
for internal routing, which does not have sufficient multi-level
support. Although IS-IS supports multi-area routing, it imposes a
strict hierarchy between backbone and sub-areas and allows only
the advertisement of a default route from the backbone area to the
sub-areas instead of specific prefixes. This restriction may be
suitable for a network with a simple sub-area topology. A sub-area
in a large network, typically a regional or access network, itself
has a complicated topology. Receiving highly abstract routing
information, such as a default route, would affect the sub-area's
ability to make route selections required for traffic engineering.
It would also limit the information passed to external ASs, for
example, IGP-derived BGP Multi-Exit-Discriminator (MED)
information.

Efforts are being made to modify the IS-IS protocol to allow the
distribution of specific route from backbone area to sub-areas. A
mechanism facilitates such distribution is specified in [15]. When
implementation of such mechanism become available, implementing
multi-level IGP will be an attractive option for building routing
hierarchy within a large network.

   3) One IGP Area with BGP Hierarchy

   In lieu of multi-area IS-IS, the routing hierarchy could be
   achieved by defining one IGP domain for the entire network while
   employing a BGP hierarchy. Fortunately, the hierarchical topology
   of the network in this case helps reduce adjacencies in the
   routing domain (recall there are no connections among the second-
   level network components). In addition, improvements could be made
   to further reduce the adjacency by carefully arranging the
   adjacencies to keep them at a minimum but still achieve good
   redundancy. However, this is less than ideal since the number of
   routers remains unchanged, which increases the load on the SPF
   calculation. Moreover, instability within any regional network
   would still affect the entire network (that is, there would be no
   fault isolation).

   Even with one IGP domain, it is possible to build BGP hierarchy to
   make I-BGP more scalable in the network. BGP Reflectors and BGP
   Confederations are existing mechanisms to address the scaling
   problem of full-mesh I-BGP.

   Further, a BGP reflector provides the ability to build more than
   two levels of hierarchy, as long as the interactions among the
   different levels of the hierarchy are carefully arranged to avoid
   the possibility of creating routing loops.

Questions worth asking are: "Are two levels of routing hierarchy
sufficient for handling scaling issues?" "Is there really a need for
more than two levels of hierarchy?"

When a second-tier sub-domain of a large network, such as a regional
network, grows too big for routing protocols to handle, either
another layer of hierarchy needs to be introduced or the sub-domain
needs to be split into multiple second-tiered sub-domains.

Keeping two levels of hierarchy and adding more sub-domains appears
to be more manageable than adding another level to the hierarchy.
However, one concern is to avoid adding more nodes to the top-level
or transit core network to make it less scalable. Connecting the
split sub-areas to the same core router would eliminate the need to
add more nodes in the core area than is recommended.

Having more than two levels of hierarchy would exceed the capability
of IGPs as they are defined today. In OSPF, for example, all the
areas must be connected via the backbone area, which eliminates the
possibility of having more than two levels of hierarchy. IS-IS has
the same limitation. Therefore, the protocols need to be redefined
should more than two hierarchical layers in IGP be desirable.

The complexity of protocols and management will increase with the
number of levels added to the hierarchy. According to [6], most of
the OSPF protocol bugs found over the years are related to routing
area support. Because the interaction among the multiple levels
increases management and debugging complexity, it is desirable to
keep the levels within a hierarchy to a minimum.

6.2. Compartmentalization

A scalable routing design of a large network should be able to
localize problems or failures, thus preventing them from spreading to
the entire network, consuming resources of network routers, and
causing network wide instability. This is compartmentalization.
Network compartmentalization makes fault isolation possible which
contributes the stability of a large network.

To achieve compartmentalization in routing design for a large
network, one needs to avoid a design where the whole large network is
one flat routing system or routing domain. This is the reason for the
architecture of dividing interior and exterior routing in the global
routing system. Within a network, it is best to divide the network
into multiple routing domains or multiple routing areas. For example,
in OSPF, only summary route SLAs, rather than individual area routes,
are flooded beyond the area. When an area border router aggregates
the routes in its sub-area, instability of any route included in the
summary route would not cause flooding of SLAs to other areas. As a
result, router resources in other areas would not be consumed for
handling flooding and the SPF recalculation. In other words,
instability within each individual area would be prevented from
spreading to the entire routing domain.

Since building a routing hierarchy essentially divides a big routing
area into smaller areas or domains, it help achieve the goal of
compartmentalization.

6.3. Making Proper Trade-offs

When designing routing for a large network, the overall goal should
be set with considerations of routing scalability and stability. The
trade-offs between conflicting goals should be taken into account.
Examples of such trade-offs are redundancy vs. scalability and
convergence vs. stability.

Redundancy introduces complexity and increased adjacencies to the
network topology. Redundancy also imposes the need for as many
alternative paths as possible for each route, which increases route

processing and storage burdens. Because of these problems, it may be
necessary to sacrifice absolute redundancy in favor of a reasonable
level that scales better for the routing system.

Fast convergence requires that changes in network topology be
propagated to the network as quickly as possible. Such action
increases routing updates and, consequently, the route processing
burden. The burden is aggravated when a network carries full Internet
routing information, as large networks usually do, and topology
changes happen frequently. Route dampening may be necessary to
achieve stability at the expense of absolute fast convergence.

6.4. Reduce Burdens of Routing Information Processing

The tasks of reducing routing processing burdens includes: i)
strategically place the routing intelligence within the network, ii)
avoid carrying unnecessary routing information and iii) reduce the
impact of route flapping.

6.4.1. Routing Intelligence Placement

A router that executes routing policies, performs route filtering and
dampening is said to posses routing intelligence. Routing
intelligence is needed for a network i) to enforce the business
agreement between network entities in the form of routing policies;
ii) to protect the integrity of the routing information within the
network and sometimes iii) to shield a network from instability
happening elsewhere in the Internet.

The more routing intelligence a router has, the more resources of the
router are needed to perform those tasks. It is logical, then, to
place as little routing intelligence as possible on routers that
already are heavily burdened with other tasks.

Usually, traffic is heavily concentrated in the core of the network.
Because traffic aggregates from the edge of the network toward the
core, traffic is less concentrated near the edge of the network.
Consequently, to build a scalable routing system, it is wise to place
routing intelligence at the edge of the network, especially in the
networks deployed with routers that do not sufficiently decouple
forwarding and routing. In addition, pushing routing intelligency as
close to the edge of the network as possible also serves the purpose
of distributing computational and configuration burdens across all
routers.

It is also desirable to move the heavy burden of processing routes to
out-of-band processors, freeing more resources in network routers for
packet forwarding and handling.

6.4.2. Reduce Routes and Routing Information

   As discussed in Section 4.1, a large number of routes in the system
   is one of the major culprits in route scaling problems. Therefore, it
   is best to reduce the number of routes in the system without losing
   necessary routing information.

6.4.2.1. CIDR and Route Aggregation

   CIDR as specified in [10] provides a mechanism to aggregate routes
   for efficiently utilizing IP address space as well as reducing the
   number of routes in the global routing table. CIDR offers a way to
   summarize routing information, which is one of the keys for routing
   scalability in today's Internet.

   Route aggregation would not only help global Internet scalability but
   would also contribute to scalability in local networks. The overall
   goal is to keep the routes in the backbone to a minimum.

   To achieve better aggregation within the network; that is, to reduce
   the number of routes in the network, a block of consecutive IP
   addresses should be allocated to each access or regional network so
   that when a regional network announces its routes to the transit core
   network, they can be aggregated. This way, the core and other
   regional networks would not need to know the specific prefixes of any
   particular access network. Although assignment of customer addresses
   from a provider block would have to be planned to support
   aggregation, the effort would be worthwhile.

6.4.2.2. Utilize Default Routing When Possible

   The use of a default route achieves ultimate route summarization,
   which reduces routing information to minimum. Route summarization
   also masks the instability associated with an individual route, for
   example, in the case of route flapping. It's beneficial for a network
   to utilize default routing when appropriate. For example, if a
   second-tiered regional network is a stub and there is no connected
   customer requesting full Internet routing information, the regional
   network can simply point default to its connected core network.
   However, over-summarization of routing information has the danger of
   losing routing granularity and as a result, management of network
   such as traffic engineering would be adversely affected. Therefore,
   caution needs to be exercised when using default routing.

6.4.2.3. Reduce Alternative Paths

   Due to the requirement of reliability, the connectivity in the
   Internet is rich, resulting in many paths toward a particular
   destination. In other words, there are many alternate paths in the
   BGP routing table towards the same destination, which consumes router
   memory and adds to the routing processing burden.

   To make routing scale, it is desirable to reduce alternate paths
   while preserving reasonable redundancy. For example, on a given
   border router (such as a NAP router), one primary path plus an
   alternate path should provide reasonable redundancy. In this case, a
   third or a fourth alternate route could be discarded for the sake of
   scaling.  This is a trade-off decision every network administrator
   needs to make based on the particular needs of her network.

6.4.3. Use Static Route at Edges

   As mentioned earlier, one of the scaling issues in large networks is
   that a single router may fan out to hundreds of customer routers. As
   a result, resource consumption will be very intensive if all the
   customer routers communicate via BGP with the edge router. Is it
   necessary for the edge router to BGP with all of its attached
   customer routers?

   At first glance, it seems necessary for a customer network in a
   different Autonomous System(AS) to exchange routing information with
   the provider network via BGP. However, this is not necessarily the
   case. When a customer network is single-homed (that is, if the sole
   network connection for a customer is via its provider network), BGP
   is not necessary and static routing can work. Since the customer
   network is single-homed, static routing will not have any negative
   impact on services. The advantages are that the customer aggregation
   router will have fewer E-BGP sessions to handle, and no route
   flapping can result from the statically configured customer routes.

   Configuration of the customer's static routes on the provider's
   aggregation router may add management overhead, especially if a
   customer advertises a large number of routes. On the other hand, the
   set of routes a customer announces to the provider usually changes
   infrequently; thus it requires low maintenance once it is configured.

6.4.4. Minimize the Impact of Route Flapping

   As discussed earlier, route flapping is largely caused by link
   instability and/or BGP session instability that results in excessive
   routing updates across the Internet. Route flapping can originate
   anywhere in the global Internet and affect every network in the

Internet routing mesh (BGP mesh). Given that there are over 70,000
routes known to the Internet and there is little isolation for route
flapping, handling route flapping could be overwhelming to routers in
any network.

One way to reduce the effect of route flapping is to turn on route
dampening as specified in [10]. Essentially, dampening suppresses an
unstable route until it becomes stable. The current practice is for
each ISP to enable route dampening on its border routers. This way,
excessive routing updates can be stopped at the border.

An ideal model is to suppress the announcement of a flapping route
right at the source. One way to implement this is to have a router
recognize instability associated with its directly connected links
and suppress the announcement of the route. So far, there is no such
implementation. This approach should be explored.

Route aggregation often masks route flapping since components of an
aggregated route (more specific routes) would not cause the
aggregated route to flap. Therefore using CIDR can also help to
alleviate route flapping.

6.5. Scalable Routing Policy and Scalable Implementation

Routing policy involves routing decisions about acceptance and
advertisement of certain routes to or from other networks and about
routing preference when more than one route becomes available.
Routing policy enforces business agreements between network entities
and is largely governed by non-technical criteria. In essence,
routing policy involves defining criteria for route filtering and
route selection.

One aspect of route filtering has to do with traffic control between
routing domains or between different provider networks. Making policy
based on individual prefixes should be avoided in this case because,
with the large number of prefixes in the Internet, it does not scale.
Making policy based on ASs that administratively represent a set of
prefixes scales better.

Another purpose of route filtering is to protect the integrity of
routing information by preventing the acceptance of falsely
advertised routing information that would lead traffic to 'black
holes'. In this case, only prefix-based filtering will sufficiently
achieve the goal. Prefix-based filtering needs to occur at the
borders between a network and its direct customers or peer networks.
The filtering is harder to manage at the boundary of the peer
networks since a peer network usually advertises a large amount of
prefixes. As mentioned earlier, there are about 70,000 routes known

to the Internet. This means a large default-free network would need
to filter on the order of hundred of thousands of prefixes or even
more since a route could be advertised by more than one sources. The
sheer amount of the prefixes to be filtered imposes challenges for
router configuration memory and configuration management. To make it
scale, one would need to rely on the help from an out-of-band process
to sort out which prefixes should be accepted or denied from which
source. IRR [11] and DNS [12] are among the current proposed
mechanisms for implementing prefix-based filtering.

Route selection policy determines which path should be used to send
traffic toward a certain destination. This is important, for example,
when a network has two connections to another network and learns
routes from both connections. The decision involves which path to
select to send traffic to the customers behind the other network. The
choices are typically:

  o Directing traffic to the closest interconnection point for
    traffic to exit the network. This policy is also known as Hot-
    Potato-Routing

  o Directing traffic to the optimal network exit point. The optimal
    exit point is determined based on certain criteria by the
    network administrator and is not necessary the closest exit
    point

  o Always preferring routes advertised by directly connected
    customers

  o Allowing other network or customer to determine the path

When a policy is defined, its implications for scalable
implementation need to be considered. For example, if the policy
allows customers to determine which paths traffic follows, customers,
not the provider, should be required to set routing parameters to
make the routing favor their preferred path. Customers can use the
BGP community or mechanisms such as MED to set routing preferences in
a much more scalable way. This avoids putting such routing management
burdens solely on the provider. Distributing the routing management
burden makes the policy implementation more scalable.

Another scaling measure is to avoid making complex policy. When
routing policy is complex, management, such as configuration of the
router and debugging, would be a problem. The ultimate goal is to
make the network manageable.

The following basic principles would help scale the routing policy
management.

     o Making policies as simple as possible but meet the requirements

     o Automating as much as possible to avoid error-prone manual work

     o Avoiding policy based on individual prefixes as much as possible
       with the exception of prefix-based route filtering for
       protecting routing integrity

     o Avoiding making exceptions

     o Using out-of-band routing policy processing where possible

6.6. Out-of-Band Process

   A typical router assumes both routing and forwarding functions.
   However, conceptually, routing and forwarding are two separate
   processes. A router's ultimate task is to forward packets based on
   its forwarding table, which is derived from routing information. One
   of the main causes of route scaling problems is that routers run out
   of processing power because routing requires too much processing.
   While a router has to forward packets, it does not necessarily have
   to exchange and process routing information or execute routing
   policy; these tasks can be performed elsewhere. Thus the question
   should be: Would it be possible to remove the routing process from a
   router to reduce its burden? Moving the routing process from the
   routers to other systems is referred to as out-of-band route
   processing.

   Out-of-band route processes would, in short, perform the heavy-duty
   routing tasks. They would build a forwarding table for the router,
   select routes based on pre-defined policy, filter routes, and shield
   the router from route flapping attacks.

   The shortcomings of out-of-band route processing are the possible
   introduction of delays in routing changes; the de-coupling of routing
   and forwarding paths, which could introduce inaccurate routing
   information; and the cost of extra equipment.

   Appendix A presents a current example of out-of-band route
   processing. It also suggests other possible solutions.

7. Conclusion and Discussion

   How routing scales has a direct impact on network stability and
   performance. With the fast growth of the Internet and consequent
   expansion of providers' networks, routing scaling become increasingly

an important issue to address. This document identifies the major
factors that affect route scalability and establishes basic
principles for designing scalable routing in large networks.

The major routing scaling issues we are facing today are excessive
router resource consumption due to routing processing burdens causing
routing convergency difficulties thus introducing network
instability; and routing complexity resulting in difficulties of
management and trouble shooting causing degradation of service.

The outlined principles for designing a scalable routing system are
building routing hierarchy; introducing fault isolation; reducing
routing processing burden where possible; defining manageable routing
policies and using the assistance of available out-of-band routing
process.

The use of out-of-band resources to assist routing processing is a
concept only been used in the Internet Exchange Points (IXPs).
However, it could potentially be used to advantage within a network
to help addressing routing scaling issues. This is a topic worthy of
further exploration.

Routing protocols and/or their implementations can still be improved
or enhanced for better handling of the scaling issues. For example,
the IS-IS multiple level mechanism is needed in order to scale the
IGP in large network. Also, using multicast or a reliable flooding
mechanism for I-BGP updates instead of pairwise full mesh peering is
something worth investigating.

It is our belief that even with the deployment of new technologies
such as DWDM, MPLS and others in the future, the fundamental routing
scheme will remain the current IGP/BGP paradigm.  Therefore, the
scalable routing design principles outlined in this document should
still apply with the deployment of new technologies.

8. Security Considerations

This document deals with routing scaling issues and thus is unlikely
to have a direct impact on security.

However, certain routing scaling improvement mechanisms suggested in
the document, such as network compartmentalization, will possibly
alleviate network outages caused by denial-of-service attacks since
it would help prevent such outages from spreading to the entire
network.

Although the mechanisms described in this document do not enhance or
weaken the security aspect of routing protocols, it is worth
indicating here that security enhancement of routing protocols or
routing mechanisms may impact routing scalability. Therefore, when
applying security enhancement in routing, one has to be aware of the
implications on scalability.

For example, TCP MD5 signature option is proposed to be a mechanism
to protect BGP sessions from being spoofed [13]. It is done on a
per-session basis and the overhead of MD-5 extensions are minimal
thus has no direct impact on scalability. There have been concerns
about doing per-prefix AS path verification as any one ISP along a
path could have forged or modified information (maliciously or not).
One extreme solution is to have a signature for each prefix which
gives very strong security but presents enormous scaling issues in
terms of processing, memory and administrative overhead.

9. Acknowledgement

Special thanks to Curtis Villamizar and Dave Katz for the extensive
review of the document and many helpful comments. Many thanks to
Yakov Rekhter, Noel Chiappa and Rob Coltun for their insightful
comments. The author also like to thank Susan R. Harris for the much
needed polishing of English language in the document.

The author was made aware after the publication of this document that
there is a relevant and independent presentation made by Enke Chen on
the subject. The presentation is thus referenced in [14].

10. References

   [1]   "Intermediate System to Intermediate System Intra-Domain
         Routeing Exchange Protocol for use in Conjunction with the
         Protocol for Providing the Connectionless-mode Network Service
         (ISO 8473)", ISO DP 10589, February 1990.

   [2]   Callon, R., "Use of OSI IS-IS for Routing in TCP/IP and Dual
         Environments", RFC 1195, December 1990.

   [3]   Moy, J., "OSPF Version 2", RFC 2328, April 1998.

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

   [5]   C. Labovitz, R. Malan, F. Jahanian, "Origins of Internet Routing
         Instability," in the Proceedings of INFOCOM99, New York, NY,
         June, 1999

   [6]  J. Moy, "OSPF-Anatomy of an Internet Routing Protocol",
        Addison-Wesley, January 1998.

   [7]  Bates, T., Chandra, R. and E. Chen, "BGP Route Reflection - An
        alternative to full mesh IBGP", RFC 2796, April 2000.

   [8]  Traina, P., "Autonomous System Confederation Approach to Solving
        the I-BGP Scaling Problem", RFC 1965, June 1996.

   [9]  Curtis, V., Chandra, R. and R. Govindan, "BGP Route Flap
        Damping", RFC 2439, November 1998.

   [10] Fuller, V., Li, T., Yu, J. and K. Varadhan "Classless Inter-
        Domain Routing (CIDR): an Address Assignment and Aggregation
        Strategy", RFC 1519, September 1993.

   [11] Villamizar, C., Alaettinoglu, C., Govindan, R. and D. Meyer,
        "Routing Policy System Replication", RFC 2769, February 2000.

   [12] Bates, T., Bush, R., Li, T. and Y. Rekhter, "DNS-based NLRI
        origin AS verification in BGP", Work in Progress.

   [13] Heffernan, A., "Protection of BGP Sessions via the TCP MD5
        Signature Option", RFC 2385, August 1998.

   [14] E. Chen, "Routing Scalability in Backbone Networks." Nanog
        Presentation: http://www.nanog.org/mtg-9901/ppt/enke/index.htm

   [15] T. Li, T. Przygienda, H. Smit,  "Domain-wide Prefix Distribution
        with Two-Level IS-IS", Work in Progress.

Author's Address

   Jieyun (Jessica) Yu
   CoSine Communications
   1200 Bridge Parkway
   Redwood City, CA  94065

   EMail: jyy@cosinecom.com

Appendix A. Out-of-Band Routing Processes

   The use of a Route Server(RS) at NAPs is an example of achieving
   routing scalability through an out-of-band routing process. A NAP is
   a public inter-connection point where ISP networks exchange traffic.
   ISP routers at a NAP establish BGP peer sessions with each other. The
   result is full mesh E-BGP peering with a complexity of O(N^2) system
   wide. When the RS is in place, each router peers only with the RS
   (and its backup) to obtain necessary routing information (or more
   precisely, the necessary forwarding information). In addition, the RS
   also filters routes and executes policy for each provider's router,
   which further reduces the burden on all routers involved.

   The concept of the Route Server can also be used to help address
   routing scalability in a large network.

   1) RS Assisted Peering between Customer Aggregation Router and
   Customer Routers

   Currently, in a typical large provider network, it's not unusual that
   a customer aggregation router connects up to hundreds of customer
   routers. That means the router has to handle hundreds of E-BGP
   sessions and filter a large number of prefixes. These tasks impose a
   heavy burden on the aggregation router. Reducing the number of
   customer routers per aggregation router is not an optimal option,
   since this would introduce more routers in the routing system of the
   whole network, which is neither scalable for backbone routing, nor
   cost efficient. Using an RS between customers and the providers'
   customer aggregation router become an attractive option to reduce the
   burden on the router.

   Figure 1 shows one way of incorporating an RS router between a
   provider's customer aggregation router and customer routers.

```
                    -------------------------- LAN Media in a POP
                       |             |
                     -----         -----
                    |CR |         |RS |
                     -----         -----
                    / | \
                   /  |  \
                 C1  C2..Cn
```
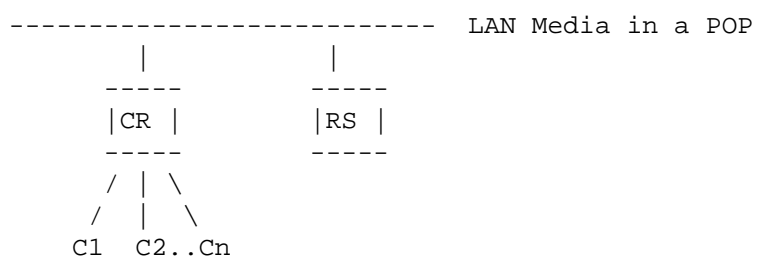
           Figure 1: RS serving customer aggregation router connecting
                   customer routers

In a scenario without an RS, the customer aggregation router(CR) has
to peer with customer routers C1, C2 ... Cn (where n could be in the
hundreds). When an RS router is introduced, CR, C1, C2 ... Cn peer
with the RS router instead, and the RS passes the processed routing
information (or forwarding information) to all of them, according to
policy and filters.

The advantages are obvious:

   o The customer aggregation router peers only with the RS router
     instead of with hundreds of customer routers.

   o The customer aggregation router does not need to filter prefixes
     or process routing policies, which frees resources for packet
     forwarding and handling.

One general concern with the use of an RS router is the possibility
of a mismatch of routing connectivity and the physical connectivity.
For example, if the link between the CR and C1 is down and if the RS
router is not aware of the outage, it will continue to pass routes
from C1 to the CR, and the traffic following these routes will be
black holed. However, this is not a problem in the specific
application described here. This is because the RS router has to go
through the CR to peer with C1, C2 ... Cn. When the link is down, C1
is inaccessible from the RS router, and no routing information can be
exchanged between the two. Consequently, the RS will announce no
routes related to C1.

Another concern is the creation of single point of failure. If the RS
router is down, no routing information can be exchanged between the
customer aggregation router and C1, C2 ... Cn, and no traffic will
flow between them. This problem could be addressed by adding a second
RS router as a backup.

In this scenario, since RS peers with C1 ... Cn via CR, it requires
that when the RS router passes routing information to C1...Cn, it
designates the IP address of the CR as the next hop. Likewise, when
the RS router passes routes from each customer router to the customer
aggregation router, it needs to place the correct next hop on the
route. Modifications need to be made to the RS code to include this
function.

2) Private RS Router at InterExchange Point

A large provider network often has many BGP peers at the
Interexchange Point, NAP or private interconnection. This means a
border router has to handle many E-BGP sessions. Since an

Interconnect points is usually the administrative boundary between
ISPs, policy and route filtering are very demanding. This imposes a
scaling problem on the border router.

Deploying many routers to distribute the load among them is an
expensive solution: extra hardware and extra ports cost money.
Shifting the routing burden to an RS router is a promising
alternative solution. In the case of using RS for multiple peers at a
private interexchange point, the scenario is similar to RS used
between customer aggregation router and customer routers as described
in 1) above. In the case of such peering at a NAP, the private RS
could be placed either on the same NAP media or a private media
between the ISP's NAP router and the RS.

3) RS Routers at Each POP in a Large Network

Even in a network with a hierarchical routing structure, a sub-area
may become too large, and I-BGP full meshing may impose a scaling
problem. One way to address this would be to split the sub-area or
add yet another tier of I-BGP reflector structure. Another possible
solution would be to use an RS router as an I-BGP Server. Depending
on the topology of a POP, this solution may or may not be suitable.
The use of RS routers at network POPs need to be investigated
further.

Full Copyright Statement

Acknowledgement