Building a global PKGng CDN with ZFS for PCBSD

http://pkg.cdn.pcbsd.org

Allan Jude

Goals and Requirements

PCBSD faced a number of problems:

- Two package sets (-RELEASE and -STABLE), each updated twice a month
  - mirrors kept rsync server pegged at 100mbps 24/7 for 2-3 weeks after each update
- Repositories must be updated atomically
- Manual mirror selection is cumbersome and restrictive (mirrors go away/get stale)

Solving the Problems

- Problem: Package sets take too long to get to the mirrors
- Solution: Replace rsync master with beefier server with 1Gbps pipe

- Problem: Mirrors take too long to converge
- Solution: Replace donated mirrors with CDN nodes with 1Gbps pipes

Atomic Updates

This is accomplished in two ways

1. rsync --delay-update
   - Files are uploaded to .~tmp~/ rather than overwriting existing files. Only once ALL files are uploaded are these new files renamed into place
   - rsync updates symlinks too soon (apparently not a bug, just unexpected behaviour)
   - Solved by doing 2 passes, first skips links
2. ZFS snapshots and replication
   - Master snapshots every 15 minutes
   - Slaves replicate on staggered schedule

ZFS Replication

```
zxfer -dFkv -g 375 \
    -o readonly=on \
    -D 'bar -s %%size%% -bl 1m -bs 256m'
    -O "-i /usr/home/pcbsd/.ssh/id_rsa \
    -oPort=1122 -oTcpRcvBuf=2560 \
    -oNoneEnabled=yes -oNoneSwitch=yes \
    pcbsd@pcbsd-master.scaleengine.net" \
    -R zstore/m/pcbsd/pkg zstore/m/pcbsd
```

ZFS Replication System Elements

- Uses security/openssh-portable for newer SSHd with NONE Cipher enabled
- Uses patched sysutils/zxfer that supports delta estimate, progress bar, ETA etc. https://github.com/allanjude/zxfer/
- Uses zfs allow to accomplish all replication without root access (requires vfs.usermount)
- Uses sysutils/zfs-snapshot-mgmt to create and age snapshots

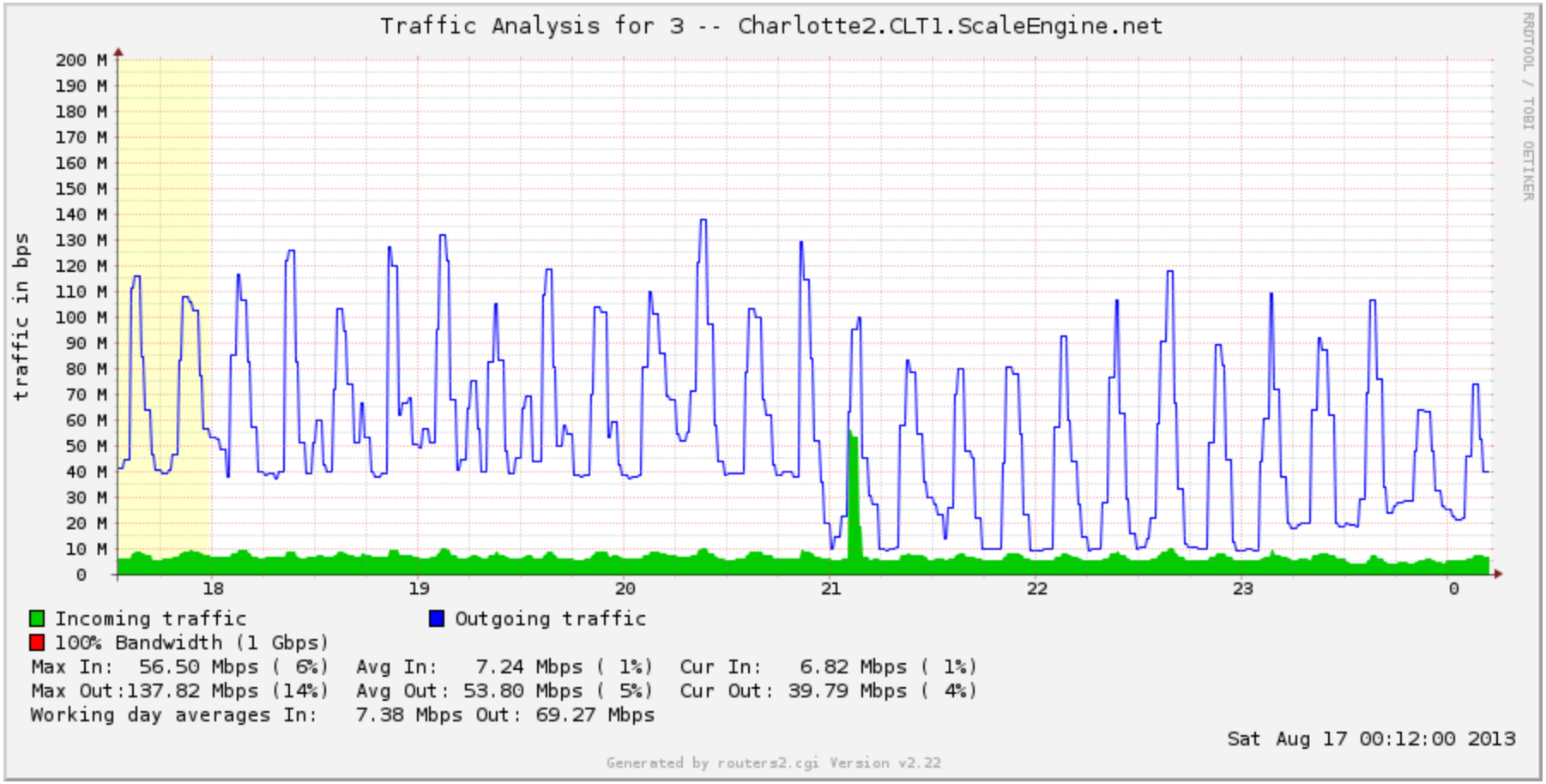
zxfer

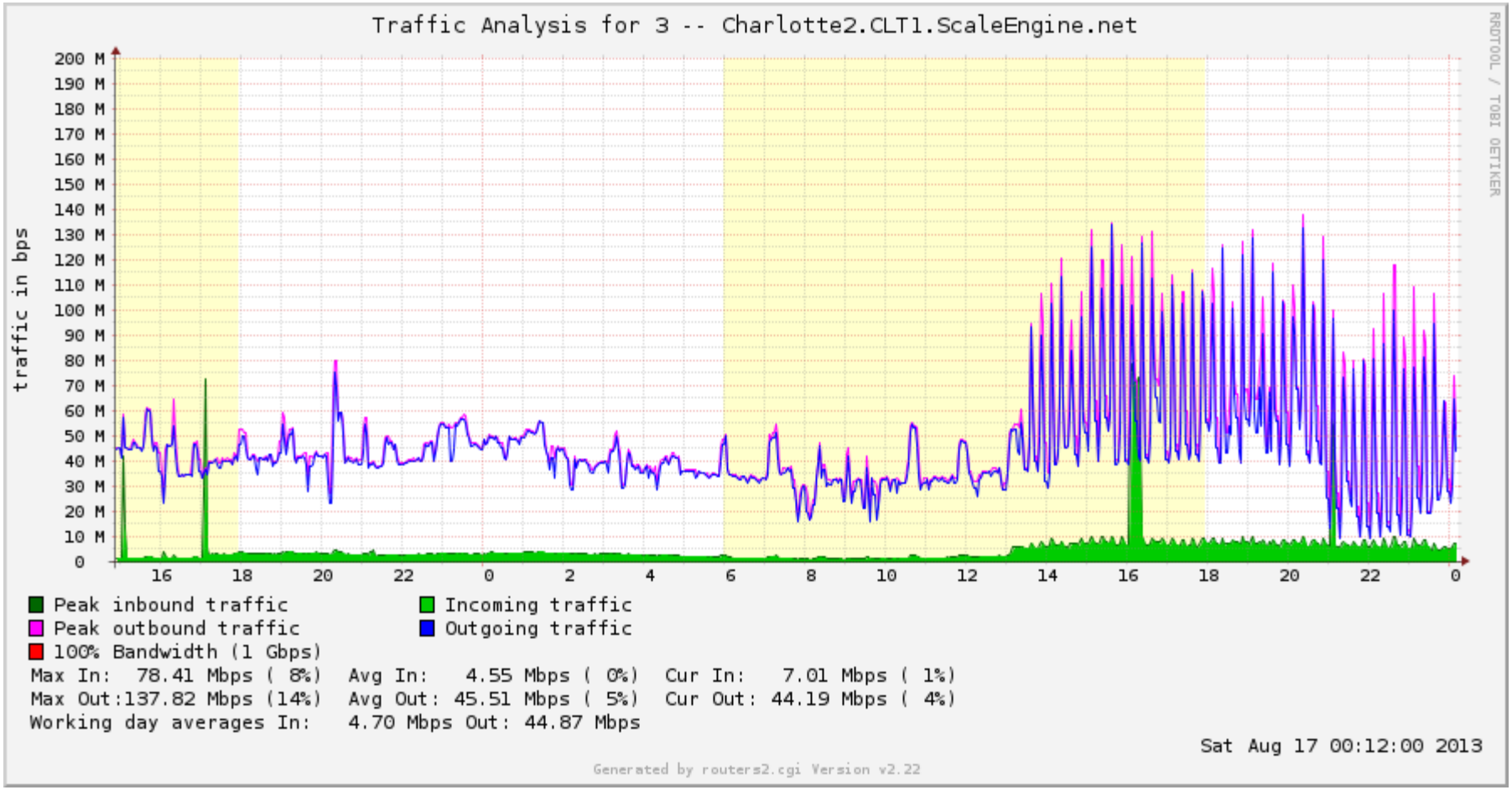
- Handy shell script by:
  - Constantin Gonzalez
  - Ivan Nash Dreckman
- Lives at http://code.google.com/p/zxfer/
- Last updated for FreeBSD 8.2
- Fails on later version of ZFS due to new read-only properties that it tries to replicate
- Missing some error checking
- I've created a github fork of it, that solves these issues and am adding new features
- Hope to eventually integrate fdpv

```
#zfs-snapshot-mgmt.conf
snapshot_prefix: auto-
filesystems:
  zstore/m:
    recursive: true
    creation_rule:
      at_multiple: 15
      offset: 0
# Keep all snapshots for the first 240 minutes (4 hours)
# those created at 60 min intervals for 24 hours
# then keep 12h intervals for 7 days
    preservation_rules:
     - { for_minutes:   240, at_multiple:0, offset:   0 }
     - { for_minutes:   720, at_multiple:   60, offset:    0 }
     - { for_minutes: 10080, at_multiple:  720, offset:0 }
```

What Replication Looks Like

- Kris uploads from the buildfarm at his house
- He only has ~5mbps of upstream
- Snapshot every 15 minutes is replicated to edge servers on a staggered schedule
- Because our master and edges are on full 1Gbps pipes they replicate what Kris can upload in 15 minutes in under 30 seconds
- Because ZFS is replicating the partial files, then rsync renames them, edges converge 15 minutes after Kris has finished uploading


Traffic Analysis for 3 -- Charlotte2.CLT1.ScaleEngine.net
traffic in bps
200 M
190 M
180 M
170 M
160 M
150 M
140 M
130 M
120 M
110 M
100 M
90 M
80 M
70 M
60 M
50 M
40 M
30 M
20 M
10 M
0
18
19
20
21
22
23
0
RRDTOOL / TOBI OETIKER
Incoming traffic
Outgoing traffic
100% Bandwidth (1 Gbps)
Max In: 56.50 Mbps (6%) Avg In: 7.24 Mbps (1%) Cur In: 6.82 Mbps (1%)
Max Out:137.82 Mbps (14%) Avg Out: 53.80 Mbps (5%) Cur Out: 39.79 Mbps (4%)
Working day averages In: 7.38 Mbps Out: 69.27 Mbps
Sat Aug 17 00:12:00 2013
Generated by routers2.cgi Version v2.22


Traffic Analysis for 3 -- Charlotte2.CLT1.ScaleEngine.net
traffic in bps
200 M
190 M
180 M
170 M
160 M
150 M
140 M
130 M
120 M
110 M
100 M
90 M
80 M
70 M
60 M
50 M
40 M
30 M
20 M
10 M
0
16
18
20
22
0
2
4
6
8
10
12
14
16
18
20
22
0
RRDTOOL / TOBI OETIKER
Peak inbound traffic
Peak outbound traffic
100% Bandwidth (1 Gbps)
Incoming traffic
Outgoing traffic
Max In: 78.41 Mbps (8%) Avg In: 4.55 Mbps (0%) Cur In: 7.01 Mbps (1%)
Max Out:137.82 Mbps (14%) Avg Out: 45.51 Mbps (5%) Cur Out: 44.19 Mbps (4%)
Working day averages In: 4.70 Mbps Out: 44.87 Mbps
Sat Aug 17 00:12:00 2013
Generated by routers2.cgi Version v2.22

Automatic Mirror Selection

- ScaleEngine Global Server Load Balancer
- dns/gdnsd with MaxMind GeoIP Database
- Monitors mirrors for freshness and health
- Attempts to always return more than 1 IP
- Uses EDNS0-Client-Subnet to get IP of requestor from recursive nameservers
- See my talk from EuroBSDCon 2012
- http://youtu.be/WF75lGx9svM

Conclusions

- ZFS makes it easier to replicate atomically
- If a zfs send/receive is interrupted, it is rolled back (also zfs receive resuming is on the road map for the future)
- zfs-snapshot-mgmt makes and ages snapshots automatically
- Can take manual snapshots, or use zfs hold
- zxfer manages incremental transfers, removes snapshots that have been deleted
- zxfer-patch gives progress bar with ETA

Statistics

PCBSD

- 2013-07: 19 TiB
- 2013-08: 30 TiB
- 2013-09 Projected: 36 TiB

pkg.cdn: ~85 GiB/day (surge on update)

iso.cdn: ~1 TiB/day

FreeNAS

- 2013-08: 14 TiB
- 2013-09 Projected: 16 TiB