Cacheboy

An Open-Source Content Delivery Network

Adrian Chadd - Xenion Pty Ltd
<adrian@xenion.com.au>

Overview

- A platform for open CDN development
- Hosting open source/free software content “right”
- Provide a distribution method which is easily deployable in hard to reach places (eg Africa)
- Provide an experimental platform base
- Bridge networks, systems and application layers

Statistics

- .. pushing 400-700mbit aggregate at the moment
- .. around 4TB a day; 30TB a week
- .. 13 node sites, mostly in Europe
- .. 4 DNS sites
- .. around 1 million DNS lookups a day
- .. mirroring a few projects at the present
  - mozilla, videolan

Components

- Operating Systems - FreeBSD; Linux
- Caches - Squid/Lusca
- DNS servers - PowerDNS
- Origin servers - Lighttpd
- BGP edges; BGP core - Quagga
- Monitoring
- Statistics

History

- Pilosoft donated and hosts the management, monitoring and backend storage servers currently in use out of New York
- Thanks to Pilosoft for continued support!
- Started with three major content nodes:
  - Pilosoft - http://www.pilosoft.net/
  - 5Nines - http://www.5ninesdata.com/
  - David Bannister - http://www.mojo.net/

Sponsors

- **USA**
  Pilosoft, Inc.
  5nines Data, Inc. (MADIX)
  Cluetrust, Inc.
  Netwurx
  IRIS Networks, Inc

- **Canada**
  Priority Colo (TORIX)

- **Japan**
  Mozilla Foundation

- **Australia**
  Xenion Pty Ltd (WAIX)

- **Belgium**
  Bart Champagne (PANAP)

- **Sweden**
  David Bannister

- **United Kingdom**
  TENET (LINX)
  UK Broadband
  Alex Smith

- **South Africa**
  TENET (South Africa)

Sponsors

- All equipment is donated
- All bandwidth is donated
- All hosting, colocation fees are donated
- All of Adrian's time is donated (for the most part)
- **There's no revenue being generated for the project by this endeavour**
    - But whatever revenue nodes generate on the traffic exchange is up to them!

Capacity (on good days)

- Canada (100mbit to TOR-IX)
- USA (100mbit + 300mbit + 100mbit)
- Australia (100mbit to WAIX)
- Sweden (1000mbit); UK (1000mbit+300mbit +500mbit); Italy (100mbit); Belgium (100mbit)
- South Africa - TENET - (1000mbit)
- Japan - Mozilla Foundation - (1000mbit)

How it works

- Traffic redistribution is being primarily achieved via DNS "smart replies"
  - .. and all of the issues that entails
- Experimenting with BGP and network "clue"
- Lusca/Squid nodes cache frequently accessed content
  - .. and achieve 99.9% hit rates
- Content is small, but traffic levels are high!

Backend Servers

- Content is kept on large backend servers
- Optimised for fast revalidation of content
- Lighttpd works “fine enough” for this workload (low connection counts, high revalidation count)
- Later on:
  - multiple backend sites around the internet
  - nodes talking to close nodes with content

GeoIP DNS

- Custom PowerDNS backend
- “nerd.dk” GeoIP IPv4 mapping database
- DNS servers map client IP to geoip zone
- mapping of {server, weight} list -> geoip
- “geoip daemon” returns fast geoip lookups; reloads geoip maps in background
- Works “well enough” to distribute traffic

BGP setup

- Quagga running on most nodes
- BGP to ISPs; partial or full tables
- BGP route reflectors - Australia and New York
- Manual configuration at the moment
- No filtering done at the moment!
- More automation needed!
- More flexibility needed - communities, etc

BGP selection

- Currently toying with BGP info for selection
  - Explicit over-rides for certain sites - WAIX, TOR-IX
  - “bgp daemon” handles fast IP lookups, avoiding hacking quagga/openbgpd/etc
  - Uses nexthop lookup in iBGP mesh
  - Feeds into the DNS decision path
- More to do before BGP is used everywhere

Problems with BGP!

- Well, the big problem is BGP + DNS
- Client IP doesn't always exist in the same "network" (ASN, network paths, etc) as the DNS server they use
- So sub-optimal decisions are often made
- Short-term goal: building tools to get statistics on precisely this

Traffic (re)direction

- Some stuff "breaks" when doing naive temporary or permanent HTTP redirection to "better" nodes
  - .. buggy mozilla/firefox download updaters
- Asymmetric networks are not few and far between
  - So "transit" vs "peering" traffic for CDNs start to matter more and more

TCP behaviour

- Can we get away with smaller socket buffers with CDN nodes closer to clients?
- Throughput important? Concurrent client important?
- Can we track "internet goodness" based on sampling TCP behaviour to IPs/networks?
- Any point in "pacing" TCP? (eg delay pools)
- Read the academic literature!

Live traffic patterns

- Mozilla Peak load - 10-40x the average requests sustained over 24-36 hours
- How does one design and run a network to handle this?
  - Over-engineer and be very pragmatic about what is available
  - Shuffle traffic to less-preferred nodes to handle overflow traffic
  - Throttle per-client traffic rates

Node behaviour

- Optimising for throughput versus client counts?
  - Dynamically changing this?
- Building {cost, space, power} efficient nodes:
  - Goal: 100mbit on < $200 hardware
  - Goal: 1000mbit on < $1000 hardware
- Node performance is linked to content distribution!

Storage Engineering

- Current goals -specifically- limits content to be able to focus on other issues
- Building high-throughput servers for TB's of content is out of "reach" of the project at the moment
- Distributing different projects to different cache nodes to maximise caching/locality
- Mozilla "language" downloads on different nodes based on GeoIP/BGP

Global destinations

From Sun Aug 23 00:00:00 2009 to Sun Aug 30 00:00:00 2009

us - 7358609.49 - 22.08%
de - 2486308.45 - 7.46%
jp - 1758617.32 - 5.28%
ca - 1588227.23 - 4.76%
in - 1523049.96 - 4.57%
au - 1440753.45 - 4.32%
id - 1298216.36 - 3.89%
fr - 1277184.75 - 3.83%
it - 1047863.53 - 3.14%
uk - 997535.46 - 2.99%
ph - 993934.75 - 2.98%
br - 712077.28 - 2.14%
es - 567520.71 - 1.70%
ru - 552972.59 - 1.66%
se - 549062.75 - 1.65%
my - 503825.05 - 1.51%
nl - 488462.23 - 1.47%
pl - 426771.39 - 1.28%
mx - 385363.79 - 1.16%
cn - 362017.91 - 1.09%

Global destinations

From Sun Aug 23 00:00:00 2009 to Sun Aug 30 00:00:00 2009

AS3320 - 1085 gbytes - 3.26% - DTAG Deutsche Telekom AG
AS17974 - 946 gbytes - 2.84% - TELKOMNET-AS2-AP PT Telekomunikasi Indonesia
AS7132 - 761 gbytes - 2.28% - SBIS-AS - AT&T Internet Services
AS19262 - 621 gbytes - 1.87% - VZGNI-TRANSIT - Verizon Internet Services Inc.
AS9829 - 473 gbytes - 1.42% - BSNL-NIB National Internet Backbone
AS3215 - 455 gbytes - 1.37% - AS3215 France Telecom - Orange
AS4713 - 440 gbytes - 1.32% - OCN NTT Communications Corporation
AS1221 - 439 gbytes - 1.32% - ASN-TELSTRA Telstra Pty Ltd
AS9299 - 412 gbytes - 1.24% - IPG-AS-AP Philippine Long Distance Telephone Company
AS3269 - 412 gbytes - 1.24% - ASN-IBSNAZ TELECOM ITALIA
AS4788 - 371 gbytes - 1.11% - TMNET-AS-AP TM Net, Internet Service Provider
AS22773 - 323 gbytes - 0.97% - ASN-CXA-ALL-CCI-22773-RDC - Cox Communications
AS9121 - 309 gbytes - 0.93% - TTNET TTnet Autonomous System
AS24560 - 307 gbytes - 0.92% - AIRTELBROADBAND-AS-AP Bharti Airtel Ltd.
AS3209 - 295 gbytes - 0.89% - VODANET International IP-Backbone of Vodafone
AS12322 - 290 gbytes - 0.87% - PROXAD AS for Proxad/Free ISP
AS812 - 282 gbytes - 0.85% - ROGERS-CABLE - Rogers Cable Communications Inc.
AS8151 - 278 gbytes - 0.84% - Uninet S.A. de C.V.
AS3352 - 276 gbytes - 0.83% - TELEFONICA-DATA-ESPANA TELEFONICA DE ESPANA
AS6327 - 274 gbytes - 0.82% - SHAW - Shaw Communications Inc.

Query content served: 9058 gigabytes; 11874433 requests.
Total content served: 33333 gigabytes; 38585579 requests.

WAIX Destinations

From Sun Aug 23 00:00:00 2009 to Sun Aug 30 00:00:00 2009

AS7545 - 123730.91 - 31.95% - TPG-INTERNET-AP TPG Internet Pty Ltd
AS4802 - 84778.92 - 21.89% - ASN-IINET iiNet Limited
AS4739 - 56267.21 - 14.53% - CIX-ADELAIDE-AS Internode Systems Pty Ltd
AS4854 - 16989.37 - 4.39% - NETSPACE-AS-AP Netspace Online Systems
AS9543 - 15451.91 - 3.99% - WESTNET-AS-AP Westnet Internet Services
AS17746 - 12402.16 - 3.20% - ORCONINTERNET-NZ-AP Orcon Internet
AS9443 - 7397.02 - 1.91% - INTERNETPRIMUS-AS-AP Primus Telecommunications
AS9822 - 7390.72 - 1.91% - AMNET-AU-AP Amnet IT Services Pty Ltd
AS7657 - 6414.81 - 1.66% - VODAFONE-NZ-NGN-AS Vodafone NZ Ltd.
AS17435 - 4702.44 - 1.21% - WXC-AS-NZ WorldxChange Communications LTD
AS17412 - 3797.84 - 0.98% - WOOSHWIRELESSNZ Woosh Wireless
AS7543 - 3352.20 - 0.87% - PI-AU Pacific Internet (Australia) Pty Ltd
AS9790 - 2806.56 - 0.72% - CALLPLUS-NZ-AP CallPlus Services Limited
AS9889 - 2703.36 - 0.70% - MAXNET-NZ-AP Auckland
AS24313 - 2532.96 - 0.65% - NSW-DET-AS NSW Department of Education and Training
AS18359 - 1819.89 - 0.47% - H3GA-AP Hutchison 3G Australia
AS24093 - 1740.05 - 0.45% - BIGAIR-AP BIGAIR. Multihoming ASN
AS1221 - 1663.50 - 0.43% - ASN-TELSTRA Telstra Pty Ltd
AS17705 - 1498.02 - 0.39% - INSPIRENET-AS-AP InSPire Net Ltd
AS17808 - 1411.16 - 0.36% - VODAFONE-NZ-AP AS number for Vodafone NZ

Query content served: 358 gigabytes; 165161 requests.
Total content served: 387 gigabytes; 193648 requests.

Short-term goals

- Distributed monitoring/management
- Automated configuration
- Tie in "network quality" into redistribution metrics
- Flesh out BGP network "stuff"
- Acquire ASN/IPv4/IPv6 space to run own services where possible
- **Need more North America nodes!**

Short-term goals (ctd)

- Start fleshing out interfaces for node administrators
- Control BGP decision process via communities (ie, "standard stuff")
- API to node admins - real time control
  - set bandwidth per-AS/per-network
  - gather per-AS/per-network statistics in real-time

Medium-term goals

- Add other content services
  - Eg streaming audio/video via VLC
  - Eg flash content
- Add other content
  - Eg mirroring *NIX distributions, etc
- .. both of which require more hardware and solve different problems
- Better statistics and analysis tools!

Long-term goals

- Build some Anycast service(s) as a platform for others to be able to experiment
- Bring up IPv6 support!
- Push (more) cache nodes into the third world
- Begin pushing "cloud" edge processing nodes into the network to deploy applications
  - For experimentation, **not** competition!

Open Source?

- It'll all be eventually 100% open sourced
- Most likely under Affero GPL licence
  - Closing the "ASP Loophole"
- Alternatives may be made available for-fee
- More data will be published as gathered
- Updates - http://cacheboy.blogspot.com/
- Ideas/suggestions are welcome!

Thanks!

- Project - http://www.cacheboy.net/
- Blog - http://cacheboy.blogspot.com/
- Personal - http://www.creative.net.au/
- Company - http://www.xenion.com.au/
- Adrian Chadd <adrian@xenion.com.au>