A Tour of Basho's Source at GitHub

Scott Lystig Fritchie

Senior Software Engineer

scott@basho.com
@slfritchie

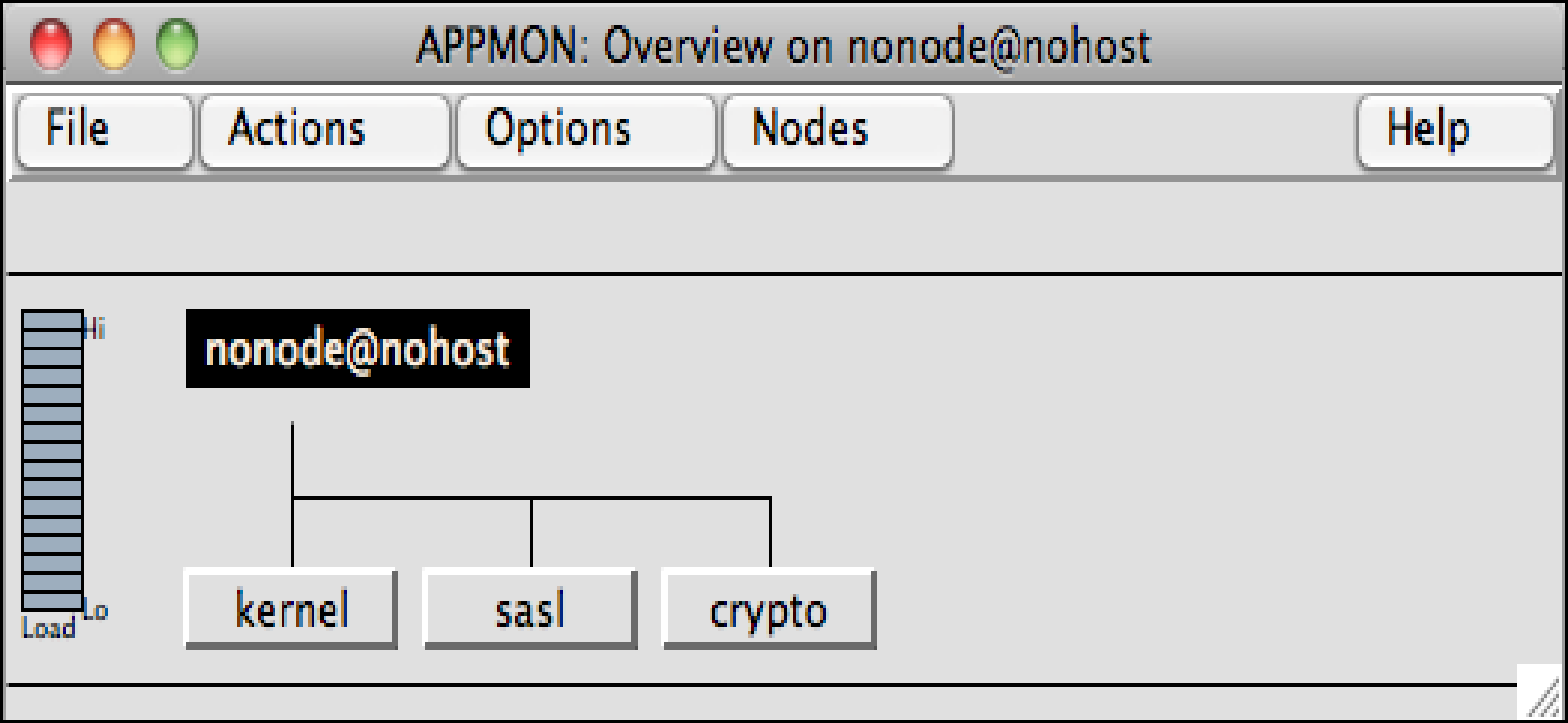
TL;DR

- See slides 3-57.

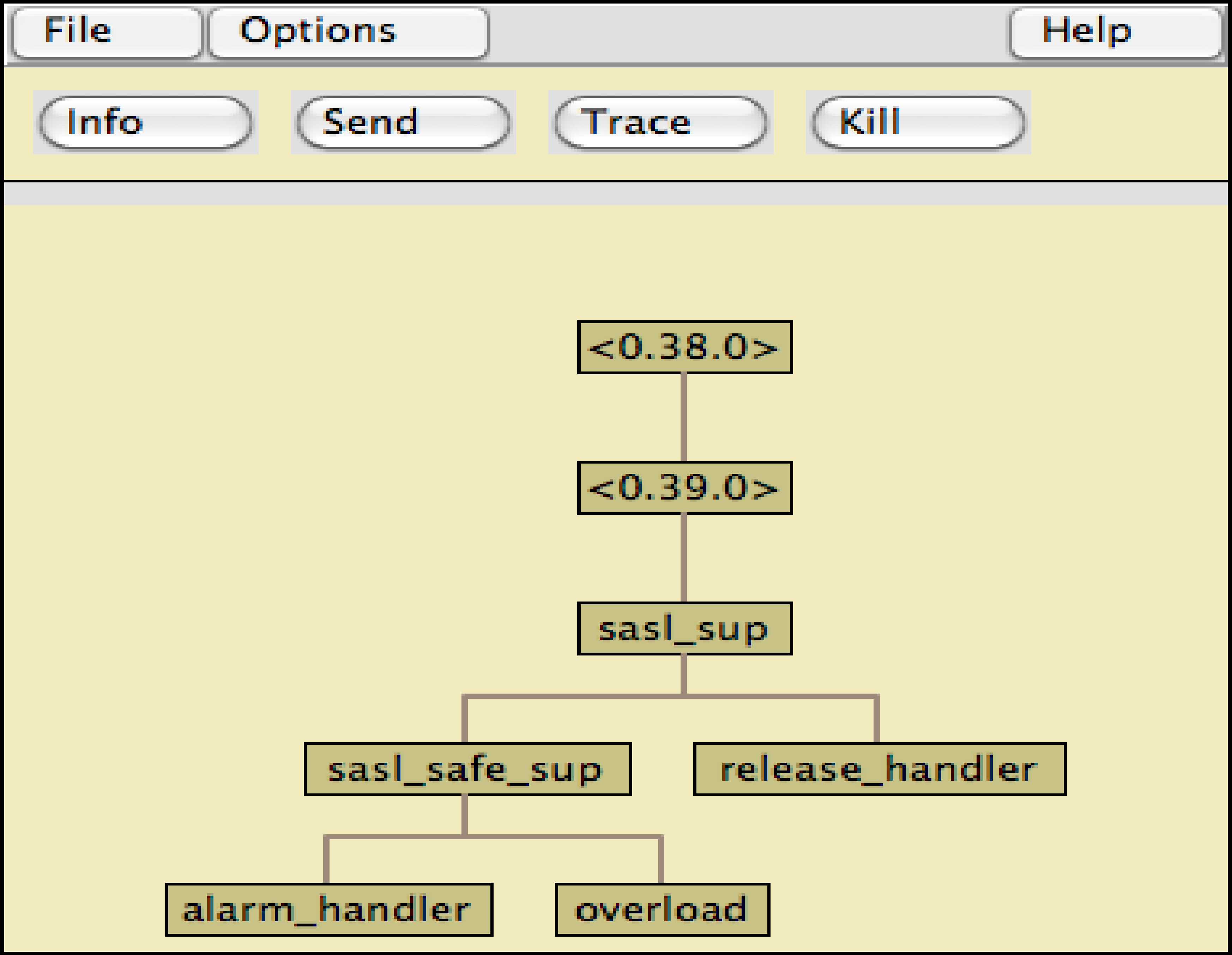
Goals

- You know what an OTP application is.
- You know what OTP apps Basho has @ GitHub.
- You know how Basho's apps might help *your* app.
- You don't mob me demanding beer....

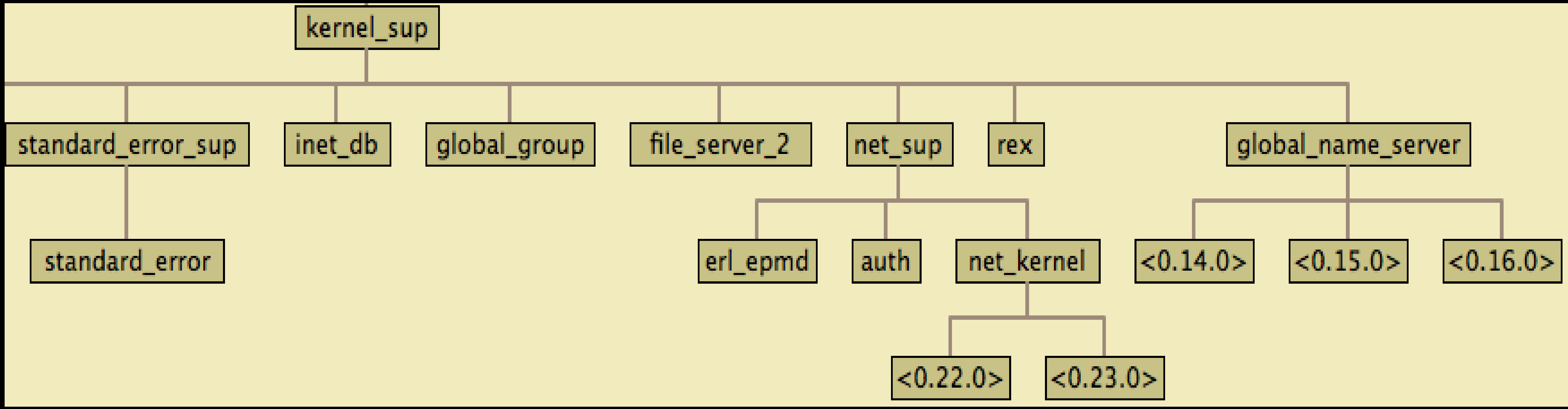
The 'appmon' GUI



A View of 'sasl' ...



... and 'kernel'


kernel_sup
standard_error_sup
inet_db
global_group
file_server_2
net_sup
rex
global_name_server
standard_error
erl_epmd
auth
net_kernel
<0.14.0>
<0.15.0>
<0.16.0>
<0.22.0>
<0.23.0>

OTP Application Properties

- Version number
- BEAM files
- Scripts: application dependencies, upgrade and downgrade scripts, ...
- Processes
  - Supervisors
  - Workers

Starting & Stopping

- application:start(AppName).
- application:stop(AppName).
- application:which_applications().

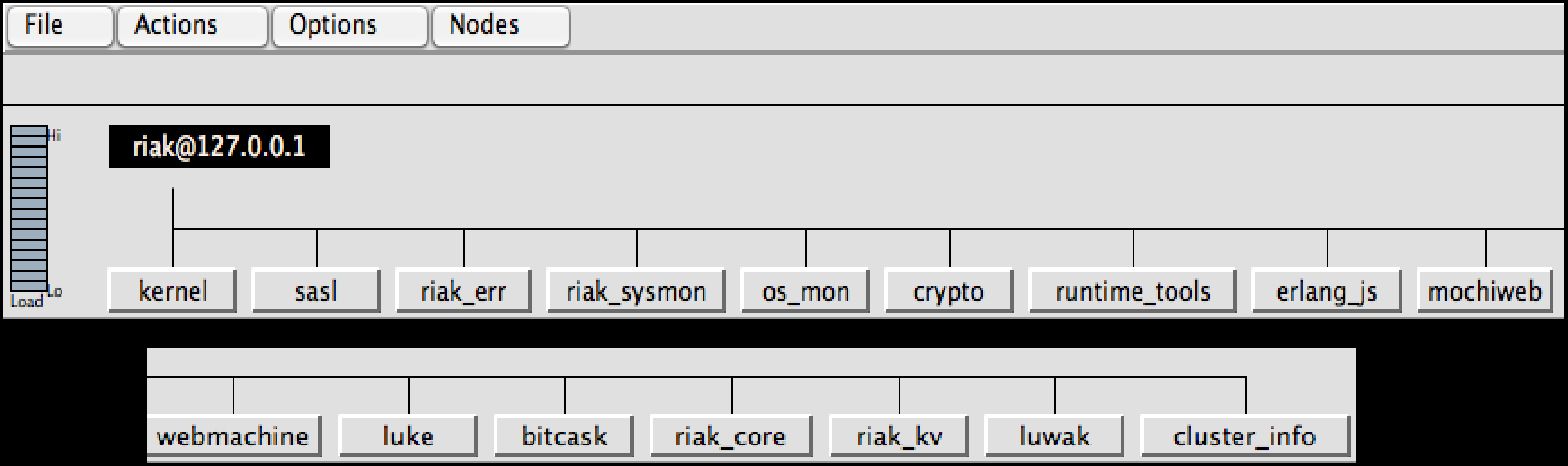
```
[{basho_stats,"Basic Erlang statistics library","1.0.1"},

 {bitcask,[],"1.1.5"},
 {cluster_info,"Cluster info/postmortem app","1.1.0"},
 {crypto,"CRYPTO version 1","1.6.4"},
 {erlang_js,"Interface between BEAM and JS","0.5.0"},
 {kernel,"ERTS  CXC 138 10","2.13.5"},
....
```

What Does This Have to Do With GitHub?

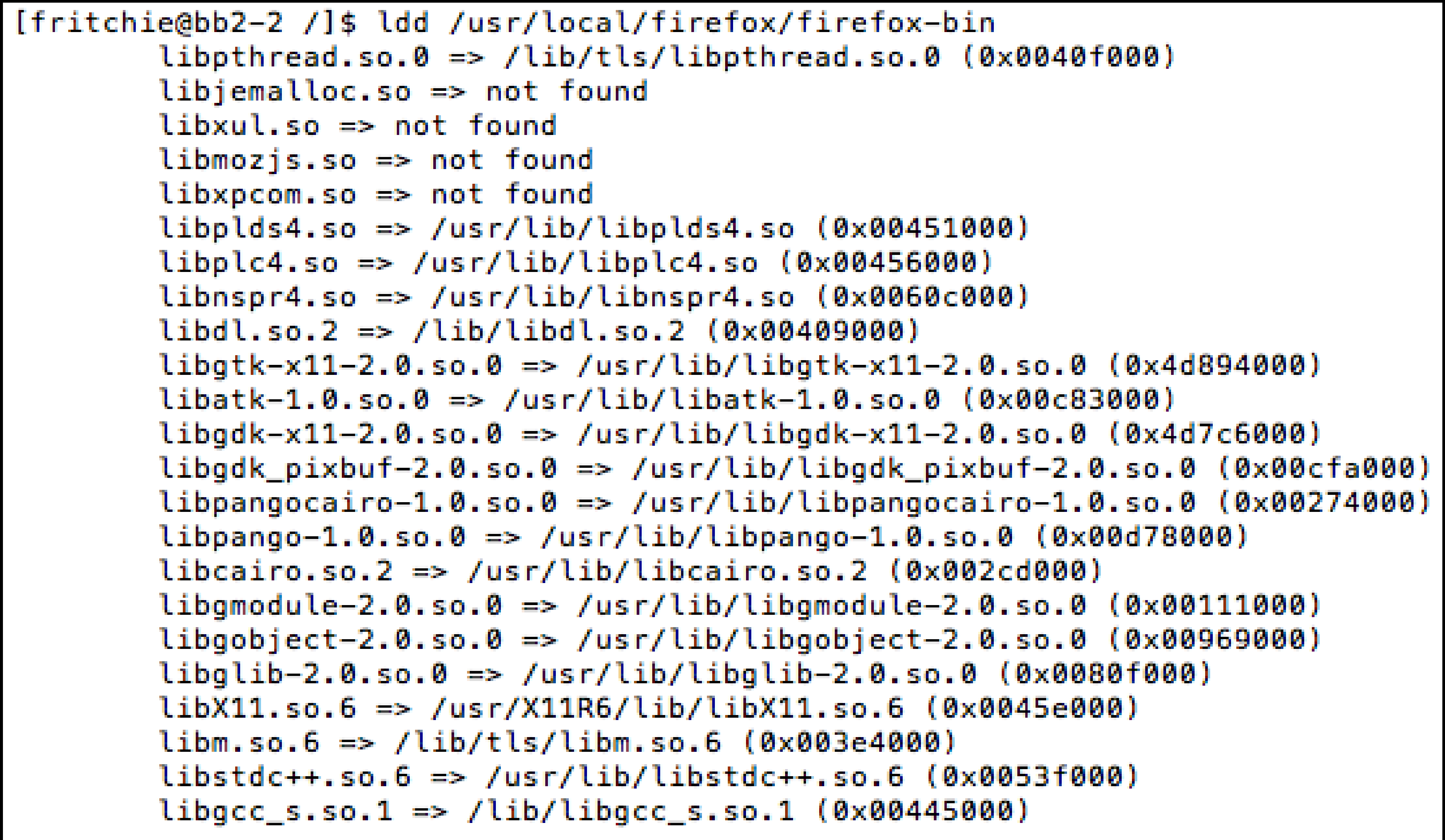
- Yeah, I'm getting there....

Riak as Seen by 'appmon'



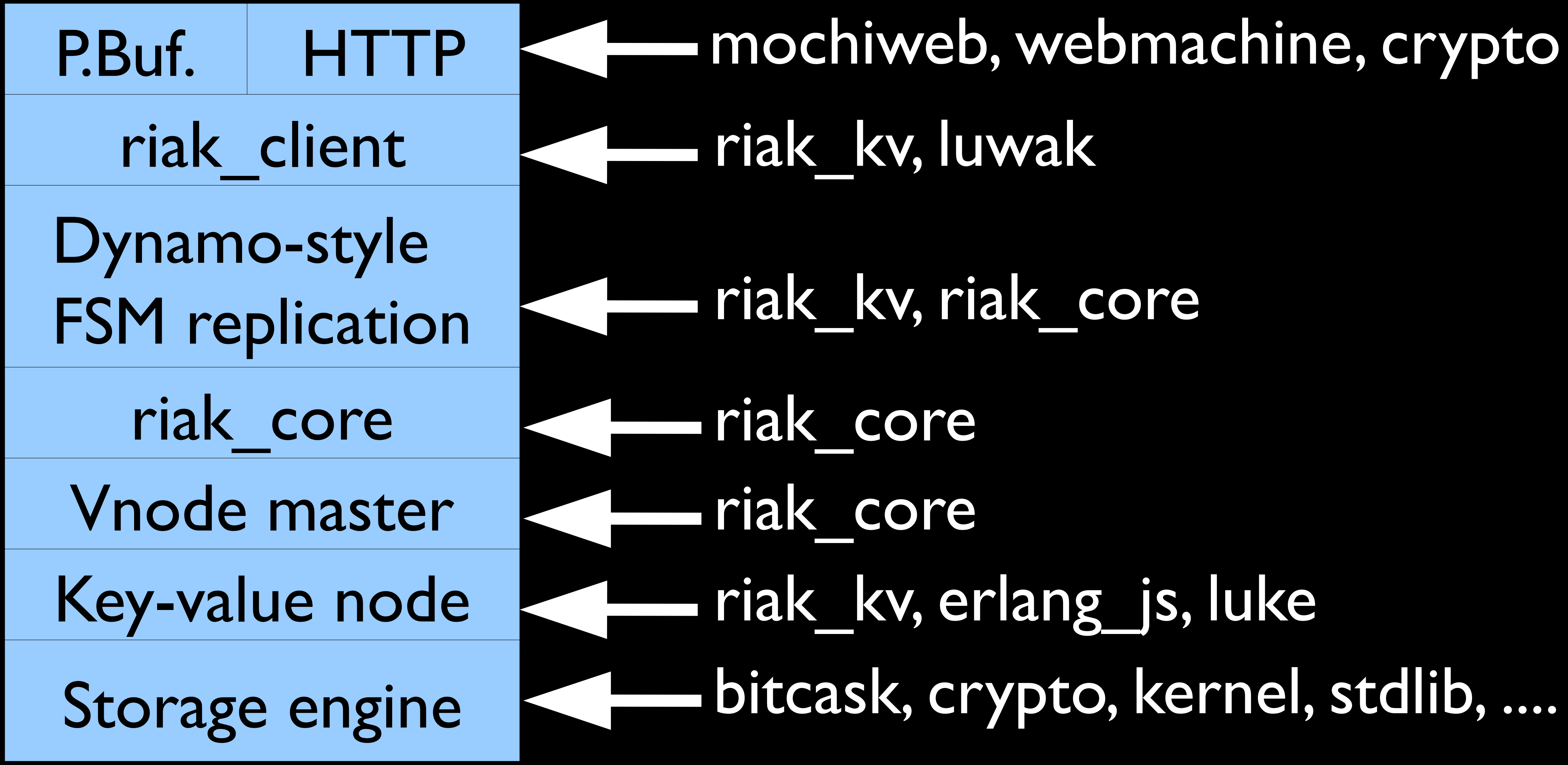
Why so many apps?

- Riak has many parts, different from C packaging.

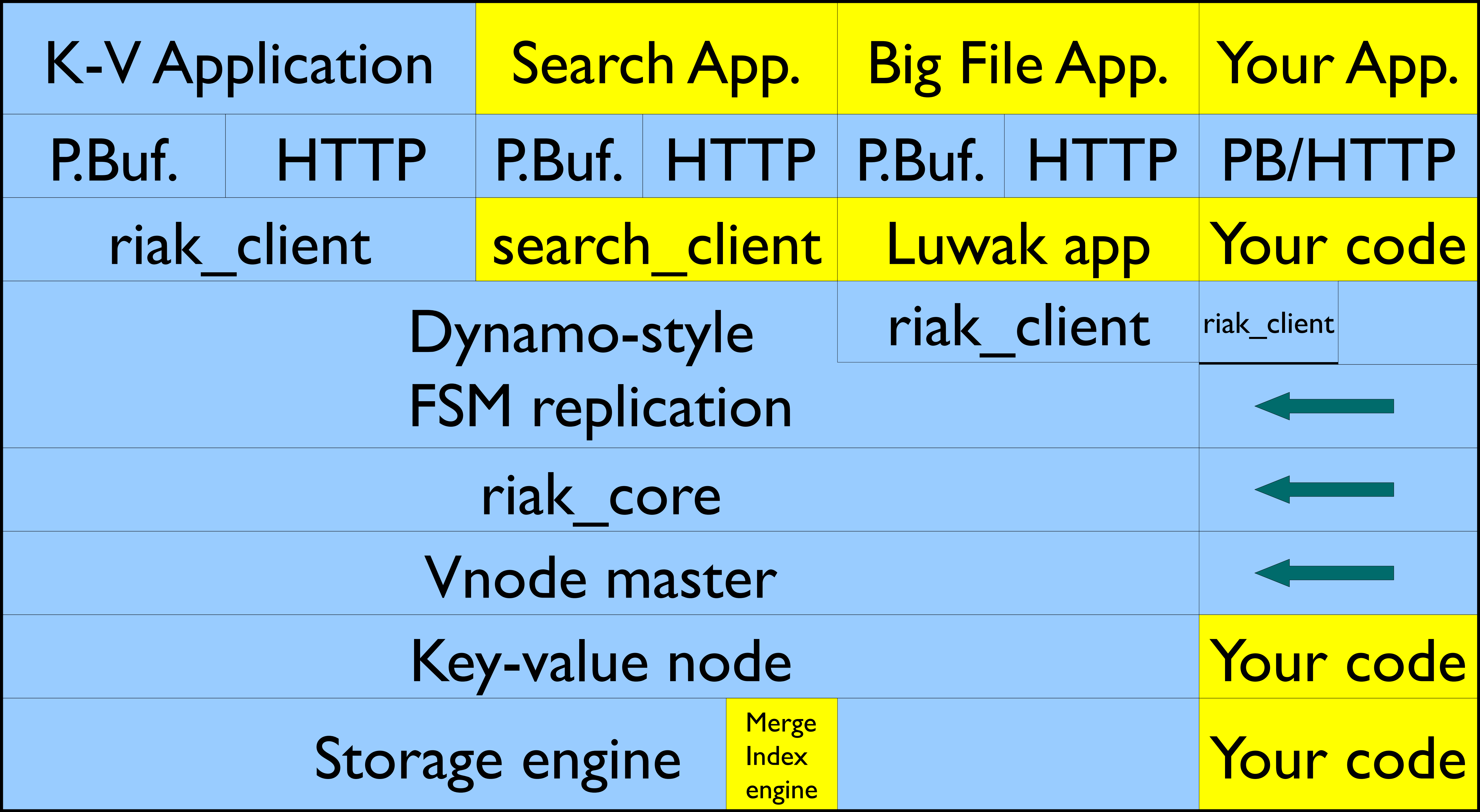


```
[fritchie@bb2-2 /]$ ldd /usr/local/firefox/firefox-bin
        libpthread.so.0 => /lib/tls/libpthread.so.0 (0x0040f000)
        libjemalloc.so => not found
        libxul.so => not found
        libmozjs.so => not found
        libxpcom.so => not found
        libplds4.so => /usr/lib/libplds4.so (0x00451000)
        libplc4.so => /usr/lib/libplc4.so (0x00456000)
        libnspr4.so => /usr/lib/libnspr4.so (0x0060c000)
        libdl.so.2 => /lib/libdl.so.2 (0x00409000)
        libgtk-x11-2.0.so.0 => /usr/lib/libgtk-x11-2.0.so.0 (0x4d894000)
        libatk-1.0.so.0 => /usr/lib/libatk-1.0.so.0 (0x00c83000)
        libgdk-x11-2.0.so.0 => /usr/lib/libgdk-x11-2.0.so.0 (0x4d7c6000)
        libgdk_pixbuf-2.0.so.0 => /usr/lib/libgdk_pixbuf-2.0.so.0 (0x00cfa000)
        libpangocairo-1.0.so.0 => /usr/lib/libpangocairo-1.0.so.0 (0x00274000)
        libpango-1.0.so.0 => /usr/lib/libpango-1.0.so.0 (0x00d78000)
        libcairo.so.2 => /usr/lib/libcairo.so.2 (0x002cd000)
        libgmodule-2.0.so.0 => /usr/lib/libgmodule-2.0.so.0 (0x00111000)
        libgobject-2.0.so.0 => /usr/lib/libgobject-2.0.so.0 (0x00969000)
        libglib-2.0.so.0 => /usr/lib/libglib-2.0.so.0 (0x0080f000)
        libX11.so.6 => /usr/X11R6/lib/libX11.so.6 (0x0045e000)
        libm.so.6 => /lib/tls/libm.so.6 (0x003e4000)
        libstdc++.so.6 => /usr/lib/libstdc++.so.6 (0x0053f000)
        libgcc_s.so.1 => /lib/libgcc_s.so.1 (0x00445000)
```

Some of Riak's major OTP apps



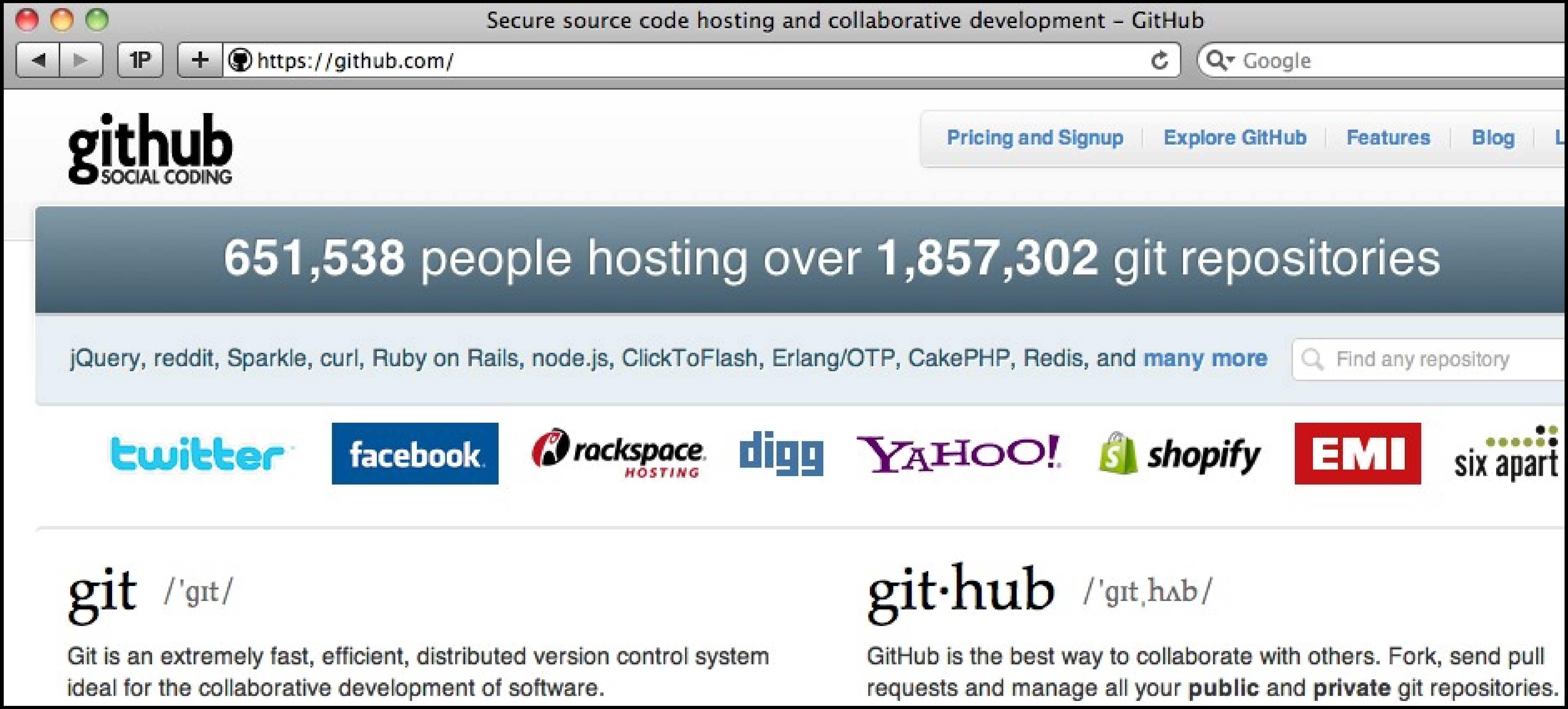
Flexible App Packaging: KV, Search, Luwak, custom



Reality Check

- You probably know a bit more about:
  - What OTP applications are.
  - Why OTP applications are useful
- Questions?

GitHub



Secure source code hosting and collaborative development - GitHub

https://github.com/

651,538 people hosting over 1,857,302 git repositories

jQuery, reddit, Sparkle, curl, Ruby on Rails, node.js, ClickToFlash, Erlang/OTP, CakePHP, Redis, and many more

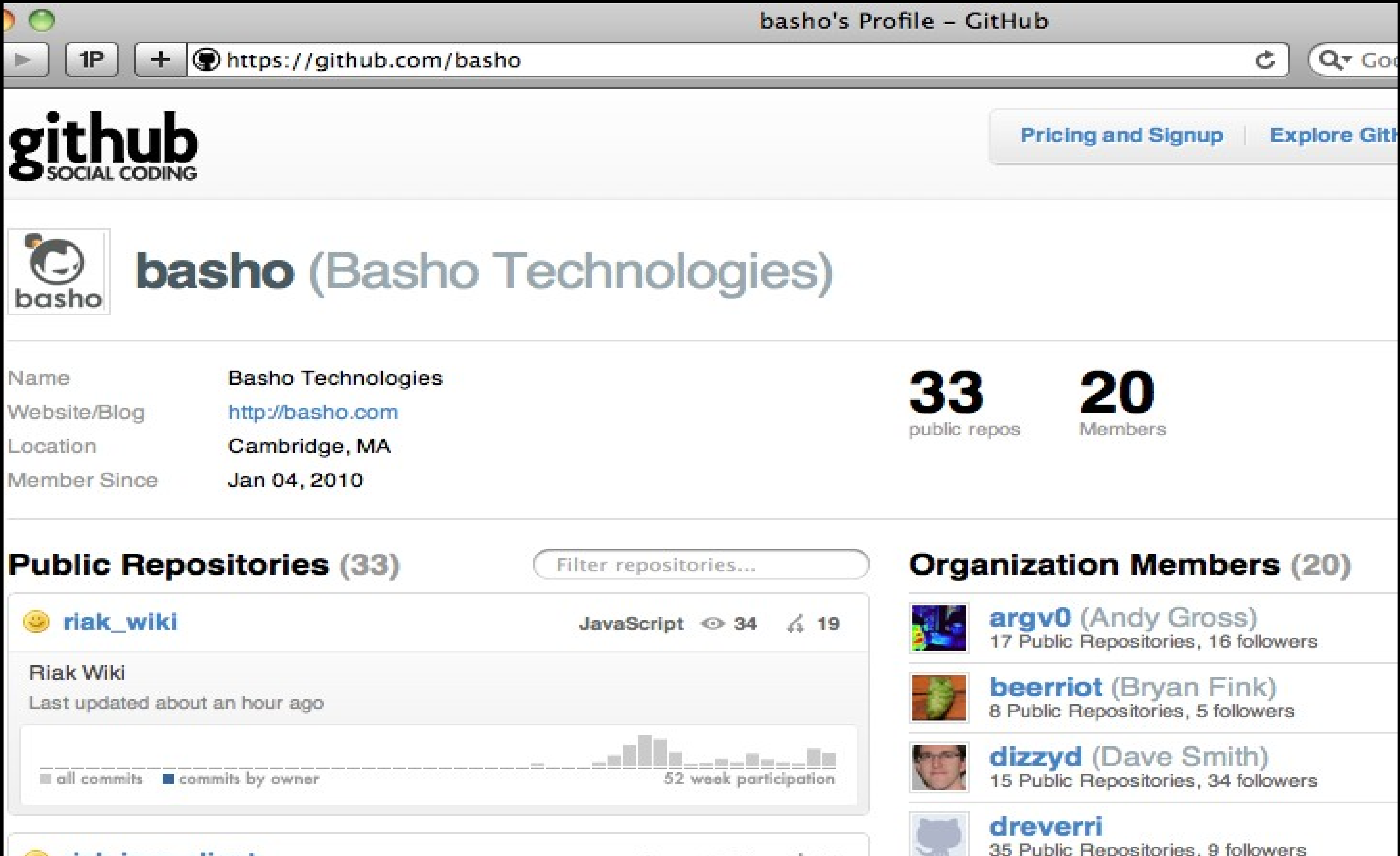
git /ˈgɪt/

Git is an extremely fast, efficient, distributed version control system ideal for the collaborative development of software.

git·hub /ˈgɪtˌhʌb/

GitHub is the best way to collaborate with others. Fork, send pull requests and manage all your **public** and **private** git repositories.

Basho @ GitHub



basho's Profile – GitHub

https://github.com/basho

Pricing and Signup | Explore Git

basho (Basho Technologies)

Name	Basho Technologies
Website/Blog	http://basho.com
Location	Cambridge, MA
Member Since	Jan 04, 2010

33 public repos

20 Members

Public Repositories (33)

riak_wiki — JavaScript 34 19

Riak Wiki

Last updated about an hour ago

all commits · commits by owner · 52 week participation

Organization Members (20)

argv0 (Andy Gross)
17 Public Repositories, 16 followers

beerriot (Bryan Fink)
8 Public Repositories, 5 followers

dizzyd (Dave Smith)
15 Public Repositories, 34 followers

dreverri
35 Public Repositories, 9 followers

Basho's Public Repos

- Riak major apps & client protocols
- Riak client apps, demos, & documentation
- **HTTP servers**
- **Local data stores**
- **Benchmarking**
- **Utilities**
- **Testing libraries**

Riak Major Apps

- riak: top-level packaging for Riak (see also: rebar)
- riak_core: Riak's distributed systems logic
- riak_kv: Riak's key-value & MapReduce logic
- riak_search: full-text search engine for Riak
- luwak: Large-object storage interface for Riak
- erlang_js: linked-in driver to Mozilla's Spidermonkey
- luke: Dataflow/MapRed. coordination framework

Riak Client Protocols

- riak-erlang-client
- riak-erlang-http-client
- riak-java-client
- riak-javascript-client
- riak-php-client
- riak-python-client
- riak_function_contrib

Riak Apps, Demos, Docs

- riaktant: node.js app: stores syslog messages in Riak Search
- wriaki: wiki-like app fully embedded into Riak
- riak_wiki: content for http://wiki.basho.com/
- bashobot: bot for the #riak IRC channel

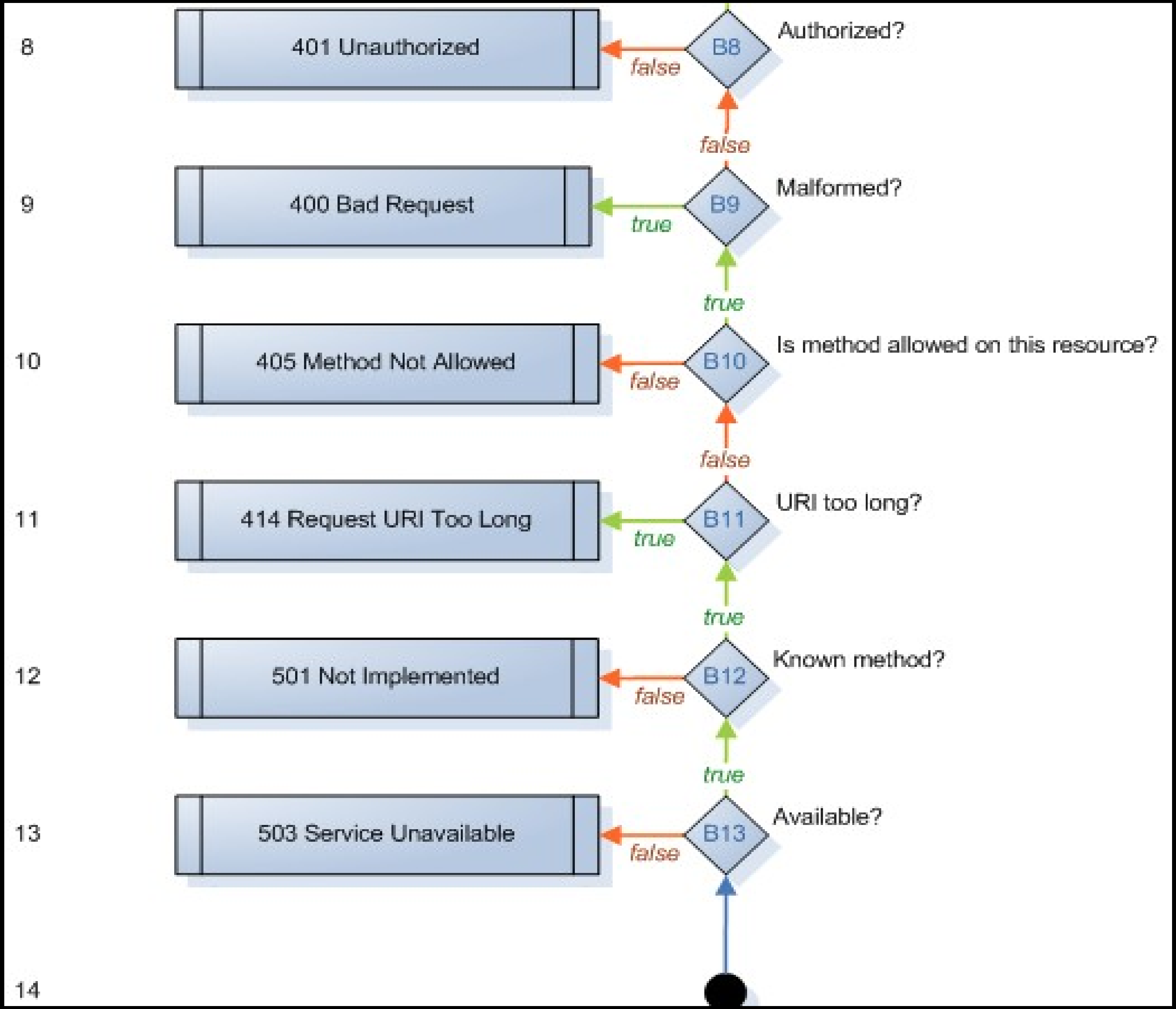
HTTP Servers

- MochiWeb: forked from Mochi's excellent HTTP server
- WebMachine: a REST-based system for building Web apps

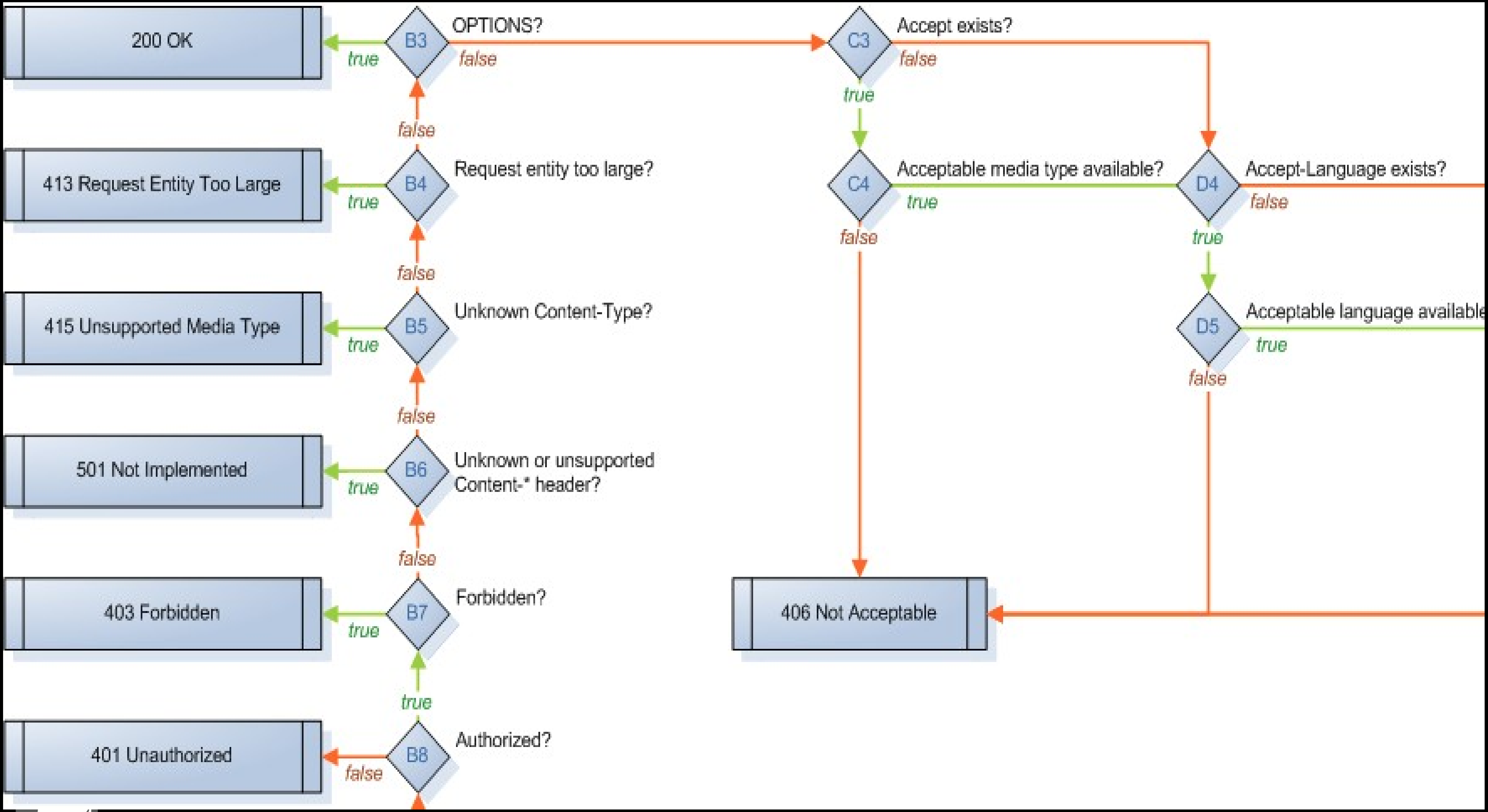
WebMachine

- Maps an HTTP server's logic onto a flowchart
- Complete control over every stage of HTTP request processing.
- More docs: http://webmachine.basho.com/

WebMachine

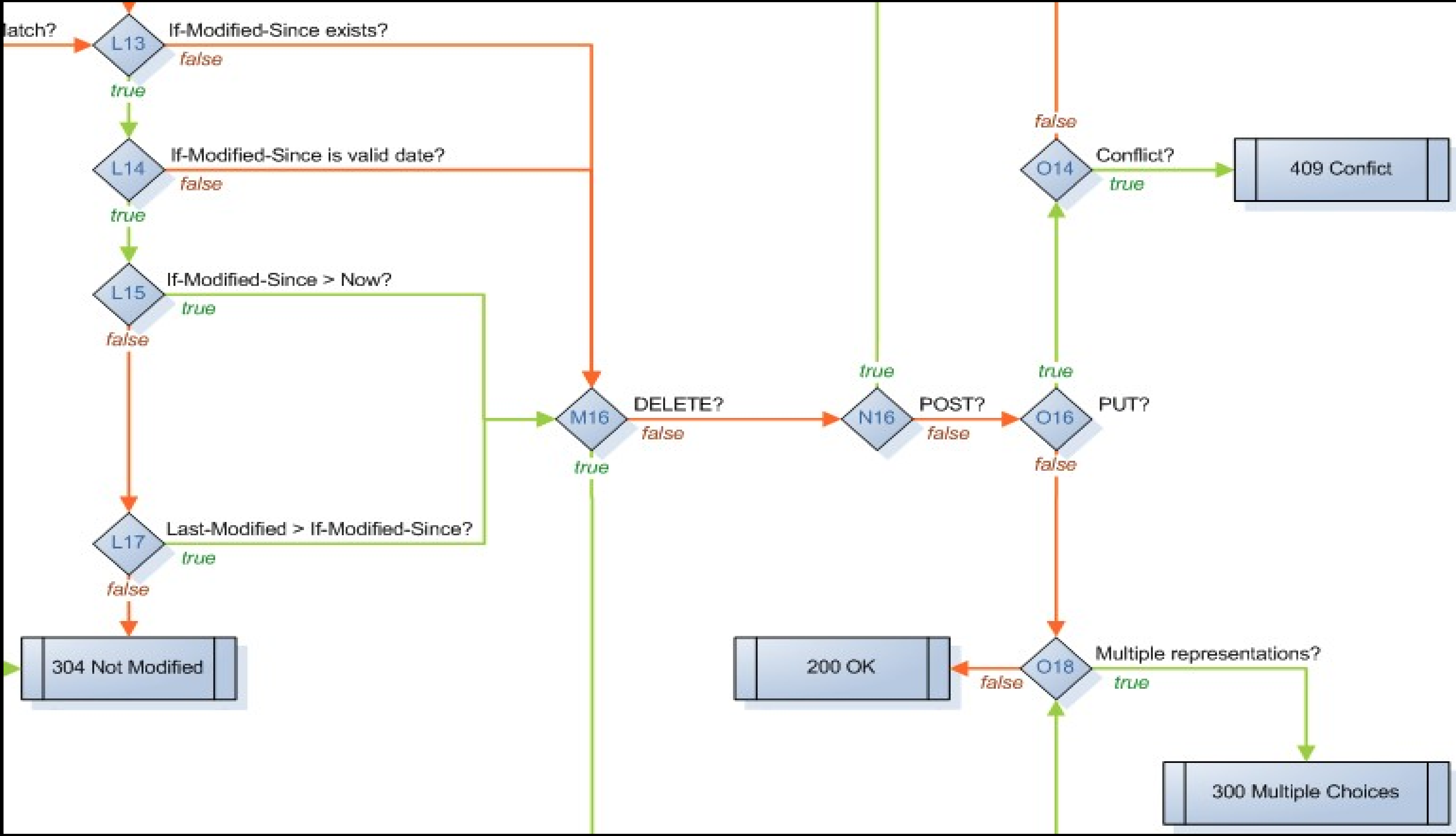

8
9
10
11
12
13
14
401 Unauthorized
400 Bad Request
405 Method Not Allowed
414 Request URI Too Long
501 Not Implemented
503 Service Unavailable
B8
Authorized?
B9
Malformed?
B10
Is method allowed on this resource?
B11
URI too long?
B12
Known method?
B13
Available?
true
false

WebMachine

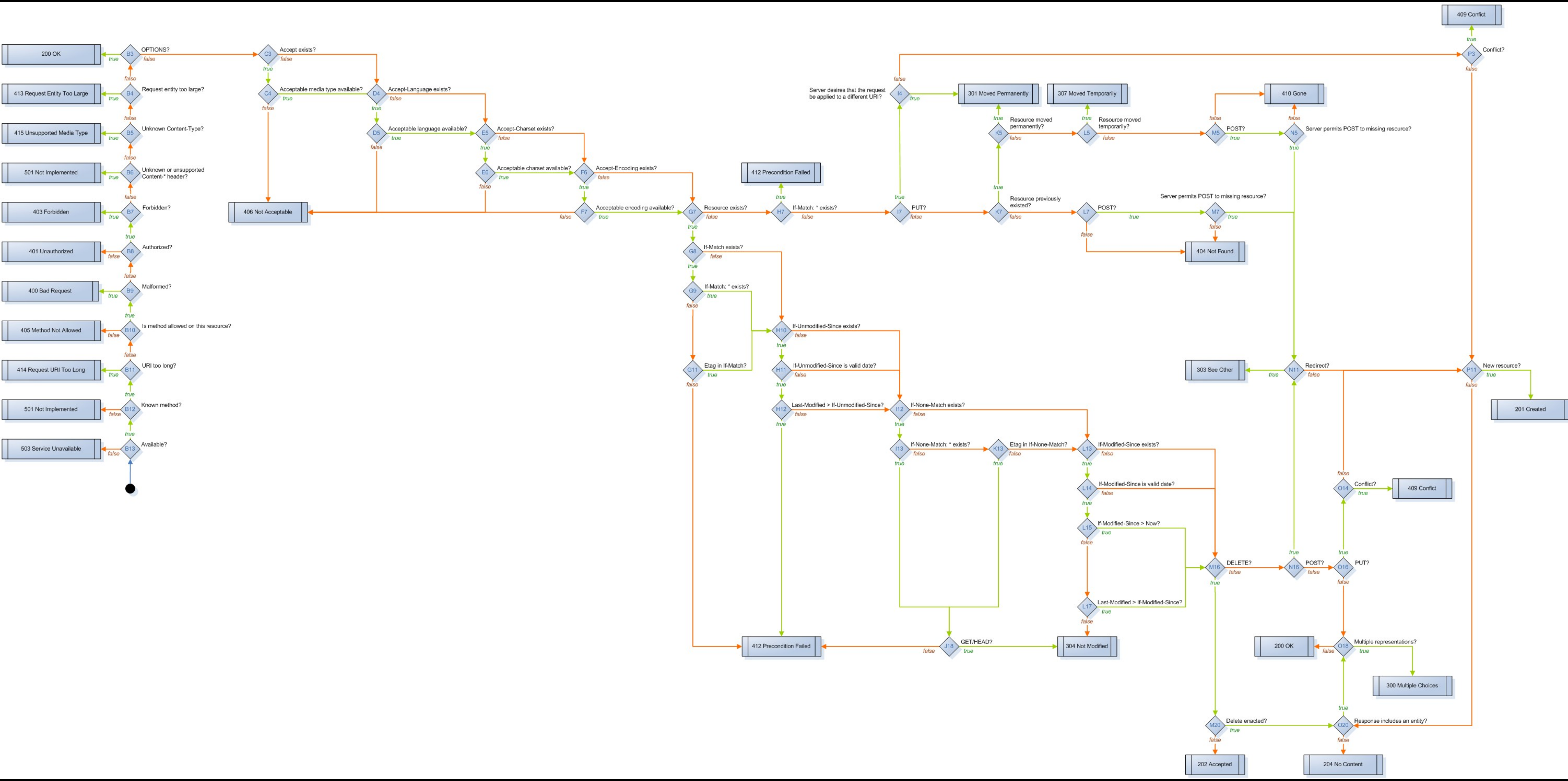


basho

WebMachine


latch?
L13
If-Modified-Since exists?
false
true
L14
If-Modified-Since is valid date?
false
true
L15
If-Modified-Since > Now?
true
false
L17
Last-Modified > If-Modified-Since?
true
false
304 Not Modified
M16
DELETE?
false
true
N16
POST?
false
true
O16
PUT?
true
false
O14
Conflict?
true
false
409 Conflict
O18
Multiple representations?
true
false
200 OK
300 Multiple Choices

WebMachine


409 Conflict
200 OK
B3 OPTIONS?
C3 Accept exists?
P3 Conflict?
413 Request Entity Too Large
B4 Request entity too large?
C4 Acceptable media type available?
D4 Accept-Language exists?
Server desires that the request be applied to a different URI?
I4
301 Moved Permanently
307 Moved Temporarily
410 Gone
415 Unsupported Media Type
B5 Unknown Content-Type?
D5 Acceptable language available?
E5 Accept-Charset exists?
K5 Resource moved permanently?
L5 Resource moved temporarily?
M5 POST?
N5 Server permits POST to missing resource?
501 Not Implemented
B6 Unknown or unsupported Content-* header?
E6 Acceptable charset available?
F6 Accept-Encoding exists?
412 Precondition Failed
403 Forbidden
B7 Forbidden?
406 Not Acceptable
F7 Acceptable encoding available?
G7 Resource exists?
H7 If-Match: * exists?
I7 PUT?
K7 Resource previously existed?
L7 POST?
M7 Server permits POST to missing resource?
401 Unauthorized
B8 Authorized?
G8 If-Match exists?
404 Not Found
400 Bad Request
B9 Malformed?
G9 If-Match: * exists?
405 Method Not Allowed
B10 Is method allowed on this resource?
H10 If-Unmodified-Since exists?
414 Request URI Too Long
B11 URI too long?
G11 Etag in If-Match?
H11 If-Unmodified-Since is valid date?
303 See Other
N11 Redirect?
P11 New resource?
501 Not Implemented
B12 Known method?
H12 Last-Modified > If-Unmodified-Since?
I12 If-None-Match exists?
201 Created
503 Service Unavailable
B13 Available?
I13 If-None-Match: * exists?
K13 Etag in If-None-Match?
L13 If-Modified-Since exists?
O14 Conflict?
409 Conflict
L14 If-Modified-Since is valid date?
L15 If-Modified-Since > Now?
M16 DELETE?
N16 POST?
O16 PUT?
L17 Last-Modified > If-Modified-Since?
412 Precondition Failed
J18 GET/HEAD?
304 Not Modified
200 OK
O18 Multiple representations?
300 Multiple Choices
M20 Delete enacted?
O20 Response includes an entity?
202 Accepted
204 No Content
true
false

Local Data Stores

- bitcask
- innostore

bitcask

- A log-structured hash table for fast (and predictable latency) key/value data
  - All keys are stored in RAM
  - All values are stored on disk
    - All writes: append-only, sequential I/O
    - All reads: at most one open(), lseek(), read()
  - Garbage collection via log file merge (sequential disk I/O)

innostore

- A driver for Embedded InnoDB
  - A transactional engine for MySQL
  - InnoDB's API not covered 100%
    - API is intentionally slim
  - http://www.innodb.com/

Benchmarking

- … is hard to do well.
- … these apps can help:
  - basho_bench: an extendable benchmarking tool
  - casbench: utility library for basho_bench, interfacing to Cassandra via Thrift

basho_bench

- Throughput:
  - number of operations per unit of time
  - aggregated across all operation types
- Latency:
  - time to complete single operations
  - captured in quantiles per-operation and 95%, 99%, and 99.9%
- Graphs created by R (external package)

basho_bench

- Drivers: bitcask, cassandra, DETS, Hibari, HTTP (use/abuse as you wish), Innostore, null, Riak (HTTP), Riak (Protocol Buffers)
  - Very easy for an Erlang novice to write a new driver.
- Configurable key distribution: sequential_int, partitioned_sequential_int, uniform_int, pareto_int, truncated_pareto_int, user-defined
- Control run time, # of concurrent worker procs, worker proc operation rate

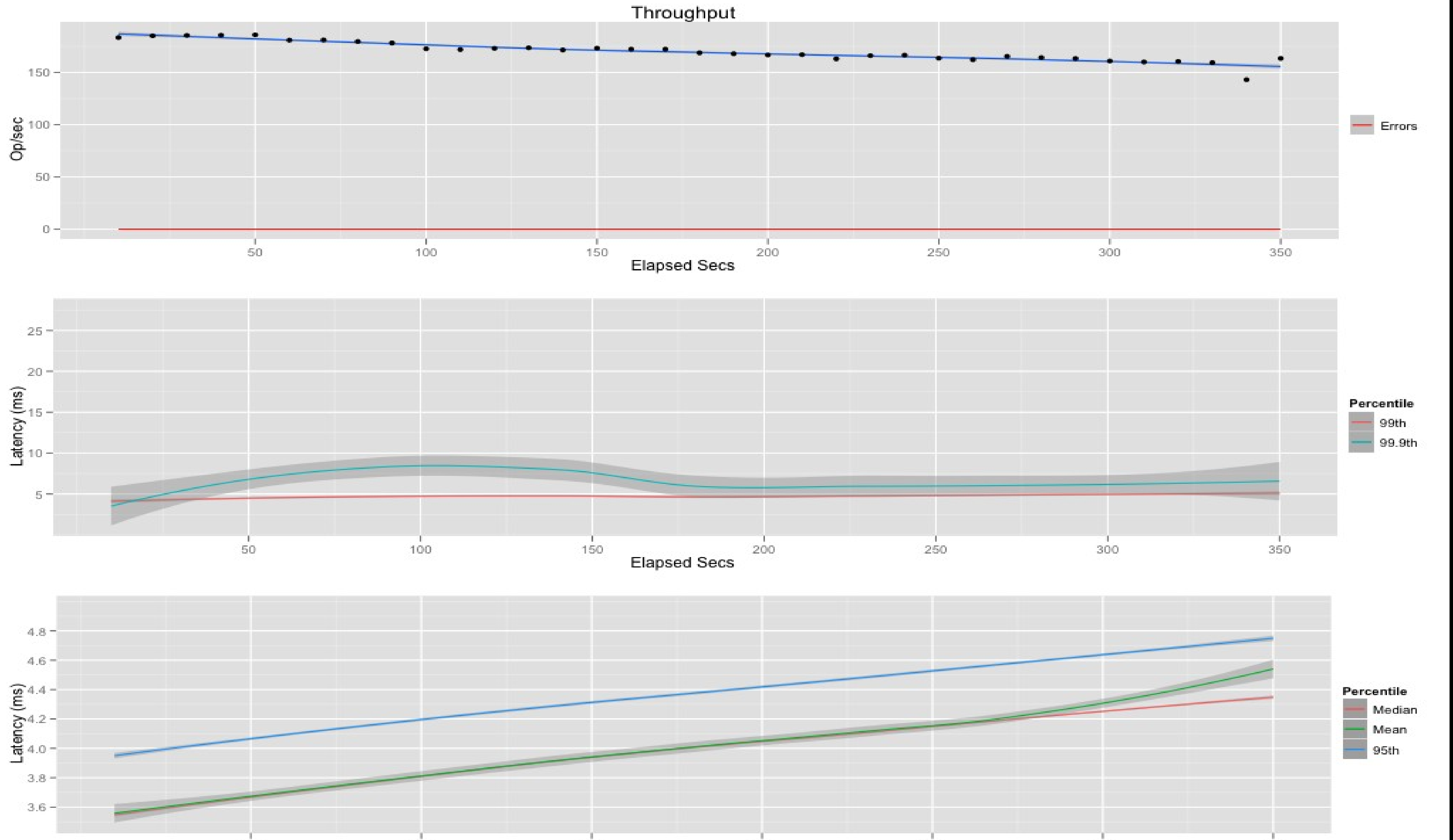
basho_bench

```
run(get, KeyGen, _ValueGen, State) ->
  {NextUrl, S2} = next_url(State),
  case do_get(url(NextUrl, KeyGen, State#state.path_params)) of
    {ok, _Url, _Headers} ->
      {ok, S2};
    {error, Reason} ->
      {error, Reason, S2}
  end.

next_url(S = #state{base_urls = Base, base_urls_index = BaseIndex)
 when BaseIndex > tuple_size(Base)
  {element(1, Base), S#state{base_urls_index = 1}};
next_url(S = #state{base_urls = Base, base_urls_index = BaseIndex) ->
  {element(BaseIndex, Base), S#state{base_urls_index = BaseIndex+1}}.

%% do_get() uses the 'ibrowse' HTTP client.
%% do_get() also takes care of HTTP persistent connections and
%% mapping of status 200/300    ok, 404    not_found, 5xx    error
```

basho_bench


Throughput
Op/sec
150
100
50
0
50
100
150
200
250
300
350
Elapsed Secs
Errors
Latency (ms)
25
20
15
10
5
Percentile
99th
99.9th
4.8
4.6
4.4
4.2
4.0
3.8
3.6
Median
Mean
95th

Utilities

- basho_stats
- cluster_info
- ebloom
- erlang_protobuffs
- rebar
- riak_err
- riak_sysmon
- skerl

basho_stats

- Basic Erlang statistics library
- Used by Riak for latency statistics
  - Min, max, mean, variance, standard deviation, median, quartiles, histogram
- NOTE: the EUnit tests with QuickCheck will occasionally fail @ basho_stats_histogram:qc_quantile_test()
  - It's OK for that test to fail occasionally, QuickCheck is evil (in a good way)

cluster_info

- How many times have you needed more info about an Erlang system in the field?
  - Memory usage, # of processes, RAM used, ETS table sizes, # of ports in use, …
  - You need it for all nodes in the cluster.
  - You need it simple. Single-command simple.

cluster_info

- cluster_info:dump_all_connected("/tmp/out.txt").
- All nodes' output → one file
- All the info bits mentioned earlier
  - … and very easy to extend.

cluster_info

```
% egrep '^==* ' /tmp/out.txt
== Node: 'riak@127.0.0.1'
= Generator name: Current time and date
= Generator name: VM statistics
= Generator name: erlang:memory() summary
= Generator name: Top 50 process memory hogs
= Generator name: Registered process names
= Generator name: Registered process name via regs()
= Generator name: Non-zero mailbox sizes
= Generator name: Ports
= Generator name: Applications
= Generator name: Timer status
= Generator name: ETS summary
= Generator name: Nodes summary
= Generator name: net_kernel summary
= Generator name: inet_db summary
= Generator name: Alarm summary
= Generator name: Global summary
= Generator name: erlang:system_info() summary
= Generator name: Loaded modules
[… output truncated …]
```

ebloom

- NIF driver for a Bloom filter
  - http://en.wikipedia.org/wiki/Bloom_filter

```
1> PredictedElementCount=5.
2> FalsePositiveProbability=0.01.
3> RandomSeed=123.
4> {ok, Ref} = ebloom:new(PredictedElementCount,
                             FalsePositiveProbability, RandomSeed).
5> ebloom:insert(Ref, <<"abcdef">>).
ok
6> true = ebloom:contains(Ref, <<"abcdef">>).
true
7> false = ebloom:contains(Ref, <<"zzzzzz">>).
false
```

erlang_protobuffs

- An implementation of Google's Protocol Buffers for Erlang
- Based on Nick Gerakines code
  - https://github.com/ngerakines/erlang_protobuffs
- For when you can't use Joe Armstrong's / Gemini Mobile's UBF protocol for Erlang/JavaScript/Java/Python/.... :-)
  - https://github.com/norton/ubf

rebar

- A “make” replacement that's aware of OTP design principles.
- Aware of dependencies on 3rd-party source repositories.
- 80% of what you need → dead simple
  - The next 15% isn't too hard.
- Won't download the entire Internet before compiling our project.

riak_err

- Goal: make the SASL `error_logger` really difficult to crash an Erlang VM.
- Default SASL error handler is a memory pig.
- If `error_logger` crashes the entire Erlang VM:
  - Your customers are unhappy
  - You lost the error message that triggered the crash.
  - Unhappiness increases geometrically...
    - … if not exponentially …

riak_err

```
9> StrLen = 128*1024.                  %% 128KB
131072
10> Big = lists:duplicate(StrLen, 131).
[131,131,131,131,...]
11> Formatted = io_lib:format("Big string: ~p\n", [Big]).
[...]
12> erts_debug:flat_size(Formatted).    %% 1.53MByte
1605628
13> 1605628 / StrLen.
12.249969482421875
```

- One byte → four characters: 1, 3, 1, \,
- One ASCII char → one cons cell → 2 words → 8/16 bytes (32bit vs. 64bit platform)

riak_err

- Easy to use: drop-in replacement for default SASL error logger event handler
- Configurable max string length
    - If string > limit, then truncate
- Not perfect, but *much less* likely to hog memory

riak_sysmon

- How many times have you wondered?
  - Is garbage collection causing latency problems?
  - Are some processes hogging too much memory?
  - Are some sockets blocked by fast producers/slow consumers?

riak_sysmon

- The VM can help answer those questions:
    - Process GC exceeds N milliseconds
    - Process heap size exceeds N bytes
    - Ports are busy
    - Network distribution ports are busy
- But very few Erlang apps subscribe to these events.
- VM allows only a single process to subscribe

riak_sysmon

- Easy to use: a self-contained OTP application
- Add your own event handler (`gen_event` style)
- Multiple OTP apps can manage their own event handlers
  - … and share an “unsharable” system resource

skerl

- NIF interface for Skein hash function
- Supports 256, 512, and 1024 bit hash values

Testing Libraries

- mapred_verify: exercise Riak's MapReduce
- QuickCheck & PropEr tests
- basho_expect (coming soon!)
- Protocol simulator (coming soon!)

QuickCheck & PropEr tests

- Scattered throughout the code
  - bitcask, erlang_protobuffs, riak_core, riak_kv, riak_search
- If you aren't using property-based testing, *you should think again*.
  - Hard-core evil QA genius in a box

Protocol Buffers Test

```
prop_encode_decode1() ->
  ?FORALL({FieldNum,Data,Type}, protobuff_data(),
      begin
          {{N, RData}, <<>>} = protobuffs:decode(
                                  protobuffs:encode(FieldNum, Data,
                                                    Type), Type),
          FieldNum =:= N andalso
               (compare(Data, RData) orelse
               foreign_type(Type, Data, Rdata))% true|false <=> 1|0
      end).
```

Protocol Buffers Test

```
protobuff_data() ->
  oneof([
      {field_num(), int(32), int32},
      {field_num(), uint(32), uint32},
      {field_num(), int(64), int64},
      {field_num(), uint(64), uint64},
      {field_num(), bool(), bool},
      {field_num(), sint(32), sint32},
      {field_num(), sint(64), sint64},
      {field_num(), real(), float},
      {field_num(), real(), double},
      {field_num(), list(char()), string},
      {field_num(), binary(), bytes}
  ]).
```

basho_expect

- Python-based tool (uses Pexpect & SSH)
    - Erlang driver for Pexpect is partially written
- Used for testing all Riak packages before release
    - Single node regression tests, client protocol tests
    - Multi-node cluster tests
    - Cluster agnostic: Xen, VMware, EC2, real machines, ...
- Testing Solaris, OpenSolaris, RedHat EL, Ubuntu, and Fedora Core nodes in the same cluster
- Will release in 2011 (hopefully well before December)

Messaging Simulator

- Goal: test message-based protocols...
  - when process scheduling may be *very unfair.*
  - when network partitions happen arbitrarily
- Same semantics & behavior as Erlang
  - Some message ordering rules, otherwise "send and pray"
- QuickCheck friendly
- Releasing soon, contact me directly you're interested in early access.

Plug

Support & consulting? Enterprise features, EE pricing for startups?

Email info@basho.com or go to http://www.basho.com/contact.html to talk with us.

www.basho.com

basho

Questions