Grafana Labs

Kube, Where's My Metrics?

The Challenges of Scaling Multi-Cluster Prometheus

19 March 2024

Iain Lane
Senior Software Engineer

Niko Smeds
Senior Software Engineer

Prometheus 101

- Collects and stores **metrics** as **time-series data**
- Metrics are **scraped** over **HTTP requests**
- Data traditionally stored on **local disks**
- Support for **remote-write** and **remote-read**
- **PromQL** is the Prometheus query language

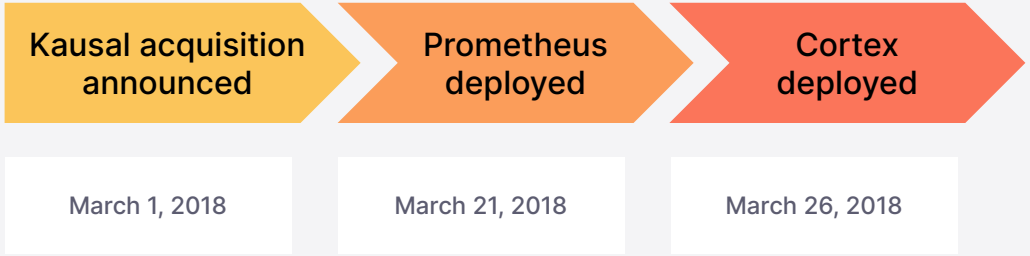
History

Internal monitoring timeline

2016

Internal monitoring timeline

Deploying Prometheus + Cortex



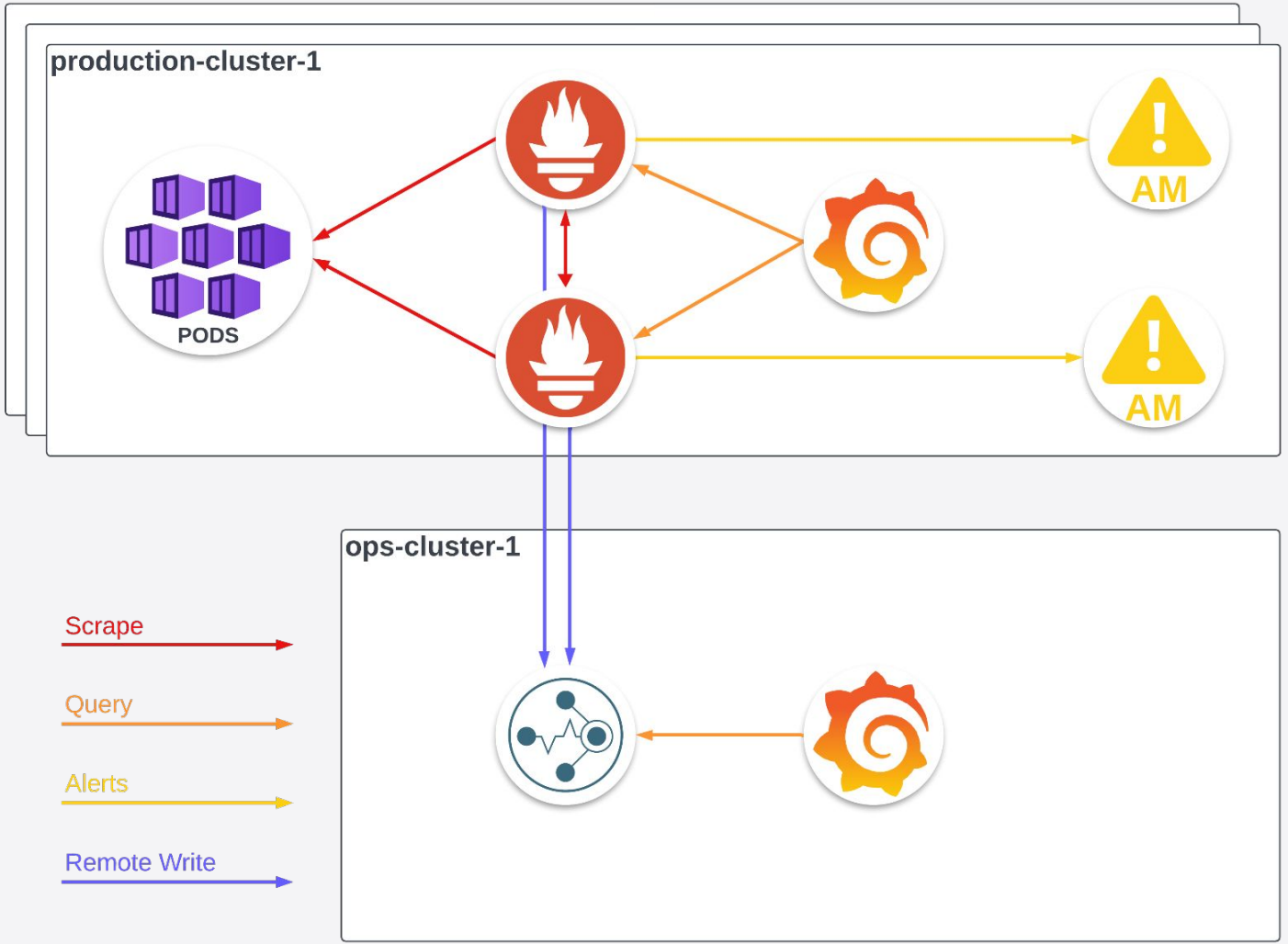
default/prometheus

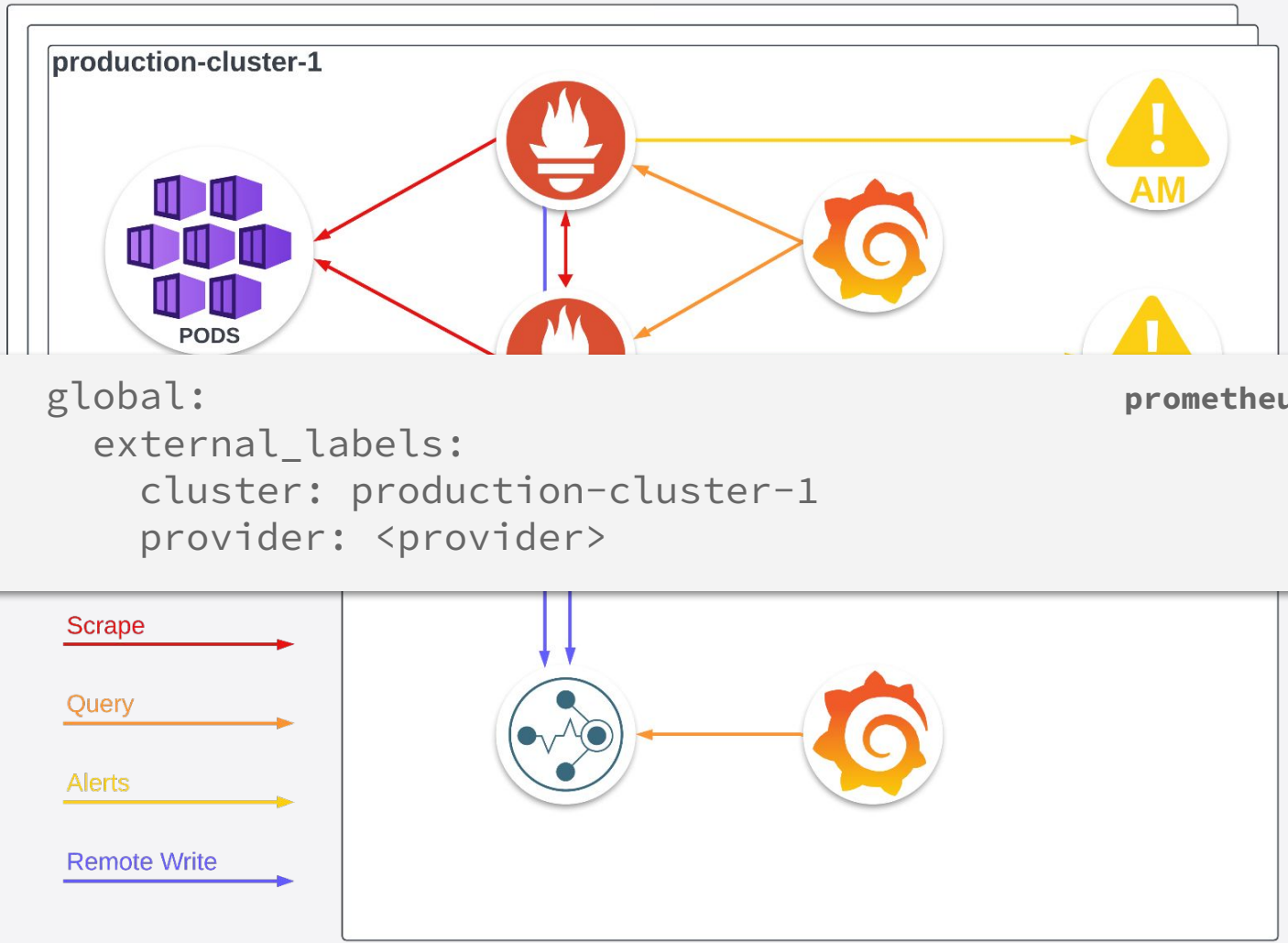
Prometheus pair

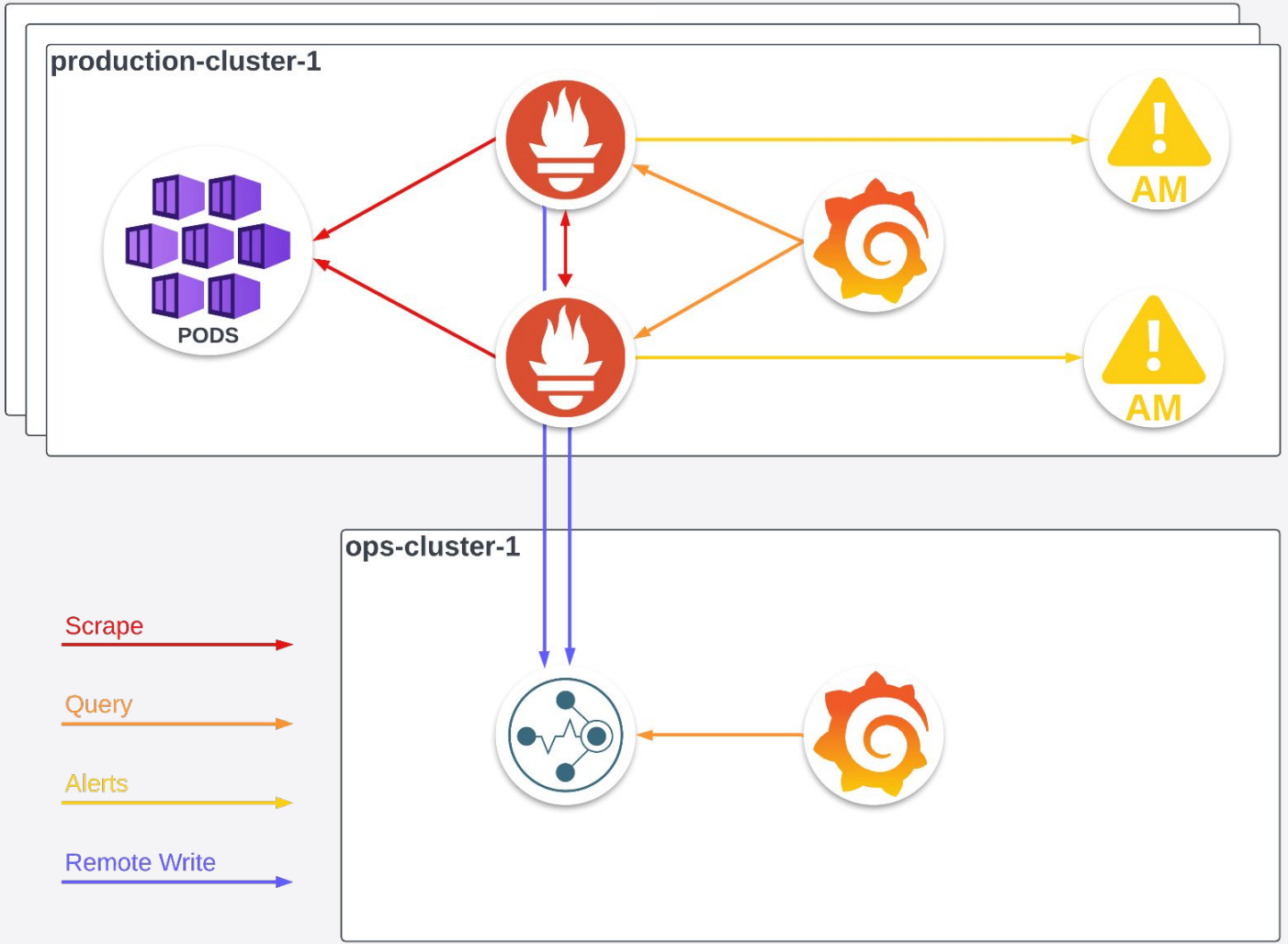
- Deployed in every cluster
- Scrapes all pod and cluster metrics
- Remote-write to central Cortex
- Alert evaluation

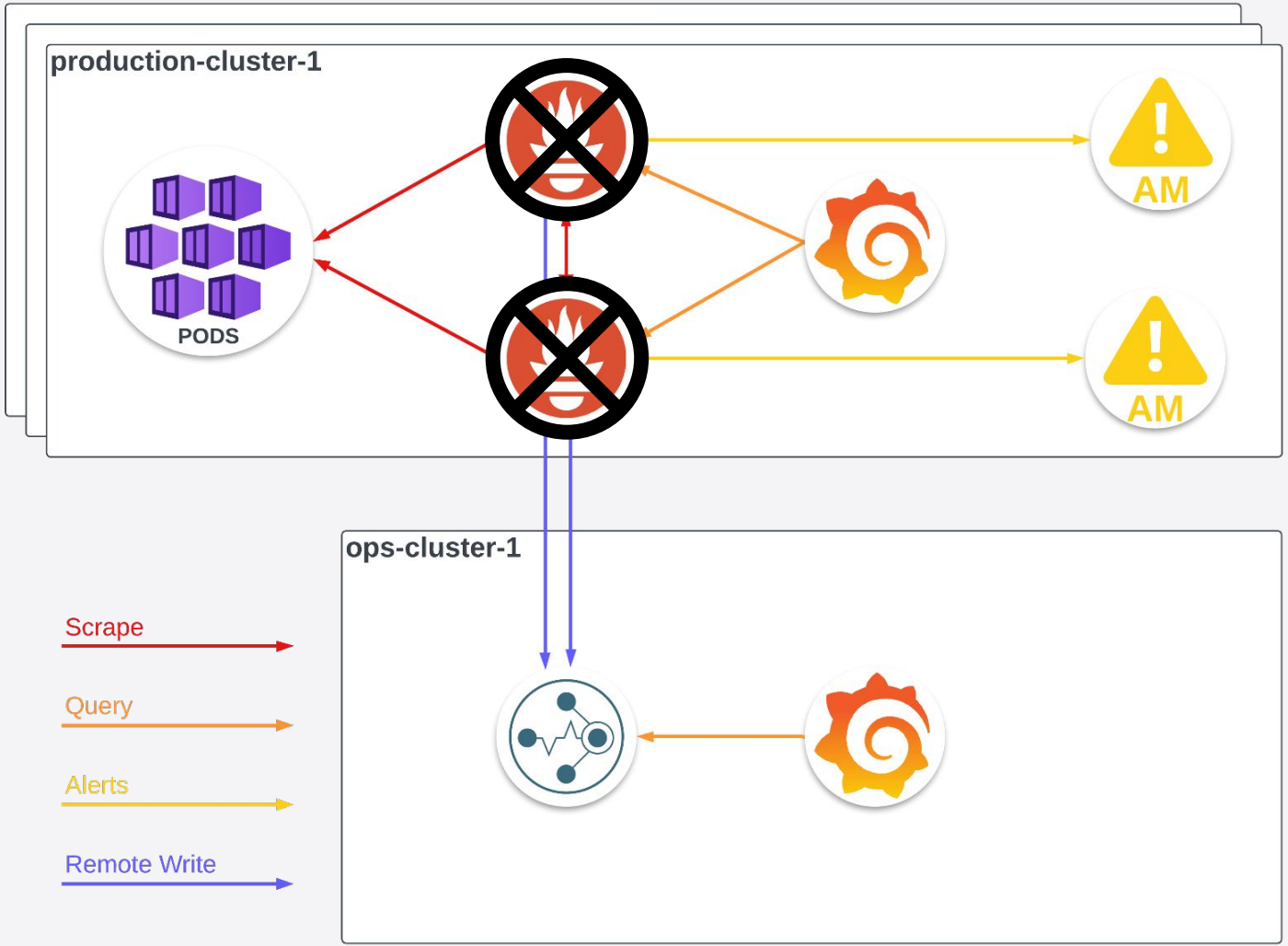
Alertmanager pair

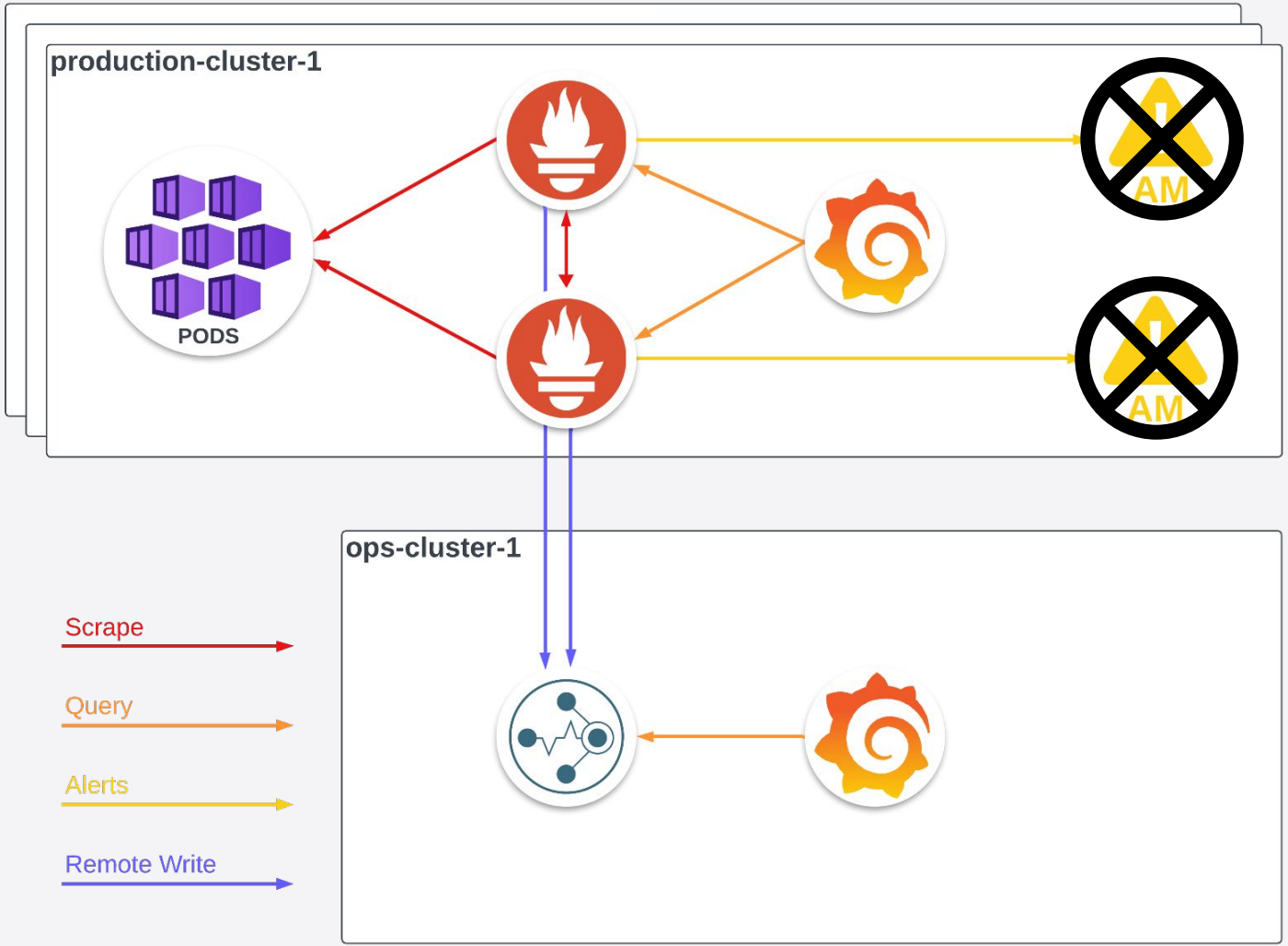
- Deployed in every cluster
- Alert deduplication
- Forward alerts to receivers
- Silence alerts

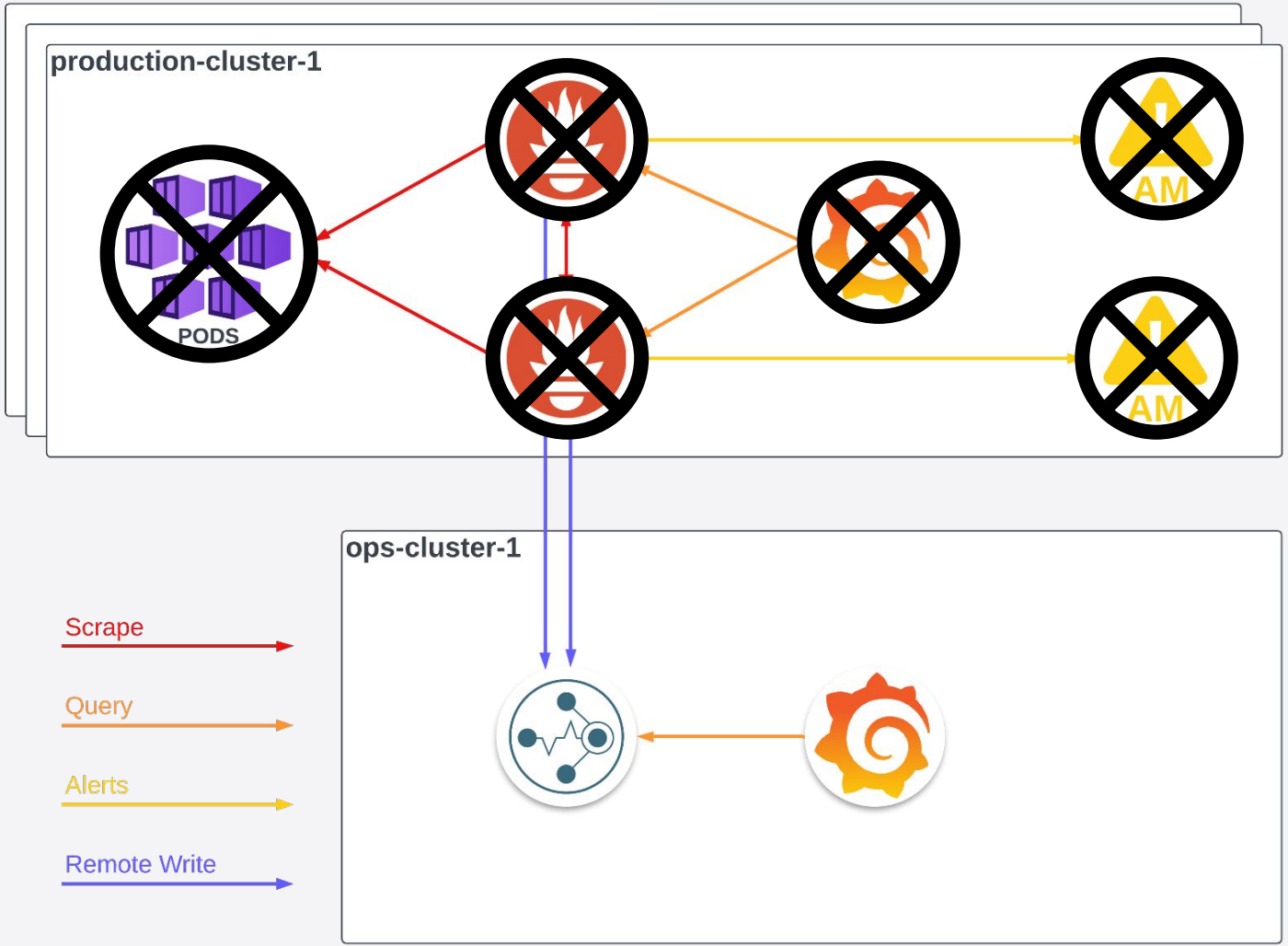

production-cluster-1
PODS
AM
AM
ops-cluster-1
Scrape
Query
Alerts
Remote Write


production-cluster-1
PODS
AM
global:
external_labels:
cluster: production-cluster-1
provider: <provider>
prometheus.yml
Scrape
Query
Alerts
Remote Write

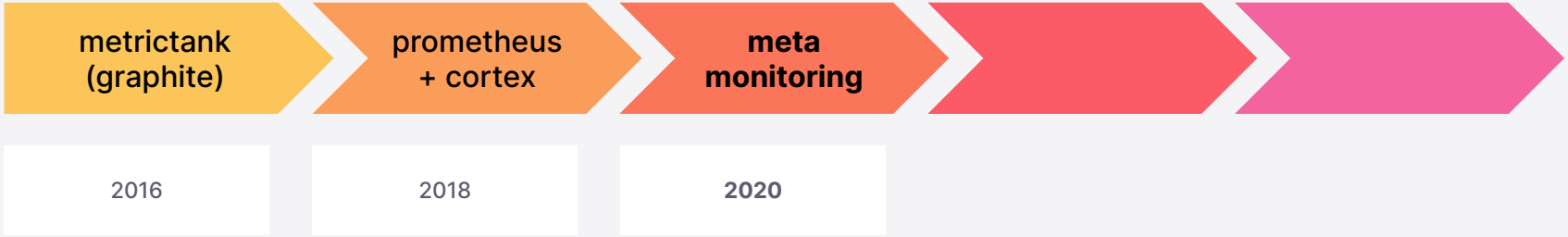

production-cluster-1
PODS
AM
AM
ops-cluster-1
Scrape
Query
Alerts
Remote Write


production-cluster-1
PODS
AM
AM
ops-cluster-1
Scrape
Query
Alerts
Remote Write


production-cluster-1
PODS
AM
AM
ops-cluster-1
Scrape
Query
Alerts
Remote Write


production-cluster-1
PODS
AM
AM
ops-cluster-1
Scrape
Query
Alerts
Remote Write

Internal monitoring timeline



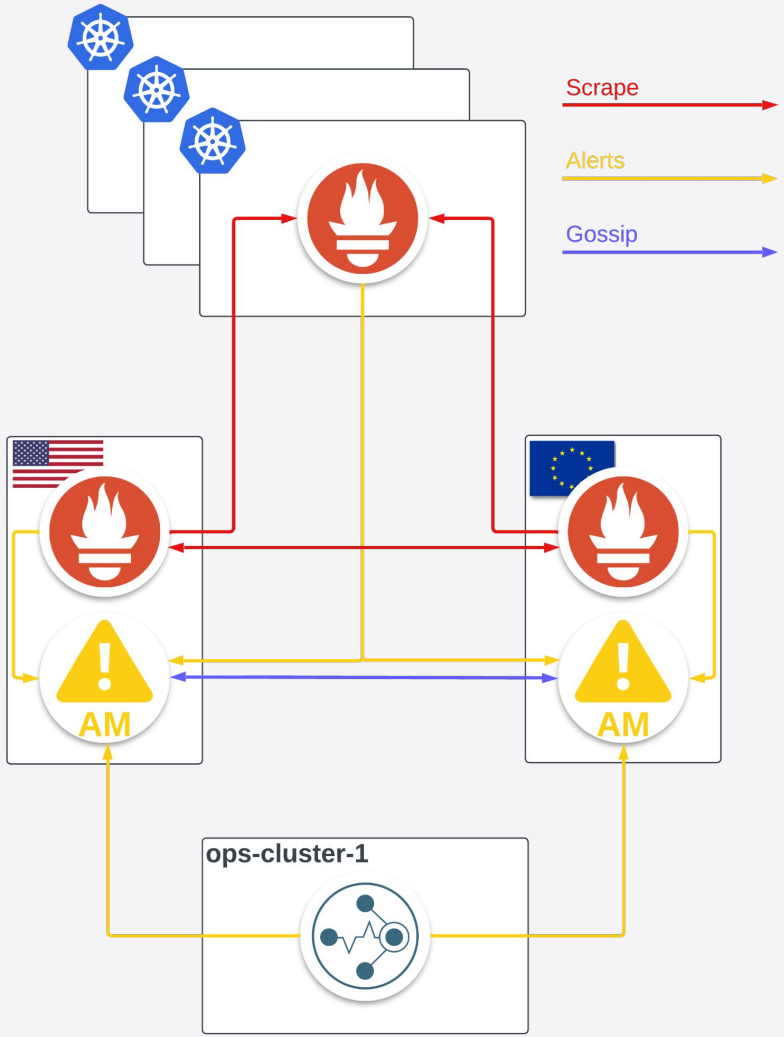
metamonitoring/prometheus

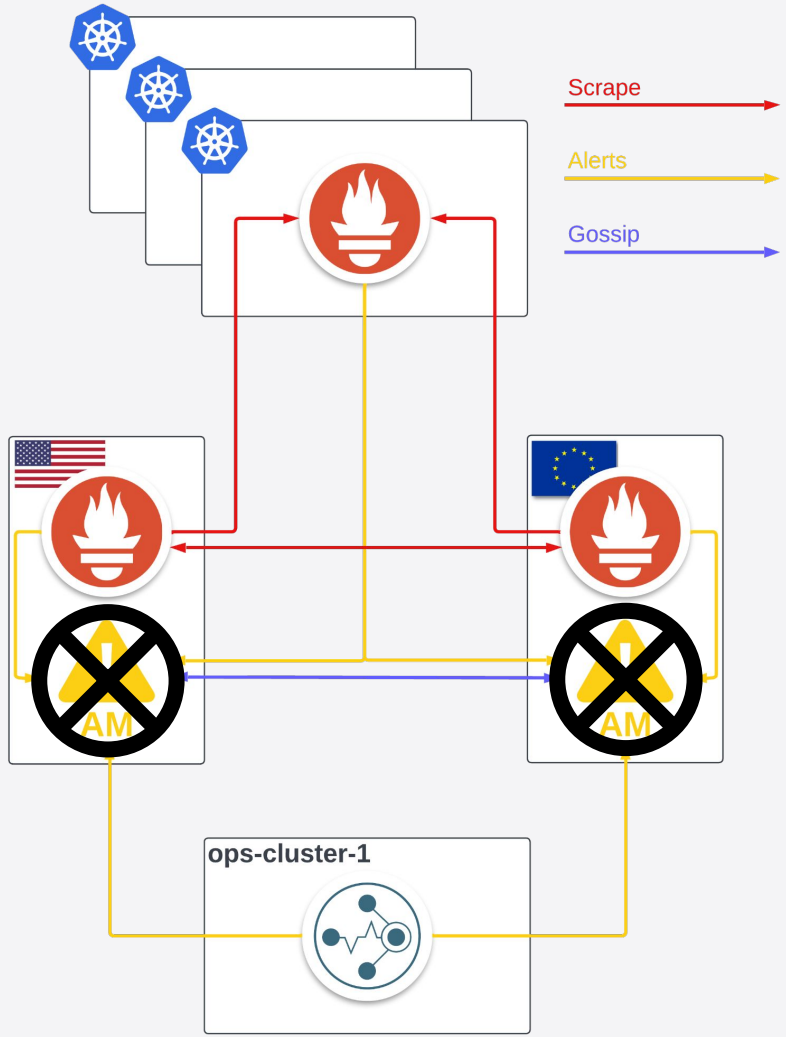
Prometheus pairs

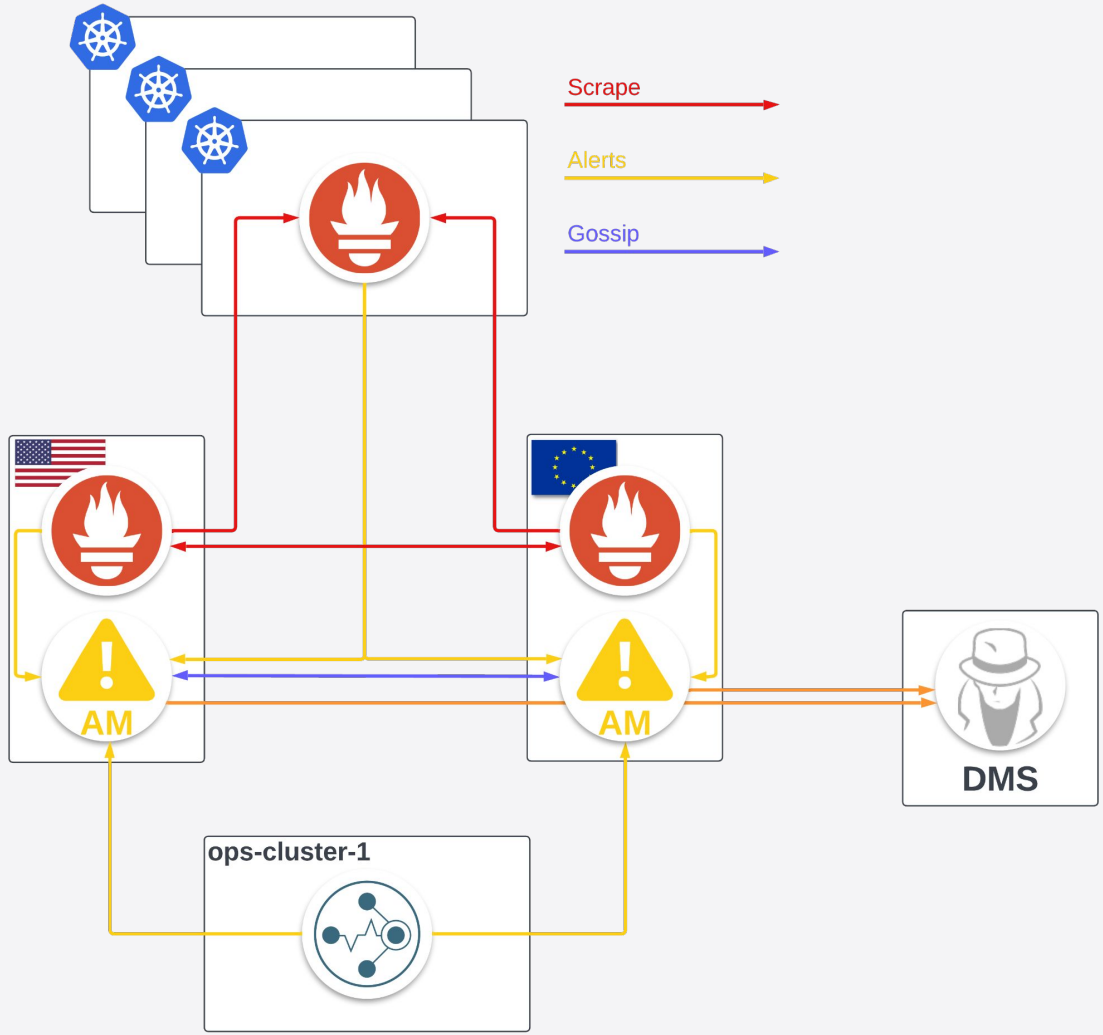
- **Deployed in two continents**
- **Scrapes all Prometheus endpoints**
- Remote-write to central Cortex
- Alert evaluation

Alertmanager pairs

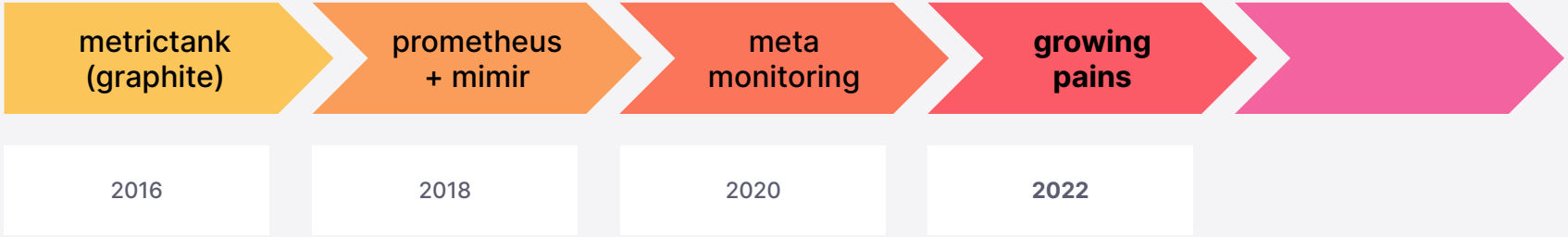
- **Deployed in two continents**
- Alert deduplication
- Forward alerts to receivers
- Silence alerts
- **Replaced all per-cluster pairs**


Scrape
Alerts
Gossip
AM
AM
ops-cluster-1

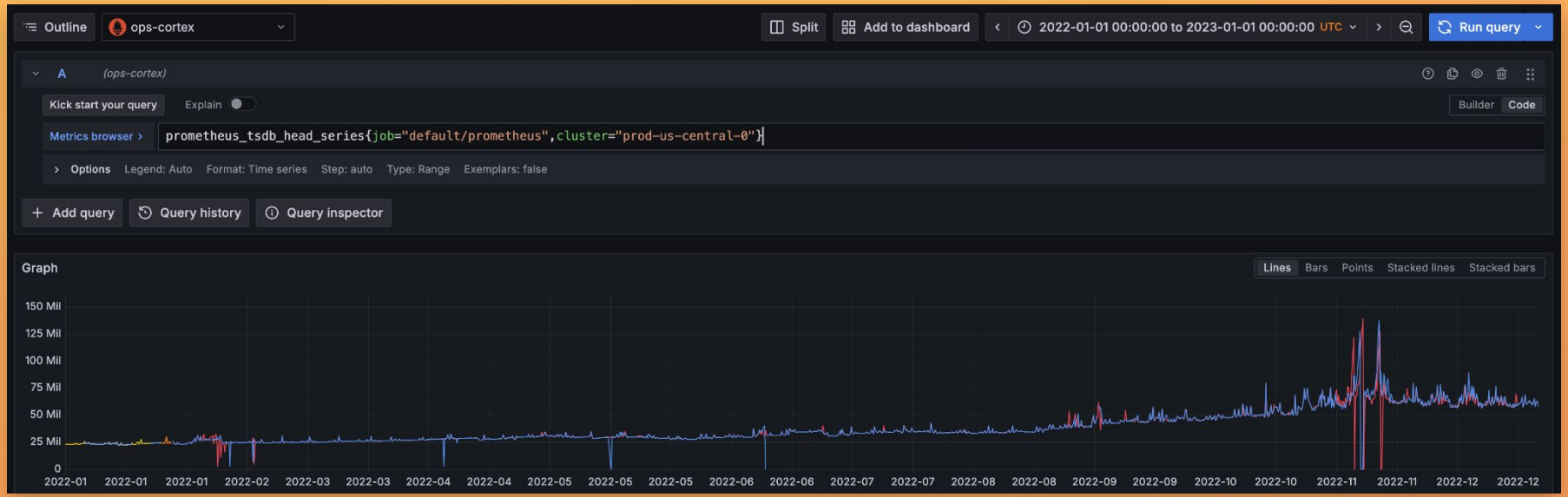

Scrape
Alerts
Gossip
AM
AM
ops-cluster-1


Scrape
Alerts
Gossip
AM
AM
DMS
ops-cluster-1

History of monitoring at Grafana



Time series growth over one year



November 16-26, 2022

Scaling Prometheus

Clusters scaled from 10s to 100s to 1000 nodes; metrics ingestion increased.

Basic levers:

- Reduce storage retention
- Increase scrape interval
- Drop unused metrics
- Deploy more Prometheus! (e.g. shard by subset of namespaces)
- Use `weaveworks/watch` for live reloads

Scaling Prometheus

Advanced lever: **GOGC**

- Controls aggressiveness of the Go garbage collector
- Default value of 100
  - Garbage collection triggered once heap has grown by 100% since the previous collection
- We **reduced GOGC value** to **40**
  - Reduces memory usage but increases CPU

Scaling Prometheus

Advanced lever: **Hashmod relabelling**

- Scale Prometheus horizontally by sharding the scraped targets

Hash on the instance label to 1..8

```
- action: hashmod
  source_labels: [instance]
  modulus: 8
  target_label: __tmp_hashmod
```

Keep series with label 5

```
- action: keep
  source_labels: [__tmp_hashmod]
  regex: 5
```

Conversation 11 | Commits 1 | Checks 0 | Files changed 2

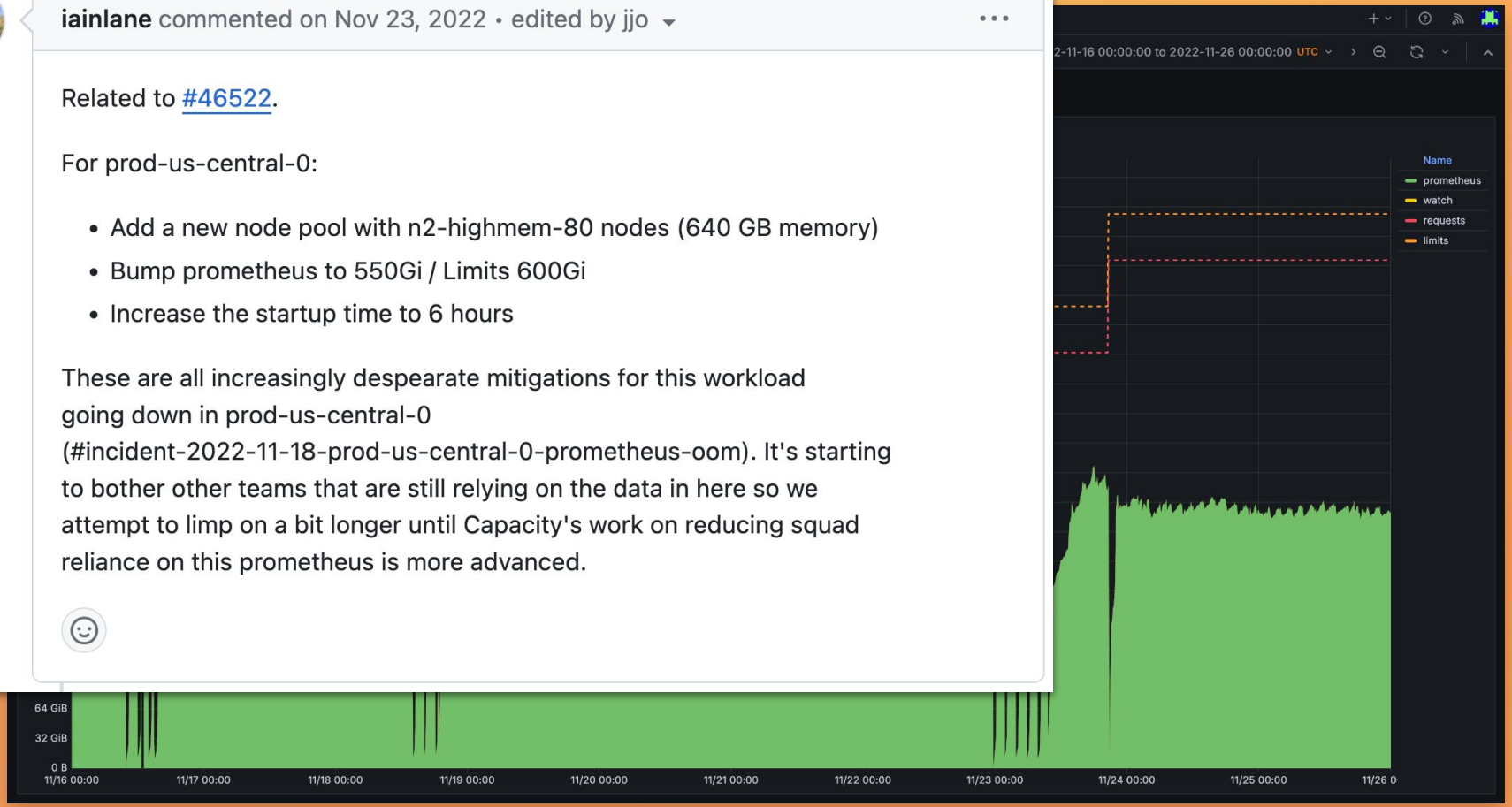
iainlane commented on Nov 23, 2022 • edited by jjo

Related to #46522.

For prod-us-central-0:

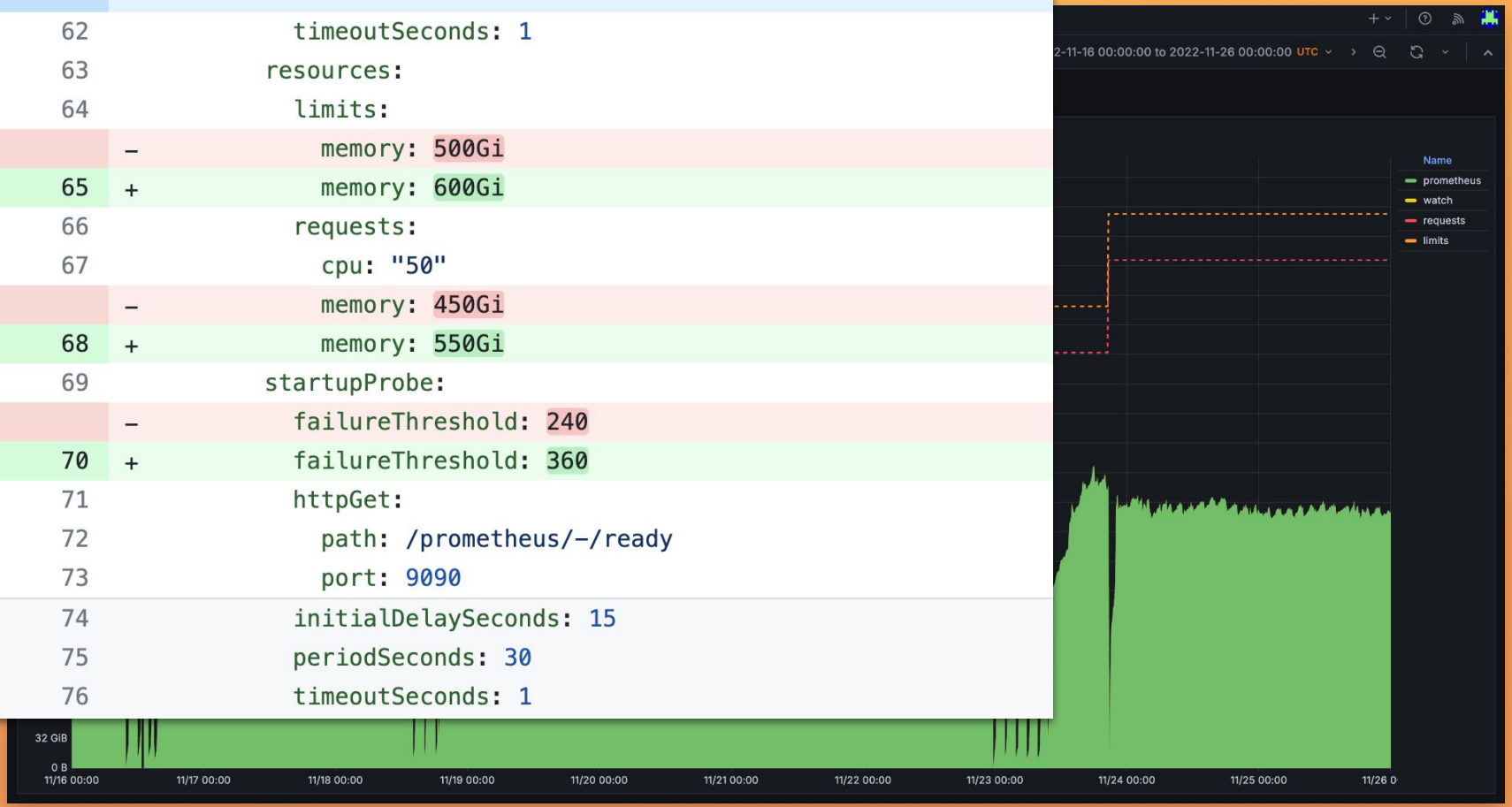
- Add a new node pool with n2-highmem-80 nodes (640 GB memory)
- Bump prometheus to 550Gi / Limits 600Gi
- Increase the startup time to 6 hours

These are all increasingly despearate mitigations for this workload going down in prod-us-central-0 (#incident-2022-11-18-prod-us-central-0-prometheus-oom). It's starting to bother other teams that are still relying on the data in here so we attempt to limp on a bit longer until Capacity's work on reducing squad reliance on this prometheus is more advanced.



6 flux/prod-us-central-0/default/StatefulSet-prometheus.yaml

```
@@ -62,12 +62,12 @@ spec:
62 62           timeoutSeconds: 1
63 63         resources:
64 64           limits:
65    -           memory: 500Gi
   65 +           memory: 600Gi
66 66           requests:
67 67             cpu: "50"
68    -           memory: 450Gi
   68 +           memory: 550Gi
69 69         startupProbe:
70    -           failureThreshold: 240
   70 +           failureThreshold: 360
71 71           httpGet:
72 72             path: /prometheus/-/ready
73 73             port: 9090
74 74           initialDelaySeconds: 15
75 75           periodSeconds: 30
76 76           timeoutSeconds: 1
```



prometheus/prod-us-central-0: More memory, more startup tim #49337

Merged by grafanabot master ← ~~iainlane/bench-prometheus-prod-us-central-0~~ on Nov 23, 2022

Conversation 11 · Commits 1 · Checks 0 · Files changed 2

iainlane commented on Nov 23, 2022 • edited by jjo

Related to #46522.

For prod-us-central-0:

- Add a new node pool with n2-highmem-80 nod
- Bump prometheus to 550Gi / Limits 600Gi
- Increase the startup time to 6 hours

These are all increasingly despeparate mitigations fo
going down in prod-us-central-0
It's starting
to bother ot
attempt to li
reliance on t

prod-us-central-0/prometheus: yet more resources #56326

Merged by master ← ~~yet-another-default-prometheus-resources-increase~~ on Feb 2, 2023

chore(prom): bump ops-us-east-0 pvc size #53640

Merged master ← ~~/plat-o11y/bump-ops-us-east-0-prom-pvc-size~~ on Jan 11, 2023

platform-o11y/prod-us-central-0: bump startup timeout #55754

Merged master ← ~~/platform-o11y/prod-us-central-0-increase-startup-probe-timeout~~ on Jan 30, 2023

chore: adjust prod-us-east-0 default/prometheus resources #52338

Merged master ← ~~adjust-prod-us-east-0-prometheus-resources~~ on Dec 22, 2022

Increase query.max-concurrency for prod-us-central-0 prom #52495

Merged master ← ~~/prom_prod_us_central_0_query_concurency~~ on Dec 26, 2022

Increase CPU request and GOMAXPROCS for pr #52380

Merged master ← ~~/tune_prom_prod_us_central_0~~ on Dec 23, 2022

platform-o11y/prod-us-central-0: increase PVC

Merged master ← ~~/platform-o11y/bump-prod-us-central-0-pvc~~ on Jan 30, 2023

increase prom resources in prod-eu-west-0 #563

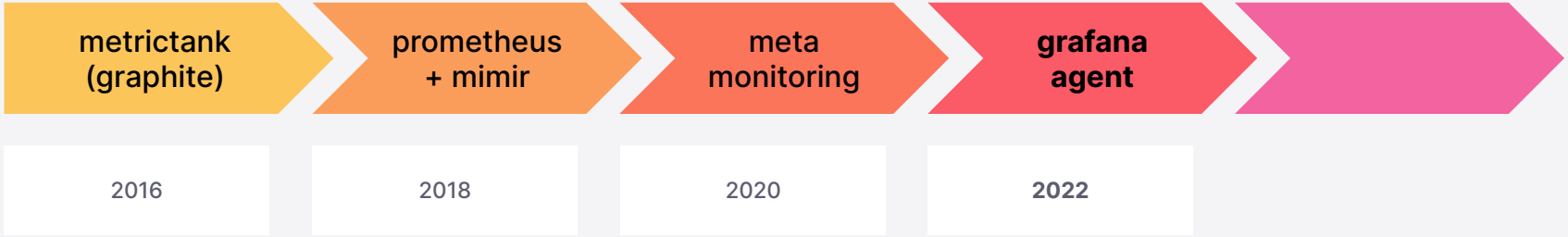
Merged by grafanabot master ← ~~/increase-euwest0-resources~~ on Feb 2, 2023

Conversation 10 · Commits 2 · Checks 3 · Files changed 2

commented on Feb 2, 2023

prometheus needs more resources in eu-west-0

History of monitoring at Grafana



Grafana Agent

default/prometheus pair

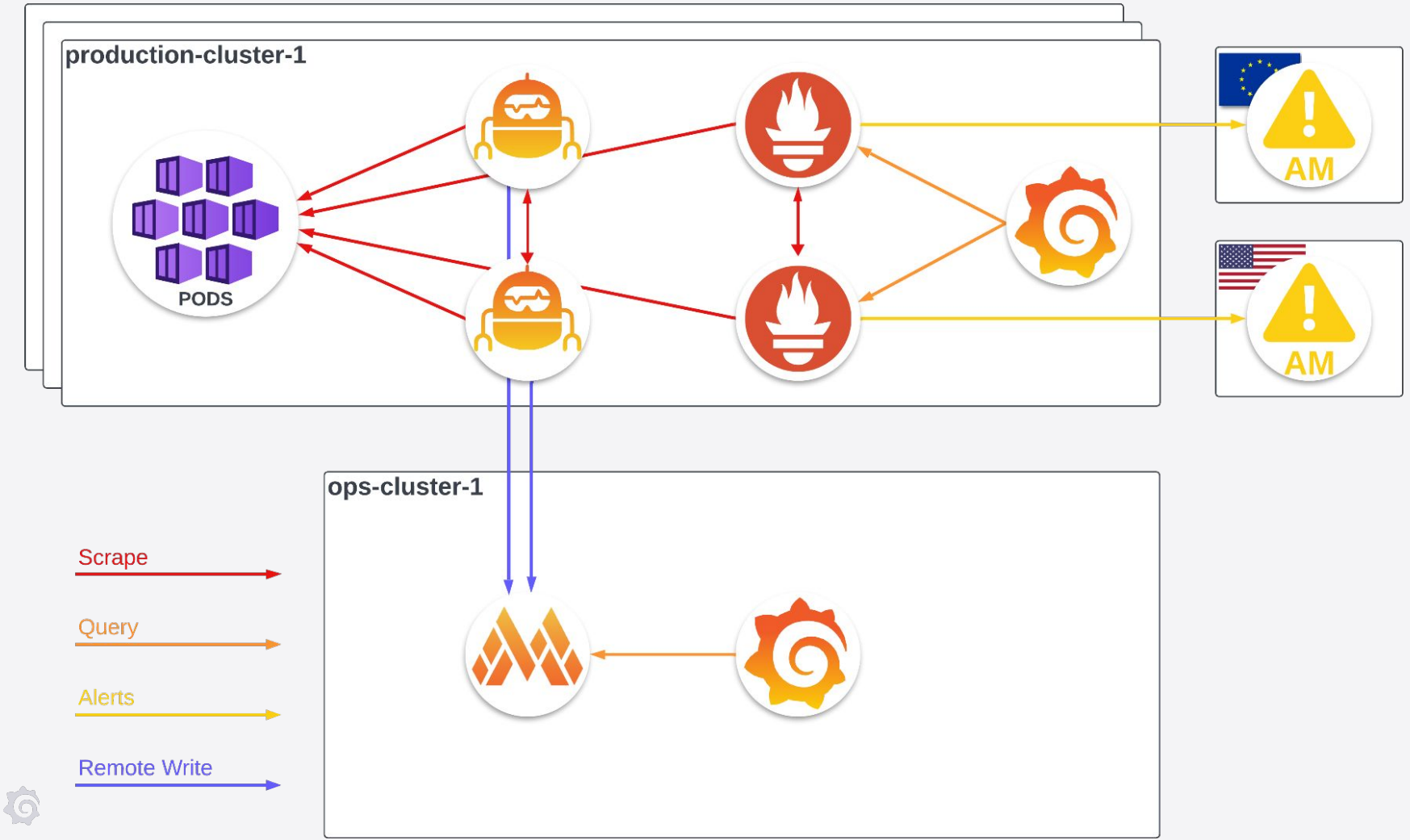
- Deployed in every cluster
- Scrapes all pod and cluster metrics
- ~~Remote-write to central Mimir~~
- Alert evaluation

grafana-agents

- Deployed in every cluster
- Scrapes all pod and cluster metrics
- **Remote-write to central Mimir**
- **No data retention**

mimir

- Queries


production-cluster-1
PODS
AM
AM
ops-cluster-1
Scrape
Query
Alerts
Remote Write

Alerting

default/prometheus pair

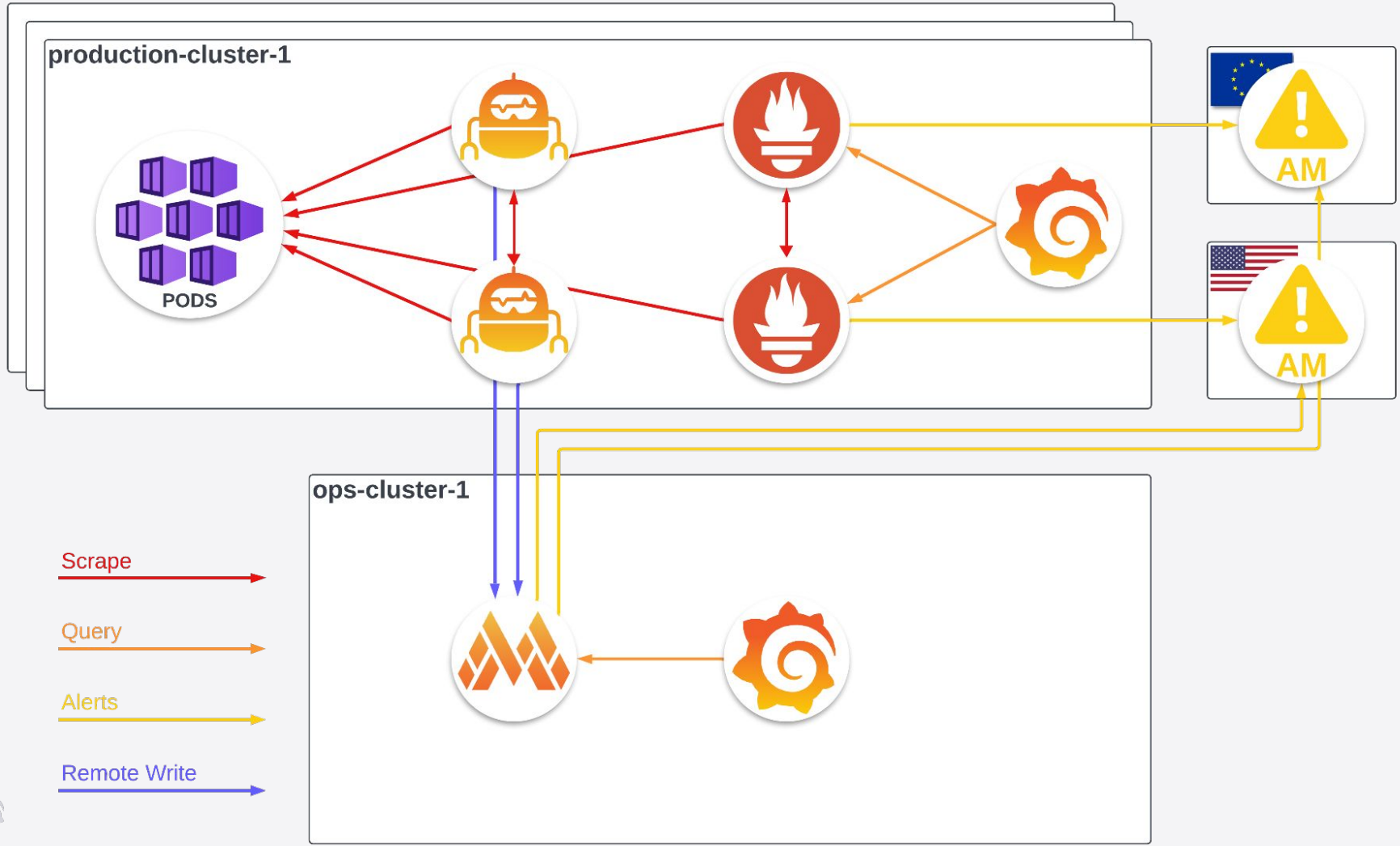
- Deployed in every cluster
- Scrapes all pod and cluster metrics
- Alert evaluation

grafana-agents

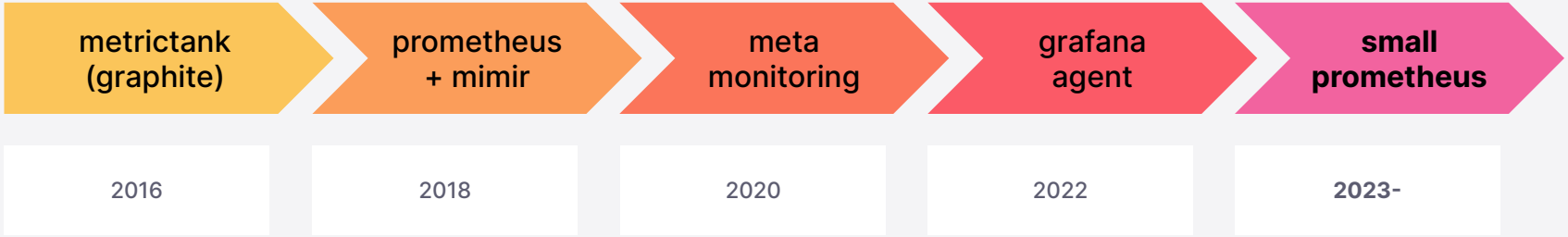
- Deployed in every cluster
- Scrapes all pod and cluster metrics
- Remote-write to central Mimir
- No data retention

mimir

- Queries
- **Alert evaluation**


production-cluster-1
PODS
AM
AM
ops-cluster-1
Scrape
Query
Alerts
Remote Write

Internal monitoring timeline



Small Prometheus

mimir-prometheus pair

- Scrapes **internal Mimir metrics**
- Owned by Mimir product team

critical-prometheus pair

- Deployed in every cluster
- Scrapes **KEDA autoscaling metrics**
- Owned by Platform team

remove default prometheus from all clusters #67875

Merged `master` on May 3, 2023

Conversation 9 | Commits 5 | Checks 5 | Files changed 10

commented on May 3, 2023 • edited

ref: #65266

This PR realizes the shutdown of default/Prometheus in prod.

Thereby we have no more default/Prometheus in our fleet.

Closing thoughts

Prometheus alert annotations

- Dashboard URL
- Runbook URL

[FIRING:1] pop-dev-aws-oregon-0: SyntheticMonitoringSuccessRateByClusterLow ()
Firing alerts:
- Synthetic Monitoring check success rate has dropped below the 95% of the median of the last 7 days for over 10 minutes.

cluster: pop-dev-aws-oregon-0
provider:

Runbook (internal) | Runbook | Source | Silence | Dashboard

Grafana Labs

Thank you