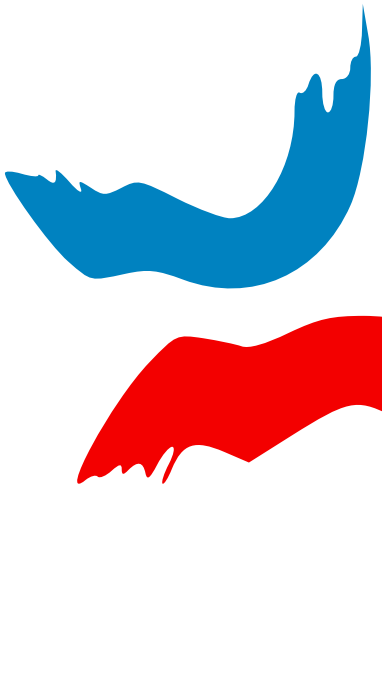
CirrusSearch

How we've replaced a great search engine with an awesome search engine

Nik Everett, Software Engineer
Chad Horohoe, Software Engineer
Dan Garry, Product Manager



WIKIMEDIA FOUNDATION

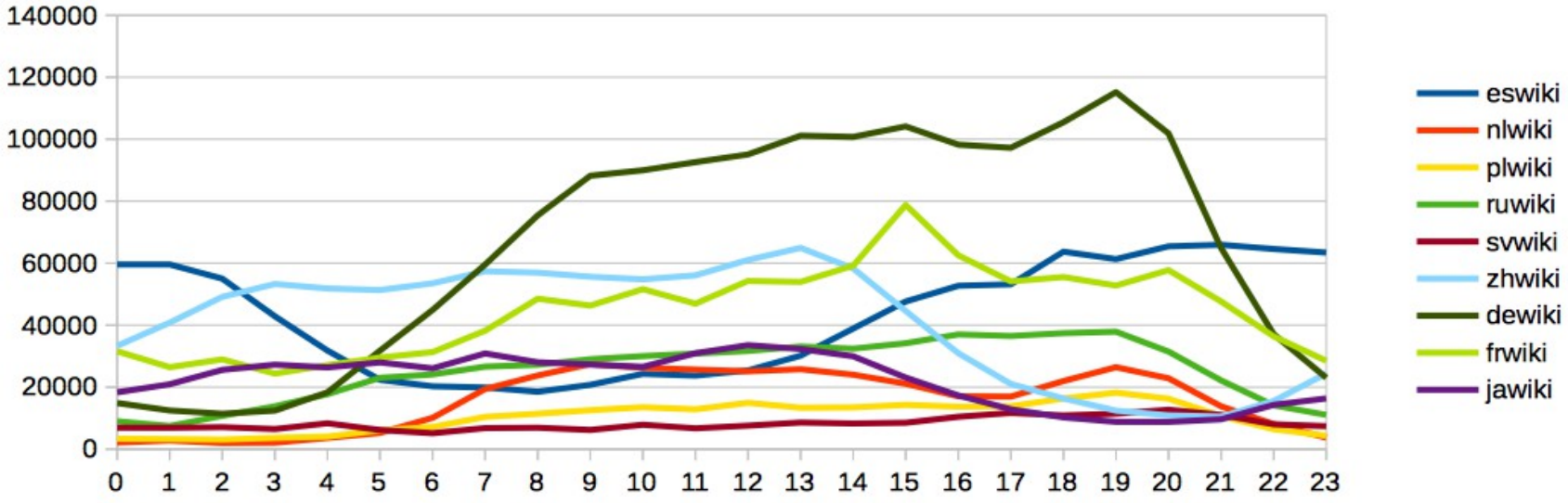
WIKIMANIA 2014

Who are we?

- Nik
- Dan
- Chad

Search at Wikimedia

- Search is really important
- People search a lot



History

- Starting in 2004, Wikipedia and sister sites needed better search
- Community wrote lucene-search and lucene-search-2
  - First in Java, then C#, then Java again
  - Based on Apache Lucene, major open source search backend
- MWSearch written as MediaWiki extension to provide integration

History

- Very feature rich
- Written by Wikipedians for Wikipedia sites, so it was (mostly) perfect for us
- Works surprisingly well and scaled great for years

Late 2012 to Early 2013

- As volunteers and staff came and went, institutional knowledge was lost
  - No third party users – no community support
- System became increasingly unstable
- A number of emergency fixes were made, but alternative was clearly needed

CirrusSearch

- "Cirrus" is a type of cloud
  - Clouds are web 2.0 buzzwords
  - Wanted a unique name too
- Set out with 3 goals in mind
  - Feature parity with existing search
  - Provide near-real-time results
  - Easy for third parties to use/extend as well

First steps

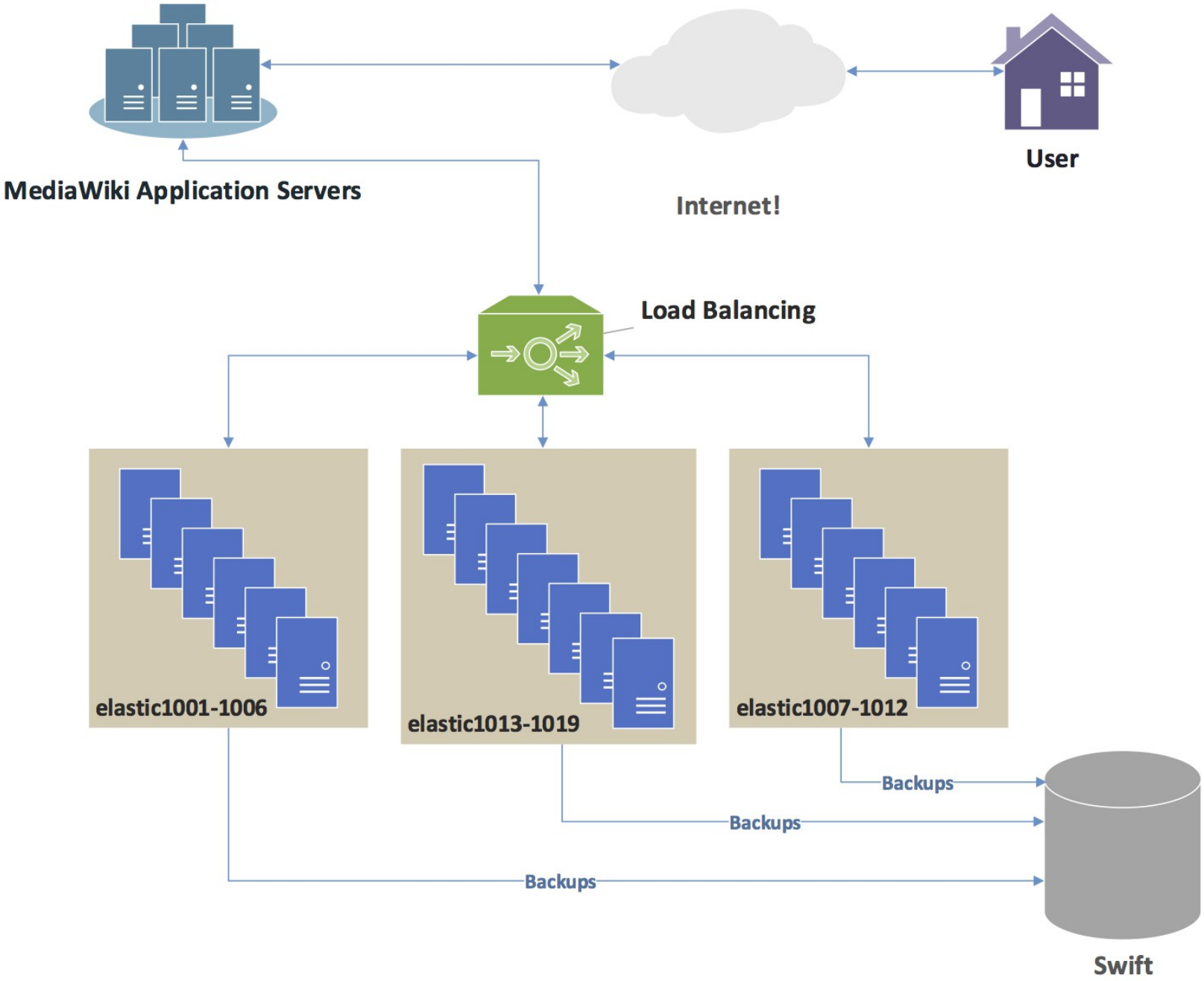
- Started by investigating Solr
  - Based on Lucene, which we'd used before
  - Has nice new cluster features
  - Had used before in limited areas (GeoData & Translation Memory)
- Prototyped this in about 3-4 weeks, mixed results
  - Worked, but also depended on Zookeeper
  - Schema management was hard

Second attempt

- Then investigated Elasticsearch
  - Also based on Lucene
  - Operations suggested it
- Prototyped in about 3-4 weeks, loved it
  - RESTful API was easy to develop against
  - Schema and cluster management were simple
  - Upstream very friendly & easy to work with

Elasticsearch it is

- After trying both and discussing it extensively, we stuck with Elasticsearch
- Began building a small cluster and deploying to a few test wikis
  - Italian and Catalan Wikipedians are early adopters. Provide valuable feedback


MediaWiki Application Servers
Internet!
User
Load Balancing
elastic1001-1006
elastic1013-1019
elastic1007-1012
Backups
Backups
Backups
Swift

Feature parity

- This is hard
  - Scoring algorithm is undocumented and untested
  - Features are undocumented
- Lots of expert syntax
  - Don't want to change it much, breaks backwards compatibility
- Think we got it pretty good

Things we did different

- (Near) Real time indexing
  - Old search updated via daily-ish batch job
  - Cirrus updates about every 30 seconds (usually)
- Expanding wikitext
  - Old search had bad behavior with templates
  - Better for readers, template-heavy wikis like Wiktionary
- Support other content models
  - Javascript and Wikidata aren't indexed like wikitext

Features you might expect

- intitle:
- incategory:
- Namespace-specific searching
- Prefix searching
- Fuzzy searching
- Multiple search terms

Extra things we did

- New syntax features
  - hastemplate:
  - insource:
- Searches inside of PDF/DjVu text as well as file description page
- Make it possible for wikis to curate the best—or worst—of their content in search results (think Featured Articles or Quality Pictures or Stubs)

Extra things we did

- Commons search results included for File namespace
  - All the time now
- Interwiki search results
  - Coming soon. Needs major UI work
- Tightened up logic when combining multiple query parameters
  - Category intersections on search work as expected

Things we're doing now

- Finish rolling out to all Wikimedia users
  - Already running most WMF wikis; only as beta on the largest Wikipedias & Commons
  - Next few weeks for them – more hardware on the way
- File metadata searching
  - File sizes/dimensions, mime types, Exif/XMP data, etc.
- More performance tuning

Things we want to do soon

- Improved support for less-common languages
- Faster faster faster!
- Nicer looking search results
- Search data replicated to labs like the databases

What do you want search to do?

Thank you!

Project page – mediawiki.org/wiki/Search

Slide deck – is.gd/VuIhHk