SPARKPOST

The Day the DNS Died

Jeremy Blosser, Principal Operations Engineer
jblosser@sparkpost.com
@SparkPost

https://tinyurl.com/spdnstalk

Introduction

SparkPost, aka Message Systems, is a high-volume, transactional email software and services vendor.

Introduction

SparkPost, aka Message Systems, is a high-volume, transactional email software and services vendor.

We send a lot of email:

- Over 30% of the world's non-spam email is sent using our software.
- 15B messages/month sent via our cloud offering.

Introduction

SparkPost, aka Message Systems, is a high-volume, transactional email software and services vendor.

We send a lot of email:

- Over 30% of the world's non-spam email is sent using our software.
- 15B messages/month sent via our cloud offering.

That requires a lot of DNS:

- 8,000 queries/second.
- 20Mb/s+ sustained traffic just for DNS queries.
- Several different resolution paths.

Introduction

But DNS is easy, right?

Introduction

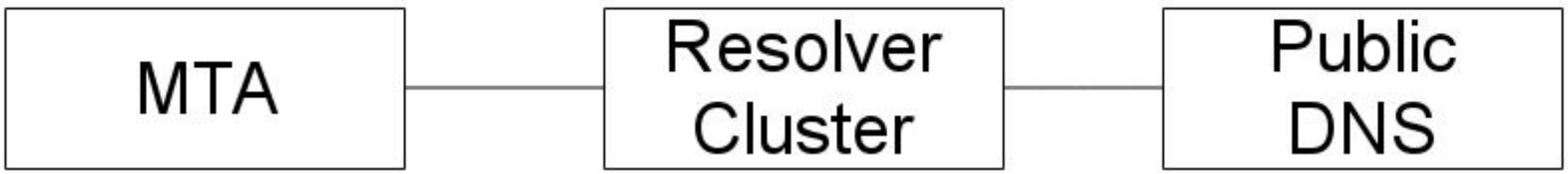
But DNS is easy, right?

Outline

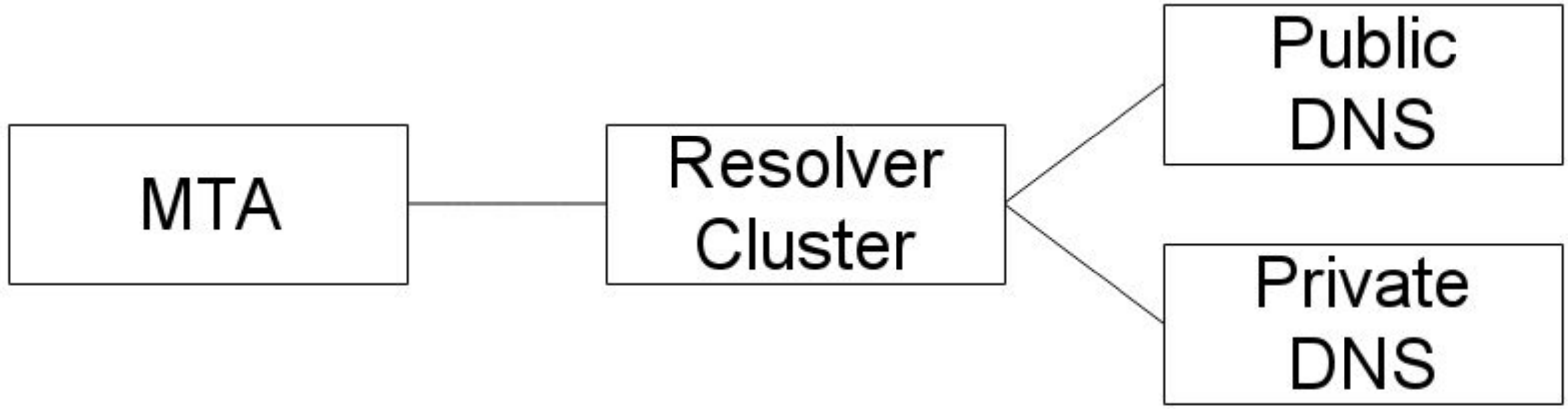
- Introduction
- Previous DNS Design(s)
- May 2017 Outage
- New DNS Design
- Lessons Learned / Remembered
- References
- Questions?

Previous DNS Design(s)

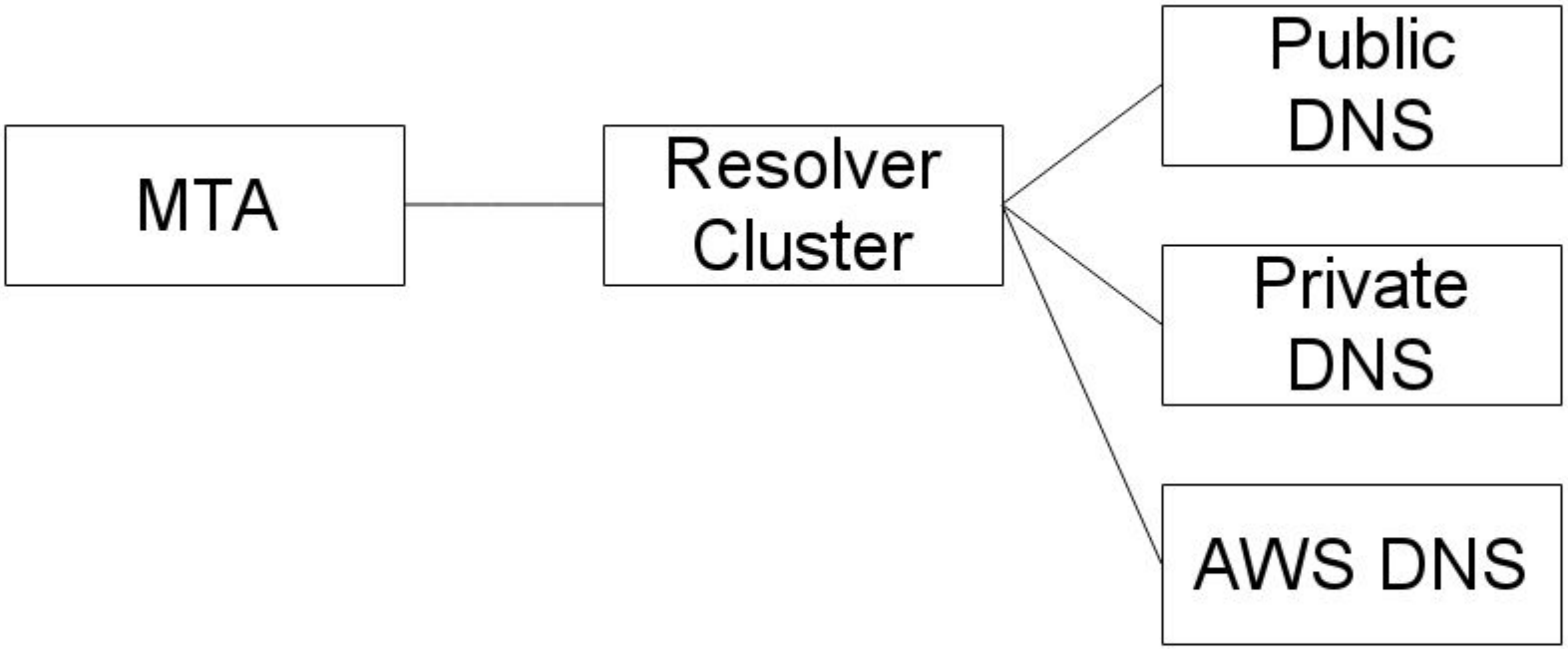
Version 1, Centralized Internal Resolver Cluster



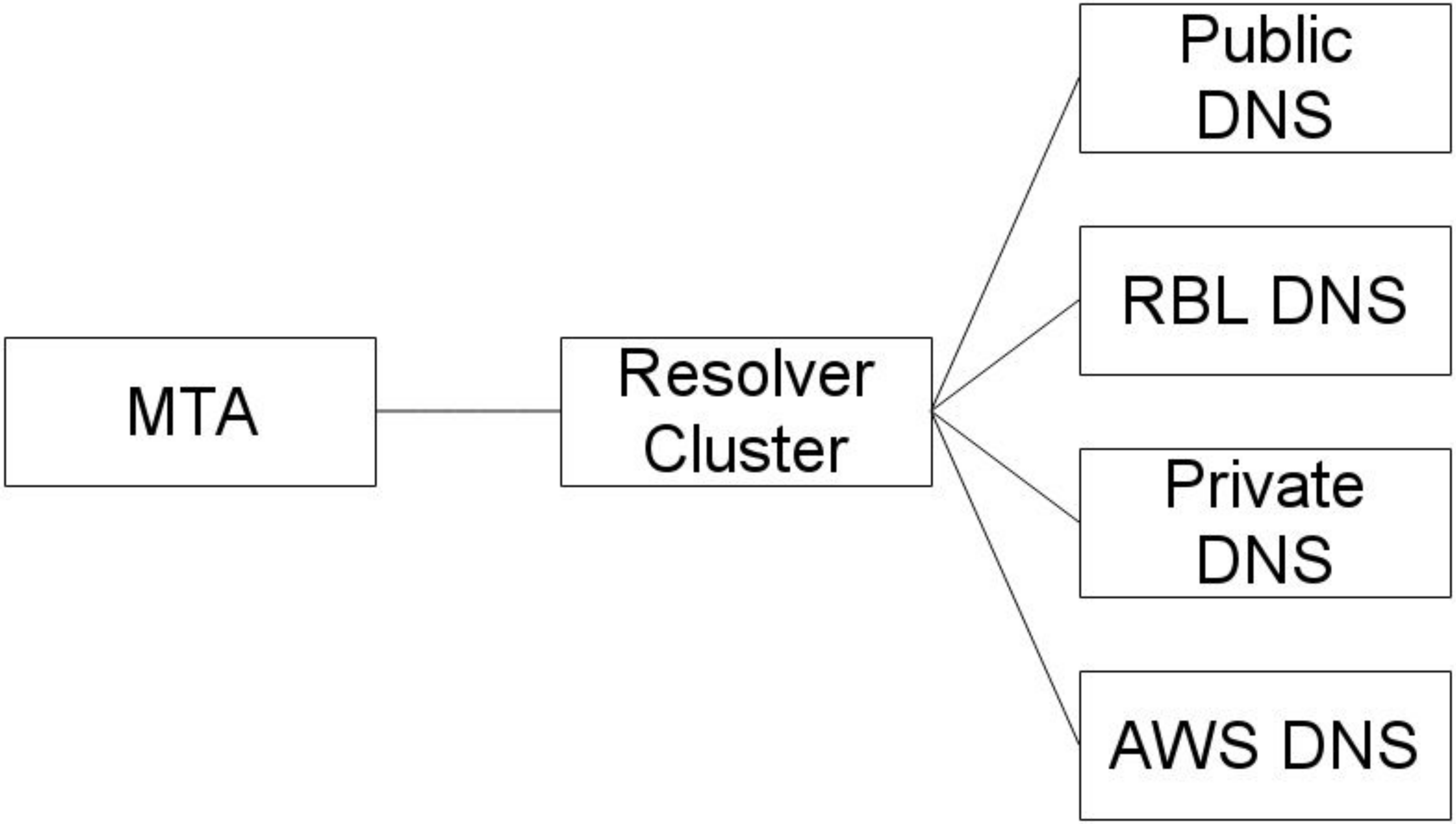
Version 1, Centralized Internal Resolver Cluster



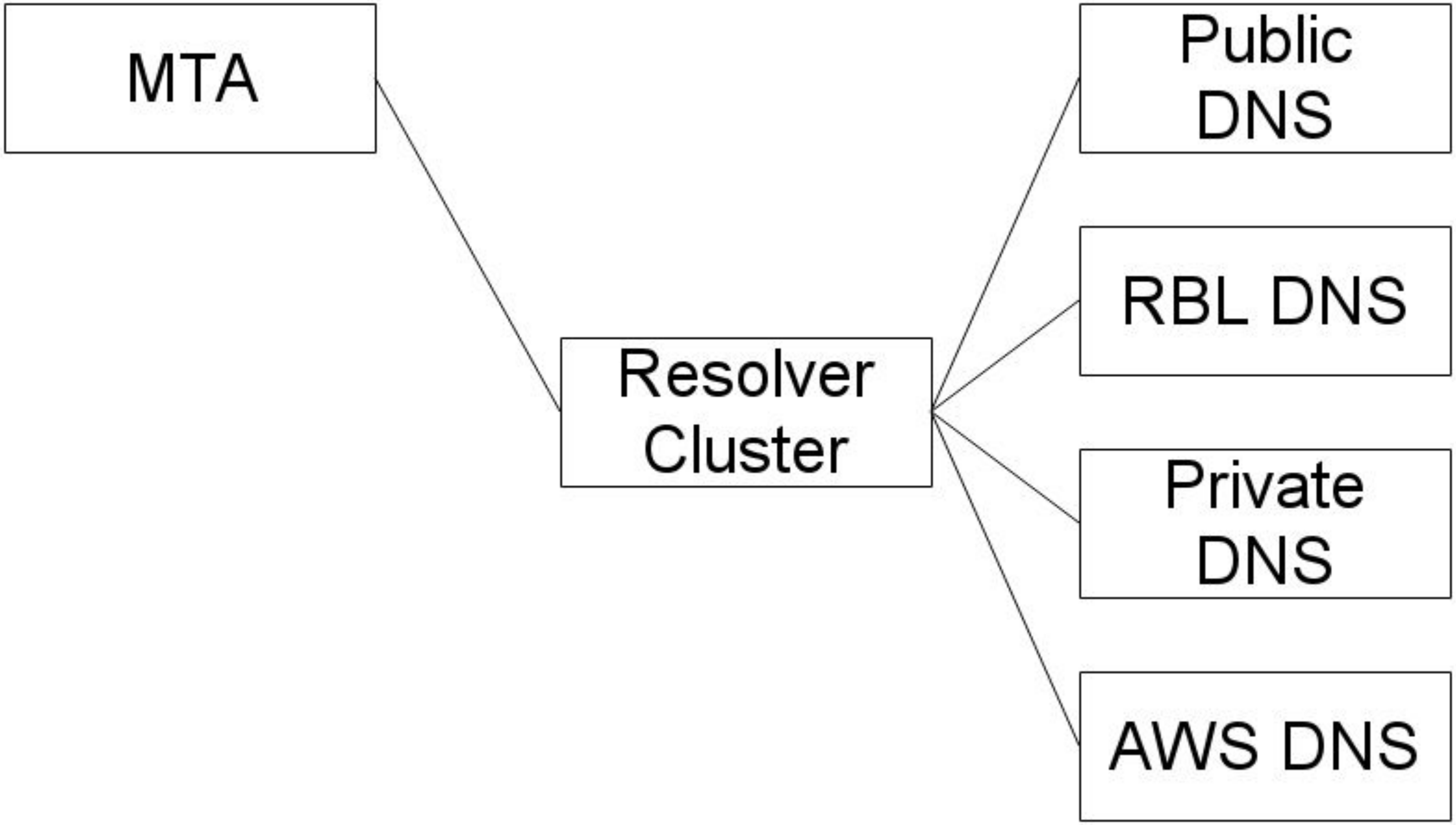
Version 1, Centralized Internal Resolver Cluster



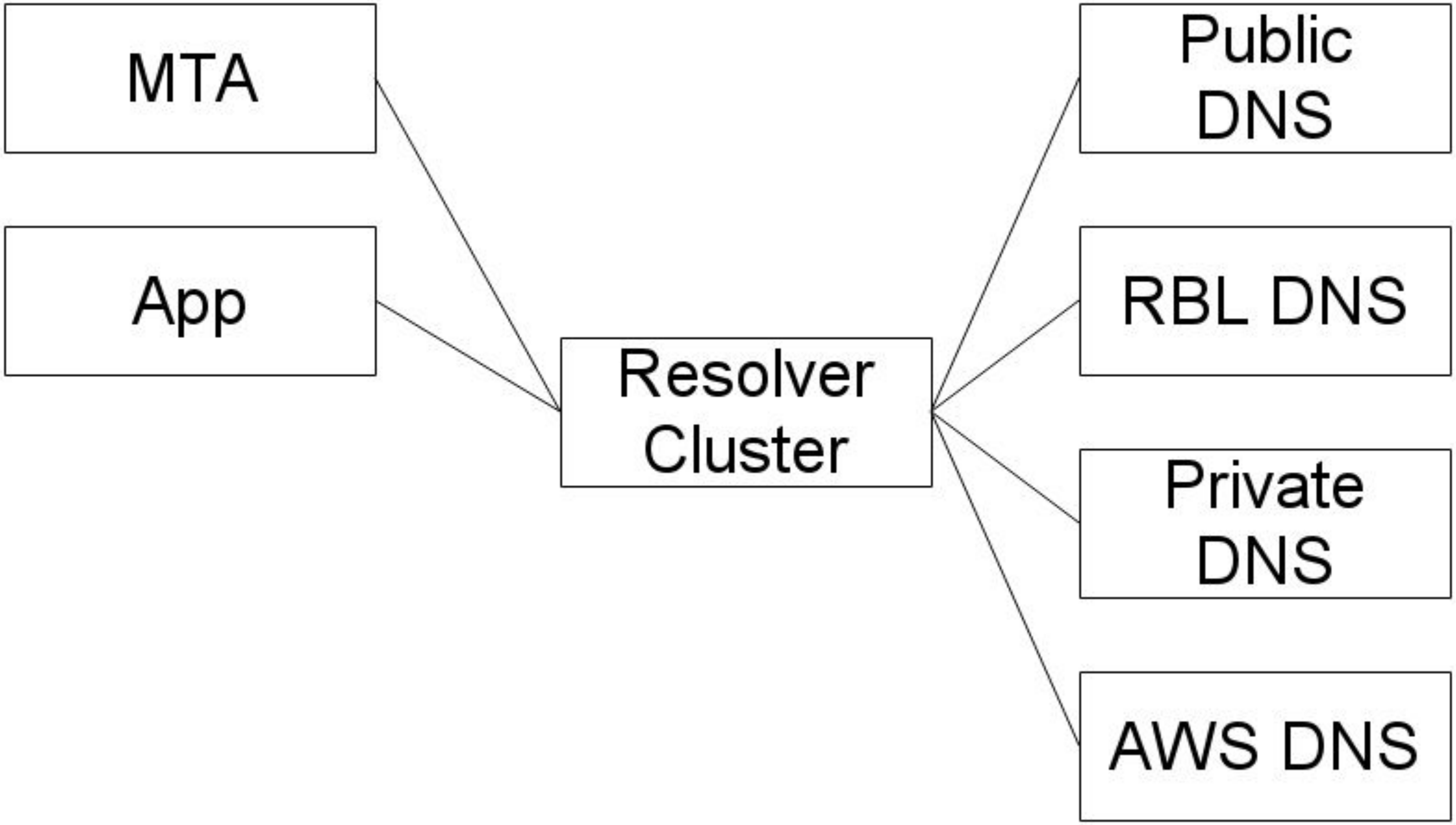
Version 1, Centralized Internal Resolver Cluster



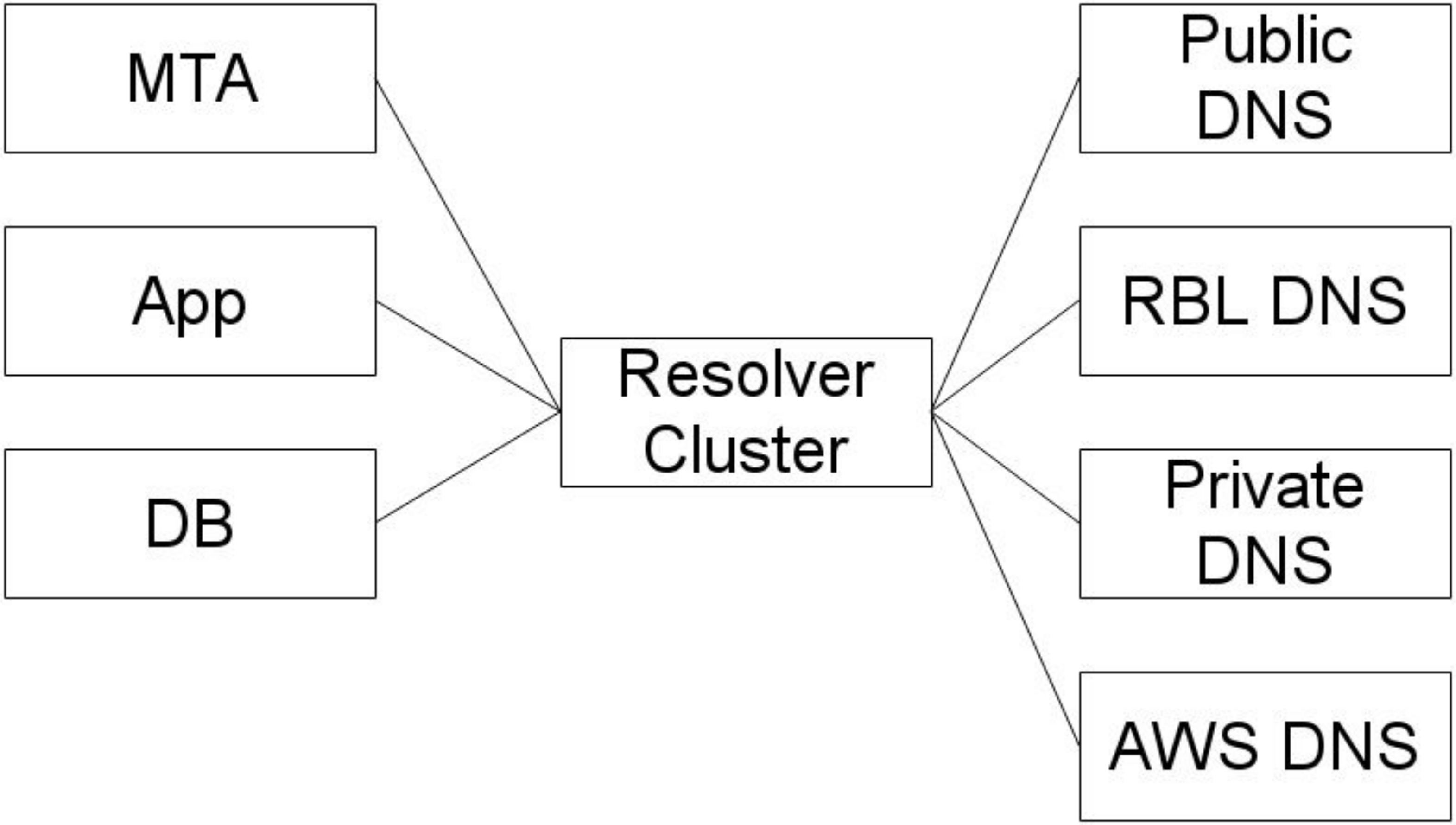
Version 1, Centralized Internal Resolver Cluster



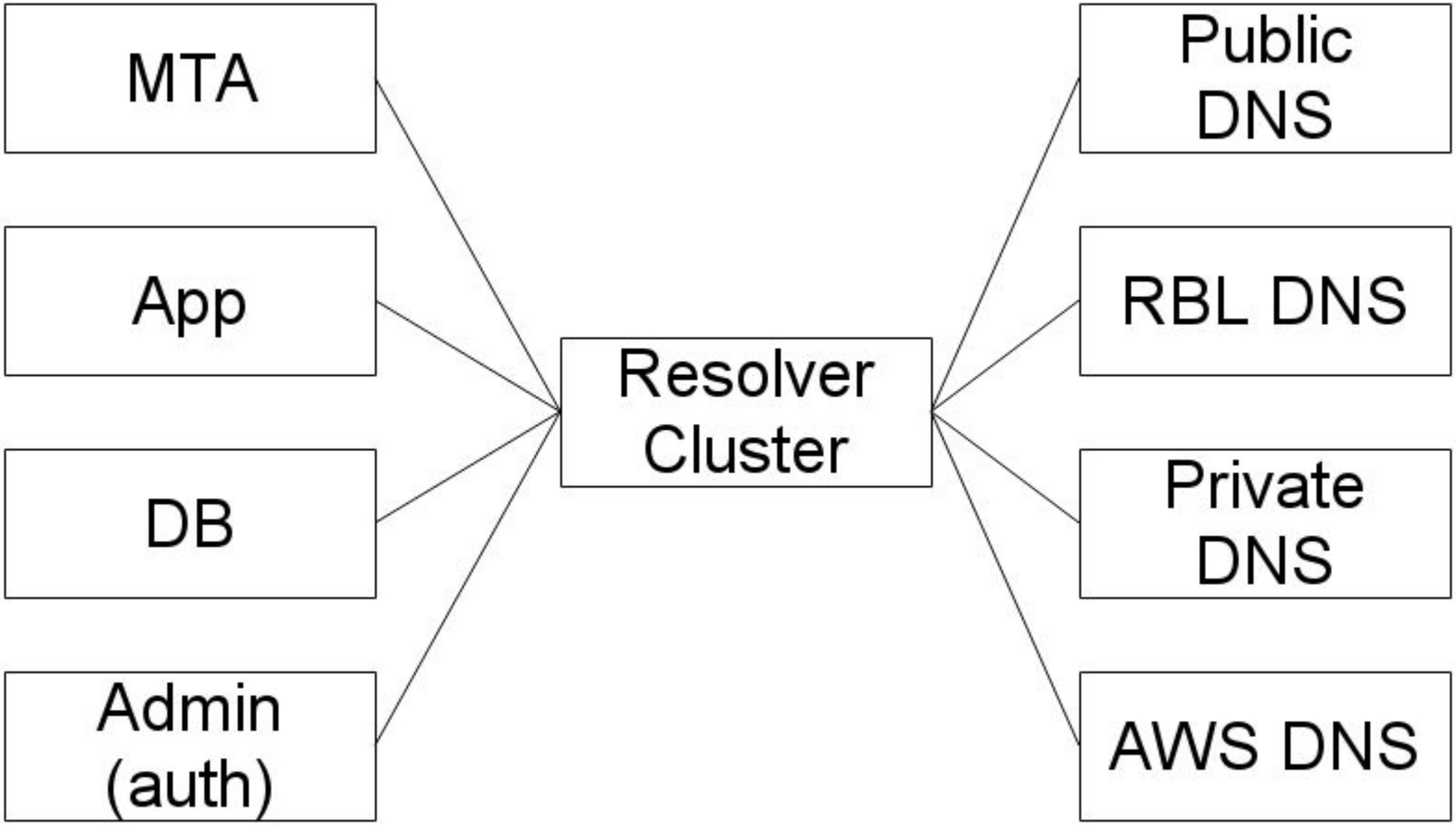
Version 1, Centralized Internal Resolver Cluster



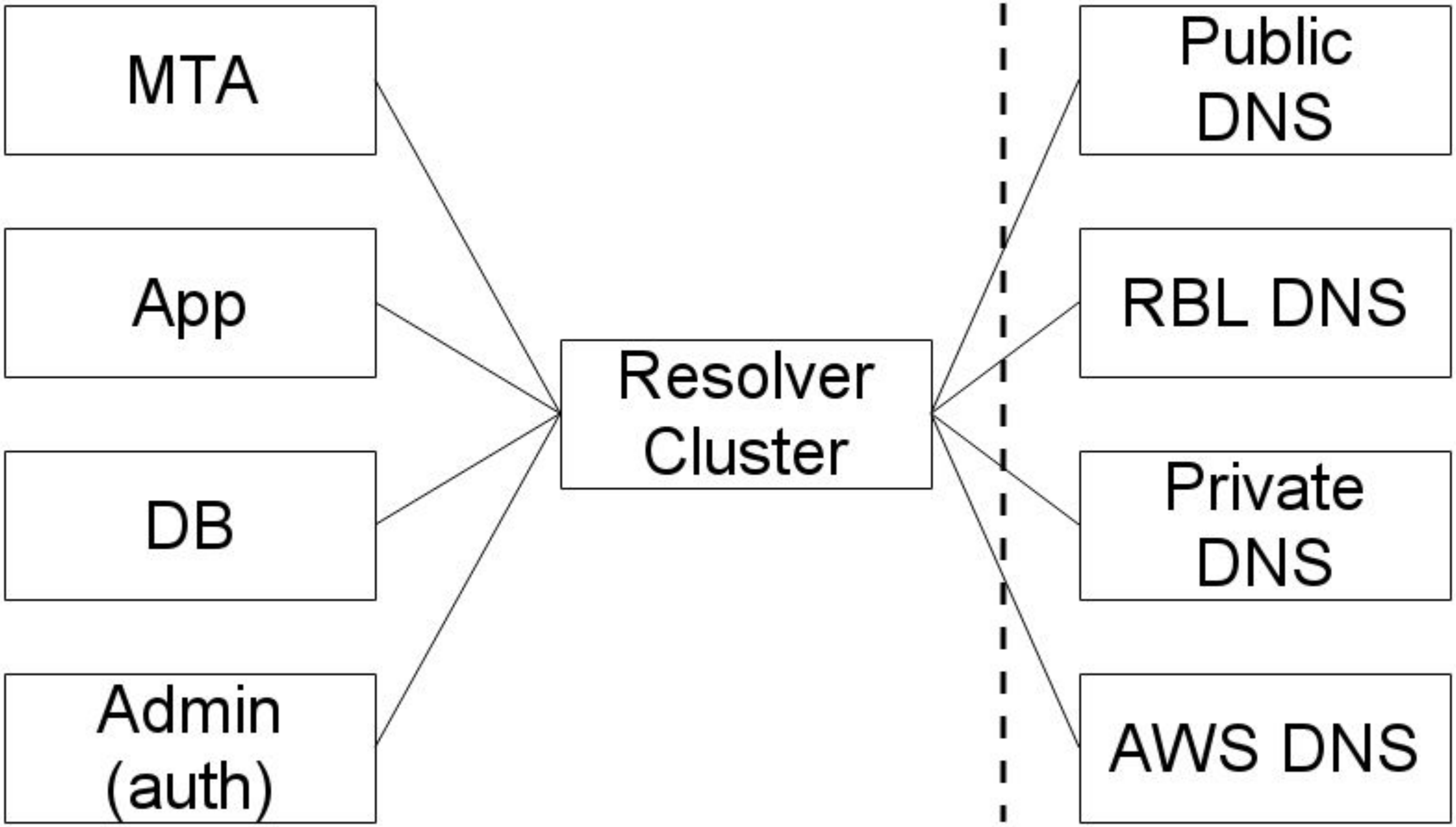
Version 1, Centralized Internal Resolver Cluster



Version 1, Centralized Internal Resolver Cluster



Version 1, Centralized Internal Resolver Cluster



MTA

App

DB

Admin (auth)

Resolver Cluster

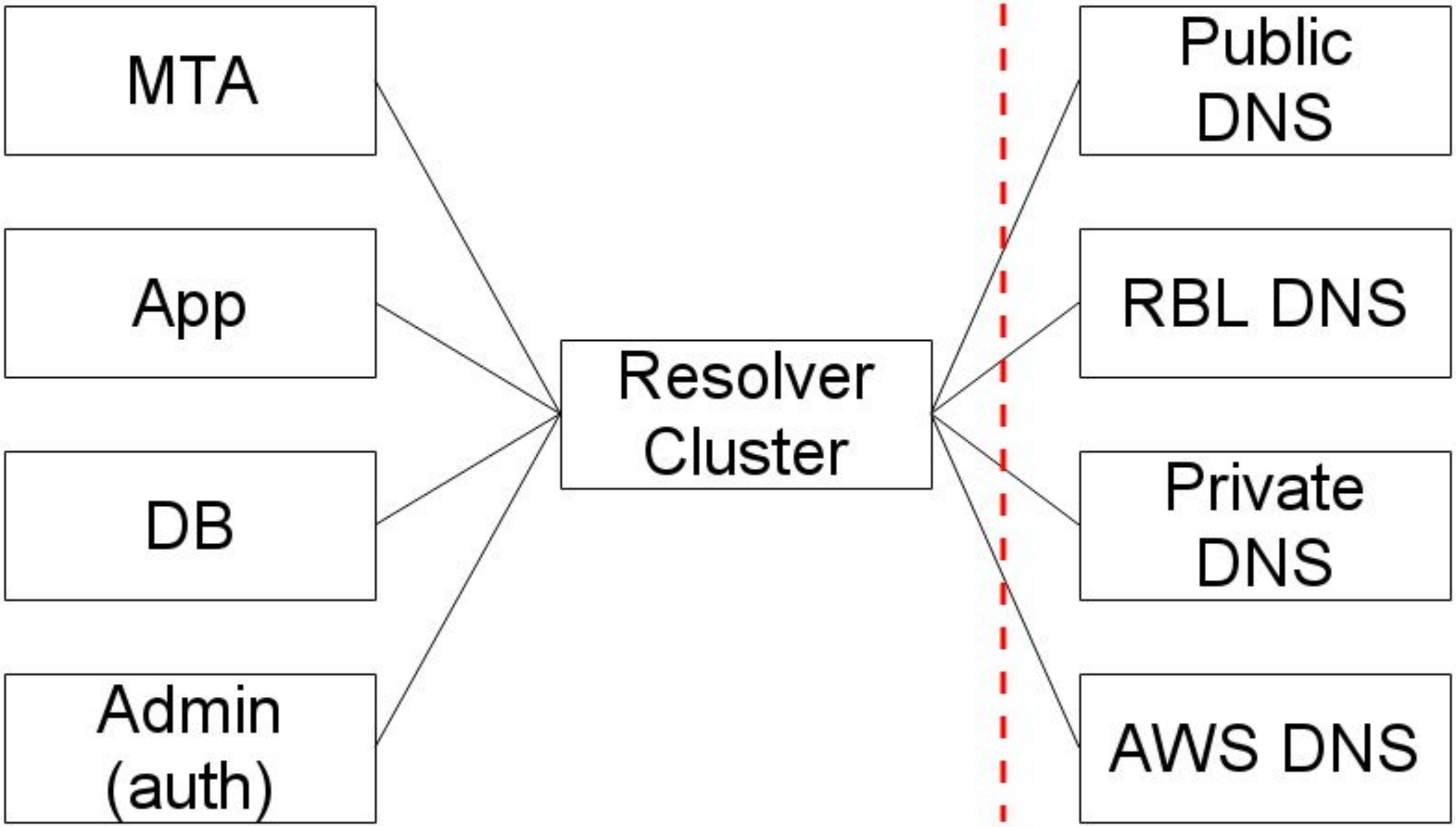
Public DNS

RBL DNS

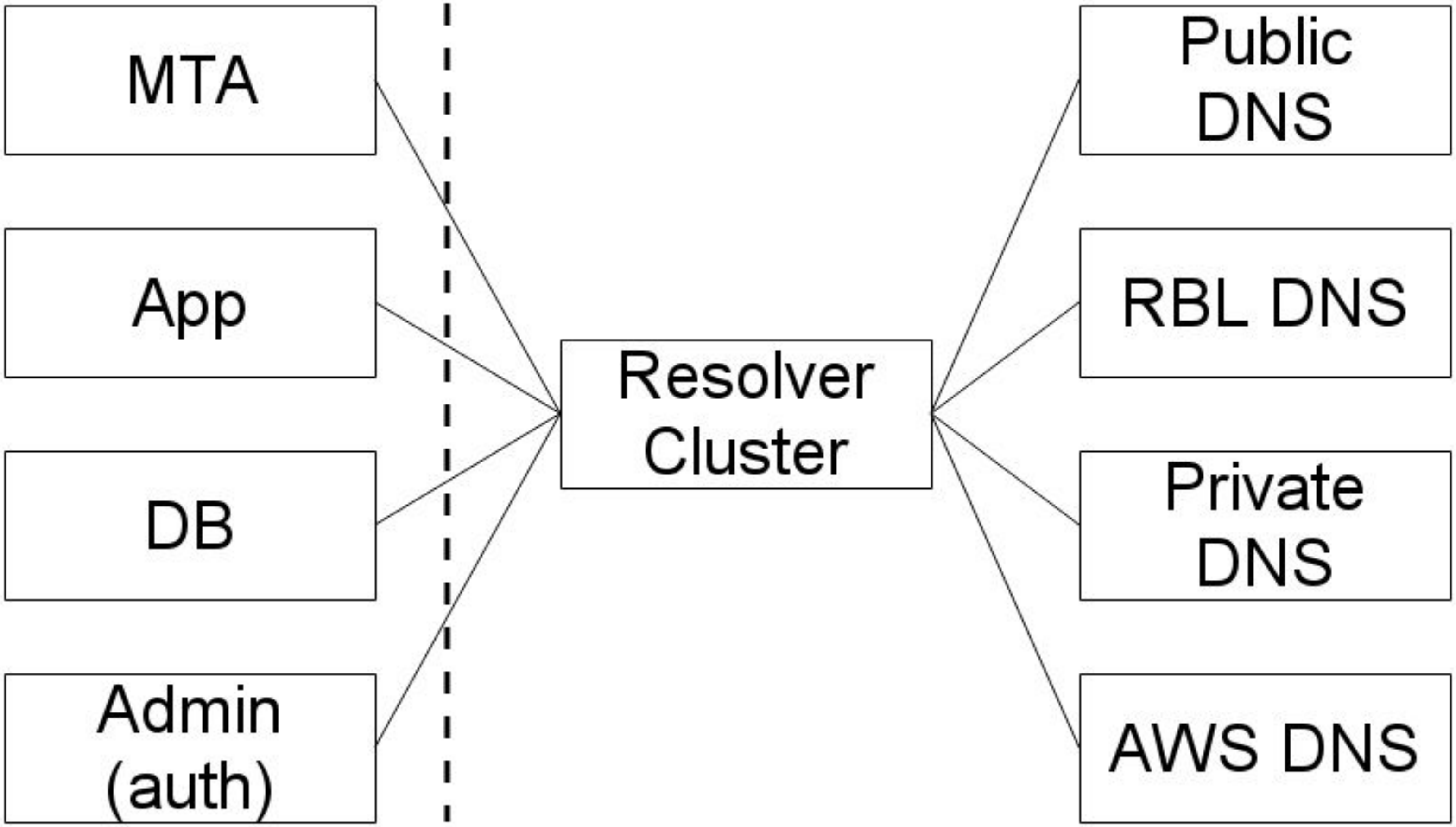
Private DNS

AWS DNS

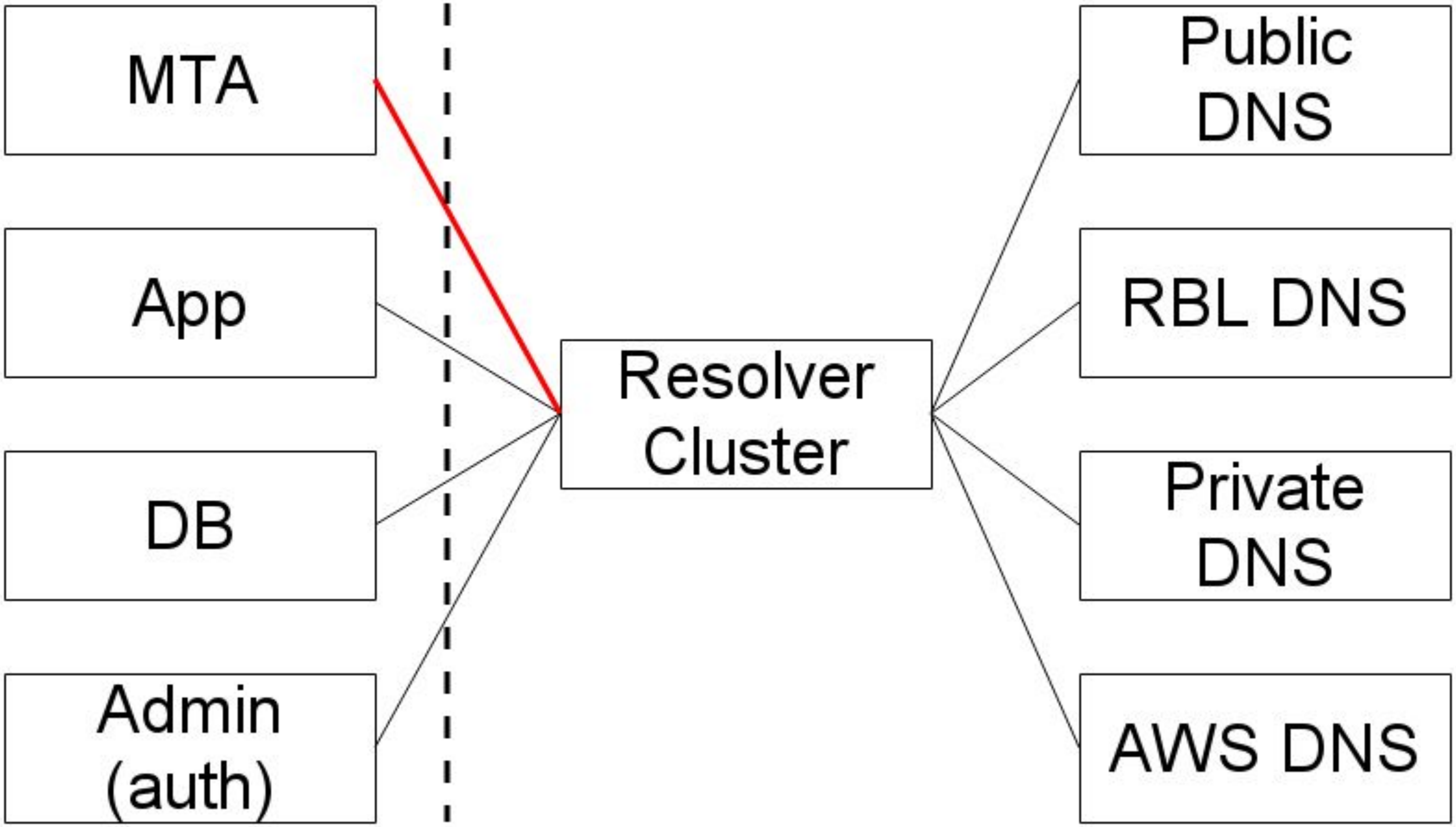
Version 1, Centralized Internal Resolver Cluster



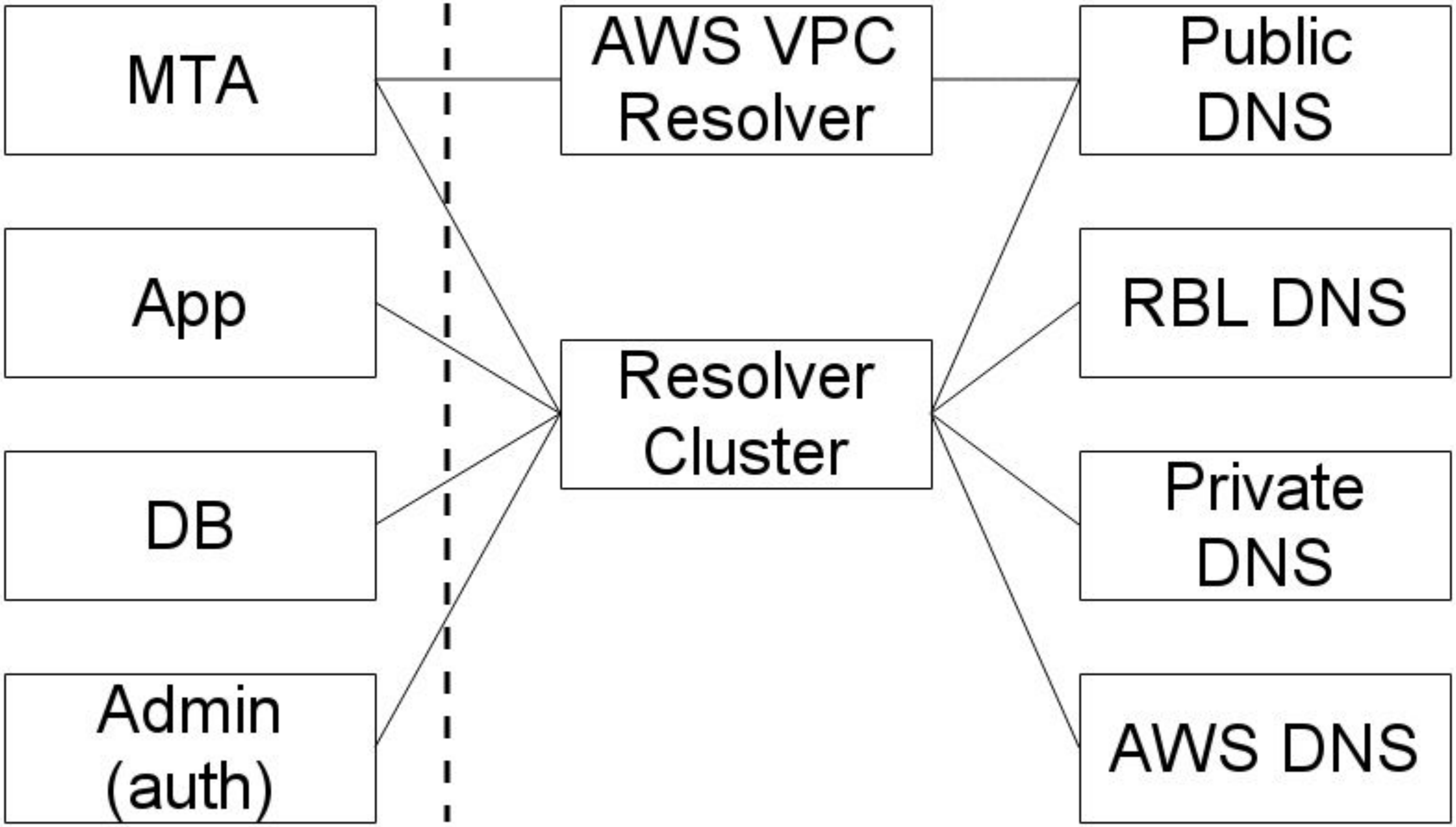
Version 1, Centralized Internal Resolver Cluster



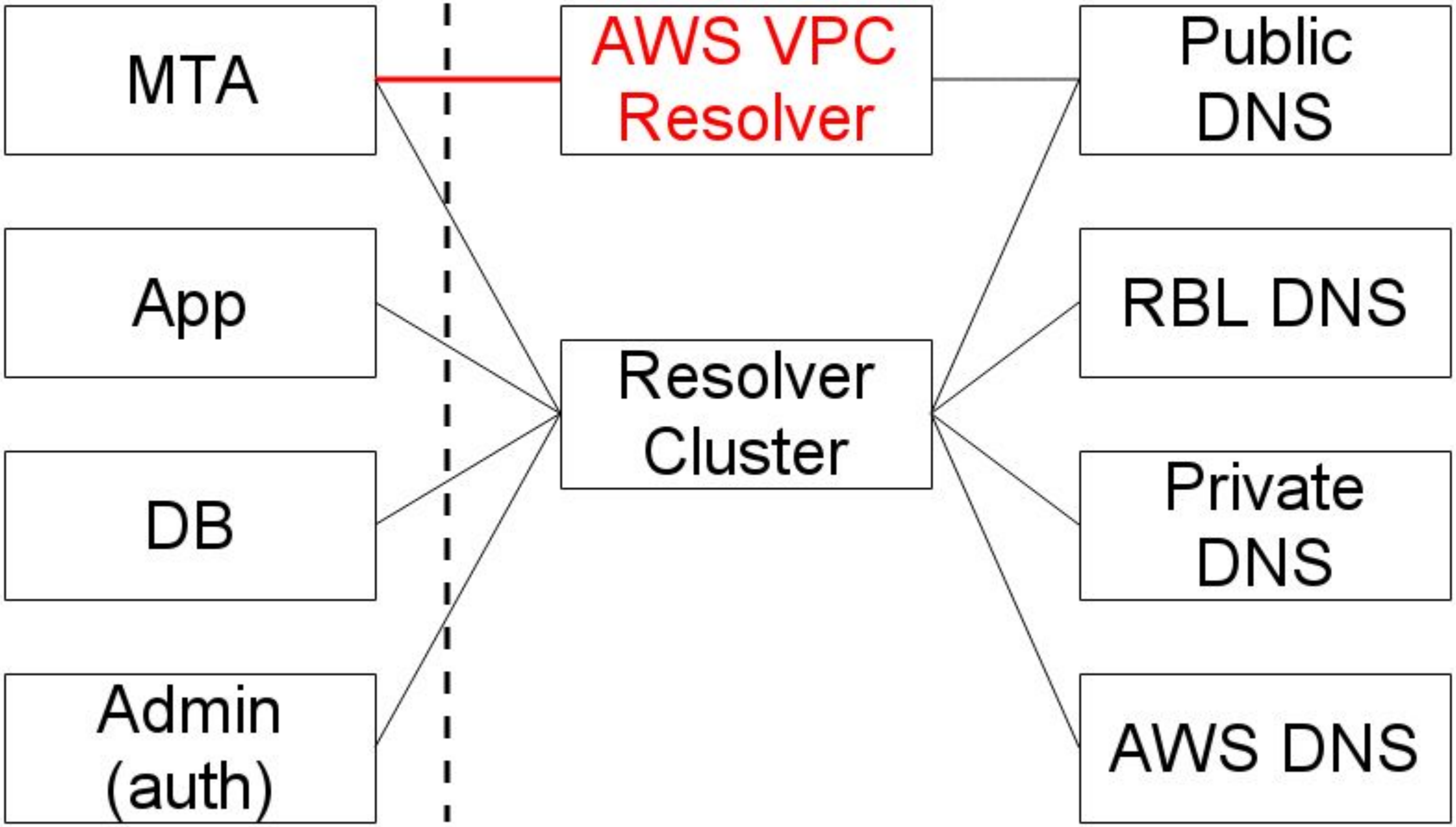
Version 1, Centralized Internal Resolver Cluster



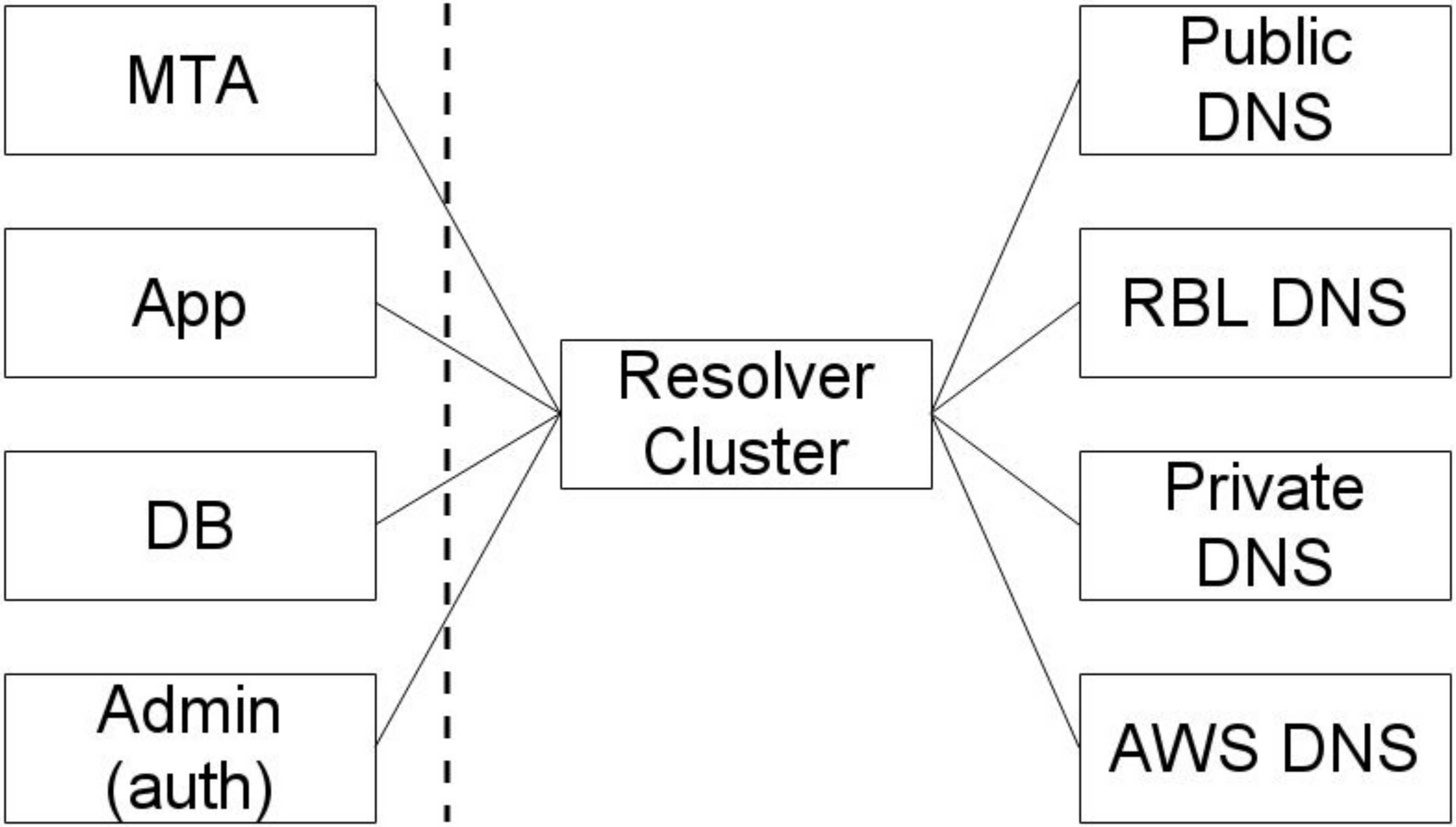
Version 2, AWS VPC Resolver



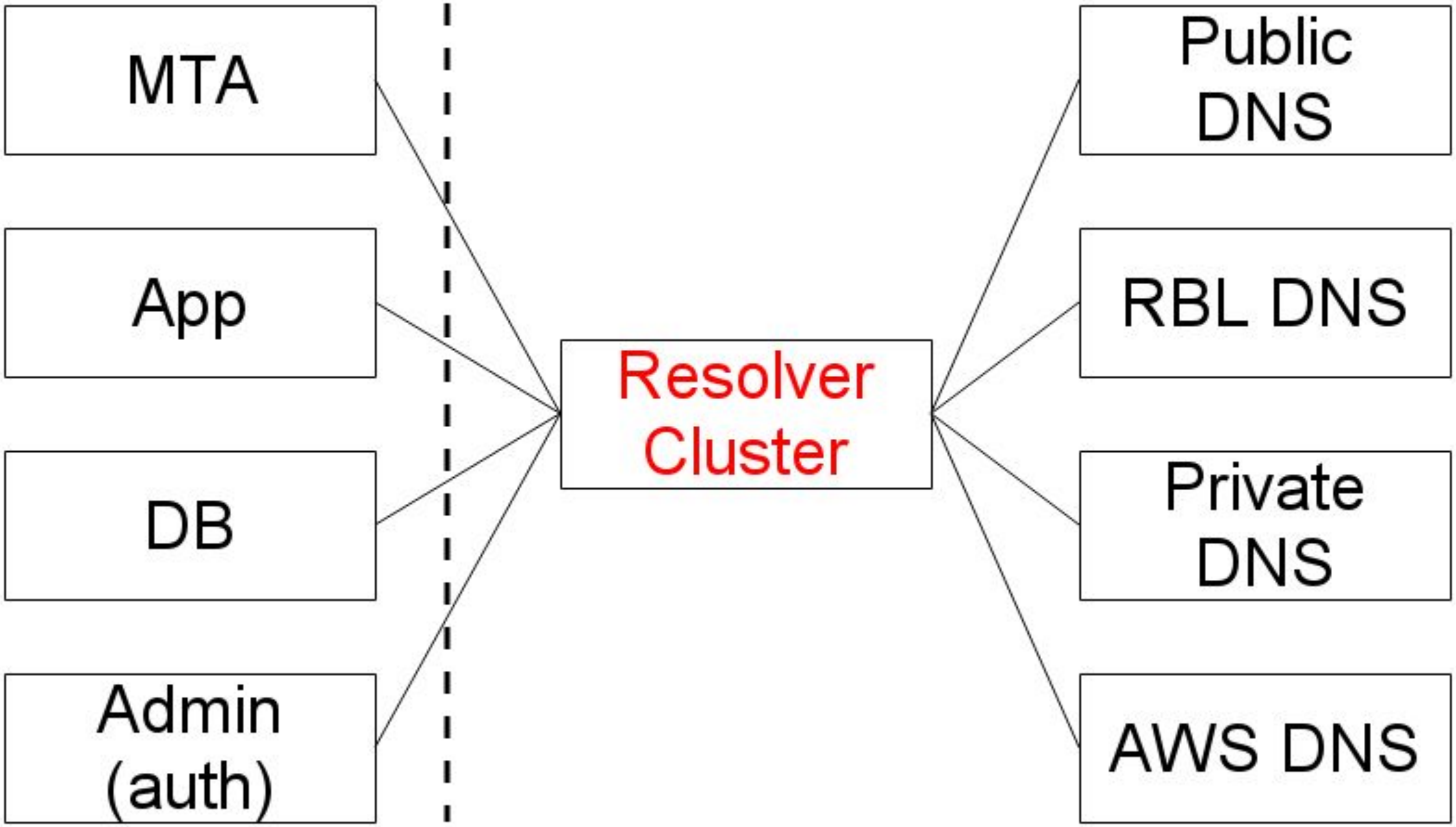
Version 2, AWS VPC Resolver



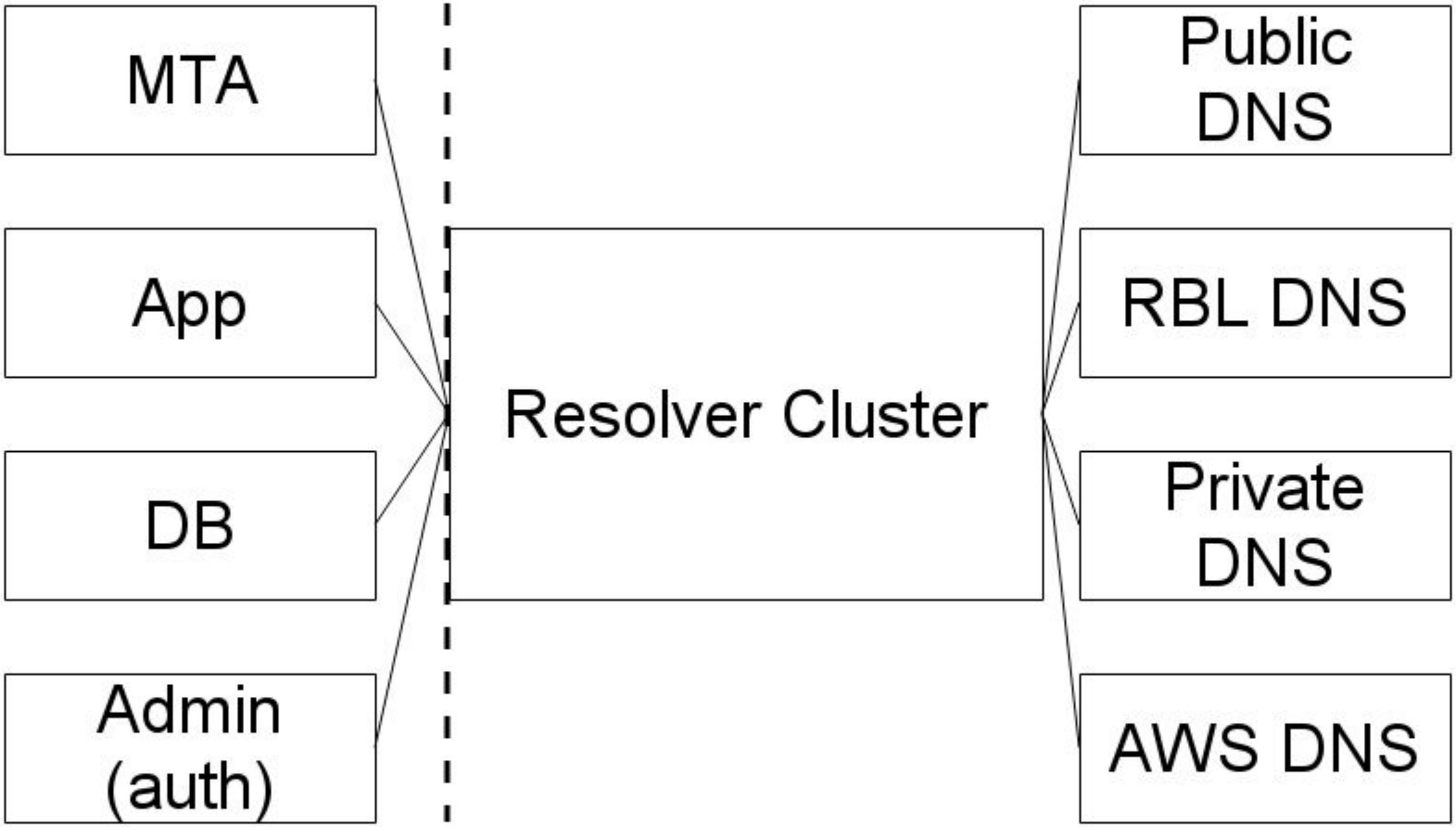
Version 1.5, Centralized Internal Resolver Cluster



Version 1.5, Centralized Internal Resolver Cluster



Version 3.14, Centralized Internal Resolver Cluster



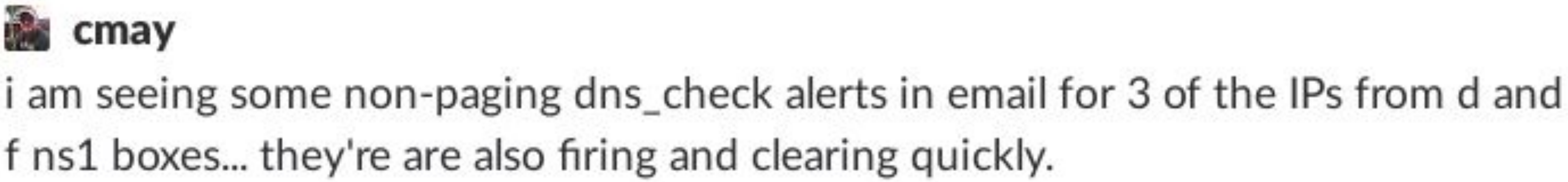
May 2017

May 2017 Outage

- A day like any other day until…

May 2017 Outage

- A day like any other day until…

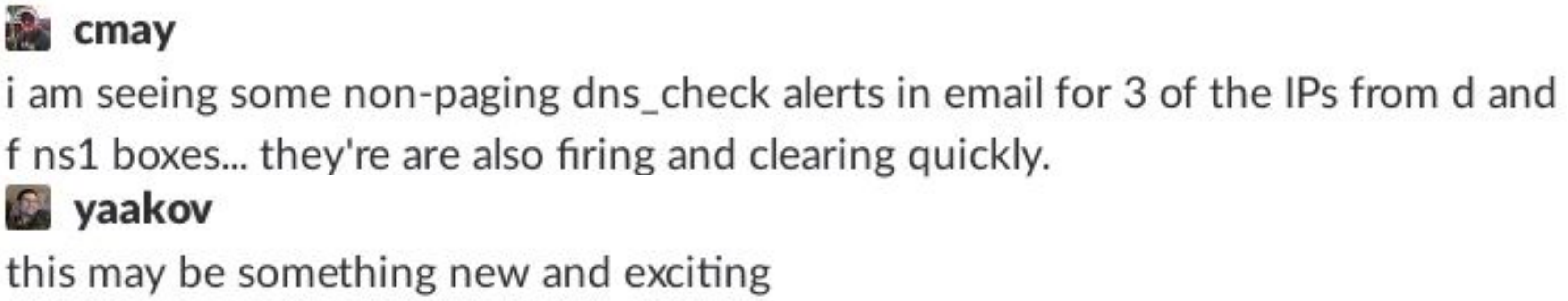


cmay

i am seeing some non-paging dns_check alerts in email for 3 of the IPs from d and f ns1 boxes... they're are also firing and clearing quickly.

May 2017 Outage

- A day like any other day until…



cmay

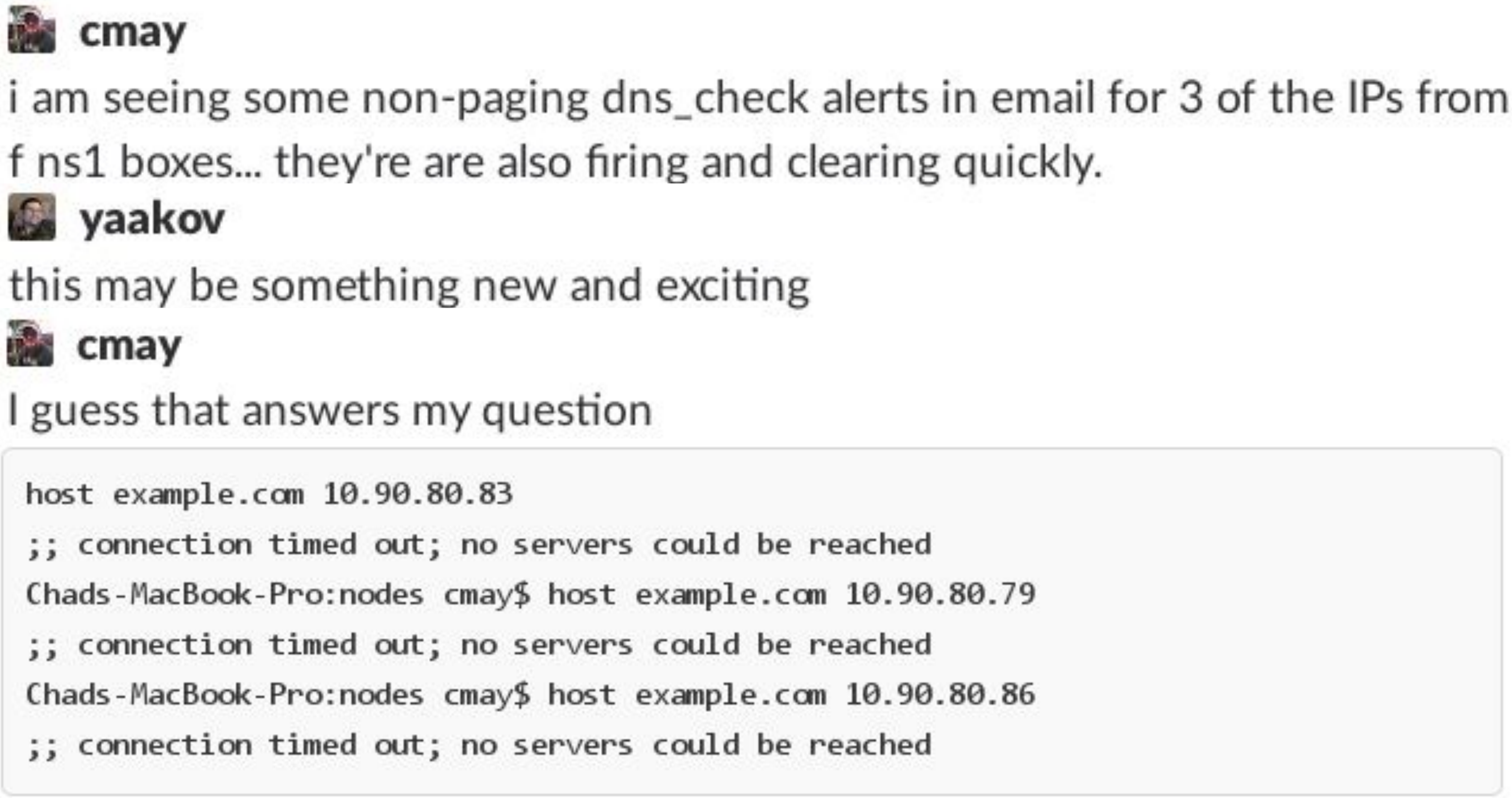
i am seeing some non-paging dns_check alerts in email for 3 of the IPs from d and f ns1 boxes... they're are also firing and clearing quickly.

yaakov

this may be something new and exciting

May 2017 Outage

- A day like any other day until...



cmay

i am seeing some non-paging dns_check alerts in email for 3 of the IPs from d and f ns1 boxes... they're are also firing and clearing quickly.

yaakov

this may be something new and exciting

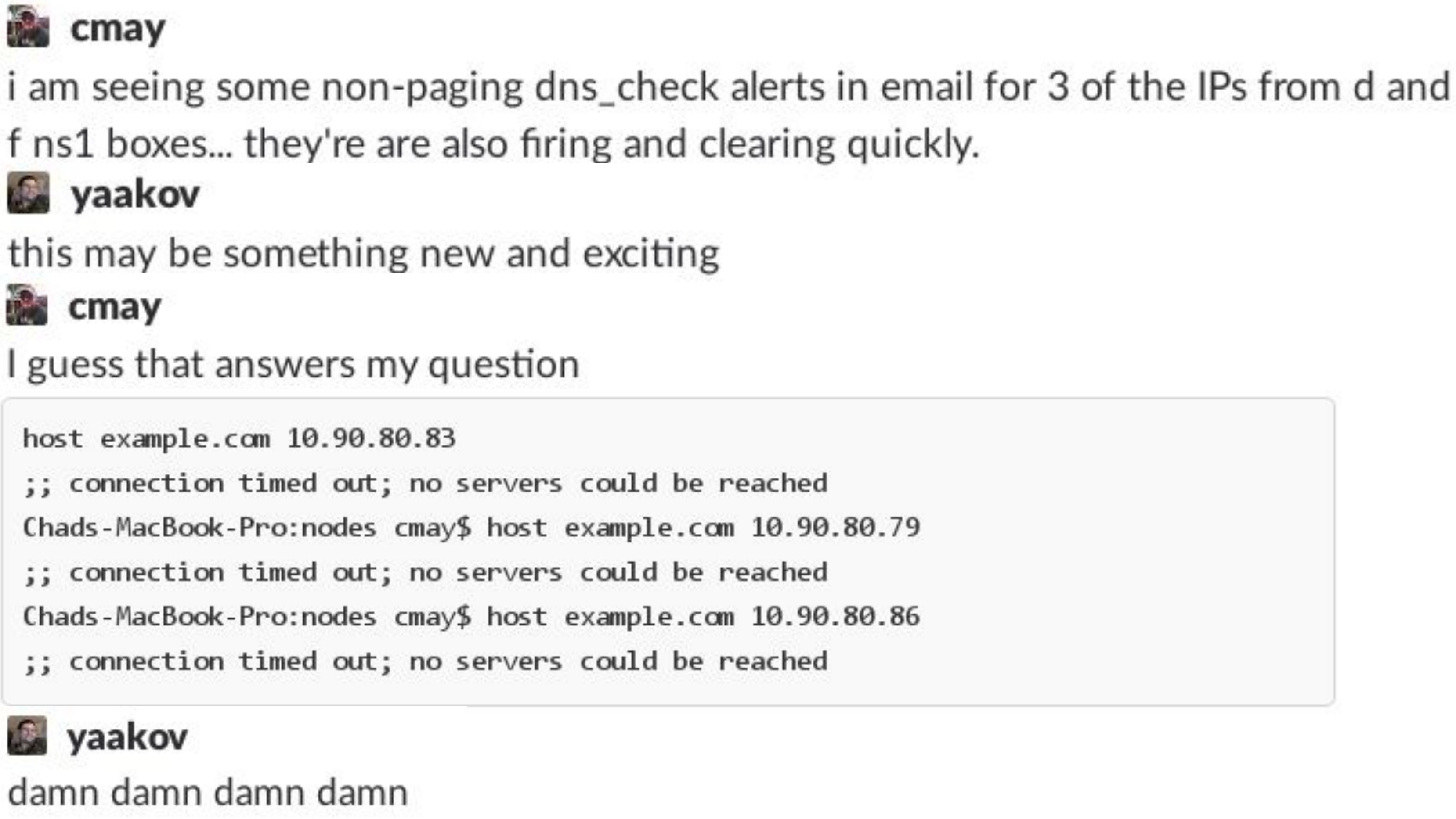
cmay

I guess that answers my question

```
host example.com 10.90.80.83
;; connection timed out; no servers could be reached
Chads-MacBook-Pro:nodes cmay$ host example.com 10.90.80.79
;; connection timed out; no servers could be reached
Chads-MacBook-Pro:nodes cmay$ host example.com 10.90.80.86
;; connection timed out; no servers could be reached
```

May 2017 Outage

- A day like any other day until…



cmay

i am seeing some non-paging dns_check alerts in email for 3 of the IPs from d and f ns1 boxes... they're are also firing and clearing quickly.

yaakov

this may be something new and exciting

cmay

I guess that answers my question

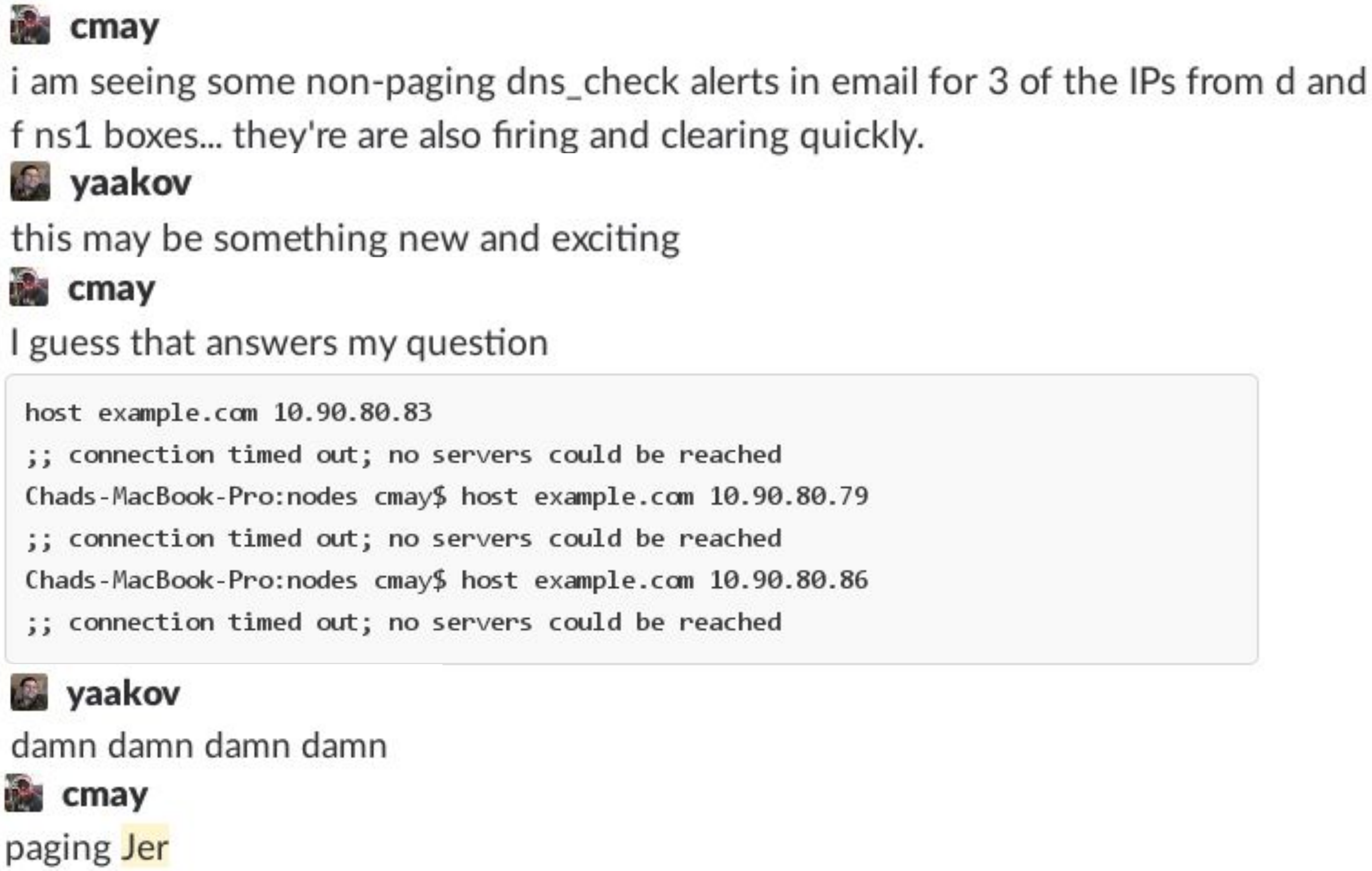
```
host example.com 10.90.80.83
;; connection timed out; no servers could be reached
Chads-MacBook-Pro:nodes cmay$ host example.com 10.90.80.79
;; connection timed out; no servers could be reached
Chads-MacBook-Pro:nodes cmay$ host example.com 10.90.80.86
;; connection timed out; no servers could be reached
```

yaakov

damn damn damn damn

May 2017 Outage

- A day like any other day until…



cmay

i am seeing some non-paging dns_check alerts in email for 3 of the IPs from d and f ns1 boxes... they're are also firing and clearing quickly.

yaakov

this may be something new and exciting

cmay

I guess that answers my question

```
host example.com 10.90.80.83
;; connection timed out; no servers could be reached
Chads-MacBook-Pro:nodes cmay$ host example.com 10.90.80.79
;; connection timed out; no servers could be reached
Chads-MacBook-Pro:nodes cmay$ host example.com 10.90.80.86
;; connection timed out; no servers could be reached
```

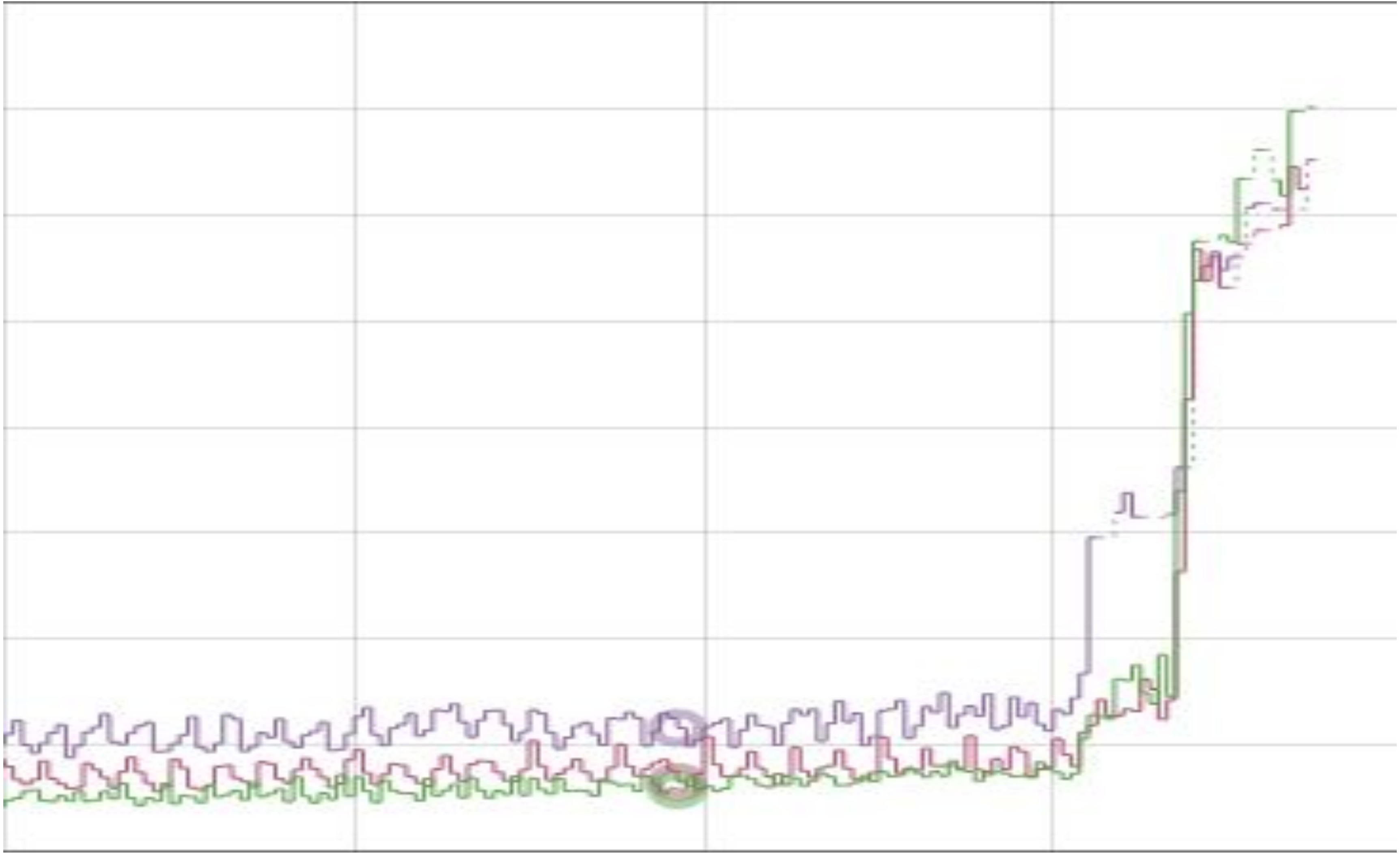
yaakov

damn damn damn damn

cmay

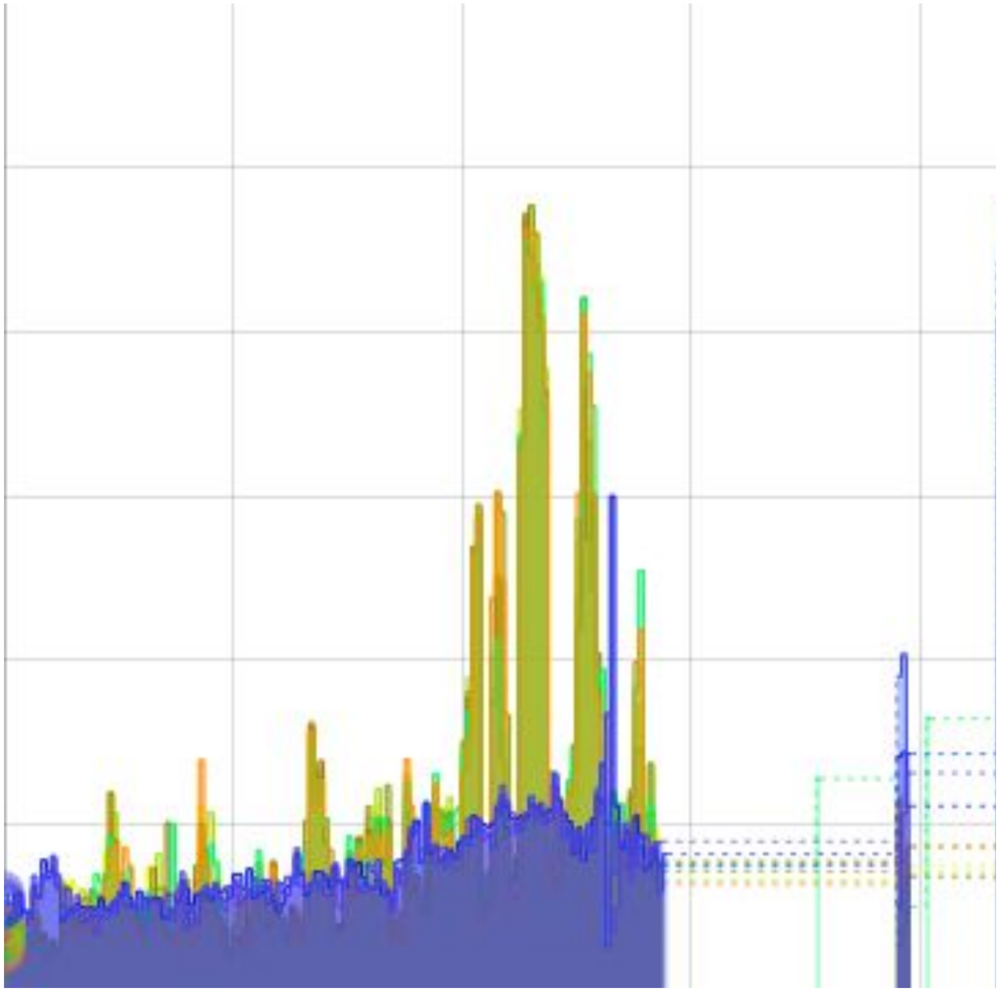
paging Jer

May 2017 Outage



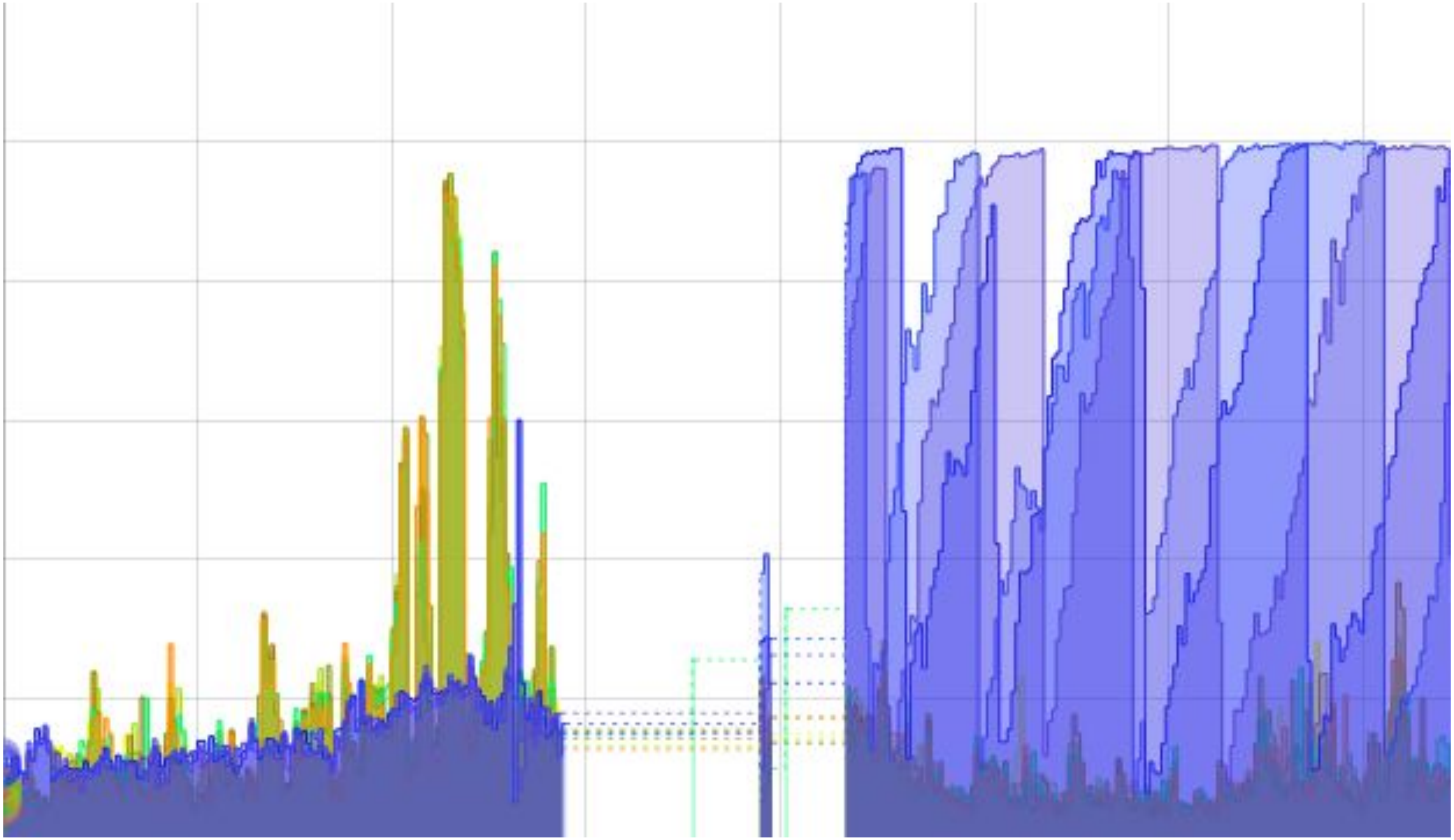
DNS Cluster Aggregate CPU

May 2017 Outage



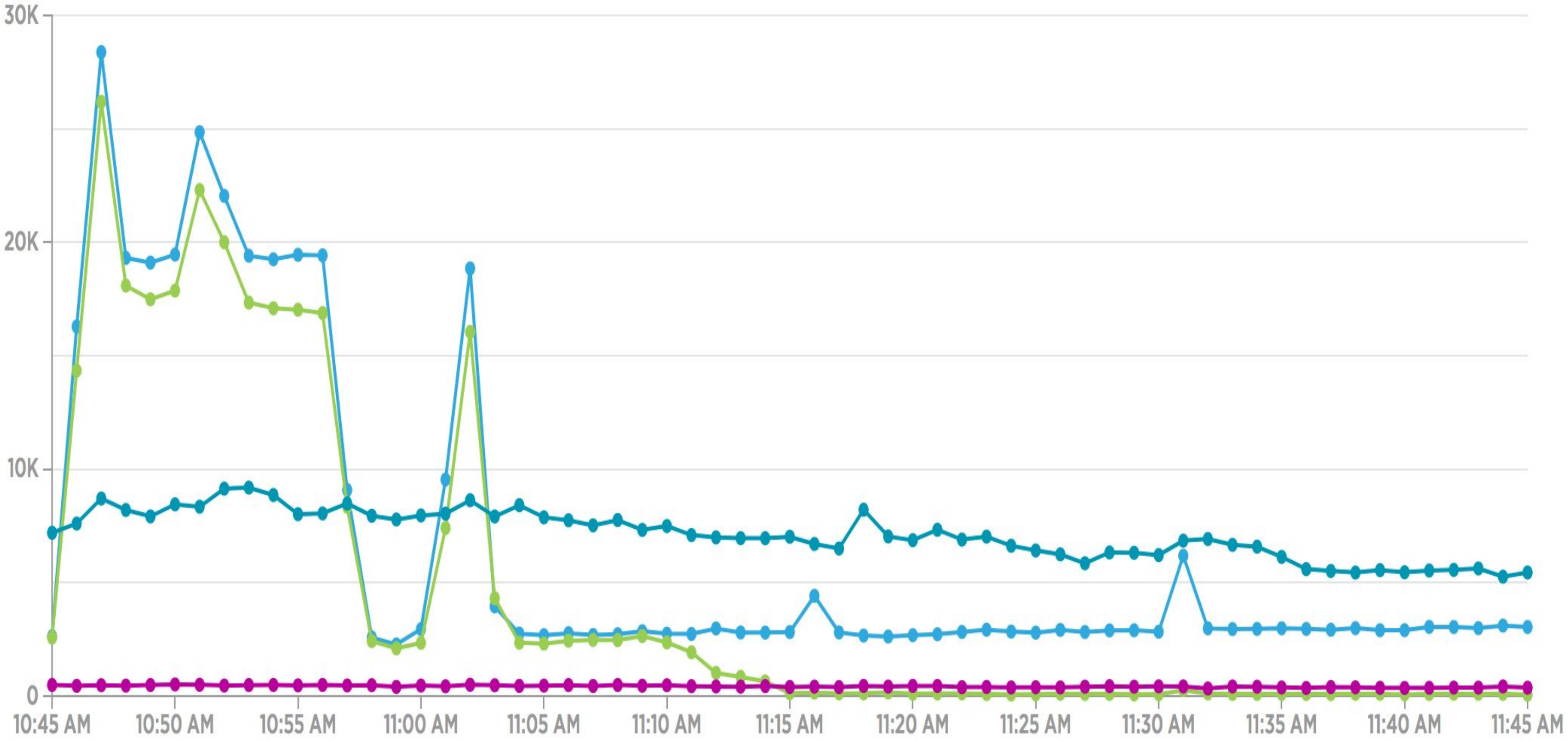
MTA Cluster Aggregate CPU

May 2017 Outage



MTA Cluster Aggregate CPU

May 2017 Outage



**Mail Delivery
(one customer)**

May 2017 Outage

(Near) Total Impact

- Sending mail

May 2017 Outage

(Near) Total Impact

- Sending mail
  - (most) customer mail injection not impacted

May 2017 Outage

(Near) Total Impact

- Sending mail
  - (most) customer mail injection not impacted
- App/DB traffic

May 2017 Outage

(Near) Total Impact

- Sending mail
  - (most) customer mail injection not impacted
- App/DB traffic
- Metrics

May 2017 Outage

(Near) Total Impact

- Sending mail
  - (most) customer mail injection not impacted
- App/DB traffic
- Metrics
- Config management (partial)

May 2017 Outage

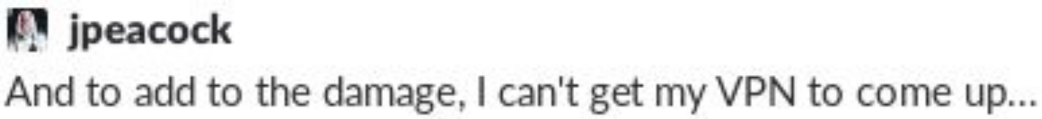
(Near) Total Impact

- Sending mail
  - (most) customer mail injection not impacted
- App/DB traffic
- Metrics
- Config management (partial)
- Admin logins

May 2017 Outage

(Near) Total Impact

- Sending mail
  - (most) customer mail injection not impacted
- App/DB traffic
- Metrics
- Config management (partial)
- Admin logins



jpeacock

And to add to the damage, I can't get my VPN to come up...

May 2017 Outage

May 2017 Outage

Diagnosing Blind

May 2017 Outage

Diagnosing Blind

- Lack of insight into our DNS
- Unable to reach support systems

May 2017 Outage

Diagnosing Blind

- Lack of insight into our DNS
- Unable to reach support systems
- Is it throttling (again)?
- Is it capacity (again)?

May 2017 Outage

Diagnosing Blind

- Lack of insight into our DNS
- Unable to reach support systems
- Is it throttling (again)?
    - Central forward to VPC Resolver
        - Immediately overrun
- Is it capacity (again)?
    - Add capacity
        - Immediately affected

May 2017 Outage

Mitigation

- Repoint individual instances to VPC Resolver
    - Edit resolv.conf

May 2017 Outage

resolv.conf

- Limited to 3 entries
- Always tried top to bottom
- Limited practical retry
- Read on app startup
  - Changes require restarts

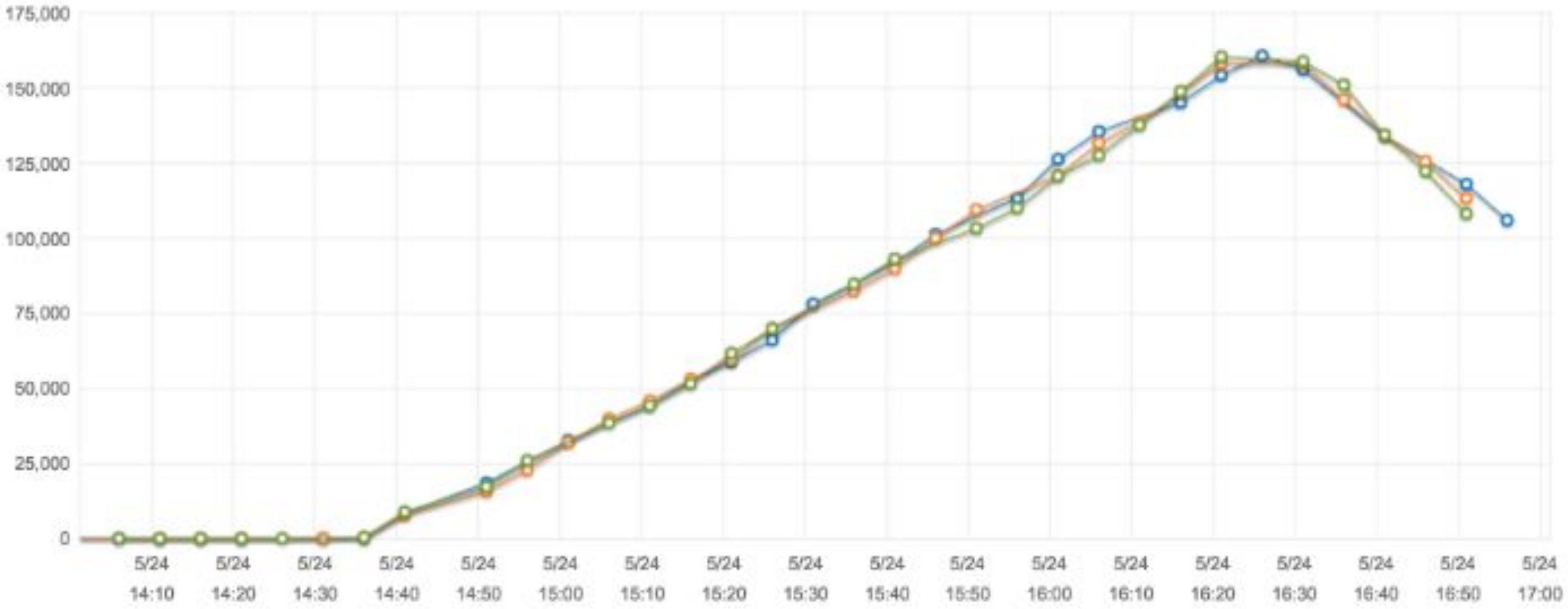
May 2017 Outage

Mitigation

- Repoint individual instances to VPC Resolver
  - Edit resolv.conf, with restarts
  - Provided breathing room
- Main resolver cluster recovered as load was removed
- App tier recovery: 2 hours
- Major customer mail recovery: 4-5 hours
- Time to full recovery: 7 hours

May 2017 Outage

Mitigation

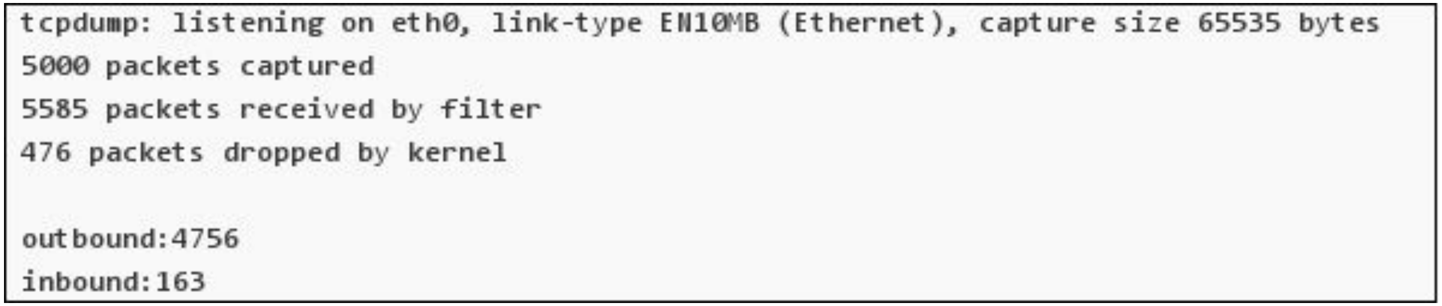


Webhook SQS Queued Messages

May 2017 Outage

Diagnosis

- Asymmetric DNS packet flow
  - Tcpdump
  - AWS Network Flow Logs



```
tcpdump: listening on eth0, link-type EN10MB (Ethernet), capture size 65535 bytes
5000 packets captured
5585 packets received by filter
476 packets dropped by kernel

outbound:4756
inbound:163
```

- Average 300 responses per 5000 queries (94% failure)

May 2017 Outage

The Cause?

May 2017 Outage

The Cause?

Connection Tracking

May 2017 Outage

The Cause?

[Undocumented] Connection Tracking

May 2017 Outage

The Cause?

[Undocumented] Connection Tracking

May 2017 Outage

After Action Conclusions

- Incident response process was functional
- Ability to respond via the process was compromised
- Limits of iteration
- New DNS design required

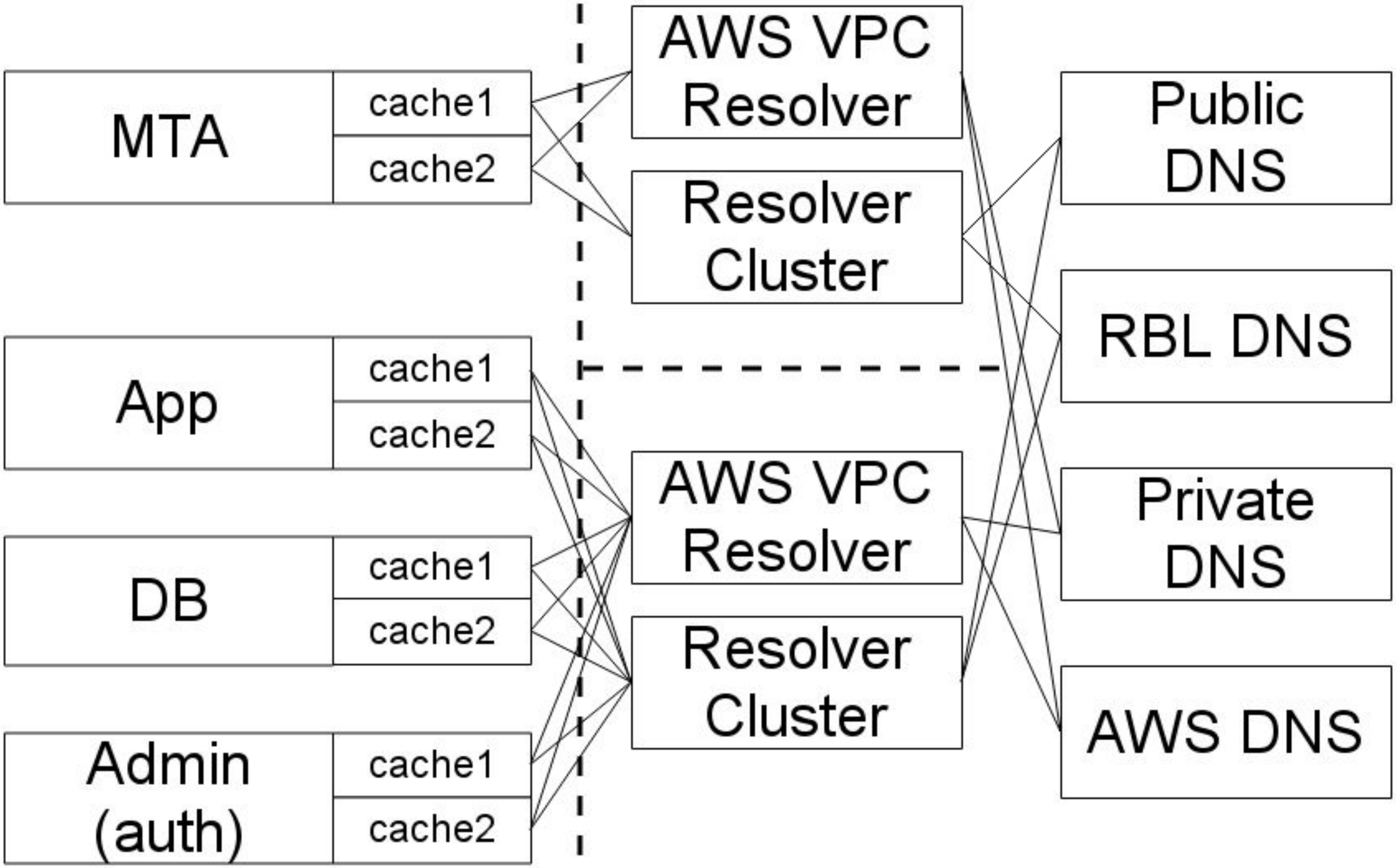
New DNS Design

New DNS Design

Requirements

- Resolve all needed name sources
- Modifiable without changing resolv.conf
- Avoid throttling
- No conntrack
- Multi cluster / isolate components
- Distributed across resolver clusters
- Minimize latency
- Effective caching
- Respect TTLs
- Increase DNS profiling and monitoring

New DNS Design

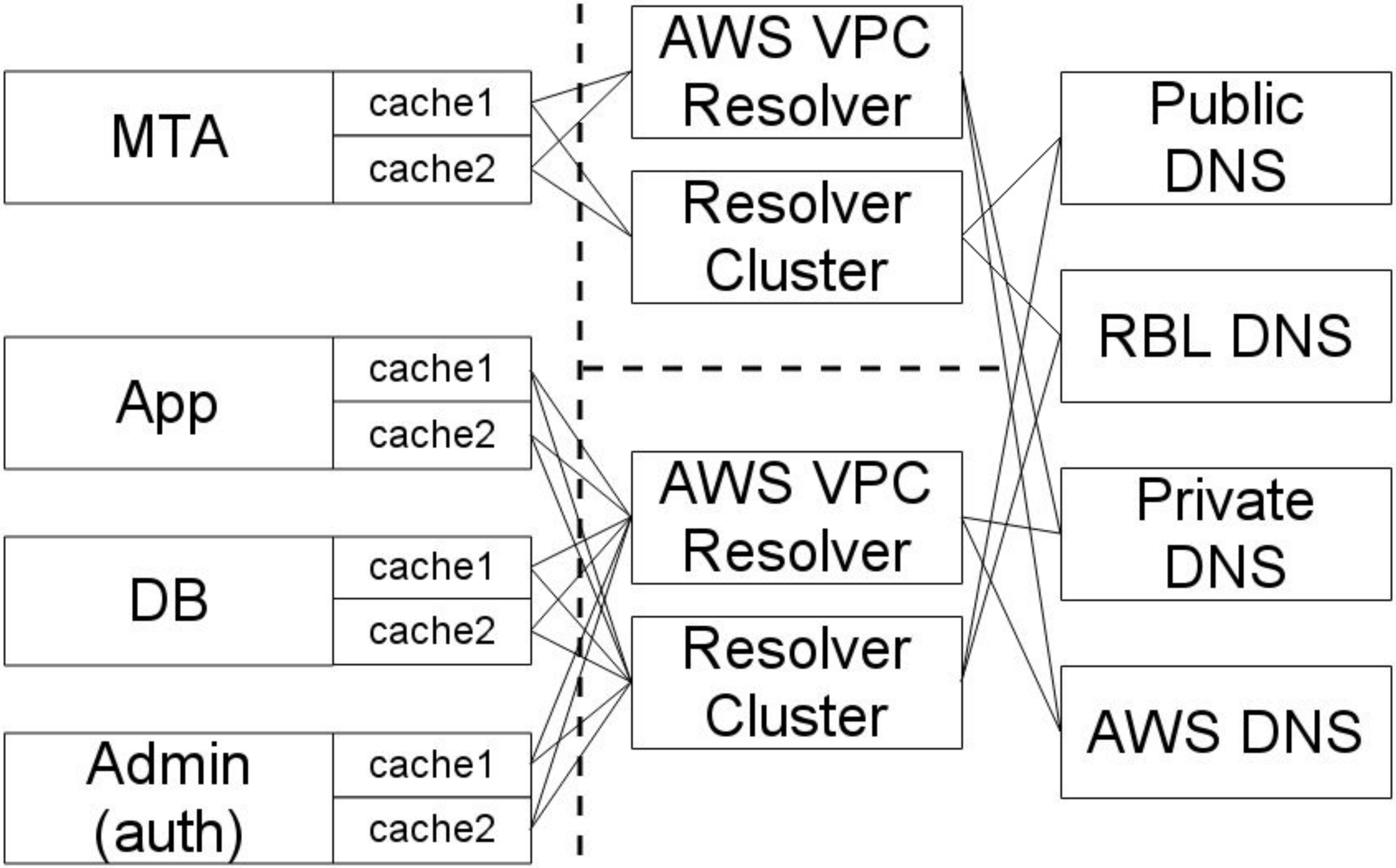


New DNS Design

Network Configuration

- Dedicated VPC for isolation
- Open Security Groups with stateless ACLs
- Separate resolver clusters to isolate impacts
- Query traffic favors same Availability Zone

New DNS Design



New DNS Design

Resolver (Unbound) Configuration

- Instance and service tuning
- Multiple network interfaces per instance
- Multiple IPs per interface
- “serve-expired” enabled

New DNS Design

OS Configuration

- Two local cache services
- 127.0.0.1 routes to resolvers in same AZ
- 127.0.0.2 routes to resolvers in other AZs

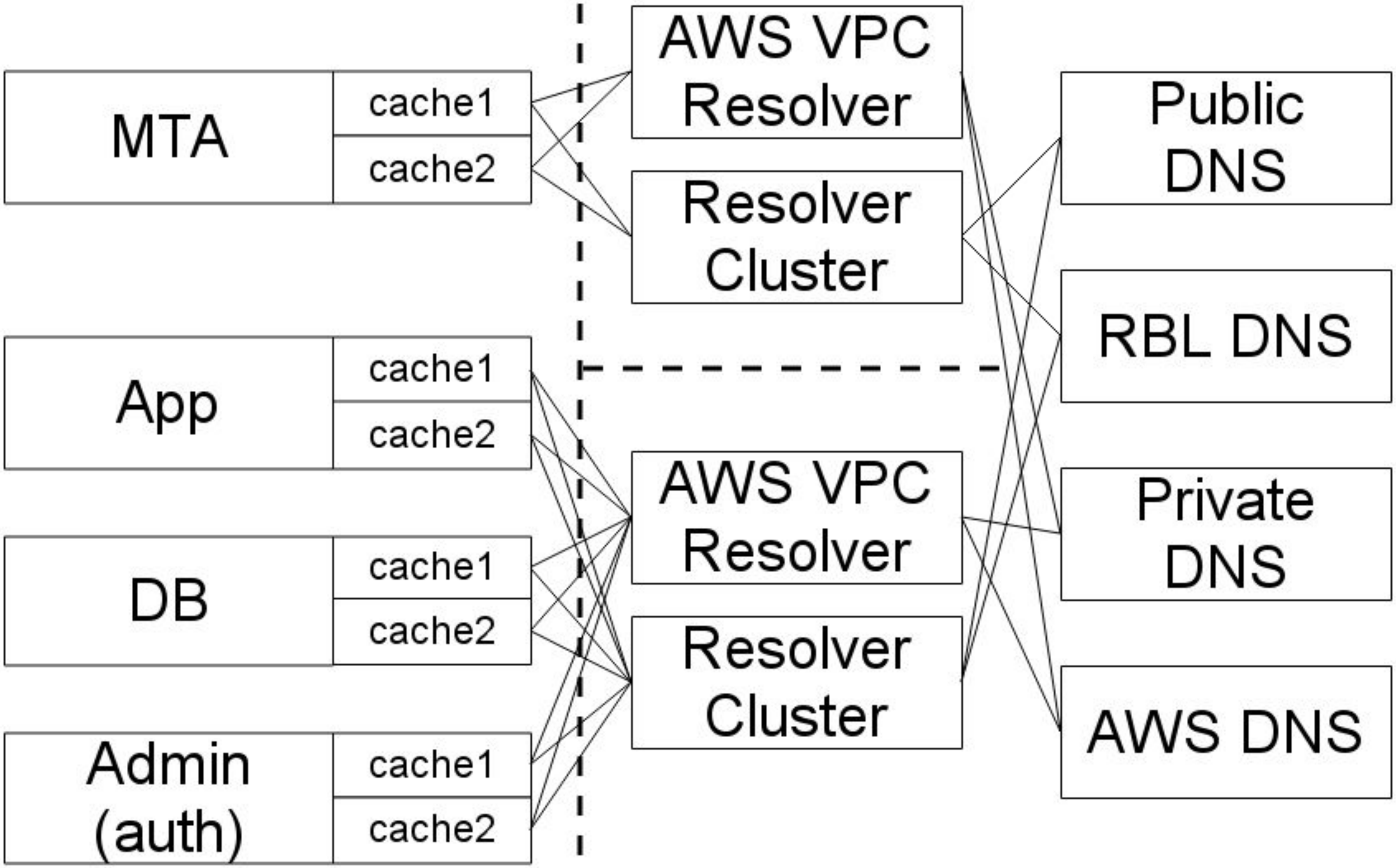
dnsmasq Configuration

- Max concurrency
- Max cache size

/etc/resolv.conf points to:

- 127.0.0.1
- 127.0.0.2
- direct resolver IP

New DNS Design



MTA

cache1

cache2

App

cache1

cache2

DB

cache1

cache2

Admin (auth)

cache1

cache2

AWS VPC Resolver

Resolver Cluster

AWS VPC Resolver

Resolver Cluster

Public DNS

RBL DNS

Private DNS

AWS DNS

Lessons Learned / Remembered

- AWS' main service model is pull, not push

Lessons Learned / Remembered

- AWS’ main service model is pull, not push
- Not all cloud provider limits are apparent
  - make sure they understand your business

Lessons Learned / Remembered

- AWS' main service model is pull, not push
- Not all cloud provider limits are apparent
  - make sure they understand your business
- Instrument your support services
  - and protect them from each other

Lessons Learned / Remembered

- AWS’ main service model is pull, not push
- Not all cloud provider limits are apparent
    - make sure they understand your business
- Instrument your support services
    - and protect them from each other
- resolv.conf is not agile
    - not even eventually consistent

Lessons Learned / Remembered

- AWS’ main service model is pull, not push
- Not all cloud provider limits are apparent
  - make sure they understand your business
- Instrument your support services
  - and protect them from each other
- resolv.conf is not agile
  - not even eventually consistent
- Iteration doesn’t solve it all

Lessons Learned / Remembered

- It’s always a DNS problem

Lessons Learned / Remembered

- It’s always a DNS problem
  - unless it’s a firewall problem

References

- https://d1.awsstatic.com/whitepapers/hybrid-cloud-dns-options-for-vpc.pdf
- https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-network-security.html#security-group-connection-tracking
- http://unbound.net/
- https://docs.aws.amazon.com/AmazonVPC/latest/UserGuide/vpc-dns.html
- http://www.thekelleys.org.uk/dnsmasq/doc.html

Questions?

jblosser@sparkpost.com