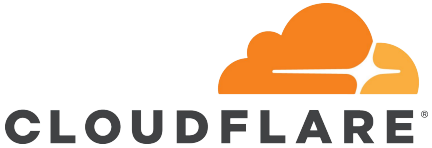


Monitoring Cloudflare's planet-scale edge network with Prometheus

Matt Bostock

@mattbostock
Platform Operations

Prometheus for monitoring

- Alerting on critical production issues
- Incident response
- Post-mortem analysis
- Metrics, but not long-term storage

CLOUDFLARE

What does Cloudflare do?

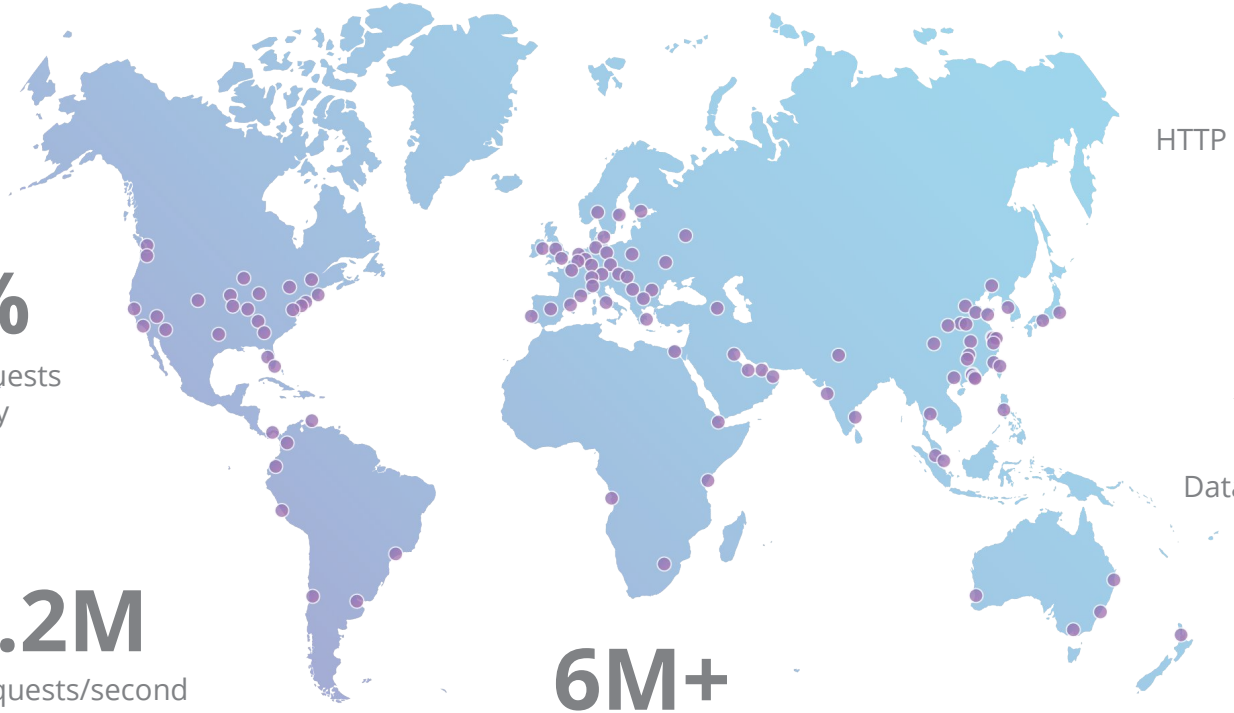
CDN
Moving content physically closer to visitors with our CDN.

Website Optimization
Caching
TLS 1.3
HTTP/2
Server push
AMP
Origin load-balancing
Smart routing

DNS
Cloudflare is one of the fastest managed DNS providers in the world.

CLOUDFLARE

Cloudflare’s anycast edge network



5M
HTTP requests/second

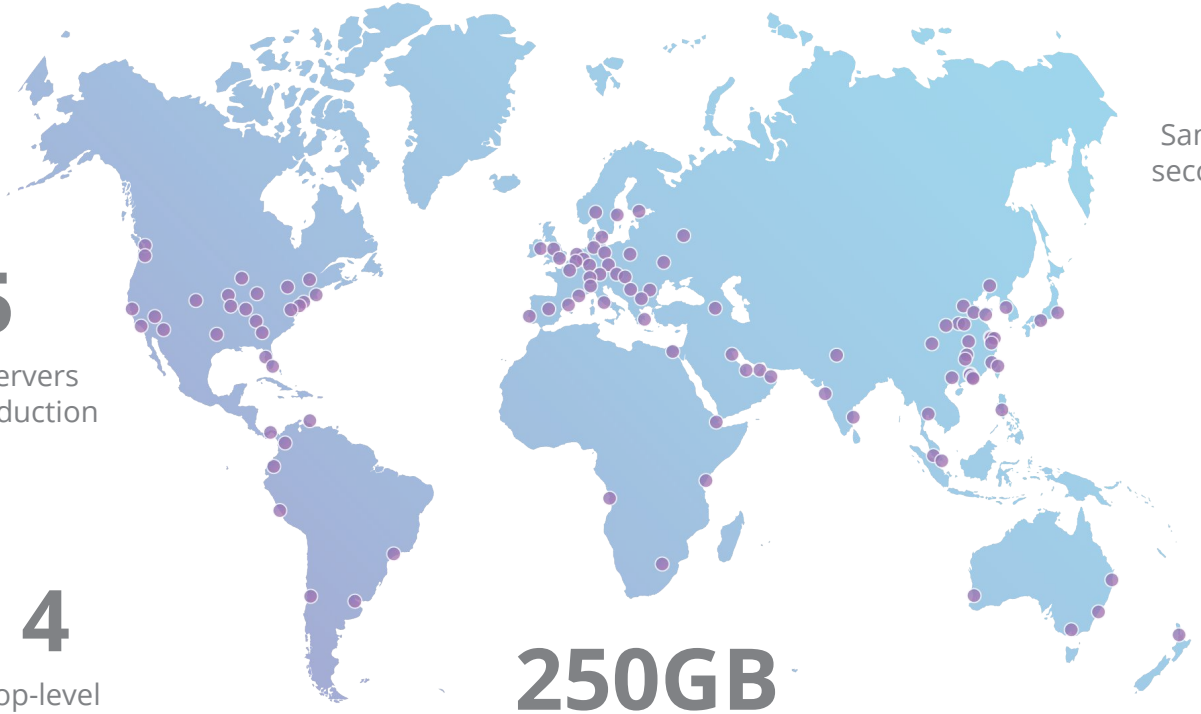
10%
Internet requests every day

115+
Data centers globally

1.2M
DNS requests/second

6M+
websites, apps & APIs in 150 countries

Cloudflare's Prometheus deployment



185
Prometheus servers currently in Production

4
Top-level Prometheus servers

250GB
Max size of data on disk

72k
Samples ingested per second max per server

4.6M
Time-series max per server

Edge Points of Presence (PoPs)

- Routing via anycast
- Configured identically
- Independent

Services in each PoP

- HTTP
- DNS
- Replicated key-value store
- Attack mitigation

Core data centers

- Enterprise log share (HTTP access logs for Enterprise customers)
- Customer analytics
- Logging: auditd, HTTP errors, DNS errors, syslog
- Application and operational metrics
- Internal and customer-facing APIs

Services in core data centers

- PaaS: Marathon, Mesos, Chronos, Docker, Sentry
- Object storage: Ceph
- Data streams: Kafka, Flink, Spark
- Analytics: ClickHouse (OLAP), CitusDB (shared PostgreSQL)
- Hadoop: HDFS, HBase, OpenTSDB
- Logging: Elasticsearch, Kibana
- Config management: Salt
- Misc: MySQL

Prometheus queries

```
node_md_disks_active / node_md_disks * 100
```

```
count(count(node_uname_info) by (release))
```

```
rate(node_disk_read_time_ms[2m]) /
rate(node_disk_reads_completed[2m])
```

Metrics for alerting

```
sum(rate(http_requests_total{job="alertmanager", code=~"5.."}[2m])) /
    sum(rate(http_requests_total{job="alertmanager"}[2m]))
                                * 100 > 0
```

CLOUDFLARE

```
count(
  abs(

(hbase_namenode_FSNamesystemState_CapacityUsed /
hbase_namenode_FSNamesystemState_CapacityTotal)

  - ON() GROUP_RIGHT()

(hadoop_datanode_fs_DfsUsed / hadoop_datanode_fs_Capacity)

  ) * 100
    > 10
      )
```

CLOUDFLARE

Prometheus architecture

Before, we used Nagios

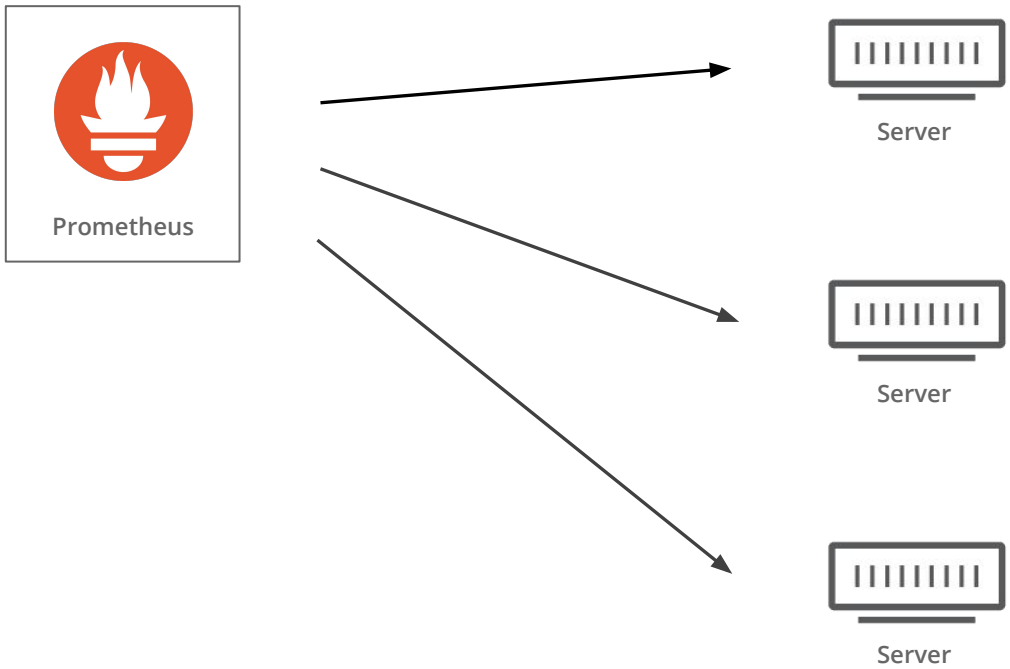
- Tuned for high volume of checks
- Hundreds of thousands of checks
- One machine in one central location
- Alerting backend for our custom metrics pipeline

Specification

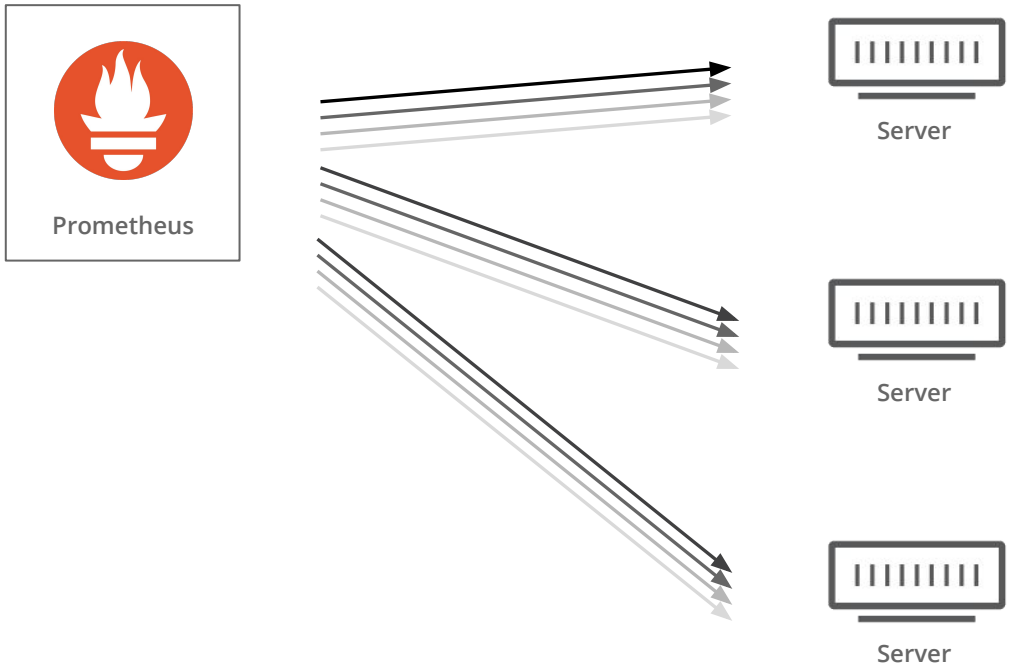
Comments

CLOUDFLARE

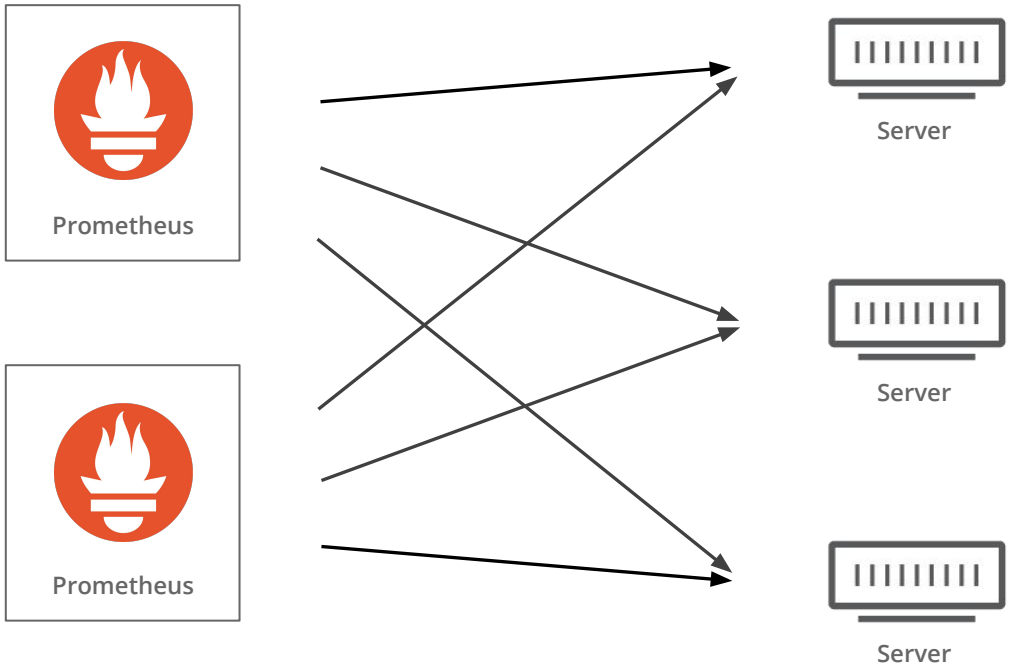
Inside each PoP



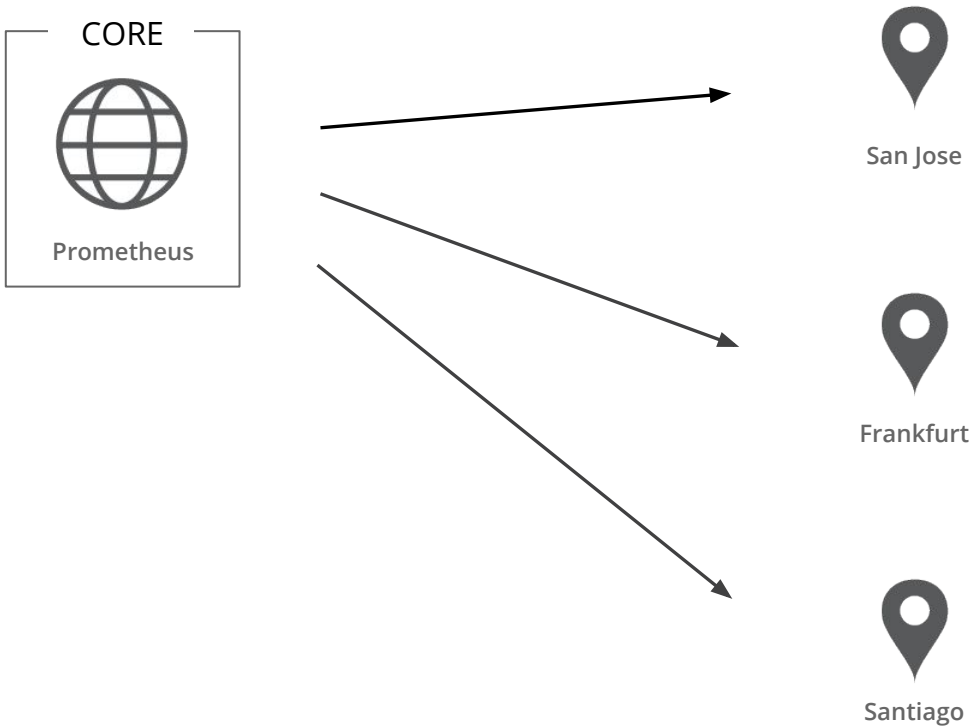
Inside each PoP



Inside each PoP: High availability



Federation



Federation configuration

```
- job_name: 'federate'
  scheme: https
  scrape_interval: 30s
  honor_labels: true
  metrics_path: '/federate'
  params:
    'match[]':
      # Scrape target health
      - '{__name__="up"}'

      # Colo-level aggregate metrics
      - '{__name__=~"colo(?:_.+)?:.+"}'
```

CLOUDFLARE

Federation configuration

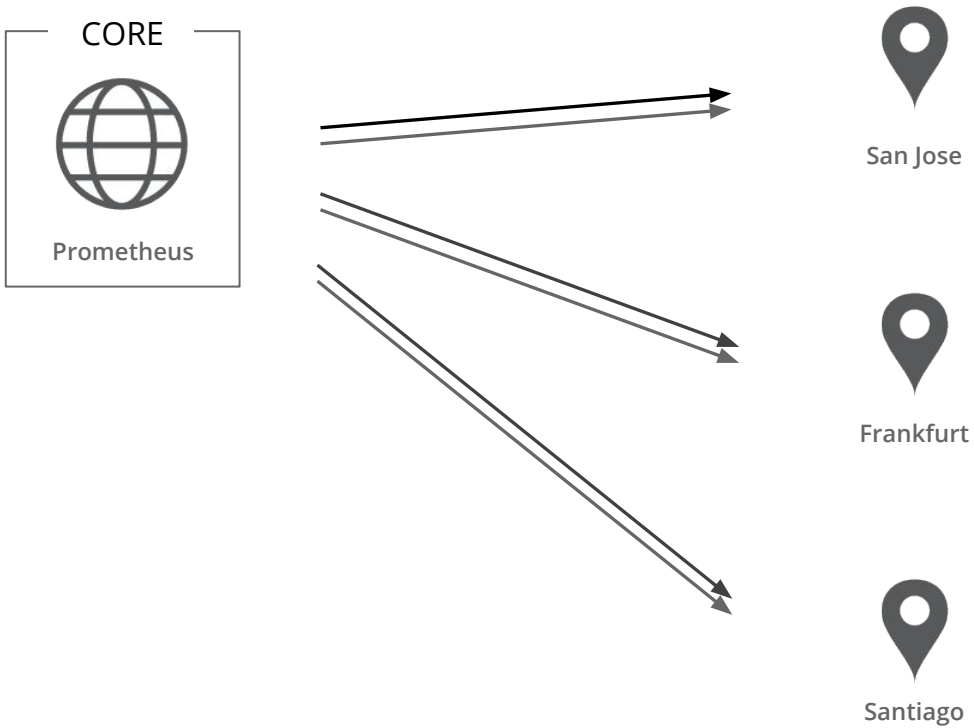
```
- job_name: 'federate'
  scheme: https
  scrape_interval: 30s
  honor_labels: true
  metrics_path: '/federate'
  params:
    'match[]':
      # Scrape target health
      - '{__name__="up"}'

      # Colo-level aggregate metrics
      - '{__name__=~"colo(?:_.+)?:.+"}'
```

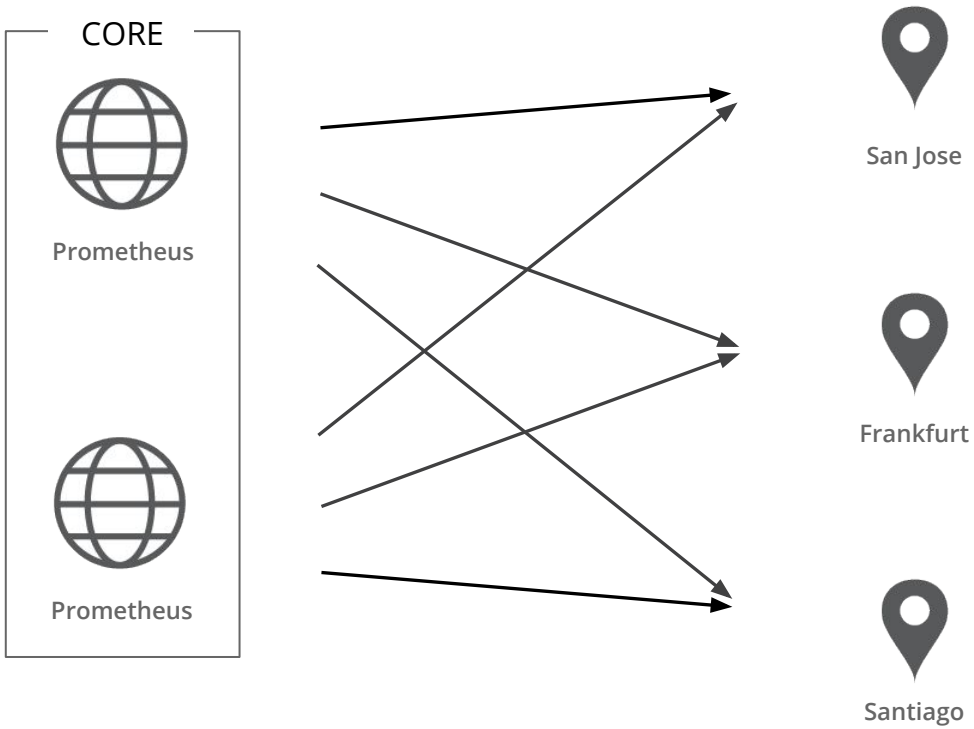
```
colo:*
colo_job:*
```

CLOUDFLARE

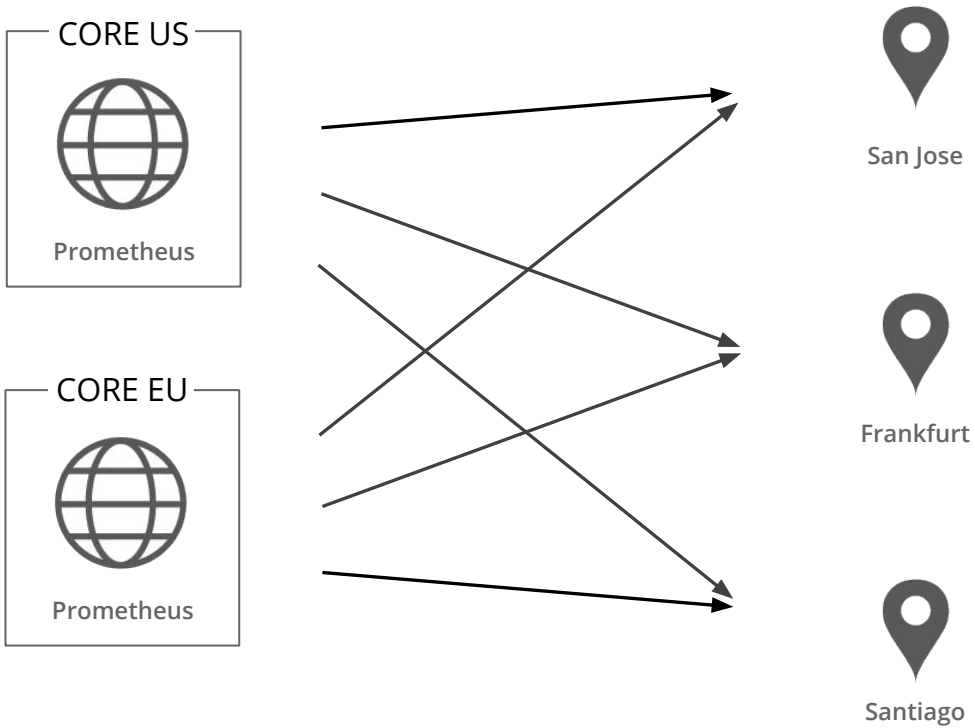
Federation



Federation: High availability



Federation: High availability



Retention and sample frequency

- 15 days’ retention
- Metrics scraped every 60 seconds
  - Federation: every 30 seconds
- No downsampling

Exporters we use

Purpose	Name
System (CPU, memory, TCP, RAID, etc)	Node exporter
Network probes (HTTP, TCP, ICMP ping)	Blackbox exporter
Log matches (hung tasks, controller errors)	mtail

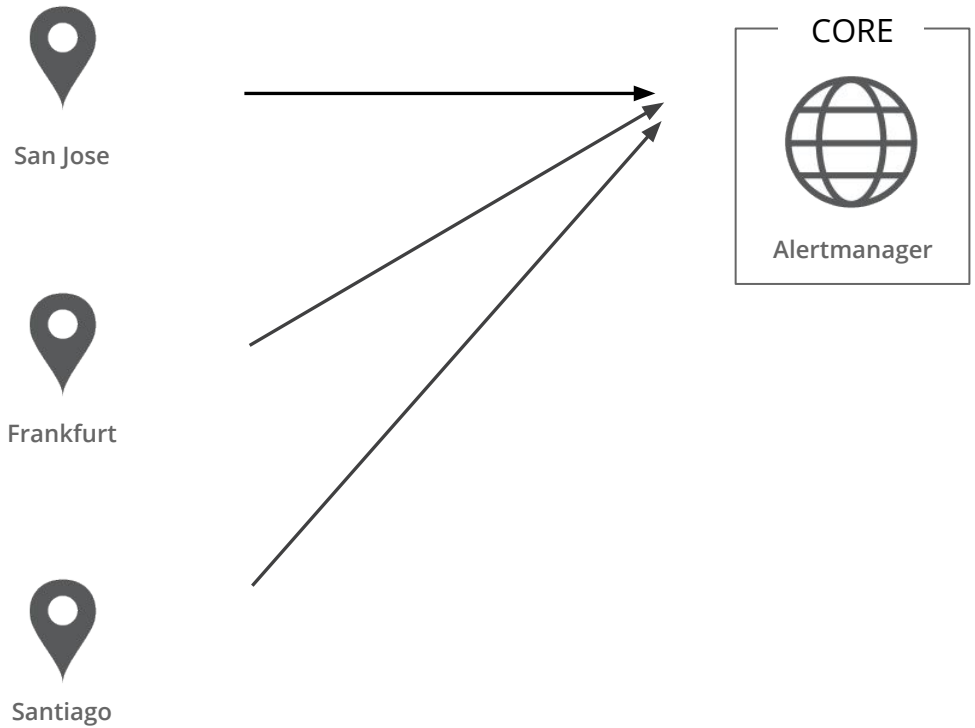
Deploying exporters

- One exporter per service instance
- Separate concerns
- Deploy in same failure domain

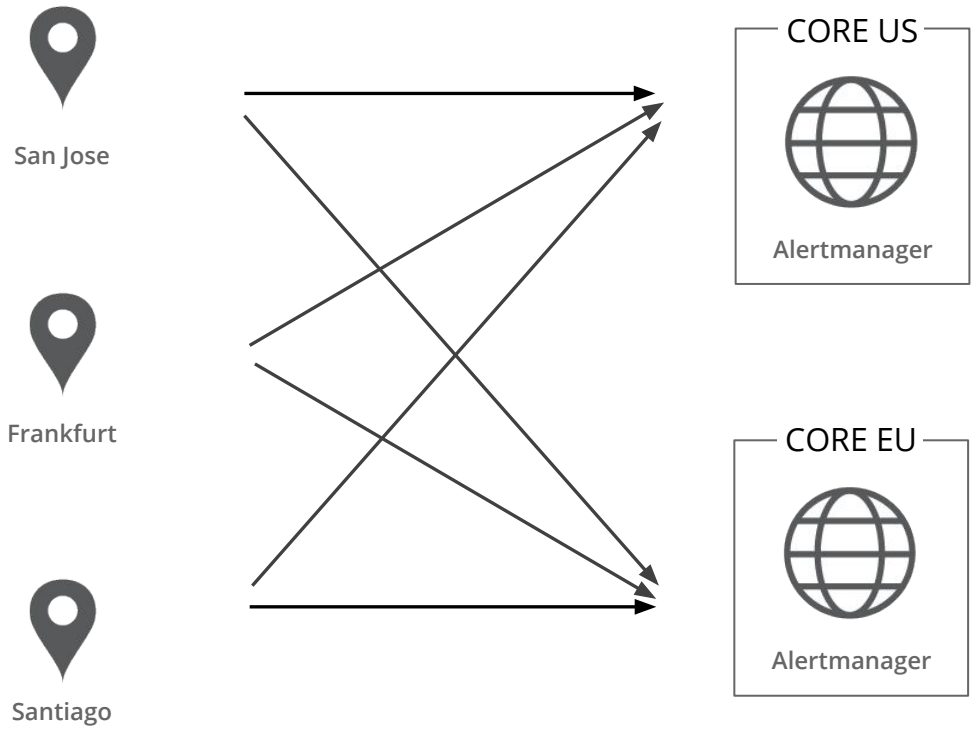
CLOUDFLARE

Alerting

Alerting



Alerting: High availability (soon)



Writing alerting rules

- Test the query on past data

Writing alerting rules

- Test the query on past data
- Descriptive name with adjective or adverb

RAID_Array

RAID_Health_Degraded

Writing alerting rules

- Test the query on past data
- Descriptive name with adjective/adverb
- Must have an alert reference

Writing alerting rules

- Test the query on past data
- Descriptive name with adjective/adverb
- Must have an alert reference
- Must be actionable

Writing alerting rules

- Test the query on past data
- Descriptive name with adjective/adverb
- Must have an alert reference
- Must be actionable
- Keep it simple

Example alerting rule

```
ALERT RAID_Health_Degraded
  IF node_md_disks - node_md_disks_active > 0
  LABELS { notify="jira-sre" }
  ANNOTATIONS {
    summary = `{{ $value }} disks in {{ $labels.device }} on {{ $labels.instance }} are faulty`,
    Dashboard = `https://grafana.internal/disk-health?var-instance={{ $labels.instance }}`,
    link = "https://wiki.internal/ALERT+Raid+Health",
  }
```

CLOUDFLARE

Monitoring your monitoring

PagerDuty escalation drill

```
ALERT SRE_Escalation_Drill
  IF (hour() % 8 == 1 and minute() >= 35) or (hour() % 8 == 2 and minute() < 20)
  LABELS { notify="escalate-sre" }
  ANNOTATIONS {
     dashboard="https://cloudflare.pagerduty.com/",
     link="https://wiki.internal/display/OPS/ALERT+Escalation+Drill",
     summary="This is a drill to test that alerts are being correctly escalated.
     Please ack the PagerDuty notification."
     }
```

Monitoring Prometheus

- Mesh: each Prometheus monitors other Prometheus servers in same datacenter
- Top-down: top-level Prometheus servers monitor datacenter-level Prometheus servers

Monitoring Alertmanager

- Use Grafana’s alerting mechanism to page
- Alert if notifications sent is zero even though notifications were received

Monitoring Alertmanager

```
(
sum(rate(alertmanager_alerts_received_total{job="alertmanager"}[5m]))
without(status, instance) > 0
and
sum(rate(alertmanager_notifications_total{job="alertmanager"}[5m]))
without(integration, instance) == 0
)
or vector(0)
```

CLOUDFLARE

Alertmanager

Alertmanager in Marathon | Alertmanager status page | Alertmanager logs

Overview

Is Alertmanager up?

UP

HTTP 5XX

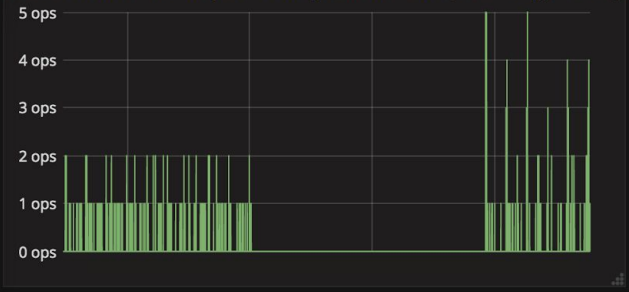
0%

Notification failures

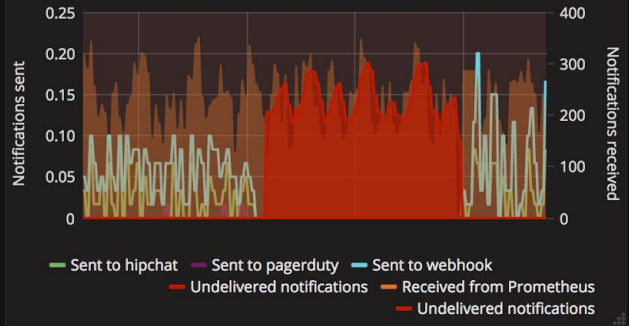
0%

Alerts

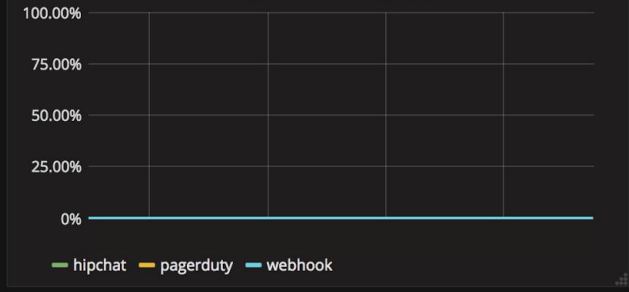
Alert notifications sent per second (in Elasticsearch; Y-axis capped at 5)



AlertManager notifications per second



AlertManager notification failure rate



HTTP

HTTP request latency (99 percentile)

	min	max	avg
del_silence	0 µs	7.13 s	446 ms
alert_groups	0 µs	221 ms	114 ms
add_alerts	0 µs	123 ms	94 ms
list_alerts	0 µs	106 ms	61 ms
alert_groups	0 µs	222 ms	38 ms
add_alerts	0 µs	119 ms	21 ms
list_alerts	0 µs	71 ms	12 ms

HTTP requests per second by handler and status code

	min	max	avg
add_alerts 200	0 ops	7 ops	5 ops
add_alerts 200	0 ops	7 ops	1 ops
prometheus 200	0 ops	0 ops	0 ops
index 200	0 ops	0 ops	0 ops
prometheus 200	0 ops	0 ops	0 ops
index 200	0 ops	0 ops	0 ops
list_silences 200	0 ops	0 ops	0 ops

Alert routing

Alert routing

```
notify="hipchat-sre escalate-sre"
```

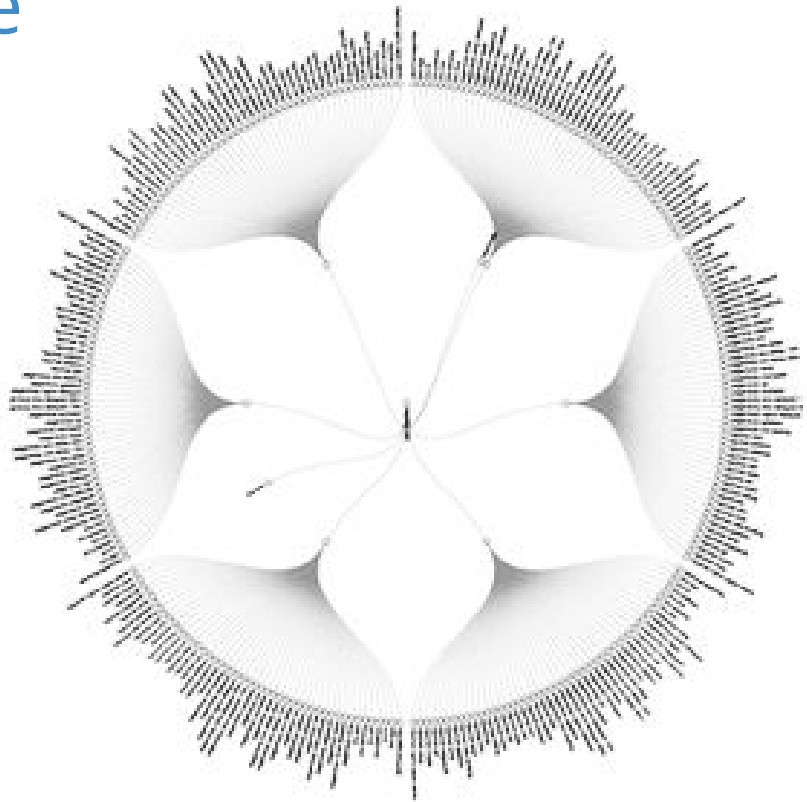
CLOUDFLARE

Alert routing

```
- match_re:
    notify: (?:.*\s+)?hipchat-sre(?:\s+.*)?
  receiver: hipchat-sre
  continue: true
```

CLOUDFLARE

Routing tree



CLOUDFLARE

fabxc / jiralerts

Watch 4 Star 22 Fork 5

Code Issues 3 Pull requests 2 Projects 0 Insights

Prometheus Alertmanager integration for JIRA

3 commits 1 branch 0 releases 1 contributor

Branch: master New pull request Find file Clone or download

fabxc Fix wrong parameter naming Latest commit c3d5c7c on Mar 2, 2016

README.md	Fix wrong parameter naming	a year ago
main.py	Add README	a year ago
requirements.txt	first commit	a year ago

README.md

jiralerts

This is a basic JIRA integration for Alertmanager. It receives Alertmanager webhook messages and files labeled issues for it. If an alert stops firing or starts firing again, tickets are closed or reopened.

Given how generic JIRA is, the integration attempts several different transitions that may be available for an issue.

Consider this an opinionated example snippet. It may not fit your use case without modification.

Running it

AlertManager - JIRA integration test, please ignore

Edit | Comment | Assign | More | Close | Start Work | Workflow | Export

Details

Type:	Chore	Status:	ACTION NEEDED (View Workflow)
Priority:	Normal	Resolution:	Unresolved
Affects Version/s:	None	Fix Version/s:	None
Component/s:	None		
Labels:	alert, alert_group_key=18252849132189650816, bot_jiralerts		

Description

Description

testhost.local:9100 of job node has been down for more than 5 minutes.

Summary

Instance testhost.local:9100 down

Active alerts (updated in realtime)

node_down: Instance testhost.local:9100 down

Description: testhost.local:9100 of job node has been down for more than 5 minutes.
Summary: Instance testhost.local:9100 down

```
alertname = node_down
env = prod
instance = testhost.local:9100
job = node
monitor = prometheus
severity = critical
```

Breakdown — Source

node_down: Instance othertesthost.local:9100 down

Dashboard: https://example.com
Description: othertesthost.local:9100 of job node has been down for more than 5 minutes.
Link: https://example.com
Summary: Instance othertesthost.local:9100 down

```
alertname = node_down
```

People

Assignee:	Matt Bostock
Reporter:	JIRA Alerts via AlertManager
Votes:	0 Vote for this issue
Watchers:	1 Start watching this issue

Dates

Created:	2 hours ago
Updated:	Just now

Zendesk

There are no tickets linked to this issue.

Development

Create branch

Agile

View on Board

HipChat discussions

Dedicated room: Create a room | Choose a room

cloudflare / alertmanager2es

Watch 11 Star 29 Fork 1

Code Issues 2 Pull requests 0 Insights

Receives HTTP webhook notifications from AlertManager and inserts them into an Elasticsearch index for searching and analysis

alerting monitoring prometheus alertmanager elasticsearch analytics

5 commits 1 branch 2 releases 2 contributors Apache-2.0

Branch: master New pull request Find file Clone or download

prymitive committed with mattbostock Update supported webhook version to 4 ... Latest commit 162379f on May 10

vendor	Send AlertManager notifications to Elasticsearch	4 months ago
.gitignore	Send AlertManager notifications to Elasticsearch	4 months ago
.travis.yml	Add .travis.yml	3 months ago
CONTRIBUTING.md	Send AlertManager notifications to Elasticsearch	4 months ago
LICENSE	Send AlertManager notifications to Elasticsearch	4 months ago
Makefile	Send AlertManager notifications to Elasticsearch	4 months ago
README.md	Make a note that this only work with Alertmanager 0.6.x	3 months ago
main.go	Update supported webhook version to 4	3 months ago
main_test.go	Update supported webhook version to 4	3 months ago

README.md

alertmanager2es

alertmanager2es receives HTTP webhook notifications from AlertManager and inserts them into an Elasticsearch

cloudflare / unsee

Watch 14 Star 107 Fork 8

Code Issues 1 Pull requests 1 Insights

Alert dashboard for Prometheus Alertmanager

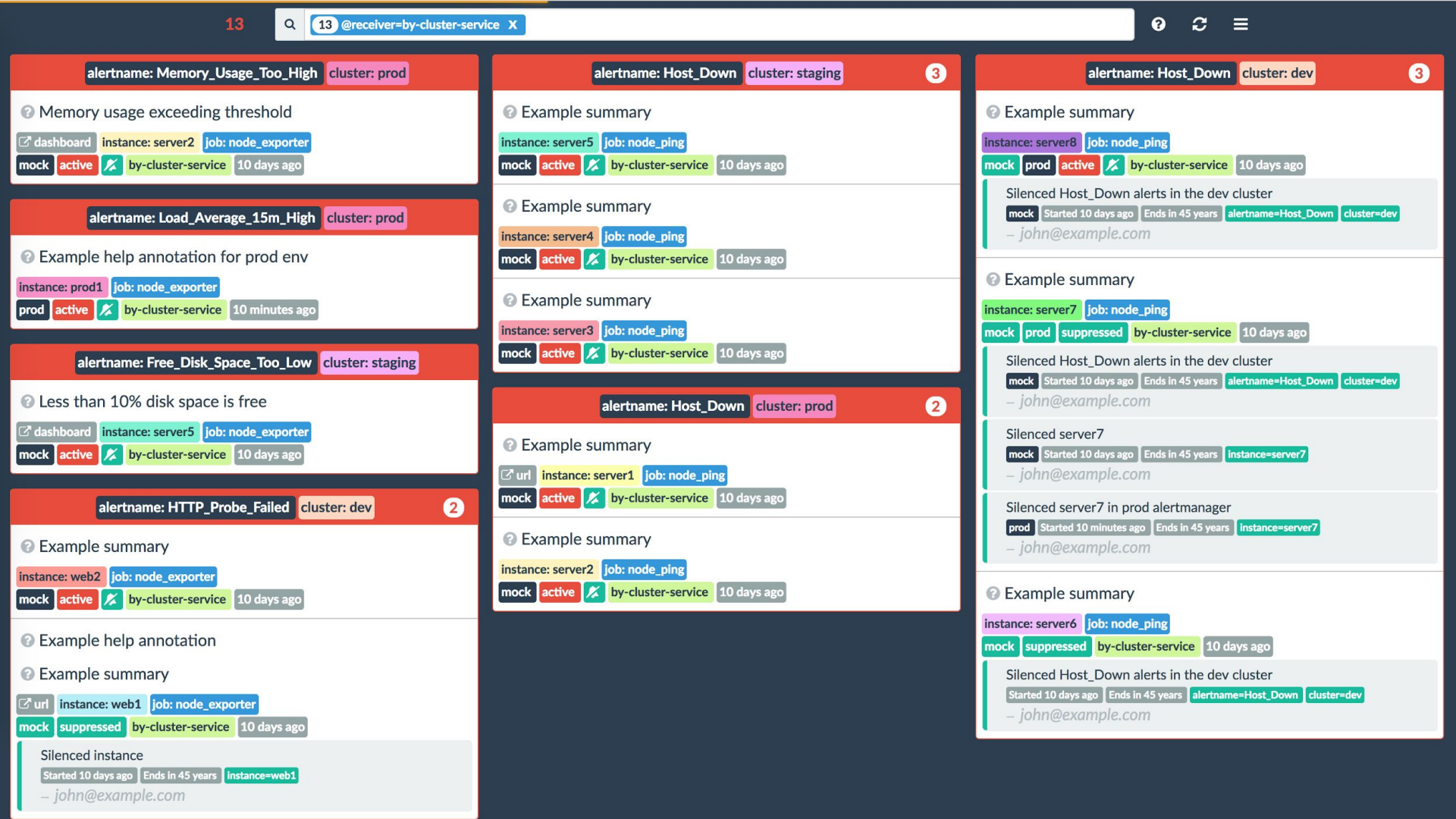
monitoring alerting dashboard prometheus alertmanager

455 commits 10 branches 11 releases 6 contributors Apache-2.0

Branch: master New pull request Find file Clone or download

prymitive committed on GitHub Merge pull request #149 from cloudflare/fix-silence-form-colors ... Latest commit 9ea3236 7 days ago

alertmanager	Fix test failing with 0.8.0 mock data	11 days ago
assets	Show correct label colors in the silence form	7 days ago
config	Vendor renamed Sirupsen/logrus to sirupsen/logrus, fix imports	a month ago
filters	Vendor renamed Sirupsen/logrus to sirupsen/logrus, fix imports	a month ago
hooks	Put the branch name first on the version string for docker master builds	3 months ago
mapper	Correctly seed alert fingerprints	26 days ago
mock	Add mock data from Alertmanager 0.8.0	11 days ago
models	Refactor group fingerpint tests	15 days ago
slices	Move all *InSlice functions into a slices package	a month ago
transform	Move all *InSlice functions into a slices package	a month ago


13
13 @receiver=by-cluster-service X
alertname: Memory_Usage_Too_High cluster: prod
Memory usage exceeding threshold
dashboard instance: server2 job: node_exporter
mock active by-cluster-service 10 days ago
alertname: Load_Average_15m_High cluster: prod
Example help annotation for prod env
instance: prod1 job: node_exporter
prod active by-cluster-service 10 minutes ago
alertname: Free_Disk_Space_Too_Low cluster: staging
Less than 10% disk space is free
dashboard instance: server5 job: node_exporter
mock active by-cluster-service 10 days ago
alertname: HTTP_Probe_Failed cluster: dev 2
Example summary
instance: web2 job: node_exporter
mock active by-cluster-service 10 days ago
Example help annotation
Example summary
url instance: web1 job: node_exporter
mock suppressed by-cluster-service 10 days ago
Silenced instance
Started 10 days ago Ends in 45 years instance=web1
– john@example.com
alertname: Host_Down cluster: staging 3
Example summary
instance: server5 job: node_ping
mock active by-cluster-service 10 days ago
Example summary
instance: server4 job: node_ping
mock active by-cluster-service 10 days ago
Example summary
instance: server3 job: node_ping
mock active by-cluster-service 10 days ago
alertname: Host_Down cluster: prod 2
Example summary
url instance: server1 job: node_ping
mock active by-cluster-service 10 days ago
Example summary
instance: server2 job: node_ping
mock active by-cluster-service 10 days ago
alertname: Host_Down cluster: dev 3
Example summary
instance: server8 job: node_ping
mock prod active by-cluster-service 10 days ago
Silenced Host_Down alerts in the dev cluster
mock Started 10 days ago Ends in 45 years alertname=Host_Down cluster=dev
– john@example.com
Example summary
instance: server7 job: node_ping
mock prod suppressed by-cluster-service 10 days ago
Silenced Host_Down alerts in the dev cluster
mock Started 10 days ago Ends in 45 years alertname=Host_Down cluster=dev
– john@example.com
Silenced server7
mock Started 10 days ago Ends in 45 years instance=server7
– john@example.com
Silenced server7 in prod alertmanager
prod Started 10 minutes ago Ends in 45 years instance=server7
– john@example.com
Example summary
instance: server6 job: node_ping
mock suppressed by-cluster-service 10 days ago
Silenced Host_Down alerts in the dev cluster
Started 10 days ago Ends in 45 years alertname=Host_Down cluster=dev
– john@example.com

amtool

```
matt➜~» go get -u github.com/prometheus/alertmanager/cmd/amtool

matt➜~» amtool silence add \

     --expire 4h \

     --comment https://jira.internal/TICKET-1234 \

     alertname=HDFS_Capacity_Almost_Exhausted
```

CLOUDFLARE

Pain points

Storage pressure

- Use -storage.local.target-heap-size
- Set -storage.local.series-file-shrink-ratio to 0.3 or above

Alertmanager races, deadlocks, timeouts, oh my

Cardinality explosion

```
mbostock@host:~$ sudo cp /data/prometheus/data/heads.db ~
mbostock@host:~$ sudo chown mbostock: ~/heads.db
mbostock@host:~$ storagetool dump-heads heads.db | awk '{ print $2 }' | sed 's/{.*//' |
sed 's/METRIC=//' | sort | uniq -c | sort -n
...snip...
 678869 eyom_eyomCPTOPON_numsub
 678876 eyom_eyomCPTOPON_hhiinv
 679193 eyom_eyomCPTOPON_hhi
2314366 eyom_eyomCPTOPON_rank
2314988 eyom_eyomCPTOPON_speed
2993974 eyom_eyomCPTOPON_share
```

CLOUDFLARE

Standardise on metric labels early

- Especially probes: source versus target
- Identifying environments
- Identifying clusters
- Identifying deployments of same app in different roles

Next steps

Prometheus 2.0

- Lower disk I/O and memory requirements
- Better handling of metrics churn

Integration with long term storage

- Ship metrics from Prometheus (remote write)
- One query language: PromQL

More improvements

- Federate one set of metrics per datacenter
- Highly-available Alertmanager
- Visual similarity search
- Alert menus; loading alerting rules dynamically
- Priority-based alert routing

More information

blog.cloudflare.com

github.com/cloudflare

Try Prometheus 2.0: prometheus.io/blog

Questions? @mattbostock

CLOUDFLARE

Thanks!

blog.cloudflare.com

github.com/cloudflare

Try Prometheus 2.0: prometheus.io/blog

Questions? @mattbostock

CLOUDFLARE