

Healthy On-call: How To Get Your Beauty Sleep



JCON Slovenia 2024

About me



Aljaž Blažej

Tech lead and manager at Outbrain

Achieved **incident reduction of 80%**

Load increased by **10x**

Outbrain (Zemanta)

- Ad tech
- **8 million** RPS
- 5 data centers
- **600+** Java microservices







Possible problems

- A sudden **spike** in traffic
- **Hardware** fault
- **Network** partition
- **VPN** problems
- **Unexpected data** that your application can't handle
- A **bug** in the code
- **Int overflow** in one of the tables
- **Memory leak**
- **External services** you depend on down
- **DDOS attack**
- Data center **fire**
- Ship hits a network cable under the bridge

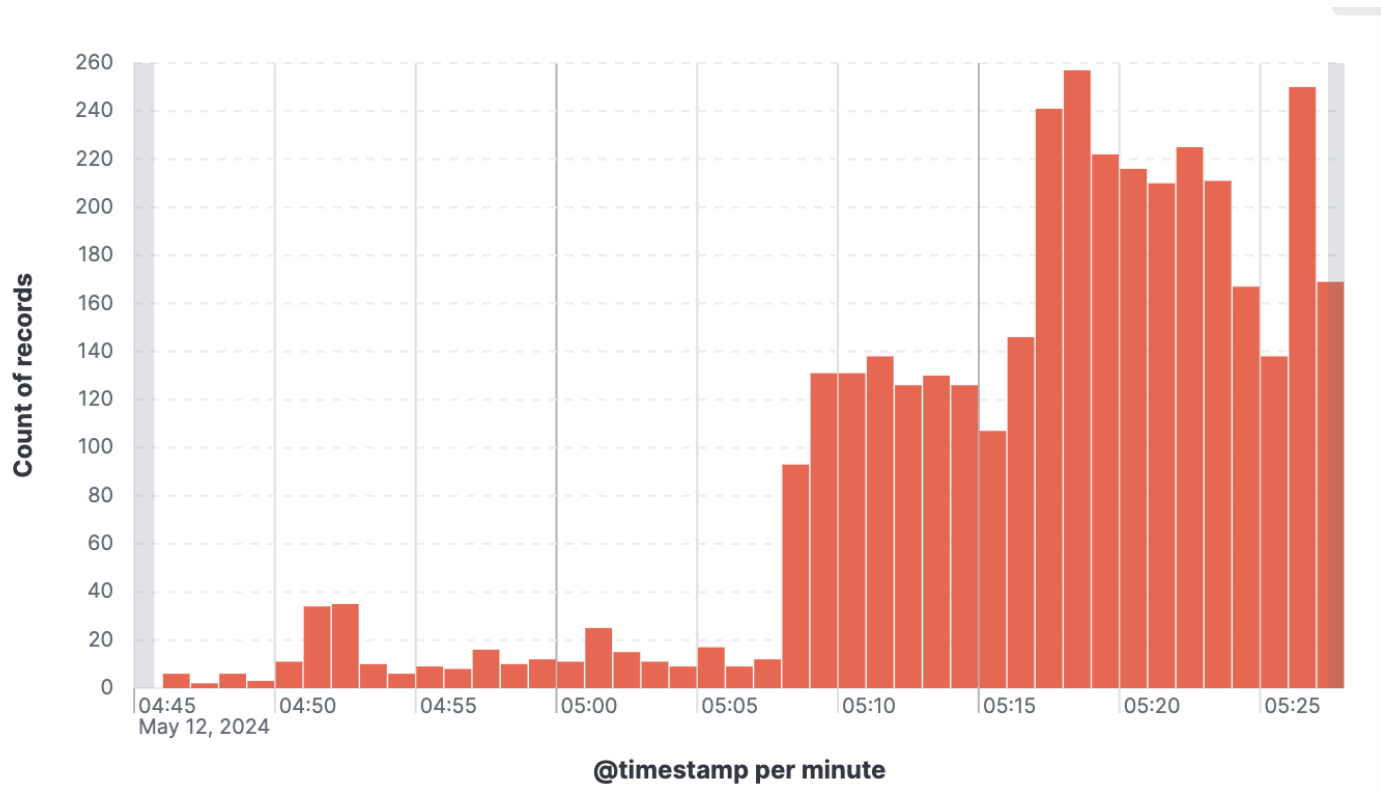
Possible problems

- A sudden **spike** in traffic
- **Hardware** fault
- **Network** partition
- **VPN** problems
- **Unexpected data** that your application can't handle
- A **bug** in the code
- **Int overflow** in one of the tables
- **Memory leak**
- **External services** you depend on down
- **DDOS attack**
- Data center **fire**
- Ship hits a network cable under the bridge



Reality

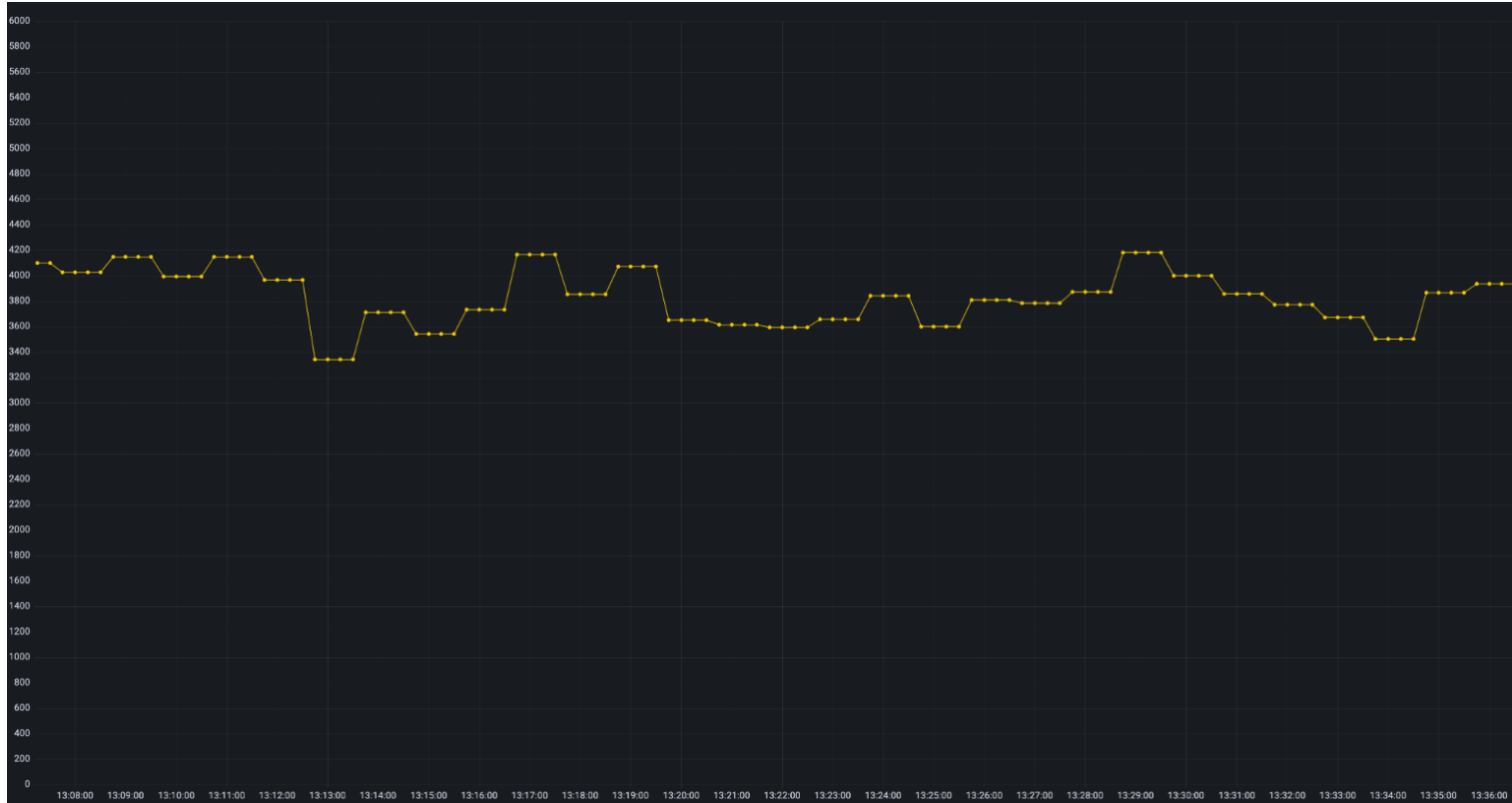
- Murphy's law
- **Nobody** will notify you
- It's **not** a **rare** event



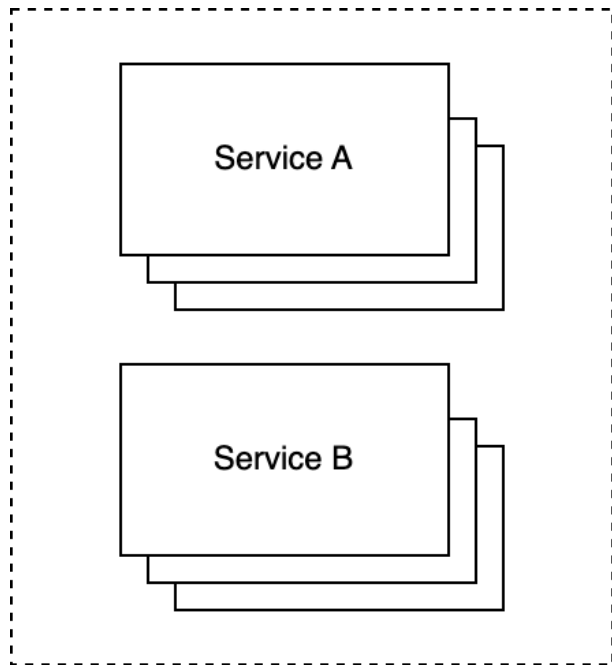
Blameless post mortem / take in



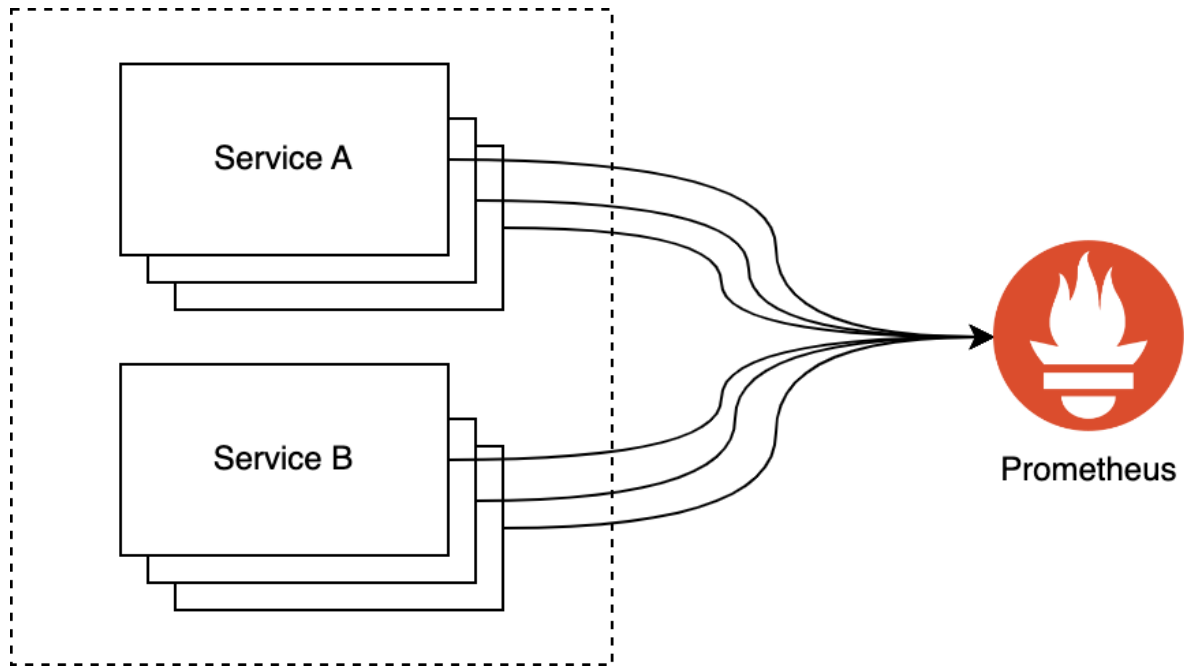
Metrics



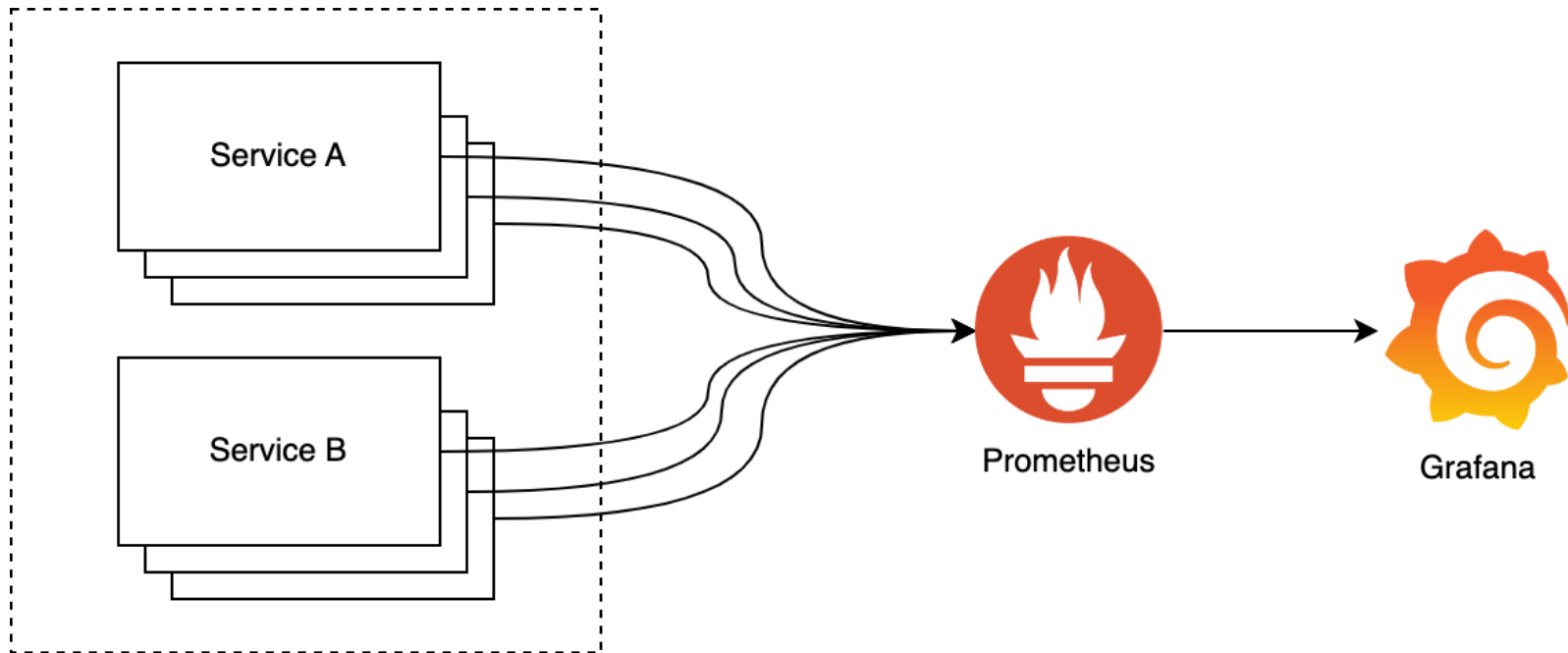
Example observability stack



Example observability stack



Example observability stack



693 K

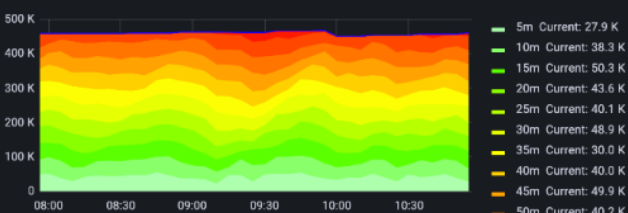
Expected full sync cycle

23.4 min

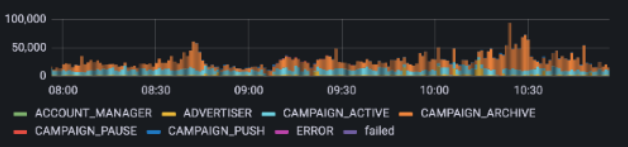
Average full sync cycle of last ...

56.0 min

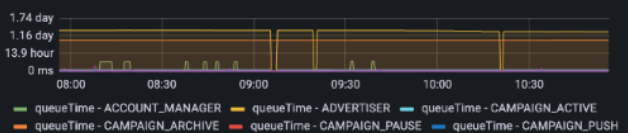
Active campaign sync age



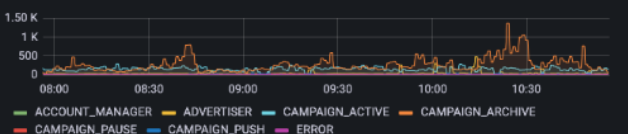
Neo syncs per minute



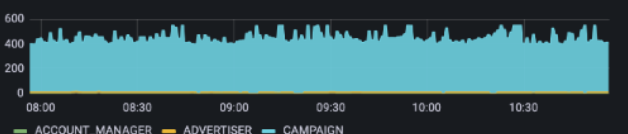
Campaign queue delay



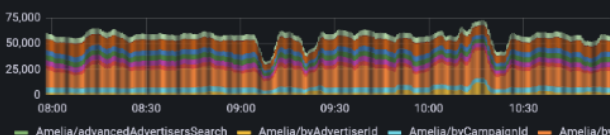
sync counters - success



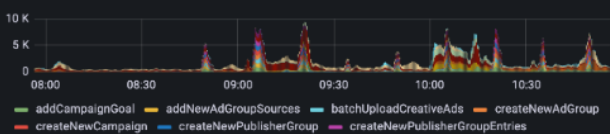
Syncs in progress



Outbrain API calls per minute



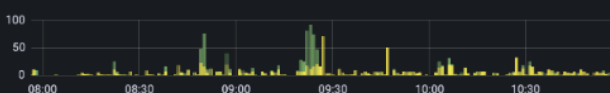
z1 update calls



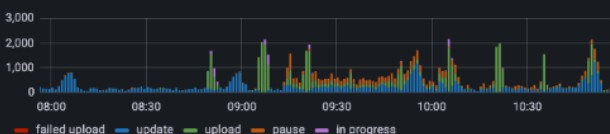
Syncs in Progress



Link logs



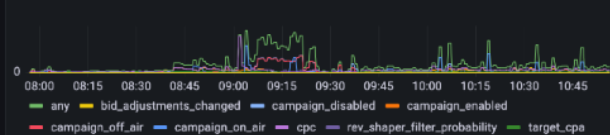
Ads Sync



Queue Sizes



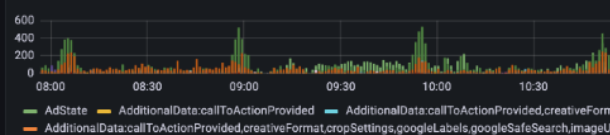
Push notifications processed by Link per minute



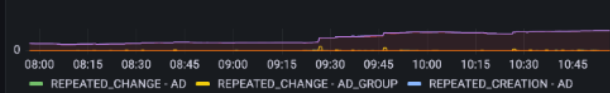
Delay between getting a push to adding campaign to queue



Ad property change



Link sync anomalies



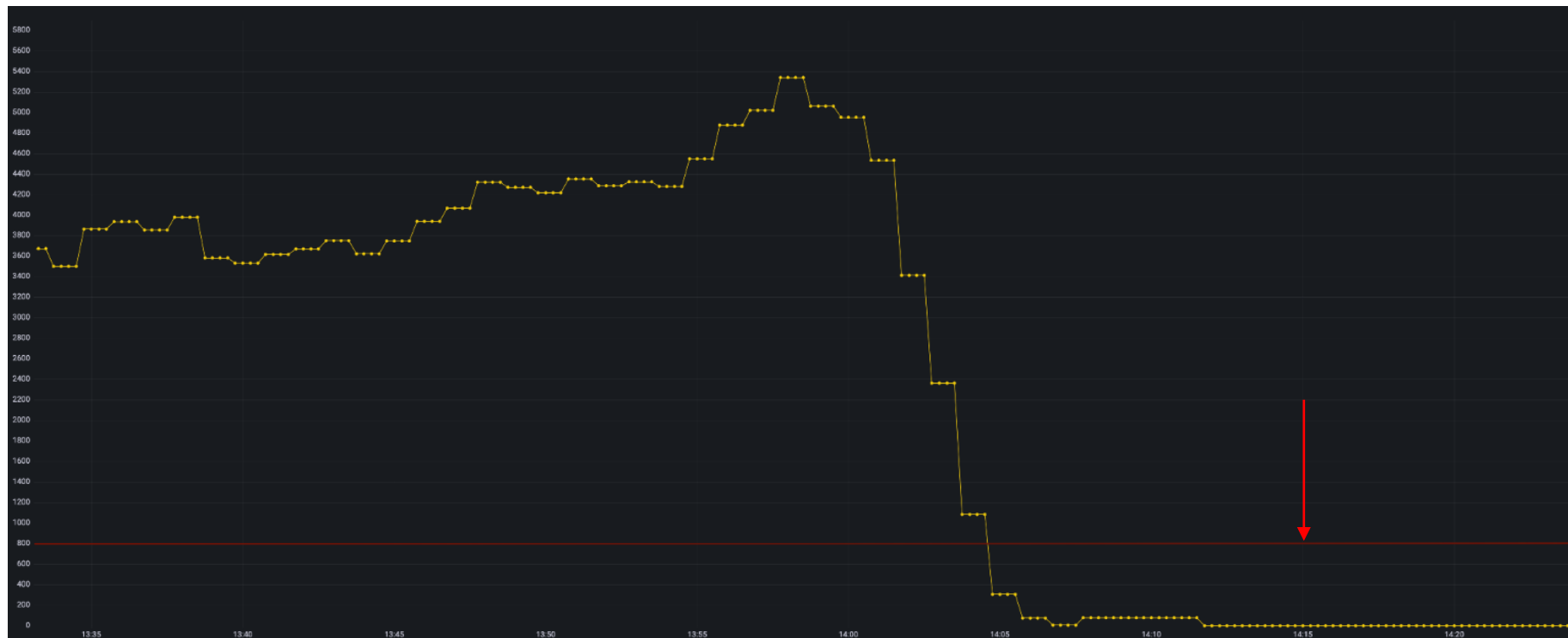
Push queue delay



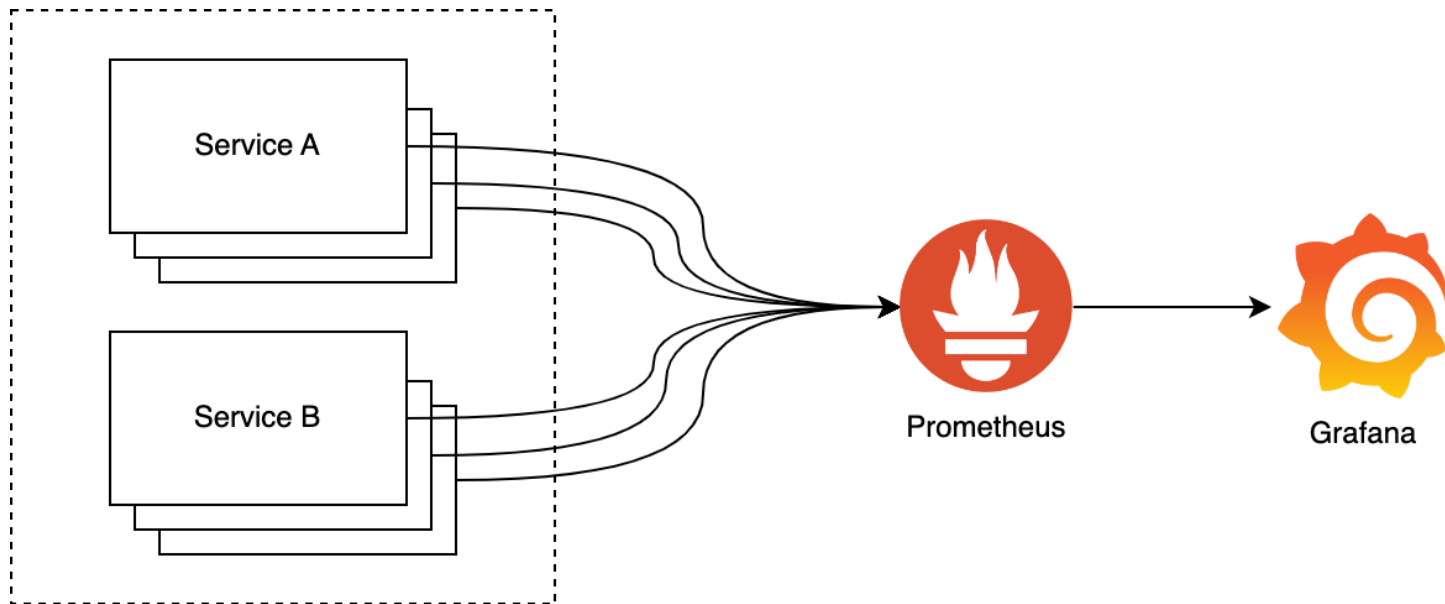
Delay between getting the push and syncing the campaign



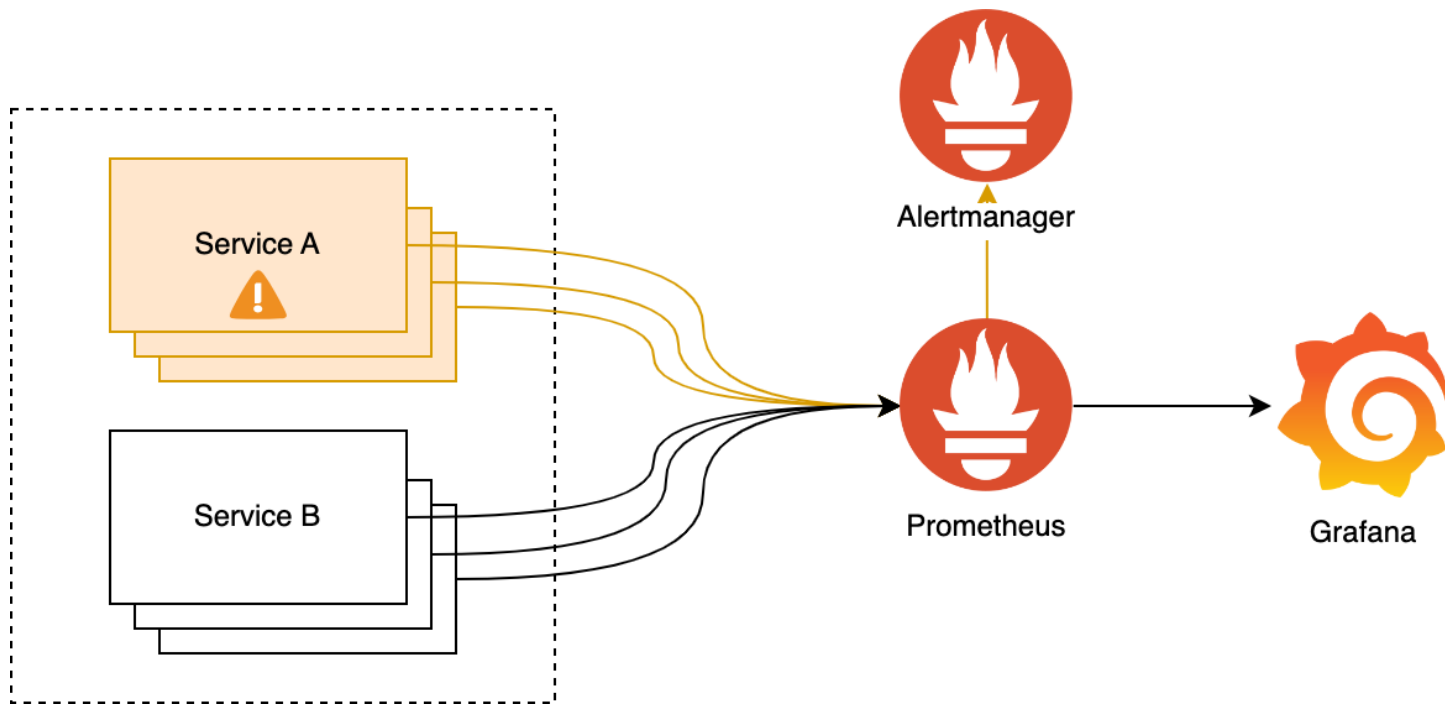
Alerting



Example observability stack



Example observability stack



Critical alert

- Not all critical
- Send an **email**?
- Bot post in **slack**?
- **Who** to notify? The whole team?

On-Call Alerting Tools


- PagerDuty
- Opsgenie
- Splunk
- XMatters
- Grafana

PagerDuty example







When a high-urgency incident is assigned to me...

 **Immediately** after it's assigned to me, push notify me on **iPhone**



 **1 minute** after it's assigned to me, push notify me on **iPhone**




 **1 minute** after it's assigned to me, sms me at **+386 xxxxxxxx (Mobile)**



 **3 minutes** after it's assigned to me, phone me at **+386 xxxxxxxx (Mobile)**



 **4 minutes** after it's assigned to me, phone me at **+386 xxxxxxxx (Mobile)**



[+ Add Notification Rule](#)

Prog Supply Shift 1

Time Zone: Ljubljana

Today



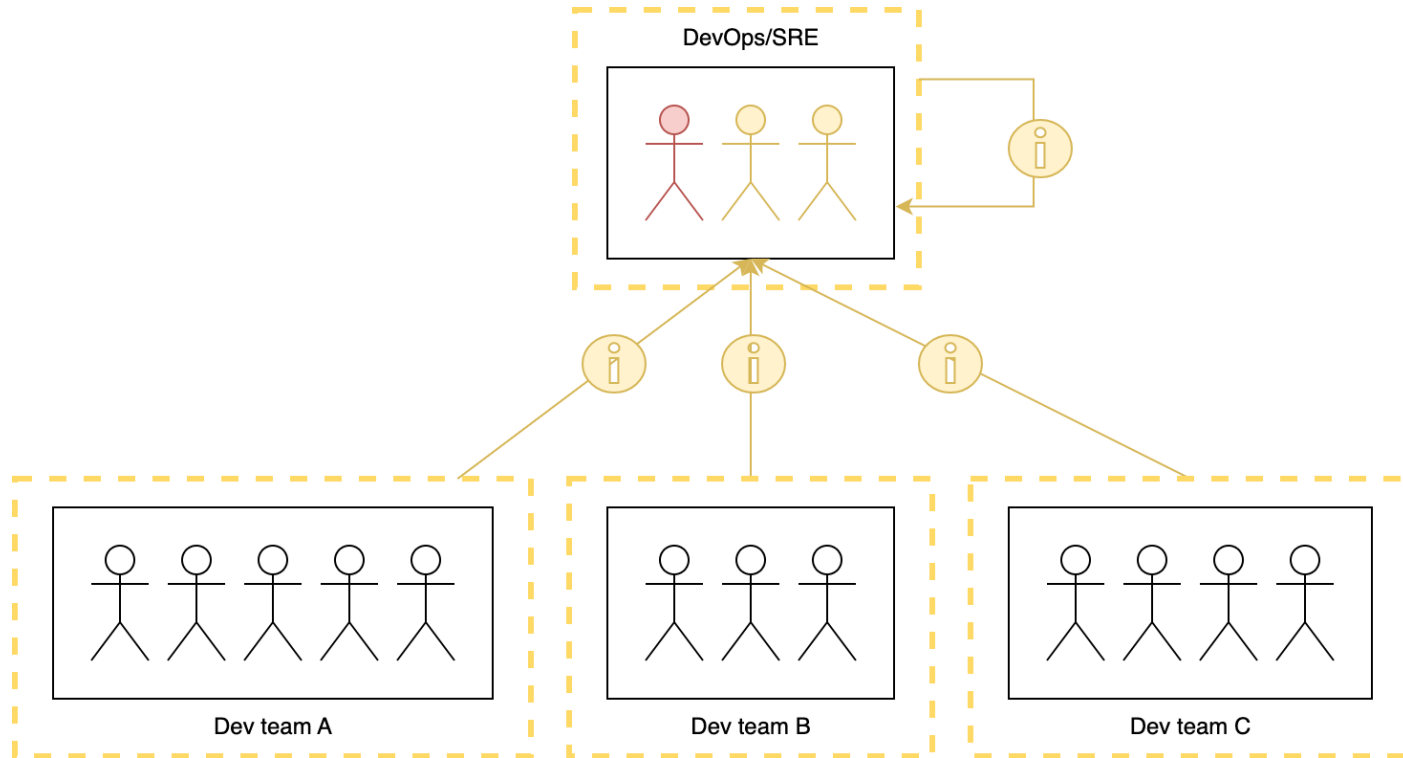
June 2024

Timeline View

Calendar View

SUN	MON	TUE	WED	THU	FRI	SAT
26	27	28	29	30	31	1
Mark Volk	Jan Zeman					
2	3	4	5	6	7	8
Jan Zeman	Aljaz Blazej					
9	10	11	12	13	14	15
Aljaz Blazej	Domen Kek					
16	17	18	19	20	21	22
Domen Kek	Jernej Žust					
23	24	25	26	27	28	29
Jernej Žust	Matjaz Brečko					
30	1	2	3	4	5	6
Matjaz Brečko	Klemen Đukić					

Dedicated team

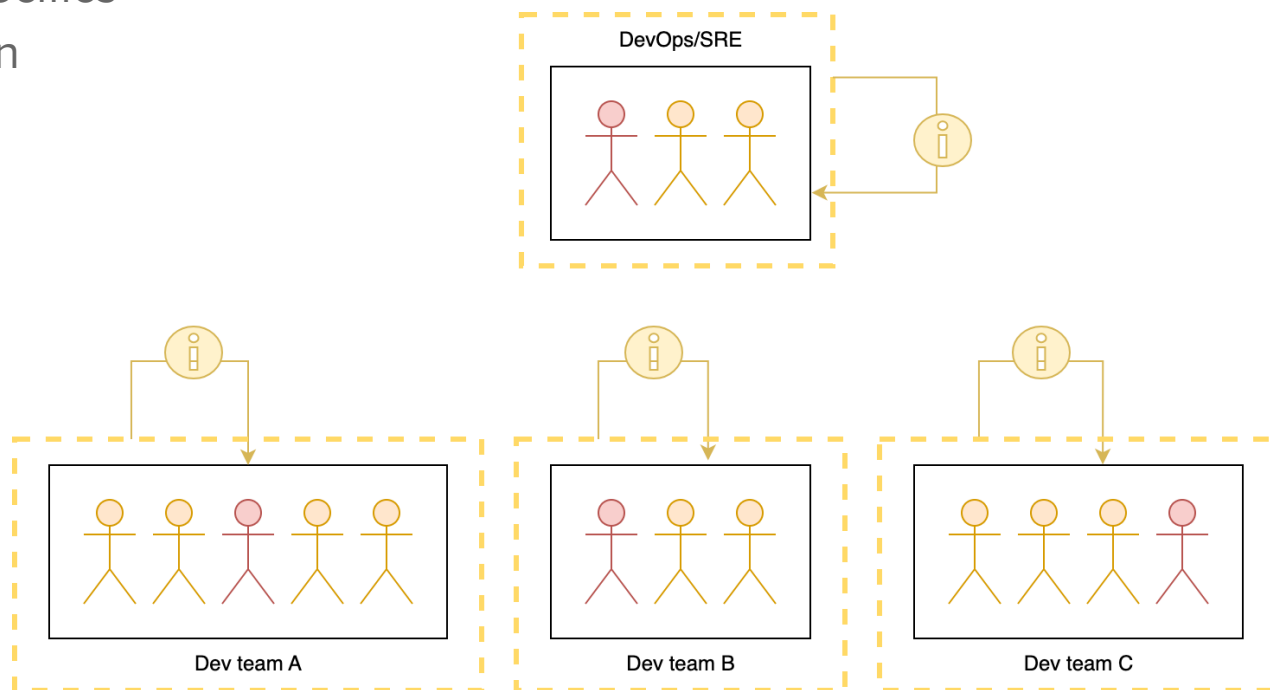


Problems

- Unfamiliarity with systems
- No context
- Conflicting interests

Each team handles their own incidents

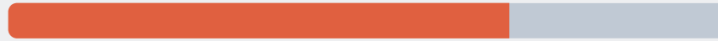
- Knowledge of specifics
- Distributes burden
- Ownership





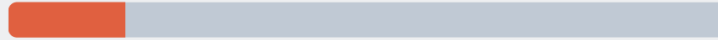
Which best describes how on-call is structured?

Each team is responsible for their own on-call



69% 137 respondents

One dedicated team is responsible for all on-call



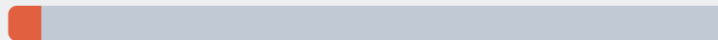
16% 32 respondents

Multiple dedicated teams are responsible for all on-call



10% 19 respondents

Other

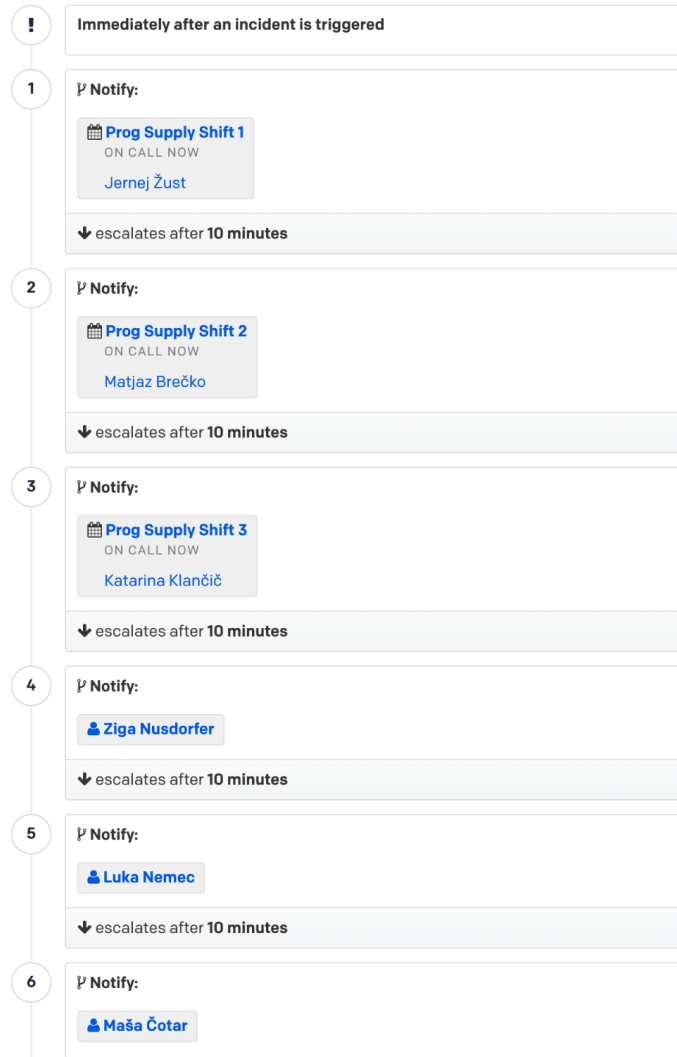


5% 9 respondents

Escalations

- Unresponsiveness
- Acknowledging alerts
- Help needed

Advice: test alerts on Mondays



Prog Supply Shift 1

First line on-call schedule

AB Aljaž Blažej Now

J 2 Escalation Policies
 🗨 Programmatic Supply Eng

Export ▾ ...

Fri 14	Sat 15	Sun 16	Mon 17	Tue 18	Wed 19	Thu 20	Fri 21	Sat 22	Sun 23	Mon 24	Tue 25	Wed 26	Thu 27
Mark Volk			Matjaz Brečko							Domen Kek			

Prog Supply Shift 2

Escalation on-call schedule

JZ Jan Zeman Now

J rnd-zemanta-progsupply-policy
 🗨 Programmatic Supply Eng

Export ▾ ...

Fri 14	Sat 15	Sun 16	Mon 17	Tue 18	Wed 19	Thu 20	Fri 21	Sat 22	Sun 23	Mon 24	Tue 25	Wed 26	Thu 27
Matjaz Brečko			Domen Kek							Jernej Žust			

Prog Supply Shift 3

KK Katarina Klančič Now

J rnd-zemanta-progsupply-policy
 🗨 Programmatic Supply Eng

Export ▾ ...

Fri 14	Sat 15	Sun 16	Mon 17	Tue 18	Wed 19	Thu 20	Fri 21	Sat 22	Sun 23	Mon 24	Tue 25	Wed 26	Thu 27
Katarina Klančič													



Robust alerting

- Good visibility
- Automatic alerts
- Alert delivery
- Schedule / rotation of on-call engineers
- Escalations / backups
- Structure

New problems

- Colleague wrote this service
- 3 am brain
- Where is the dashboard
- Onboarding new members

Playbook / Runbook

Pages / ... / Incident Playbooks

Edit Save for later Watching Share ...

Click processing issues

Created by Žiga Stopničnik, last modified by Aljaž Blažej on Dec 11, 2023

Summary	Rate of valid clicks dropped or invalid clicks increased, impacting our costs. Often a symptom of syncing delays.
Owner team	Brokers
Escalation	Depends, potentially Buyers, Solutions, Z1 group or Bidder group.

Every invalid click is revenue lost, and every missing valid click is a lost opportunity. If the rate is high, OEN bidding might need to be stopped.

Sati service
Sati is our validation service. It consumes clicks and impressions and does a lot of different checks in realtime (was CPC calculated correctly, are clicks coming from a campaign that is off air, wrong geo locations,...).

Clicks flow
OEN clicks first go to the Zemanta redirector (like non-OEN clicks) but are then redirected to Outbrain's redirector. None of the brokers' services are involved in the clicks processing. But our Sati service does consume clicks from a kafka (from outbrain's redirector) just to check if the clicks were valid and to do some additional validations.

There are 2 different conditions that can trigger this PD. Either we aren't getting clicks anymore (number of all clicks is lower than threshold) or the number of invalid clicks we are getting is higher than threshold. Check the [grafana metric](#) to see which one triggered the PD and only read the relevant section of the two below.

Rate of valid clicks drops below threshold

Check [grafana panel](#) linked to the PD.

[There is a label in the grafana panel which shows how many clicks are **validated**. Only the most simple checks on clicks are done synchronously when the click is consumed from kafka. All additional validations are done in a separate thread asynchronously. We call clicks that go through the additional validations "**validated**" clicks. Ideally, all clicks should be validated, but if sati has problems, some may not be. This is not a reason for the PD, it just means that we didn't do most checks on these clicks and we don't know if they are correct]

- If only the **Total OEN clicks** metric falls, but **Total Outbrain clicks** remain high, there is a problem on programmatic side:
 - check the [Rate1-prog-datasci-protops-pagerduty](#) channel if bidder is having issues. If there aren't any alerts there, check [bidder's grafana main dash](#). If you see any problems there, escalate to bidder.
 - try restarting [Sati service](#) if you see [kafka errors in logs](#).
 - escalate to Brokers
 - If it's not just a Sati issue but we actually aren't getting clicks, escalate the issue immediately to bidder
- If the **Total Outbrain clicks** metric falls dramatically (not just by the amount of OEN clicks), there is a problem in Outbrain clicks processing ([their grafana dashboard](#)).
 - check in [#rtd-buyers](#) and [#potential_issues](#) slack channels if there are issues with processing clicks
 - try restarting [Sati service](#) if you see [kafka errors in logs](#).
 - escalate to Brokers

Rate of invalid clicks above threshold

Sati doesn't decide which clicks are invalid, it already gets this information in the click when it consumes it from Outbrain's redirector. Outbrain's redirector makes this decision based on different conditions:

- if the click comes from a campaign, which is already out of budget or disabled. Type **STOPPED** in [grafana](#)
- if the click comes from a geo region which a campaign is not targeting. Type **GEO** in [grafana](#)
- something else is the reason. Type **OTHER** in [grafana](#)

You can see which type of clicks increased the most by looking at the top 3 panels in this [dashboard](#). Also look at this [dashboard](#), which gives you a more detailed type.

To assess the severity of the situation, you can also check how much money we are losing because of the failed clicks in this [dashboard](#). Increase the time range to see what the normal numbers look like.

These invalid clicks are also called spilled clicks and we can't charge them to the advertiser (so we pay all of the traffic needed to generate them out of pocket)

If most of the spilled clicks are of type STOPPED

This means we don't stop campaigns in bidder even though they are disabled in Amplify.

Campaigns are stopped through two flows:

Benefits

Reduced

- Stress
- Escalations
- MTTR
- human errors

Playbook content

- Alert **description** and **urgency**
- **Services** involved
- Relevant **links**
- **Team dependencies** and communication
- **Steps** to identify the problem
- **Action items** and **tooling**



Playbooks

- Playbook for each alert
- Benefits
- Content

Not healthy yet

- **Many alerts**
- **Terrible** experience
- **Unproductive** during on-call

Follow up

- **Prioritize**
- Assign a ticket
- **Update** on resolution
- **Post mortem** / take in
- Only **off hours** incidents

Reducing noise

- **False** and **noncritical** alerts
- **Many alerts** for one incident
- Clear **SLOs**
- Alerts unaligned with **other teams**

Monthly review

- **Accountability**
- **Acknowledging pain** and making **action items**
- **Knowledge** sharing
- Increased **confidence**

Joining teams

Double down on:

- **Playbooks**
- **Follow up**
- Aim for **8 - 12** on-call engineers

Monitor the on-call metrics

Over Time

[View report](#)

Total Incidents

470 ↑84



MTTA

1m 1s ↑20s



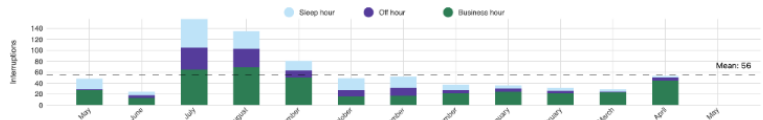
MTTR

58m ↑24m



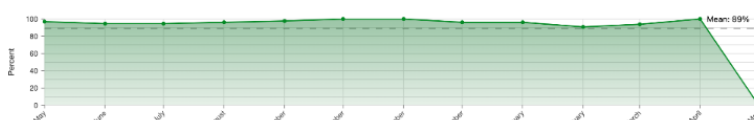
Interruptions

732 ↑138



Acknowledgment Rate %

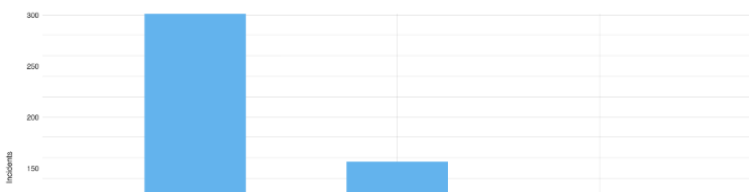
97% -



Services

[View report](#)

Incidents by Service

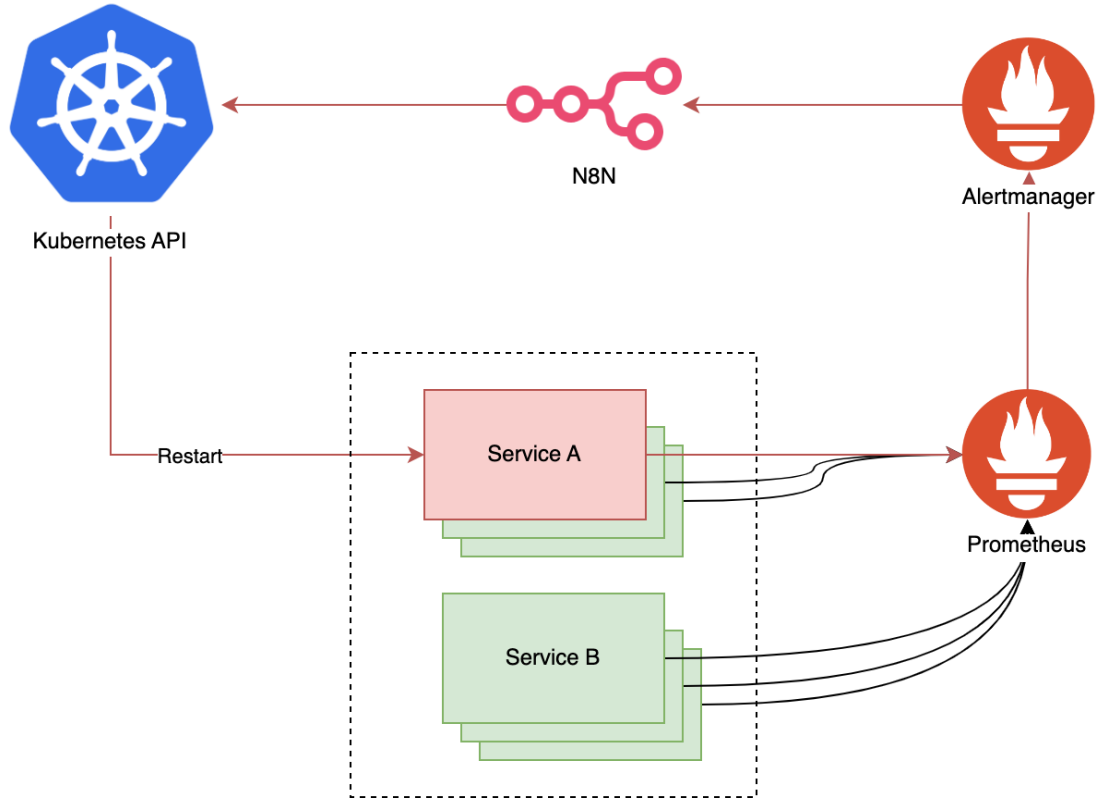


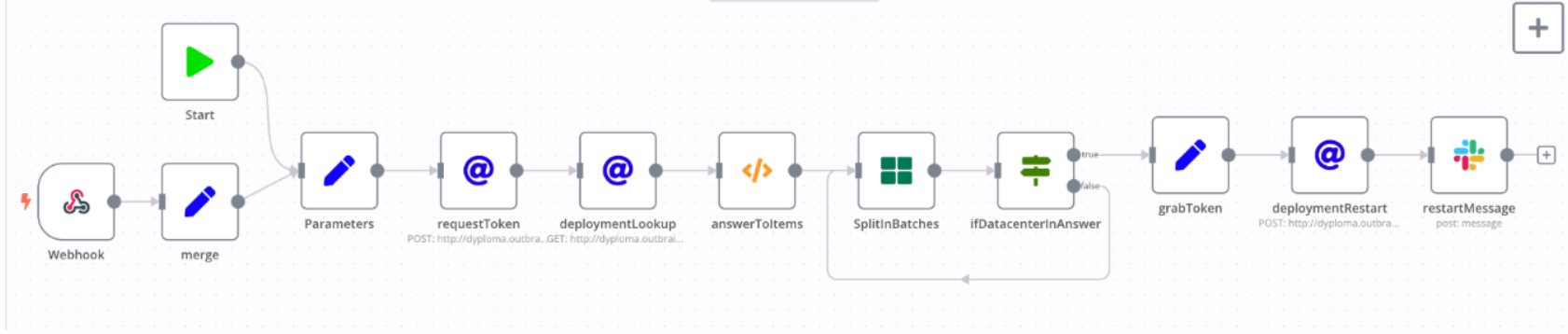
Uptime by Service

Service	Uptime %
Zemanta Z1 Campaigns	100%
prometheus-brokers	100%
Solutions Services	100%
Zemanta Operations	100%
Zemanta grafana alerts	100%

Playbook automation

- Separate automation system
- Same trigger
- Best effort







Using AI to help solve incidents


- Slack bot
- Slack chat history
- Company documentation
- Playbook data

Thread


 **PagerDuty** APP 8 hours ago

 [FIRING:1] Z1 - Ad group sync - queue size too large

Service: [Zemanta Z1 Campaigns](#)
Urgency: ↑ High

 Resolved by [Prometheus](#) | Today at 7:23 AM | via [AlertC](#)

4 replies

 **Devel** APP 8 hours ago

Information from ChatGPT that may help you




Note: Generated information might be incorrect or misleading

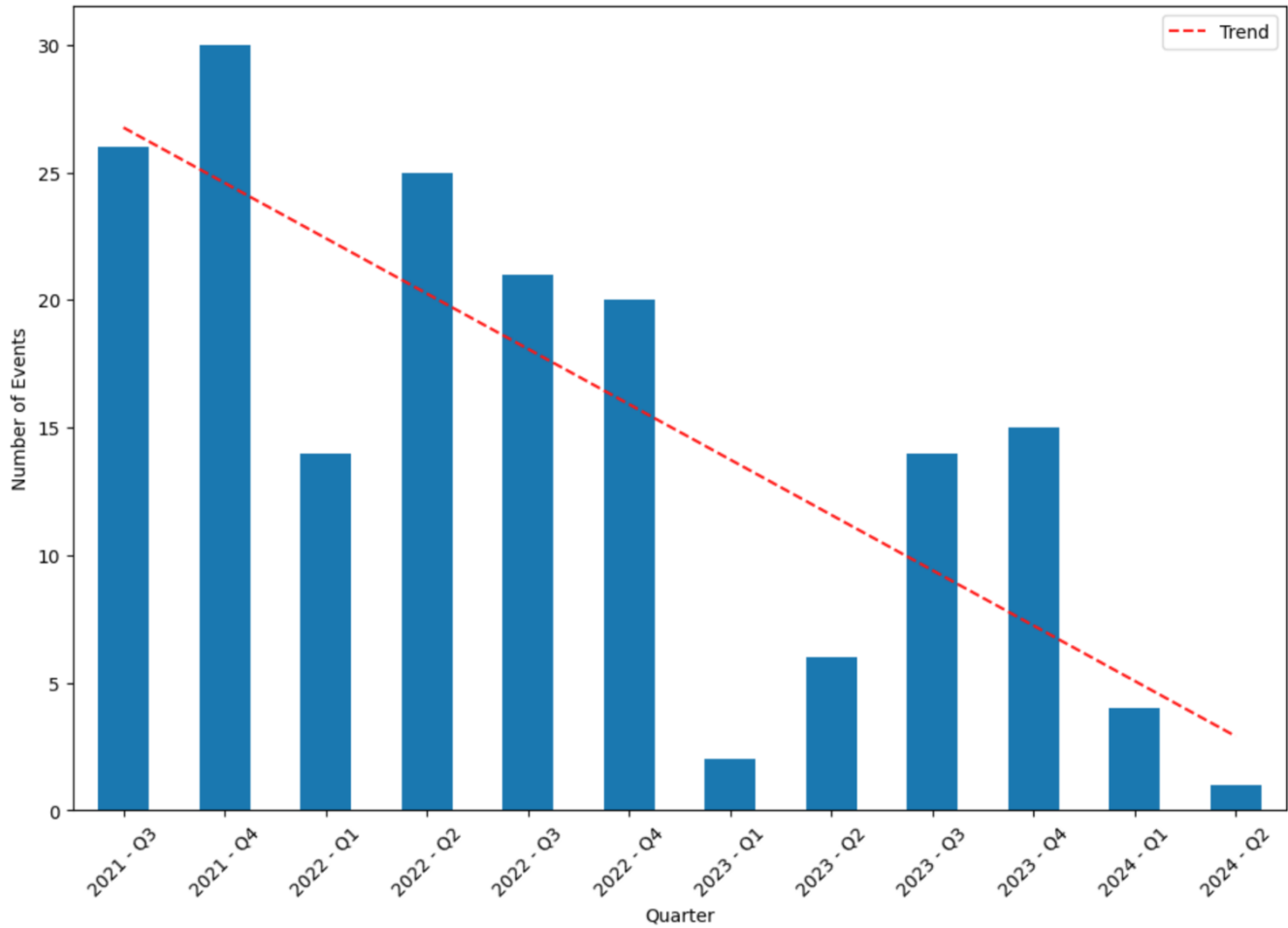
The alert for **Z1 - Ad group sync - queue size too large** indicates that the queue for syncing ad groups is larger than expected, which could lead to delays in processing. This should be addressed as soon as possible to avoid potential impact on campaign performance.

According to the playbook for **Z1** ...
[See more](#)

References: [Confluence page](#) | [Confluence page](#) | [Confluence page](#)

Was this info helpful? use the emojis or tell us in [#ask-platform-products](#)

 1  1 





Making it healthy

- Following up
- Reducing noise
- Monthly review
- Joining teams
- Monitoring on-call
- Playbook automation
- AI help

Recap

- Why we need on-call?
- Robust alerting
- Playbooks
- Making it healthy

Questions?