gMonitor: The itsy bitsy spider

By: Gerhard van Andel

Web crawler

- Rest Endpoint,
  - POST -- `"/api/url"`
    - body:

      ```
      { "url": "https://asynchronousgillz.github.io/" }
      ```

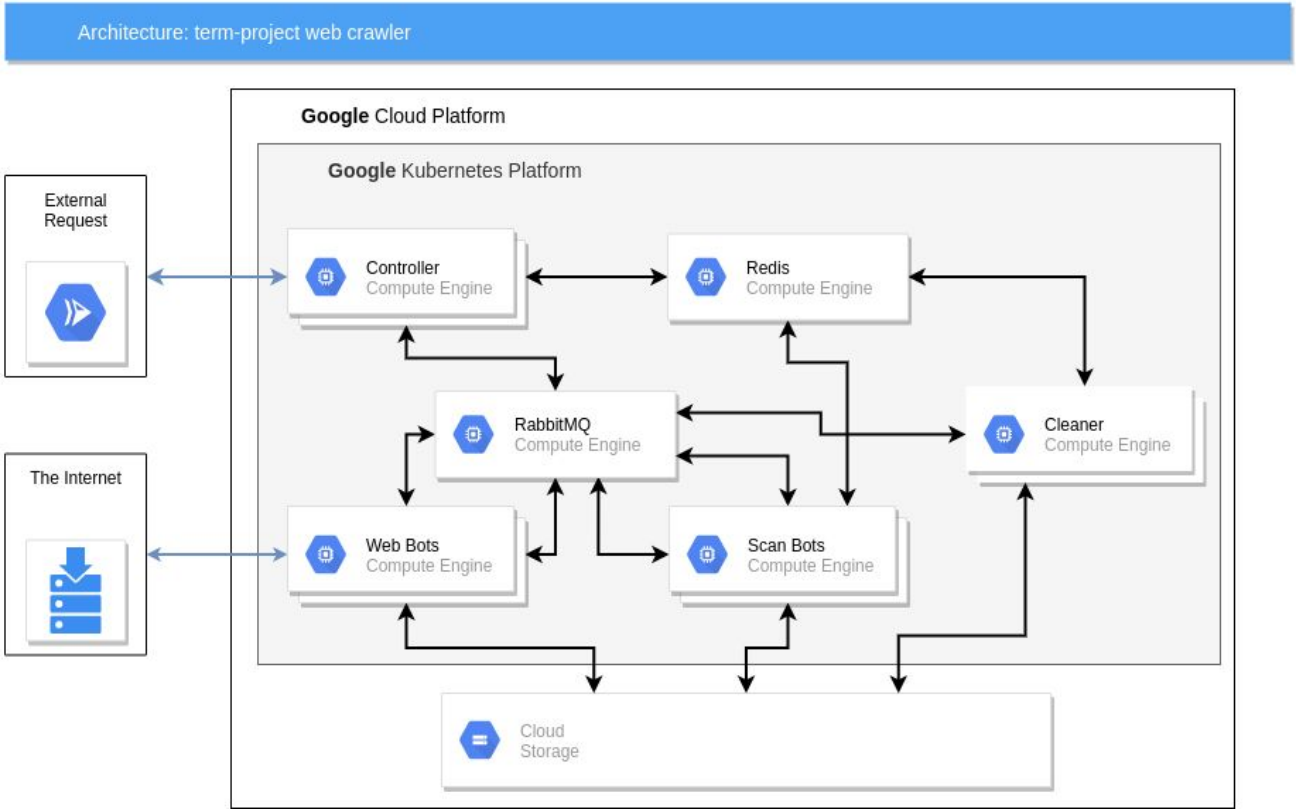
    - response:

      ```
      {"uuid": "1234-1234-1234-1234"}
      ```

Components

- Google Kubernetes Engine
- Google Cloud Storage
- RabbitMQ
- Redis
- Docker / python3 services
  - python packages:
    - Flask, Pika, Redis, Requests, BeautifulSoup, Google Cloud Libraries

Architectural Diagram



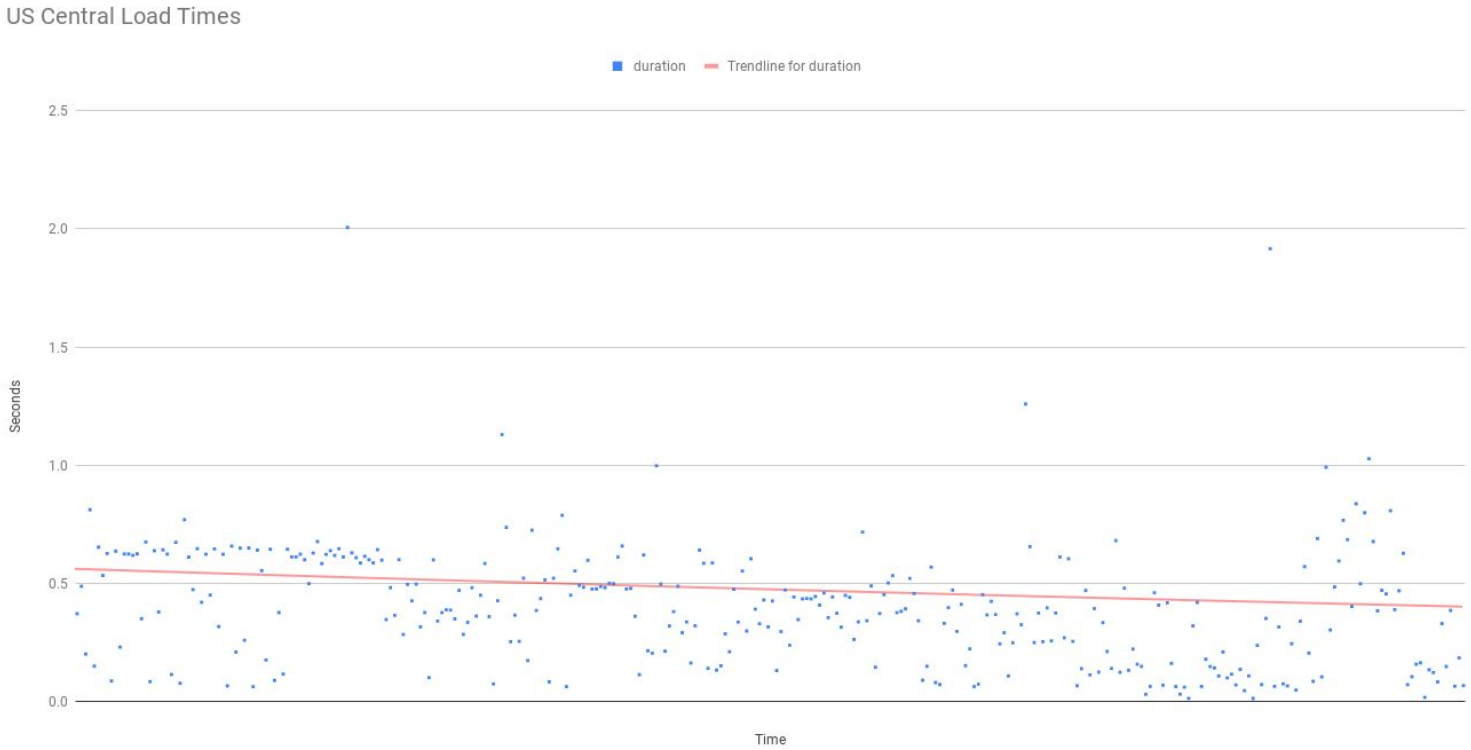
Debugging

- **Logging with containers**
  - Using docker-compose locally to test massaging / cache interactions
- **Cluster communication**
  - RabbitMQ messaging
- **Rubber Duck on my desk**
  - Doesn't give up any secrets of the universe but can helped me work through the problems I encountered.

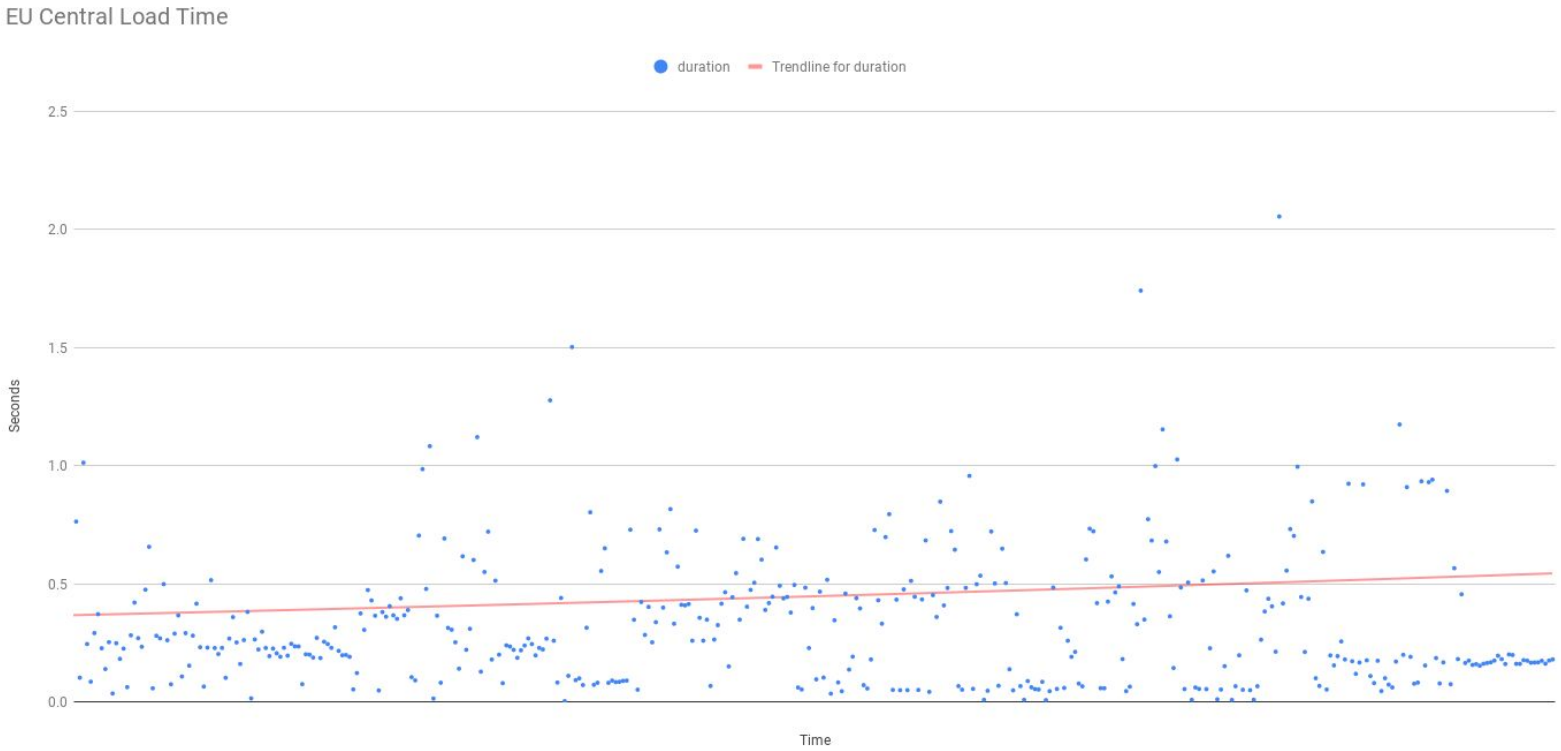
Learnings

- **Resource limiting to not overload web servers**
  - distributed lock manager
    - one worker at a time per domain
- **Following the rules of robots.txt**
  - crawl delay
    - time between requests
  - sitemap
    - use a sitemap to provide information about specific types of content on your pages
- https://support.google.com/webmasters/answer/7451184

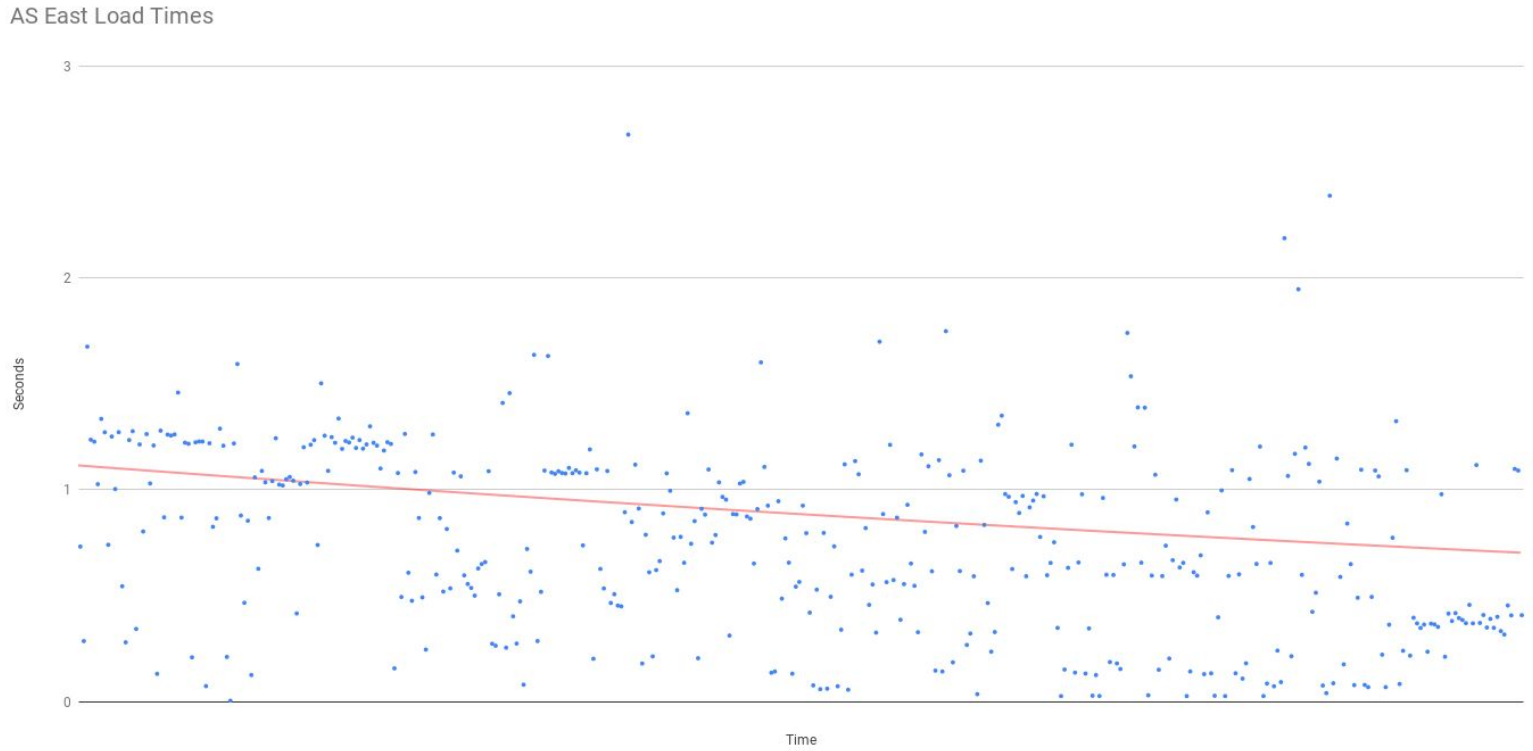
Results United-States



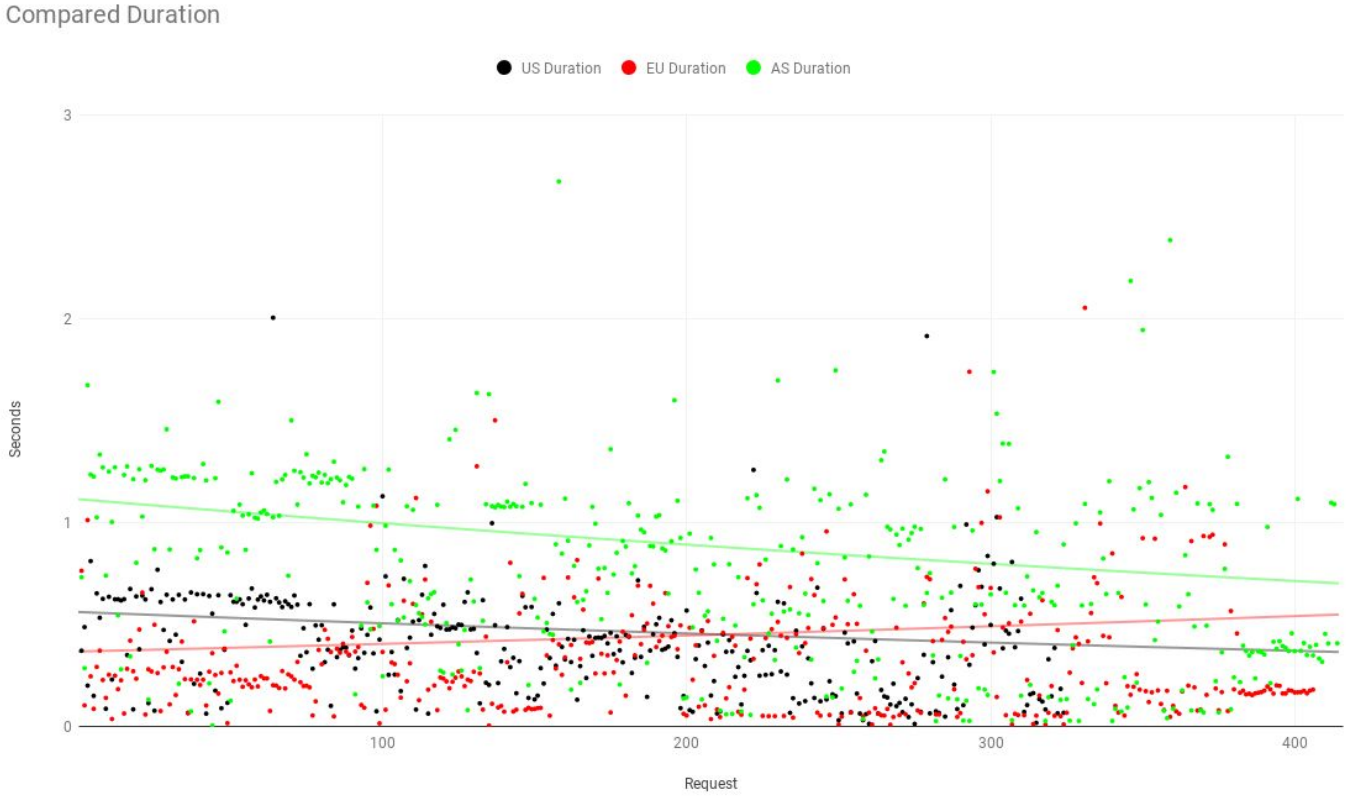
Results Europe



Results East Asia



Results Compared



Thank You!

contact information
email: geva2368@colorado.edu