BGPDNS

Using BGP topology information for DNS RR sorting
a scalable way of multi-homing

André Oppermann
oppermann@pipeline.ch

Claudio Jeker
jeker@n-r-g.com

RIPE 41 Meeting Amsterdam, 15. January 2002

What is BGPDNS about?

- BGPDNS is a *concept* and a *protocol* for doing AS-less and PI IP-less server multi-homing
- Use the topology information contained in the global BGP routing table to sort the multiple DNS resource records

because

- Traditional AS-based server multi-homing is a burden to the global Internet routing system
- Excessive consumption of AS number and non-aggregated prefixes

The demand

- In many cases it is desired by the customers to connect (important) servers to more than one upstream ISP. Reasons include:
  - Acquire redundancy in case (the link to) one upstream ISP fails
  - Balance/share load over more than one upstream ISP
  - Become independent from individual ISP's

Today‘s solution

- Today these objectives have to be satisfied by:
  - requesting PI IP space
  - obtaining an AS number
  - participate in the global BGP routing

Shortcomings of today‘s solution

- Whilst some advantages, this approach has several drawbacks to the Internet at large and to the newly multi-homed customer:
  - Large fragmentation of IP address space (bad)
  - Excessive memory and computing power requirements in the Internet core routers (good for Cisco and Juniper)
  - Exhaustion of current AS number space requires upgrade with transition to 32 bit AS numbers (very bad)

Shortcomings to customers

- Running and tuning BGP requires significant knowledge and experience as well as continued monitoring and adjustments
- BGP without knowledgeable tuning quickly leads to unintended asymmetric traffic patterns through the upstream ISP's

- Unqualified modification on BGP router quickly leads to disconnection from the global Internet because of missing route announcements or route flap dampening
- Misconfiguration of the routing table entries quickly lead to bogus route announcements (like a full /8 or multiple /24 instead of an aggregate) and can cause serious traffic interruptions (*hello Teleglobe Europe!*)

- ISP's have to employ very strict filters towards their multi-homed customers because of these frequent problems
- These filters in turn decrease flexibility and increase complexity while representing a significant source of errors in themselves.

Summary of shortcomings

- Many times the requestors of non-aggregated PI IP address space and AS numbers are not aware of these implications and lack sufficient technological background knowledge to qualify and quantify the impact on themselves and the Internet in general

- Many times only one or a subset of one of the reasons for AS based multi-homing is given by the requestor

- Unfortunatly in these cases the cure of AS based multi-homing is often worse than the disease of being single-homed

An alternative is needed

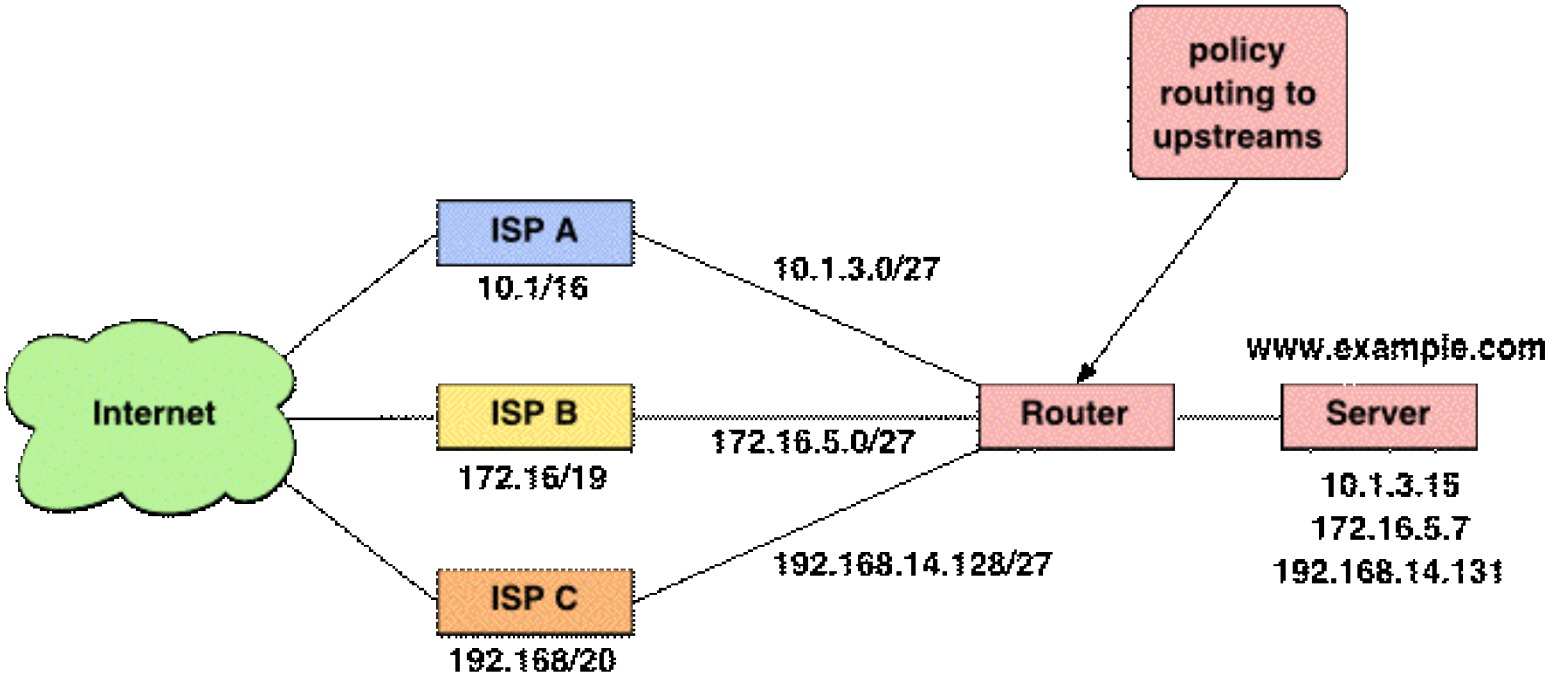
Overview of BGPDNS

- The BGPDNS approach combines the power of BGP with the ease of DNS
- BGPDNS does BGP but does *not* require an unique AS number on the customer side
- BGPDNS does *not* need PI IP address space and fully maintains aggregates

Details of BGPDNS

- In BGPDNS multi-homing, the server operator has upstream links to more than one ISP
- From each of these ISP's the operator also receives a reasonable IP prefix out of their aggregates
- One IP address of each prefix of the ISP's is assigned to the multi-homed server
- The router connecting to all these ISP's does policy routing to direct the outgoing packets to the ISP where the prefix belongs to.

Details of BGPDNS



Details of BGPDNS

- All of the IP addresses of this server are put into DNS as multiple “A” records to the same name

```
www.example.com.        IN A    10.1.3.15
                        IN A    172.16.5.7
                        IN A    192.168.14.131
```

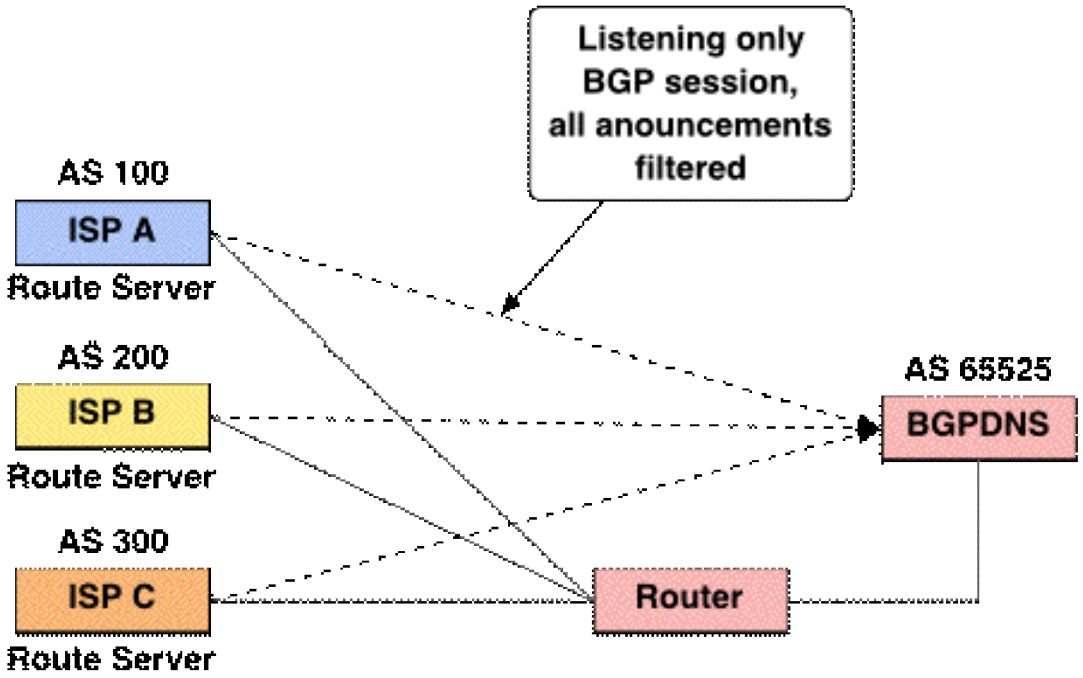
- Instead of round-robin we‘re going to use BGP for ordering!

Details of BGPDNS

- The BGPDNS server establishes a BGP *listening-only* session with each of the ISP's route-servers to get a comprehensive view of the Internet topology from it's own perspective

- It does normal best-path evaluation either subject to the default rules or custom crafted metrics as in normal AS-based multi-homing

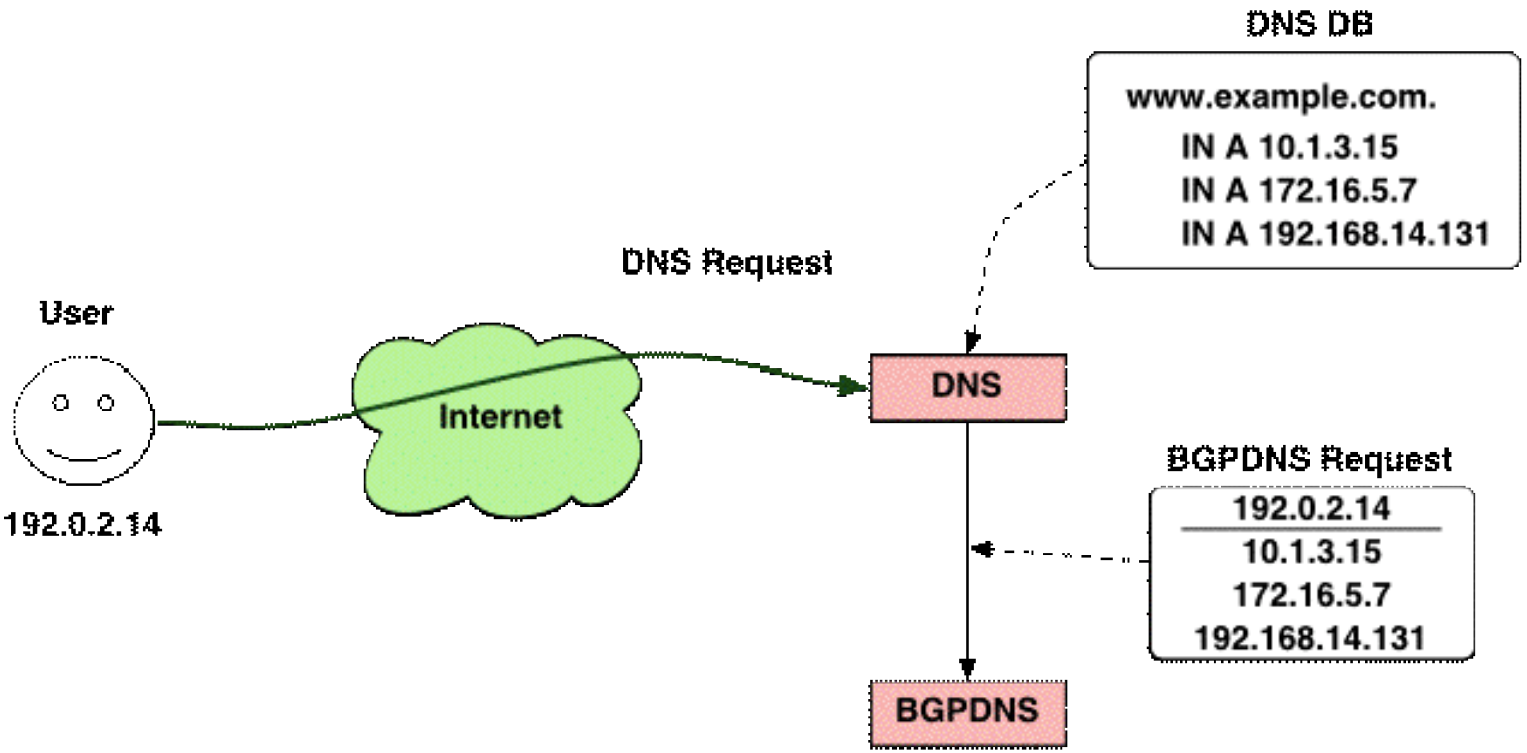
Details of BGPDNS



Details of BGPDNS

- User types www.example.com into her browser (her IP address is 192.0.2.43)
- Authorative DNS server at BGPDNS site receives request for www.example.com
- DNS server finds multiple „IN A“ records for www.example.com
- DNS server has to find out which „IN A“ is best reachable for this user → Ask BGPDNS server!

Details of BGPDNS



Details of BGPDNS

- BGPDNS server receives request from DNS server containing:
  - source IP of DNS request (192.0.2.43)
  - list of possible answers (10.1.3.15, 172.16.12.7 and 192.168.14.31)
- BGPDNS server looks up the best path to the requestor via the BGP topology information
  - „show ip bgp 192.0.2.43“

Details of BGPDNS

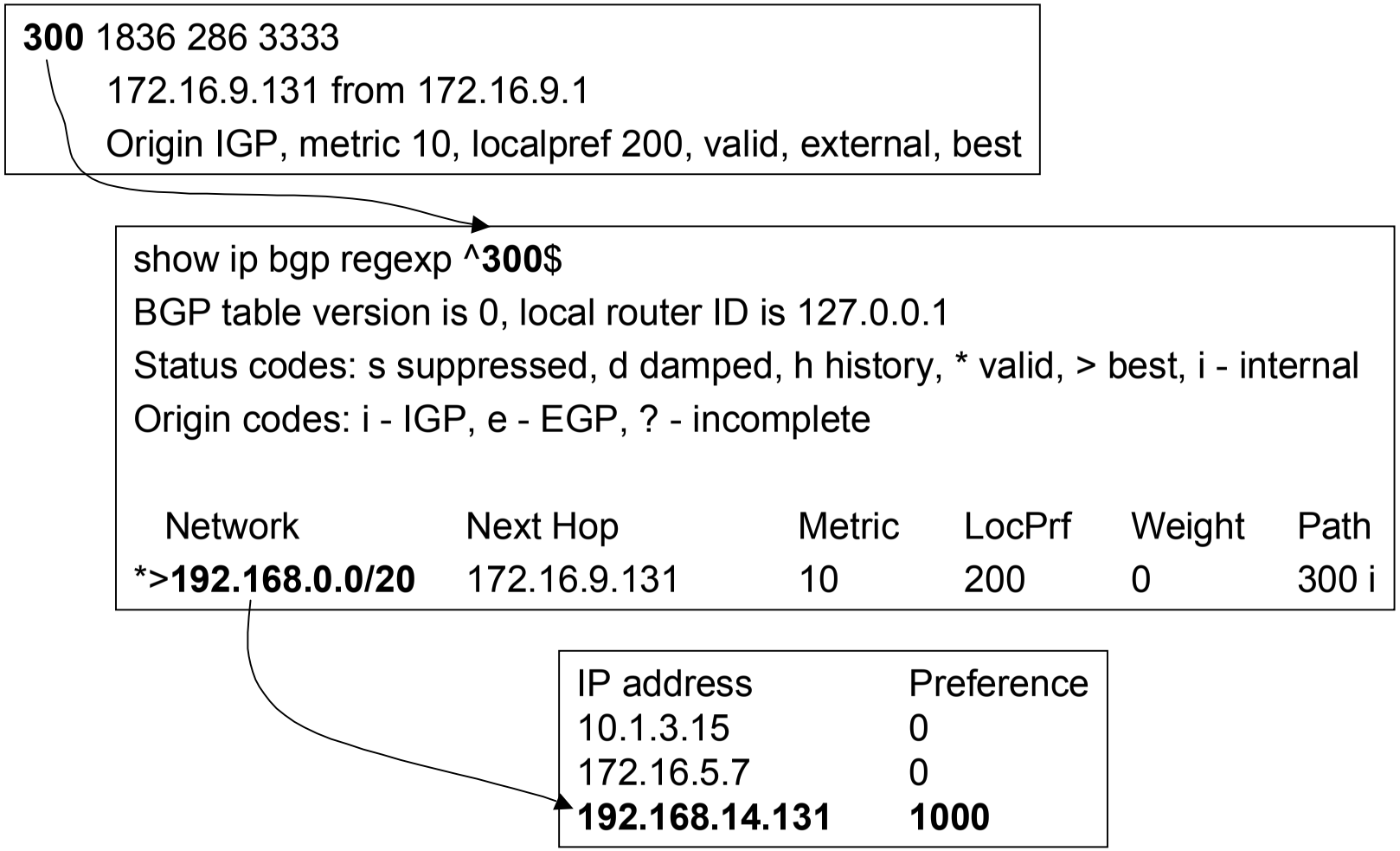
show ip bgp **192.0.2.43**
BGP routing table entry for 192.0.2.0/21
Paths: (3 available, **best #2**, table Default-IP-Routing-Table)

100 8235 1836 286 3333
    10.1.1.18 from 10.1.1.1
    Origin IGP, localpref 200, valid, external
300 1836 286 3333
    172.16.9.131 from 172.16.9.1
    Origin IGP, metric 10, localpref 200, valid, external, best
200 9177 8210 3333
    192.168.5.56 from 192.168.5.56
    Origin IGP, metric 10, localpref 100, valid, external

Details of BGPDNS

- BGPDNS takes leftmost AS number of best path (which is the active path from our point of view)
- BGPDNS takes a list of all prefixes of the upstreams
  - “show ip bgp regexp ^leftmost-as$”
- BGPDNS loops over the „IN A“ records to find the local IP that is within one of these upstream prefixes and assigns the highest weight to it
  - It has to be because all the „IN A“ IP‘s are form our upstreams
  - If not, normal round-robin applies as usual

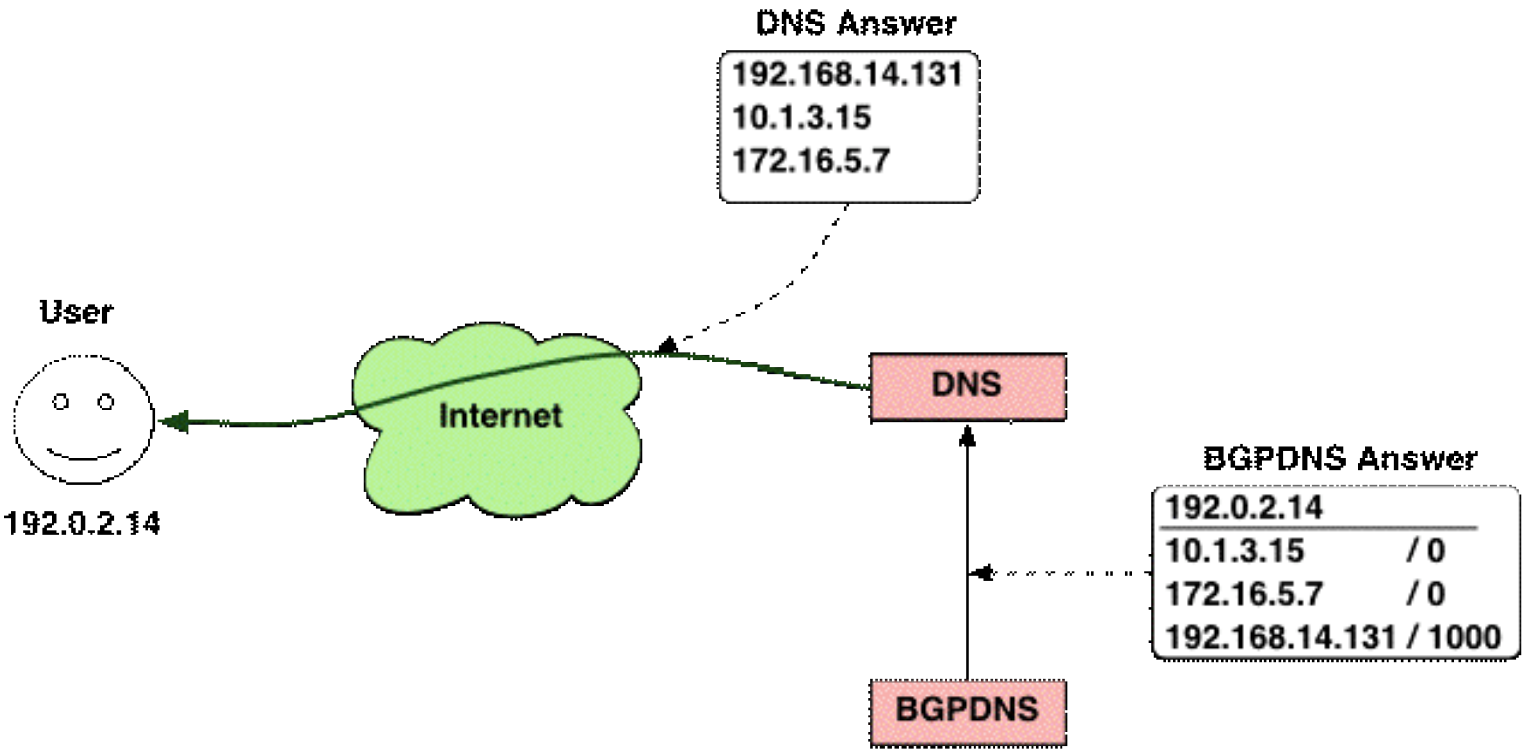
Details of BGPDNS



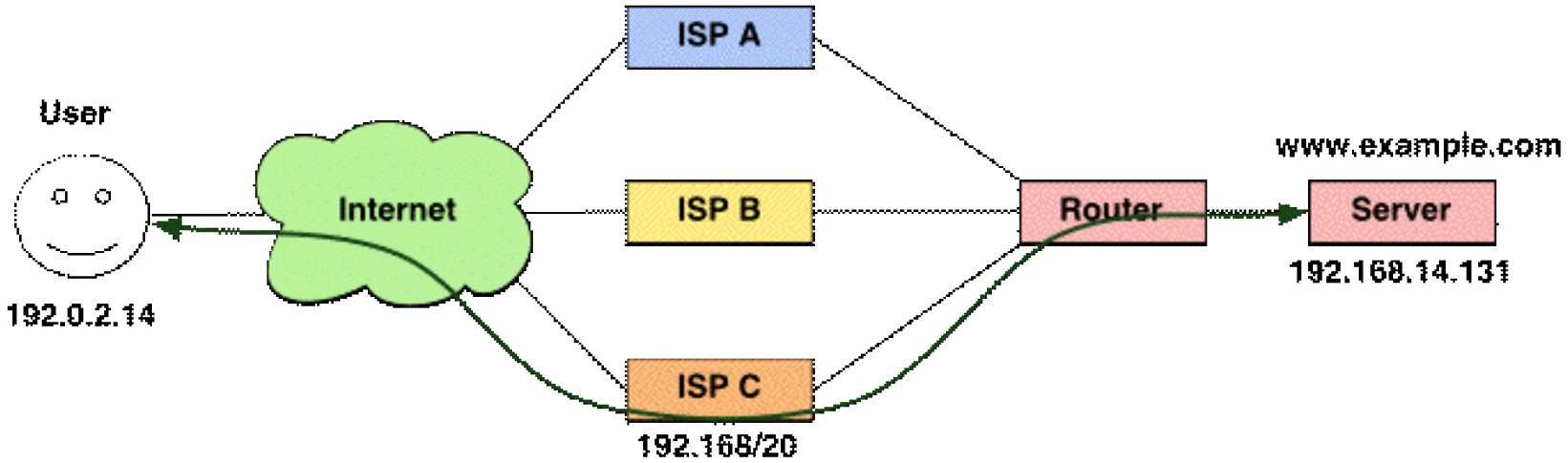
Details of BGPDNS

- BGPDNS returns the packet to DNS server with best path IP set to highest preference
- DNS server sorts response with highest preference first and answers to the user
- User uses the first IP address of the DNS response to establish a connection with the target server

Details of BGPDNS



Details of BGPDNS



Comparison to normal BGP

- Responsiveness to link state or topology changes is immediate for new requests
  - If a link fails, the corresponding IP won't be chosen for new requests anymore because all BGP NLRI information is gone, hence no more best path for this upstream
- In the case of a upstream link failure, the maximum black hole time for a particular requestor is the configured DNS resource record expiration timeout
  - Affects only requestors who had this particular upstream link as best path
  - The DNS RR expiration timeout has to be chosen carefully!

Convergence timers

- BGP today has a global convergence and propagation time of approx. 3 minutes as shown by recent research
- Determining the optimal DNS RR expiration timeout is balancing between two opposite tradeoffs:
  - If the expiry timeout is low this will add latency during a session leading to poor responsiveness experience by the user because of multiple successive DNS requests
  - if the expiration timeout is high this may reuse cached values with now sub-optimal path information or, if the link of the preferred IP has gone down in the meantime, to partial unreachability until the cached DNS RR expires
- We recommend DNS RR expiry timeouts between 20 seconds and 2 hours

Advantages of BGPDNS

- Link and ISP redundancy
- Load balancing over more than one link and ISP
- Independence of a single ISP
- Connection path symmetry between the server and the client (because of prefix based routing)
- No impact on global BGP routing system

Disadvantages of BGPDNS

- Each server requires as many IP addresses as it is connected to ISP's
  - Which is relative, AS-based multi-homing needs at least a /24, here in this example with three ISP's I can get away with 3x /27 or so
- For cached DNS resource records; only timeout-based convergence
- Additional load on the DNS system
  - Which doesn't seem to be a problem (Thanks Akamai!)

Rationale for BGPDNS

- IP Address space depletion is not as fast and does not have the same impact as AS number space depletion and Internet core router memory consumption
- With next generation IP numbering (IPv6?) address space depletion is no longer a issue
- Content delivery networks like Akamai have proven that DNS RR based global load balancing is working on a large scale and does not have a negative impact on the Internet at large nor on the individual user

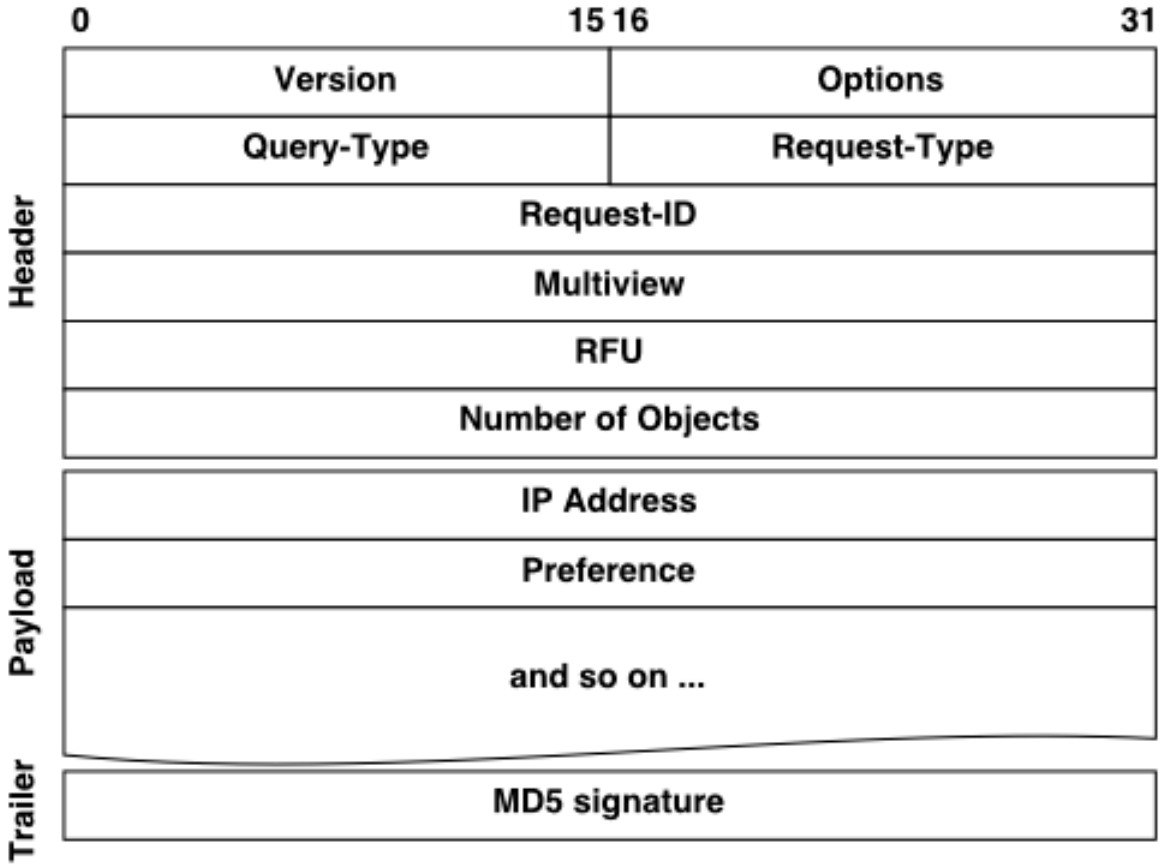
Rationale for BGPDNS

- For all services that are either stateless, have only short-lived or restartable sessions BGPDNS is well suited and provides equal results as true AS-based multi-homing
  - applies to HTTP, HTTPS, SMTP, POP, IMAP, FTP (in part), NNTP (client sessions)...
- By keeping IP address space aggregation intact and the positive effects on AS numbers and router memory by far outweigh the negative effects of BGPDNS by requiring one IP address per server and upstream ISP

The BGPDNS protocol

- The BGPDNS protocol is spoken between the DNS server and the BGP listener
- The communication is stateless and uses the UDP datagrams for communication
- Because of its close relationship to the BGP the port number 179/UDP is chosen for the BGPDNS task on the BGP listener

The BGPDNS protocol



BGPDNS reference implementation

- DJBDNS‘ tinydns as DNS server
  - Has nice non-threaded internal design
  - It‘s impossible to follow bind9 code without suffering serious brain damange
- Zebra bgpd as BGPDNS server
  - More or less structured internal design
  - Any other recent and stable OpenSource BGP daemon?
- Available on www.BGPDNS.org

Important in implementation

- Response time is critical
  - The BGPDNS must do fast lookups to not delay the DNS RTT too much
  - Lookup times must be far below 1 second
- Timeouts
  - If the DNS server receives no response within 1 second from the BGPDNS server it will send out random order

BGPDNS performance

```
cvs.pipeline.ch> sh dnssort statistics

DNSsort statistics
------------------
Received packets: 139909174
sent packets:     138057953
dropped packets:  0
currently queued: 0 / 128

DNSsort cache table for view     0
---------------------------------
AS 5378 :  697113 hits
AS 15623:  377189 hits
AS 8237 :   11598 hits
AS 15517:   77078 hits
AS 8220 : 4415231 hits
AS 6667 :       0 hits
AS 15600:  132782 hits
AS 6893 :  475114 hits
AS 13646:   53162 hits
AS 1836 :  8518605 hits
AS 9044 :   515931 hits
AS 6776 :    66926 hits
AS 12350:   394278 hits
AS 12520:   148551 hits
AS 8758 :    78749 hits
AS 20940:        0 hits
AS 9177 : 97422320 hits
AS 8327 :   130593 hits
AS 8235 :     5372 hits
AS 12429:   607428 hits
AS 8833 :   579531 hits
AS 13250:    51576 hits
AS 13030:    53294 hits
AS 15667:   161435 hits
AS 8271 :    39818 hits
```

145 hits/second average

2.2% CPU on a PIII-550/FreeBSD4.4

Failure cases

- BGPDNS server does not respond
  - If the DNS server does not receive answers for some time it should mark the BGPDNS server as defektive and answer with random sorting
- Loss of all BGP sessions
  - If the BGPDNS looses all BGP session it will simply answer all requests without any preference set
- Network not in table
  - If the network prefix of the end-user is not in the BGP routing-table it will simply answer without any preference set

Security considerations

- Authorization of BGPDNS requestors
  - MD5 shared secret (like in OSPF)
- DoS/Overloading attacks
  - Can‘t be done much, BGPDNS should not fall over but provide some form of overload protection
- Spoofing of requests/answers
  - MD5 shared secret
  - Filter 179/UDP on border router / firewall
- Information leakage
  - The DNS information is public anyway

Recommendation

- More scrunity for AS number requests
- In case of services that are stateless, have only short-lived or restartable sessions
- A „BGPDNS policy“ like the RIPE “HTTP policy” or “Static Dial Up policy” should be applied
- And BGPDNS should be considered as an alternative then

Other Projects

- www.SuperSparrow.ORG
  - We found out about it quite some time after the initial idea during the research phase for the BGPDNS project
  - SuperSparrow has serious scalability issues; it uses telnets to the routers to access the BGP path information
  - The last release is 0.0.0 and over a year old from 9th January 2001

The authors

- André Oppermann, oppermann@pipeline.ch
  - Idea and concept
- Claudio Jeker, jeker@n-r-g.com
  - Implementation

- Other projects by the authors
  - qmail-ldap www.nrg4u.com

Questions and comments?

- Yes, you can grill us now!

- You‘ll find a RFC draft and patches on:

www.BGPDNS.org