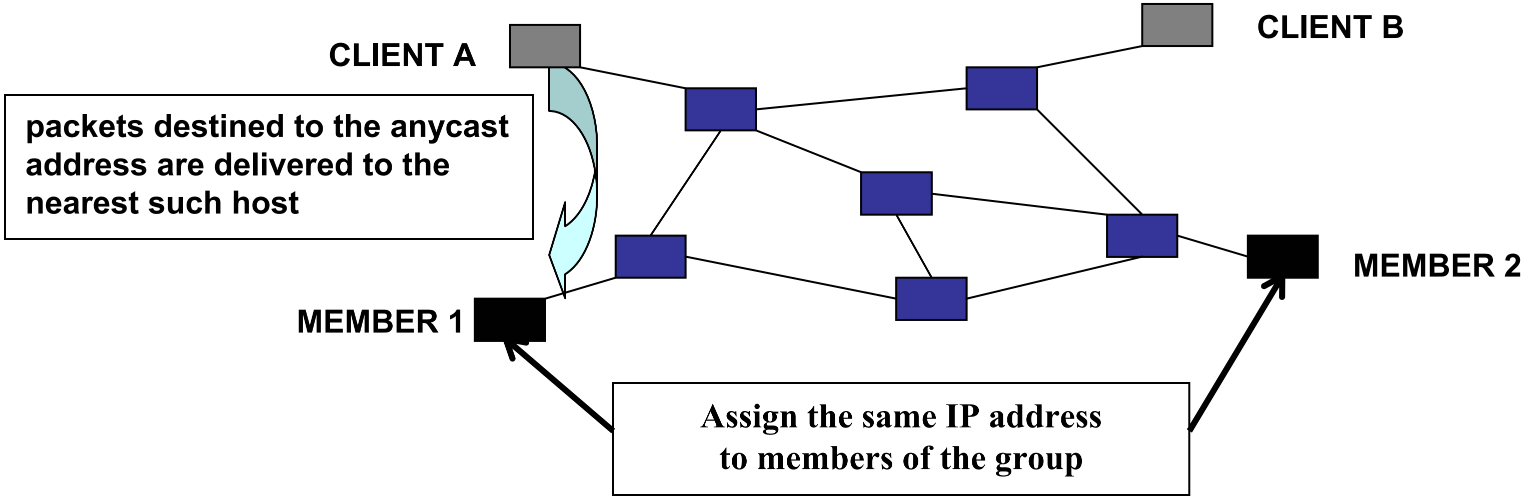
Towards a deployable IP Anycast service

Hitesh Ballani, Paul Francis
Cornell University

{hitesh, francis}@cs.cornell.edu

What is IP Anycast?

- A paradigm for communicating with any member of a group



packets destined to the anycast address are delivered to the nearest such host

Assign the same IP address to members of the group

- Offers a powerful set of tools for service discovery, routing services …
  - Ease configuration
  - Improve robustness and efficiency

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

- 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
  - no notion of load or other application layer metrics, convergence time

Application-layer anycast, typified by DNS-based load balancing, is what current applications such as content distribution make do with!

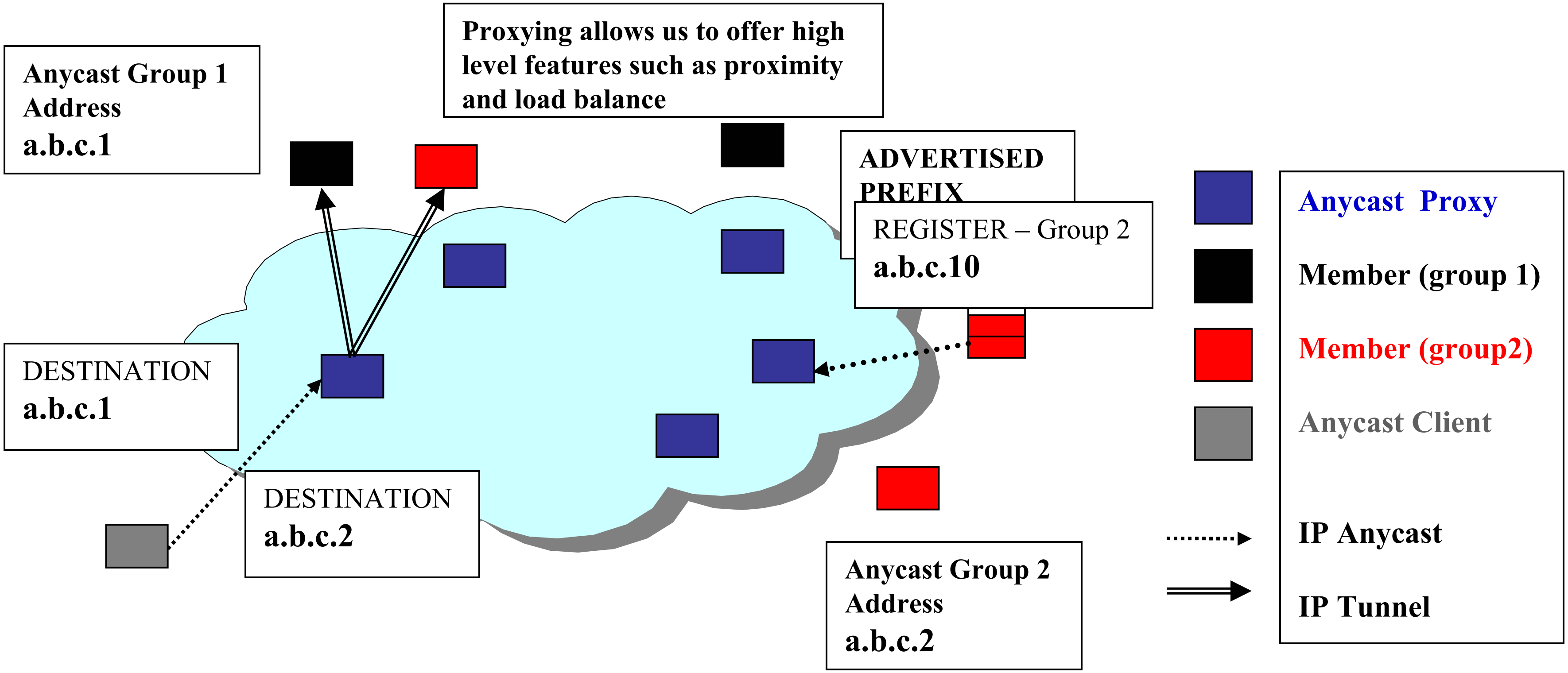
So, why bother?

IP Anycast* has a lot to offer!

- Support for low level services
  - Eg. anycasting to reach a multicast tree or to 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 group member and a client
  - Nearby neighbor discovery for P2P Multicast, network games etc.

Proxy IP Anycast Service (PIAS)

- KEY IDEA : ***Native*** IP Anycast routing is not responsible for delivering anycast packets all the way to the anycast members
  - It delivers the packets to the Anycast Proxies (AP)
  - The proxies forward the packets to the appropriate member



Proxying allows us to offer high level features such as proximity and load balance

What have we solved?

- Efficient address space usage
  - A /24 can potentially support 256 anycast groups
  - Actually, we can do much better
    - Identify anycast groups using transport adresses (<IP addr, port>)
    - Thousands of groups per IP address in the anycast block
    - Beneficial for **scaling by the number of groups**
- Pragmatic deployment model
  - Infrastructure operator obtains the address block/AS number
    - Deployment effort amortized across all supported groups
  - Group member perspective
    - Registration with a proxy to join an anycast group
    - Minimal changes at the server (group member)
    - No changes at the client

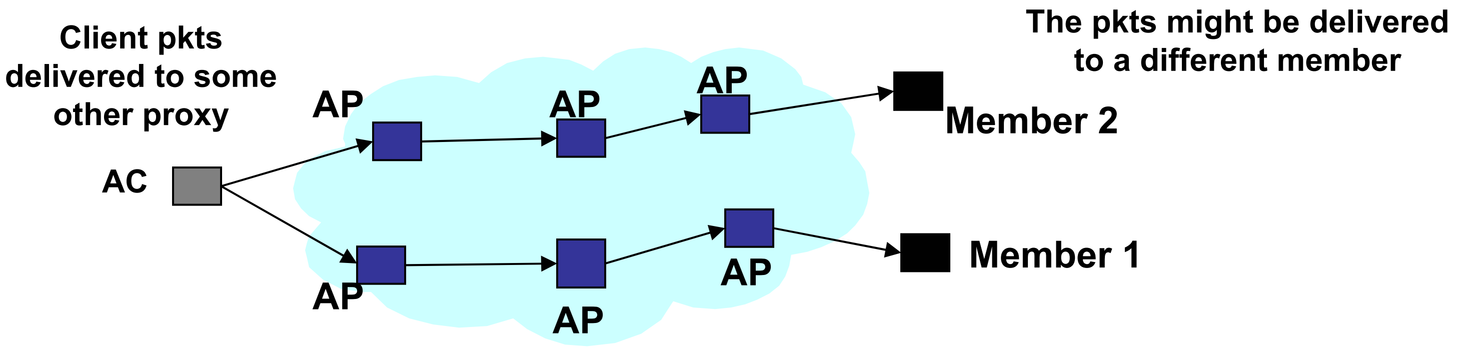
What have we solved? (Cont …)

- Scalability and addressing issues
  - Transferred them from routing to proxy infrastructure
  - Much easier to solve when isolated from IP routing!

- Solving these issues in the proxy infrastructure
  - We have designed the system to address them
  - For eg, scalability by the number of groups
    - every proxy node cannot keep state for every group
    - use consistent hashing to achieve this
  - Other issues
    - scalability by group size
    - scale to groups with high churn
    - efficiency of traversing the proxy infrastructure
  - Details in the paper

What about the connection affinity?

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


Client pkts delivered to some other proxy
The pkts might be delivered to a different member
AC
AP
AP
AP
AP
AP
AP
Member 2
Member 1

- What kind of affinity is offered by ***native*** IP anycast?
  - Measured the affinity offered by IP routing against anycasted DNS root-servers
  - Over 9 days, probed the 6 anycast groups from 40 sources at a probe/minute
    - Probability that a 2 minute connection breaks = 1 in 13000
  - Perceived notion of **lack of affinity** in IP anycast seems to be **overly pessimistic**

- Working on approaches that allow PIAS to:
  - bear some native IP anycast vagaries
  - **provide E2E affinity**

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
- Deployment efforts are underway
  - 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

- Routing issues
  - Minimize routing changes
    - The AS-path for the anycast prefix should be stable
  - Achieve fast fail–over
    - BGP is notorious for high convergence times, in rare cases ~15 minutes
- 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?

Conclusion

- A ‘practical’ proposal for IP anycast deployment
  - Solves the major problems afflicting ***native*** IP anycast
  - Combines the advantages of application layer and native IP anycast
- Next frontier : system deployment
  - Will help us answer the research issues
  - Looking for volunteers who would be interested in supporting the deployment effort and who have ideas for applications which might benefit from such a primitive

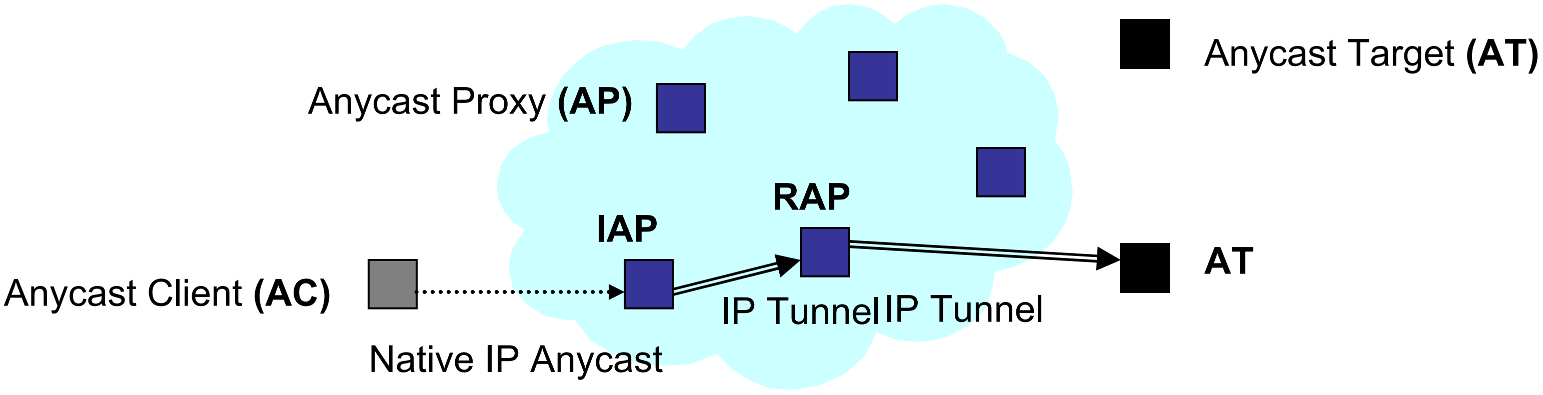
Details : www.cs.cornell.edu/~hitesh/anycast.html

THANKS!

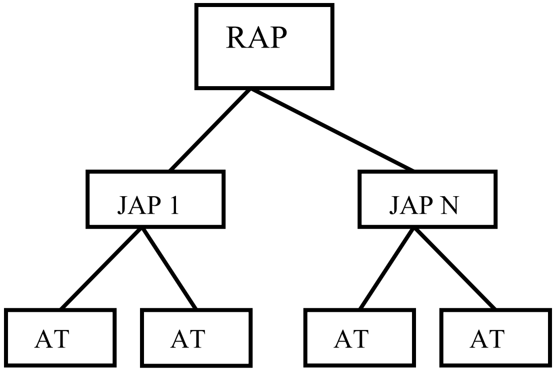
Backup slides!!!

A few details ….

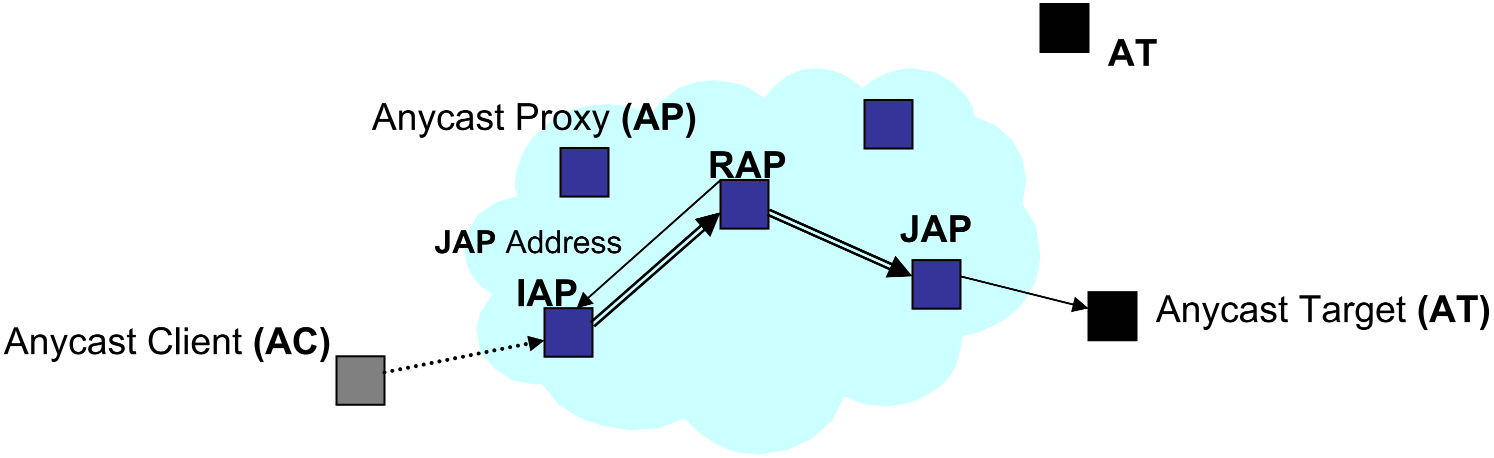
- Scale by the number of groups
  - All proxies cannot keep state for all groups
  - Each group's membership is tracked by a few designated proxies – **Rendezvous Anycast Proxy (RAP)** for the group



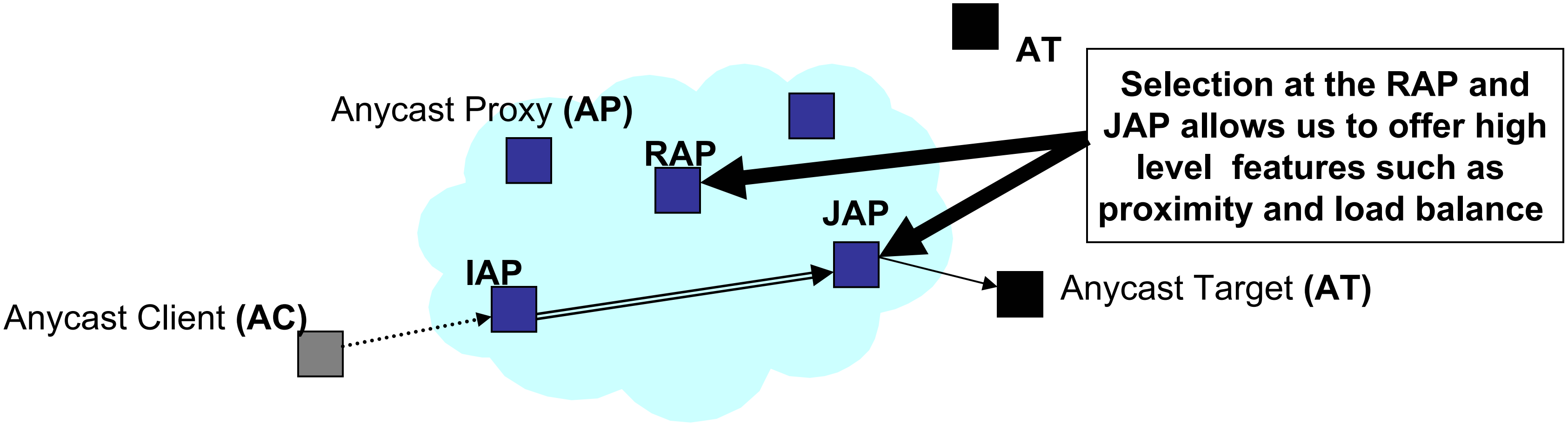
- Scale by group size and group churn
  - Add a tier to the membership management hierarchy
  - **Join Anycast Proxy** – the proxy contacted by the target when it joins the group
  - Feeds approximate number of targets associated with it to the group RAPs



A few details …. (cont.)



INITIAL PACKET PATH – 4 SEGMENTS LONG



Selection at the RAP and JAP allows us to offer high level features such as proximity and load balance

SUBSEQUENT PACKET PATH – 3 SEGMENTS LONG