JPNIC Public Forum

Paul Vixie

Chairman,
Internet Software Consortium

January 21, 2003

- Paul Vixie has been contributing to Internet protocols and UNIX systems as a protocol designer and software architect since 1980. Early in his career, he developed and introduced *sends, proxynet, rtty, cron* and other lesser-known tools. Today, Paul is considered the primary modern author and technical architect of BINDv8 the Berkeley Internet Name Domain Version 8, the open source reference implementation of the Domain Name System (DNS). He formed the Internet Software Consortium (ISC) in 1994, and now acts as Chairman of its Board of Directors. The ISC reflects Paul's commitment to developing and maintaining production quality open source reference implementations of core Internet protocols.
- More recently, Paul cofounded MAPS LLC (Mail Abuse Prevention System), a California nonprofit company established in 1998 with the goal of hosting the RBL (Realtime Blackhole List) and stopping the Internet's email system from being abused by spammers. Vixie is currently the Chief Technology Officer of Metromedia Fiber Network Inc (MFNX.O).
- Along with Frederick Avolio, Paul co-wrote "Sendmail: Theory and Practice" (Digital Press, 1995). He has authored or co-authored several RFCs, including a Best Current Practice document on “Classless IN-ADDR.ARPA Delegation” (BCP 20). He is also responsible for overseeing the operation of F.root-servers.net, one of the thirteen Internet root domain name servers. ⇒ *http://www.isc.org/ISC/vixie.html*

Agenda

1. Internet Software Consortium
   History - Status - Plans
2. Load Balanced DNS Service
3. IPv6 Transport Issues in BIND
4. Q&A

Agenda

1. Internet Software Consortium

   History - Status - Plans

Brief History of DNS and BIND

- DNS introduced in late 1980's to overcome limitations in HOSTS.TXT approach
- BIND4 was released with 4.2BSD
- BIND4.9 was released by DECWRL
- BIND8, BIND9 released by ISC

Brief History of ISC

- Founded by Paul Vixie (after DECWRL) and Rick Adams (using UUNET funding)
- Current board: Paul Vixie (chairman), Evi Nemeth (CAIDA), Teus Hagan (NLNet), Stephen Wolfe (Cisco), Jun Murai (WIDE)
- HQ is Redwood City, California; also has employees in .AU, .CA, .CO.US, .VA.US
- Staff: 10 people (7 technical; 3 admin)
- 2003: Paul Vixie is now full time President

BIND Development Plans

- BIND4 is deprecated (security reasons)
- BIND8 is nearing end of life (architectural stress wrt new protocol features)
  - Even so, we're going to add IPv6 transport
- BIND9 is in production, development:
  - Performance (queries per second) is too low
  - Tracking DNSSEC is very difficult/expensive
  - IPv6 on root servers requires some support

BIND Forum

- Our BSD-style license encourages product bundling, derivative works, and wide use
- System/software vendors, and major DNS operators who depend on our software need a direct, two-way relationship with ISC
  - Secure channel for bug reporting, attacks
  - Forum for discussing feature priorities
  - Way to help prevent code forks (support ISC)

F-root Anycast

- Historically present only at PAIX Palo Alto
  - Used OSPF ECMP for local load balancing
  - Palo Alto: 6 hosts; San Francisco: 3 hosts
- DDoS volume can be >10Gbits/second
- Solution: BGP Anycast
  - BGP Anycast “keeps local attacks local”
  - F is now live in Madrid and Hong Kong
  - Adding 7..10 more cities in 2003
  - Need another 20 cities by the end of 2004

Other ISC Activities

- Hosting: NetBSD, OpenLDAP, XFree86, Kernel.ORG, distributed.net, many others
- Software: INN, DHCP, cron, rtty, OpenReg
- Operations: Usenet Moderated Aliases, Usenet Official Newsgroup List
- Data: Domain Survey
- Coming soon: `dns-isac.org`

Agenda

2. Load Balanced DNS Service

Necessary Distinctions

- Local load balancing
  - Sometimes called clustering
  - Maybe using an appliance
- Distributed load balancing
  - Might just be a diverse set of NS RRs
  - Or it might be that a single NS RR is global
- Policy based (directed) load balancing
  - Different answers in different regions
  - We call this “Stupid DNS tricks” (don’t do it!)

Local Load Balancing (1)

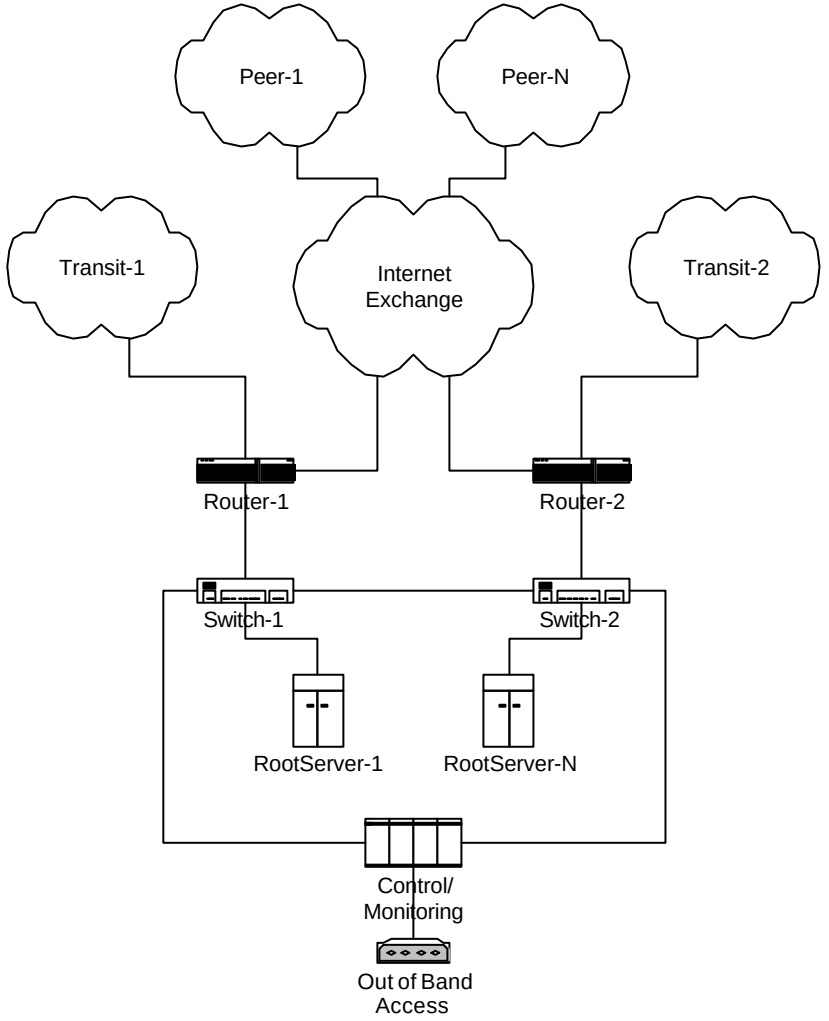
- An L4 switch with health monitoring can distribute query load across a cluster
- This “extra powered box” is a failure point
- Sometimes requires that all TCP land on a single host instance (with fallback)
- Sometimes requires that a single MAC address be used by all cluster members
- This is really the wrong approach

Local Load Balancing (2)

- Using routers and switches that you probably already have in the data path…
- Use GateD/Zebra for host-based OSPF
- Assign a single service address as an “lo0” alias on all members of the cluster
- OSPF “stub host” logic advertises it
- Modern Cisco (CEF) and Juniper (IP-II) routers will do flow hashed load sharing

F-Root

1/19/2003



Distributed Load Balancing (1)

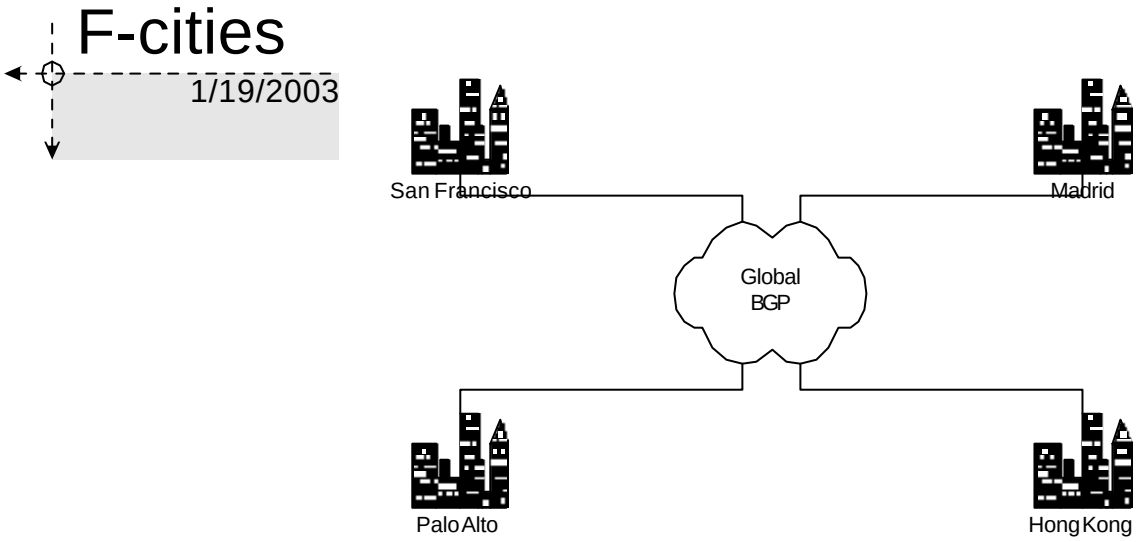
- Core internet routing protocol is BGP, which is loosely distance-vector based
- When multiple paths exist, one is chosen, usually based on AS-path length
- This is not useful for actual load balancing:
  - Geography != Topology
  - Too coarse-grained
  - Depends on other ISP's policies

Distributed Load Balancing (2)

- BGP routes can be tagged “no-export” to ensure that there is no “accidental transit”
  - F-root only has transit at PAIX Palo Alto
- Thereby one can collect traffic from a deliberately restricted part of the topology
- For example, all peers at an exchange point
- This is especially useful for partitioning DDoS attacks and keeping them “local”

Distributed Load Balancing (3)

- Each wide area F-root has its own AS number and its own “management” /24
- The management /24 gets transit from multiple ISPs over private crossover-ether
- The “F” /24 is advertised through the public exchange point, tagged with “no-export”
- Attacks are localized, and do not interfere with network management
- Exchange point fabric means no “DDoS bottleneck”


F-cities
1/19/2003
San Francisco
Madrid
Global
BGP
Palo Alto
Hong Kong

Other Advantages

- Fluidity: add or drop servers or cities at will
  - To upgrade a host or city, drop then add
  - Failures are local and meaningless
  - Add capacity or shift load during attacks
  - Headroom, headroom, headroom!
- Measurement: triangulate on DDoS sources
  - Source routing and source spoofing don’t mix

Agenda

3. IPv6 Transport Issues in BIND

Table of Contents

- Hearing queries sent to an AAAA
- Adding AAAA to a full delegation packet
- Selecting between AAAA and A RR glue
- If you only have one IP stack

Hearing queries sent to an AAAA

- The advanced IPv6 BSD API has no conformance test or conformance pressure
- Listening on only one IPv6 address on a multihomed host is sometimes impossible
- This makes BIND9's "view" feature unusable on some systems
- The fix, so far, is putting pressure on vendors (KAME gets this right, of course!)

Adding AAAA to a full delegation packet

- Classic DNS only allows for 512 octets of UDP payload (IPv4's minimum maximum)
- Root name server response is: a copy of the query, the list of NS RRs, and "glue" RRs
- There's no space for 13 AAAA RRs today!
  - EDNS fixes this, but deployment will take time
- "Prefer AAAA" vs. "Prefer A" feature
  - Today: server selected; Tomorrow: EDNS1?

Selecting between AAAA and A RR glue

- On a dual stack host, a recursive DNS operation can follow both AAAA and A
- If a zone has both, which one to prefer?
- Right now BIND’s resolver prefers AAAA
- Needs to be made configurable

If you only have one IP stack

- On non-dualstack hosts, you might only have IPv6 connectivity, so how to reach a zone which is only served by IPv4 servers?
- For now, we assume that a relay will exist
- New “6to4-servers” configuration option
- “f.6to4-servers.net” is available for use