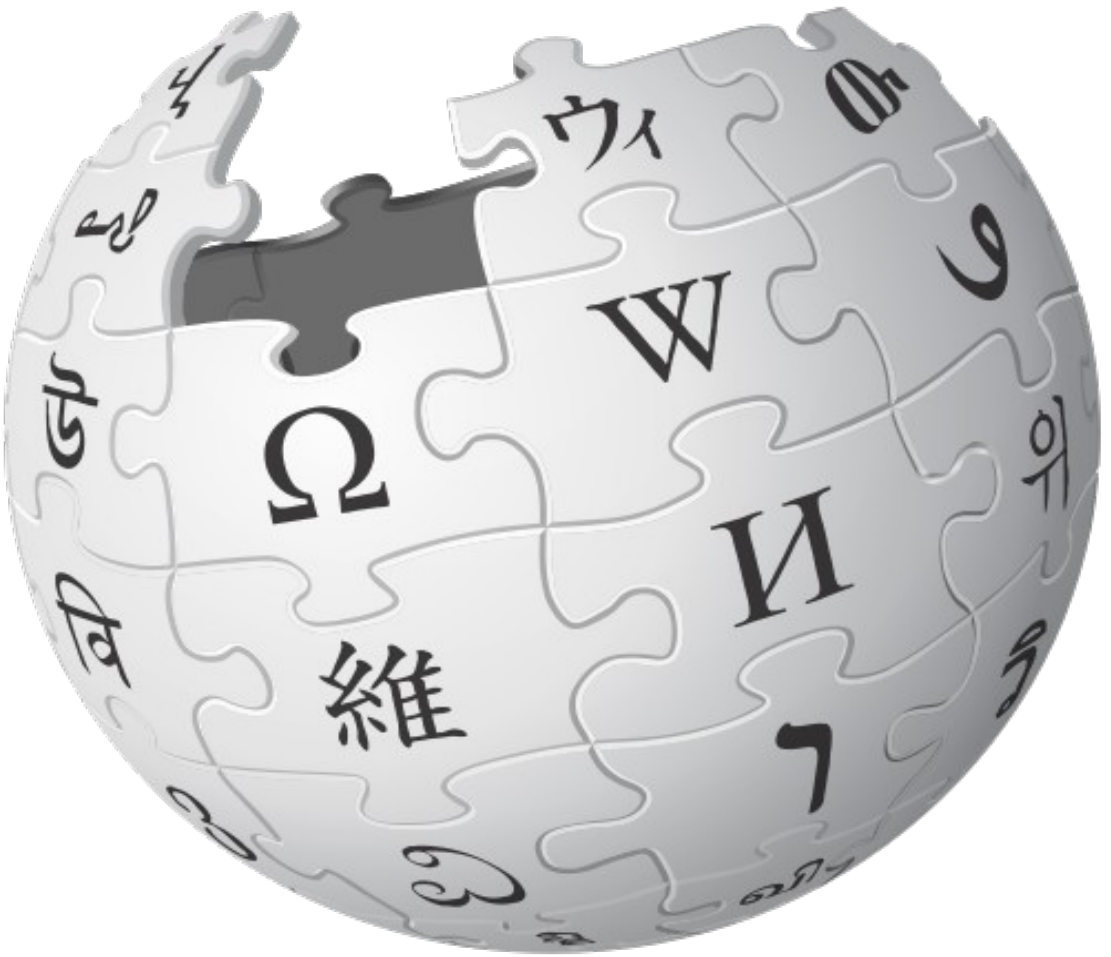
# Powering the world's favorite wiki - Wikipedia

Tomasz Finc – Engineering Program Manager

MediaWiki



# Three Questions

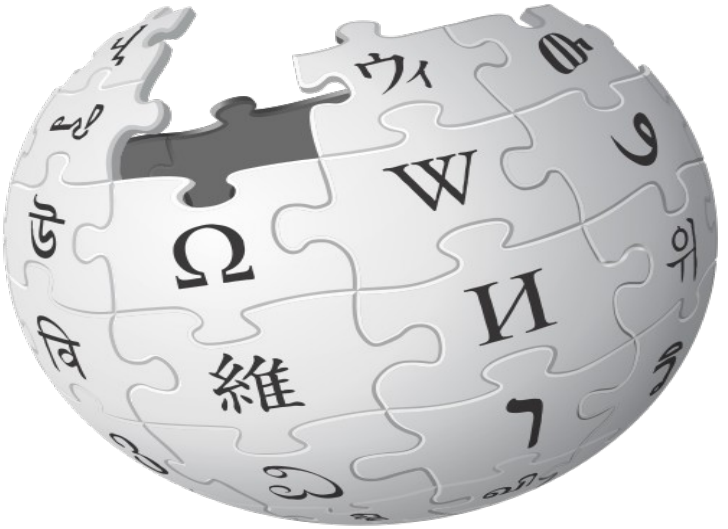
How many of you have used Wikipedia?

How many of you have clicked on the edit button?

How many of you clicked save?

# It all started in 2001

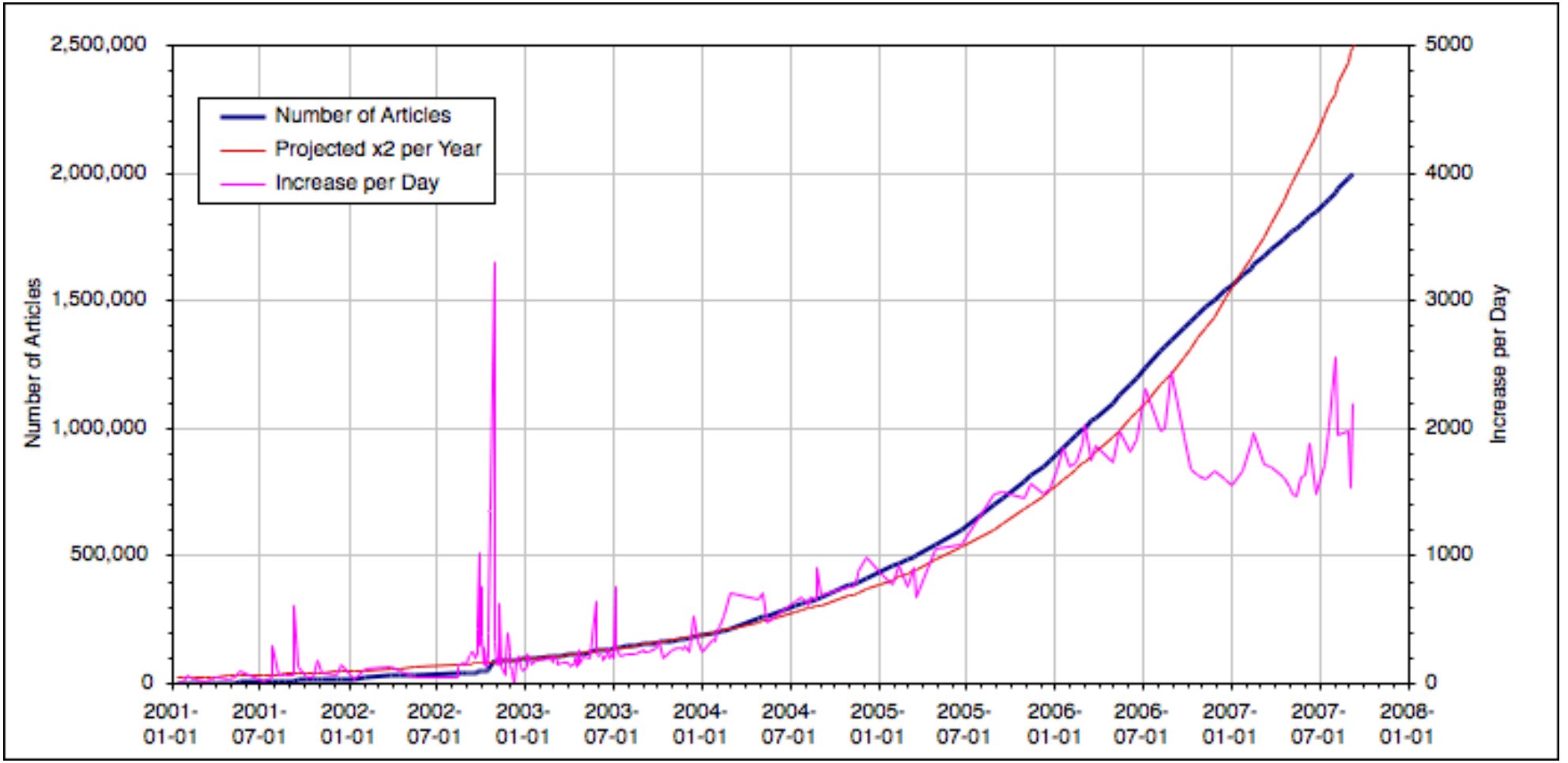
UPEDIA.COM
THE OPEN CONTENT ENCYCLOPEDIA



"Anyone else (yes, absolutely anyone else) can come along and make absolutely any changes to it that he wants to."

- Larry Sanger

# Edit Demo


Number of Articles
Projected x2 per Year
Increase per Day
Number of Articles
Increase per Day
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
5000
4000
3000
2000
1000
0
2001-01-01
2001-07-01
2002-01-01
2002-07-01
2003-01-01
2003-07-01
2004-01-01
2004-07-01
2005-01-01
2005-07-01
2006-01-01
2006-07-01
2007-01-01
2007-07-01
2008-01-01

“Imagine a world in which every single person on the planet is given free access to the sum of all human knowledge. That's what we're doing.”

- Jimmy Wales

# And we've grown since

And we've grown since … a lot

# 5th largest website in global web traffic

- Comscore

But how much traffic is that really?

12,933,000,000 / month

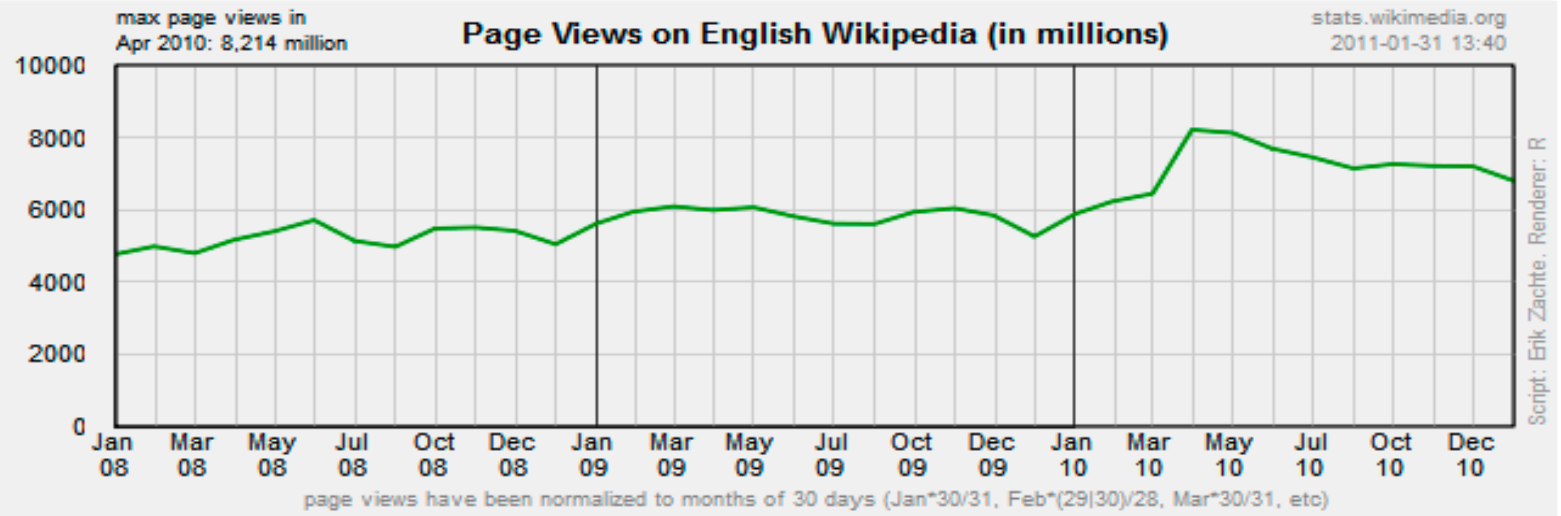
431,000,000 / day

18,000,000 / hour

299,000 / minute

**5000 / second**

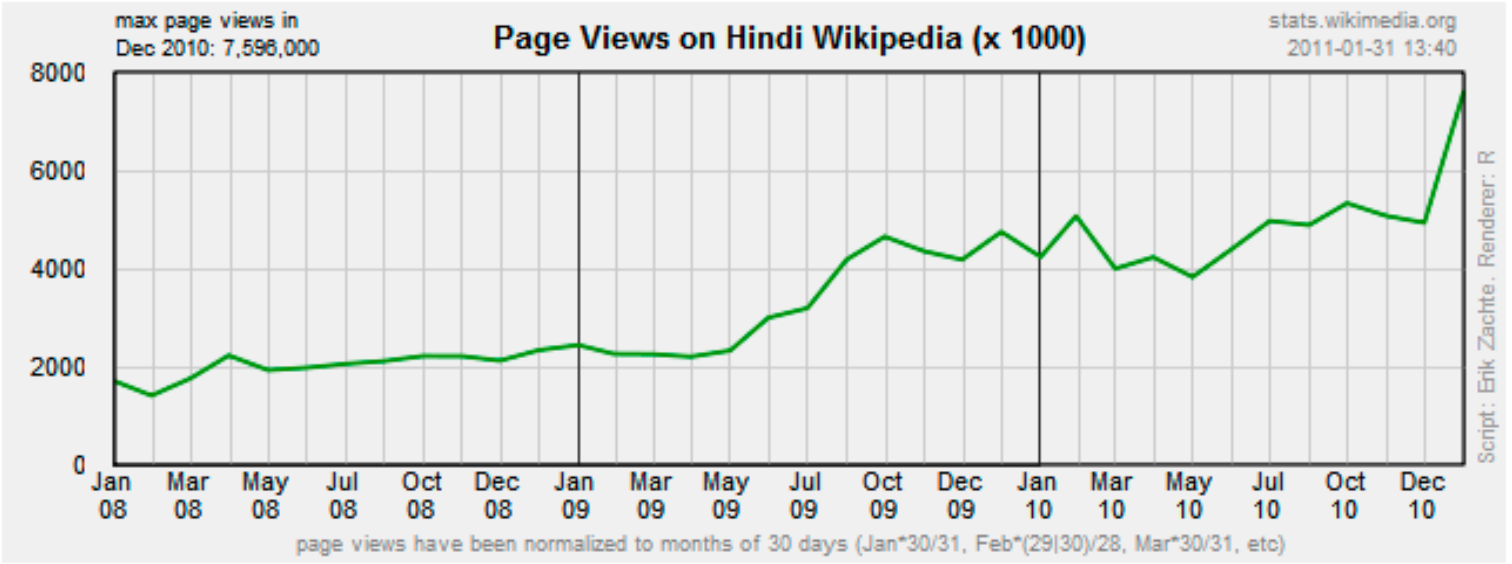
Source: http://stats.wikimedia.org/EN/TablesPageViewsMonthly.htm




page views: 7,141 million per month = 238 million per day = 9.9 million per hour = 165 thousand per minute = 2.8 thousand per second

*Source stats.wikimedia.org / Published 31 January 2011*

---




page views: 4.5 million per month = 151 thousand per day = 6.3 thousand per hour = 105 per minute = 1.7 per second

*Source stats.wikimedia.org / Published 31 January 2011*

But isn't Wikipedia done?

Aren't all the articles written?

Not even close

Is your home town on Wikipedia? … it should be

But remember, Wikipedia is an encyclopedia. No original work is done on Wikipedia and citations are required.

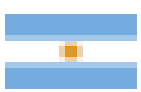
But how do you support it ?

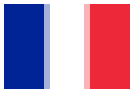
- 501(C) Non Profit
- 57 Staff members
- San Francisco Based
- Volunteer supported
- Growing

“Our mission is to empower a global volunteer community to collect and develop the world's knowledge and to make it available to everyone for free, for any purpose. We work together with a network of chapters in many different countries to achieve this goal.”

But its not just the foundation ....

# 29 Chapters and growing

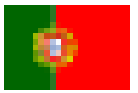




HK

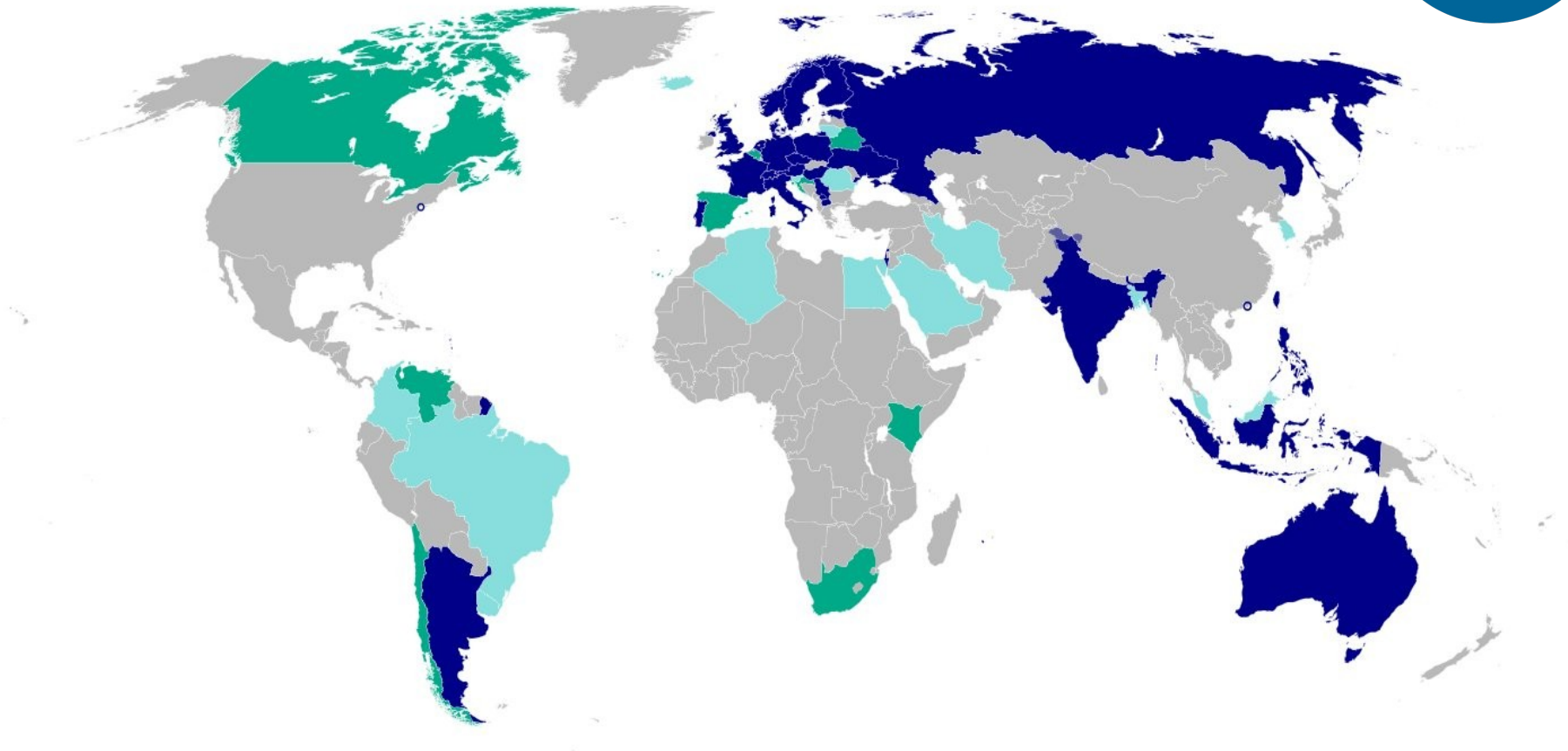
Argentina: Wikimedia Argentina
Australia: Wikimedia Australia
Austria: Wikimedia Österreich
Czech Republic: Wikimedia Česká republika
Denmark: Wikimedia Danmark
Estonia: Wikimedia Eesti
Finland: Wikimedia Suomi
France: Wikimedia Suomi
Germany: Wikimedia Deutschland
Hong Kong: 香港維基媒體協會
Hungary: Wikimédia Magyarország
India: Wikimedia India
Indonesia: Wikimedia Indonesia
Israel: Wikimedia Israel
Italy: Wikimedia Italia
Macedonia: Викимедија Македонија

Netherlands: Wikimedia Nederland
Norway: Wikimedia Norge/Noreg/Norga
Philippines: Wikimedia Philippines
Poland: Wikimedia Polska
Portugal: Wikimedia Portugal
Russia: Викимедиа РУ
Serbia: Wikimedia Србије
Sweden: Wikimedia Sverige
Switzerland: Wikimedia CH
Taiwan: 中華民國維基媒體協會
Ukraine: Вікімедіа Україна
United Kingdom: Wikimedia UK
United States, New York City: Wikimedia New York City



RUS

# Global Movement



- Founded
- In Discussion
- Planned
- Approved

And its not just the foundation and the chapters

Its YOU!

But how do we make this work?

# Open Source

php

MySQL®

SQUID

civiCRM

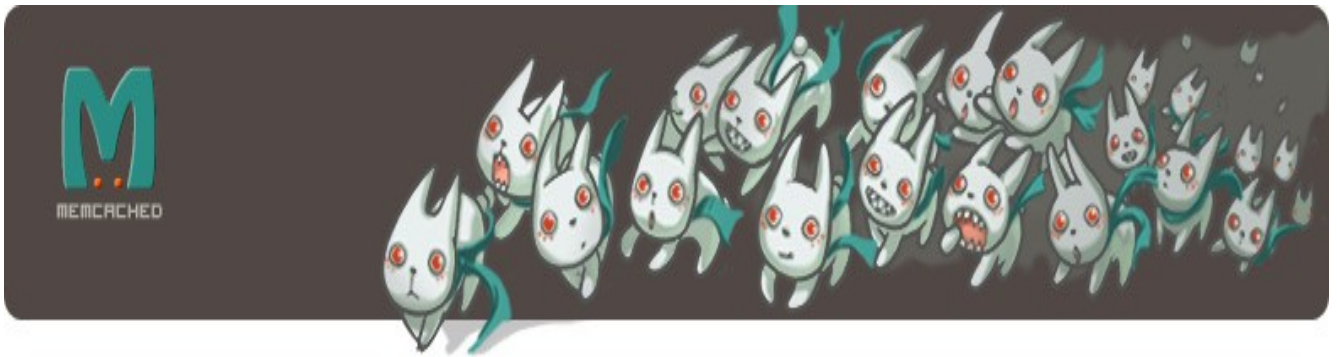
MediaWiki

LIGHTTPD
fly light.

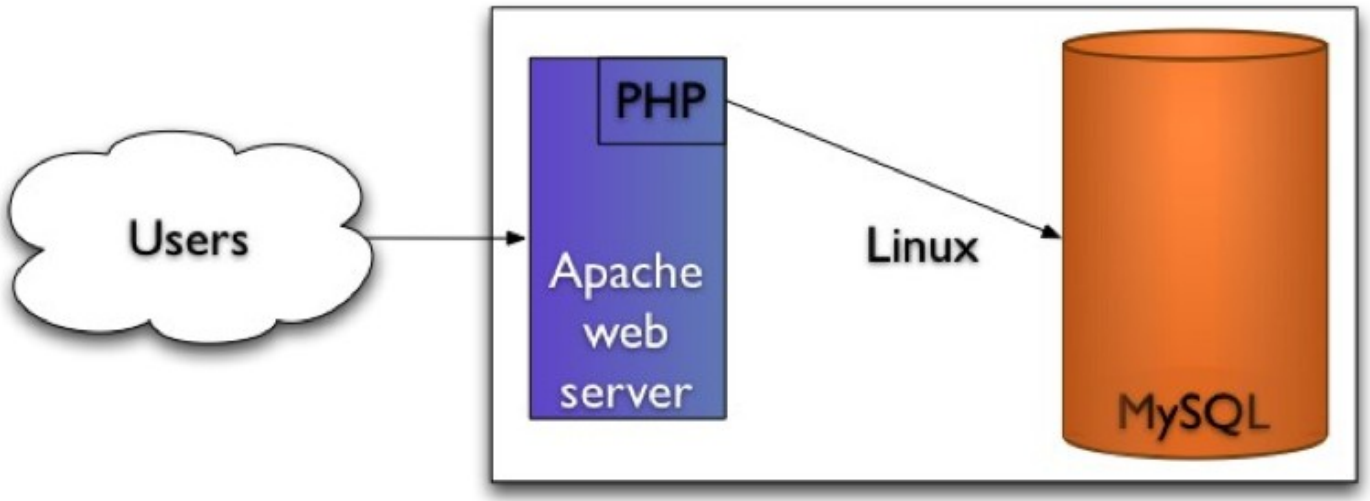
Lucene™

python™

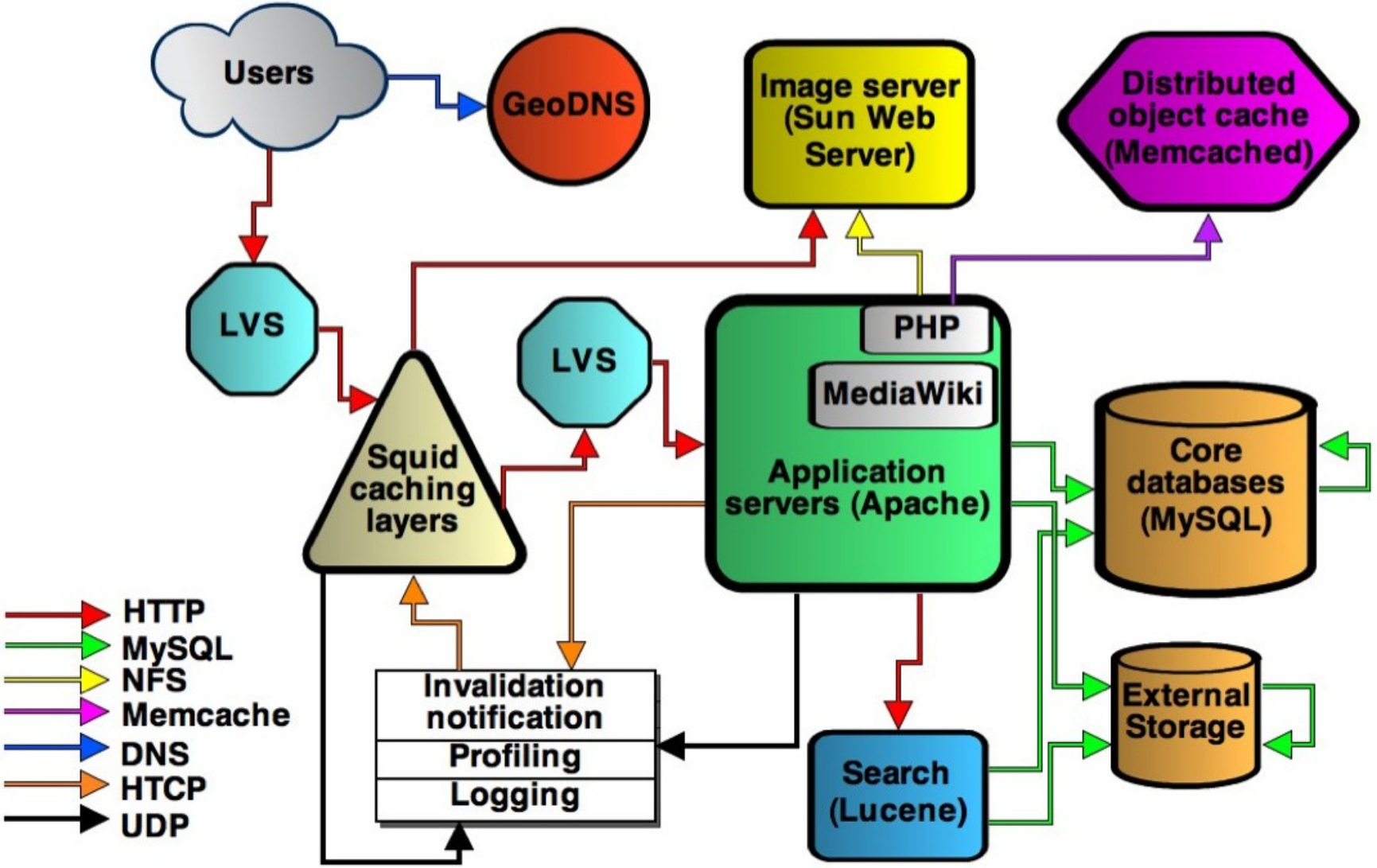
VARNISH
CACHE


MEMCACHED

# Architecture: LAMP...



# ...on steroids.


Users
GeoDNS
Image server (Sun Web Server)
Distributed object cache (Memcached)
LVS
LVS
PHP
MediaWiki
Application servers (Apache)
Squid caching layers
Core databases (MySQL)
Invalidation notification
Profiling
Logging
Search (Lucene)
External Storage
HTTP
MySQL
NFS
Memcache
DNS
HTCP
UDP

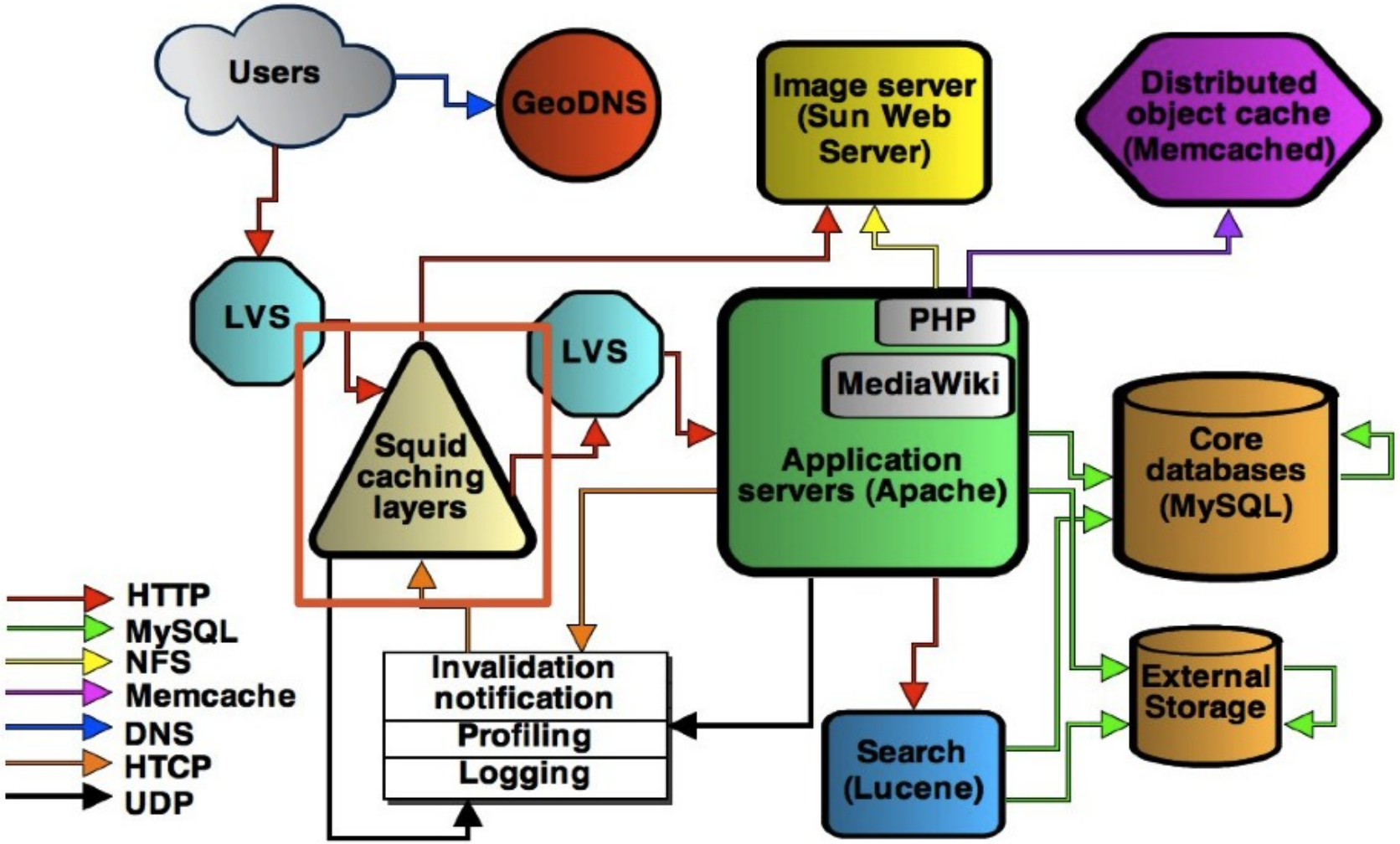
How many servers does it take to run all of the Wikimedia Projects?

~ 400 servers

But how can you scale your site to so many queries with so few machines?

# Optimization

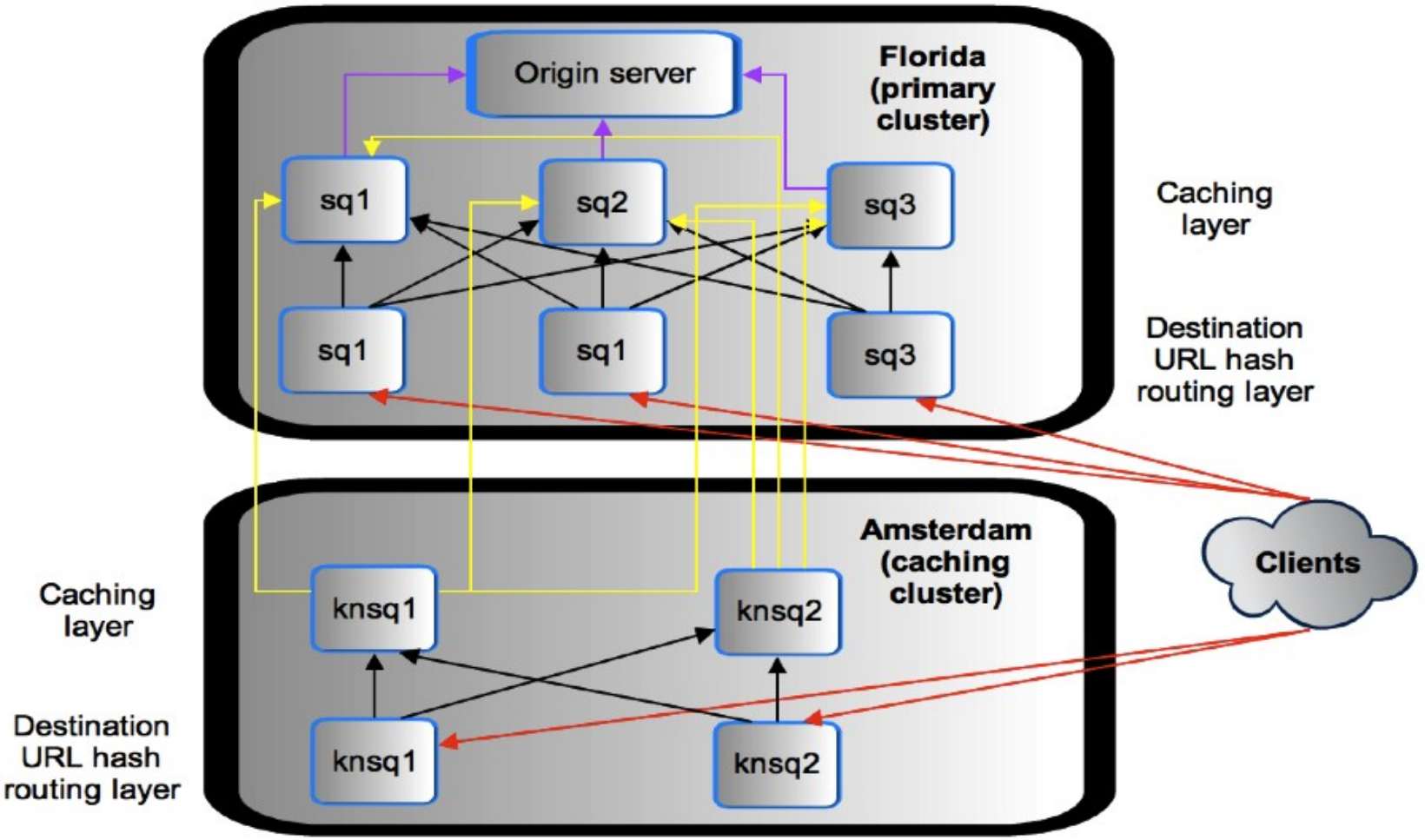
- Not doing anything inefficiently
- Aggressively profile code
- Cache, Cache, Cache


Users
GeoDNS
Image server (Sun Web Server)
Distributed object cache (Memcached)
LVS
LVS
Squid caching layers
PHP
MediaWiki
Application servers (Apache)
Core databases (MySQL)
External Storage
Search (Lucene)
Invalidation notification
Profiling
Logging
HTTP
MySQL
NFS
Memcache
DNS
HTCP
UDP

# Squid Caching

- Cache reverse proxy
- ~ 100 machines
- Serves most of our traffic
- Split into multiple groups
- Hit rates: 85% for Text, 98% for Media, since the use of CARP

# CARP

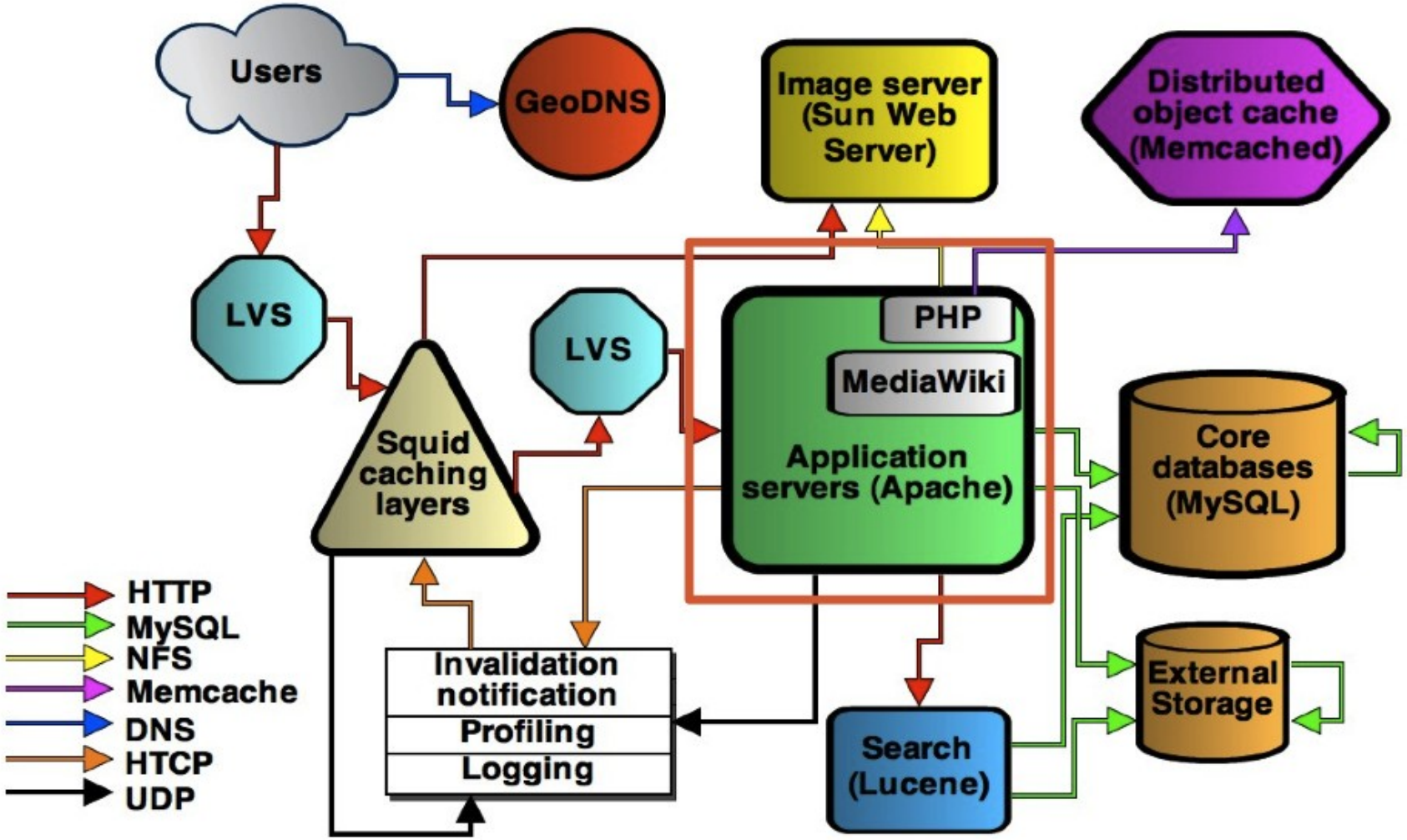

Origin server
Florida
(primary
cluster)
sq1
sq2
sq3
sq1
sq1
sq3
Caching
layer
Destination
URL hash
routing layer
Amsterdam
(caching
cluster)
knsq1
knsq2
knsq1
knsq2
Caching
layer
Destination
URL hash
routing layer
Clients

# Squid Cache Invalidation

- Wiki pages are edited at an unpredictable rate
- Users should always see the latest version
- Invalidation through expiry not acceptable
- Purge implemented using multicast UDP based HTCP protocol

# Varnish Caching

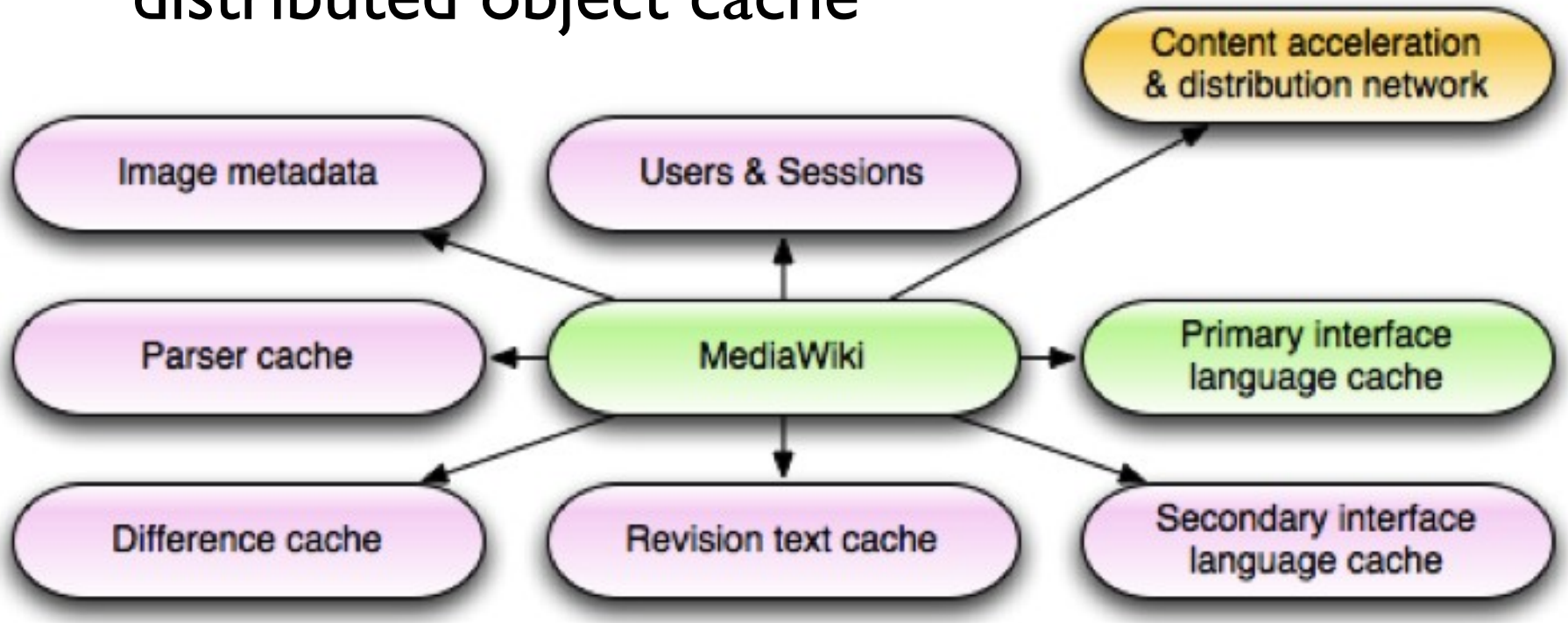
- Currently used for delivering static content such as javascript and css
- Will eventually replace squids
- 2-3 time more efficient than Squid


Users
GeoDNS
Image server (Sun Web Server)
Distributed object cache (Memcached)
LVS
LVS
PHP
MediaWiki
Application servers (Apache)
Squid caching layers
Core databases (MySQL)
Invalidation notification
Profiling
Logging
Search (Lucene)
External Storage
HTTP
MySQL
NFS
Memcache
DNS
HTCP
UDP

- Extension Framework
- Translated into 300+ languages
- Lightweight Wiki Markup
- Extensible API
- Rich content support
- High level of customization
- And much more ...

# MediaWiki Caching

- Caches are everywhere
- Most of the data is cached in Memcache, a distributed object cache

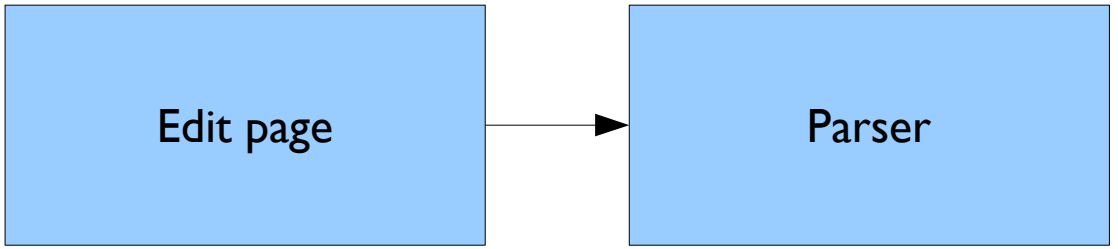


But what happens when I click save … ?

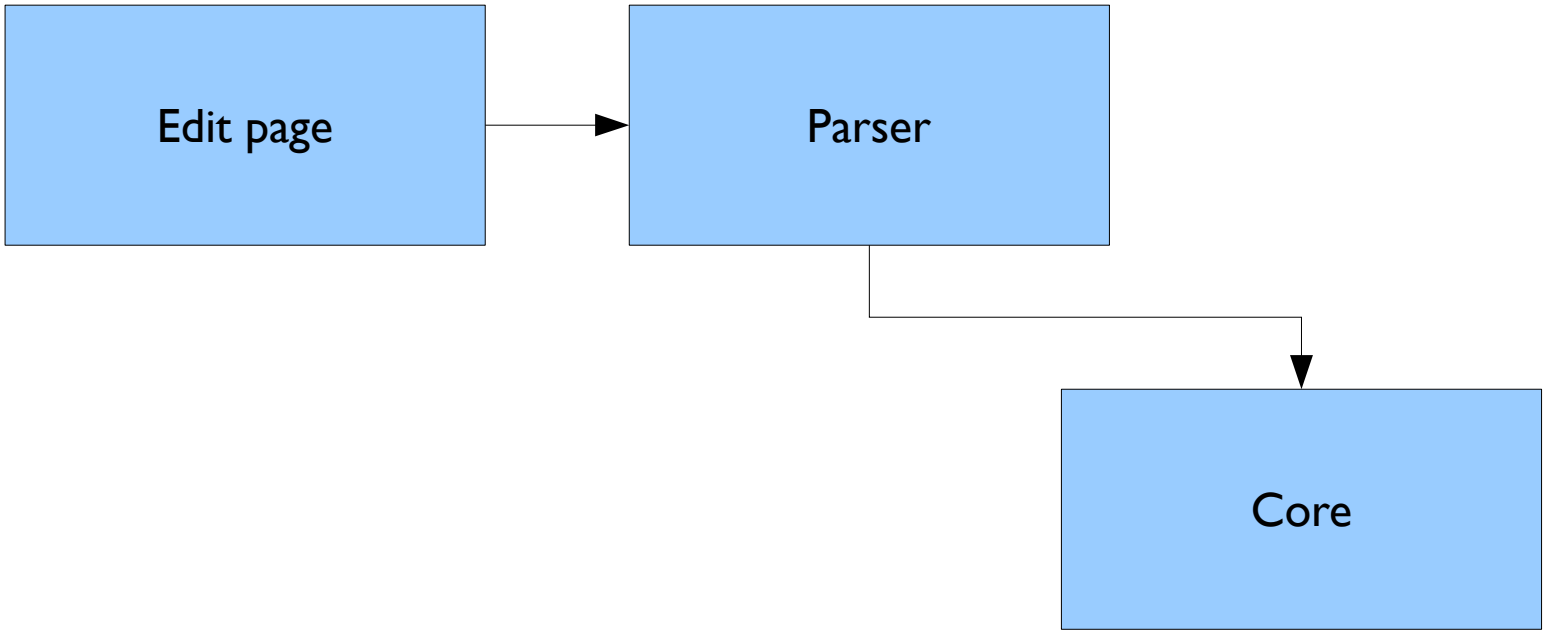
# So how does an edit work?

Edit page

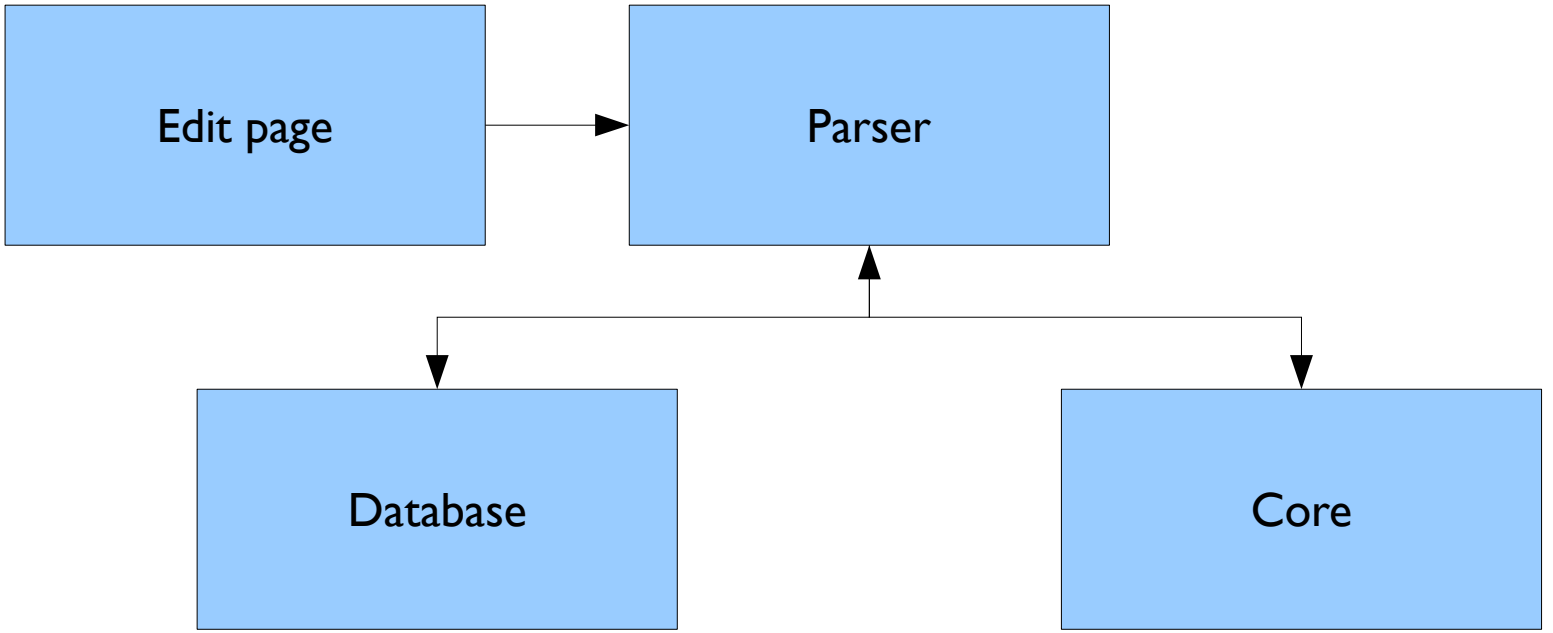
# So how does an edit work?



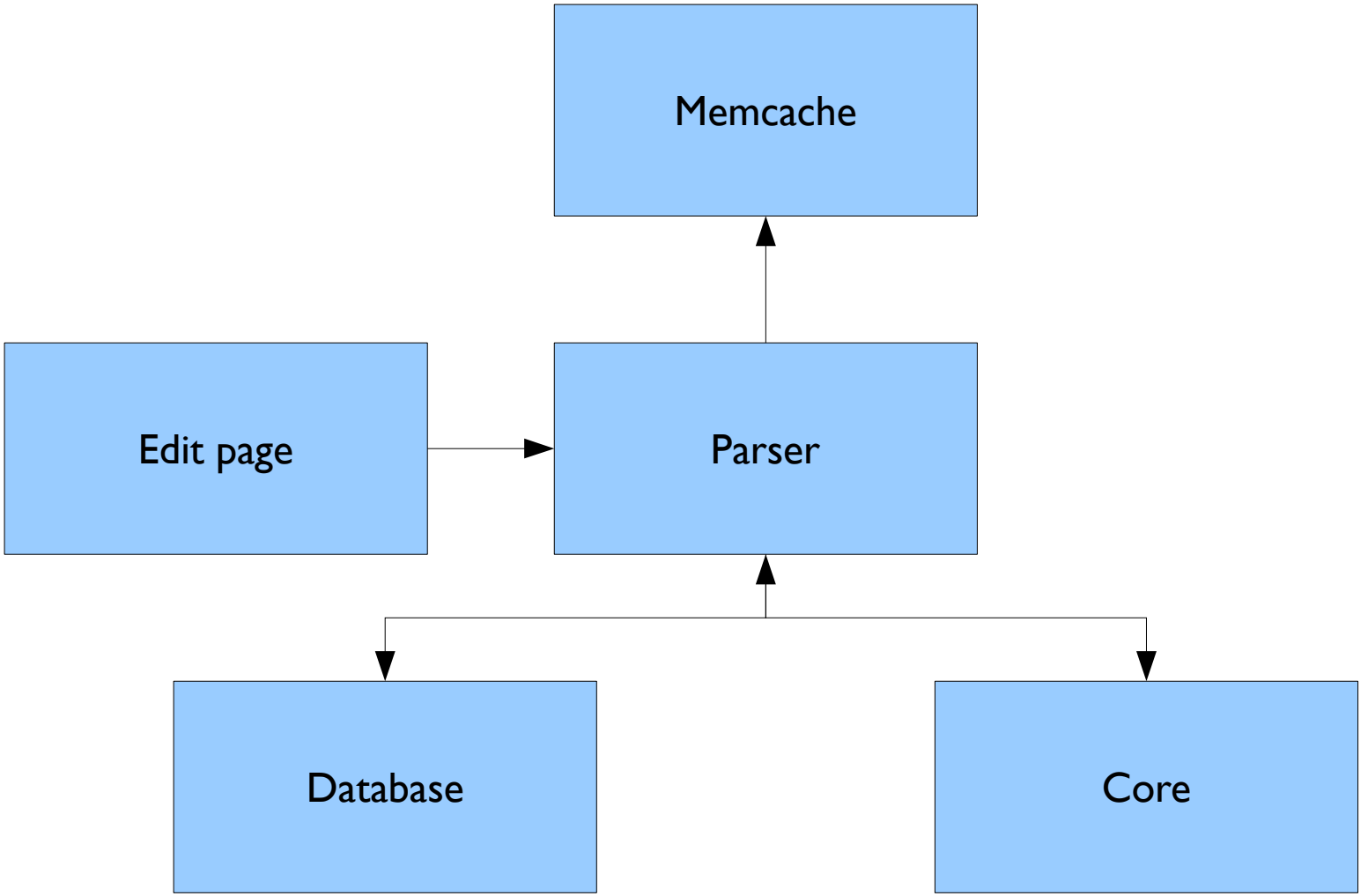
# So how does an edit work?



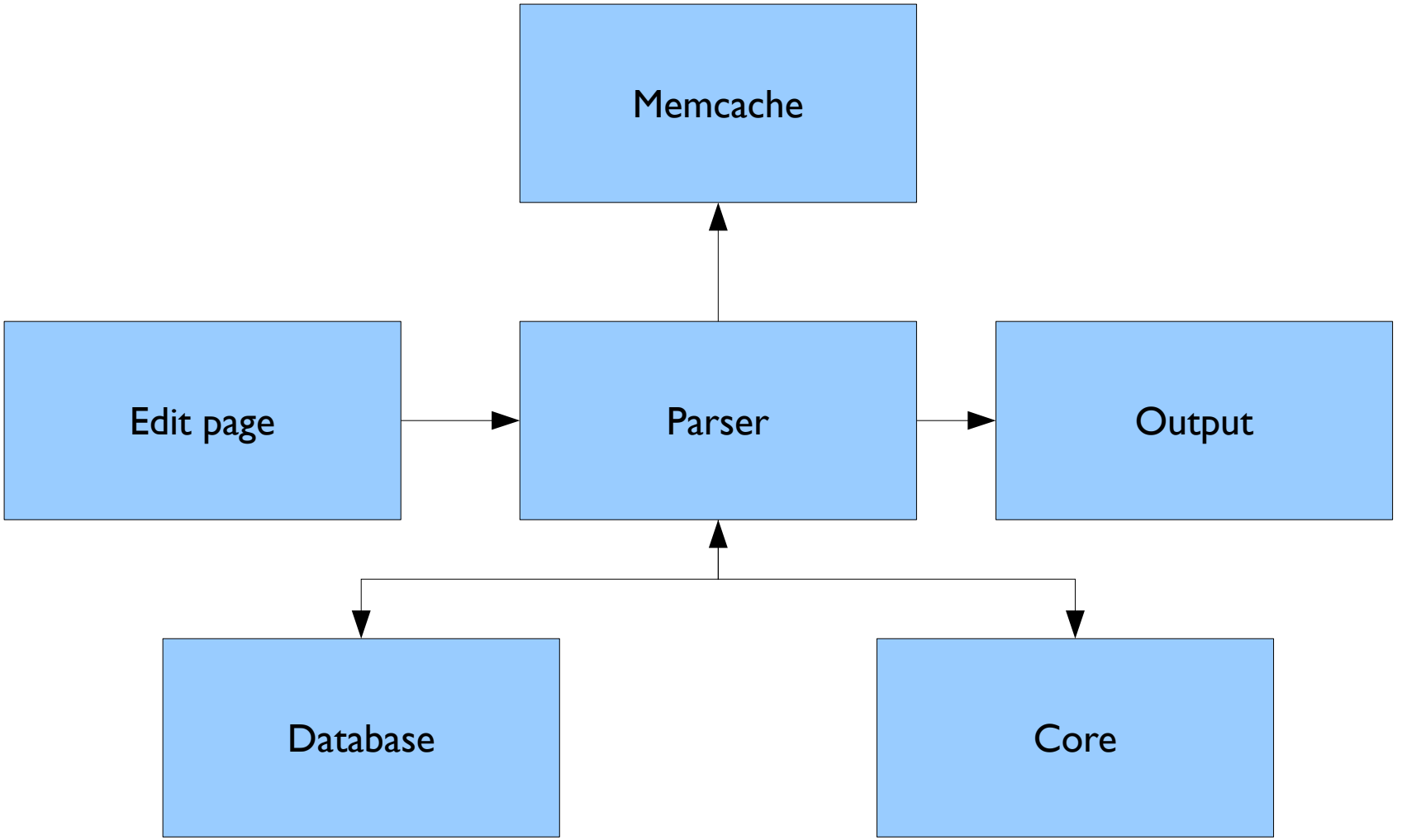
# So how does an edit work?



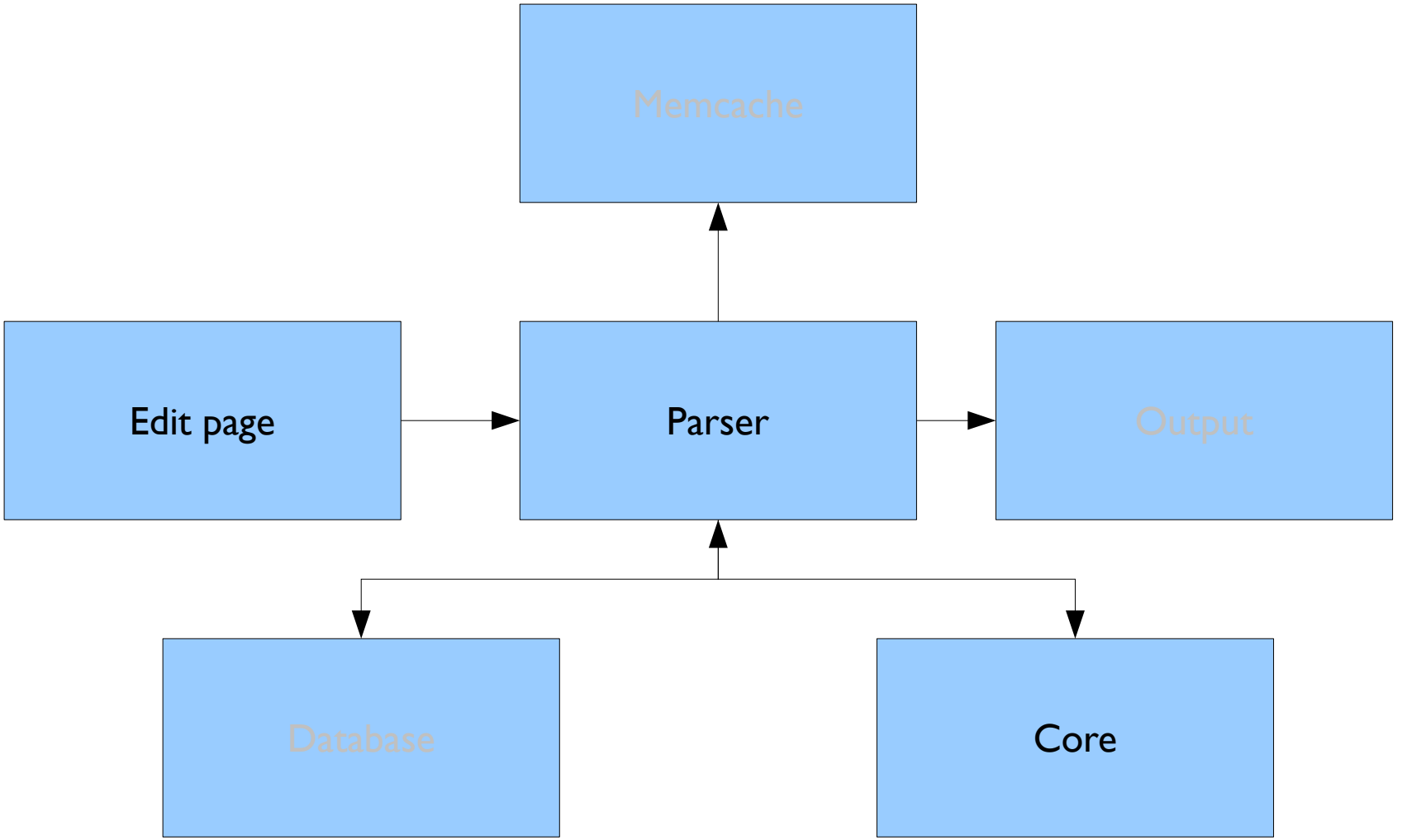
# So how does an edit work?



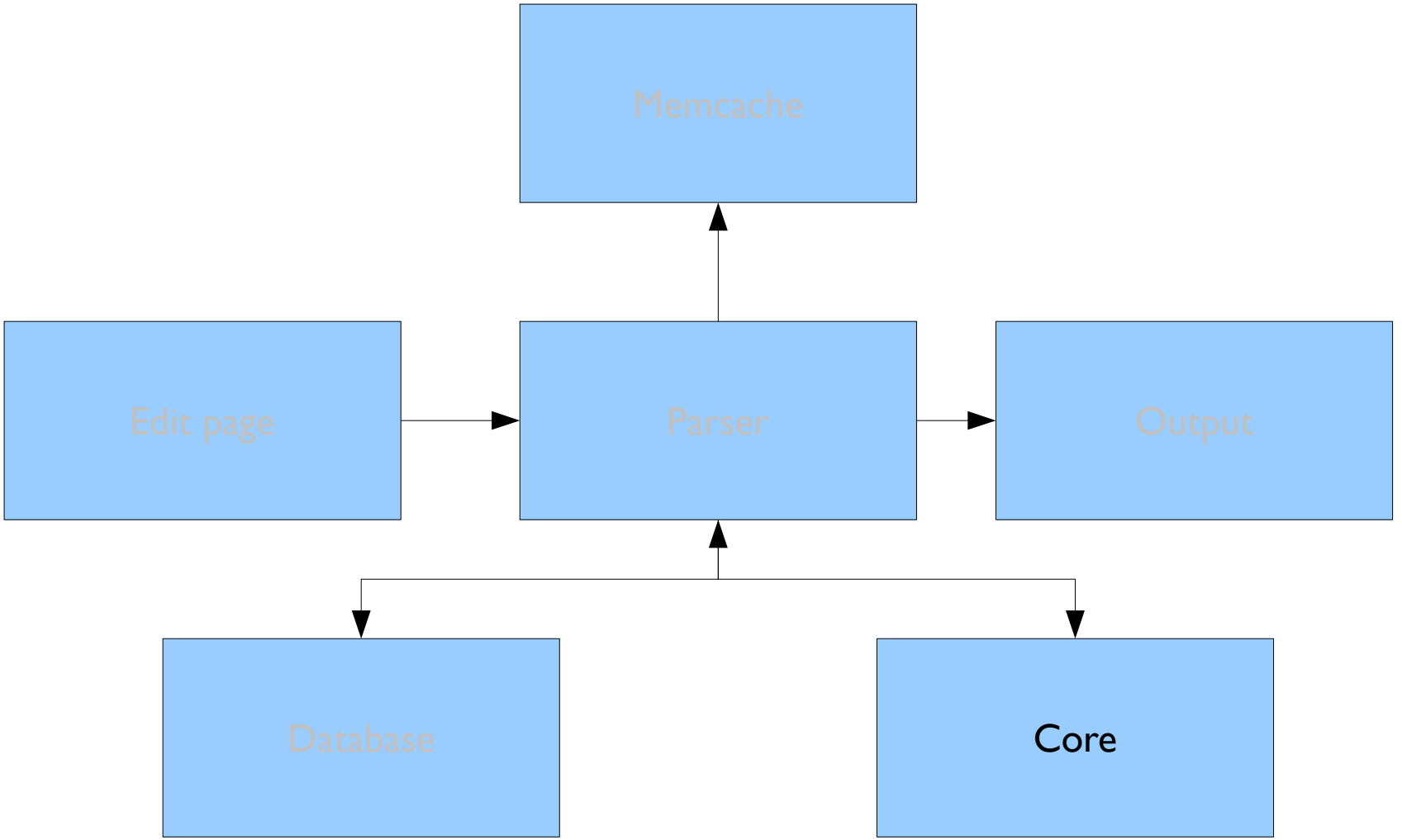
# So how does an edit work?



# So how does an edit work?



# So how does an edit work?



# MediaWiki Core

- Page
- Revision
- Archive
- Text
- User
- Logging
- Recent changes
- IP Block
- Watch lists
- Users/Groups
- Image
- File Archive
- Page links
- Category
- Redirect
- ....

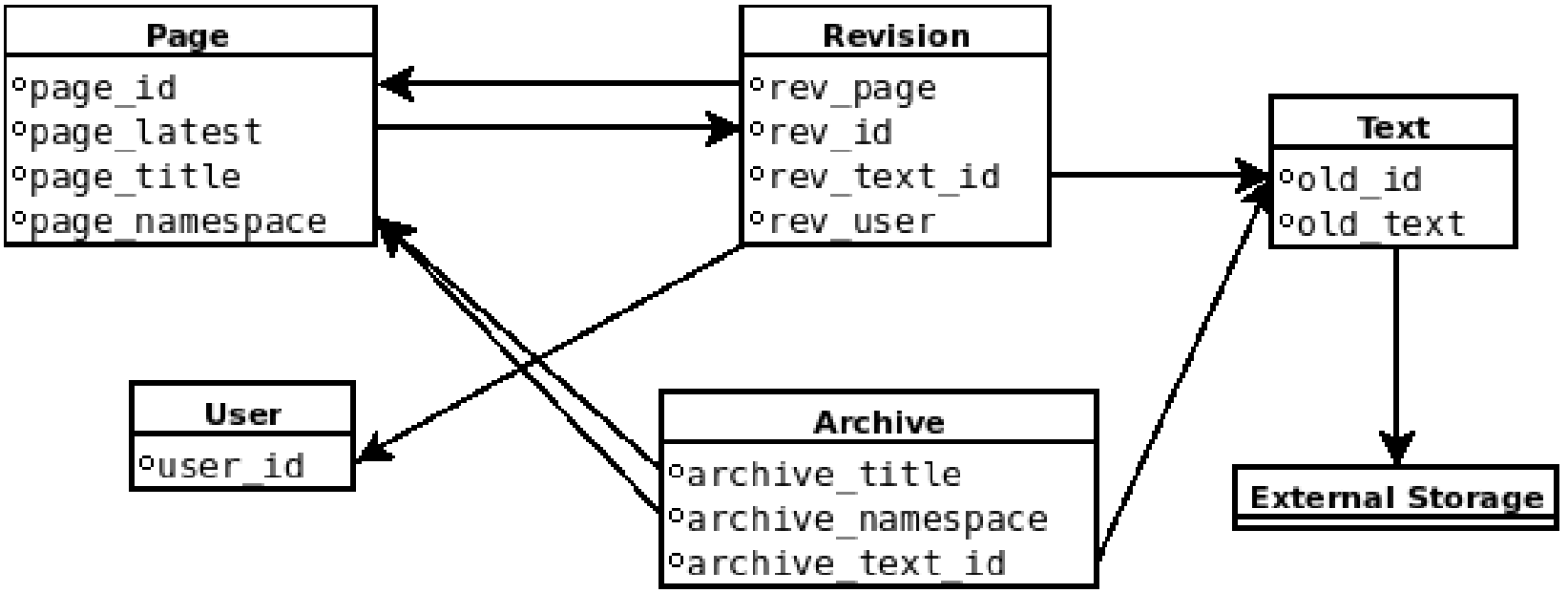
# MediaWiki Core

- Page
- Revision
- Archive
- Text
- User
- Logging
- Recent changes
- IP Block
- Watch lists
- Users/Groups
- Image
- File Archive
- Page links
- Category
- Redirect
- ....

# MediaWiki Core

- Page
- Revision
- Archive
- Text
- User
- Logging
- Recent changes
- IP Block
- Watch lists
- Users/Groups
- Image
- File Archive
- Page links
- Category
- Redirect
- ....

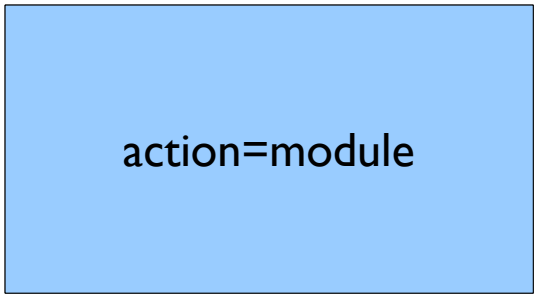
# Page Schema


Page
page_id
page_latest
page_title
page_namespace
Revision
rev_page
rev_id
rev_text_id
rev_user
Text
old_id
old_text
User
user_id
Archive
archive_title
archive_namespace
archive_text_id
External Storage

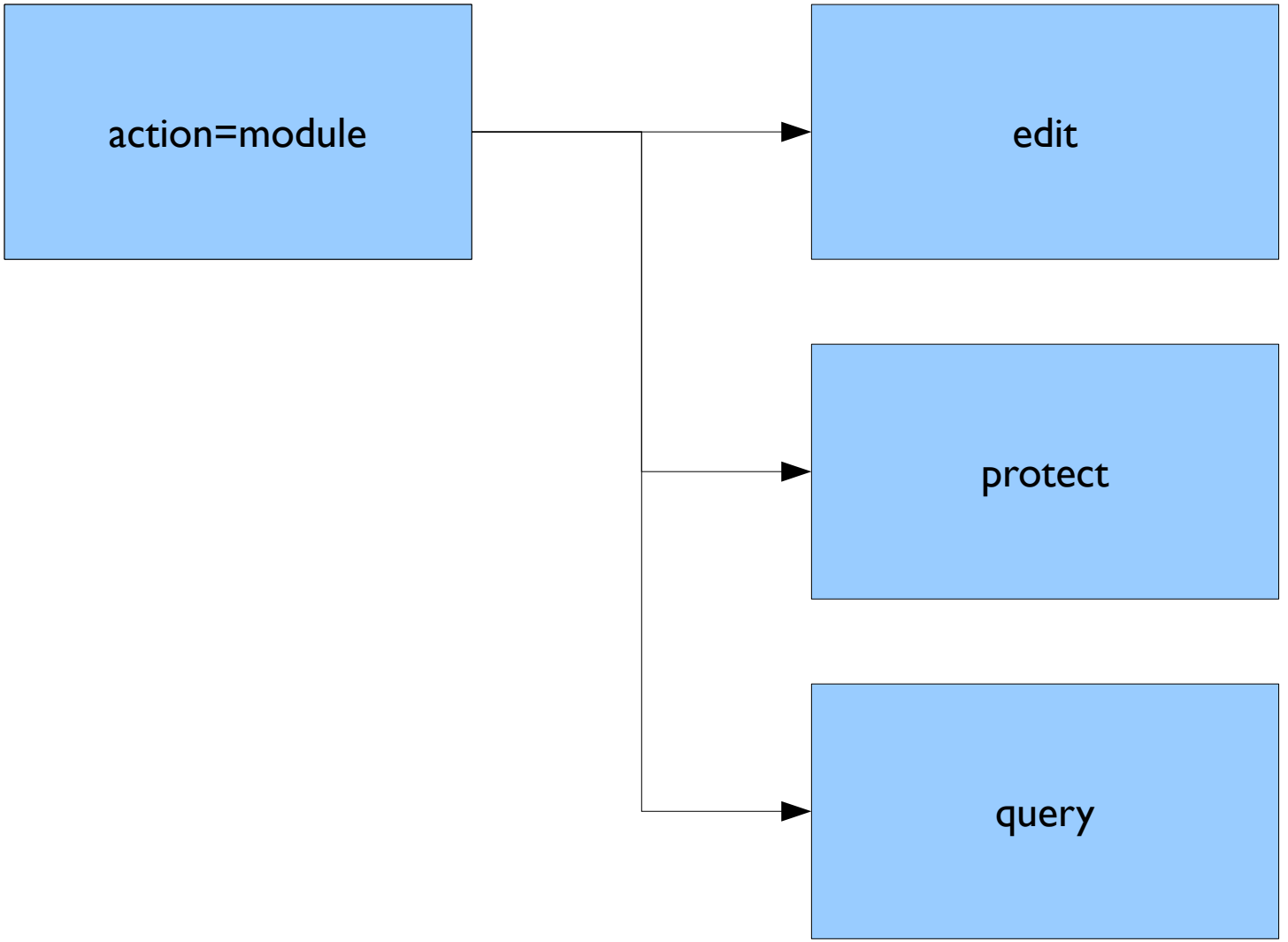
There are lots more sub components

But what if I want to script an addition to Wikipedia?

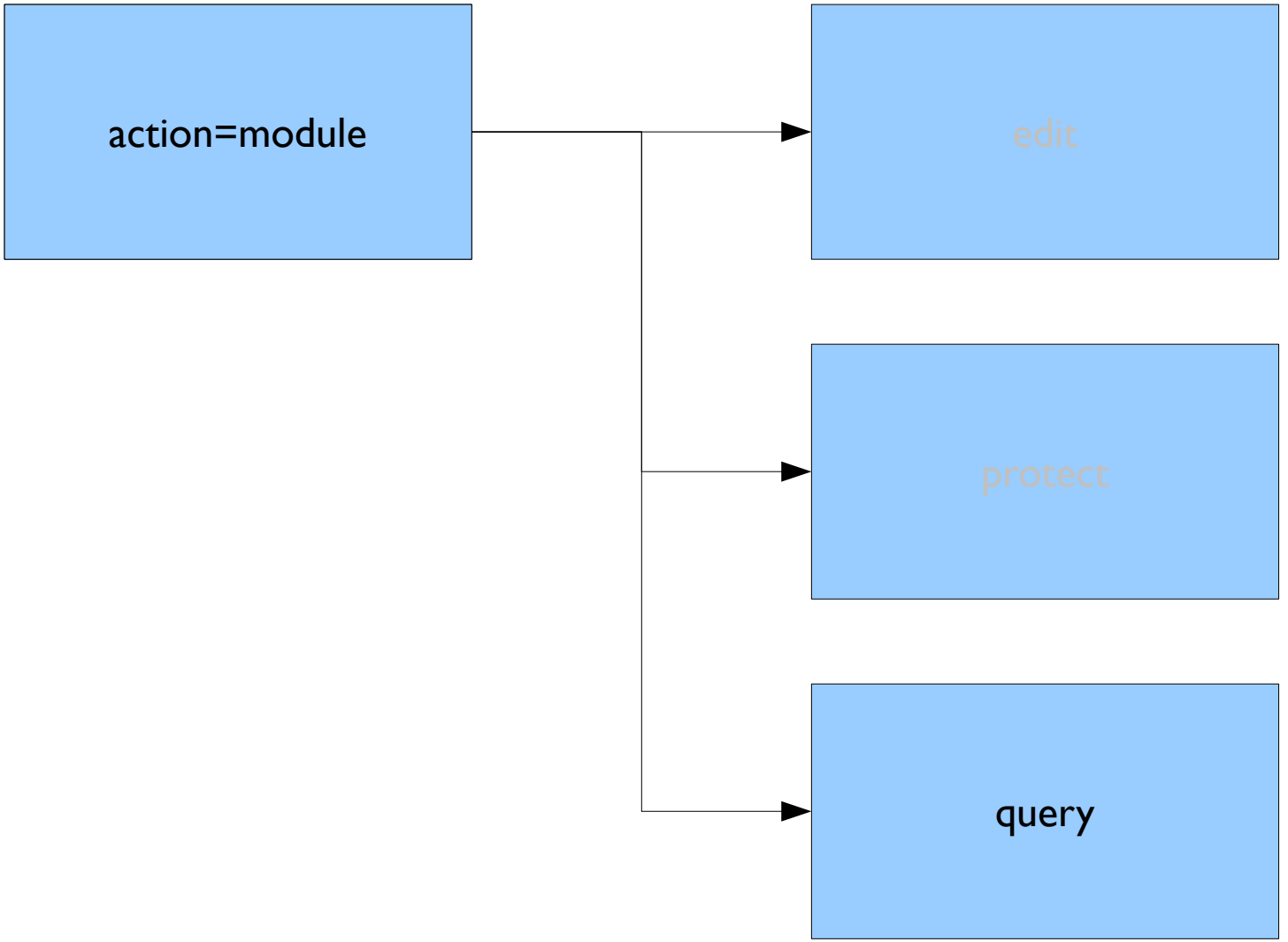
# API



# API



# API



# action = query

- Properties
  - Revisions, links, templates, revision content
- List
  - Lists of pages that are .. links, image usage, categories
- Meta
  - Everything else
- 30+ sub modules

# That's great but what about if I want to add media ?

Source: lolcats.com

# Upload Module

- action = upload
- Source
  - Url
  - File
  - Previous failed upload
- Target filename

But what if the API doesn't have a feature that I want?

# Extending the API

- Extend ApiBase class
- Register it with $wgAPIModules
- Core modules that can be overwritten
  - APIGetAllowedParams
  - APIGetParamDescription
  - APIAfterExecute

# Example

```
23 // Extension credits that will show up on Special:Version
24 $wgExtensionCredits['other'][] = array(
25     'path' => __FILE__,
26     'name' => 'ApiSVGProxy',
27     'author' => 'Roan Kattouw',
28     'url' => 'http://www.mediawiki.org/wiki/Extension:ApiSVGProxy',
29     'version' => '1.0',
30     'descriptionmsg' => 'apisvgproxy-desc',
31 );
32
33 // Set up the new special page
34 $dir = dirname( __FILE__ ) . '/';
35 $wgAutoloadClasses['ApiSVGProxy'] = $dir . 'ApiSVGProxy.body.php';
36 $wgAPIModules['svgproxy'] = 'ApiSVGProxy';
37
38 $wgExtensionMessagesFiles['ApiSVGProxy'] = dirname( __FILE__ ) . '/ApiSVGProxy.i18n.php';
```

ApiSVGProxy.php

```
30     public function execute() {
31         $params = $this->extractRequestParams();
32         $title = Title::newFromText( $params['file'] );
33         $file = wfFindFile( $title );
34
35         // Verify that the file exists and is an SVG file
36         if ( !$file ) {
37             $this->dieUsage( 'The specified file does not exist',
38                 'nosuchfile', 404 );
39         }
40         ....
```

ApiSVGProxy.body.php

# Example

```
65      public function getAllowedParams() {
66          return array(
67              'file' => null,
68          );
69      }
70
71      public function getParamDescription() {
72          return array(
73              'file' => 'Name of the file to proxy (including File: prefix)',
74          );
75      }
76
77      public function getDescription() {
78          return array(
79              'Proxy an SVG file from a (possibly remote) repository.',
80              'The file must have the .svg extension and the image/svg+xml MIME type.',
81              'If not, this module will return a 403 response. If the file doesn\'t exist,',
82              'a 404 response will be returned. If fetching the file failed, a 500 response',
83              'will be returned.',
84          );
85      }
86
87      public function getVersion() {
88          return __CLASS__ . ': $Id: ApiSVGProxy.body.php 64584 2010-04-03 23:09:39Z siebrand $';
89      }
90  }
```

ApiSVGProxy.body.php

Further info:
http://ww.mediawiki.org/wiki/API:Extensions

Come help us ...

# What are the challenges facing Wikipedia in India?

# Challenges

- Input
- Rendering
- Search

# Input Methods

- Transliteration
- InScript keyboard support
- Floating keyboard

# Rendering

- Better font support across distributions
- Real time delivery of fonts
- Multi script issues

# Search

Great solutions but none of them are open

# Further Discussion Topics

- Wikimedia Mobile – Saturday @ 1:30 in room 704
- Wikimedia Offline tools – Come find me directly

# In closing

- Wikipedia is still growing
- It's a challenge to keep up with the grown
- We need your help to support the growth

# Questions

Tomasz Finc - tomasz@wikimedia.org

~

www.wikimediafoundation.org

http://wikitech.wikimedia.org/view/Presentations