Towards a deployable IP Anycast service

Hitesh Ballani, Paul Francis
Cornell University
What is Anycast?

- A paradigm for communicating with one member of a group
- A simple concept (with a few corner cases):
  - Ease configuration
  - Improve robustness and efficiency

- Limited wide-area usage: DNS root-servers, .ORG TLD

- What limits the use of such a powerful and promising technique?
Limitations of IP Anycast

- Incredibly wasteful of addresses
  - need a block of 256 addresses even though just one is used

- Scales poorly by the number of anycast groups
  - each group requires an entry in the global routing system

- Difficult to deploy
  - obtain an address prefix and an AS number
  - requires a certain level of technical expertise

- Subject to the limitations of IP routing
  - loss of connection affinity, no notion of load, convergence time

Application-layer anycast offers an attractive alternative!
So why bother?
IP Anycast* has a lot to offer!

- Support for low level services
  - Eg. anycasting to a DNS root server or a IPv6/v4 transition device

- Redresses many problems faced by P2P and overlay technologies
  - Bootstrapping support
  - Efficient querying of DHTs or services built on top of them
  - Efficient injection of packets into overlays

- Accessing web proxies without the need for a DNS query or HTTP redirect

- If a node could be a target and a client
  - Nearby neighbor discovery for P2P Multicast, network games etc.
Proxy IP Anycast (PIAS)

- Only a prefix narrower than a /24 can be inserted into BGP
  - Imagine a set of nodes making up 256 anycast groups
  - Why have the same set of nodes part of 256 groups?
    - They could form a proxy structure for 256 actual anycast groups

![Diagram of Proxy IP Anycast (PIAS)]
What have we solved?

- **Inefficient address space usage**
  - Each address in the address block can identify an anycast group
  - Actually, we can do much better
    - Identify anycast groups using transport addresses (<IP addr, port>)
    - Thousands of groups per IP address in the anycast block
    - Beneficial for scaling by the number of groups

- **Difficult to deploy**
  - Infrastructure operator obtains the address block/AS number
    - Deployment effort amortized across all supported groups
  - User Perspective
    - Registration with a proxy to join an anycast group

- **Transferred scalability and addressing issues**
  - Much easier to solve when isolated from IP routing!
A few details ....

- Scale by the number of groups
  - All proxies cannot keep state for all groups
  - Each group’s membership is tracked by a few designated proxies – **Rendezvous Anycast Proxy (RAP)** for the group

- Scale by group size and group churn
  - Add a tier to the membership management hierarchy
  - **Join Anycast Proxy** – the proxy contacted by the target when it joins the group
  - Feeds approximate numbers of targets associated with it to the group RAPs
Selection at the RAP and JAP allows us to offer high level features such as proximity and load balance.
What about the connection affinity?

- What happens if *native* IP anycast is not sticky?

  - PIAS can be *engineered* to:
    - bear some native IP anycast vagaries
    - provide E2E affinity

- However, this might not be necessary!
  - Measured the affinity offered by IP routing against anycasted DNS root-servers
  - Over 9 days, probed the 6 anycast targets from 40 sources at a probe/minute
    - Probability that a 2 minute connection breaks = 1 in 13000
  - General notion of *affinity* offered by IP anycast seems to be *overly pessimistic*
Implementation and deployment status

- The basic PIAS system has been implemented and tested in the laboratory
  - Comprises of 2 components
    - User space - overlay management tasks
    - Kernel space - tunneling packets between proxies and NAT’ting packets forwarded to the server

- The implementation served as a sanity check for our ideas

- Slow (but steady) progress towards deployment
  - Acquired a /22 and an AS number from ARIN
  - Looking at various deployment possibilities
  - Hopefully, we will soon be able to answer some of the questions that I am going to raise next!
Research issues: large scale anycast is not well studied!

- How good is the proximity offered by native IP anycast?
  - Is the anycast node reached by a client closest node in terms of latency
  - Measured this for the anycasted DNS root-servers
    - Use of hierarchical anycast: no conclusive results
    - Good results for the few cases we know use global anycast

- BGP route flap dampening
  - Large proxy deployment ➔ frequent events ➔ BGP hold downs
  - On the brighter side
    - Short’ish AS paths with a large deployment
    - Hierarchical anycast is claimed to be beneficial!
Research issues (cont.)

- BGP convergence time
  - How to engineer the system so as to achieve fast fail-over?
    - Lots of proxies in a POP
    - Rely on IGP for convergence
    - Reduces the probability of a BGP withdrawal
  - What kind of relation to have with the hosting ISP?
    - Concerns fail-over between proxies in different POPs of the same ISP
    - E-BGP : more route control, slower convergence
    - IGP : less route-control, faster convergence

- Deployment issues
  - Interface/Hooks to the users (the anycast targets)
  - Security???
  - lots more …..
Conclusion

- A ‘practical’ proposal for IP anycast deployment
  - Solves the major problems afflicting *native* IP anycast
  - Slightly weakens the robustness and the nearness property inherent in *native* IP anycast
  - Ensures sufficient robustness while offering advanced features

- Tonnes of research issues
  - Can only be answered through an actual deployment
  - Working on understanding large-scale anycast behavior
  - Hope to have a deployment in the near future!