my mom told me that

git doesn't scale

architecture blueprints for the largest source code host in the world
Thanks for being here
Vicent Martí
1. GIVEN NAME

2. ALIAS vmg

3. COMPANY GitHub, inc

4. CURRENT POSITION Software Architect
   reckless moron

5. CONTACT ADDRESS vicent@github.com

- Extremely unprofessional.
- Nice personal hygiene.

tweets @vmg
git

Git hosting: No longer a pain in the ass
GitHub

Git hosting:
No longer a pain in the ass
for you.
Not for us.

Because, goddamnit,
if I ever find the guy who invented
this thing I’m going to hang him from
a fence by his underwear and
Let’s host some Git repos!

- .git
- src
  - file.c
  - file.h
- README.md
- COPYING.md

Bare Repository
- HEAD
- index
- objects
- refs

.git-daemon
OK, now about the web...

grit

Ruby - Git interface
OK, now about the web...

grit

Ruby - Git interface

Bare Repository

Bare Repository

Bare Repository
1VM

storage

grit

rails app
nVM -> storage (GFS)
Rails was making us slow.
Ruby on Rails — Read more

http://rubyonrails.org
Literally.
Time to move to Real Hardware
bert
(binary Erlang term)
ernie
(not an acronym)

bert
(binary Erlang term)
Horizontal Scaling

Vertical Scaling
Horizontal Scaling
problems with their gigabits

Vertical Scaling
problems with their gigahertz
them

gigabits
NetShard

alternate network
alternate network
them

gigahurtz
bottleneck: grit
bottleneck:

grit

solution:

shell out to

git
bottleneck: shell out to git
bottleneck:

shell out to git

solution:

shell out to git
bottleneck:
solution:

shell out to git

shell out to git properly
posix_spawn
posix_spawn

Seriously.

< 1ms
The issue is not in “shelling out”, the issue is in the spawned process.
GUISE
GUISE
GUISE
GUISE
GUISE
GUISE
Why don’t we take
Why don’t we take the Git binary...
Why don’t we take the Git binary... yeah?
Why don't we take the Git binary... yeah? and compile it as
Why don’t we take the Git binary... yeah? and compile it as a library
Why don’t we take the Git binary... yeah? and compile it as a library oh... go on...
Why don’t we take the Git binary... yeah?

and compile it as a library.

oh... go on...

and link that into
Why don't we take the Git binary... yeah?

and compile it as a library... oh... go on...

and link that into our server
Well, we didn’t think about
Well, we didn’t think about freeing memory, but...
Well, we didn’t think about freeing memory, but...

THIS IS THE KIND OF PROBLEM WE COULD SOLVE WITH CGI
Well, we didn’t think about freeing memory, but...

This is the kind of problem we could solve with CGI in 1995.
Memory Usage

Time

Scientific Graph™
Memory Usage

Time

Scientific
Graph™
What do you mean

the server died?
What do you mean the server died?

die("BUG: non-INDEX attr direction in a bare repo");
What do you mean the server died?

die("BUG: non-INDEX attr direction in a bare repo");

die("a bad revision is needed");
What do you mean the server died?

die("BUG: non-INDEX attr direction in a bare repo");
die("a bad revision is needed");
die("'%s' is not a valid branch name.", name);
What do you mean the server died?

die("BUG: non-INDEX attr direction in a bare repo");
die("a bad revision is needed");
die("'%s' is not a valid branch name.", name);
die("Empty patch. Aborted.");
What do you mean the server died?

die("BUG: non-INDEX attr direction in a bare repo");

die("a bad revision is needed");

die("'%s' is not a valid branch name.", name);

die("Empty patch. Aborted.");

die("unable to read index file");
libgit
the "2" means this one frees memory
NOT ENOUGH FACTORIES

NOT ENOUGH FACTORIES

the "2" means this one frees memory
JGit

the "J" means this one is in Java

...not our thing.
Java
a brief timeline

New companies don’t use Java because it’s not like Unix

New companies use Java because it’s new and shiny

New companies don’t use Java because it’s ooooooold

New companies use the JVM because

Java

a brief timeline

New companies use Java because it’s new and shiny

New companies don’t use Java because it’s not like Unix

New companies don’t use Java because it’s oooooooold

New companies use the JVM because

If you think you understand the JVM, you are either:
If you think you understand the JVM, you are either:

a) Very smart
If you think you understand the JVM, you are either:

a) Very smart
b) Very wrong
If you think you understand the JVM, you are either:

a) Very smart
b) Very wrong
Some people think that GitHub is a Rails shop or even a Ruby shop.
Some people think that **github** is a Rails shop or even a Ruby shop.

**github** is a *Unix* shop and everything else is just a detail.
So, *libgit2*
libgit2

a brief timeline

Shawn Pearce

The Past
libgit2
a brief timeline

Shawn Pearce
myself
(about to have a mental breakdown)

myself

The Past
libgit2

a brief timeline

The Past

Shawn Pearce

myself

(about to have a mental breakdown)

myself

(having a mental breakdown)
libgit2
a brief timeline

The Past

real contributors

Shawn Pearce
myself
(about to have a mental breakdown)

myself
(reaching Git nirvana)

Carlos Martín
Ben Straub

Russell Belfer
Michael Schubert

The Past

myself
(having a mental breakdown)
libgit2

a brief timeline
libgit2
a brief timeline

Microsoft

1.0 release

?
Good Heavens, just look at the time.

It's NoSQL o'clock.
...do you even
...do you even mongo?
Key-Value Stores

If you wish upon a star, and have a pure heart...
If you wish upon a star, and have a pure heart...

Anything can be a Key-Value store!
<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
<th>state</th>
<th>lat</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>San Francisco</td>
<td>CA</td>
<td>24</td>
</tr>
<tr>
<td>24</td>
<td>Phoenix</td>
<td>AZ</td>
<td>33</td>
</tr>
<tr>
<td>7</td>
<td>Denver</td>
<td>CO</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>Caribou</td>
<td>ME</td>
<td>47</td>
</tr>
<tr>
<td>2</td>
<td>Los Angeles</td>
<td>CA</td>
<td>22</td>
</tr>
</tbody>
</table>

**SELECT * FROM CITIES WHERE name = 'San Francisco'**
Git is queried like a Key-Value Store

```
git show
f3c896c1949476e85abc0d75bb2143656a9580a6
```

But it is not a Key-Value store
Intermission

a brief introduction
to the Git data model
Behold, a graph.
@@ -261,13 +261,13 @@ bool git_attr_cache__is_cached()
 static int load_attr_file(
     const char **data,
     - git_futils_file_stamp *stamp,
     + git_futils_filestamp *stamp,
             const char *filename)
 {
     int error;
     git_buf content = GIT_BUF_INIT;
     - error = git_futils_file_stamp_has_changed(stamp, filename);
     + error = git_futils_filestamp_check(stamp, filename);
     if (error < 0)
         return error;
 ...
Well that was easy.
### Nov 05, 2012

<table>
<thead>
<tr>
<th>Commit</th>
<th>Author</th>
<th>Time</th>
<th>Commit Hash</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merge pull request #1046 from libgit2/empty-remote-url-crash</td>
<td>arrbee</td>
<td>19 hours ago</td>
<td>a5e85d86b7</td>
<td>Browse code</td>
</tr>
<tr>
<td>'geterr' -&gt; 'giterr'</td>
<td>jsphrsummers</td>
<td>19 hours ago</td>
<td>clcd036e40</td>
<td>Browse code</td>
</tr>
<tr>
<td>Revert &quot;Don't expect the 'empty-remote-url' remote to be listed&quot;</td>
<td>jsphrsummers</td>
<td>20 hours ago</td>
<td>6edefa1491</td>
<td>Browse code</td>
</tr>
<tr>
<td>Don't expect the 'empty-remote-url' remote to be listed</td>
<td>jsphrsummers</td>
<td>20 hours ago</td>
<td>f358ec143c</td>
<td>Browse code</td>
</tr>
<tr>
<td>Test for GITERR_INVALID</td>
<td>jsphrsummers</td>
<td>20 hours ago</td>
<td>1fe990eee3</td>
<td>Browse code</td>
</tr>
<tr>
<td>Merge pull request #1044 from dahlbyk/repo/state</td>
<td>arrbee</td>
<td>20 hours ago</td>
<td>065be7a1d9</td>
<td>Browse code</td>
</tr>
<tr>
<td>Set GITERR_INVALID when encountering a NULL remote URL</td>
<td>jsphrsummers</td>
<td>20 hours ago</td>
<td>f8baece754</td>
<td>Browse code</td>
</tr>
<tr>
<td>Merge pull request #1034 from carlosmn/packbuilder-foreach</td>
<td>vmg</td>
<td>a day ago</td>
<td>942a76983b</td>
<td>Browse code</td>
</tr>
</tbody>
</table>

### Nov 04, 2012

<table>
<thead>
<tr>
<th>Commit</th>
<th>Author</th>
<th>Time</th>
<th>Commit Hash</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>remote: Add malformed remote load test</td>
<td>nulltoken</td>
<td>a day ago</td>
<td>a2a618948c</td>
<td>Browse code</td>
</tr>
<tr>
<td>Bail out if remote-&gt;url would be NULL</td>
<td>jsphrsummers</td>
<td>a day ago</td>
<td>8388s5891f5</td>
<td>Browse code</td>
</tr>
<tr>
<td>Fix Clar link</td>
<td>dahlbyk</td>
<td>a day ago</td>
<td>221ee54b8c</td>
<td>Browse code</td>
</tr>
<tr>
<td>repo: fix state when HEAD is not detached</td>
<td>dahlbyk</td>
<td>2 days ago</td>
<td>35d255fda6</td>
<td>Browse code</td>
</tr>
</tbody>
</table>
Oh god kill me
Little known torture methods:

death by roundtrip
<table>
<thead>
<tr>
<th>Directory</th>
<th>Date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>чная</td>
<td>a day ago</td>
<td>Add network transfer callbacks on Windows [ben]</td>
</tr>
<tr>
<td>docs</td>
<td>3 days ago</td>
<td>Fix spelling errors. [waywardmonkeys]</td>
</tr>
<tr>
<td>examples</td>
<td>3 days ago</td>
<td>Remove 'bytes' param from git_remote_download [ben]</td>
</tr>
<tr>
<td>include</td>
<td>2 days ago</td>
<td>Merge pull request #925 from nulltoken/topic/moving-branch-updates-co… [arrbee]</td>
</tr>
<tr>
<td>packaging</td>
<td>3 months ago</td>
<td>libgit2.pc.in: also switch to LIB_INSTALL_DIR [schu]</td>
</tr>
<tr>
<td>src</td>
<td>a day ago</td>
<td>Add network transfer callbacks on Windows [ben]</td>
</tr>
<tr>
<td>tests-clar</td>
<td>2 days ago</td>
<td>Merge pull request #925 from nulltoken/topic/moving-branch-updates-co… [arrbee]</td>
</tr>
<tr>
<td>.HEADER</td>
<td>4 years ago</td>
<td>Switch the license from BSD to GPL+libgcc exception [spearce]</td>
</tr>
<tr>
<td>.gitattributes</td>
<td>a year ago</td>
<td>Add git attributes settings for *c and *h to force line endings to LF. [kelly.leahy]</td>
</tr>
<tr>
<td>.gitignore</td>
<td>2 months ago</td>
<td>Ignore tags file [ben]</td>
</tr>
<tr>
<td>.travis.yml</td>
<td>17 days ago</td>
<td>tests: add a test for tag autofollow behaviour [carlosmn]</td>
</tr>
<tr>
<td>AUTHORS</td>
<td>5 months ago</td>
<td>Complete the AUTHORS list. [ben]</td>
</tr>
<tr>
<td>CMakeLists.txt</td>
<td>2 days ago</td>
<td>remove /ZM1000 from msvc builds [Edward Thomson]</td>
</tr>
<tr>
<td>CONVENTIONS</td>
<td>5 months ago</td>
<td>CONVENTIONS: Update error code names [carlosmn]</td>
</tr>
<tr>
<td>COPYING</td>
<td>8 months ago</td>
<td>Add GNU LGPL to COPYING [carlosmn]</td>
</tr>
<tr>
<td>Makefile.embed</td>
<td>4 days ago</td>
<td>Fix Makefile.embed to build src/sha1/*.c [scottig]</td>
</tr>
<tr>
<td>README.md</td>
<td>8 months ago</td>
<td>Use standard CMake variable names for installation paths [veeti]</td>
</tr>
</tbody>
</table>
warning:
the rabbit hole is pretty deep
/*
 * Copyright (C) 2009-2012 the libgit2 contributors
 *
 * This file is part of libgit2, distributed under the GNU GPL v2 with
 * a Linking Exception. For full terms see the included COPYING file.
 */

#include "git2/config.h"
#include "git2/types.h"
#include "git2/oid.h"
#include "config.h"
#include "repository.h"
#include "remote.h"
#include "fetch.h"
#include "refs.h"
#include <regex.h>

static int parse_remote.refs[1] (git_config *cfg, git_refs *refs, const char *var, bool
warning: git totally wasn’t designed for this

#include "repository.h"
#include "remote.h"
#include "fetch.h"
#include "refs.h"

#include <regex.h>

static int parse_remote_refspec(git_config *cfg, git_refspec *refspec, const char *var, bool
Git doesn’t give a #!%$ about CAP
Number of hops on a complex query

1,000,000
Number of hops on a complex query

1,000,000

Required hops for a successful query

1,000,000
Number of hops on a complex query: 1,000,000

Required hops for a successful query: 1,000,000

Replica count to ensure 100% availability: a metric shitton
We could fix it.
We could fix it. But we won’t.
GitRPC
Git

RPC

Less.
New serialization protocol

Banana Pack

MessagePack + more
New serialization protocol

Banana Pack

MessagePack + more

mochilo
New serialization protocol

MessagePack + more

mochilolo

Banana Phone
evolutionary
(disappointing?)
Summary:
## Summary:

<table>
<thead>
<tr>
<th>Ruby</th>
<th>Unix</th>
<th>Boring</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>Unix</td>
<td>Boring</td>
</tr>
<tr>
<td>Unix</td>
<td>Unix</td>
<td>Boring</td>
</tr>
<tr>
<td>Unix</td>
<td>Unix</td>
<td>Boring</td>
</tr>
<tr>
<td>Unix</td>
<td>Unix</td>
<td>Boring</td>
</tr>
</tbody>
</table>
Use the most reliable tools you know.
Challenge yourself to build the simplest thing.

Not because it’s easy, but because it works.
Innovate where it really matters.
create a revolutionary product not a revolutionary backend.
Q: Does Git scale?
Q: Does Git scale?
A: Who cares?
Thanks!

Please send your JVM-based scaling solutions to:

(bonus points if they involve Neo4J)

github.com/vmg
twitter.com/vmg