Please

Ask questions through the app

Rate Session

Thank you!

@GOTOber

gotober.com

Four Years in – How Prometheus Revolutionized Monitoring at SoundCloud

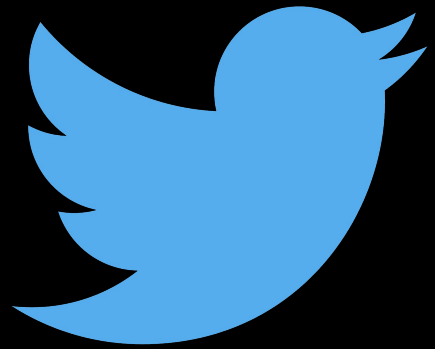
GOTO Berlin – 2017-11-16

Björn “Beorn” Rabenstein, Production Engineer, SoundCloud Ltd.

SOUNDCLOUD

Level 0

No monitoring.



Community impact

Source	Amount
Emails	4
Tweets	3
Facebook posts	
Help Community	1 post (96 views)
Total	8 direct (visibility to over 100 users)

Level 1

Nagios.

Current Network Status
Last Updated: Fri Nov 10 11:24:13 GMT 2017 - Update is PAUSED [continue]
Icinga 1.8.4 - Logged in as *admin*

Commands for checked host(s)

Select command

Submit

- View **Service Status Detail** For **All Hosts**
- View **Host Status Detail** For **All Hosts**

Host Status Details For Host/Services matching '10.33.84.91'

Page 1 of 1 Results: 50

Host	Status	Last Check	Duration	Attempt	Status Information
ip-10-42-84-91	UP	2017-11-10 11:24:01	66d 17h 45m 39s	1/3	OK - 10.42.84.91: rta 0.098ms, lost 0%

Displaying Result 1 - 1 of 1 Matching Hosts

Commands for checked services

Select command

Submit

Service Status Details For Host/Services matching '10.33.84.91'

Host	Service	Status	Last Check	Duration	Attempt	Status Information
ip-10-42-84-91	NRPE available	OK	2017-11-10 11:21:39	66d 21h 49m 50s	1/3	NRPE checks working
	check-statsd-local	OK	2017-11-10 11:19:59	142d 17h 23m 48s	1/3	OK: run: statsd: (pid 3786) 5780956s: run: log: (pid 3778) 5780956s
	nfs_mounts	OK	2017-11-10 11:20:42	388d 20h 25m 30s	1/3	NFS OK: 1 mount points avg of 0.00161 secs, max 0.00161 secs.

Page 1 of 1 Results: 50

Displaying Result 1 - 3 of 3 Matching Services

Current Network Status
Last Updated: Fri Nov 10 11:28:51 GMT 2017 - Update in 13 seconds [pause]
Icinga 1.8.4 - Logged in as *admin*

Commands for checked host(s)

Select command

Submit

- View **Service Status Detail** For **All Hosts**
- View **Host Status Detail** For **All Hosts**

Host Status Details For Host/Services matching 'go-link-redirector'

Page 1 of 1 Results: 50

Host	Status	Last Check	Duration	Attempt	Status Information
go-link-redirector	PENDING	N/A	939d 21h 41m 38s	1/3	Host check scheduled for Fri Nov 10 11:29:33 GMT 2017

Displaying Result 1 - 1 of 1 Matching Hosts

Commands for checked services

Select command

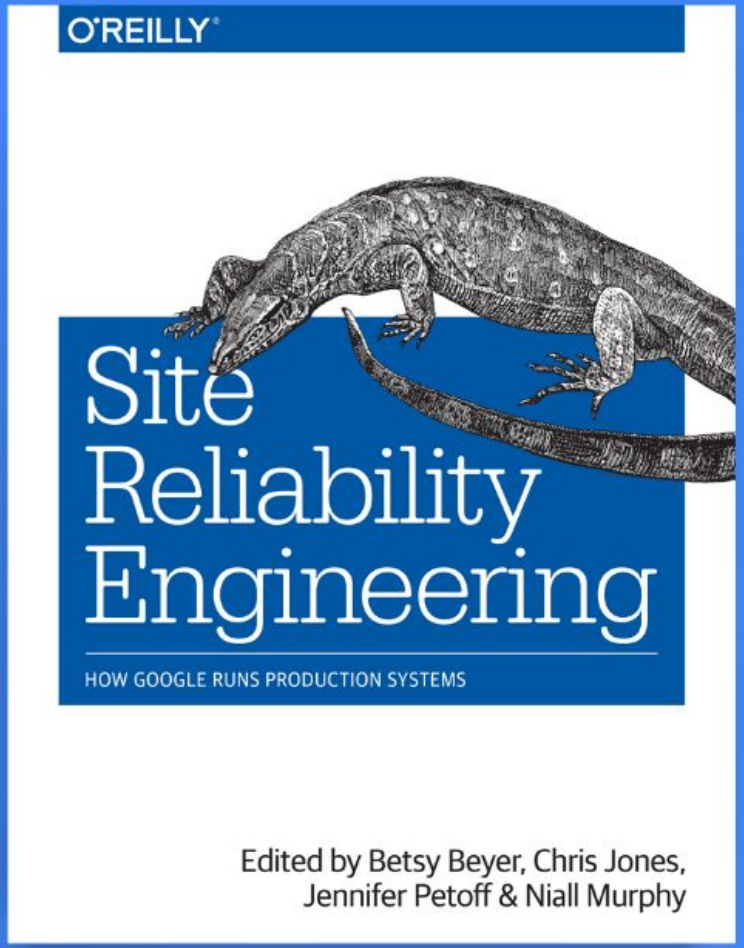
Submit

Service Status Details For Host/Services matching 'go-link-redirector'

Host	Service	Status	Last Check	Duration	Attempt	Status Information
go-link-redirector	health	OK	2017-11-10 11:27:45	226d 20h 15m 21s	1/3	HTTP OK: Status line output matched "200" - 1870 bytes in 0.019 second response time

Page 1 of 1 Results: 50

Displaying Result 1 - 1 of 1 Matching Services


O'REILLY
Site Reliability Engineering
HOW GOOGLE RUNS PRODUCTION SYSTEMS
Edited by Betsy Beyer, Chris Jones, Jennifer Petoff & Niall Murphy

Symptom-based alerting

Symptoms → pages. Causes → tickets.

Black-box monitoring FTW?

We combine *heavy use of white-box* monitoring with *modest but critical uses of black-box* monitoring. The simplest way to think about black-box monitoring versus white-box monitoring is that black-box monitoring is *symptom-oriented and represents active—not predicted—problems.* [...]

For paging, black-box monitoring has the key benefit of forcing discipline to only nag a human when a problem is both *already ongoing and contributing to real symptoms*. On the other hand, for *not-yet-occurring but imminent* problems, black-box monitoring is *fairly useless*.

Chapter 6: Monitoring Distributed Systems

Why black-box monitoring is not sufficient:

1. Your probe is not real user traffic.
2. Long-tail latency matters.
3. You still need to investigate causes.

Critical: Page Failure

Availability Test for API Mobile Track [107111]
https://api-mobile.soundcloud.id=4711424248273487234

TYPE	Web	REPORT	11/07/2017 09:30:00
MONITOR	Object	ALERT	11/07/2017 09:34:22
RUN THRESHOLD	1		
TIME THRESHOLD	default		

Performance Scatterplot Waterfall

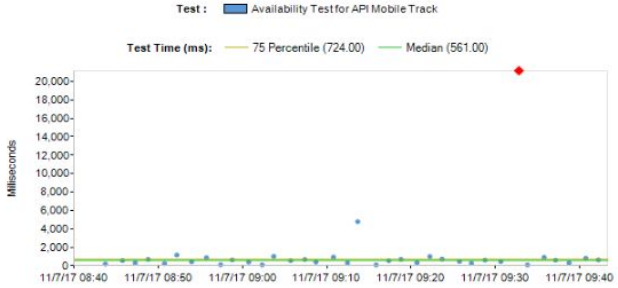
Alert Details

NODE	CAUSE	HOST IP
Dallas - Level3 4.30.68.66	Unknown	52.85.210.186

0 x 1 [2]

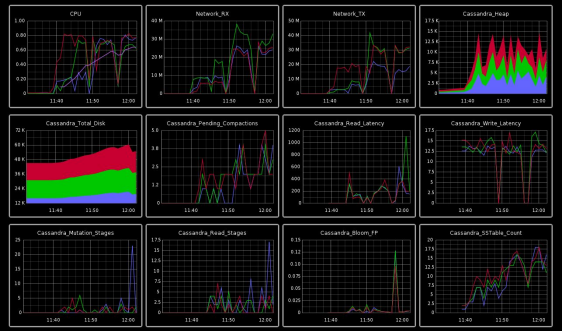
Triggered on 1 of 3 runs, expected 0 more runs in the alert interval.

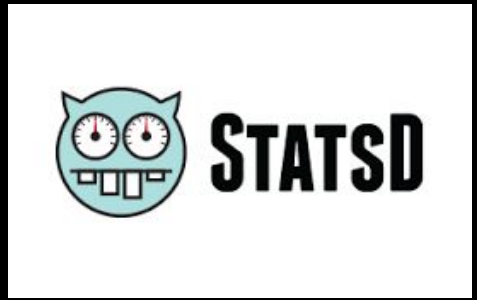
Analysis



Level 2

StatsD and Graphite.





Monitoring is many things…

- Observe. (“Shiny dashboards.”)
- Explore. (“What caused this outage?”)
- Alert. (“The machines wake me up when they need me.”)

Container orchestration

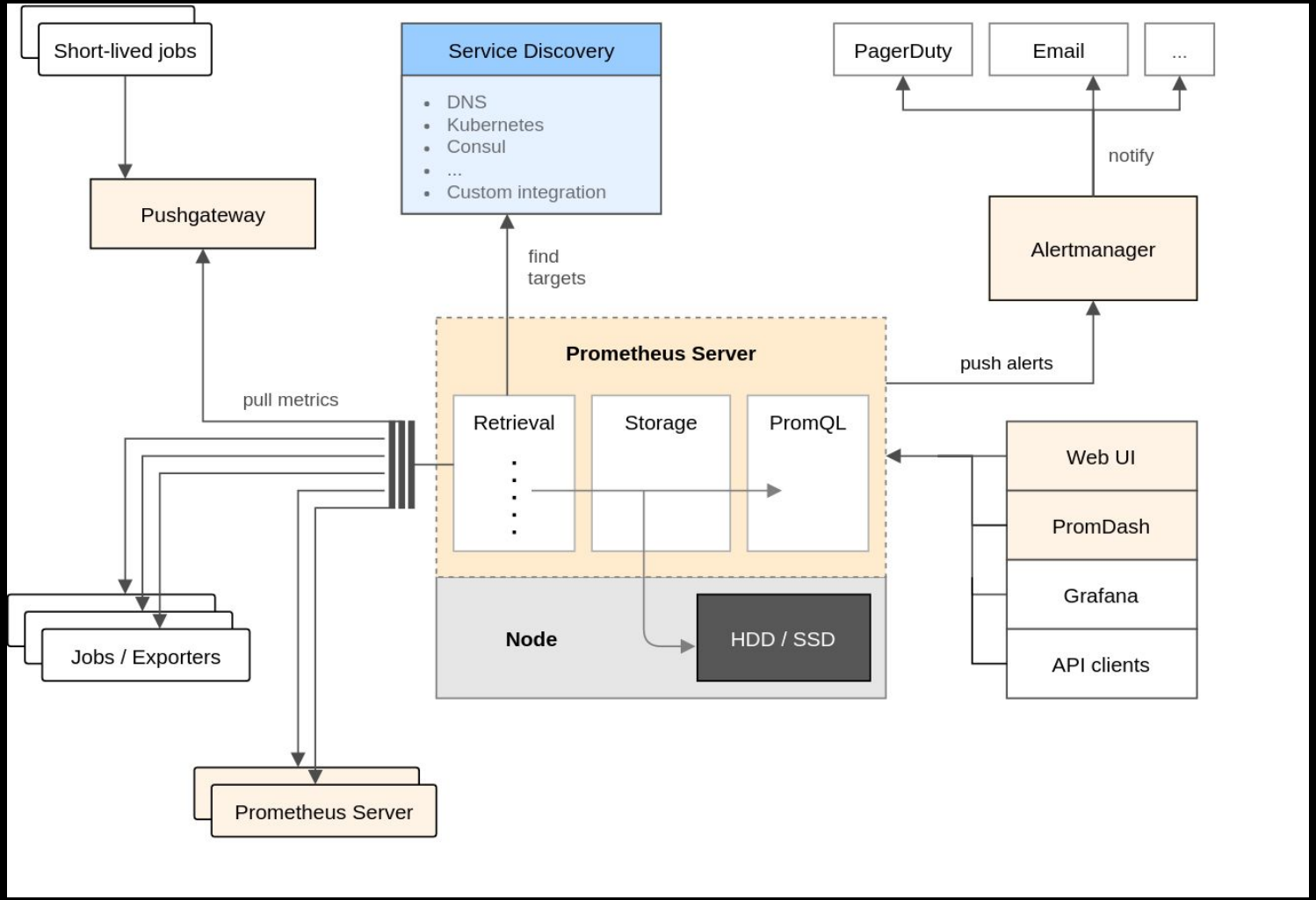
Bazooka (R.I.P.) and…

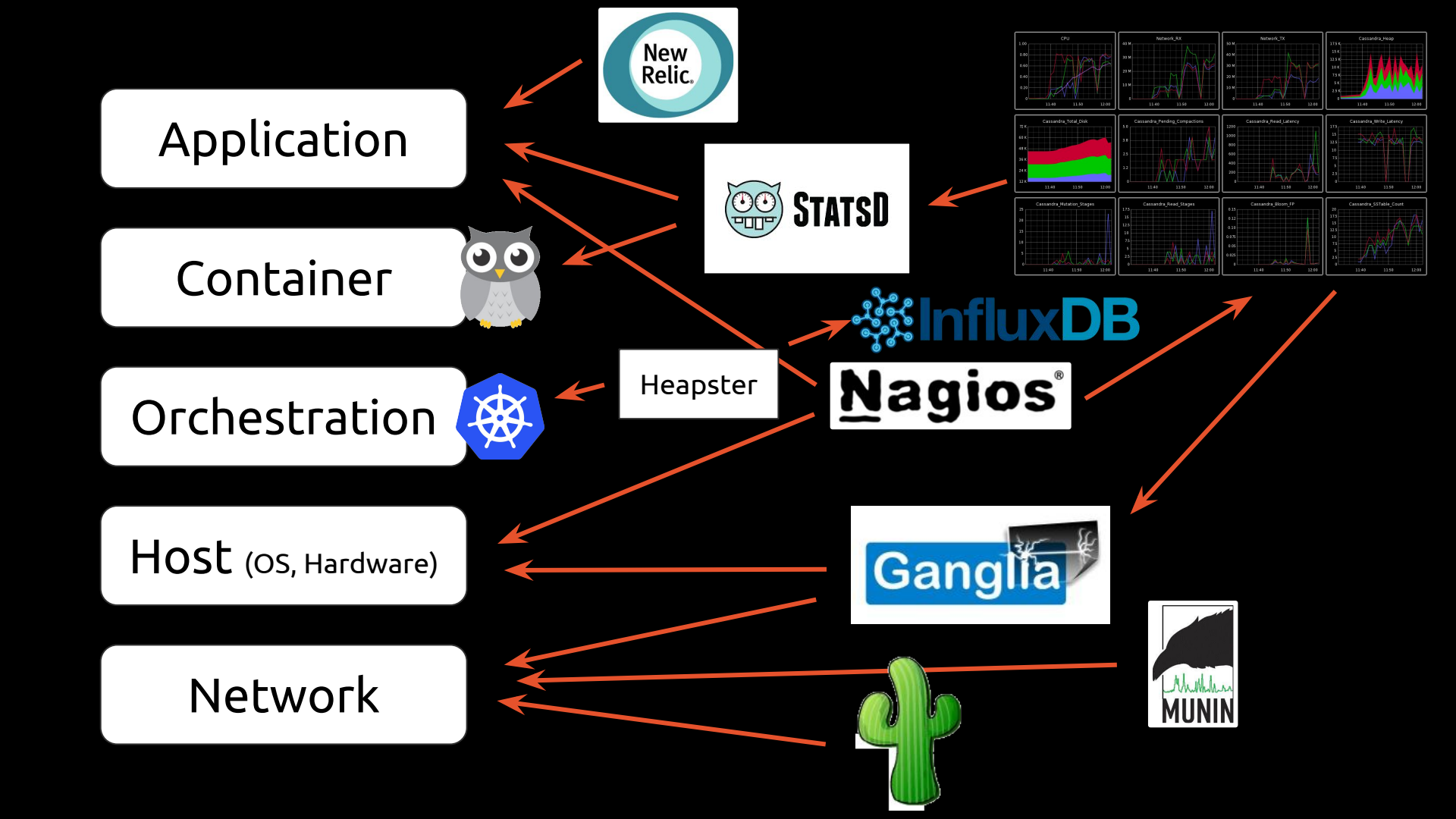
We need monitoring systems that allow us to alert for high-level service objectives, but retain the granularity to inspect individual components as needed.

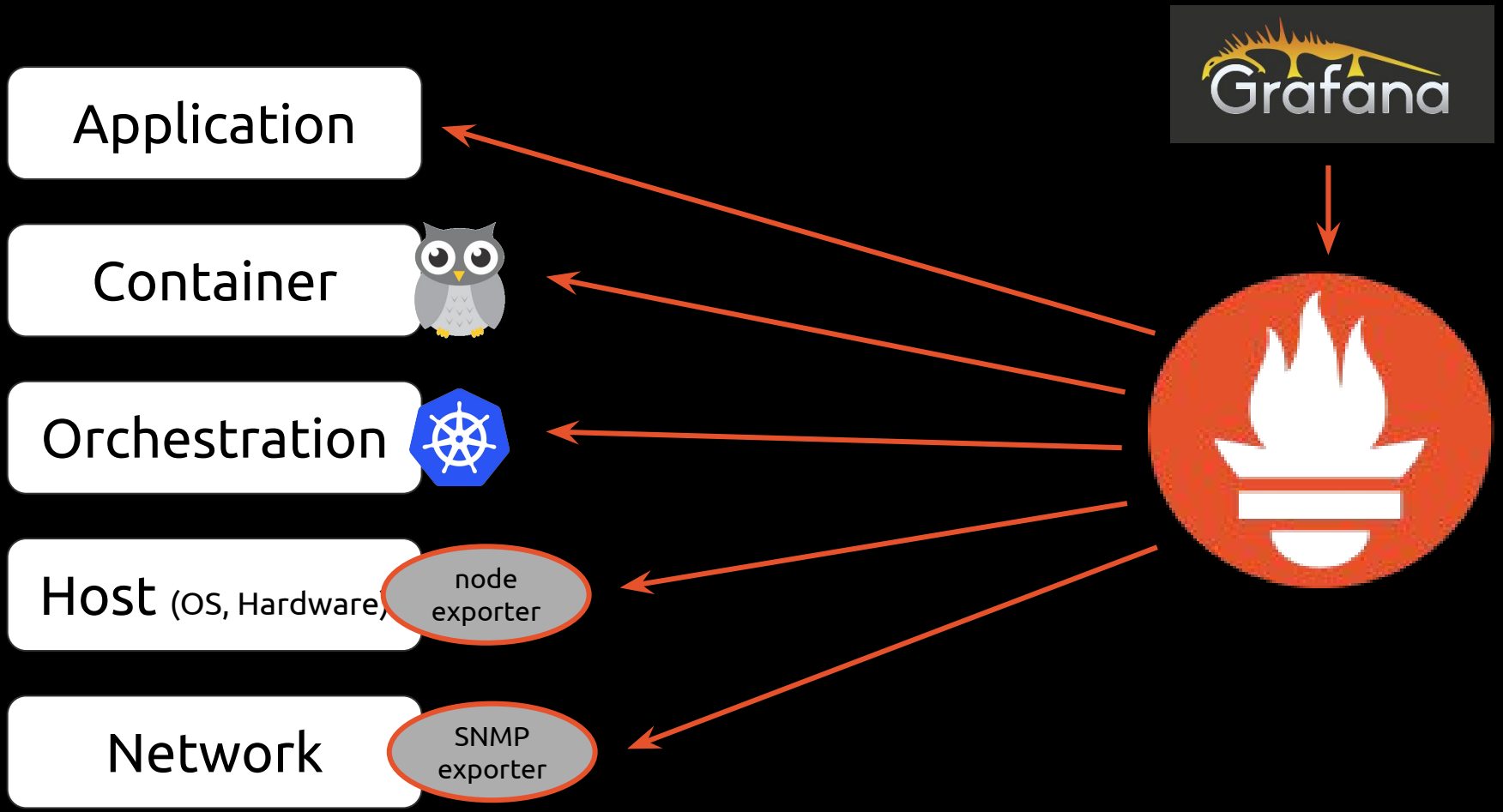
Chapter 10: Practical Alerting from Time-Series Data

Level 3

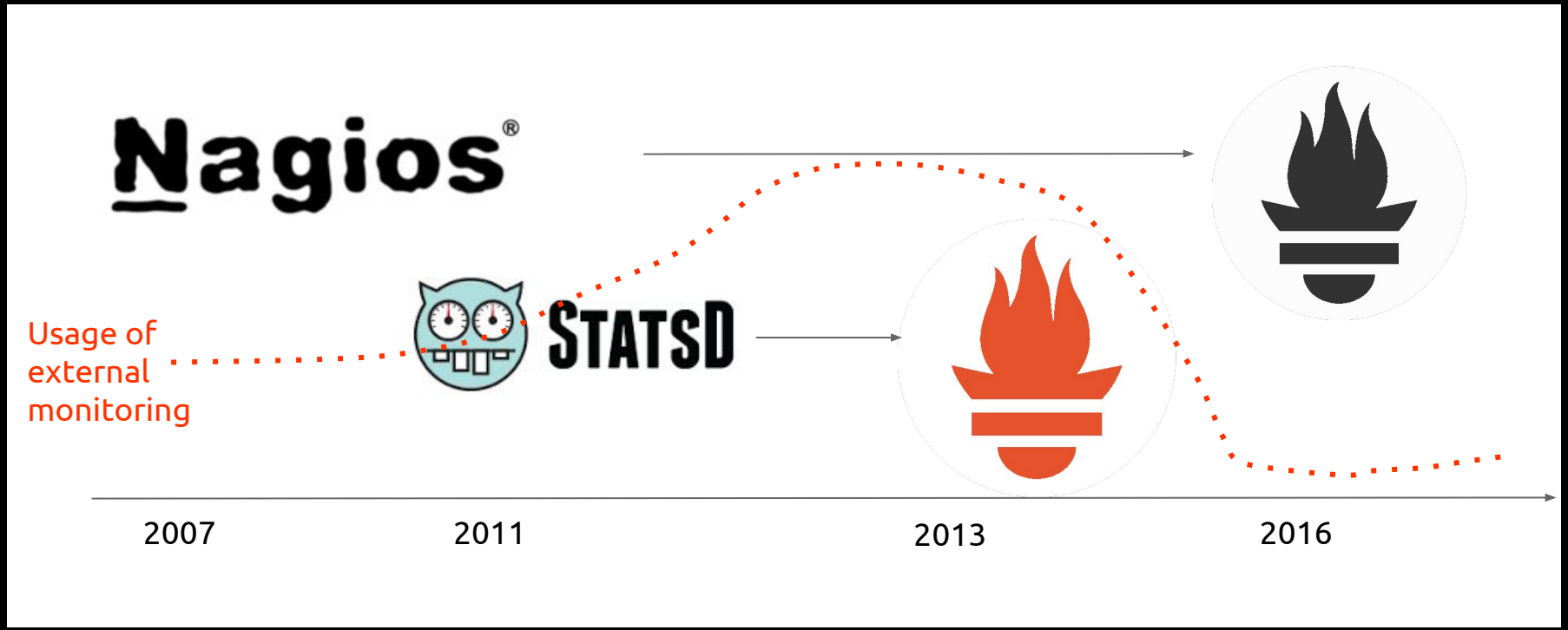
Prometheus.


Short-lived jobs
Service Discovery
DNS
Kubernetes
Consul
...
Custom integration
PagerDuty
Email
...
notify
Pushgateway
find
targets
Alertmanager
Prometheus Server
push alerts
pull metrics
Retrieval
Storage
PromQL
Web UI
PromDash
Grafana
API clients
Node
HDD / SSD
Jobs / Exporters
Prometheus Server


Application
Container
Orchestration
Host (OS, Hardware)
Network
New Relic
StatsD
InfluxDB
Heapster
Nagios
Ganglia
MUNIN


Application
Container
Orchestration
Host (OS, Hardware)
node exporter
Network
SNMP exporter
Grafana

A word about external monitoring providers



Off-premises insight for your on-premises Prometheus servers.

Latency.at measures performance and availability of your sites and services from multiple global locations and provides the results as Prometheus metrics.

Instrumentation

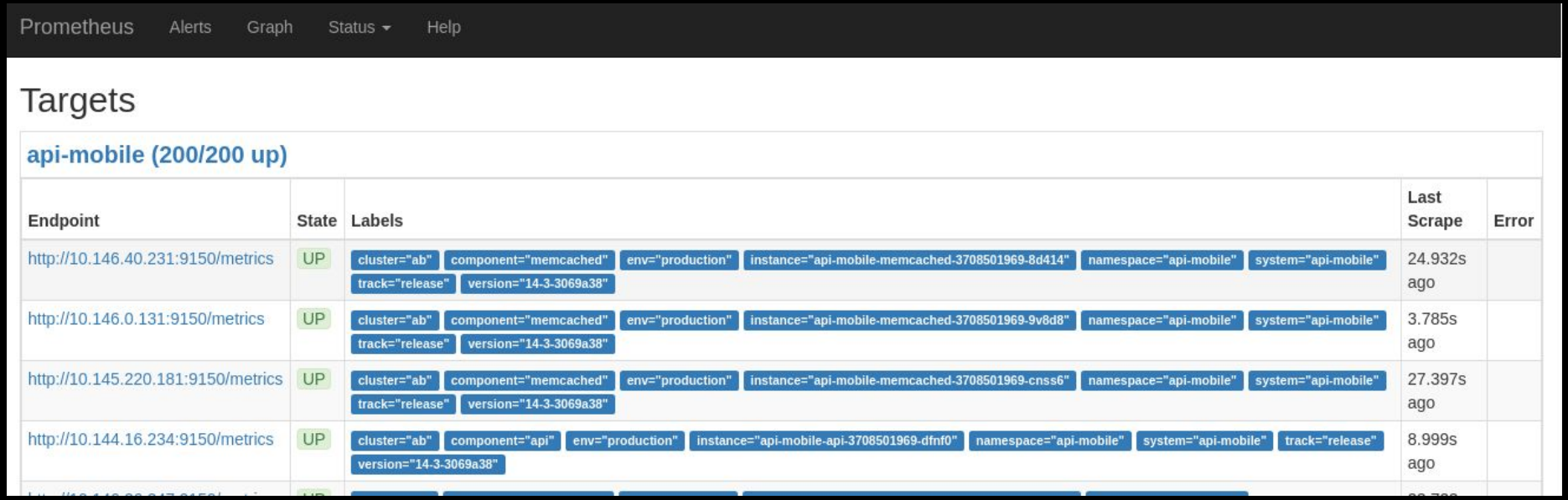
JVMKit

```
func ExampleHandler(w http.ResponseWriter, r *http.Request) {
    status := http.StatusOK
    defer func(begun time.Time) {
        duration.Observe(time.Since(begun).Seconds())
        counter.With(prometheus.Labels{
            "status": fmt.Sprint(status),
        }).Inc()
    }(time.Now())
    result, err := computeResult()
    switch err {
    case nil:
        fmt.Fprintln(w, result)
        return
    case errTimeout:
        status = http.StatusServiceUnavailable
```

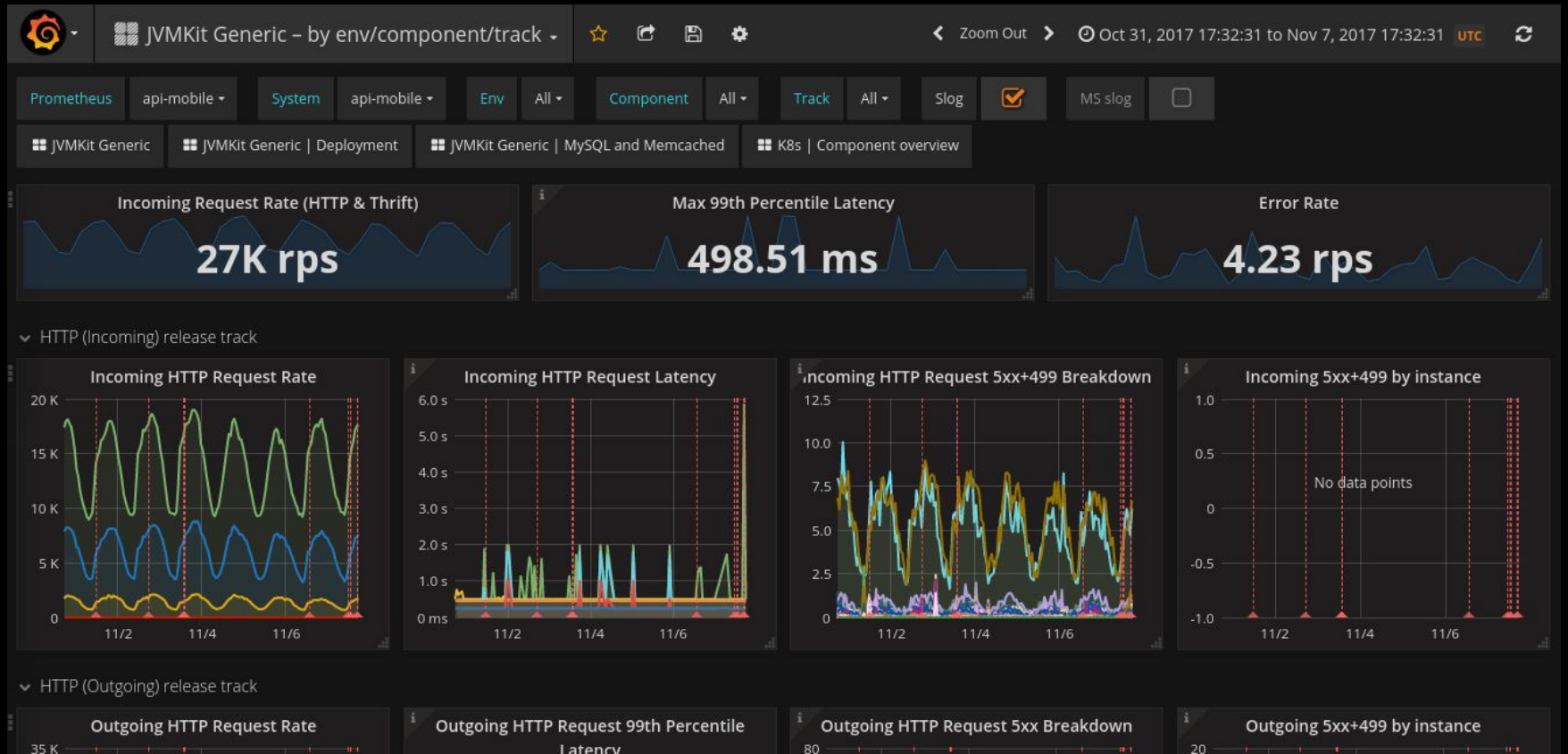
Collection

```
'prometheus' => {
   'storage' => {
     'retention' => '360h',
   },
   'owner' => 'api-mobile',
   'scrape_interval' => '30s',
   'jobs' => [
     {
       'name' => 'api-mobile',
       'k8s' => {
         'cluster' => 'ab',
         'system' => 'api-mobile',
       },
     },
   ],
   # ...
```

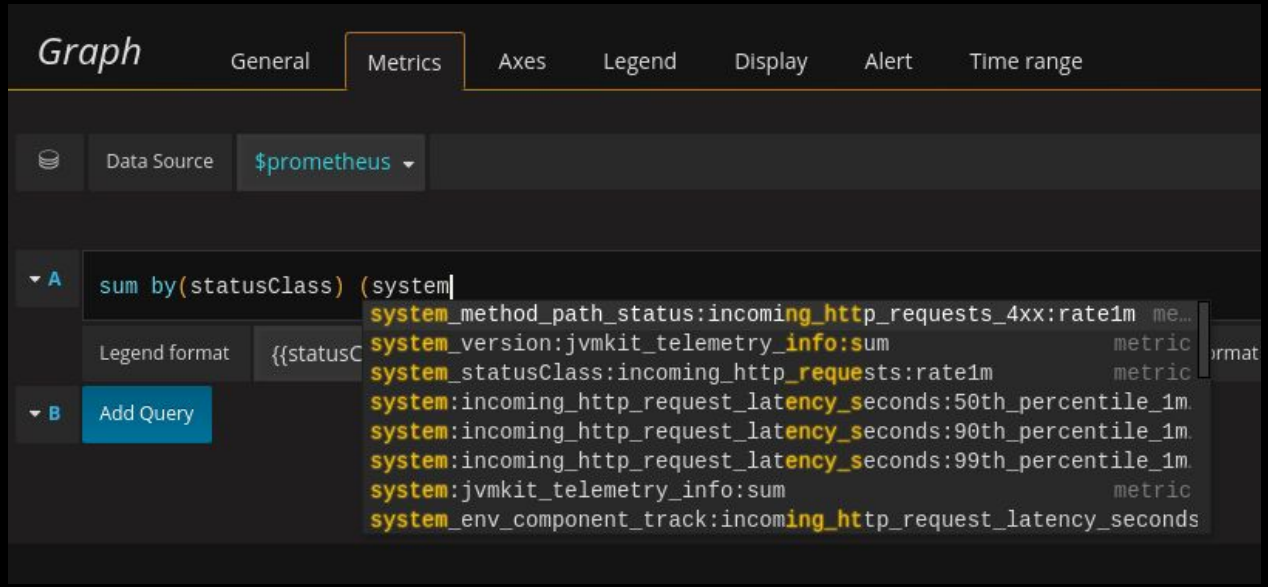
```
- job_name: api-mobile
  scrape_interval: 30s
  scrape_timeout: 10s
  metrics_path: /metrics
  scheme: http
  kubernetes_sd_configs:
  - api_server: https://api.k2.k8s.s-cloud.net
    role: pod
    tls_config:
      ca_file: /mnt/prometheus/k8s-certificates/ab/ca.crt
      cert_file: /mnt/prometheus/k8s-certificates/ab/client.crt
      key_file: /mnt/prometheus/k8s-certificates/ab/client.key
      insecure_skip_verify: false
    namespaces:
      names: []
  relabel_configs:
  - source_labels: [__meta_kubernetes_namespace, __meta_kubernetes_pod_annotation_pro
    separator: ;
    regex: .+;(?:[0-9]+,?)+|;
    replacement: $1
    action: keep
  - source_labels: [__meta_kubernetes_namespace, __meta_kubernetes_pod_label_system]
    separator: ;
    regex: .+;api-mobile|;
    replacement: $1
    action: keep
  - source_labels: [__meta_kubernetes_namespace, __meta_kubernetes_pod_label_env]
    separator: ;
    regex: .+;production|;
    replacement: $1
    action: keep
  - source_labels: [__meta_kubernetes_pod_annotation_prometheus_io_scheme]
    separator: ;
    regex: (https?)
    target_label: __scheme__
    replacement: $1
```

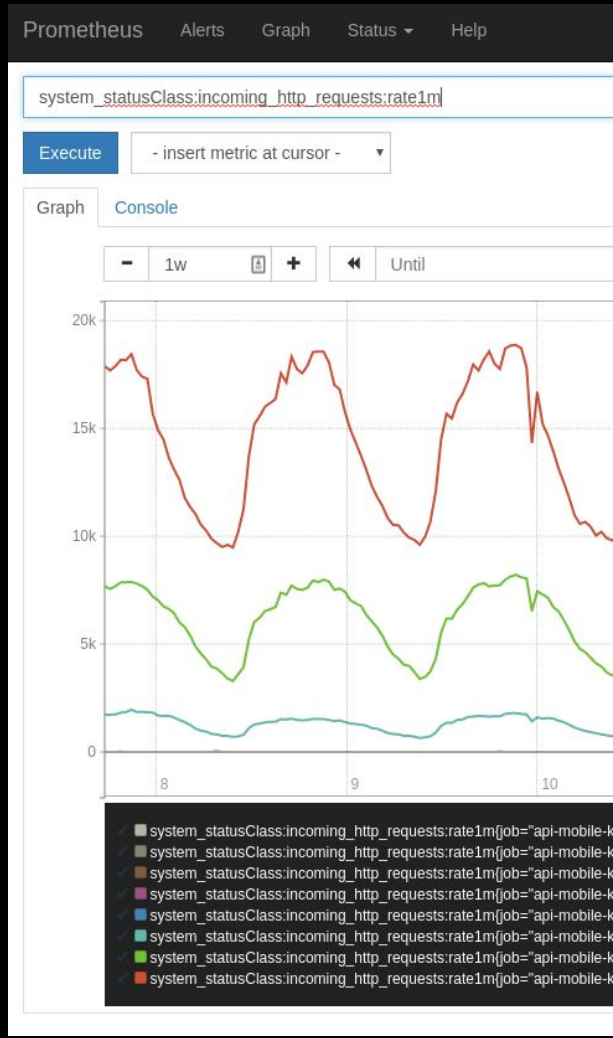

Prometheus
Alerts
Graph
Status
Help
Targets
api-mobile (200/200 up)
Endpoint
State
Labels
Last Scrape
Error
http://10.146.40.231:9150/metrics
UP
cluster="ab" component="memcached" env="production" instance="api-mobile-memcached-3708501969-8d414" namespace="api-mobile" system="api-mobile" track="release" version="14-3-3069a38"
24.932s ago
http://10.146.0.131:9150/metrics
UP
cluster="ab" component="memcached" env="production" instance="api-mobile-memcached-3708501969-9v8d8" namespace="api-mobile" system="api-mobile" track="release" version="14-3-3069a38"
3.785s ago
http://10.145.220.181:9150/metrics
UP
cluster="ab" component="memcached" env="production" instance="api-mobile-memcached-3708501969-cnss6" namespace="api-mobile" system="api-mobile" track="release" version="14-3-3069a38"
27.397s ago
http://10.144.16.234:9150/metrics
UP
cluster="ab" component="api" env="production" instance="api-mobile-api-3708501969-dfnf0" namespace="api-mobile" system="api-mobile" track="release" version="14-3-3069a38"
8.999s ago

Dashboards


JVMKit Generic – by env/component/track
Zoom Out
Oct 31, 2017 17:32:31 to Nov 7, 2017 17:32:31 UTC
Prometheus
api-mobile
System
api-mobile
Env
All
Component
All
Track
All
Slog
MS slog
JVMKit Generic
JVMKit Generic | Deployment
JVMKit Generic | MySQL and Memcached
K8s | Component overview
Incoming Request Rate (HTTP & Thrift)
27K rps
Max 99th Percentile Latency
498.51 ms
Error Rate
4.23 rps
HTTP (Incoming) release track
Incoming HTTP Request Rate
20 K
15 K
10 K
5 K
0
11/2
11/4
11/6
Incoming HTTP Request Latency
6.0 s
5.0 s
4.0 s
3.0 s
2.0 s
1.0 s
0 ms
11/2
11/4
11/6
Incoming HTTP Request 5xx+499 Breakdown
12.5
10.0
7.5
5.0
2.5
0
11/2
11/4
11/6
Incoming 5xx+499 by instance
1.0
0.5
0
-0.5
-1.0
No data points
11/2
11/4
11/6
HTTP (Outgoing) release track
Outgoing HTTP Request Rate
35 K
Outgoing HTTP Request 99th Percentile Latency
Outgoing HTTP Request 5xx Breakdown
80
Outgoing 5xx+499 by instance
20

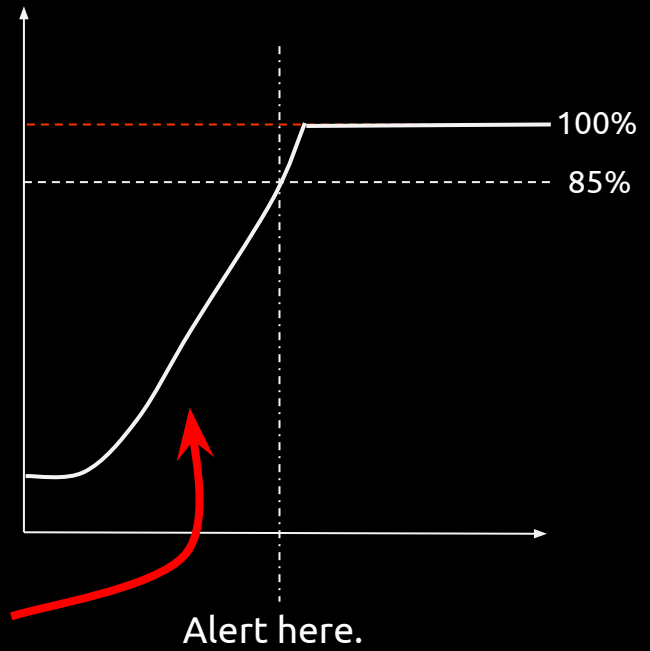
Interactive exploration



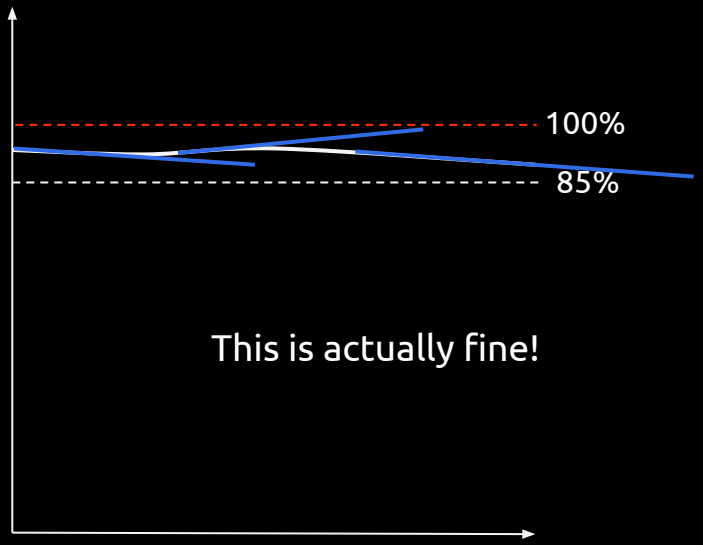


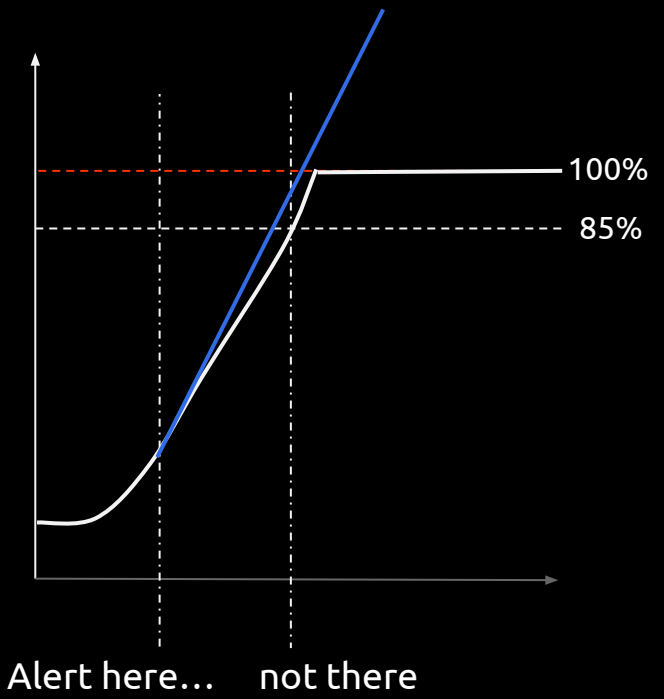
Alert creation

```
ALERT PrometheusRuleEvaluationSlow
  IF
      rate(prometheus_evaluator_iterations_missed_total[1h])
    /
      rate(prometheus_evaluator_iterations_total[1h])
    * 100
    > 5
  FOR 4h
  LABELS {
    severity = "warning",
  }
  ANNOTATIONS {
    summary = "{{$labels.job}} is evaluating rules too slowly",
    description = "In the last hour, {{$labels.job}} at {{$labels.instance}} has ...",
    runbook = "http://doc/runbooks/prometheus/#prometheusruleevaluationslow",
    dashboard = 'http://grafana/dashboard/db/prometheus-srv?var-job={{$labels.job}}...',
```



This is fine!?!

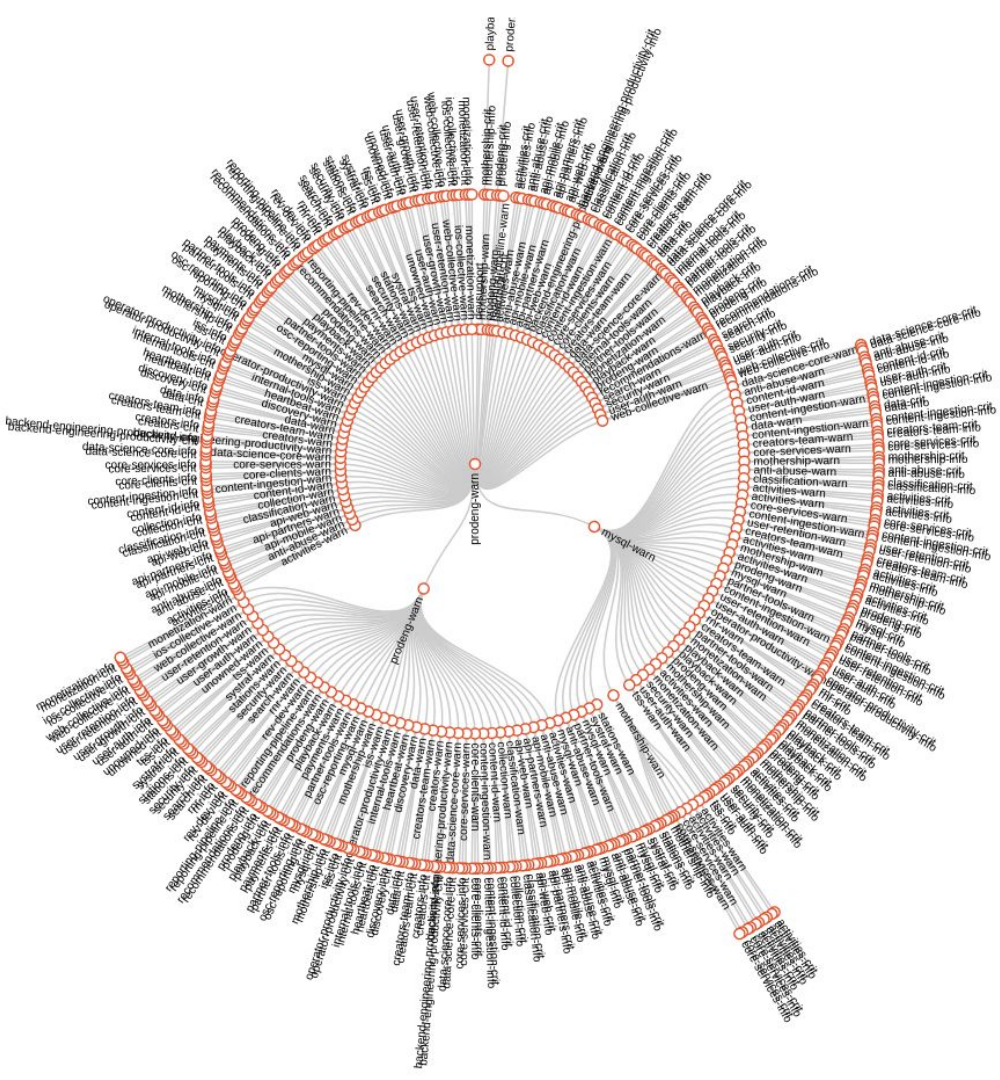

100%
85%
This is actually fine!


100%
85%
Alert here…
not there

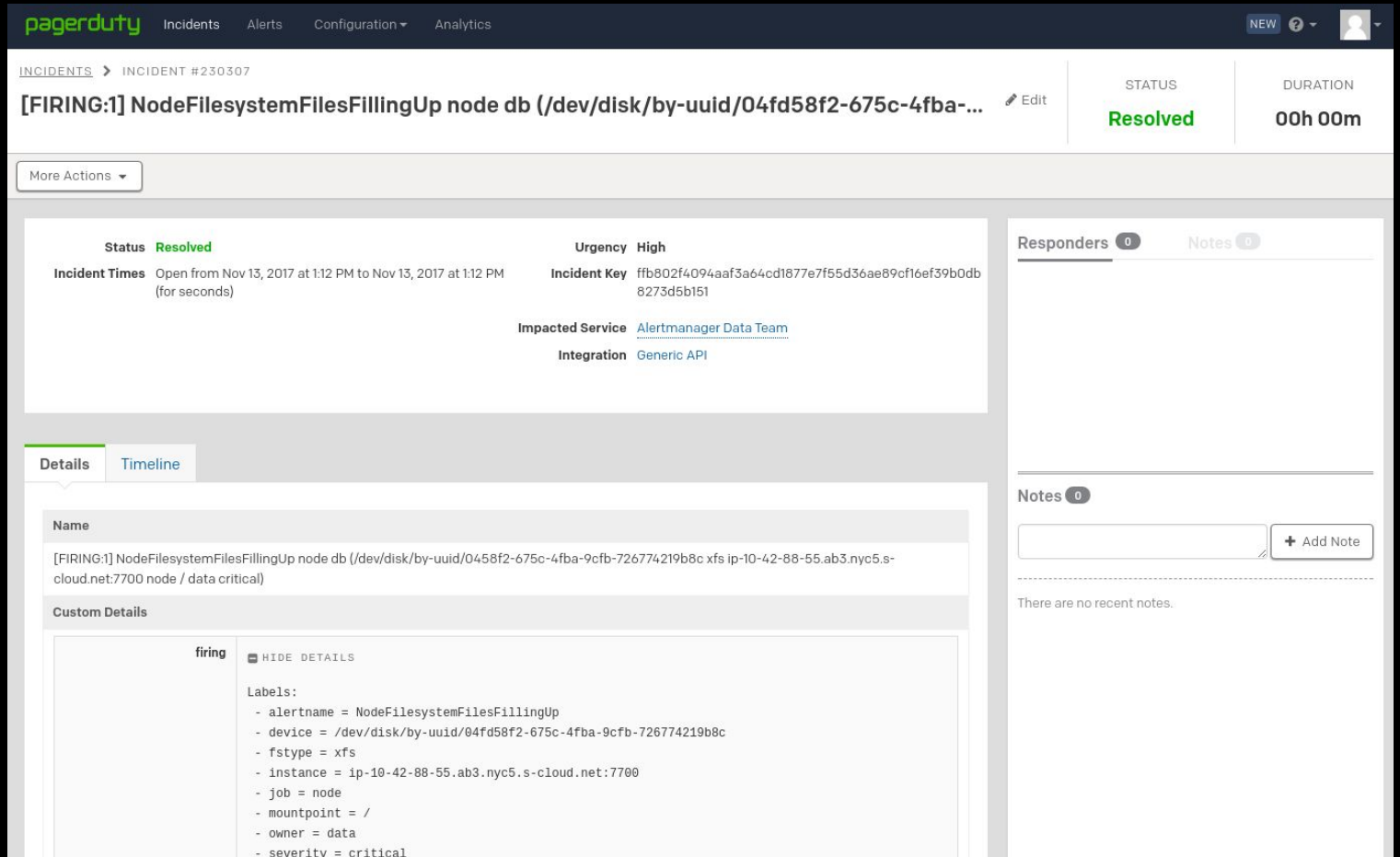
```
ALERT NodeFilesystemSpaceFillingUp
  IF
      predict_linear(node_filesystem_avail{fstype=~"ext."}[1h], 4*60*60) < 0
    AND
      node_filesystem_avail{fstype=~"ext."} / node_filesystem_size{fstype=~"ext."} < 0.2
    AND
      node_filesystem_readonly{fstype=~"ext."} == 0
  FOR 10m
  LABELS {
    severity = "critical",
  }
  ANNOTATIONS {
    summary = "Filesystem space is filling up",
    description = 'Filesystem on {{$labels.device}} at {{ $labels.instance }} is \
                   predicted to run out of space within the next 4 hours.',
    runbook = "http://doc/runbooks/node/#nodefilesystemspacefillingup",
    dashboard = 'http://grafana/dashboard/db/node?var-node={{ $labels.instance ...',
  }
```

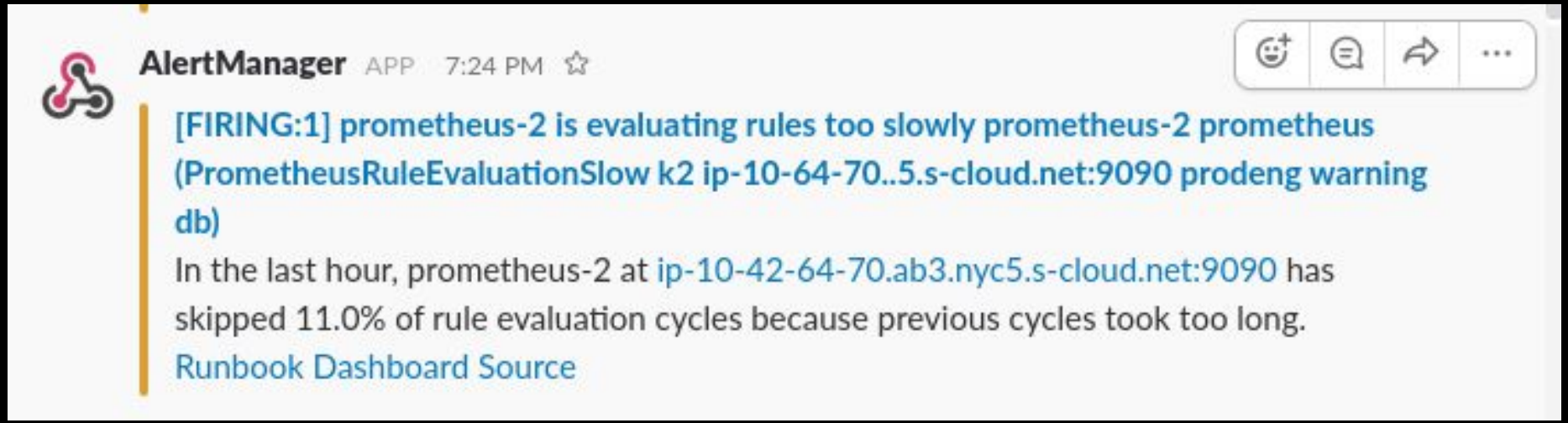
Alert grouping and routing

```
route:
  receiver: prodeng-warn
  group_by:
  - alertname
  - zone
  - system
  routes:
  - receiver: api-mobile-warn
    match:
      system: api-mobile
    routes:
    - receiver: api-mobile-crit
      match:
        severity: critical
      group_wait: 20s
      group_interval: 5m
```

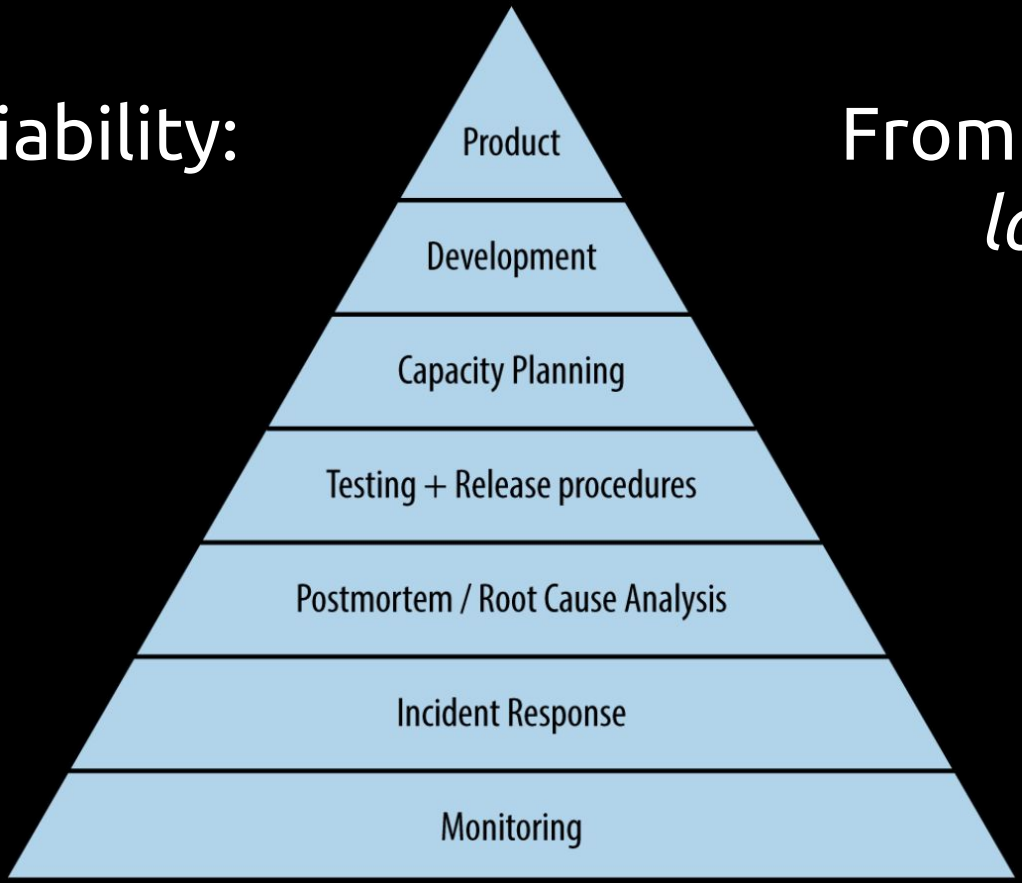


Alert delivery




AlertManager APP 7:24 PM
[FIRING:1] prometheus-2 is evaluating rules too slowly prometheus-2 prometheus (PrometheusRuleEvaluationSlow k2 ip-10-64-70..5.s-cloud.net:9090 prodeng warning db)
In the last hour, prometheus-2 at ip-10-42-64-70.ab3.nyc5.s-cloud.net:9090 has skipped 11.0% of rule evaluation cycles because previous cycles took too long.
Runbook Dashboard Source

Service reliability:



From *unknown* to *low* to *high* in four years.

Dickerson's hierarchy of service reliability
Site Reliability Engineering – How Google Runs Production Systems, B. Beyer et al. (ed.), O'Reilly 2016, p104

Slides will be linked at
https://github.com/beorn7/talks

Please

Remember to rate this session

Thank you!