Wikipedia's CDN

A Day In The Life

Emanuele Rocca

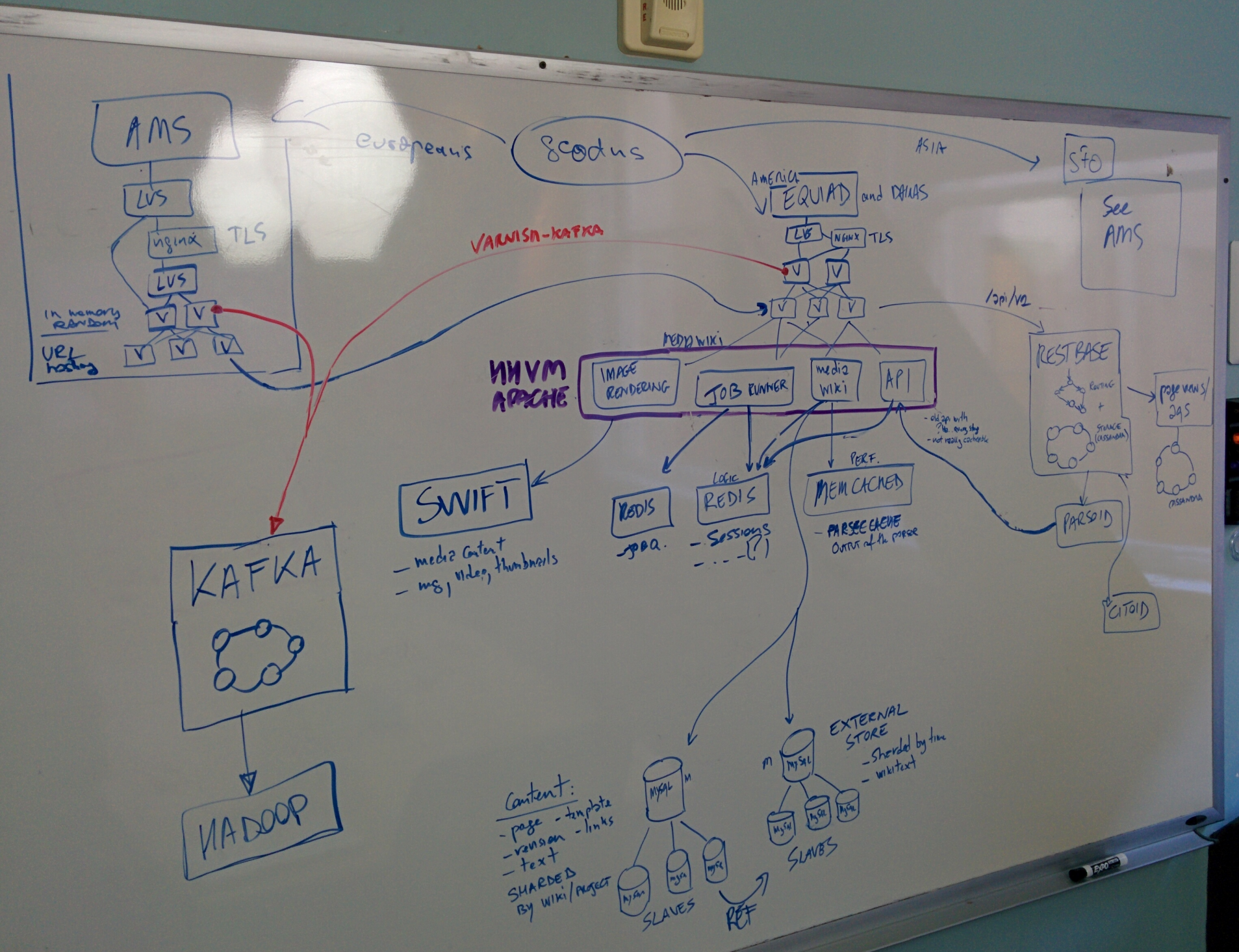
Site Reliability Engineer - Traffic

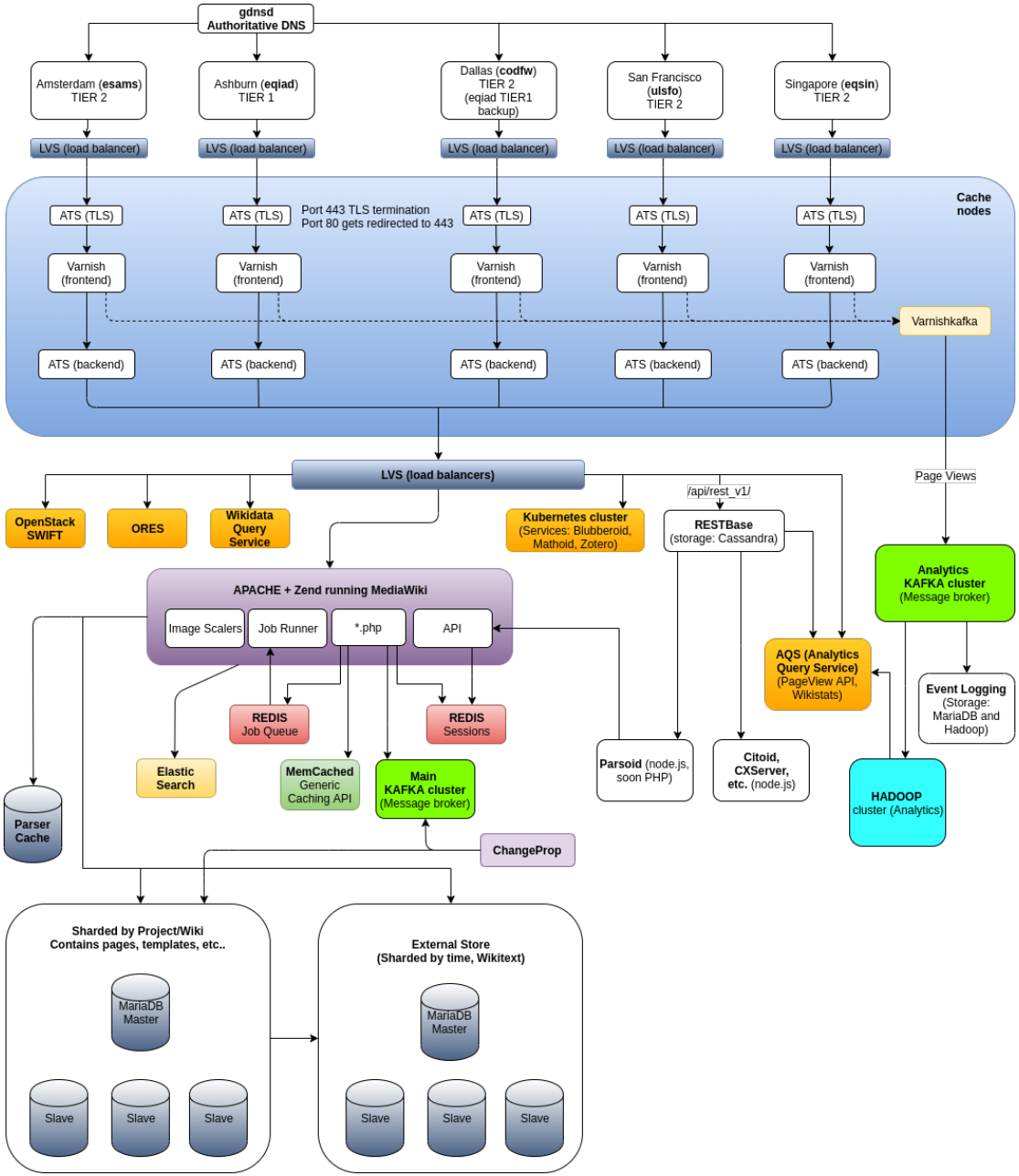
October 31st 2019

WIKIMEDIA FOUNDATION

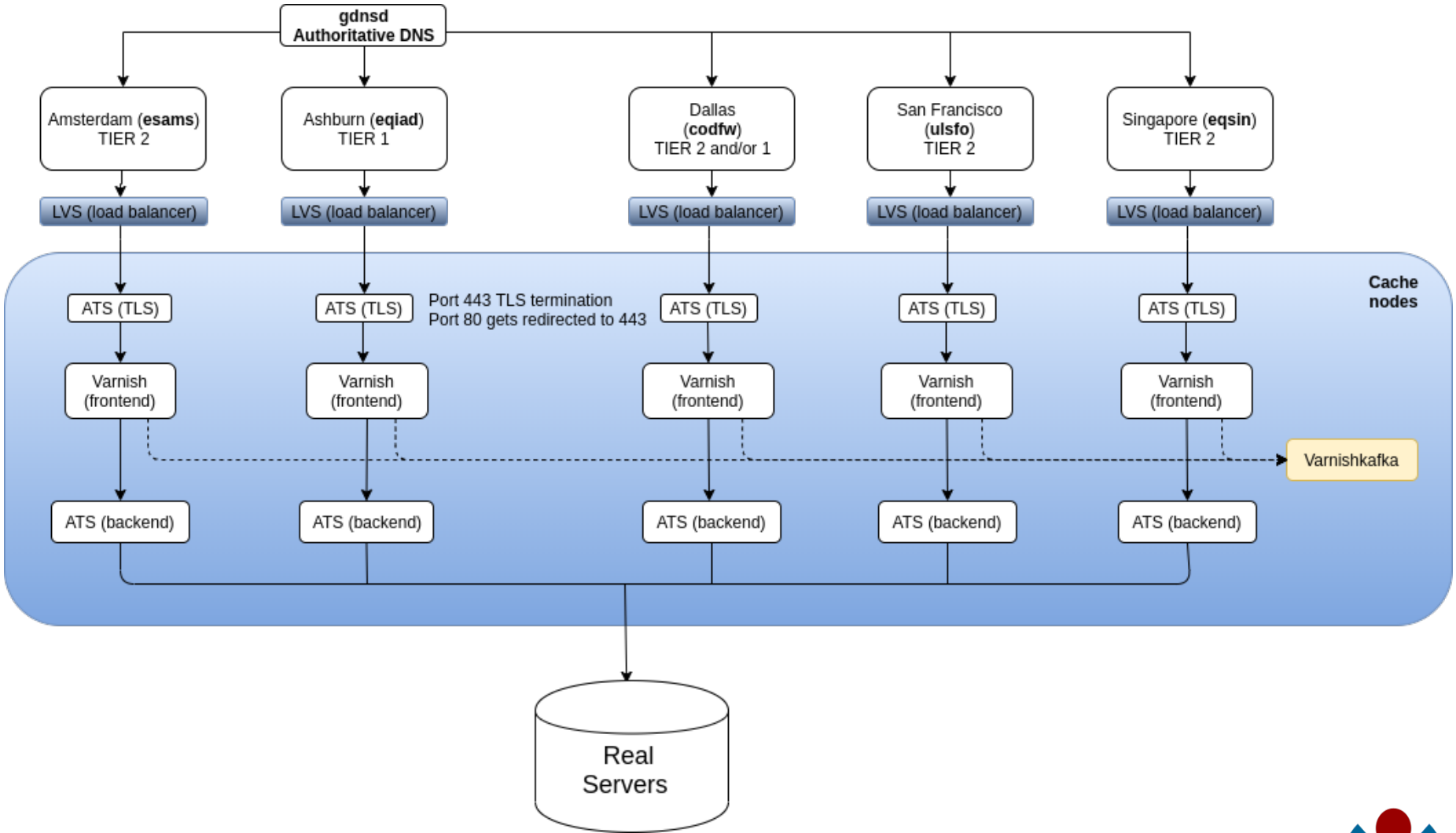
How does wikipedia end up on my screen? (partial answer)

WIKIMEDIA FOUNDATION


AMS
LVS
nginx
TLS
LVS
in memory
varnish
url hashing
eqiad
scodus
AMERICA
EQUIAD and DALLAS
ASIA
sfo
See AMS
VARNISH-KAFKA
LVS
NGINX
TLS
/api/v1
HHVM
APACHE
MEDIA WIKI
IMAGE RENDERING
JOB RUNNER
media wiki
API
RESTBASE
Routing
+
Storage (cassandra)
page views/ 295
cassandra
PARSOID
CITOID
- old api with the wrong shit
- not really cacheable
SWIFT
- media Content
- svg, video, thumbnails
REDIS
.ppa.
LOGIC
REDIS
- Sessions
- ... - (?)
PERF.
MEMCACHED
- PARSER CACHE
Output of the parser
KAFKA
HADOOP
Content:
- page - template
- version - links
- text
SHARDED
By wiki/project
MySQL
SLAVES
REF
EXTERNAL
STORE
- Sharded by time
- wikitext
MySQL
SLAVES



WIKIMEDIA
FOUNDATION



WIKIMEDIA
FOUNDATION

Outline

- What? Why?
- Geographic DNS Routing
- L4 Load Balancing and TLS termination
- HTTP Caching and L7 Load Balancing

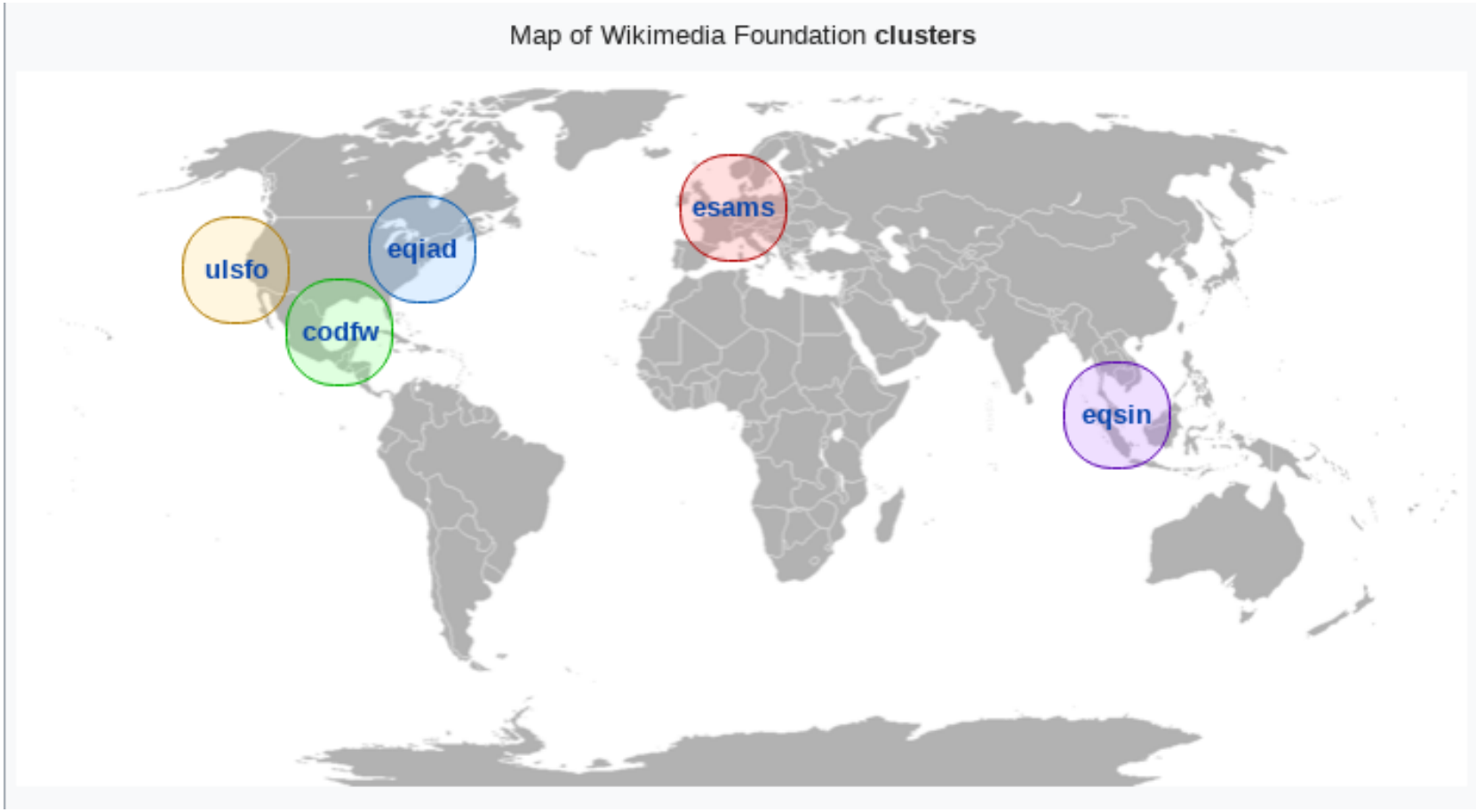
What is a CDN?

- Content Delivery Network
- Multiple servers distributed across Data Centers in various regions
- Reduce load on "real servers" by caching HTTP responses
- Reduce latency perceived by users by placing content geographically close to them (plus a few weird things)

Why our own CDN?

- Autonomy
- Privacy
- Risk of censorship

Cluster Map

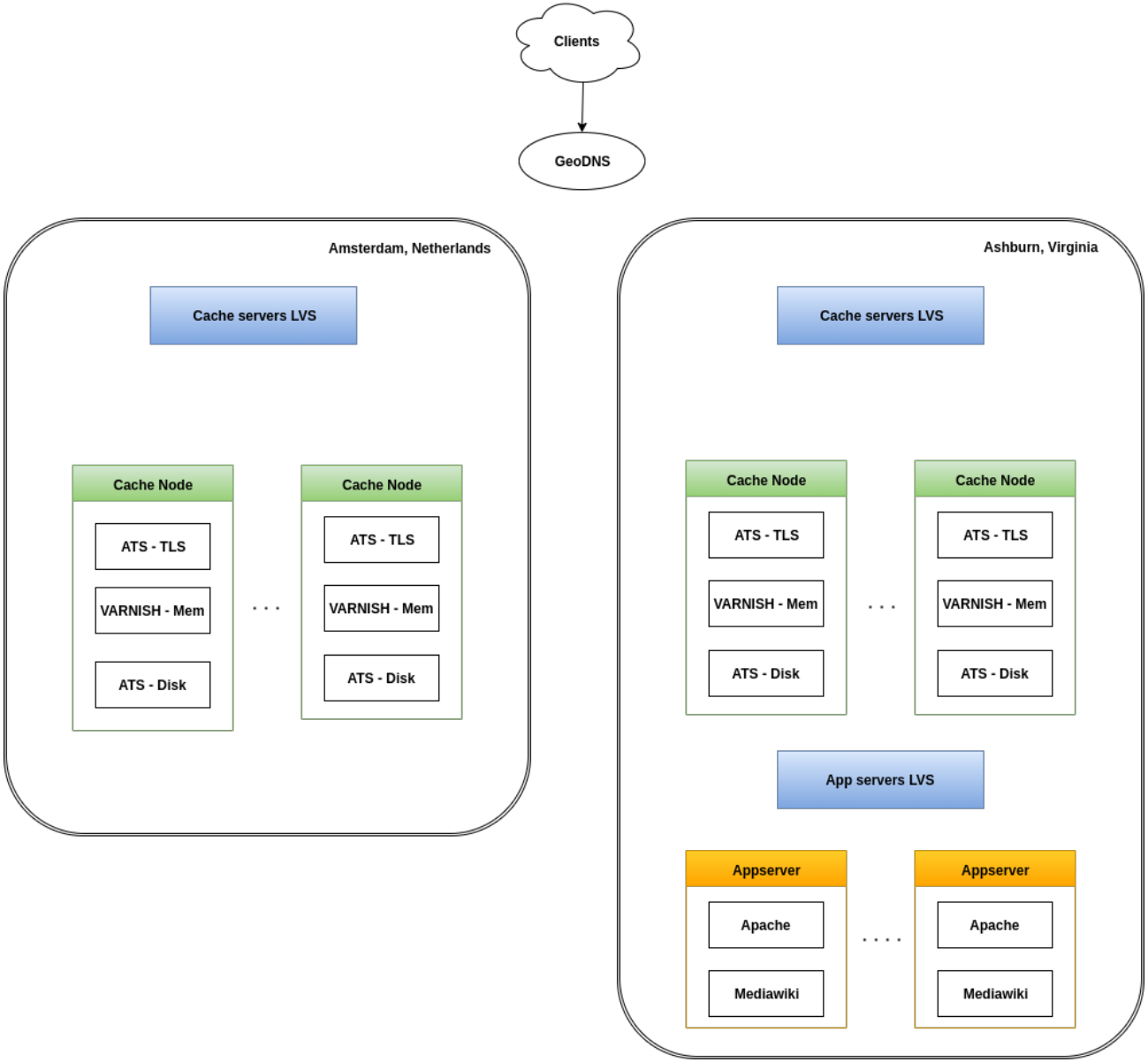


eqiad: Ashburn, Virginia - cp10xx - Tier 1
codfw: Dallas, Texas - cp20xx - Tier 1
esams: Amsterdam, Netherlands - cp30xx - Tier 2
ulsfo: San Francisco, California - cp40xx - Tier 2
eqsin: Singapore - cp50xx - Tier 2

WIKIMEDIA FOUNDATION

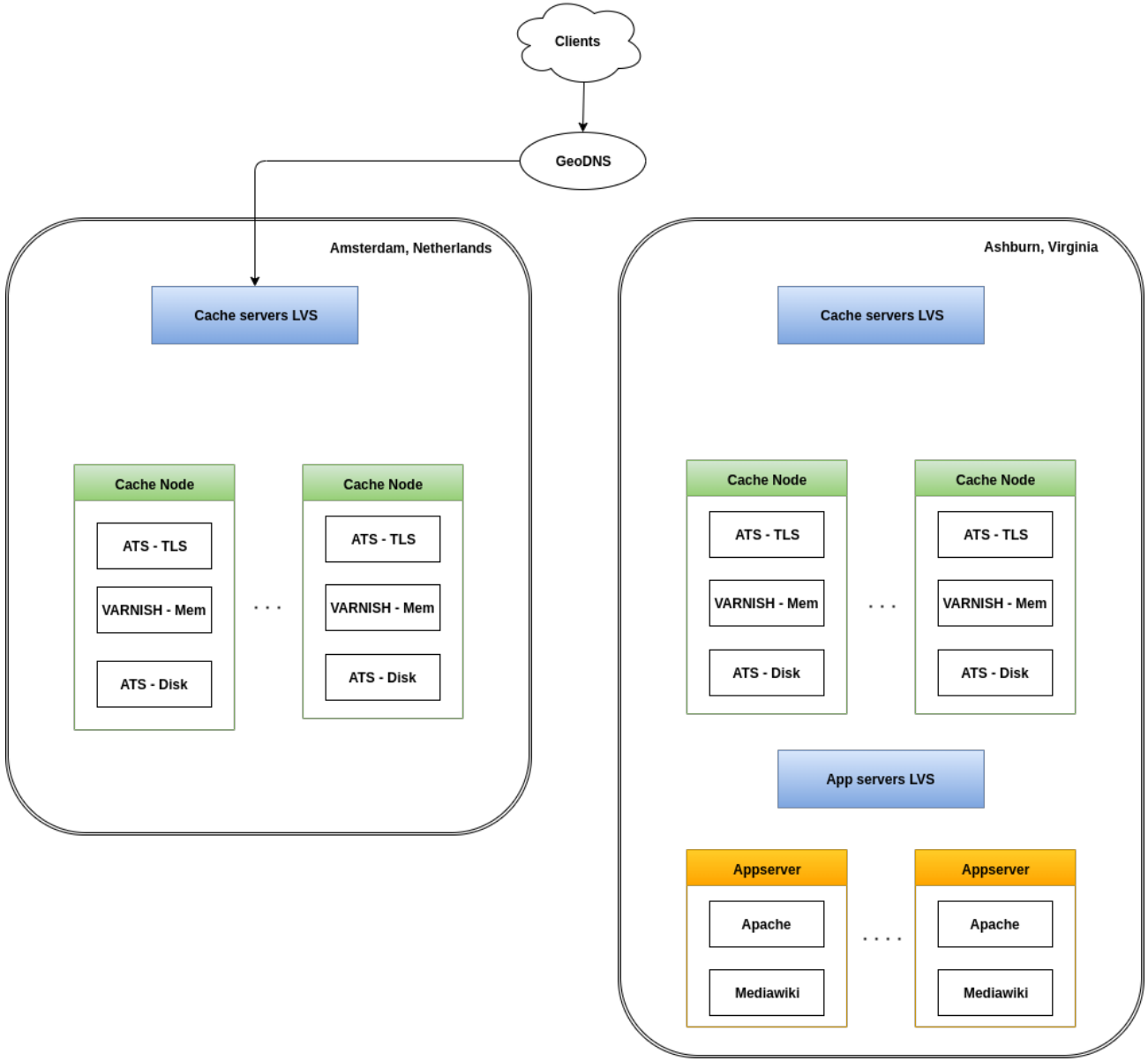
Traffic Volume

- Average: ~100k rps, peaks: ~150k rps
- esams 75k
- eqiad 35k
- eqsin 30k
- ulsfo 10k
- codfw 8k



WIKIMEDIA FOUNDATION

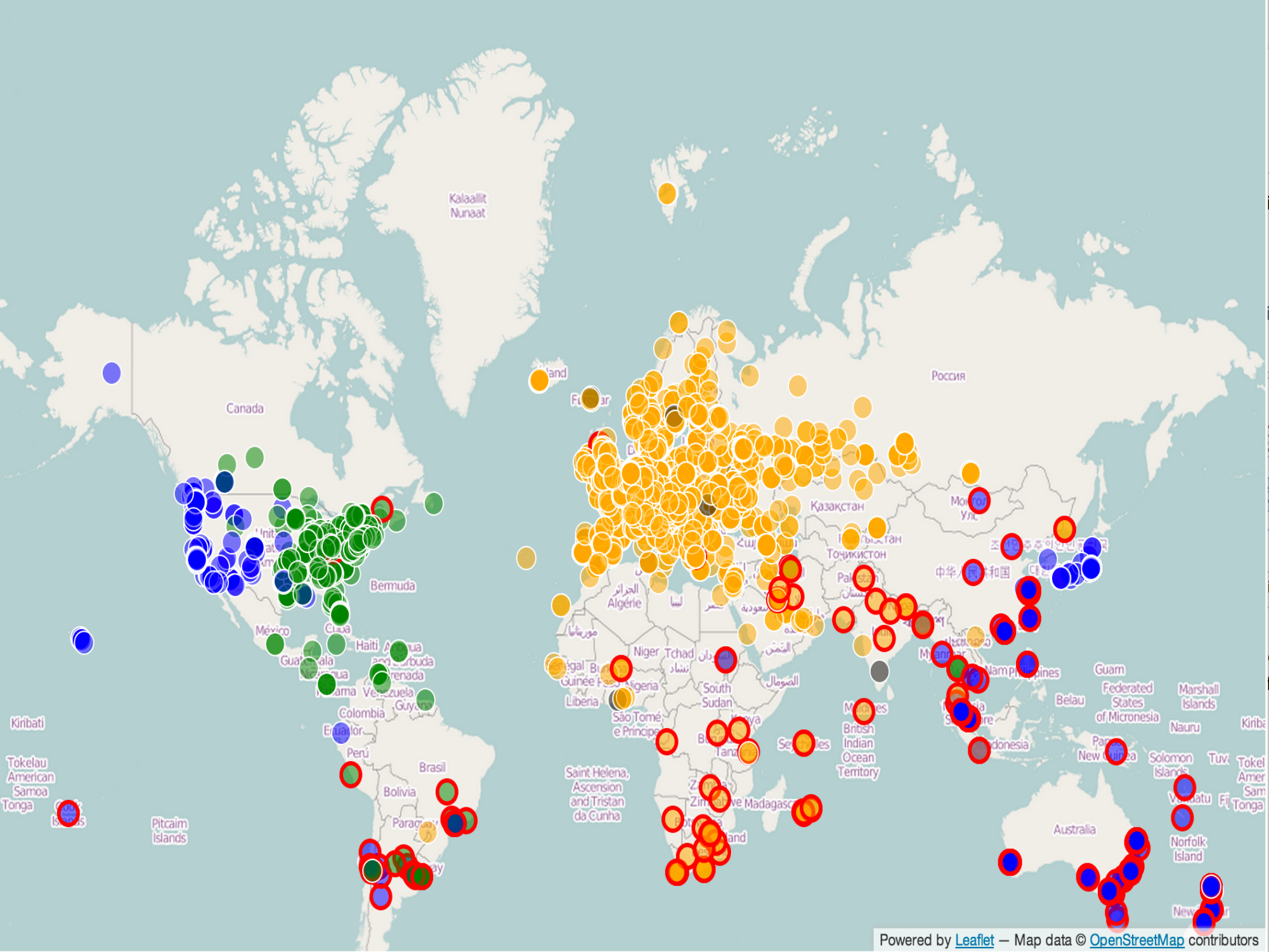
Geographic DNS Routing



WIKIMEDIA FOUNDATION

GeoDNS

- 3 authoritative DNS servers running gdnsd + geoip plugin
- GeoIP resolution, users get routed to the "best" DC
- DCs can be disabled through DNS configuration updates
- edns-client-subnet to make decisions based on the client actual IP
- RIPE Atlas probes used to define static mapping of countries to DCs


Kalaallit Nunaat
Canada
Россия
Қазақстан
Точикистон
Bermuda
México
Haiti
Colombia
Perú
Brasil
Bolivia
Pitcairn Islands
Algérie
Niger
Tchad
South Sudan
Liberia
São Tomé e Príncipe
Saint Helena, Ascension and Tristan da Cunha
British Indian Ocean Territory
Madagascar
Australia
Guam
Federated States of Micronesia
Marshall Islands
Belau
Nauru
Solomon Islands
Norfolk Island
Kiribati
Tokelau
American Samoa
Tonga
Powered by Leaflet — Map data © OpenStreetMap contributors

config-geo

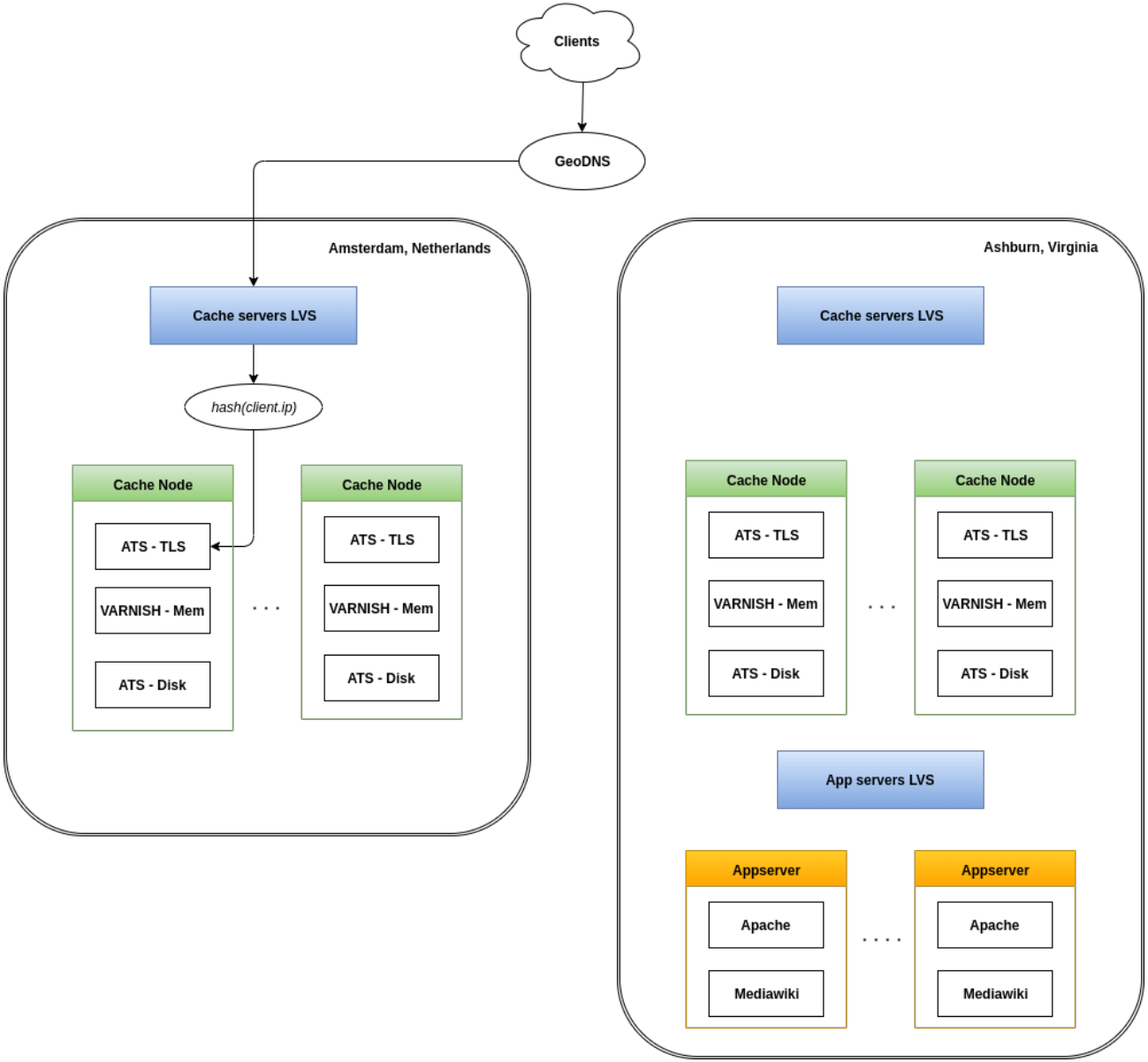
```
CA => [ulsfo, codfw, eqiad, esams, eqsin], # California
CO => [codfw, eqiad, ulsfo, esams, eqsin], # Colorado
[...]
FR => [esams, eqiad, codfw, ulsfo, eqsin], # France
JP => [eqsin, ulsfo, codfw, eqiad, esams], # Japan
```

---

https://github.com/wikimedia/operations-dns/

L4 Load Balancing and TLS termination

L4 Load Balancing



WIKIMEDIA FOUNDATION

Load balancers and cache servers

- Load balancers running Linux Virtual Server (primary/secondary)
- Configuration of servers pools via PyBal
- HTTP cache proxies running ATS and Varnish
  - ATS for TLS termination
  - Varnish In-memory: faster, smaller (~200G)
  - ATS On-disk: slower, larger (~1.5T)

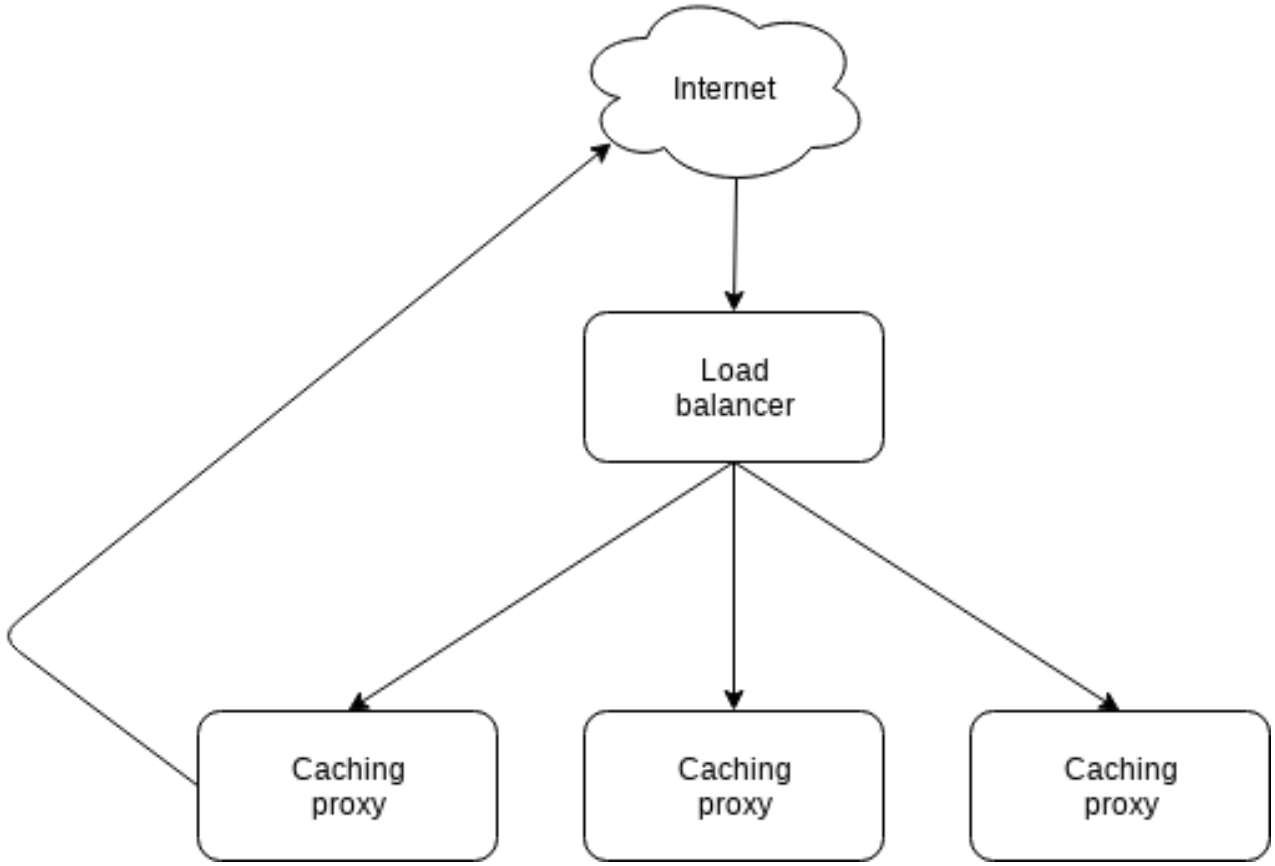
Pybal

- Checks servers health to determine which ones can be used
- Speaks BGP with the routers to announce service IPs and failover to secondary LVS
- Changes IPVS (LVS) configuration
- Gets server pools and their status (admin pooled/depooled) from etcd with HTTP Long Polling

ATS behind LVS

- ATS servers behind LVS for TLS termination
- Load-balancing hashing on client IP (TLS session reuse, TCP Fast Open)
- Direct Routing

Load balancing: direct routing

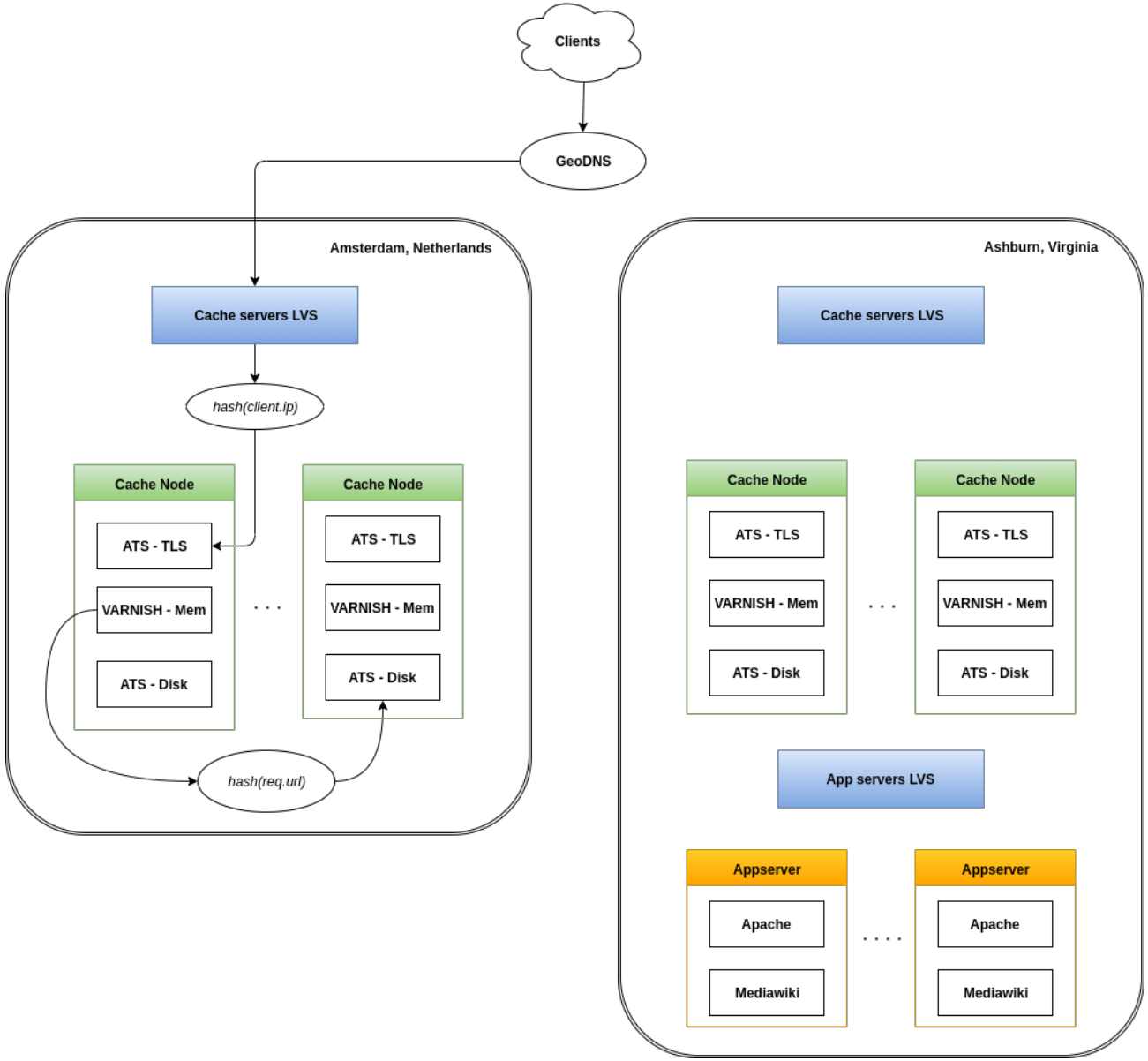


WIKIMEDIA FOUNDATION

TLS termination

- In the process of replacing nginx with ATS for TLS termination
- TLS/HTTP2 terminated as close as possible to users
- TLSv1.0+ with Perfect-Forward-Secret ciphersuites only
- On the roadmap: TLSv1.3, ESNI

HTTP Caching and L7 Load Balancing



WIKIMEDIA FOUNDATION

HTTP Caching

- Cache misses in-memory (frontend) are being served by on-disk caches (backend)
- L7 Load Balancing performed by Varnish, nodes and their pooled/depooled status defined in etcd
- Consistent hashing request URL to spread dataset among caches. Effective cache size ~sum(disk size)
- What's the effective cache size for cache frontends?

Cache miss:

```
$ curl -v https://upload.wikimedia.org/this-does-not-exist 2>&1 |
    grep x-cache:
< x-cache: cp3063 miss, cp3059 miss
```

Cache miss:

```
$ curl -v https://upload.wikimedia.org/this-does-not-exist 2>&1 |
    grep x-cache:
< x-cache: cp3063 miss, cp3059 miss
```

Cache hit:

```
$ curl -v https://upload.wikimedia.org/wikipedia/labs/4/4d/
    Infrastructure_overview.png 2>&1 | grep x-cache:
< x-cache: cp3051 hit, cp3059 hit/5
```

Cache miss:

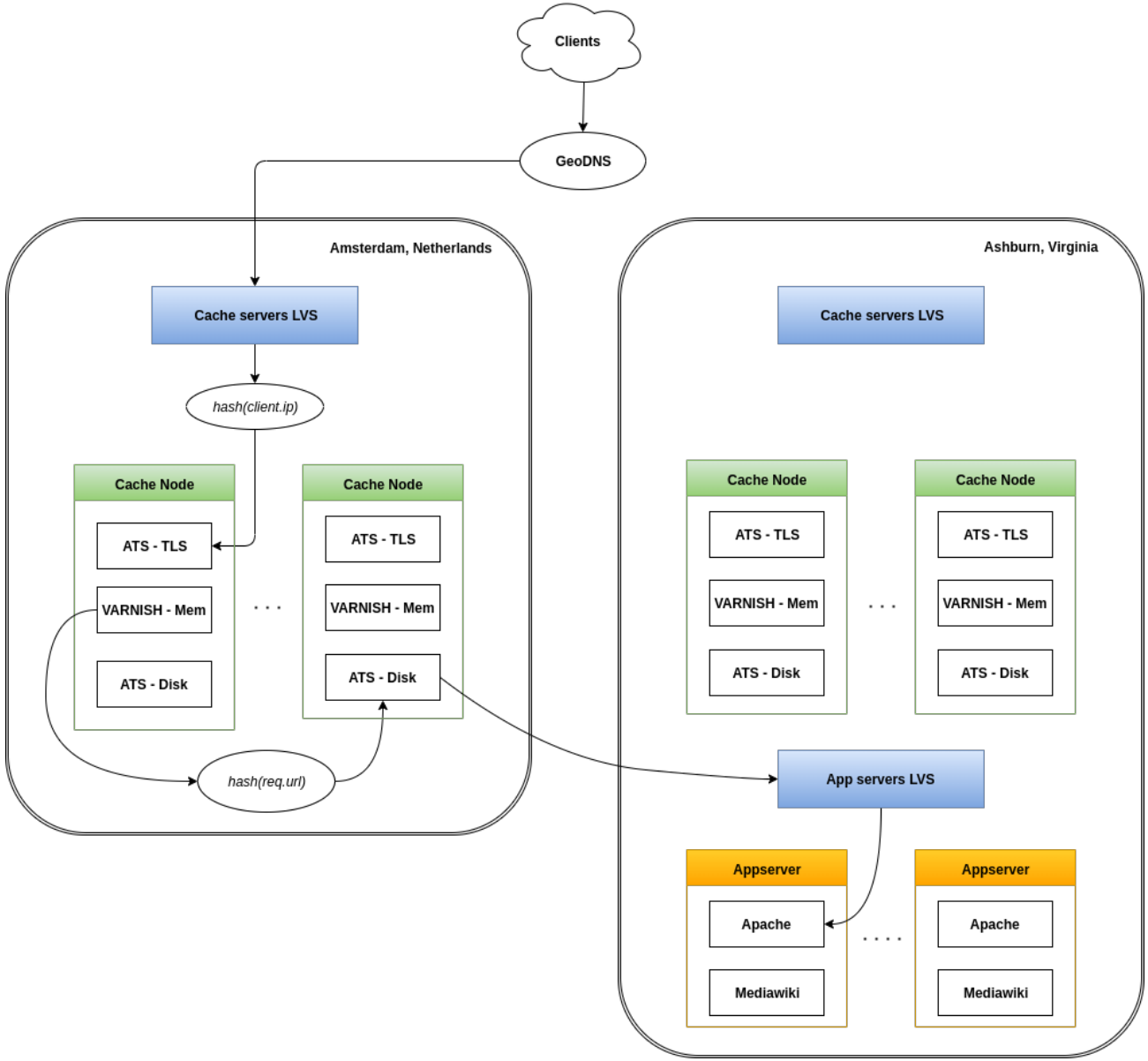
```
$ curl -v https://upload.wikimedia.org/this-does-not-exist 2>&1 |
    grep x-cache:
< x-cache: cp3063 miss, cp3059 miss
```

Cache hit:

```
$ curl -v https://upload.wikimedia.org/wikipedia/labs/4/4d/
    Infrastructure_overview.png 2>&1 | grep x-cache:
< x-cache: cp3051 hit, cp3059 hit/5
```

Forcing a specific DC:

```
$ curl -v https://upload.wikimedia.org/wikipedia/labs/4/4d/
    Infrastructure_overview.png \
    --resolve upload.wikimedia.org:443:103.102.166.240 2>&1 |
        grep x-cache:
< x-cache: cp5002 miss, cp5002 hit/1
```



WIKIMEDIA FOUNDATION

Conclusions

- WMF runs its own CDN
- Geographic DNS Routing with gdnsd
- L4 Load Balancing with LVS controlled by PyBal
- TLS terminations with ATS
- HTTP Caching and L7 Load Balancing with Varnish and ATS