LiveJournal's Backend

A history of scaling

August 2005

Brad Fitzpatrick

brad@danga.com

danga.com / livejournal.com / sixapart.com

http://www.danga.com/words/

MogileFS

- our distributed file system
- open source
- userspace
  - started on FUSE port, lost interest
- hardly unique
  - Google GFS
  - Nutch Distributed File System (NDFS)
- production-quality

MogileFS: Why

- alternatives at time were either:
  - closed, non-existent, expensive, in development, complicated, ...
  - *scary/impossible when it came to data recovery*
- because it was easy

MogileFS: Main Ideas

- MogileFS main ideas:
  - files belong to classes
    - classes: minimum replica counts
  - tracks what disks files are on
    - set disk's state (up, temp_down, dead) and host
  - keep replicas on devices on different hosts
    - Screw RAID! (for this, for databases it's good.)
  - multiple tracker databases
    - all share same MySQL database cluster
  - big, cheap disks
    - dumb storage nodes w/ 12, 16 disks, no RAID

MogileFS components

- clients
- trackers
- mysql database cluster
- storage nodes

MogileFS: Clients

- tiny text-based protocol
- Libraries available for:
  - Perl (us)
    - tied filehandles
  - Java
  - PHP
  - Python?
  - porting to $LANG is be trivial
- doesn't do database access

MogileFS: Tracker

- interface between client protocol and cluster of MySQL machines
- also does automatic file replication, deleting, etc.

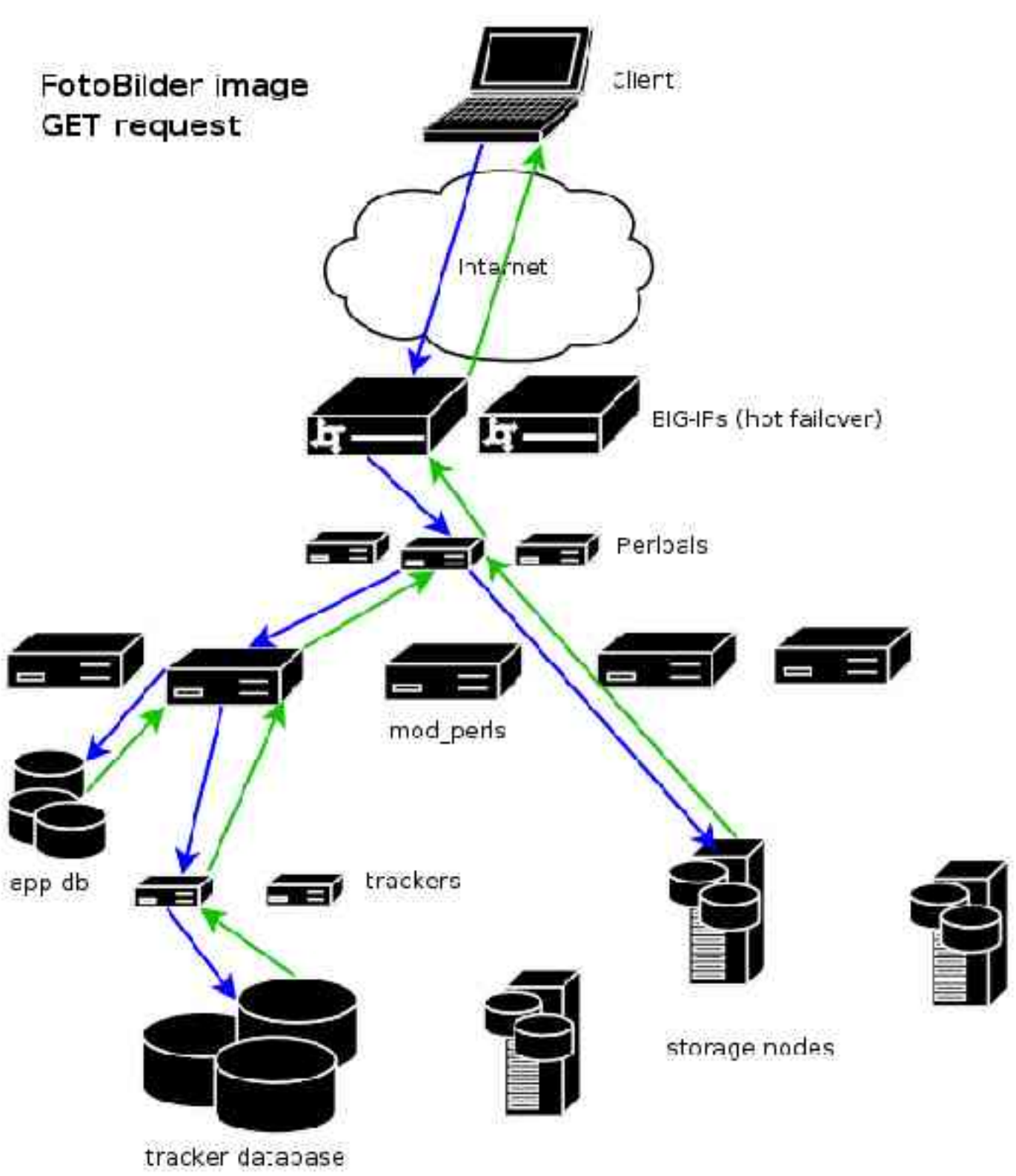
MySQL database

- master-slave or, recommended: MySQL on shared storage (DRBD/etc)

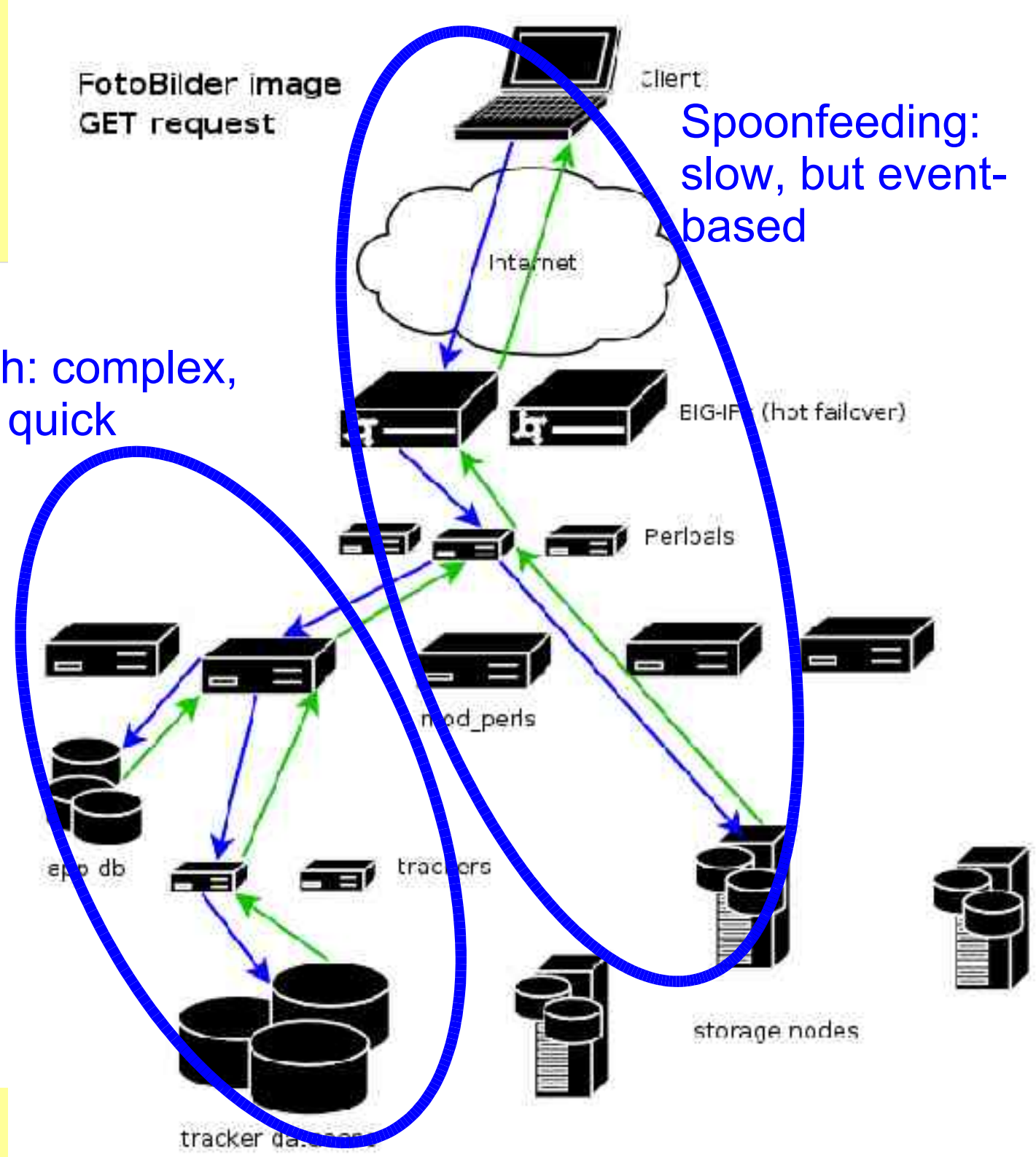
Storage nodes

- NFS or HTTP transport
  - [Linux] NFS *incredibly* problematic
- HTTP transport is either:
  - Perlbal with PUT & DELETE enabled
    - “mogstored” wrapper just does “use Perlbal;” and sets up config for you
  - Apache with WebDAV
- Stores blobs on filesystem, not in database:
  - otherwise can't sendfile() on them
  - would require lots of user/kernel copies
  - filesystem can be any filesystem

Large file GET request



Large file GET request



Auth: complex, but quick

Spoonfeeding: slow, but event-based

And the reverse...

- Now Perlbal can buffer uploads as well..
  - Problems:
    - LifeBlog uploading
      - cellphones are slow
    - LiveJournal/Friendster photo uploads
      - cable/DSL uploads still slow
  - decide to buffer to “disk” (tmpfs, likely)
    - on any of: rate, size, time
  - Big Ups to Mark “Junior” Smith

Thank you!

Questions to...
brad@danga.com

We're Hiring!
http://www.sixapart.com/jobs/