Offline Content Generator (OCG)

C. Scott Ananian <cscott@cscott.net>
Wikimedia Foundation

Quarterly Review (Q1)
October 3, 2014

Does an online encyclopedia have to be ugly to read offline?

Does an online encyclopedia have to be ugly to read offline?

(And can we better support users in the Global South?)

Context

DECEMBER 13, 2007 - ST. PETERSBURG, FLORIDA: The Wikimedia Foundation today announced a partnership that will make it possible to obtain high quality print and word processor copies of articles from Wikipedia and other wiki educational resources. The development of the underlying open source software is supported by the Open Society Institute (www.soros.org) and the Commonwealth of Learning (www.col.org), and led by PediaPress.com, a start-up company based in Germany.

"This technology is of key strategic importance to the cause of free education world-wide," said Sue Gardner, Executive Director of the Wikimedia Foundation. "It will make it possible to use and remix wiki content for a variety of purposes, both in the developing and the developed world, in areas with connectivity and without."

Time takes its toll

- The servers were installed when the wiki was young, and were unreproducible.
- The underlying wikitext translation was driven by `mwlib`, a 3rd-party wikitext parser, which was long in the tooth
- A large and growing number of open bugs filed against the service, especially with respect to non-Latin scripts
- This was the last service in the Tampa datacenter, which was costing WMF $30k/mo

Long story short

Two week-long PDF sprints in January and May

New Parsoid- and XeLaTeX-based PDF renderer

Entered production **this week** (September 29)

Team Effort: C. Scott Ananian, Matt Walker, Max Semenik, Brad Jorsch, Arlo Breault, others!

On-going maintenance / Q2 goals

The Offline Collection Generator service is robustly architected, with room for growth:

- Production
  - monitoring / robustness / documentation / bus factor
- Improvements to PDF rendering
  - tables, infoboxes, limited CSS support, CJK languages
- New backends
  - ZIM, ePub, plain text (already on beta), others...
- Next-generation PDF renderer:
  - Parsoid->PhantomJS v2->PDF

Some amount of progress on all of these expected in Q2

Q2: Next-generation PDF renderer

PDF backend based on PhantomJS v2

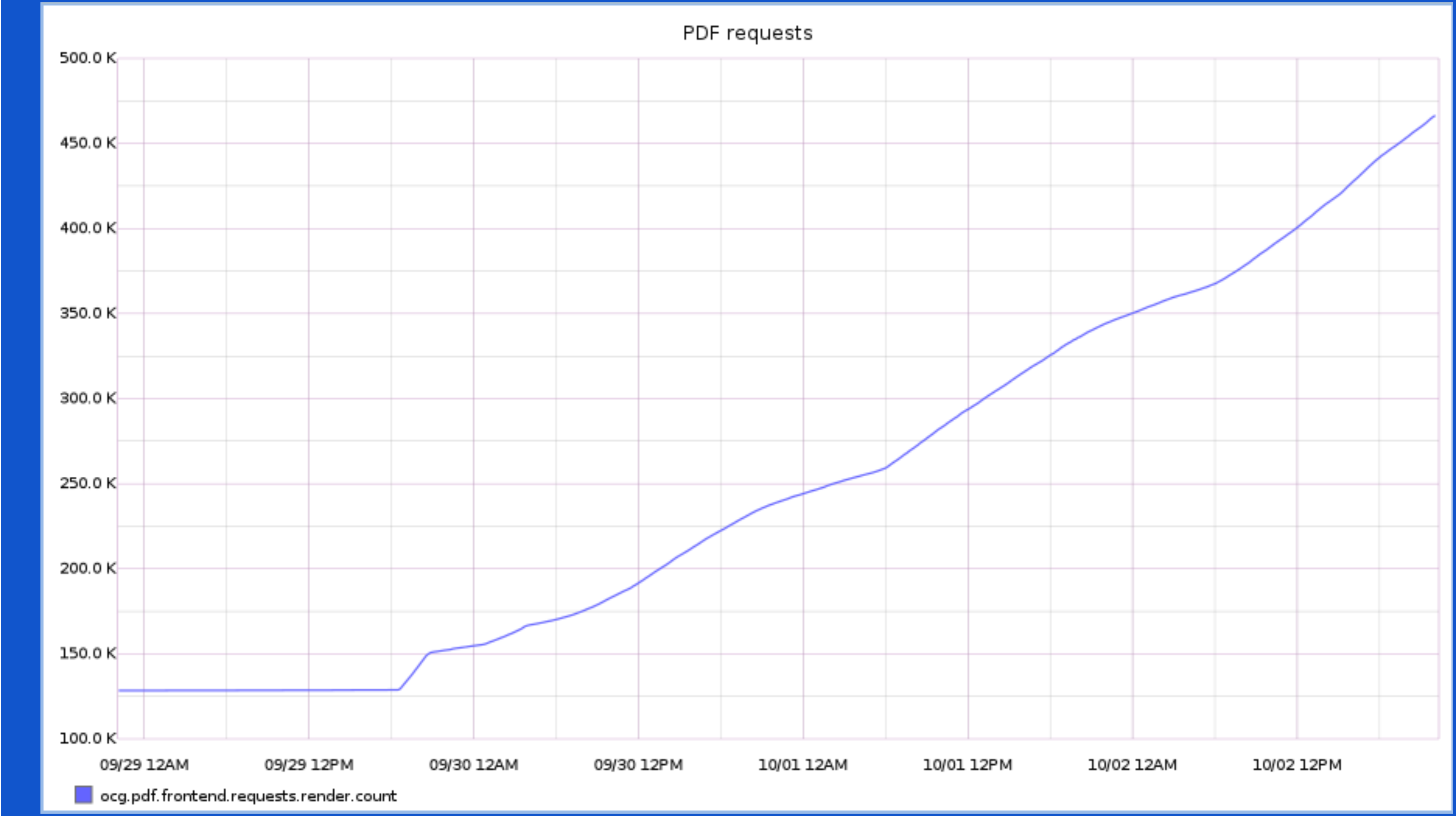
- Print CSS stylesheet for Parsoid HTML
    - could use help from design to make this beautiful
- Render Parsoid HTML + CSS directly to PDF
- Waiting on PhantomJS v2 upstream
    - but we can get started
- Simpler to maintain in the long run
    - but inferior typesetting, at least right now

Q2 goal: first draft (not production, due to uncertainties upstream)

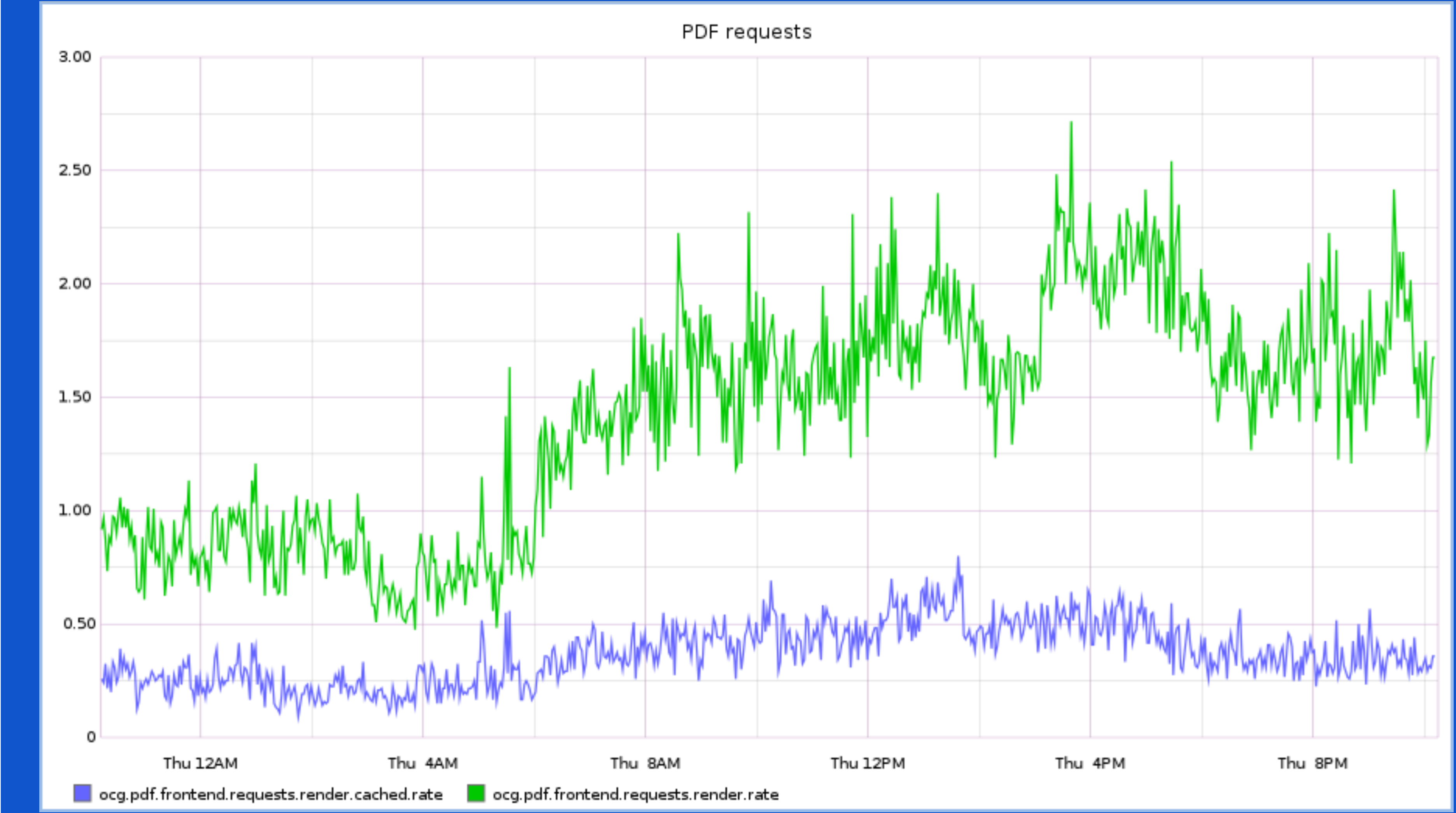
Spin-off: use Parsoid HTML + Print CSS for “Printable page” in production?

Unexpected uses

- The bundler component is a standalone article spider tool, useful for offlining
  - Talked with Kiwix folks about combining efforts here
  - Further extends the usefulness of Parsoid as a platform
- `texvcjs` component useful for Mathoid
- Plain-text backend used for DNA-based art project!

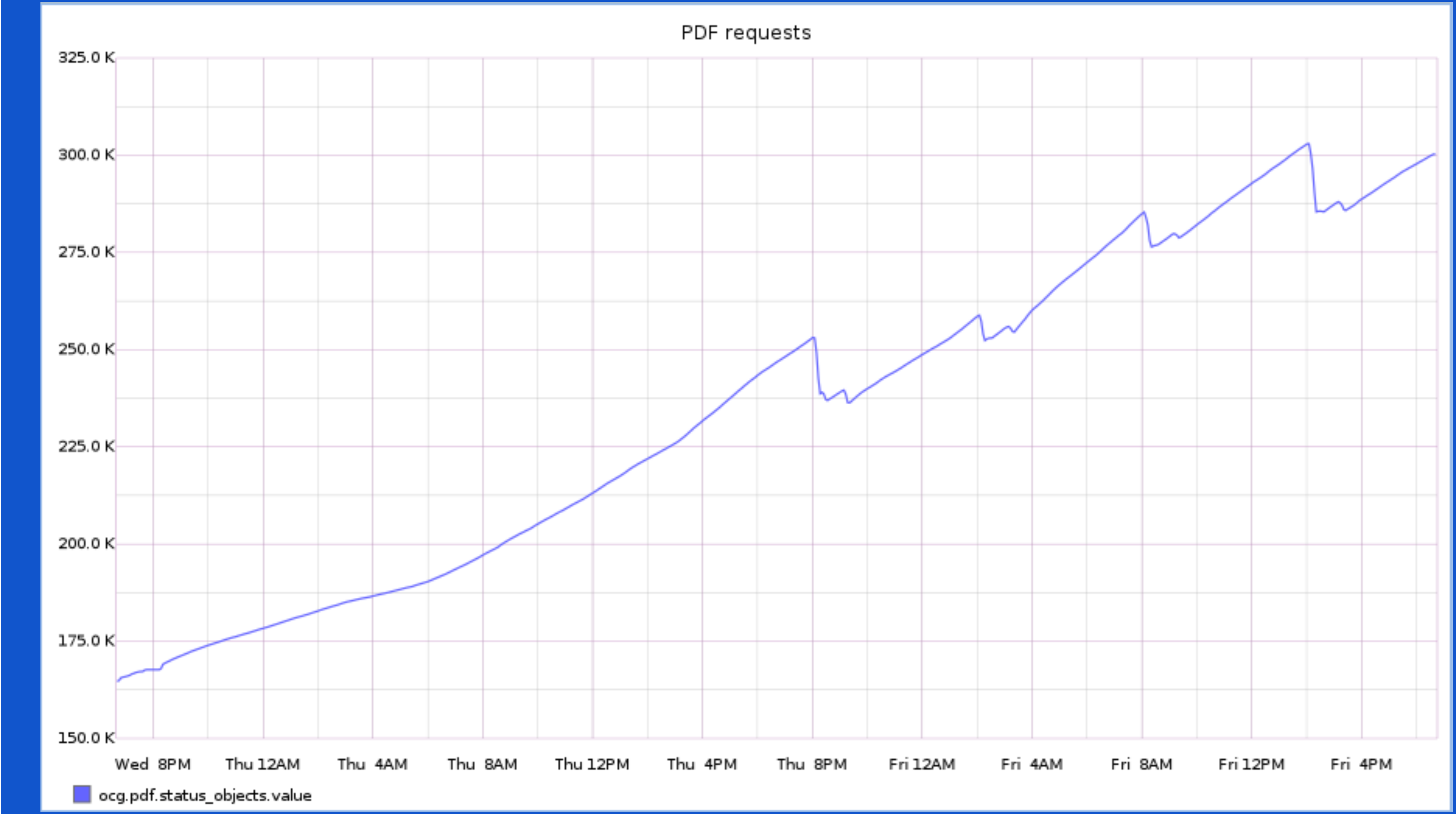


350k requests since entering production; ~110k/day



Cache hit rate is only ~25%

(request rate increase is when Indic wikis were switched over)



First gc Thu 8pm UTC! (but not yet steady state)

Thank you!

https://www.mediawiki.org/wiki/OCG