

Observability at Scale:

**How Open Systems Collects Telemetry
from Over 10'000 Edge Devices Worldwide**



Observability Day

Joel Verezhak

18.04.2023



KubeCon



CloudNativeCon

Europe 2023

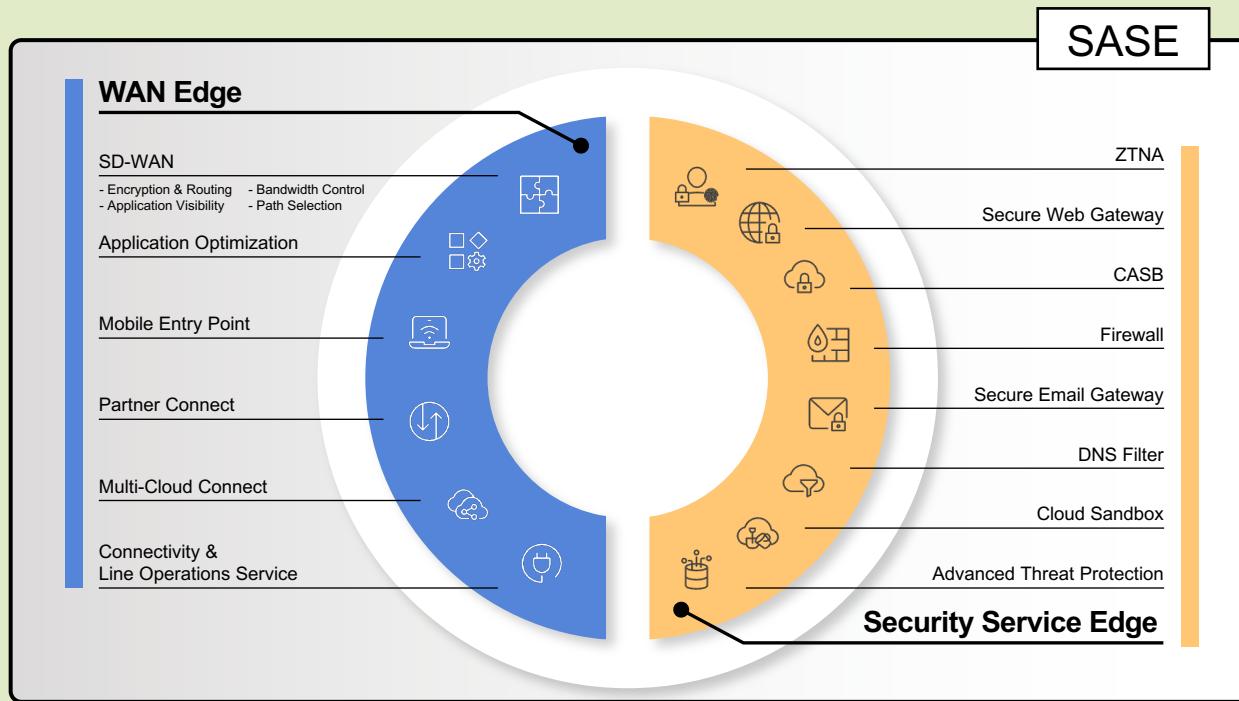


Agenda

- What we do (the *Status Quo*)
- Challenges and Action Plan
- Putting it into motion
- Outlook and Horizon

What we do: Open Systems Managed SASE

Managed SASE



Centralised policy management

24/7 Global NOC

HW and SW lifecycle

Expert Support

Data observability/analytics

Cloud and On-premises Edges

Unified data platform

Single pane of glass

The Portal

Jump To ⌘ ⌘ K

OVERVIEW TICKETS PEOPLE FILES AUDIT TRAIL

SASE Atlas Create Ticket

9 1 1

Open Systems

Dashboard Company Reports Self-Service Help

⚠ The following Open Systems service is temporarily affected:
01.02.2023 09:12:40 1190333 Testing for user context [change]

Service Map

Service Map ⊕ Service List ⊕

Availability

141	HOSTS UP
6	HOSTS DOWN
14	ISP OUTAGES

Recent Outages

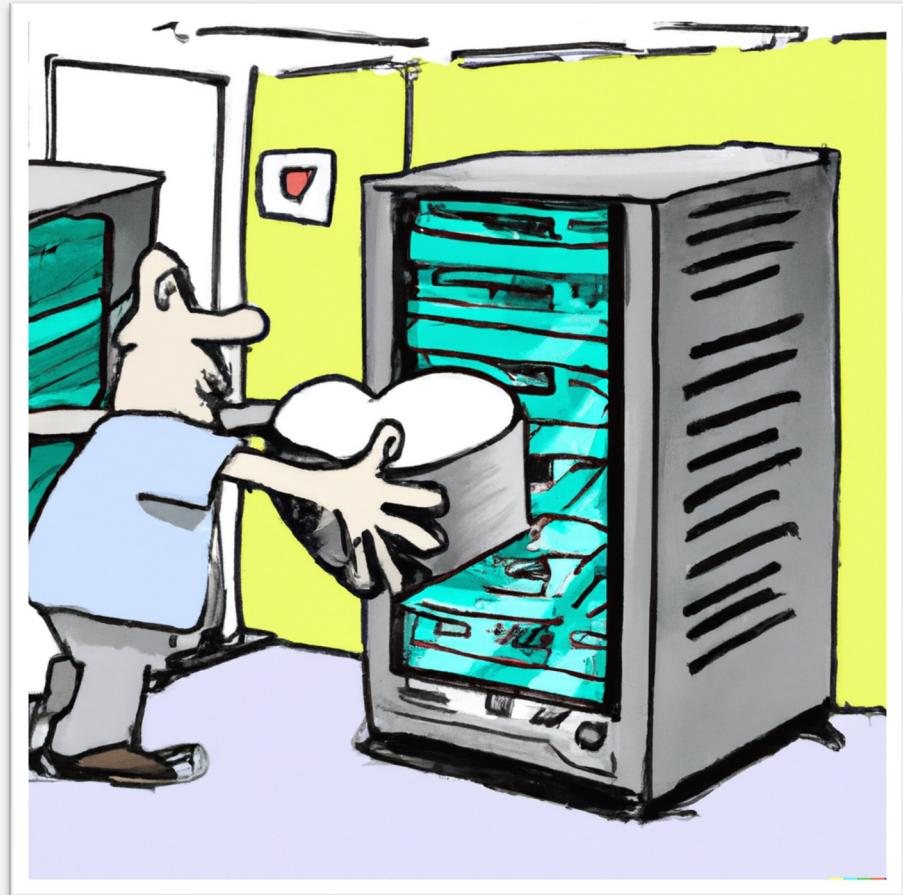
• oradev2	Zürich	for 5 days
• oradev1	Zürich	for 5 days
• open-icap001-r29-1	Zürich	for 18 days
• stc-sg005-ch-zur-aj-1	Zürich	for 18 days
• mdr-bamboo-1	Sydney	for 18 days

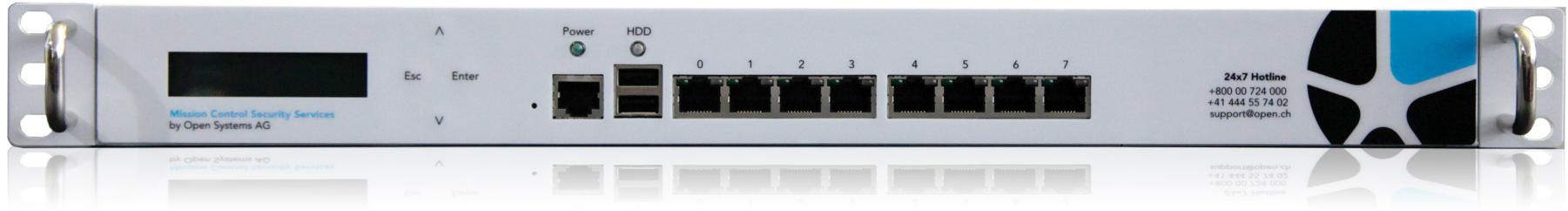
Service Availability ⊕

The status quo

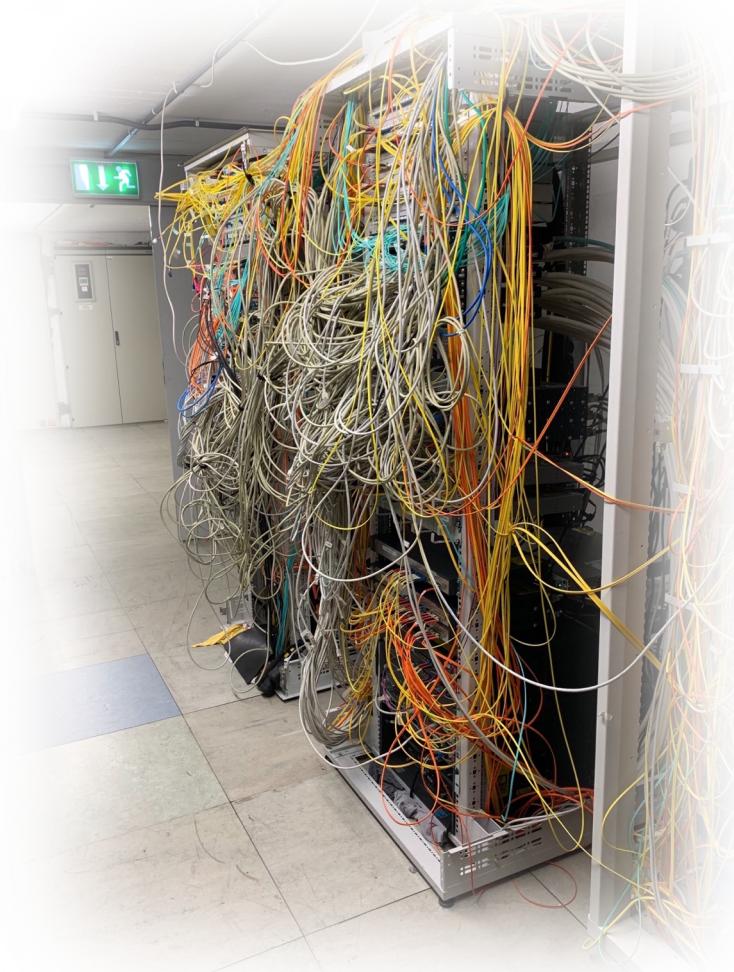
Our fleet of edge devices

- **Individually named**
- **Lovingly provisioned in-house before shipment to customer**
- **Monitored constantly in our Network Operations Centre**

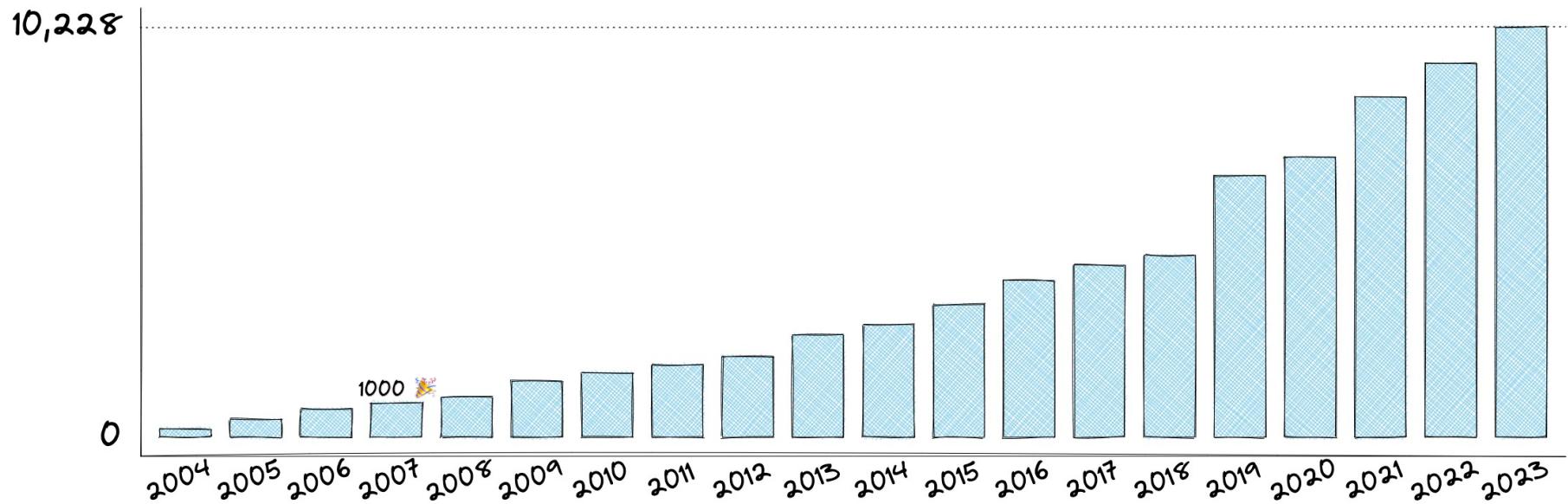








Total Active Hosts



The status quo

What happens when a device gets sick?

- **Device lovingly nursed back to health by Mission Control**
 1. ssh into the box
 2. grep the logs
 3. init.d status
 4. ...



Mission Control

Total Tickets 12 Escalation 1190154 Overdue Tickets 0 Processed Tickets 21

[Create Ticket](#) [Panic](#)

Search
Andre Baptista Aguas @aba
Incoming 0
No engineer handling this queue
Dispatch 2
No engineer handling this queue
Routine 0
No engineer handling this queue
Process 0
No engineer handling this queue
SOC 8
No engineer handling this queue
Other 0
Ticket List MC Portal

Andre Baptista Aguas

ISP outage academy-bgp001-ch-zur-2

application notification: UPSID:FEED:CYREN:ALERT open

1177246 Change application notification: UPSID:FEED:CYREN:ALERT

aba - Andre Baptista Aguas MCC Not assigned to a Queue Actions

open - Open Systems AG No Host Other No Ticket Owner

Current Status

Default All Expand All

Event	Type	Author	Date
job group 2165: skipped job 21763 - Part...	job scheduling information	Emanuele Di Nicola	17.03.2023 13:51
job group 2165: error in job 21683 - Part...	job scheduling notification	Emanuele Di Nicola	17.03.2023 13:51
job group 2165: skipped job 21707 - Part...	job scheduling information	Emanuele Di Nicola	17.03.2023 13:51
Changes were made to the Threat Prote...	configuration change	Emanuele Di Nicola	17.03.2023 13:53
Changes were made to the Threat Prote...	configuration change	Emanuele Di Nicola	17.03.2023 13:54
Changes were made to the Threat Prote...	configuration change	Emanuele Di Nicola	17.03.2023 13:58
finished job group 2165 (partial config d...	job scheduling information	Emanuele Di Nicola	17.03.2023 14:50
new job group 2166 (partial config depl...	job scheduling information	Emanuele Di Nicola	17.03.2023 16:54
Changes were made to the Threat Prote...	configuration change	Emanuele Di Nicola	17.03.2023 16:54
finished job group 2166 (partial config ...	job scheduling information	Emanuele Di Nicola	17.03.2023 16:54
new job group 70 (partial config deploy...	job scheduling information	Emanuele Di Nicola	17.03.2023 17:25
Changes were made to the Threat Prote...	configuration change	Emanuele Di Nicola	17.03.2023 17:25
finished job group 70 (partial config de...	job scheduling information	Emanuele Di Nicola	17.03.2023 17:25
new job group 71 (partial config deploy...	job scheduling information	Emanuele Di Nicola	17.03.2023 17:26
Changes were made to the Threat Prote...	configuration change	Emanuele Di Nicola	17.03.2023 17:26
finished job group 71 (partial config de...	job scheduling information	Emanuele Di Nicola	17.03.2023 17:26
Changes were made to the Threat Prote...	configuration change	Emanuele Di Nicola	17.03.2023 17:26
Update	comment	Andre Baptista Aguas	18.04.2023 11:33

[Ticket List](#) [MC Portal](#)

© 2023 Open Systems. All rights reserved.

Open
systems

11

Mission Control

Anatomy of a Ticket

When something goes wrong on a device, it ends up in a ticket

The screenshot shows a web-based ticket viewer for Open Systems. The title bar reads "Open Systems - 1193285 portals-dev-1". The URL is "mc-stable.labs.open.ch/mc/administration/tv/AdminTicketViewer/view.action?ticket=1193285". The main content area displays an incident titled "1193285 Incident: system notification: NURSE:FS:USAGE:NOTIFY".
Key details shown include:

- Status: MCC (yellow exclamation mark icon)
- Priority: MCC (checkbox checked)
- Queue: Not assigned to a Queue
- Owner: open – Open Systems AG, portals-dev-1, No Service Component, No owner
- Device: portals-dev-1 (green dot icon)
- Last 24 hours summary: -24h, -18h, -12h, -6h, 0 (green bar)
- Comments: PORTALS DE (blue speech bubble icon), PORTALS develop (blue speech bubble icon)
- Steps: A section for ticket steps.
- Add Summary: A section for adding a summary.
- Change Requests: A table showing a history of change requests.

Type	Message	Date	Owner
system notification	NURSE:FS:USAGE:NOTIFY	28.03.2023 19:09	portals-dev-1
application notification	NURSE:PORTAL_MOUNTS_CHECK:OK	30.03.2023 09:22	portals-dev-1
system failure	NURSE:FS:USAGE:ALERT	30.03.2023 09:35	portals-dev-1
escalation to os	Starting escalation to Open Systems Mission Con	30.03.2023 09:35	
system failure	SYSLOG-NG:NOSPACEONDEVICE	30.03.2023 21:28	portals-dev-1
application notification	NURSE:PORTAL_BACKEND_CHECK:OK	30.03.2023 21:30	portals-dev-1
system notification	NURSE:FS:USAGE:OK	30.03.2023 21:40	portals-dev-1
system notification	MAC address mismatch	01.04.2023 04:30	portals-dev-1
system notification	NURSE:FS:USAGE:NOTIFY	01.04.2023 06:35	portals-dev-1
system notification	MAC address mismatch	02.04.2023 04:30	portals-dev-1
system notification	MAC address mismatch	03.04.2023 04:30	portals-dev-1
system notification	NURSE:FS:USAGE:OK	03.04.2023 07:35	portals-dev-1
system notification	MAC address mismatch	04.04.2023 04:30	portals-dev-1
system notification	NURSE:FS:USAGE:NOTIFY	04.04.2023 16:45	portals-dev-1
system failure	NURSE:FS:USAGE:ALERT	05.04.2023 01:45	portals-dev-1
system notification	MAC address mismatch	05.04.2023 04:30	portals-dev-1
system notification	NURSE:FS:USAGE:OK	05.04.2023 15:25	portals-dev-1
system notification	NURSE:FS:USAGE:NOTIFY	06.04.2023 00:16	portals-dev-1
system failure	NURSE:FS:USAGE:ALERT	06.04.2023 06:36	portals-dev-1
system notification	NURSE:FS:USAGE:NOTIFY	06.04.2023 08:11	portals-dev-1
system failure	NURSE:FS:USAGE:ALERT	06.04.2023 17:16	portals-dev-1
- Buttons at the bottom: Show Demoted Events, Show All Event Details, Show History, Don't Wrap, Refresh Ticket, Download (green arrow), Upload (green arrow).

Monitoring the fleet

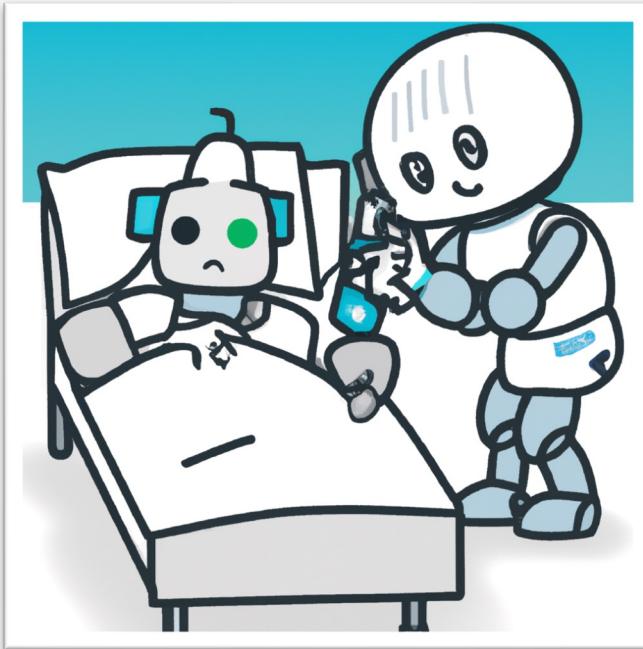
How can we tell when one of our pets
is sick?

- **Service Nurses**
- **GUMA (applications)**
- **Metrics**



Service Nurses Application Monitoring

- Small scripts that monitor critical system components
- Emit notifications (*log lines*) if predefined thresholds are met
- Each nurse is effectively a *state machine*:



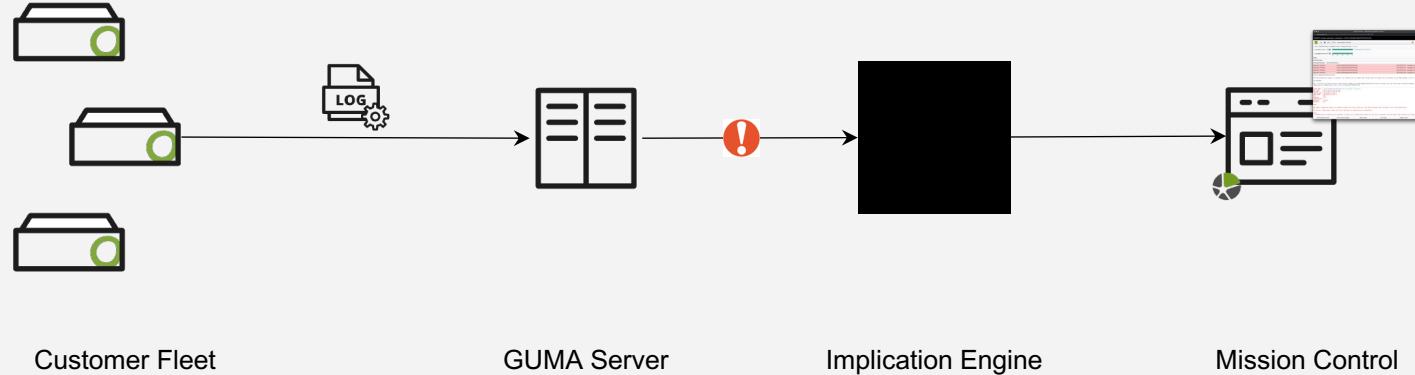
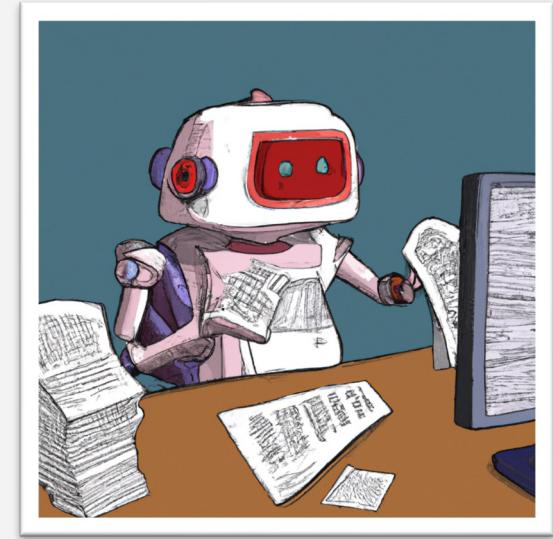
NURSE:MEM:USAGE:RAM:NOTIFY 5min average used RAM is 95% (notification limit 95%) [state=Notify, observed=900s]

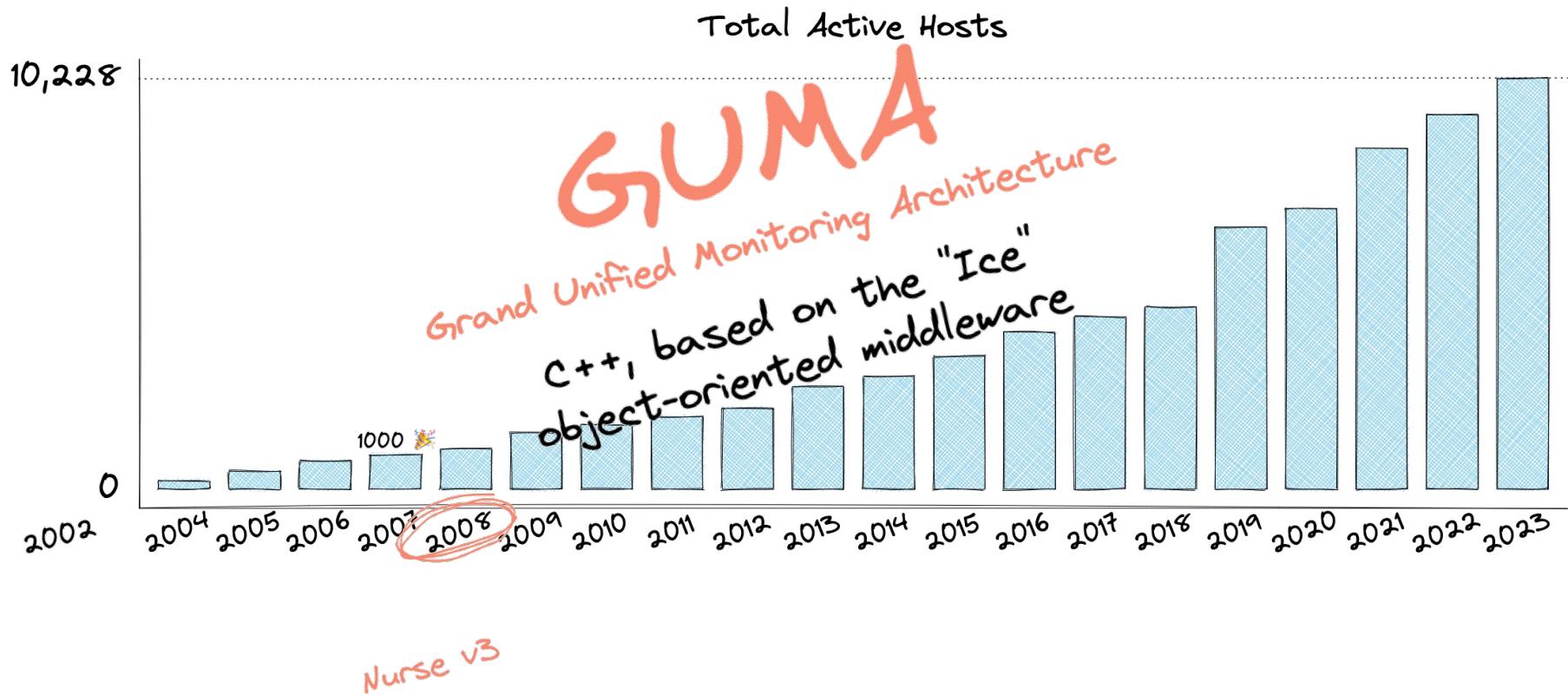
Tag Reason Meta

GUMA Application Monitoring

Grand Unified Monitoring Architecture

- Parse, classify and filter syslog
- Match log lines against a list of ~3000 regexes that map to *signatures*
- Process signatures, decide whether to alert (generate a ticket in Mission Control)

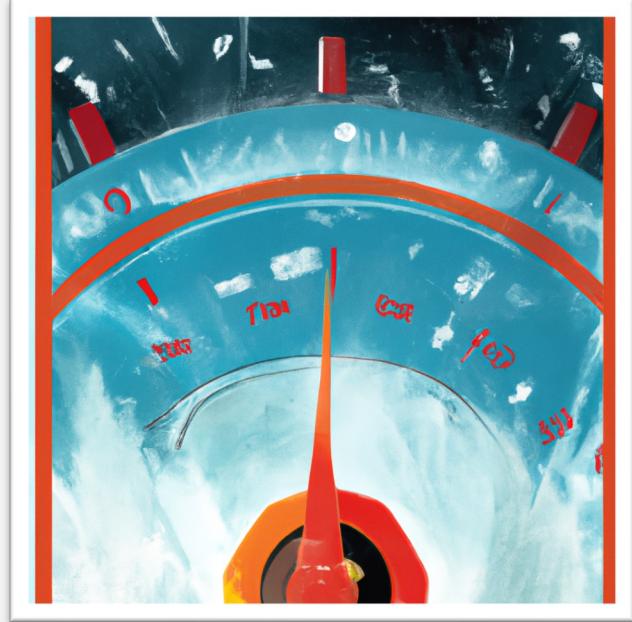




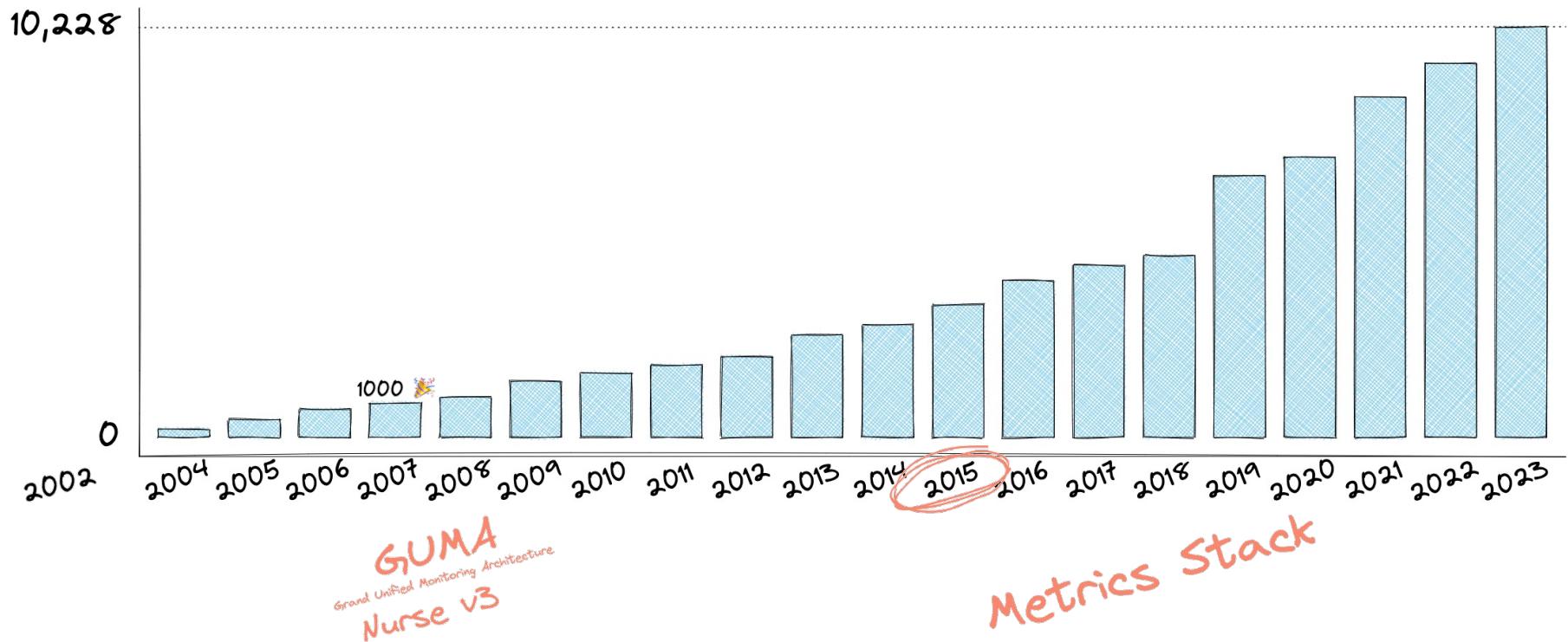
Metrics

Dashboards for the portal

- **Scrape locally on the host with Prometheus**
- **Publish to Kafka topic with Telegraf**
- **Consume into InfluxDB**
- **Continuous queries yield drill-down statistics for the Portal**



Total Active Hosts



InfluxDB

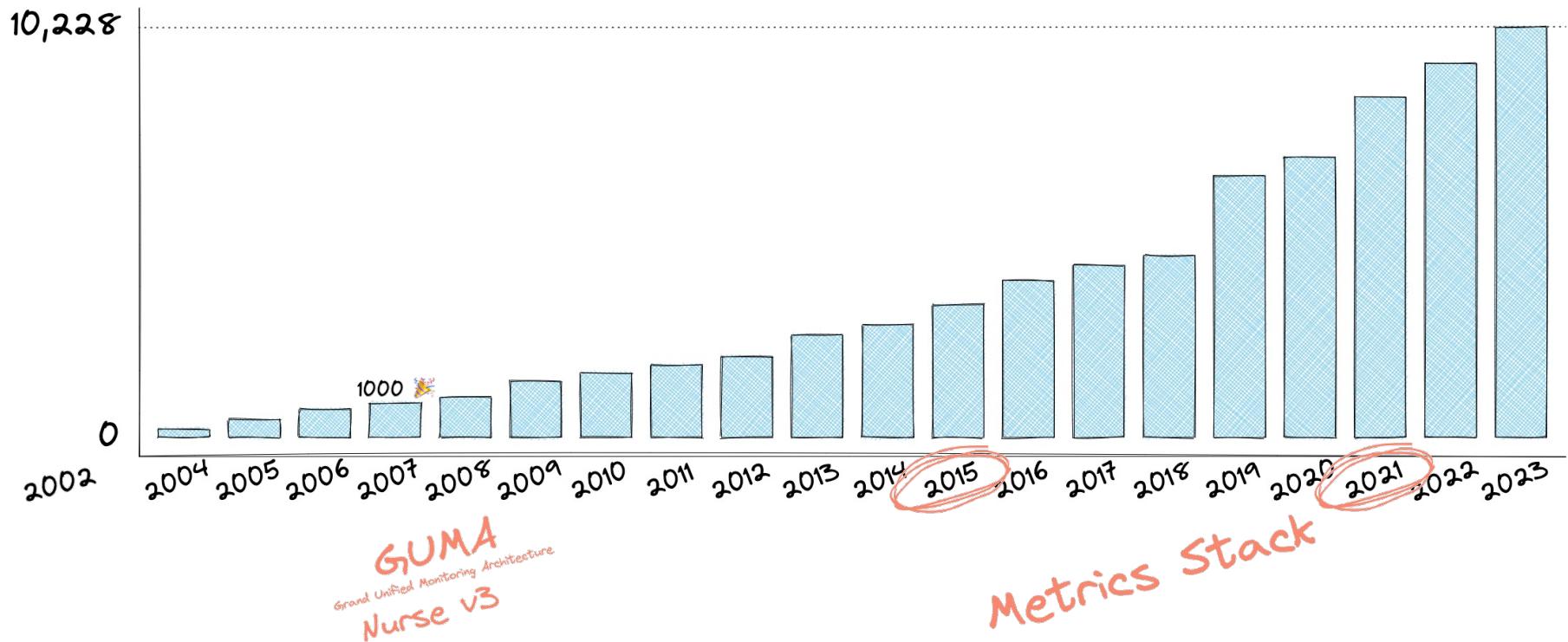
a13



i33



Total Active Hosts



InfluxDB

a13



*actual number of hosts deployed

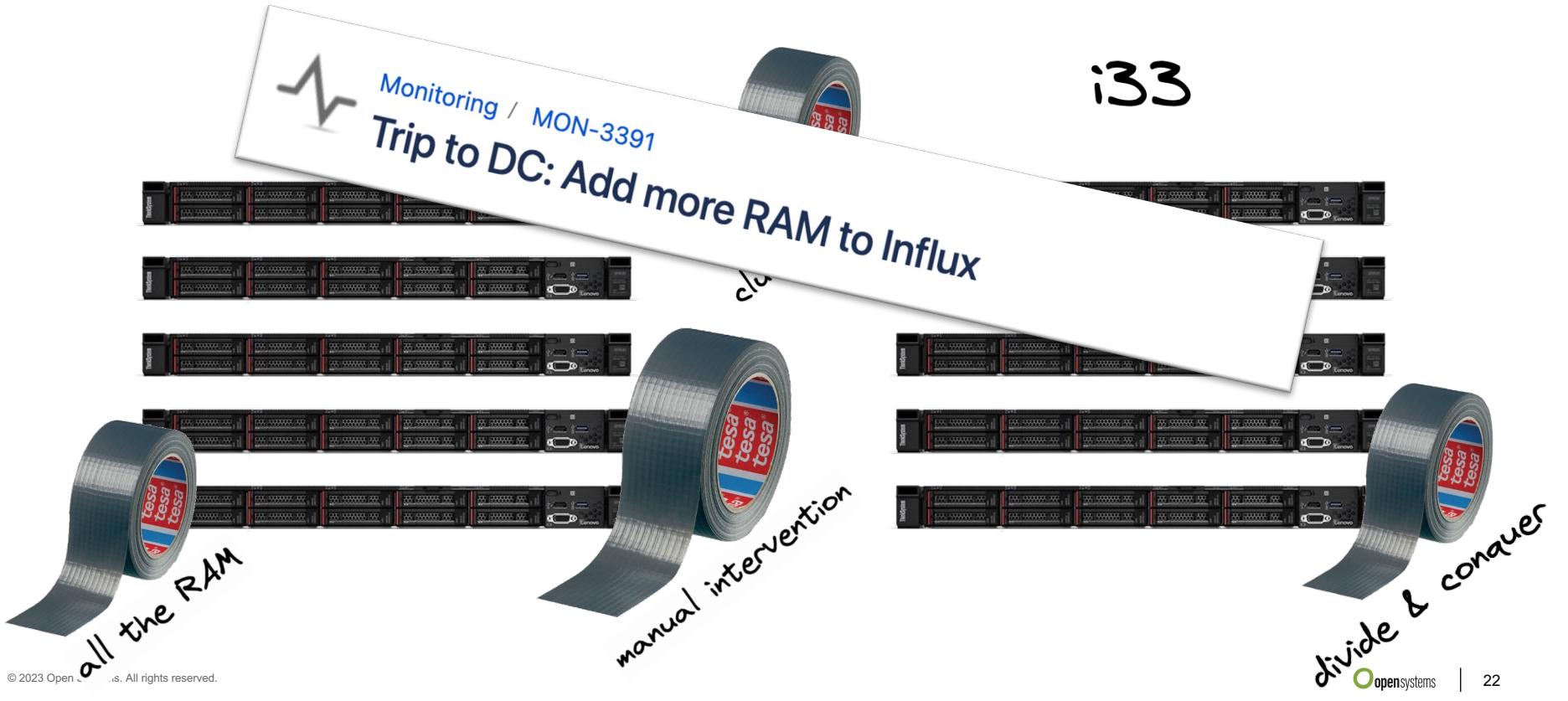


i33



divide & conquer
 opensystems

InfluxDB



The biggest challenges we need to solve

New Environments



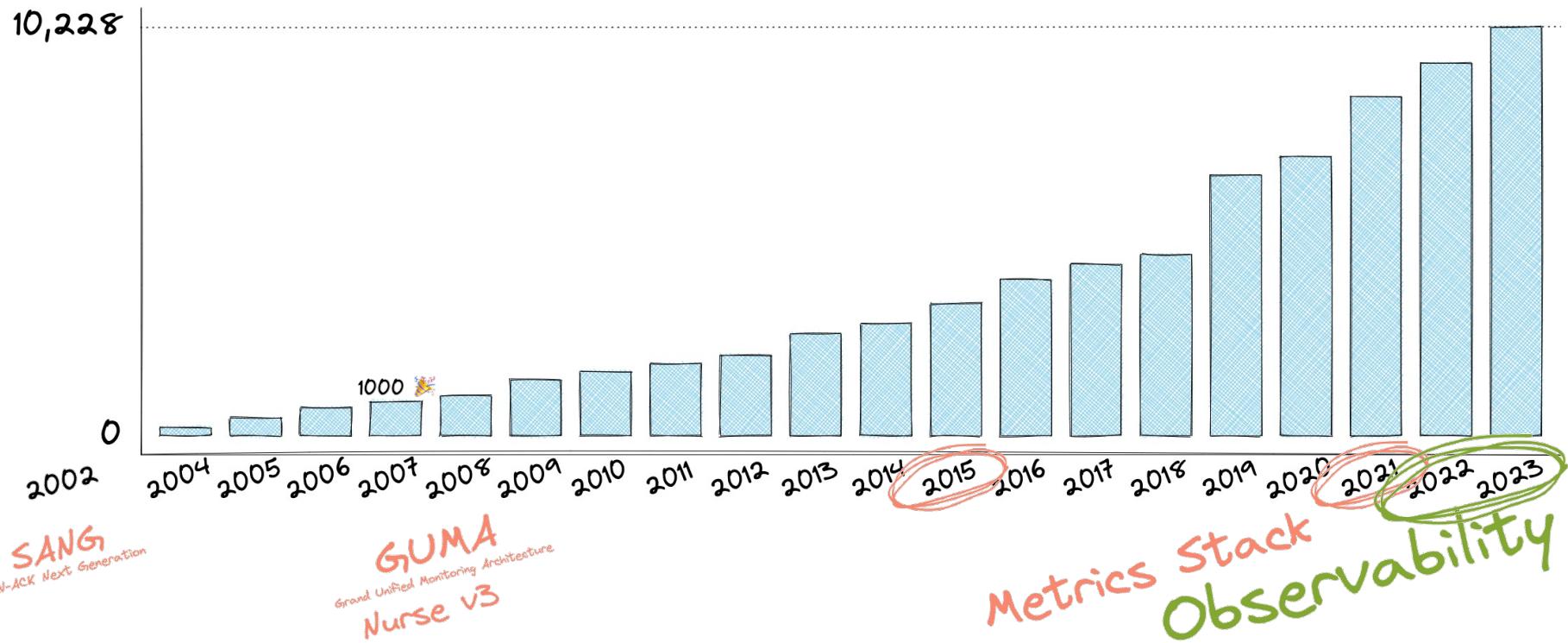
Log-centric Alerting



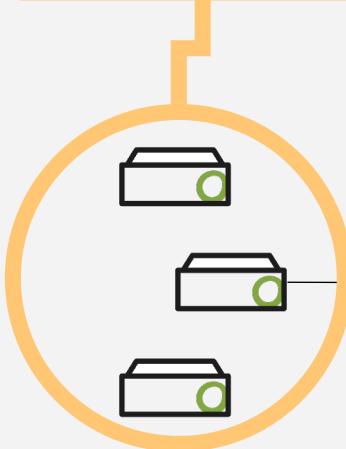
Scalability



Total Active Hosts

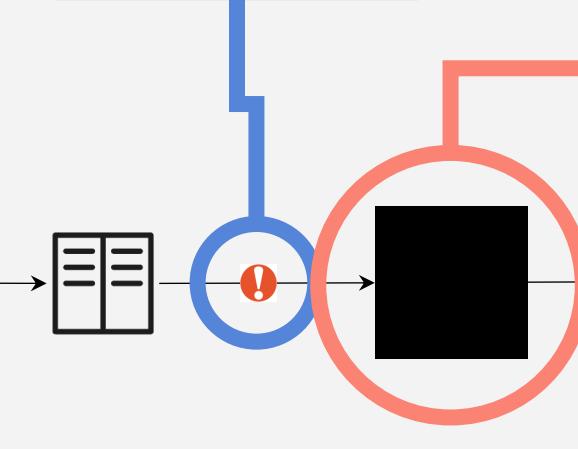


Environment agnostic Observability Platform



Customer Fleet

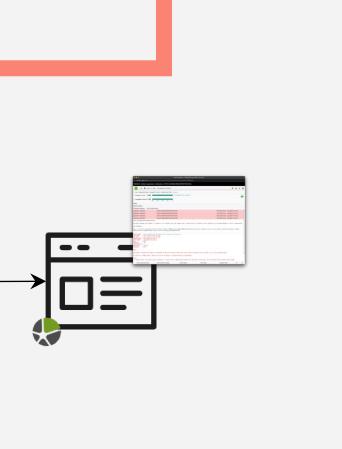
Alerting on Metrics



GUMA Server

Implication Engine

Smarter Event Handling



Mission Control

Our Telemetry Stacks at a Glance

Metrics

- 50 cores
- 800 GiB RAM
- ~370 pods
- 110 million metrics

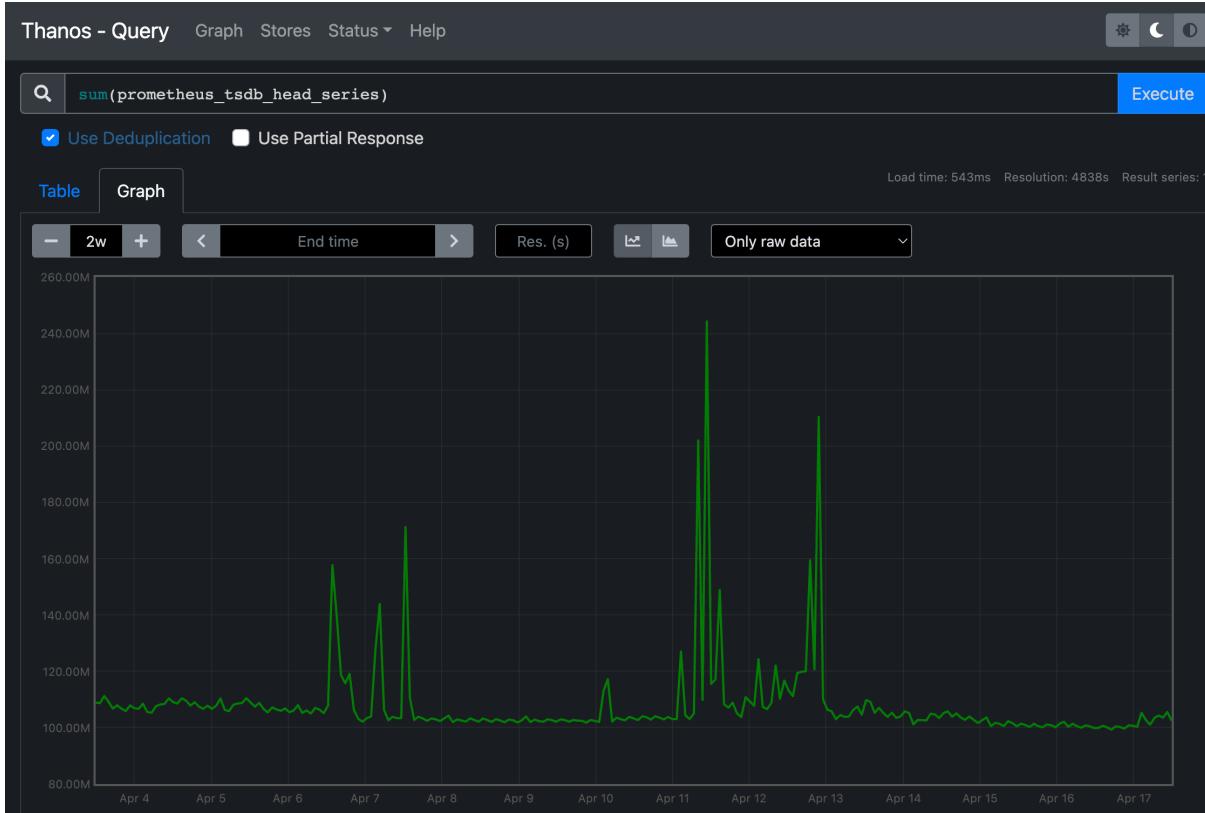


Logs

- 55 cores
- 700 GiB RAM
- ~400 pods
- 3 TiB / day



Thanos

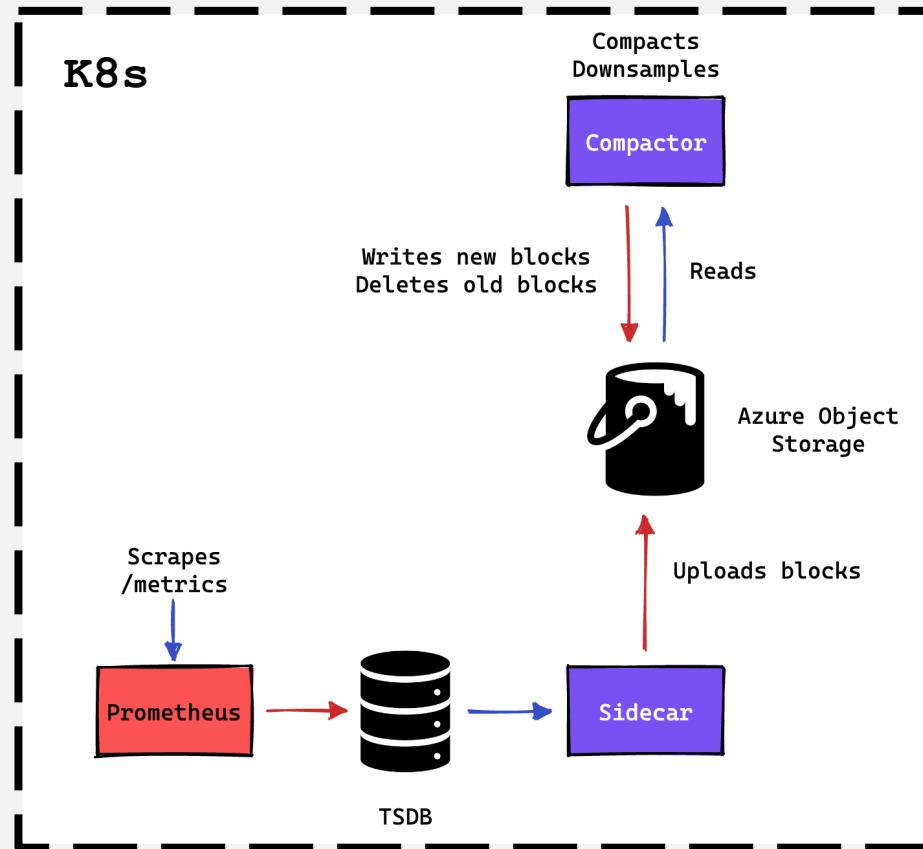


**~110 million metrics,
with bursts as high as
240m metrics**

Thanos on the cluster

Thanos Classic architecture

Thanos Sidecar: plugs Prometheus into the StoreAPI, periodically flushes blocks to object storage

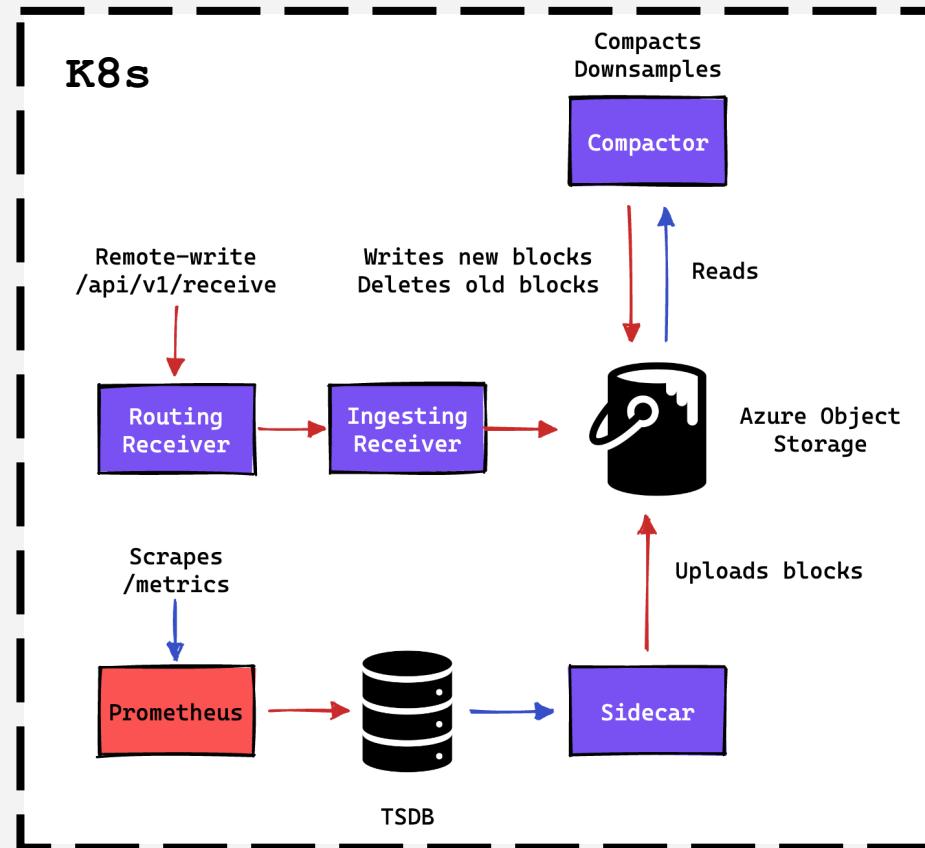


Thanos on the cluster

Remote-write Architecture

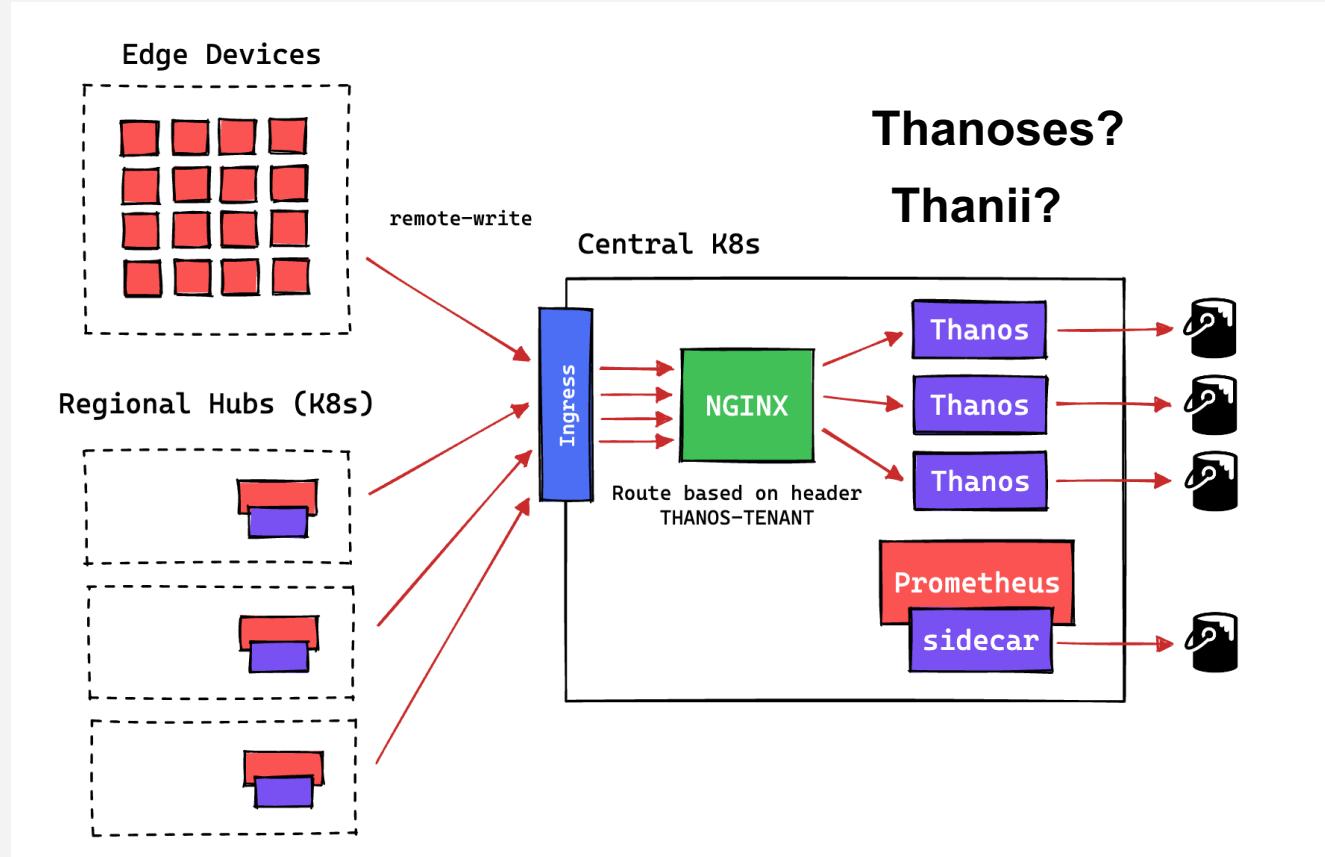
Routing receivers: process remote-write requests

Ingesting receivers: write local TSDB blocks with periodic flush to object store



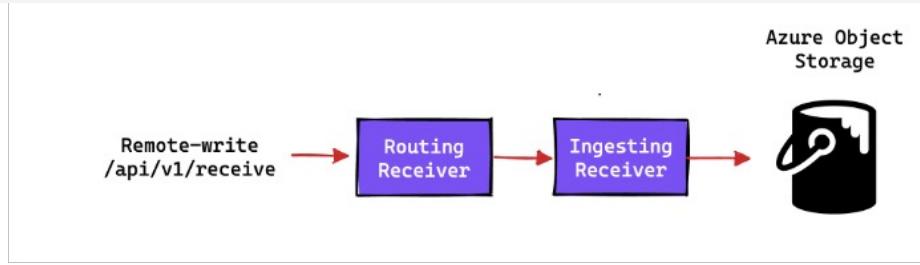
Thanos: Write-path

- Everything written centrally (compliance)
- Client certificate authentication for external requests
- NGINX gateway maps Thanos tenants to the correct ingestion pipeline



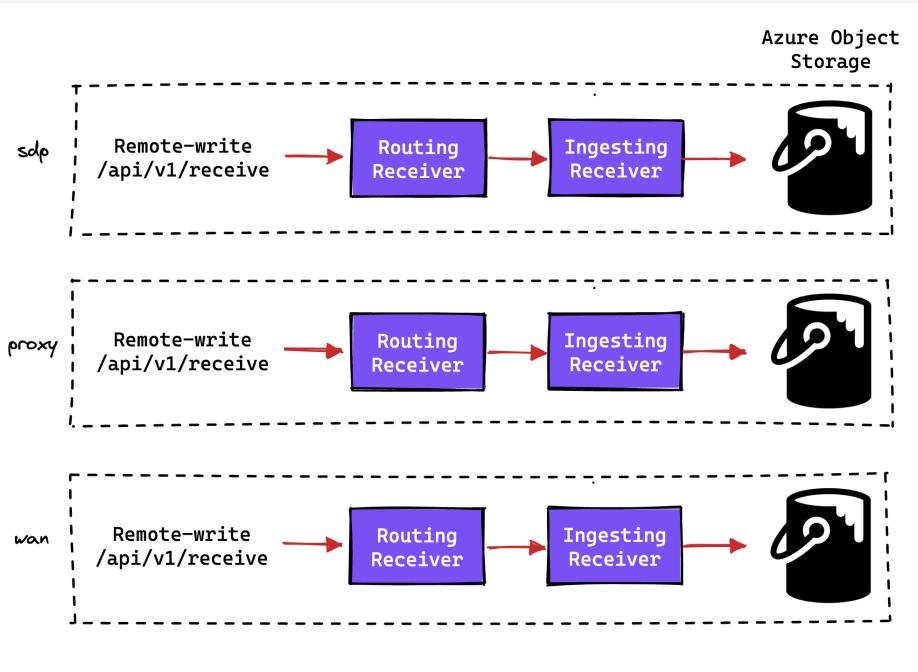
Thanos: Tenancy

What do we mean by a Tenant?



Thanos: Tenancy

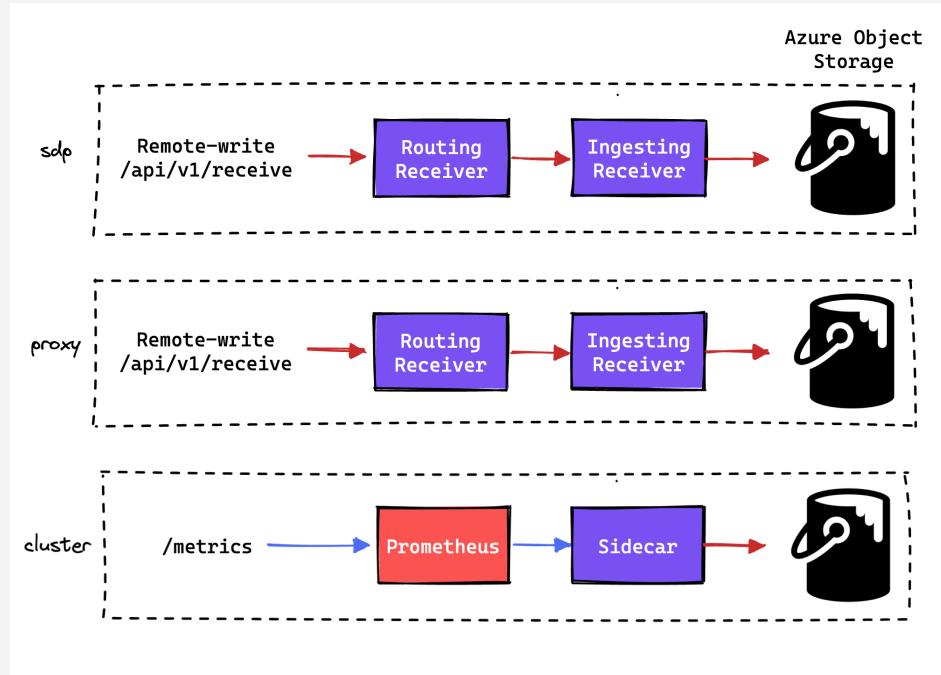
What do we mean by a Tenant?



- **Hard tenancy => one dedicated pipeline**
- **Each tenant maps to one service team**
- **Why? Mitigate the effects of unexpected *cardinality explosion***

Thanos: Tenancy

Concept extends to account for on-cluster metrics

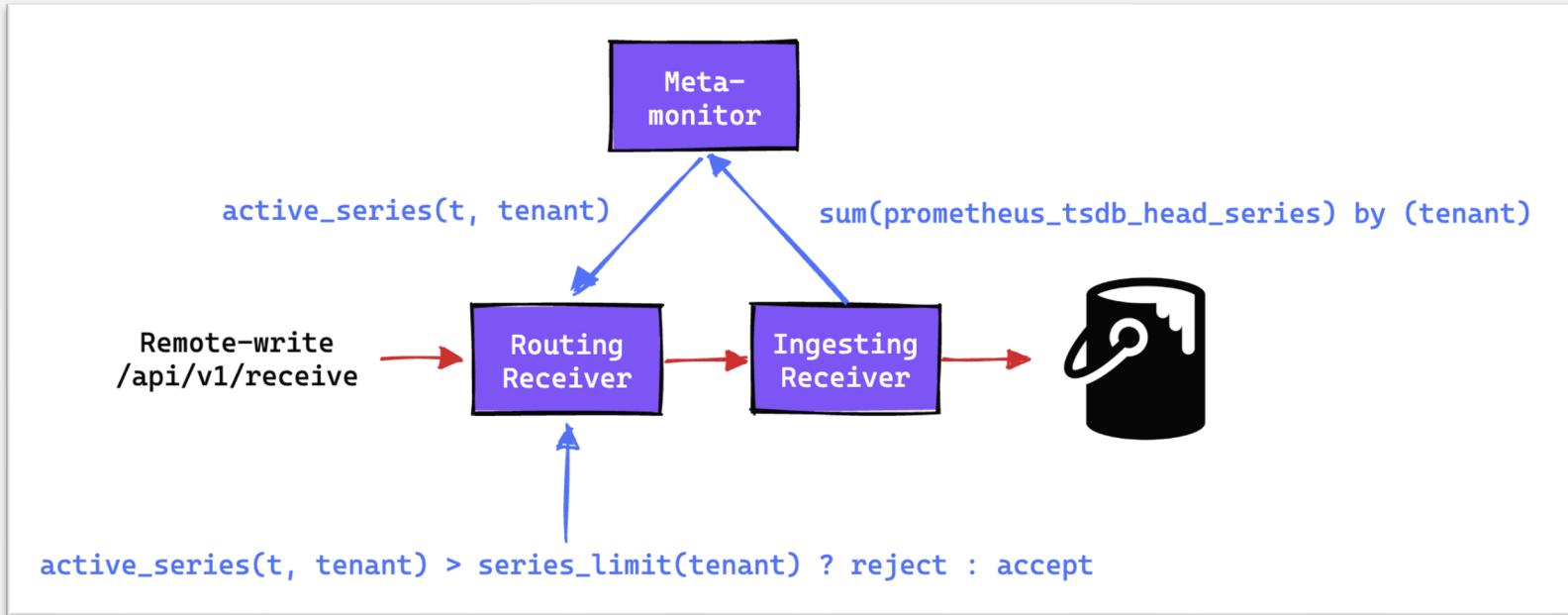


- **Cluster metrics belong to their own tenant with a dedicated storage bucket**

Cardinality Explosion: mitigation

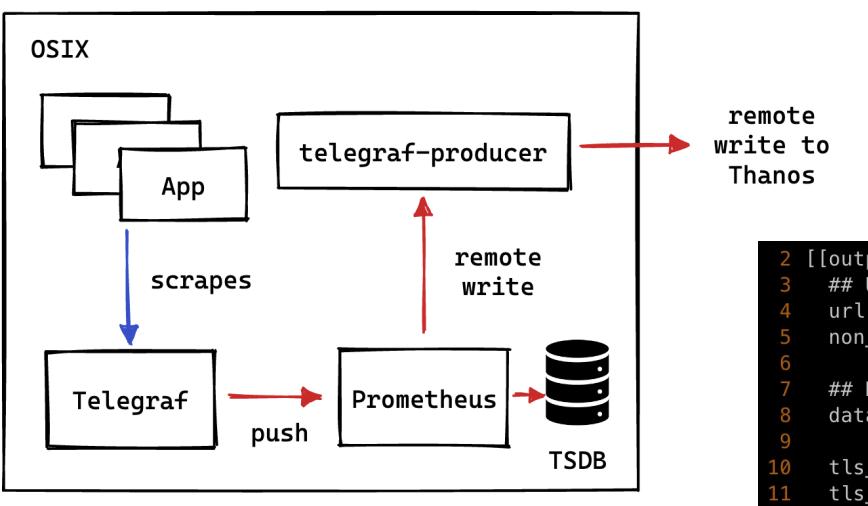
Active series limiting (available since v0.29.0)

Additional protection against 🔥 in the ingesting receivers



Collecting metrics at the Edge

Metrics mapped to tenants on edge device



Prometheus Config

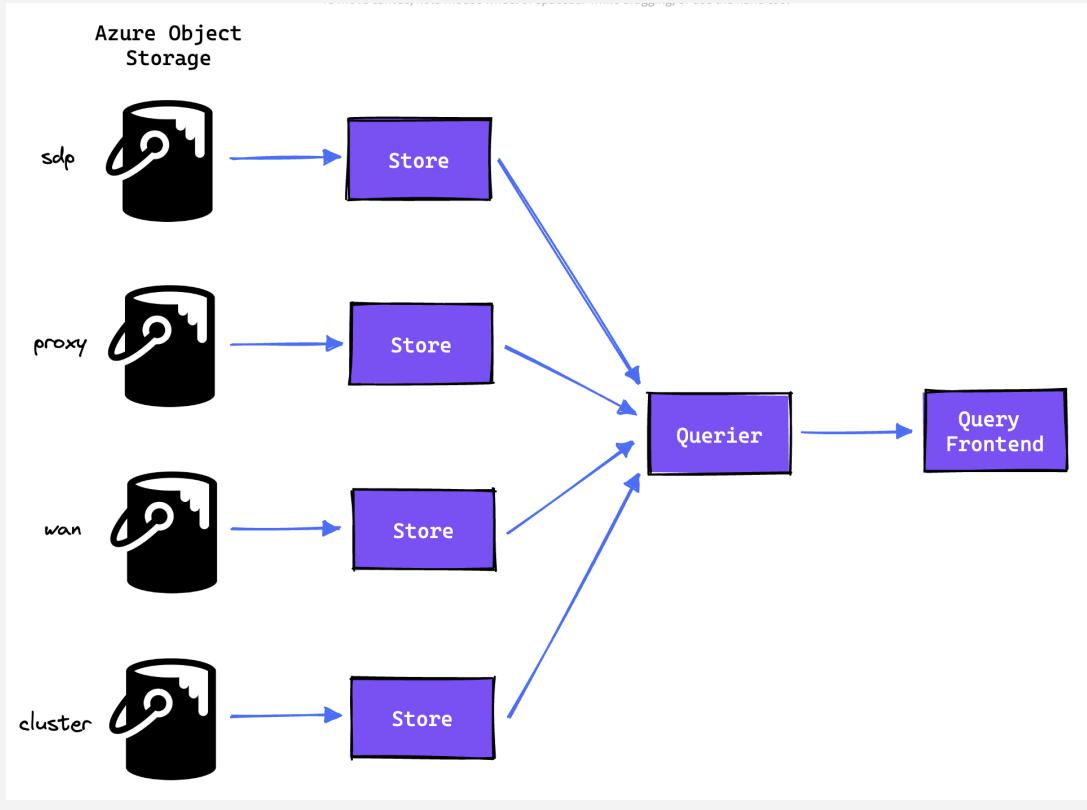
```
16      metric_relabel_configs:  
17        - source_labels: ['__name__']  
18          regex: '(?:tunnel|tunnel_interface|tmon2).*'  
19          replacement: 'wan'  
20          target_label: 'service_component'
```

Telegraf Remote-write Config

```
2 [[outputs.http]]  
3   ## URL is the address to send metrics to  
4   url = "https://thanos.test.osdp.open.ch/api/v1/receive"  
5   non_retryable_statuscodes = [409]  
6  
7   ## Data format to output.  
8   data_format = "prometheusremotewrite"  
9  
10  tls_ca = "/opt/OSAGtelegraf/etc/telegraf-kafka-producer.d/ca_server.pem"  
11  tls_cert = "/opt/OSAGtelegraf/etc/telegraf-kafka-producer.d/cert.pem"  
12  tls_key = "/opt/OSAGtelegraf/etc/telegraf-kafka-producer.d/key.pem"  
13  
14  [outputs.http.headers]  
15    Content-Type = "application/x-protobuf"  
16    Content-Encoding = "snappy"  
17    X-Prometheus-Remote-Write-Version = "0.1.0"  
18    THANOS-TENANT = "proxy"  
19  
20  [outputs.http.tagpass]  
21    service_component = ["proxy"] Match metrics with this label
```

Thanos: Read-path

Goal: global view of metrics, regardless of the source

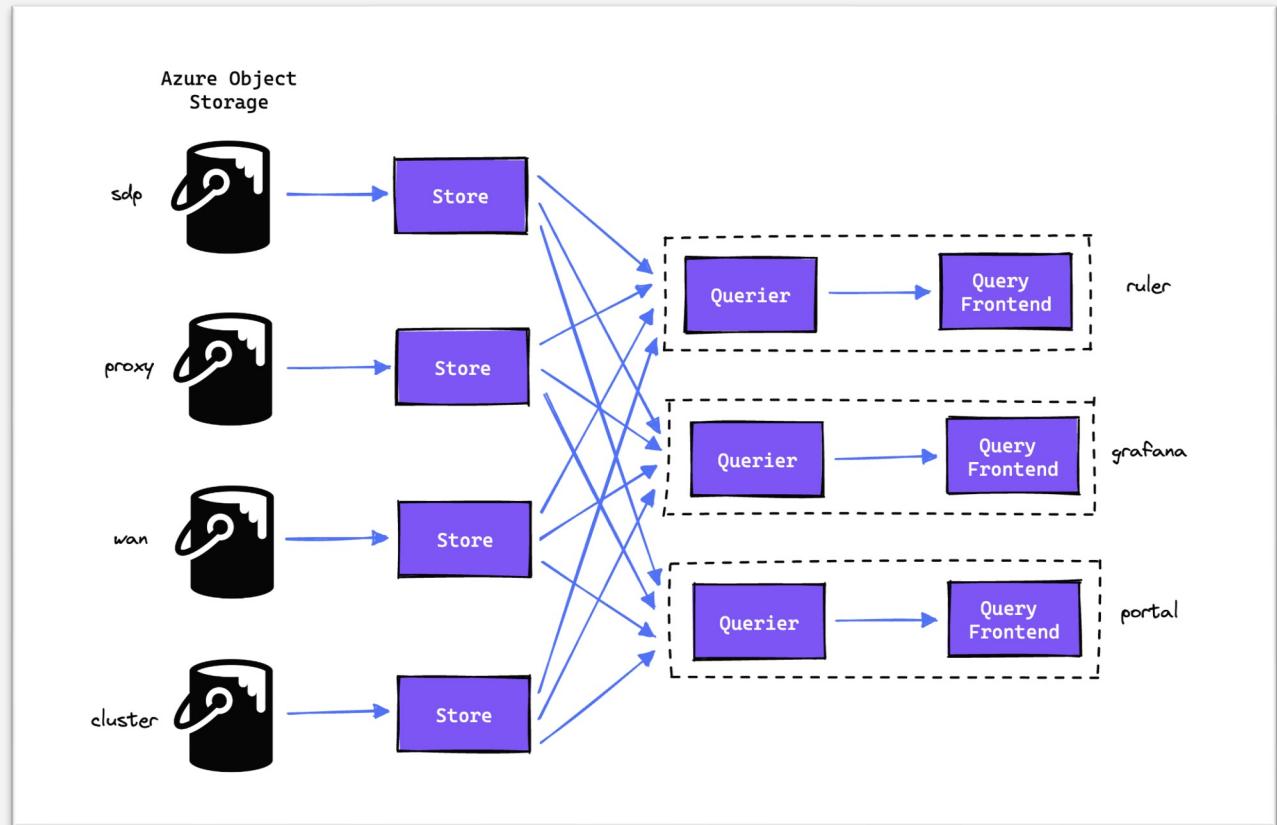


- **Store component interface to fetch data from object storage**
- **Naïve approach: plug everything into the same querier!**
- **We have many different users, with different priorities**
- **Can we prioritize query traffic?**

Thanos: Read-path

Querier quality of service

- In this architecture, all query users have the same priority
- What if we want to prioritize portal queries?

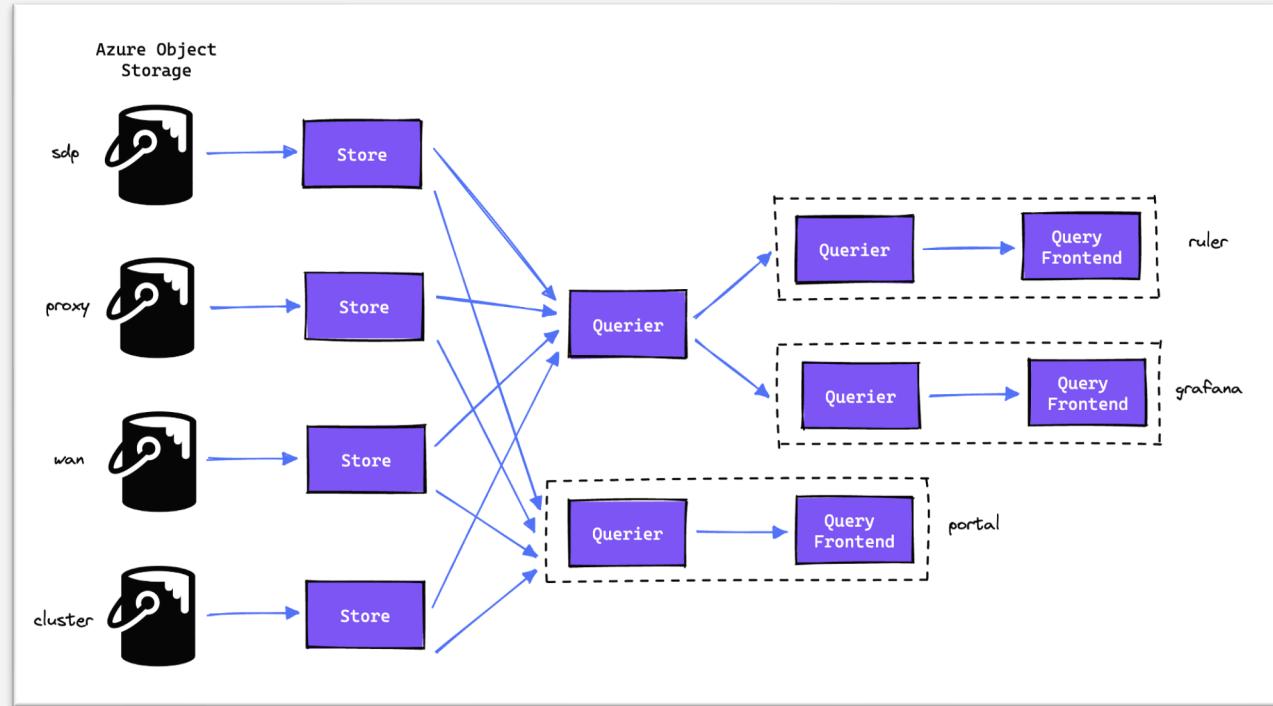


[1] <https://grafana.com/docs/loki/next/lids/0003-queryfairnessinscheduler>

Thanos: Read-path

Querier quality of service

- **Queriers at the same *level* have equal priorities**
- **By shifting the *grafana* and *ruler* users up a level, *portal* queries now have higher priority [1]**



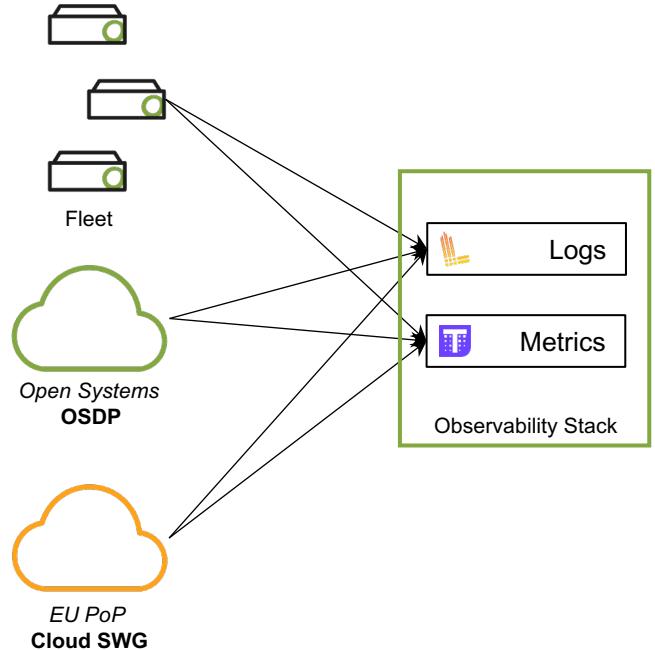
[1] <https://grafana.com/docs/loki/next/lids/0003-queryfairnessinscheduler>

Reflect

What has the new metrics pipeline given us?

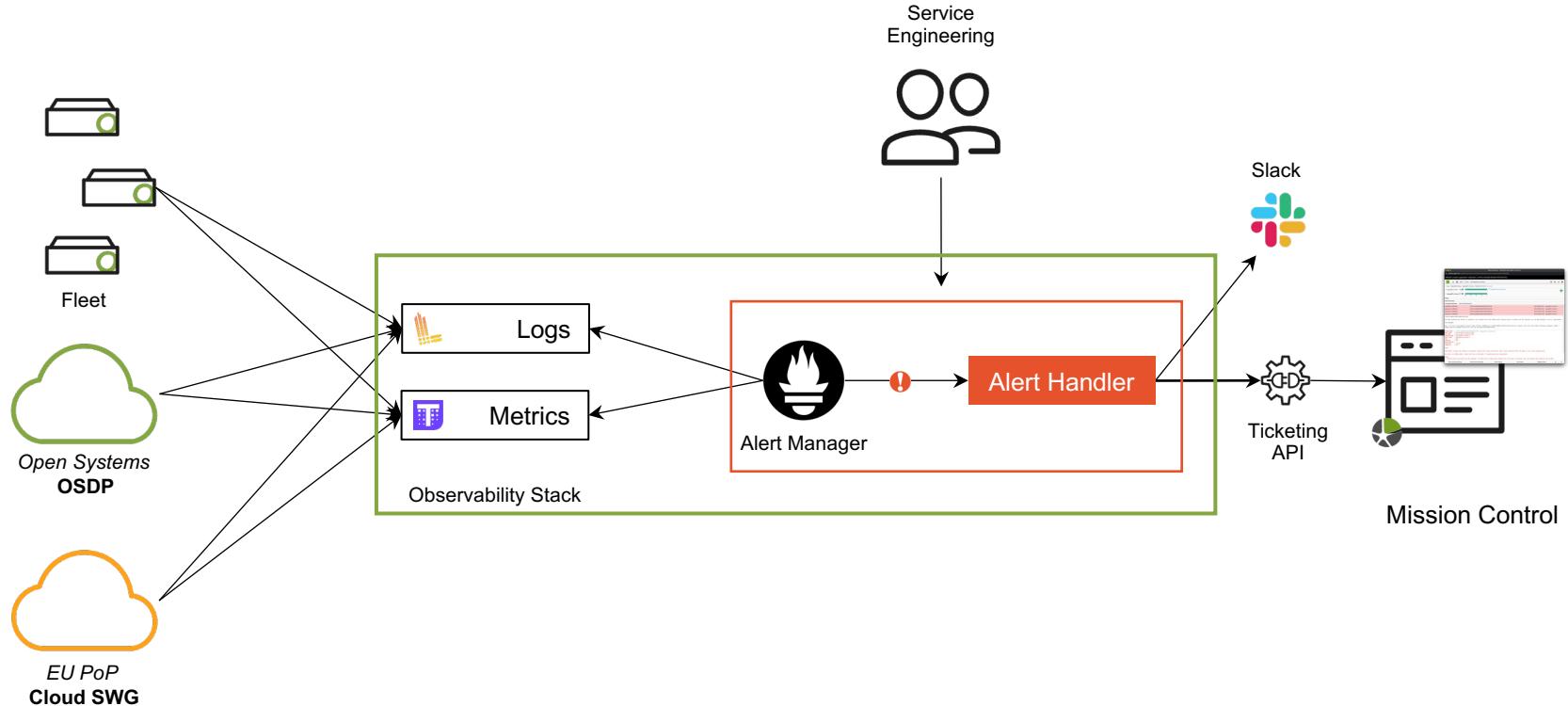
- Environment agnostic ingestion pipeline ✓
- Global metrics query view ✓
- How can we put this to use to achieve our other goals?
 - Smarter alerts (metrics, not logs)
 - Smarter alert *handling*

How do we get the relevant alerts into Mission Control?



Implication
Engine?

The unified alerting pipeline



The unified alerting pipeline

The Alert Handler

- **Replacement for the black box “implication engine”**
- **Goal is to group alerts, correlate related alerts, and enrich alerts with additional data before creating MC tickets**
- **Designed as a K8s operator that service teams can configure independently**



Conclusion & Outlook

- **The scalability dragon is at the gates, but we are working hard on implementing our Action Plan**
- **Core telemetry (logs/metrics) backends are in production and ready to roll**
- **We now go full steam ahead with making use of these data effectively, to provide real value**



100K 

Cheers to 100k+ 🍻



We are hiring!

Interested? Come say hi during the conference 😊



Site Reliability Engineer (80% - 100%, Remote)

Open Systems

Zurich, Zurich, Switzerland



Promoted



(Senior) Platform Engineer

Open Systems

Zurich, Zurich, Switzerland



2 weeks ago · 7 applicants



Platform Engineer

Open Systems

Zurich, Zurich, Switzerland



Promoted



Senior Software Systems Engineer (80-100%)

Open Systems

Zurich, Zurich, Switzerland



1 day ago

More information and job specs on [LinkedIn](#)

The background of the slide features a subtle, abstract design composed of numerous thin, light-orange lines forming a series of undulating waves across the entire frame.

Thanks for your attention!
Questions?

<https://verejoel.com>

<https://github.com/verejoel>