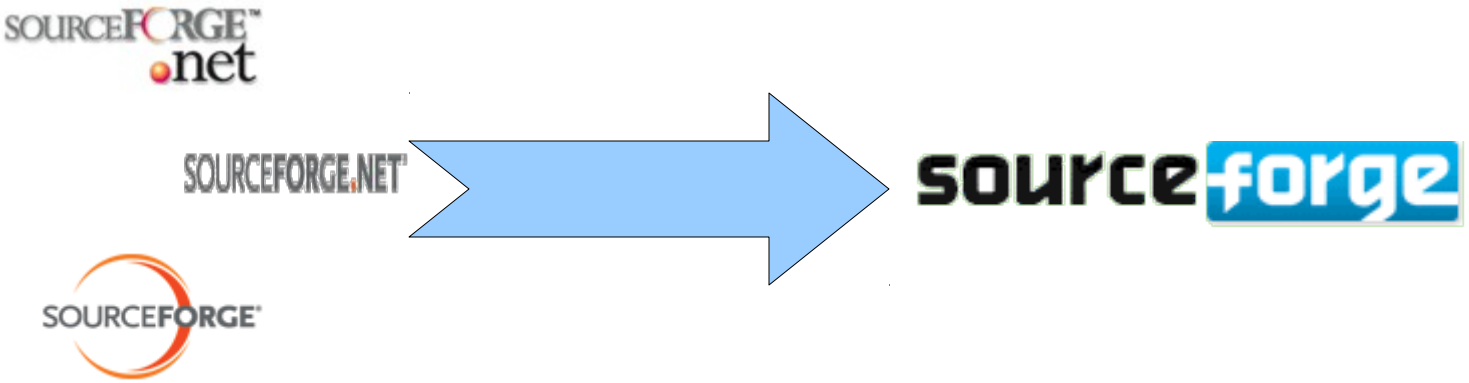
How Python, TurboGears, and MongoDB are Transforming SourceForge.net



Rick Copeland
http://blog.pythonisito.com
@rick446
rick@geek.net

Geeknet

SourceForge, GeekNet, wha...?

Geeknet

Geeknet

How do we make things exciting again?

Geeknet

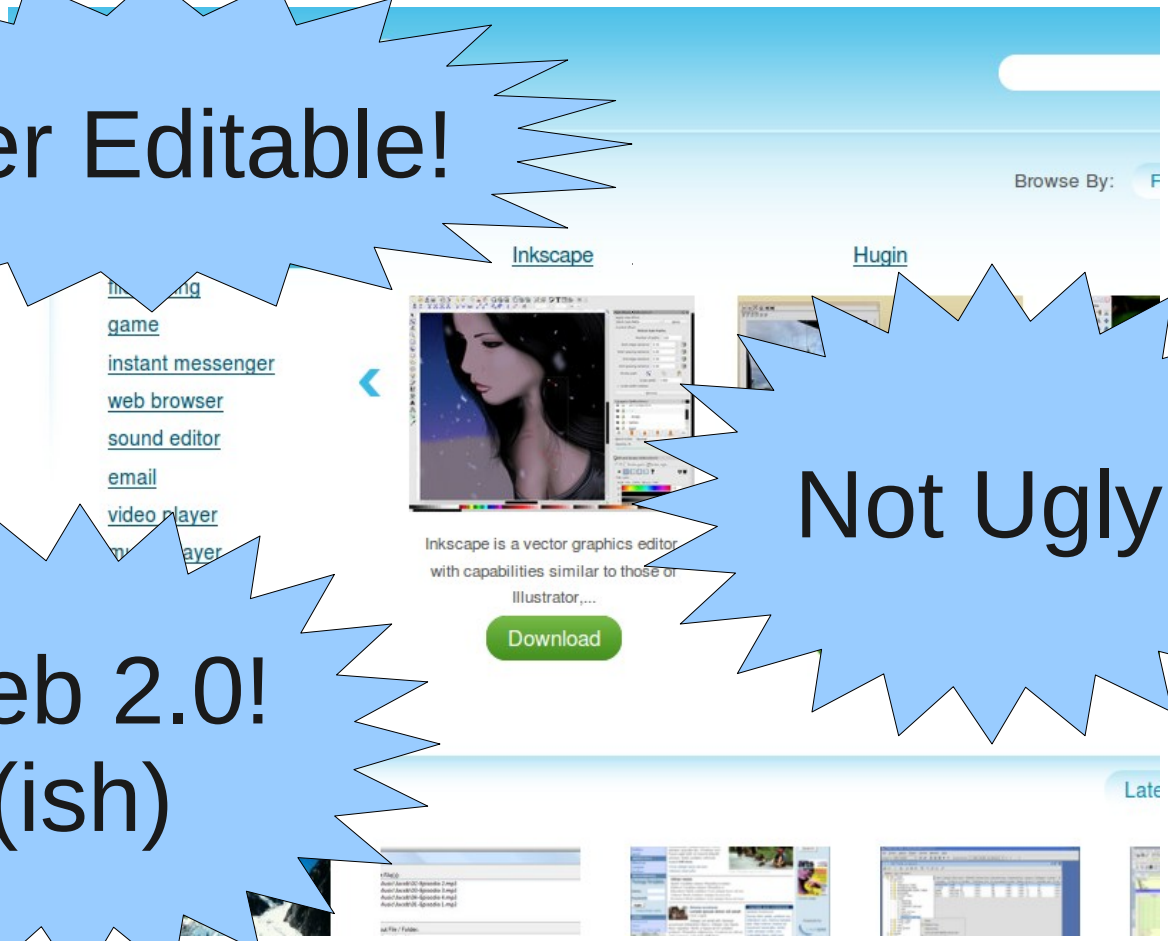
Python!

- Get it done well
- Get it done fast

Geeknet

First Project: FossFor.us

User Editable!



Not Ugly!

Web 2.0! (ish)

Geeknet

FossFor.us Technology

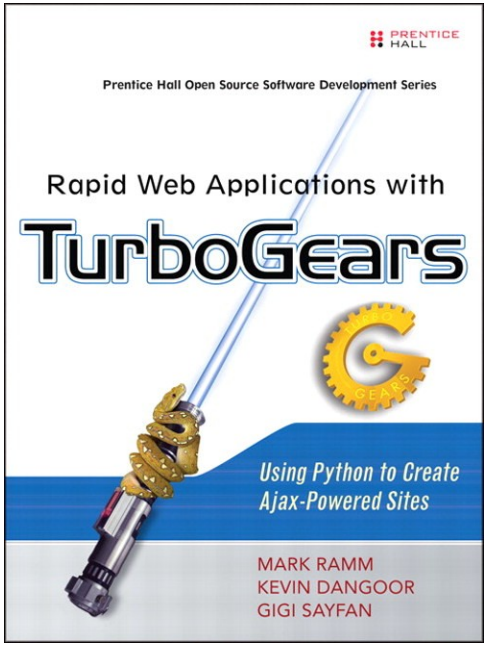
- Get it done well
- Get it done fast

Geeknet

Mark Ramm Hired Halfway Through FossFor.us

Geeknet

But isn't he the TurboGears guy?



Geeknet

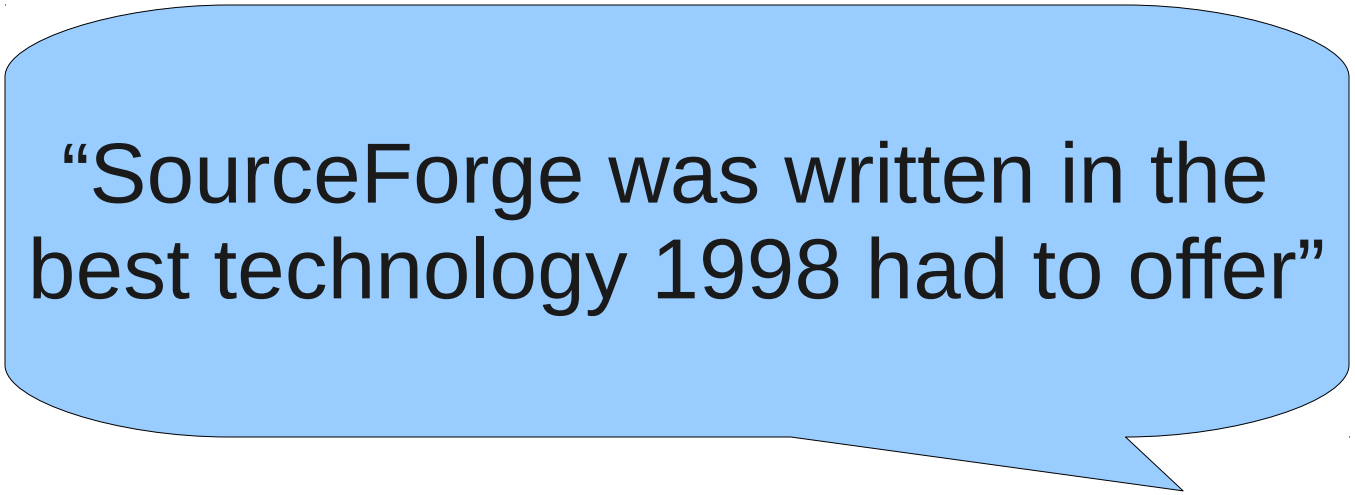
FossFor.us Went Well

- Small Team (2-3 programmers)
- Great framework/tools
- Flexibility to implement *fast*
- Could you guys give the same “spit-n-polish” to SourceForge.net itself?

Geeknet

BUT

- SourceForge.net is written in PHP... why not just upgrade the PHP?



Geeknet

So Another ~~Django~~ Site, Then?

TurboGears 2.0!

Geeknet

But Seriously, Why TG?

Easy to rip out what we don't need; most SF.net requests need:

- No authentication/authorization
- No widgets
- No admin
- No relational DB
- Lightweight session (stored in cookie)

Plays Nicely with others via WSGI middleware

Geeknet

Database: NoSQL

- Fast, fast, fast!
- Handle large data sets
- Simple replication
- No need for disconnected replication / synchronization

Geeknet

Templates

- We liked Django template syntax but we wanted....
- More speed
- A *bit* more logic in the template
- To not import all of Django for the sake of the template engine

Geeknet

BUT

- We're only doing the “consumer” side of the site
- Need to interface with legacy PHP code
- SQL-based queues are too slow

Geeknet

Everything's Going Just Fine For *Most* Projects...

Geeknet

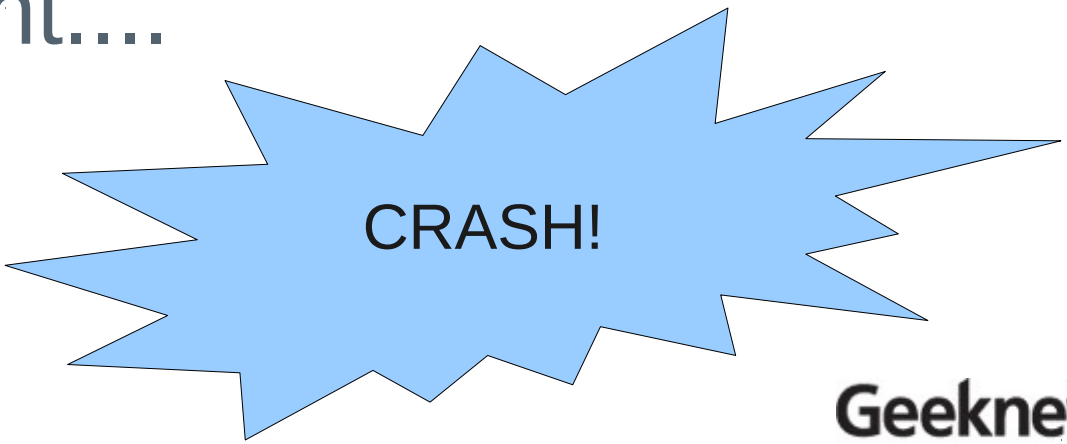
...Unless They Have a Lot of Releases

- MongoDB is document-oriented
- So we saved all the project info in one document
- When a project has lots of releases, we quickly overflow the max 4MB document size
- Also, large amounts of project data being fetched continuously
- FIX 1: Split Projects from ProjectReleases
- FIX 2: Cache Project records in memory

Geeknet

The Day the Servers Died

- Big release – Project File System replaces File Release System
- Most SF.net engineers attending OSCON
- Break away from sessions for the deployment....



Geeknet

But.... we tested it!

- Almost. Every. Part.
- Missed the download redirector – click to download took you *back* to the browse screen
- Click → redirect → click again → and again → and again
- At one point, ***5700*** open connections

Geeknet

Good News: We Didn't Fall Over

- Well, we didn't ***completely*** fall over
- Some 503s, but some people got through
- Still not what we'd like, but it let us fix the redirector online and get people their downloads sooner rather than later.

Geeknet

Scaling Lessons Learned

- Hardware load balancers are nice (but you can probably get by just fine with LigHTTPd or nginx)
- Keep a local MongoDB slave on each web server
- In-process, thread-safe memory caches for frequently hit objects. (Tough to get cache invalidation right!)
- mod_wsgi – multiprocess + multithreaded, auto restart every 2500-5000 requests to keep mem usage low
- 4 servers, 8 cores each, 8 GB each
- Run 6 TG processes per server, reserve 2 CPUs for whatever
    - We were memory-bound

Geeknet

Dead Ends? Memcached

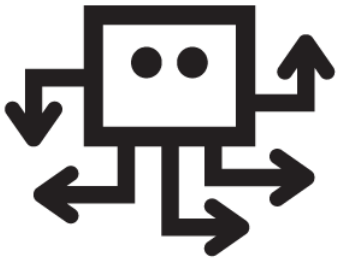
- Great idea – shared memory cache for multiple machines
- Overhead: Network
- Overhead: Serialization / Deserialization
- MongoDB is *really* fast anyway (almost as fast as Memcached)
- In-process Python dict is the fastest possible Python cache anyway

Geeknet

Where to Now?

- SourceForge releases Ming, a Python ORM-like library for MongoDB, under the MIT license
  - (http://sourceforge.net/projects/merciless)
- Continuing to rewrite more and more of our sites in Python, TG2, and MongoDB (they are our strategic tools)
- Be on the lookout for new announcements, new features, and new open source software from SourceForge.net
- “I'm not dead yet – I feel happy!”

Geeknet

Geeknet 

Questions?

Contact

Rick Copeland

http://blog.pythonisito.com

@rick446

rick@geek.net

Geeknet