Federation at Flickr: Doing Billions of Queries Per Day

Dathan Vance Pattishall

flickr

Who am I ?

- Name: Dathan Vance Pattishall
- Job: I’m the Flickr Database guy
- Things I do: A little bit of everything.

flickr GAMMA

Contents

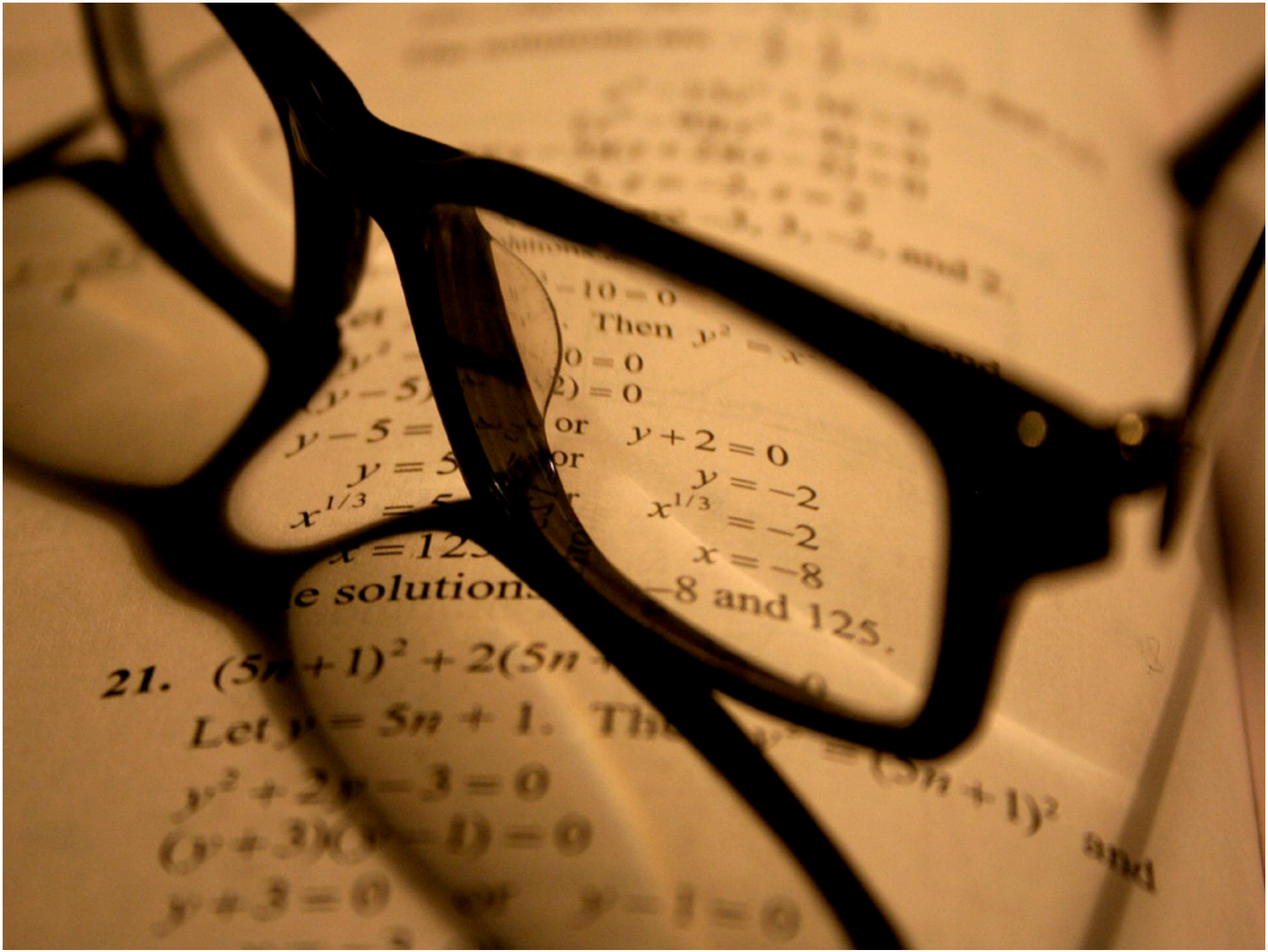
- What Flickr was Having Problems with
- Design to fix problems
- History
- Solution & Design decisions
- Stats
- Wrap Up

flickr GAMMA

What Problems was Flickr Having?

- Master Slave Topology
- Slave Lag
- Multiple SPOF
- Unable to keep up with demand
- Unable to Serve Search Traffic
- Multiple Second page load times

flickr GAMMA


Then
or
or
$y + 2 = 0$
$y = -2$
$x^{1/3} = -2$
$x = -8$
21.

Design to attain Goal

- Since write intensive, need more then 1 master, need many write points.
- To get rid of SPOFs - be redundant.
- To allow maintenance real-time, traffic needs to stick to servers, and ‘a’ server needs to be able to handle the all traffic.
- To serve pages fast with many queries need small data that fits in memory.

flickr

History

- 1999 AuctionWatch
  - ACP in a BOX
- 2003 Friendster
  - Project S00K
- 2005 Flickr
  - Federation

flickr


Federation
http://flickr.com/photos/53898331@N00/217564728/

Federation Key Components

- Shards
- Global Ring
- PHP logic to connect to the shards and keep the data consistent

flickr


Shards
http://flickr.com/photos/may_jon/254120146/

Shards

- Shards are a slice of a main database
- Shards are set up in Active Master-Master Ring Replication
  - Done by sticking a user to a server in a shard
  - Shard assignments are from a random number for new accounts
  - Migration is done from time to time
  - Can run on any hardware grade

flickr GAMMA

Global Ring

- a.k.a Lookup Ring
  - For stuff that can't be federated
  - Like where stuff is

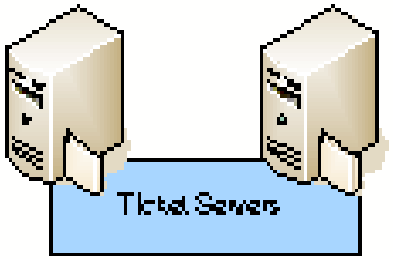
  Owner_id → SHARD-ID

  Photo_id → Owner_id

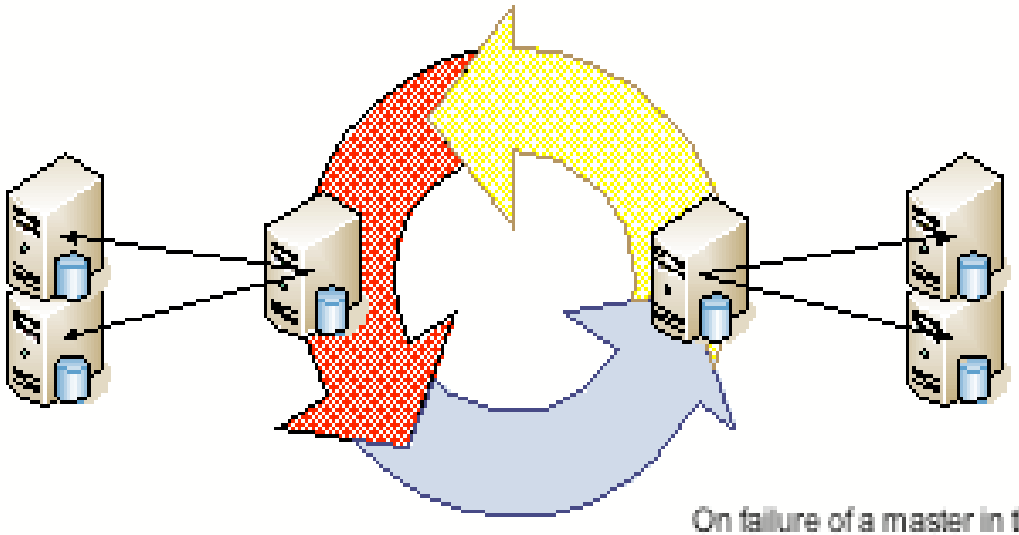
  Group_id → SHARD-ID

  Then cached in MEMCache (lasts ½ hour)

Flickr 2.0



- Circular Replication
- Each master in the ring has replication slaves
- Keep a user on a master to write to
- Randomly pick a slave to read from of the master



On failure of a master in the ring

1. Write to the master that is up
2. Those slaves are only read from
3. Sync up later when server out of the mix is recovered

flickr

what you dont know can indeed hurt you

A FAVE BLOG THIS ALL SIZES

for those interested in knowing a little more about me, i was interviewed the other day by Chris, you can check that out here

This photo has notes. Move your mouse over the photo to see them.

Comments

angelferd pro **says:**

you have talent, no doubt about it
saludos!
=D

Posted 17 months ago. (permalink)

TAugie [deleted] says:

I think these latest pictures are great...very creative. Would love to see more of your body

Posted 17 months ago. (permalink)

Uploaded on September 29, 2005
by _rebekka

_rebekka's photostream

734 photos
View as slideshow

more | browse | more

This photo also belongs to:

portfolio (Set)

50 photos
View as slideshow

more | browse | more

conceptual artsy type stuff (Set)

96 photos
View as slideshow

more | browse | more

400+ faves (Set)

31 photos
View as slideshow

more | browse | more

200+ faves (Set)

flickr

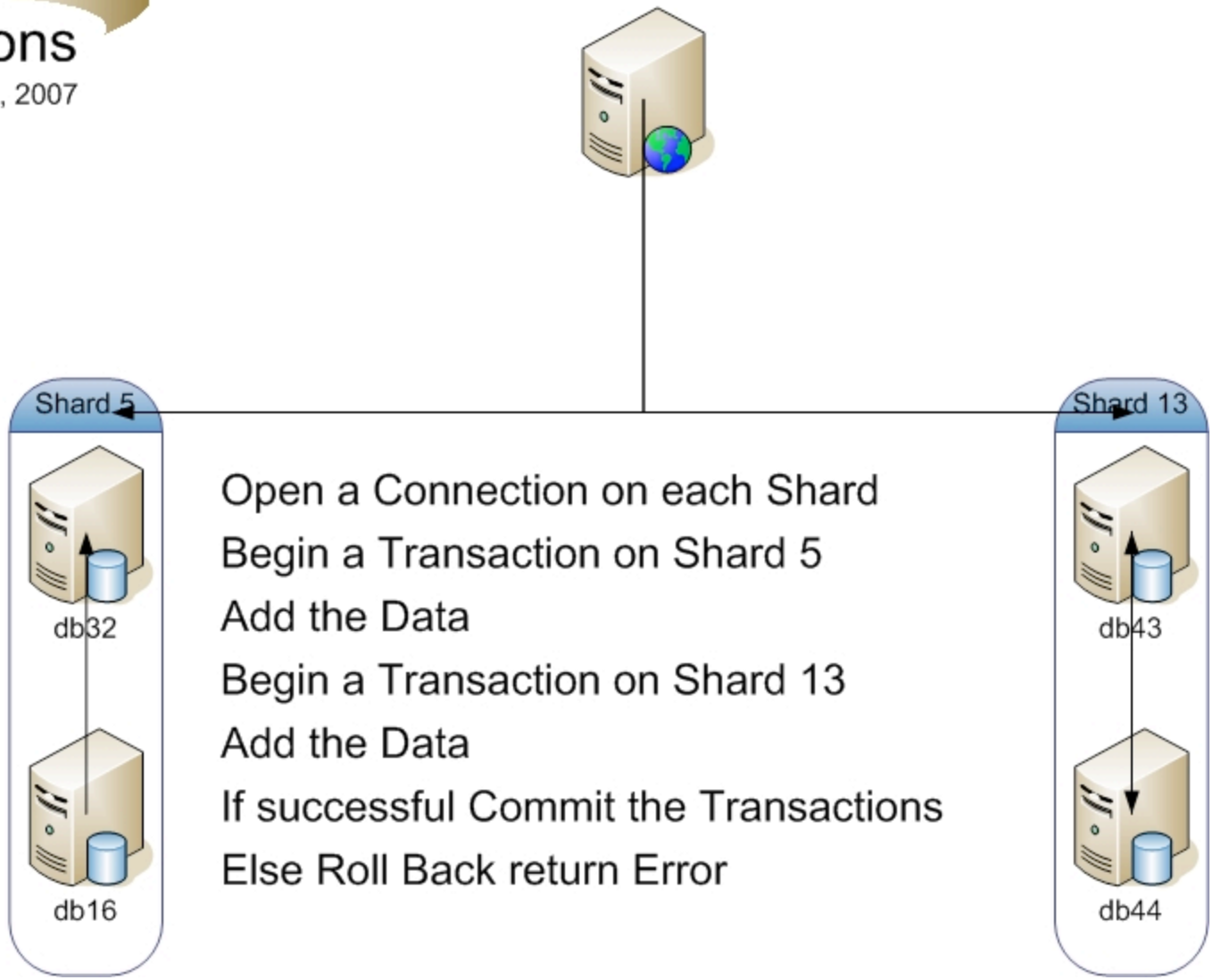
Clicking a Favorite

- Pulls the Photo owners Account from Cache, to get the shard location.
  - SHARD-5
- Pulls my Information from Cache, to get my shard location
  - SHARD-13
- Starts a “Distributed transaction”
  - To answer the question:
    - Who favorited Rebekka’s Photo?
    - What are Dathan’s Favorites

flickr

Transactions

Friday, February 09, 2007



Open a Connection on each Shard
Begin a Transaction on Shard 5
Add the Data
Begin a Transaction on Shard 13
Add the Data
If successful Commit the Transactions
Else Roll Back return Error

Getting Rid of Replication Lag

- On every page load the user is assigned to a bucket

```
$id = intval(substr($user_id, -10));
$id % $count_of_hosts_in_shard
```

- If host is down, go to the next host in the list
- If all hosts are down, display error page

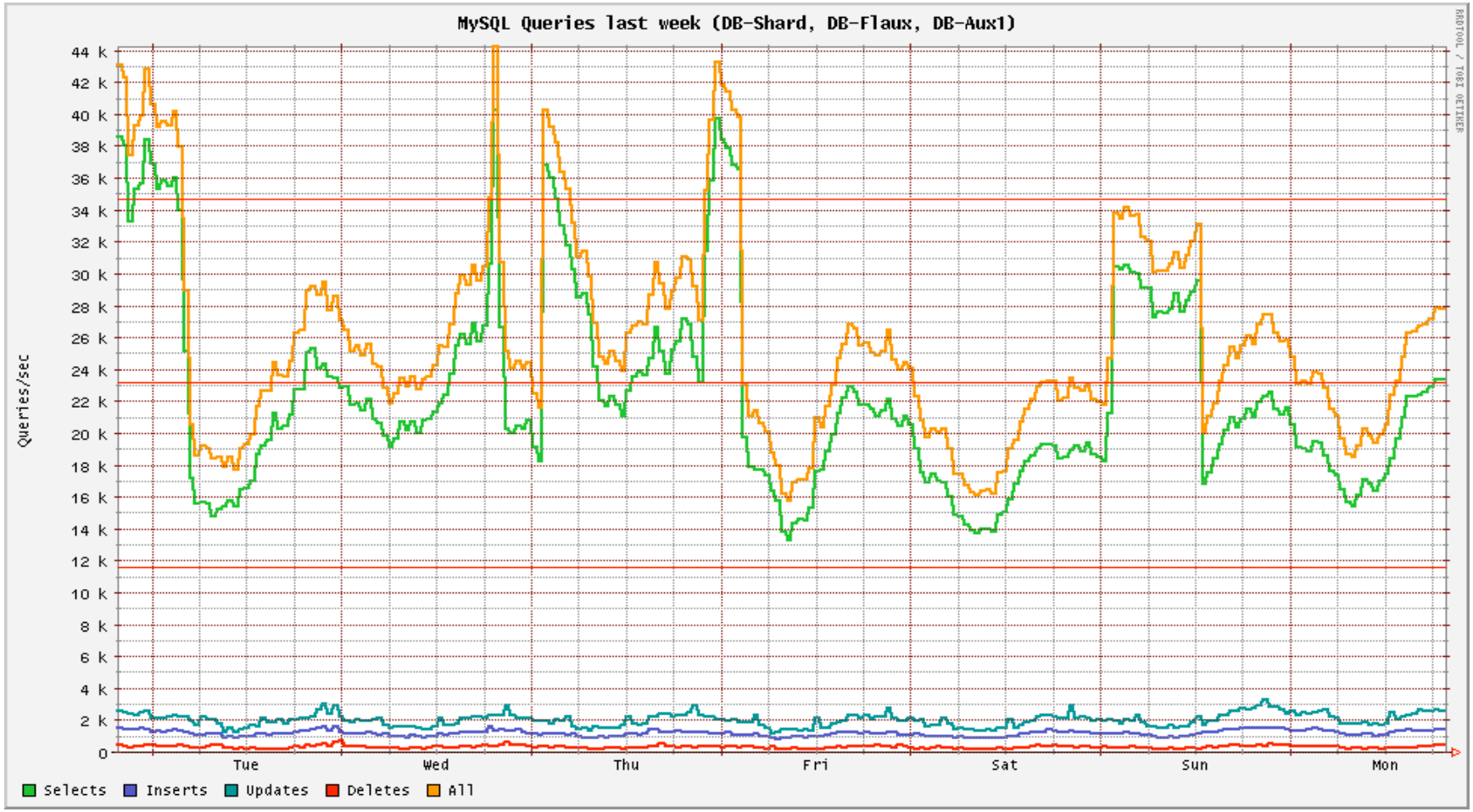
flickr

Allow for maintenance

- Each server in a Shard is 50% loaded
  - i.e. 1 server in a shard can take the full load if a server of that shard is down or in maintenance mode
- Shut down ½ the servers in each Shard
- Do the DBA thing
- Bring them up do the other ½

flickr

Stats


MySQL Queries last week (DB-Shard, DB-Flaux, DB-Aux1)
Queries/sec
44 k
42 k
40 k
38 k
36 k
34 k
32 k
30 k
28 k
26 k
24 k
22 k
20 k
18 k
16 k
14 k
12 k
10 k
8 k
6 k
4 k
2 k
0
Tue
Wed
Thu
Fri
Sat
Sun
Mon
Selects
Inserts
Updates
Deletes
All

Stats continued

- Running within capacity threshold
  - Over 36K per second
- Allow for Burst of Traffic up to
  - Over double 36K per second
- Add more Shards grow
  - 2 to 6K queries / second shard added
- Each Shard holds 400K+ users data
- 1 person (me) handles it all

flickr

What about Search

- Two search back-ends.
  - Shards 35K qps on a few shards
  - Yahoo proprietary websearch
- Owner single tag search goes to the Shards due to real-time requirements
- All other search goes to a Yahoo Search backend.

flickr

Hardware

- EMT64 Running RHEL-4 with 16GB of RAM and 6 DISK 15K RPM RAID-10
- Any class server can be used, users per shard just has to be adjusted.
- Cost per Query = soooo little it should be considered 0.
- Data Size: 12 TB of user data
- Able to serve the traffic: Priceless

flickr

If have time share some quick tips

- Swapiness set to 0
- In RHEL-4 (i.e. 2.6 Linux Kernel) run DEADLINE I/O scheduler
- Use RAID-10
- Use Battery Back cache
- Use 64-bit
- Use Memory lots of it, but leave enough for the OS to spawn many threads
- Tune Queries

Things that would Help Flickr

- Multi-Master replication
- Thread Bug in INNODB for 4.1 release already
- Optimization for OR queries, Bitwise, Sets, etc.

flickr

Questions?

flickr