Globally Distributed Content
(Using BGP to Take Over the World)

Horms (Simon Horman)

horms@vergenet.net

November 2001

http://supersparrow.org/

Introduction

Electronic content is becoming increasingly important.

Network-wise closer content provides better service to end users.

Static mirrors of sites may be used.

But manually selecting sites from a list of mirrors is a tedious process.

It makes sense for the service provider to automatically direct clients to a site that will offer them good performance.

A global load balancing infastructure.

Selecting Servers

Local load balancing algorithms may take into account load and number of connections.

It is assumed that the cost of a client connecting to any host is equal.

Global load balancing changes this.

Relative speed and bandwidth between different servers and end users become factors.

This paper will discuss methods of enabling clients to connect to servers such that network delays are minimised.

Selecting Servers: Time Constraints

Load balancing decisions must be made quickly

A long process will at best result in slugish perofmrance and at worst incoming client connections may timeout

Selecting Servers: Routing Information

Routing information determines the path that packets will take.

Changes as network topology changes.

May be used in determining the best POP for a client.

BGP

Border Gateway Protocol version 4.

Dynamic Routing Protocol.

Communicates routing information between different Internet providers.

Reflects the path traffic will take from a given point on the Internet.

BGP: Routing Protocols: Routes and Routers

route: Set of addresses and the next hop used to send traffic to.

router: Nominally a host that has more than one network interface and makes decisions about to which interface a given packet should be sent.

As networks increase, number or routes and frequency of route changes increases.

A dynamic method of managing routes is needed.

BGP: Routing Protocols

Routing protocols are a mechanism for routers to communicate routes with each other.

peers: Routers that communicate routes with each other.

When a router sends such a route it is said to be *advertising*.

Advertising routes that cannot be satisfied leads to either *routing loops* or *black-holing*.

BGP: Routing Protocols: Prefixes

prefix: Set of network addresses that a given route covers.

Older routing protocols use Classful networks for prefixes.

More recent routing protocols use or CIDR networks.

Classless Inter-Domain Routing (CIDR) is defined in RFC 1519.

CIDR networks allow networks to be defined as a network address and a netmask, enabling more flexible division of networks than classful routing.

BGP: Routing Protocols: Sessions

When peers are configured to communicate routes with each other they are said to have a *session* running.

The routers advertise routes to each other over the session.

Each router uses this information to determine the best route for each prefix.

When a session goes down the prefixes advertised by the peer in question are removed.

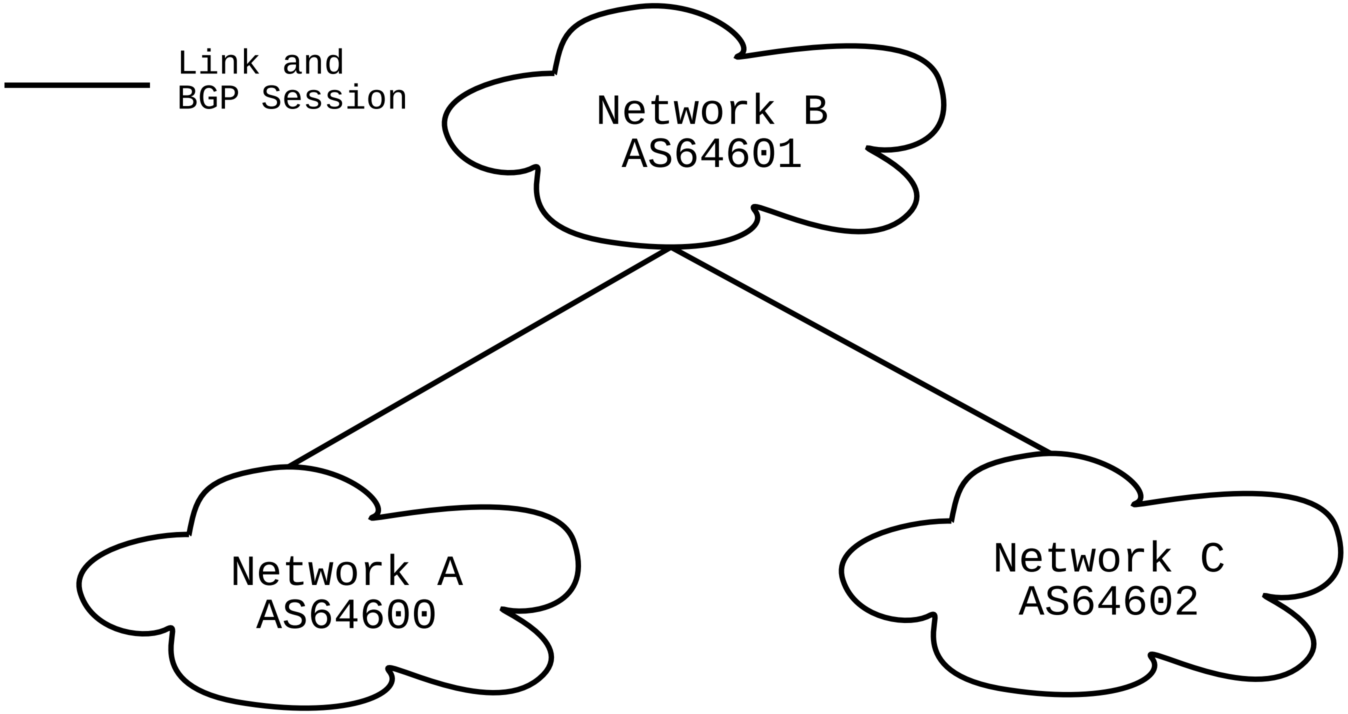
BGP: Autonomous Systems: I

When networks communicate routes using BGP, individual networks are identified using an Autonomous System (AS) Number.

AS Numbers are defined in RFC 1930.

Each route communicated using BGP contains an *AS path*, an ordered list of ASes that the route has been advertised by.

BGP: Autonomous Systems: II



BGP: Autonomous Systems: III

Networks A, B and C have the AS numbers 64600, 64601 and 64602 respectively.

Networks A and C are each directly connected to B.

BGP peering sessions are run between border routers in Networks A and B and Networks B and C.

There is no direct link between Networks A and C.

This given, the AS path on a router in Network A for a prefix advertised by Network C would be 64601 64602.

The route originated from AS64602 and was transited through AS64601.

BGP: Finding The Peer Closest to a Client: I

POP X and Y are on Networks A and C respectively.

Each POP has an AS Number.

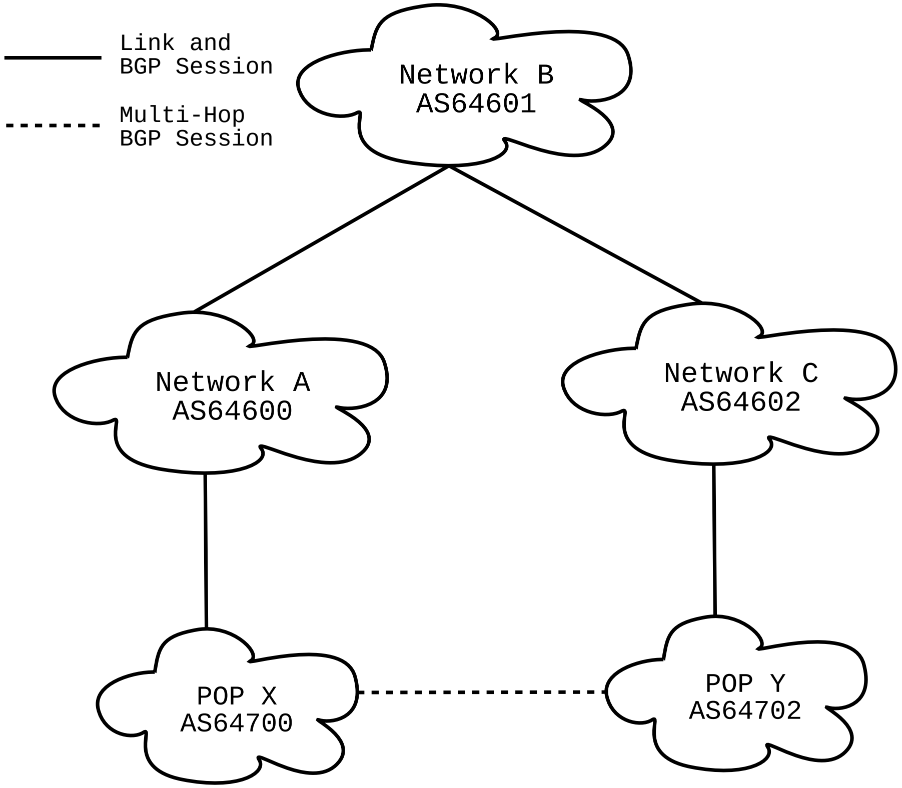
Each POP sets up a BGP session with its upstream network or networks.

POPs and upstreams should *filter* any routes originating from the POP.

Each POP has a view of all the routes that its respective upstreams have.

By establishing a *multi-hop BGP Session* between POPs, each POP can see the network that the other POP has, and in turn the view that POP's upstream has.

BGP: Finding The Peer Closest to a Client: II



BGP: Finding The Peer Closest to a Client: III

If the router running the BGP sessions to Network A from POP X is queried for the prefix used to route traffic to an address in Network C then there are two probable answers;

A prefix with the AS path 64600 64601 64602.

A prefix with AS path 64702 64602.

The latter prefix should be preferred as it has a shorter AS path.

BGP: Finding The Peer Closest to a Client: IV

As the preferred path contains the AS number of POP Y, this must be closer.

This means that if the AS number for one of the POPs appears in the AS path for a preferred prefix then the corresponding POP must be closer.

If the AS numbers of multiple POPs appear in the AS path then the last POP in the AS path must be closest.

Directing Traffic

The results of load balancing algorithms must be made available transperently to end users.

Layer 4 Switching is very effective for load balancing traffic on a LAN.

All in-bound packets and, often, all return packets passing through a single point.

Acceptable on a LAN where all packets must pass through a limited number of switches and routers.

On a WAN may significantly increas latency.

May reduce reliability as packets are traversing more hops across potentially uncontrolled networks.

Directing Traffic: A Better Way

Connections to be redirected to another site

Once the redirection has been made clients communicate directly to a POP.

To avoid a single point of failure any POP should be able to make redirections.

Two ways of achieving this are by using DNS and HTTP redirects.

Directing Traffic: DNS: I

DNS servers usually statically map a given query to a reply or list of replies.

Generally, the result changes infrequently if at all.

A DNS server may return results based on the output of some algorithm.

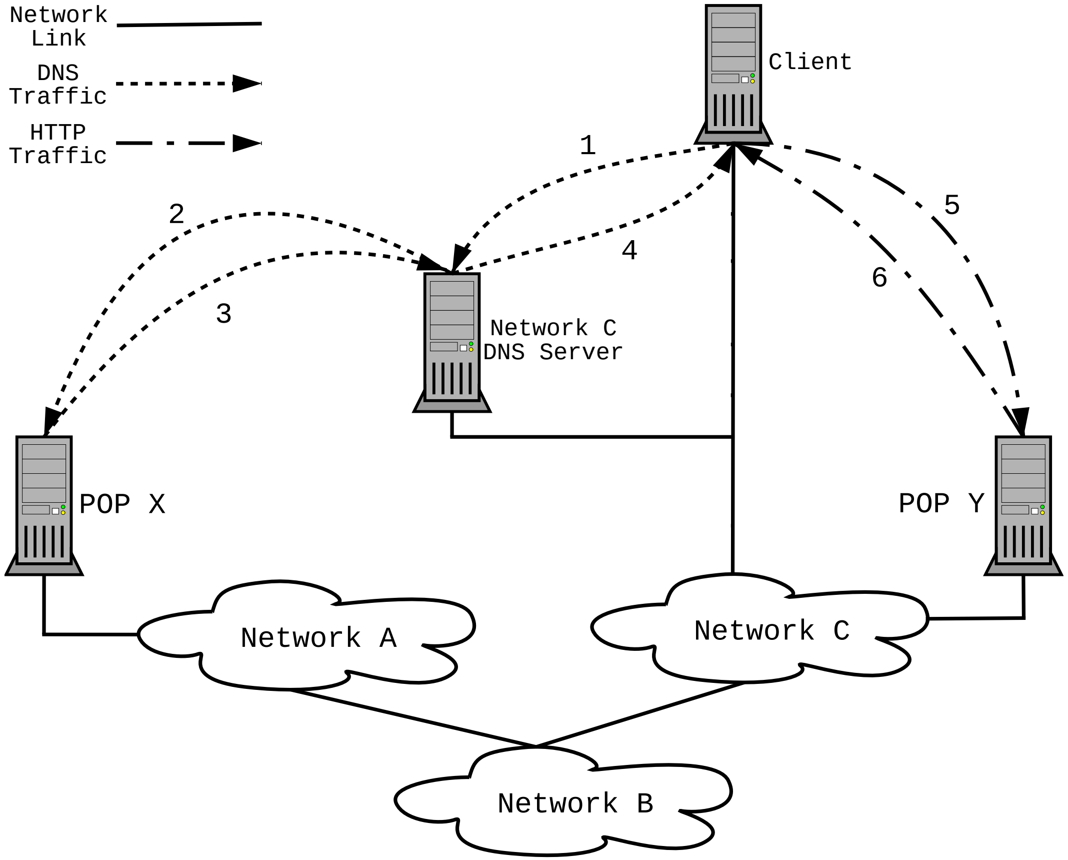
DNS lookups may be used the communicate the results of a load balancing algorithm to clients.

Multiple NS records for a domain give DNS some measure of redundancy.

Directing Traffic: DNS: I

As an example, suppose that *www.slarken.org.au* is mirrored between POP X and Y and DNS is being used to distribute traffic between these two POPs...

Directing Traffic: DNS: III



Directing Traffic: HTTP: I

HTTP Redirects may also be used to communicate load balancing information to end users.

Much finer granularity may be achieved. than for DNS

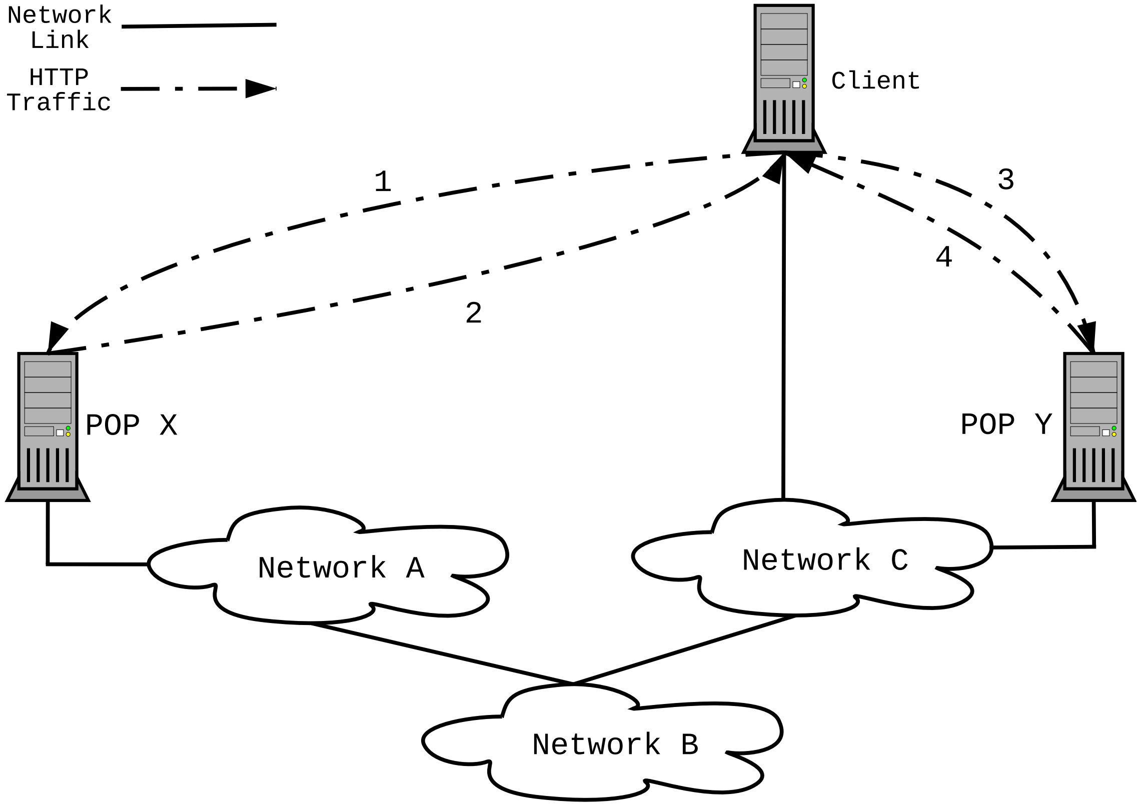
For instance all `.jpeg`, `.jpg` and `.png` URLs may be redirected while all other URL may be handled locally.

The disadvantage is that once a client is redirected to a site there may be no way of directing the client to another site.

Directing Traffic: HTTP: II

Suppose once again that *www.slarken.org.au* is mirrored between POP X and Y and that HTTP redirects are being used to distribute traffic between these two POPs...

Directing Traffic: HTTP: III



Implementation: Super Sparrow

Super Sparrow is an implementation of global load balancing.

Written in C. (Hooray :-)

Released under the GNU General Public Licence and GNU Lesser General Public Licence.

Available from *http://supersparrow.org/*.

Source and RPM packages with comprehensive online documentation.

Debian packages are being worked on.

Super Sparrow is able to load balance traffic by using BGP to find the peer closest to a client.

Implementation: Route Servers

Super Sparrow accesses BGP information by querying route servers.

A route server is a router, or host running a routing daemon, that may be queried for the preferred prefix for a IP address.

GNU Zebra, GateD and Cisco IOS are supported.

Telnet is used to access to route servers to query the preferred prefix for an IP address.

Implementation: Route Servers: Sample Session

A sample session to determine the preferred prefix for 192.168.193.15 from a route server running on 192.168.192.13 is shown.

```
jasmine> sho ip bgp 192.168.193.15
BGP routing table entry for 192.168.193.0/24
Paths: (2 available, best #2, table Default-IP-Routing-Table)
  64600 64601 64602
    192.168.192.12 from 192.168.192.12 (192.168.192.12)
      Origin IGP, metric 1, localpref 100, valid, external
      Last update: Fri Oct  6 15:47:28 2000

  64702
    192.168.193.11 from 192.168.193.11 (192.168.193.11)
      Origin IGP, metric 1, localpref 100, valid, external, best
      Last update: Fri Oct  6 15:44:05 2000
```

Implementation: *libsupersparrow*

The core functionality of Super Sparrow

Communicates with route servers.

Manages connections to multiple route servers and multiple connections to a single route server.

Caches Results read from route servers.

Manages the relationship between the AS numbers of POP and their IP addresses or hostnames.

Implementation: *mod_supersparrow* (DNS)

Dents is a modular DNS server that is intended as a drop in replacement for BIND.

Zones to be mounted in the name space.

Analogous to mounting partitions in a UNIX directory structure.

Different driver modules may be used for diffrent zones.

mod_supersparrow is a Dents driver module that returns return results based on information from BGP speaking route-servers.

The IP address returned for a hostname lookup is governed by the BGP-based global loadbalancing algorithm implemented by Super Sparrow.

Implementation: *supersparrow*

supersparrow is a stand-alone application that is linked against lib-supersparrow.

Debugging tool during development of lib_supersparrow.

May be used in conjunction with applications that are able to commumicate with other programmes using standard I/O.

supersparrow with Apache (HTTP)

Apache's mod_rewrite allows arbitary rewriting of requests received by Apache to other URLs at run time.

The rewrite is done by a map and one of the map types supported is running an external programme.

mod_rewrite communicates with the external programme via standard I/O.

The external programme is run once when apache starts, requests are written to the programme's standard in and results are read from the programme's starndard out.

The *supersparrow* stand-alone application supports a batch mode, which allows it to be used as a map for mod_rewrite.

Conclusion

Global Load Balancing there needs to take into account factors that are not apparent when load balancing traffic on a LAN.

In particular there is a need to be completely independent of other sites.

The BGP-based algorithm discussed provides a powerful mechanism for determining the network-wise POP for a client.

It does not, take into account the relative capacity or load of the POPs.

It is anticipated that in the future the implementation of Super Sparrow will be expanded to allow other, non BGP-based algorithms.