

Content distribution via BGP based anycast

Michael Horn - AS250.net Project

whois -h 193.0.0.135 MH250-RIPE

- AS250.net Project / Foundation
- Content Distribution
- Hosting / Colocation
 - Free speech projects
 - Open Source projects
 - ...many other community related projects
- PoPs in:
 - Berlin, Frankfurt, Düsseldorf, Hamburg, Cologne, Amsterdam, Brussel, Paris, London, Milano, Zurich, Tampere, Cape Town, New York...

Where is the revolution?

- We will not...
 - ...discuss a new protocol
 - ...introduce a revolutionary concept
- However we will...
 - ...use existing routing protocols in a creative way
 - ...make use of them to improve...
 - ...connectivity of services
 - ...reliability of services
 - ..."the user experience"

The mission statement.

- Content delivery
- Service Distribution
- Increase reliability and service quality
- Protocols?
 - HTTP
 - DNS
 - VoIP protocols
 - etc...

Standard solution #1

- mirrors.

Standard solution #1

- mirrors.
- more mirrors...

Standard solution #1

- mirrors.
- more mirrors...

<http://ftp.uni-erlangen.de/pub/linux/kernel/>

<ftp://ftp.de.kernel.org/pub/linux/kernel/>


<http://home.nibbler.de/~mh/attic/somestuff/linux/kernel>

etc...

Standard solution #1

- mirrors.
- more mirrors...

ugly.




<http://ftp.uni-erlangen.de/pub/linux/kernel/>
<ftp://ftp.de.kernel.org/pub/linux/kernel/>
<http://home.nibbler.de/~mh/attic/somestuff/linux/kernel>
etc...

Standard solution #1

- mirrors.
- more mirrors...

ugly.



<http://ftp.uni-erlangen.de/pub/linux/kernel/>
<ftp://ftp.de.kernel.org/pub/linux/kernel/>
<http://home.nibbler.de/~mh/attic/somestuff/linux/kernel>
etc...

- servers are temporarily out of reach
- bookmarks...
- no clever loadbalancing
- etc.

Standard solution #2

- round robin DNS
- better traffic distribution
- automatic adding/removing of mirrors
- however:
 - no failover mechanisms
 - no intelligence in mirror selection
 - no real control over the traffic distribution

BGP gives us...

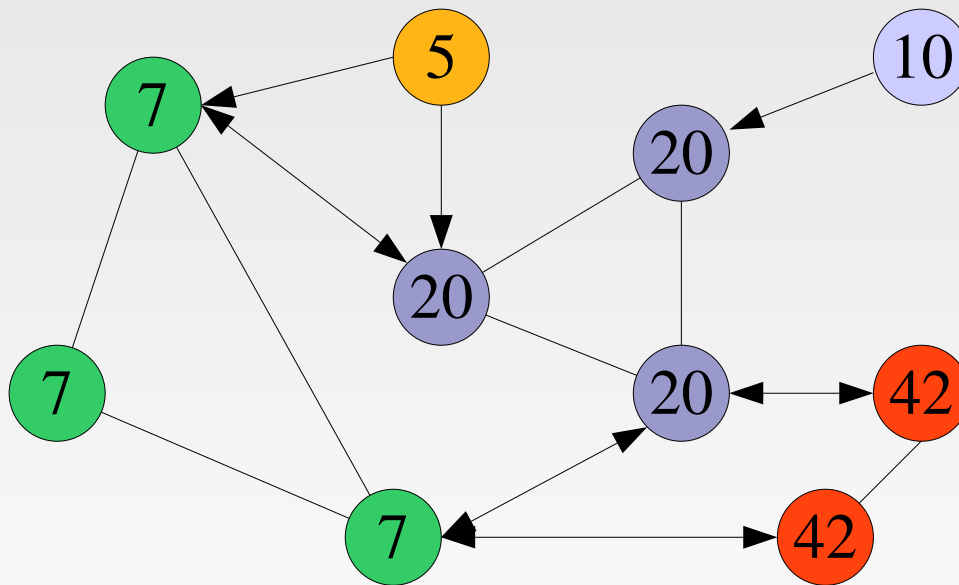
- automagic failover
- reliability through redundancy
- “best path” or “shortest path” is the goal
- minimize load on the network

Why not use BGP then?

- what happens if we advertise a prefix at two different locations?

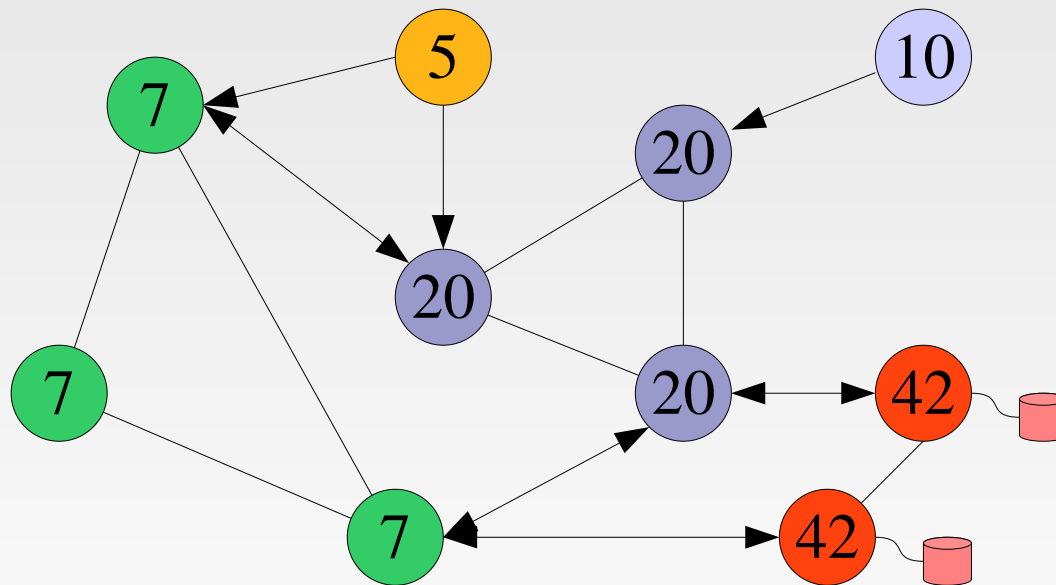
Why not use BGP then?

- what happens if we advertise a prefix at two different locations?



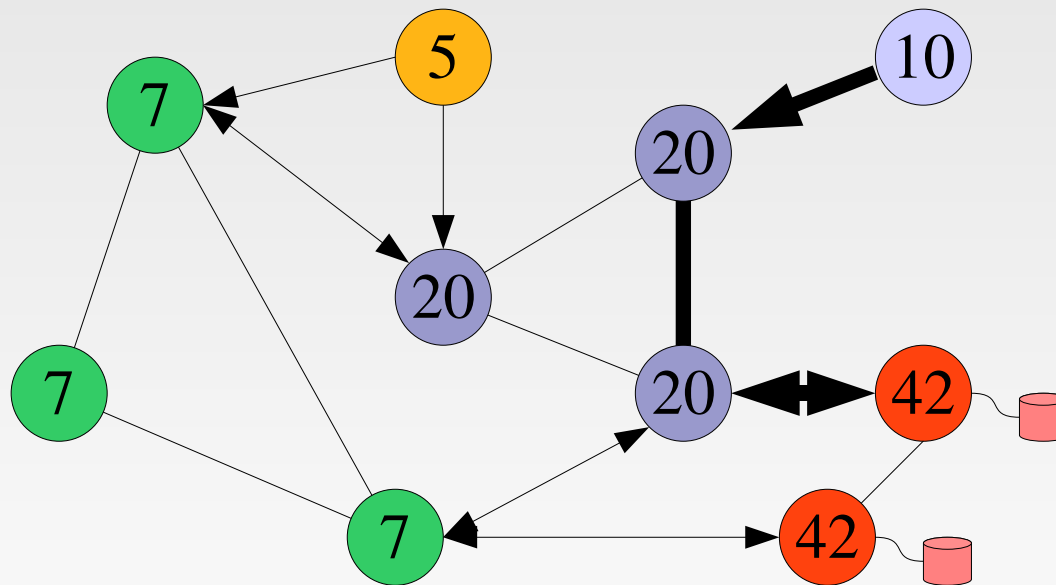
Why not use BGP then?

- what happens if we advertise a prefix at two different locations?



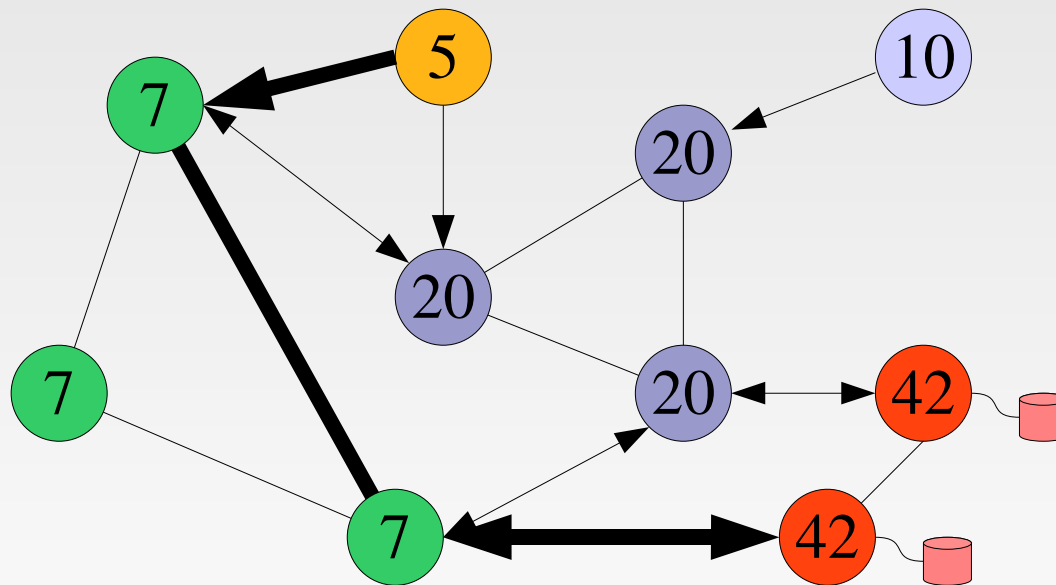
Why not use BGP then?

- what happens if we advertise a prefix at two different locations?



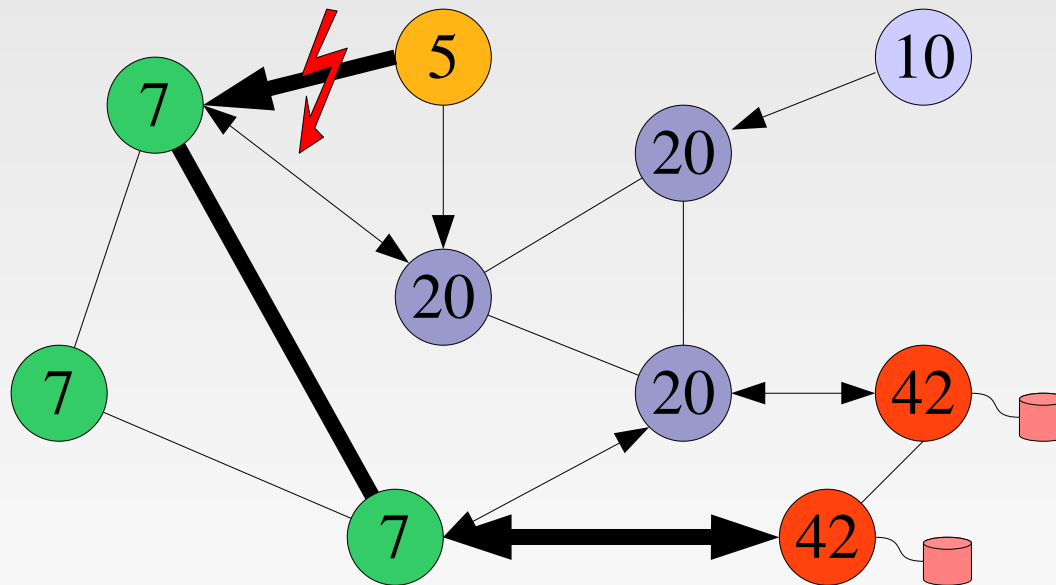
Why not use BGP then?

- what happens if we advertise a prefix at two different locations?



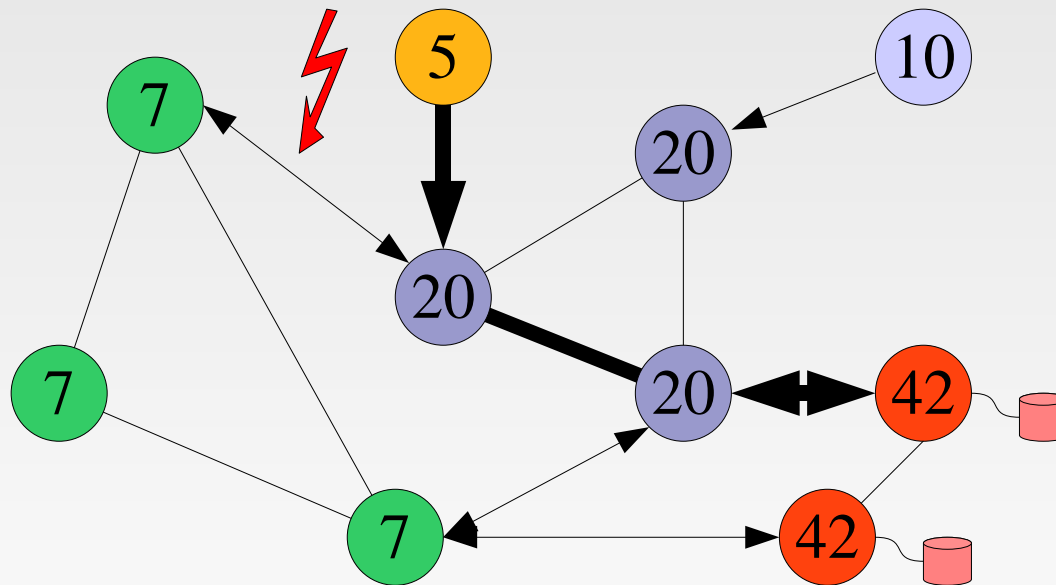
Why not use BGP then?

- what happens if we advertise a prefix at two different locations?



Why not use BGP then?

- what happens if we advertise a prefix at two different locations?



advantages of anycast BGP

- automagic failover

advantages of anycast BGP

- automagic failover
- regional aggregation of traffic

advantages of anycast BGP

- automagic failover
- regional aggregation of traffic
- global and local nodes possible

advantages of anycast BGP

- automagic failover
- regional aggregation of traffic
- global and local nodes possible
- fine-grained access to nodes via communities

advantages of anycast BGP

- automagic failover
- regional aggregation of traffic
- global and local nodes possible
- fine-grained access to nodes via communities
- no need to transport data through backbone

advantages of anycast BGP

- automagic failover
- regional aggregation of traffic
- global and local nodes possible
- fine-grained access to nodes via communities
- no need to transport data through backbone
- a better chance of surviving a DDoS attack

Fine for stateless protocols... but!

- DNS works fine
 - real world example: 141.1.1.1
 - k.root-servers.net

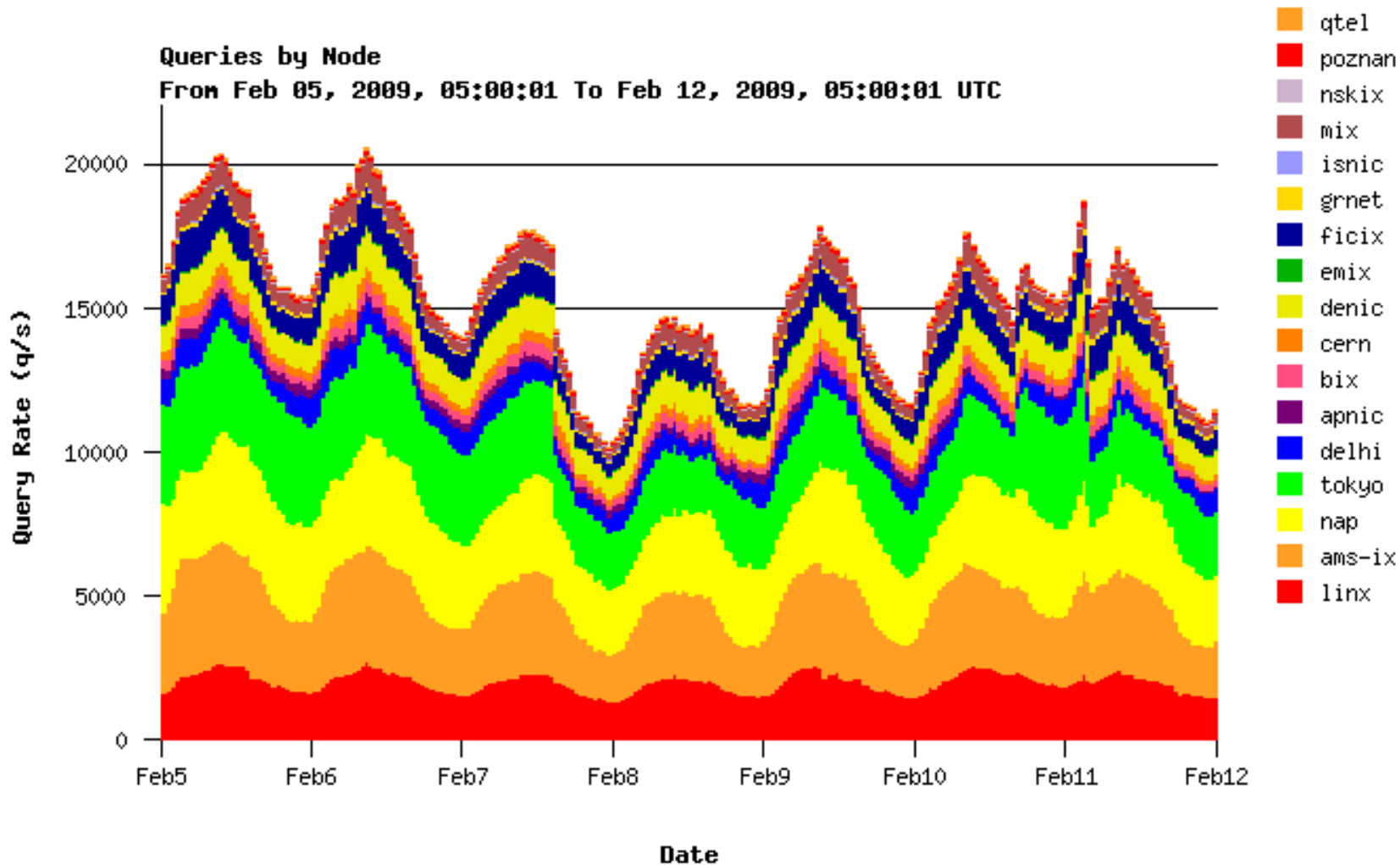
k.root-servers.net

Global Nodes: [Amsterdam, NL](#) • [London, GB](#) • [Tokyo, JP](#) • [Delhi, IN](#) • [Miami, Florida, US](#)



Local Nodes: [Budapest, HU](#) • [Milan, IT](#) • [Helsinki, FI](#) • [Reykjavik, IS](#) • [Poznan, PL](#) • [Frankfurt, DE](#) • [Geneva, CH](#) • [Athens, GR](#) • [Doha, QA](#) • [Novosibirsk, RU](#) • [Abu Dhabi, AE](#) • [Brisbane, AU](#)

k.root-servers.net



Fine for stateless protocols... but!

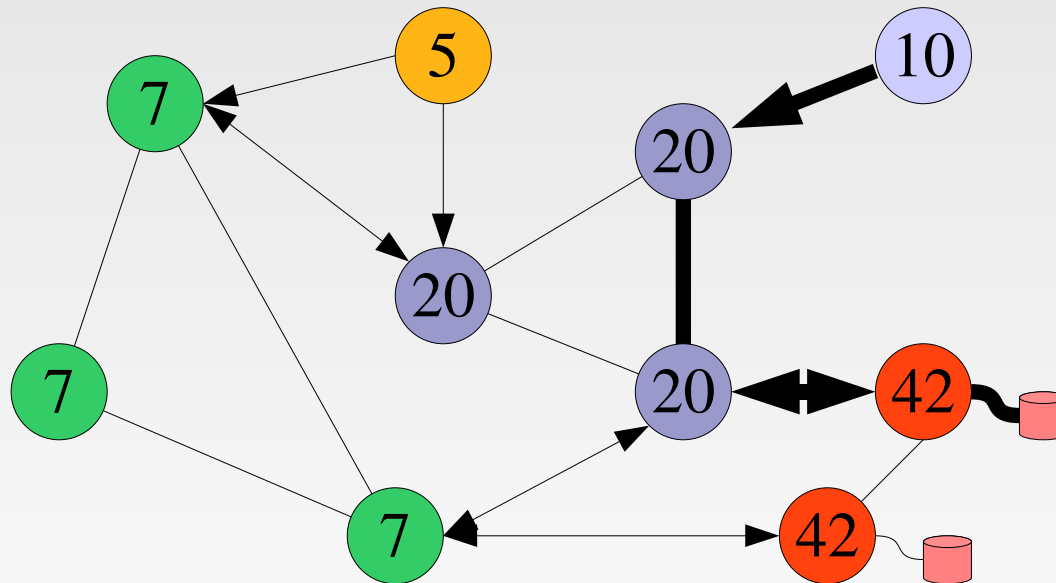
- DNS works fine
 - real world example: 141.1.1.1
 - k.root-servers.net
- but how about TCP?

Fine for stateless protocols... but!

- DNS works fine
 - real world example: 141.1.1.1
 - k.root-servers.net
- but how about TCP?
- example:
 - download of a >4 GB iso during a topology change?

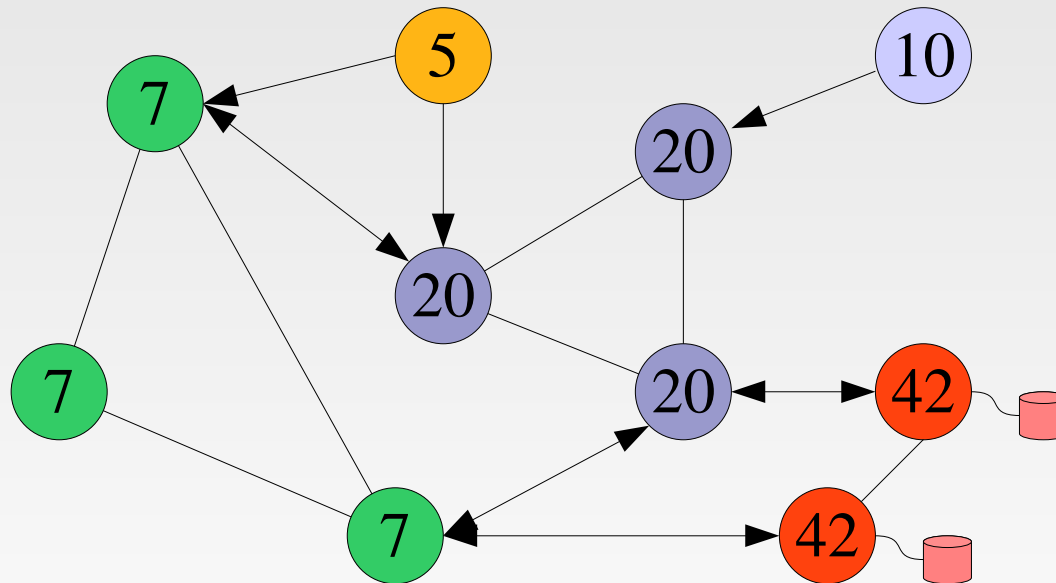
one possible solution...

- node receives DNS request for the content server
- and delivers node-local non-anycast IP as A record



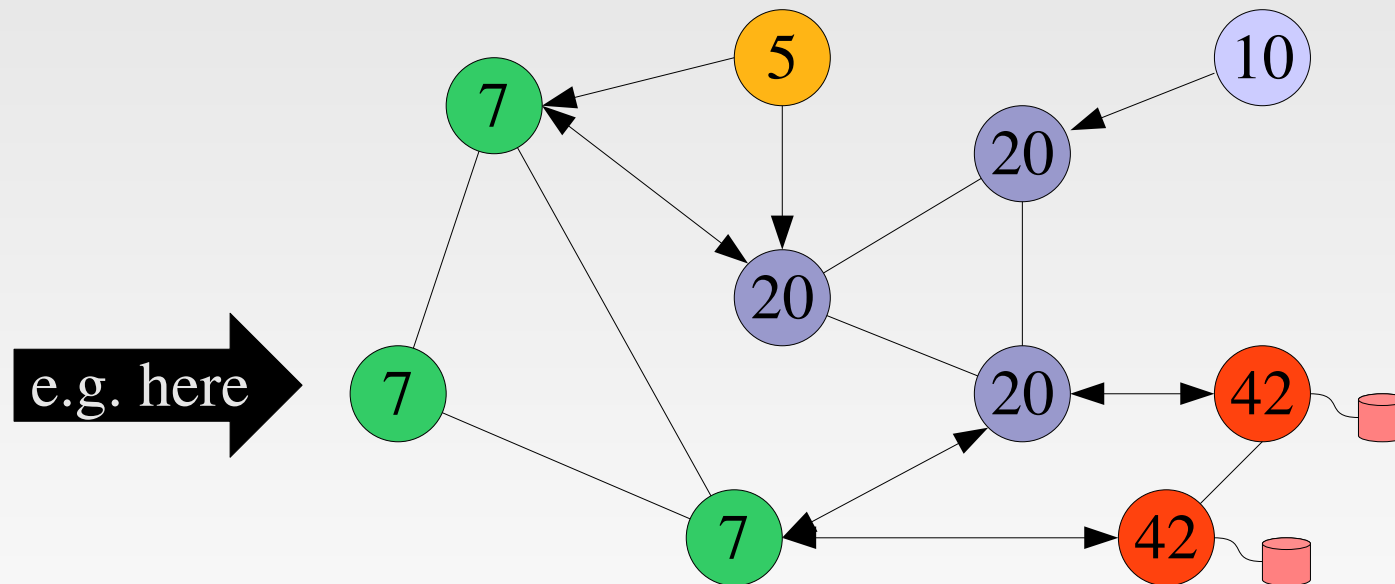
however...

- what happens if the user uses a DNS recursor that is not in his proximity?



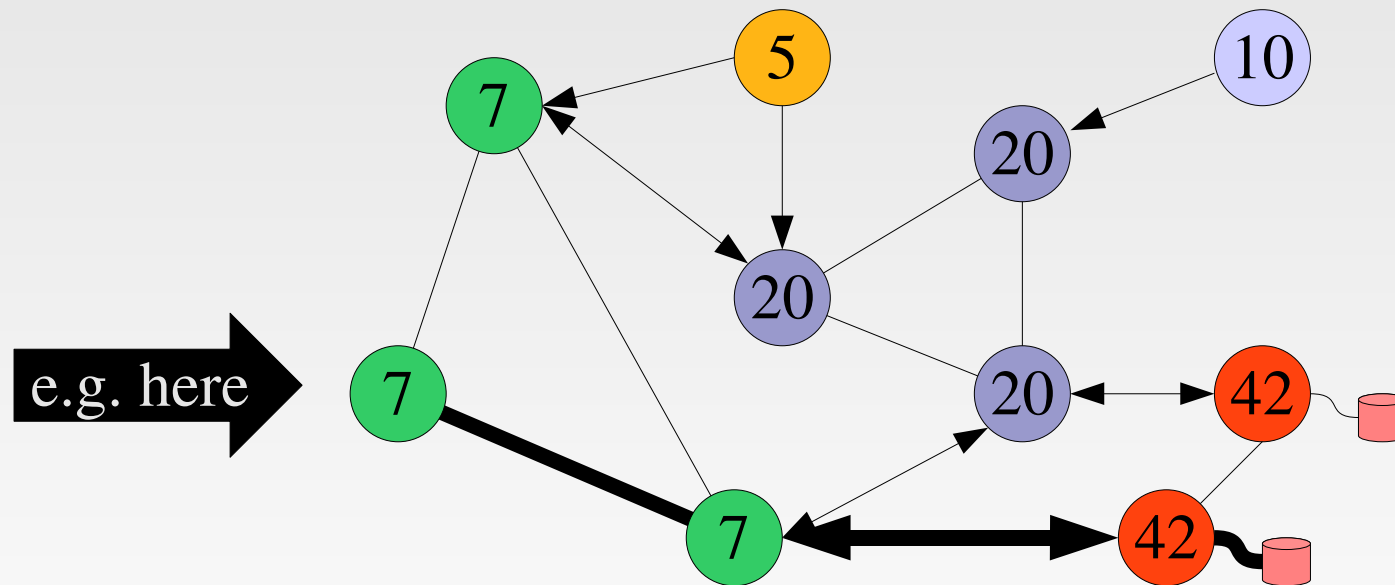
however...

- what happens if the user uses a DNS recursor that is not in his proximity?



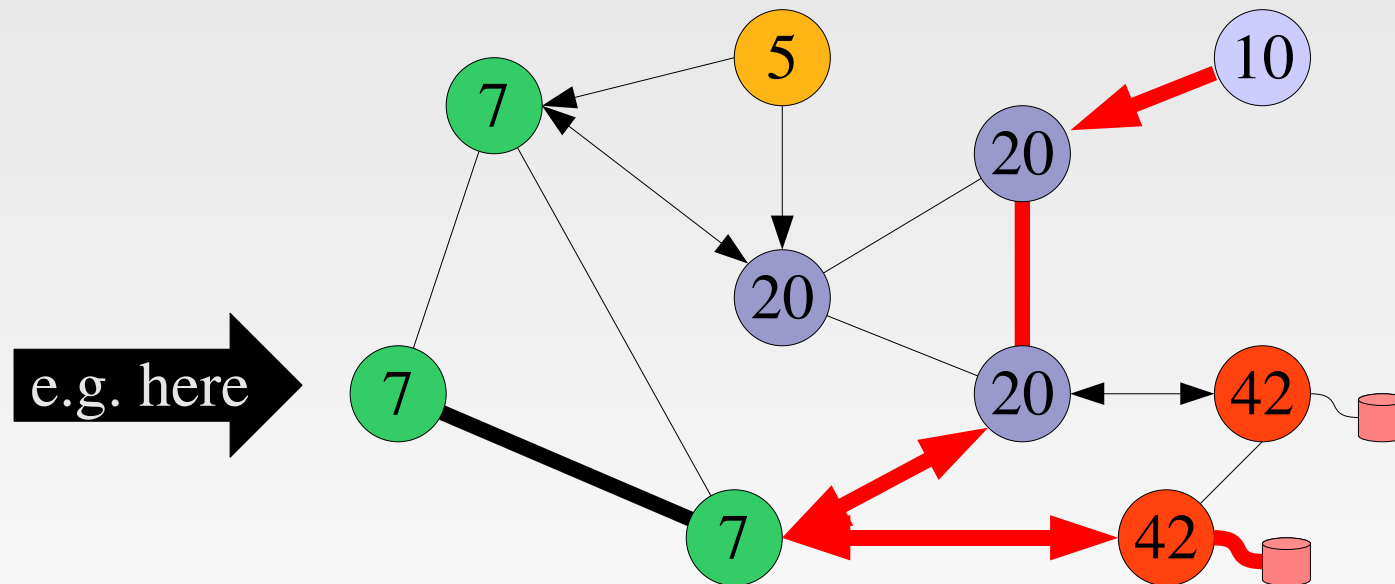
however...

- what happens if the user uses a DNS recursor that is not in his proximity?



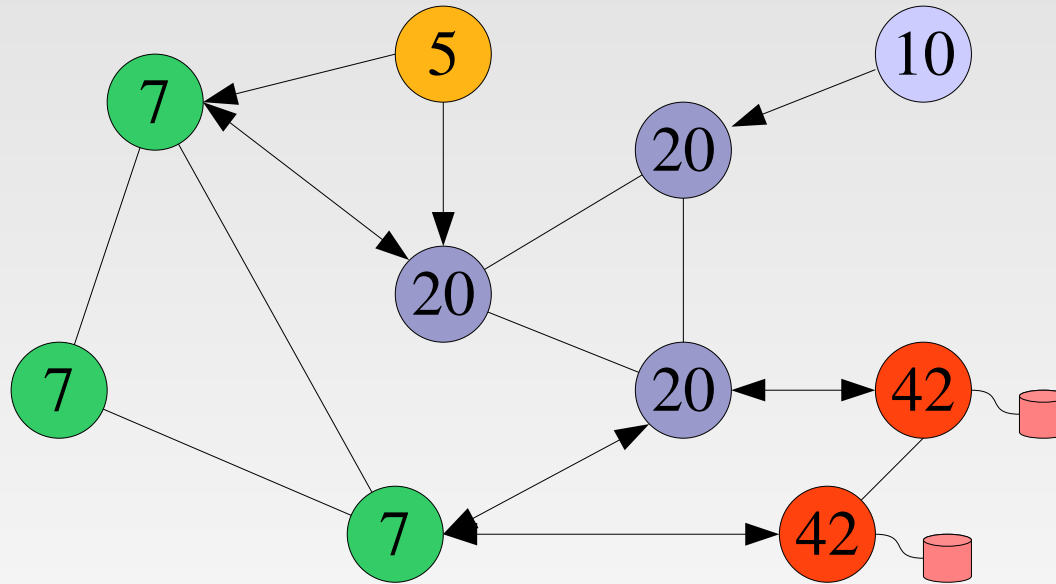
however...

- what happens if the user uses a DNS recursor that is not in his proximity?



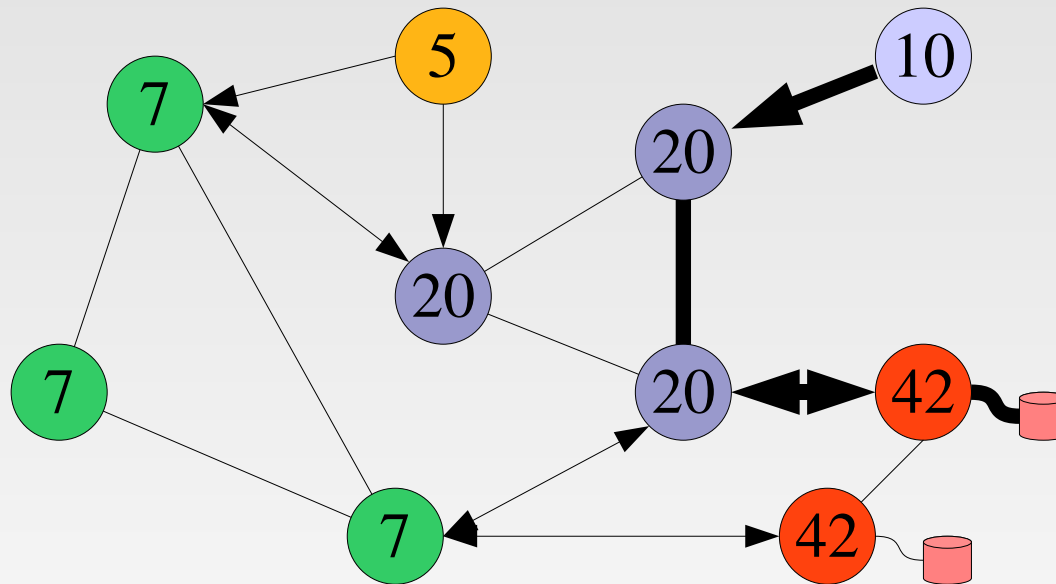
Mitigating the state problem.

- Step 1: Client asks DNS for A/AAAA record.



Mitigating the state problem.

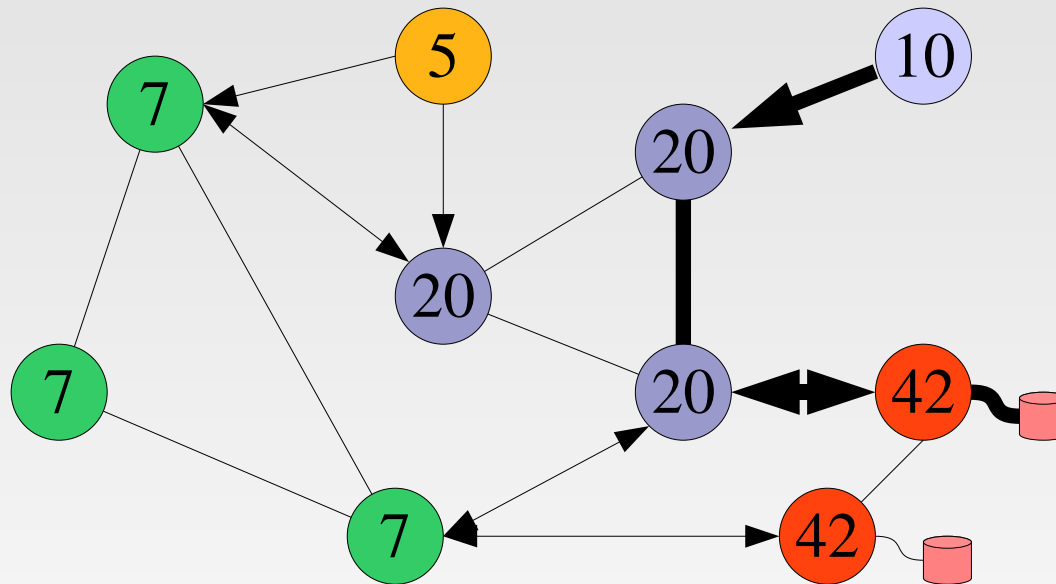
- Step 1: Client asks DNS for A/AAAA record.



- “any” DNS for the service replies with an anycasted IP address. All nodes deliver the same content for the A/AAAA record.

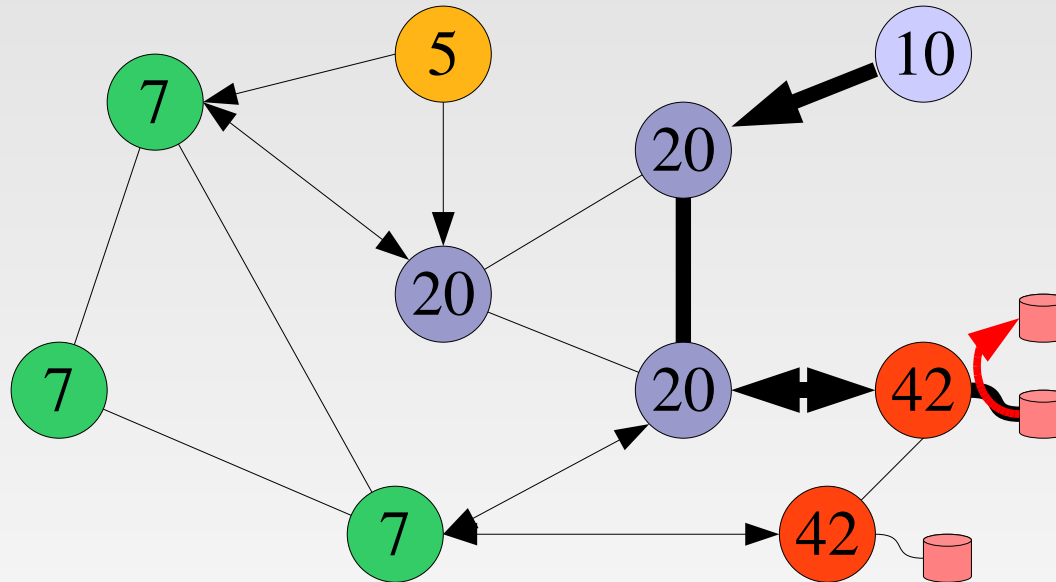
Mitigating the state problem.

- Step 2: Client connects e.g. via HTTP to that host.



Mitigating the state problem.

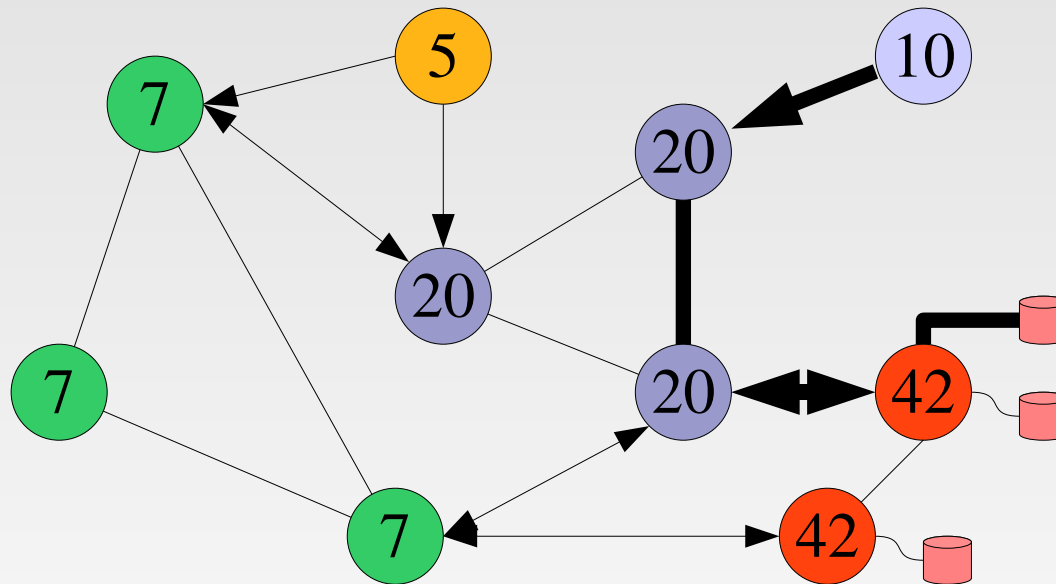
- Step 2: Client connects e.g. via HTTP to that host.



- The host replies with a brief HTTP 302 “temporarily moved” that points to a non-anycasted per-location static server.

Mitigating the state problem.

- Step 3: The server happily serves the content.



how about other protocols?

- HTTP is rather simple to handle...
- how about SIP/RTP?
- challenges for other protocols:
 - divert traffic that should stay node-local to a non-anycasted address using the protocol's features
 - how are database-writes handled?

...I got a final one for you:

- the implementation of this technique with IGPs is left as an exercise to the audience ;-)

That's it.

That's it.

questions?

E-Mail:

nibbler@ccc.de

michael@as250.net

mobile:

+491777761111

xmpp/jabber:

nibbler@jabber.berlin.ccc.de