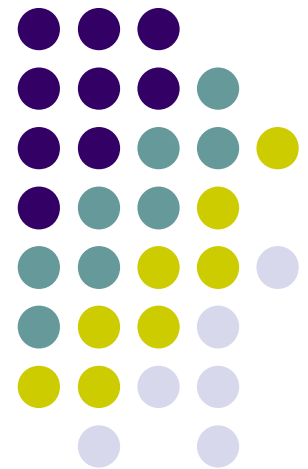


Yet Another Rails

Scaling Presentation



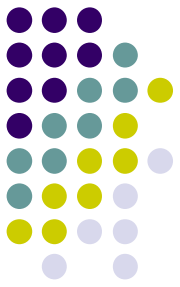
Ruby on Rails Meetup

May 10, 2007

Jared Friedman (jared@scribd.com) and

Tikhon Bernstam (tikhon@scribd.com)

Should you bother with scaling?



- Well, it depends
- But if you're launching a startup, probably
- The best way to launch a startup these days is to get it on TechCrunch, Digg, Reddit, etc.
- You don't get as much time to grow organically as you used to
- You only get one launch – don't want your site to fall over



The Predecessors

- Other great places to look for info on this
- pooocs.net The Adventures of Scaling Rails

<http://pooocs.net/2006/3/13/the-adventures-of-scaling-stage-1>

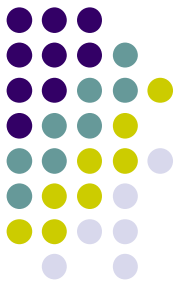
- Stephen Kaes “Performance Rails”

<http://railsexpress.de/blog/files/slides/rubyenrails2006.pdf>

- RobotCoop blog and gems

<http://www.robotcoop.com/articles/2006/10/10/the-software-and-hardware-that-runs-our-sites>

- O’reilly book “High Performance MySQL”
 - It’s not rails, but it’s really useful



Big Picture

- This presentation will concentrate on what's different from previous writings, not a comprehensive overview
- Available at <http://www.scribd.com/blog>



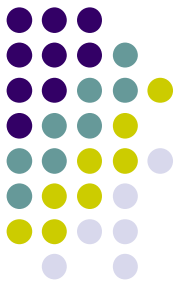
Who we are

- Scribd.com
- Like “YouTube for documents”
- Launched in March, 2007
- Handles ~1M requests per day



Key Points

- General architecture
- Use fragment caching!
- Rolling your own traffic analytics and some SQL tips



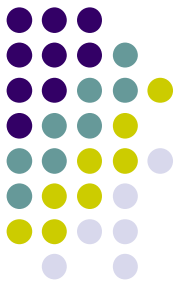
Current Scribd architecture

- 1 Web Server
- 3 Database Servers
- 3 Document conversion servers
- Test and backup machines
- Amazon S3



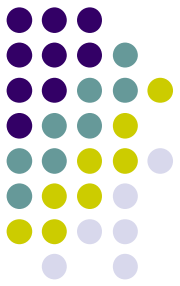
Server Hardware

- Dual, dual-core woodcrests at 3GHz
- 16GB of memory
- 4 15K SCSCI hard drives in a RAID 10
- We learned: disk speed is important
- Don't skimp; you're not Google, and it's easier to scale up than out
- Softlayer is a great dedicated hosting company



Various software details

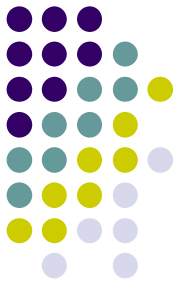
- CentOS
- Apache/Mongrel
- Memcached, RobotCoop's memcache-client
- Stefan Kaes' SQLSessionStore
 - Best way to store persistent sessions
- Monit, Capistrano
- Postfix



Fragment Caching

- "We don't use any page or fragment caching." - robotcoop
- "Play with fragment caching ... no improvement, changes were reverted at a later time." - poocs.net
- Well, maybe it's application specific
- Scribd uses fragment caching extensively, enormous performance improvement

ScreenShot



Scribd Home Browse Explore My docs My stats Profiles

Logged in as snowmaker (19,284) | (8 new) | Logout

Put your docs online. Scribd docs have been viewed 8,241,818 times. Scribd's mission is to create the world's largest open library of documents. Explore the thousands of docs already uploaded or contribute your own!

Upload Bulk upload now! No need to sign up. Formats: .pdf, .doc, .ppt, .xls, .txt, etc. Select multiple files with ctrl/cmd or shift key

Example document: [A Beginner's Guide To BitTorrent](#)

Cached for 2 minutes →

Today's popular documents

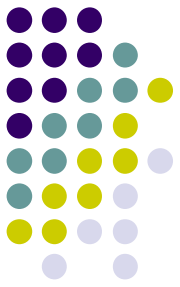
1. [\(ebook - pdf\) how to draw - female body's](#)
4,570 views - uploaded by [pj](#)
2. [David Blaine - Street Magic Revealed](#)
1,639 views - uploaded by [Truth](#)
3. [Best Company Picture Ever](#)
157 views - uploaded by [funfan](#)
4. [What to do on a first date \(If you're female\)](#)
154 views - uploaded by [Meta](#)
5. [Cristina Aguilera - Maxim](#)
519 views - uploaded by [vranghel](#)
6. [School Lunch Ideas](#)
581 views - uploaded by [evilknewit](#)
7. [The Art Of Writing](#)
131 views - uploaded by [damullan](#)
8. [Former Marijuana Smuggler Seeks Legitimate Employment \[image\]](#)
2,231 views - uploaded by [Job](#)
9. [a pdf of ink drawings \(part 2\)](#)
99 views - uploaded by [bitpic](#)
10. [Loving Freely](#)
158 views - uploaded by [dougflowd](#)
11. [\(ebook-pdf-guide\) a guide to memory increase](#)
407 views - uploaded by [pj](#)
12. [How to Make Political Cartoons with a Computer](#)
333 views - uploaded by [anon-602009](#)
13. [Frases de Aristoteles](#)
839 views - uploaded by [stampo](#)
14. [Burroooo](#)
6 views - uploaded by [anon-867701](#)
15. [Ferrari](#)
13 views - uploaded by [EXDE601E](#)
16. [Veyron a Molsheim](#)
18 views - uploaded by [EXDE601E](#)
17. [Mazda](#)
10 views - uploaded by [EXDE601E](#)
18. [Das Joint Drehbuch](#)
4 views - uploaded by [Herrengedeck](#)

Recent Documents

1. [SG KE REUTERS China, India to lead luring green, renewable energy, projects by 2012](#) uploaded 3 minutes ago by [GaneshSrinivasan](#)
2. [SG KE REUTERS Old media turns combative against new media](#) uploaded 3 minutes ago by [GaneshSrinivasan](#)
3. [SG KE REUTERS Internet encyclopedia to list all 1.8 million species](#) uploaded 3 minutes ago by [GaneshSrinivasan](#)
4. [SG KE REUTERS Thomson in talks to buy Reuters for \\$17 billion](#) uploaded 3 minutes ago by [GaneshSrinivasan](#)
5. [SG KE Openness key to growth - IBM chief](#) uploaded 4 minutes ago by [GaneshSrinivasan](#)

FAQ Blog Feedback Contact Jobs Privacy Terms Copyright © 2007

How to Use Fragment Caching

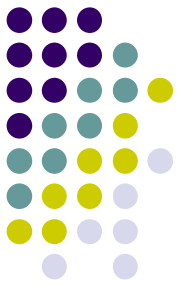


- Ignore all but the most frequently accessed pages
- Look for pieces of the page that don't change on every page view and are expensive to compute
- Just wrap them in a

```
<% cache('keyname') do %>
```

...

```
<% end %>
```
- Do timing test before and afterwards; backtrack unless significant performance gains
- We see > 10X



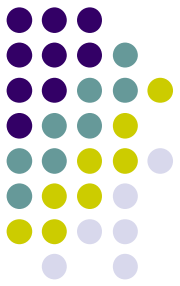
Expiring fragments, 1. Time based

- You should really use memcached for storing fragments
 - Better performance
 - Easier to scale to multiple servers
 - Most important: allows time-based expiration
- Use plugin http://agilewebdevelopment.com/plugins/memcache_fragments_with_time_expiry
- Dead easy:

```
<% cache 'keyname', :expire => 10.minutes do %>
```

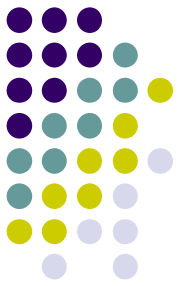
```
...
```

```
<% end %>
```



Expiring fragments, 2. Manually

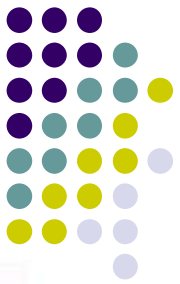
- No need to serve stale data
- Just use:
`Cache.delete("fragment:/partials/whatever")`
- Clear fragments whenever data changes
- Again, easier with memcached



Traffic Analytics

- Google Analytics is nice, but there are a lot of reasons to roll your own traffic analytics too
 - Can be much more powerful
 - You can write SQL to answer arbitrary questions
 - Can expose to users

Scribd's analytics (screenshots)



Traffic Analytics:

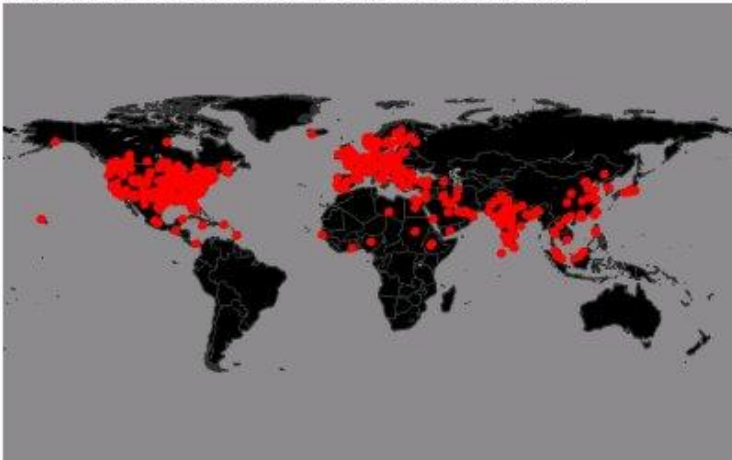
Most recent views:

- 98302. Tue May 08 21:47:33 CDT 2007 : by you
- 98301. Tue May 08 21:47:09 CDT 2007 : Albion, IL (USA)
- 98300. Tue May 08 21:46:43 CDT 2007 : Seattle, WA (USA)
- 98299. Tue May 08 21:45:27 CDT 2007 : Montreal, QC (Canada)
- 98298. Tue May 08 21:44:52 CDT 2007 : Vienna, VA (USA)

[Hide analytics](#)

[Map of users](#) [Page views](#) [Unique users](#) [View log](#)
[Referers](#) [Search engines](#) [Download CSV](#)

Click and drag to zoom. Mouse-over points for more info.



Too many dots to show! Showing latest 2000.

Traffic Analytics:

Most recent views:

- 98302. Tue May 08 21:47:33 CDT 2007 : by you
- 98301. Tue May 08 21:47:09 CDT 2007 : Albion, IL (USA)
- 98300. Tue May 08 21:46:43 CDT 2007 : Seattle, WA (USA)
- 98299. Tue May 08 21:45:27 CDT 2007 : Montreal, QC (Canada)
- 98298. Tue May 08 21:44:52 CDT 2007 : Vienna, VA (USA)

[Hide analytics](#)

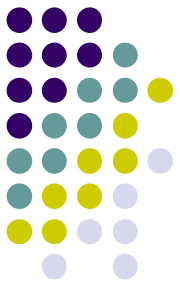
[Map of users](#) [Page views](#) [Unique users](#) [View log](#)
[Referers](#) [Search engines](#) [Download CSV](#)

Your document seems to be indexed by the following search engines

- Google (search), been here 139 times
- Yahoo, been here 157 times
- MSN, been here 49 times
- Ask Jeeves, been here 20 times
- Alexa, been here 7 times
- Sohu.com Chinese search engine, been here 1 time
- Baidu.com Chinese search engine, been here 1 time
- Goo.ne.jp Japanese search engine, been here 2 times
- Larbin multi-purpose web-crawler, been here 8 times
- Snap.com graphical search engine, been here 1 time
- Java-based (multiple sources), been here 5 times
- GenieKnows.com health search, been here 1 time
- Adobe Flash Player, been here 1 time
- Blank User-Agent, probably a mean bot, been here 44 times
- <http://www.voila.com> search engine, been here 3 times
- <http://www.tailrank.com> blog aggregator, been here 9 times
- Google (adsense), been here 4 times
- Bot for Qihoo.com - Chinese search engine, been here 4 times

Your document has not yet been indexed by the following bots





Building traffic analytics, part 1

- `create_table "page_views" do |t|
 t.column "user_id", :integer
 t.column "request_url", :string, :limit => 200
 t.column "session", :string, :limit => 32
 t.column "ip_address", :string, :limit => 16
 t.column "referrer", :string, :limit => 200
 t.column "user_agent", :string, :limit => 200
 t.column "created_at", :timestamp
end`
- Add a whole bunch of indexes, depending on queries



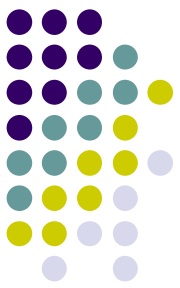
Building traffic analytics, part 2

- Create a PageView on every request
- We used a hand-built SQL query to take out the ActiveRecord overhead on this
- Might try MySQL's "insert delayed"
- Analytics queries are usually hand-coded SQL
- Use "explain select" to make sure MySQL is using the indexes you expect



Building Traffic Analytics, part 3

- Scales pretty well
- BUT analytics queries expensive, can clog up main DB server
- Our solution:
 - use two DB servers in a master/slave setup
 - move all the analytics queries to the slave



Rails with multiple databases, part 1

- "At this point in time there's no facility in Rails to talk to more than one database at a time." - Alex Payne, Twitter developer
- Well that's true
- But setting things up yourself is about 10 lines of code.
- There are now also two great plugins for doing this:
Magic multi-connections

http://magicmodels.rubyforge.org/magic_multi_connections/

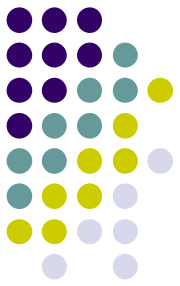
Acts as read onlyable-

http://rubyforge.org/frs/?group_id=3451



Rails with multiple databases, part 2

- At Scribd we use this to send pre-defined expensive queries to a slave
- This can be very important for dealing with lock contention issues
- You could also do automatic load balancing, but synchronization becomes more complicated (read a SQL book, not a Rails issue)



Rails with multiple databases, code

- In database.yml

```
slave1:
```

```
host: 18.48.43.29 # your slave's IP
```

```
database: production
```

```
username: root
```

```
password: pass
```

- Define a model Slave1.rb

```
class Slave1 < ActiveRecord::Base
```

```
  self.abstract_class = true
```

```
  establish_connection :slave1
```

```
end
```

- When you need to run a query on the slave, just do
Slave1.connection.execute("select * from some_table")



Shameless Self-Promotion

- Scribd.com: VC-backed and hiring
- Just 3 people so far! >10 by end of year.
- Awesome salary/equity combination
- If you're reading this, you're probably the right kind of person
- Building the world's largest open document library
- Email: hackers@scribd.com