

Cutting Metrics Traffic, Cutting Costs



KubeCon



CloudNativeCon

Europe 2026

Who we are



Rodrigo Fior Kuntzer

Staff Site Reliability Engineer



kuntzer



rodrigorfk



Iris Dyrmishi

Senior Observability Engineer



iris-dyrmishi



IrisDyr

Cutting Metrics Traffic, Cutting Costs



KubeCon



CloudNativeCon

Europe 2026

Ask us questions and we will reply in the end



Who pays for your observability?

Spoiler: You do!



KubeCon



CloudNativeCon

Europe 2026

And probably more than you think.



Cloud costs are an engineering problem



KubeCon



CloudNativeCon

Europe 2026

Not just a FinOps problem.



The era of 'free cloud money' is over



Every architecture decision has a cost dimension



'That's a networking problem' — no longer acceptable



If you're on a public cloud, costs are your liability



50-65%

of cross-AZ traffic came from observability components



70-115%

observability network costs vs compute costs ratio

Why multi-AZ? Because you have no choice



KubeCon



CloudNativeCon

Europe 2026

High availability is just the start.

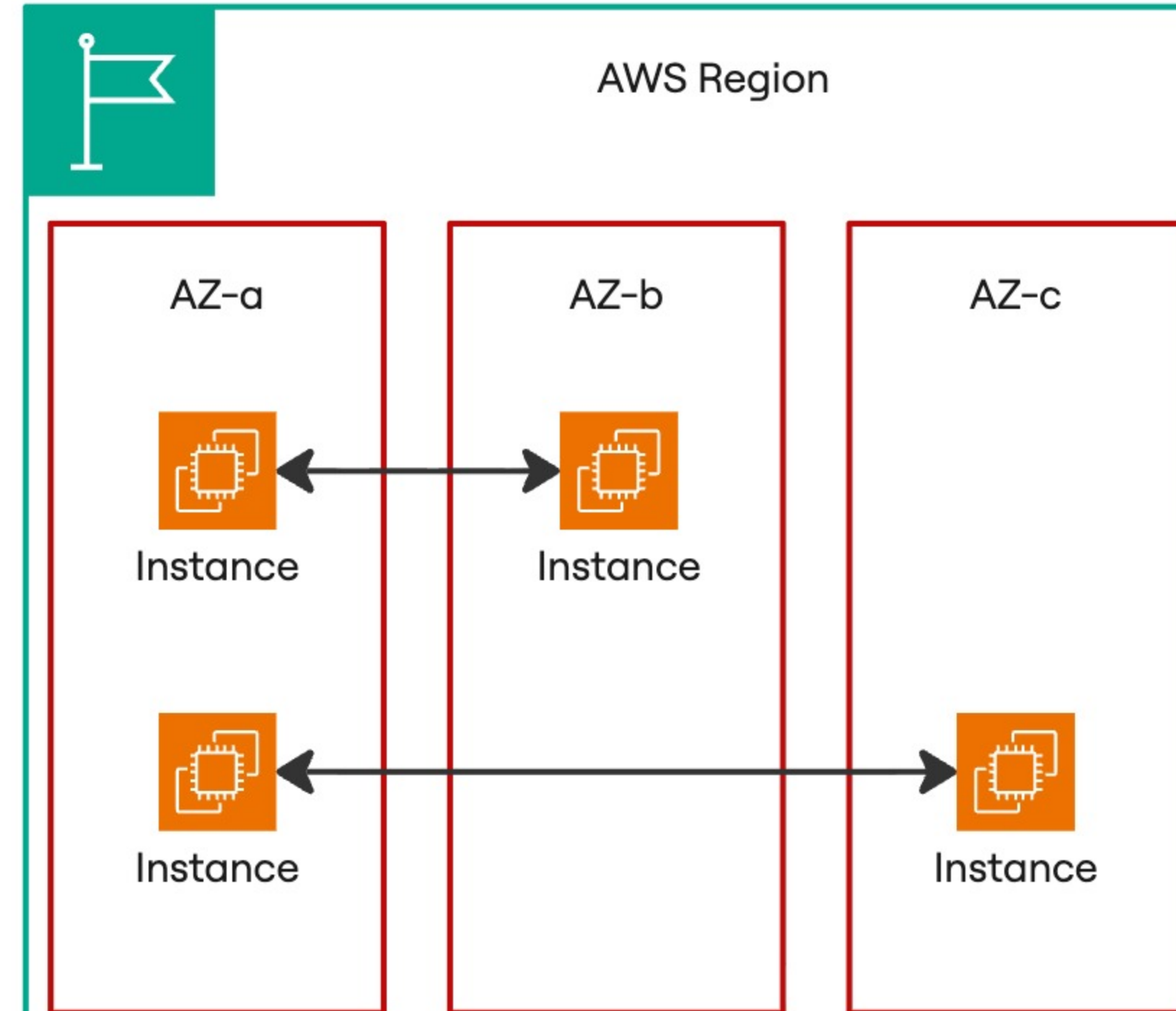
Why you need it

-  **Fault tolerance**
-  **Capacity constraints**
-  **K8s topology spread**
-  **Provider recommendations**

At scale: millions of metrics/min = hundreds of GB/day



It adds up fast — and it's invisible until it isn't.



✓ **Free within same AZ**

💰 **Cross-AZ (same region): \$0.01–0.02/GB each direction**

💰💰 **Cross Region: Even more expensive**

Observability: the heavyweight champion of traffic



KubeCon



CloudNativeCon

Europe 2026



Scraping across AZs



Ingestion pipelines



Query components

50-65%

of cross-AZ traffic

Everything else

35-50%



Paying to watch ourselves



~\$1M/year problem



Finding the culprit: cost attribution



KubeCon



CloudNativeCon

Europe 2026

From 'somewhere in the cluster' to 'that observability agent shard'



AWS Cost Explorer

Filter by EC2: Data Transfer - Inter AZ



Dedicated Node Pool

Isolate observability workloads



Cost Allocation Tags

Tag nodes for tracking



Grafana Beyla (eBPF)

Per-workload visibility



Finding the right workflow isn't easy — proper tooling is essential

Before you optimize: know your trade-offs



KubeCon



CloudNativeCon

Europe 2026

Complexity is the price of cost efficiency.

✓ Worth it if you have:

- High metric cardinality and volume
- Multi-AZ Kubernetes clusters
- Hundreds+ scrape targets per AZ
- Cross-AZ costs already visible on your bill

✗ Not worth it if:

- Small cluster, low cardinality
- Single-AZ or on-prem
- Operational simplicity is the priority
- Cross-AZ cost is negligible

This talk gives you the blueprint — you decide if it fits your context.



Observability Network Topologies

Metrics collection: pull vs push



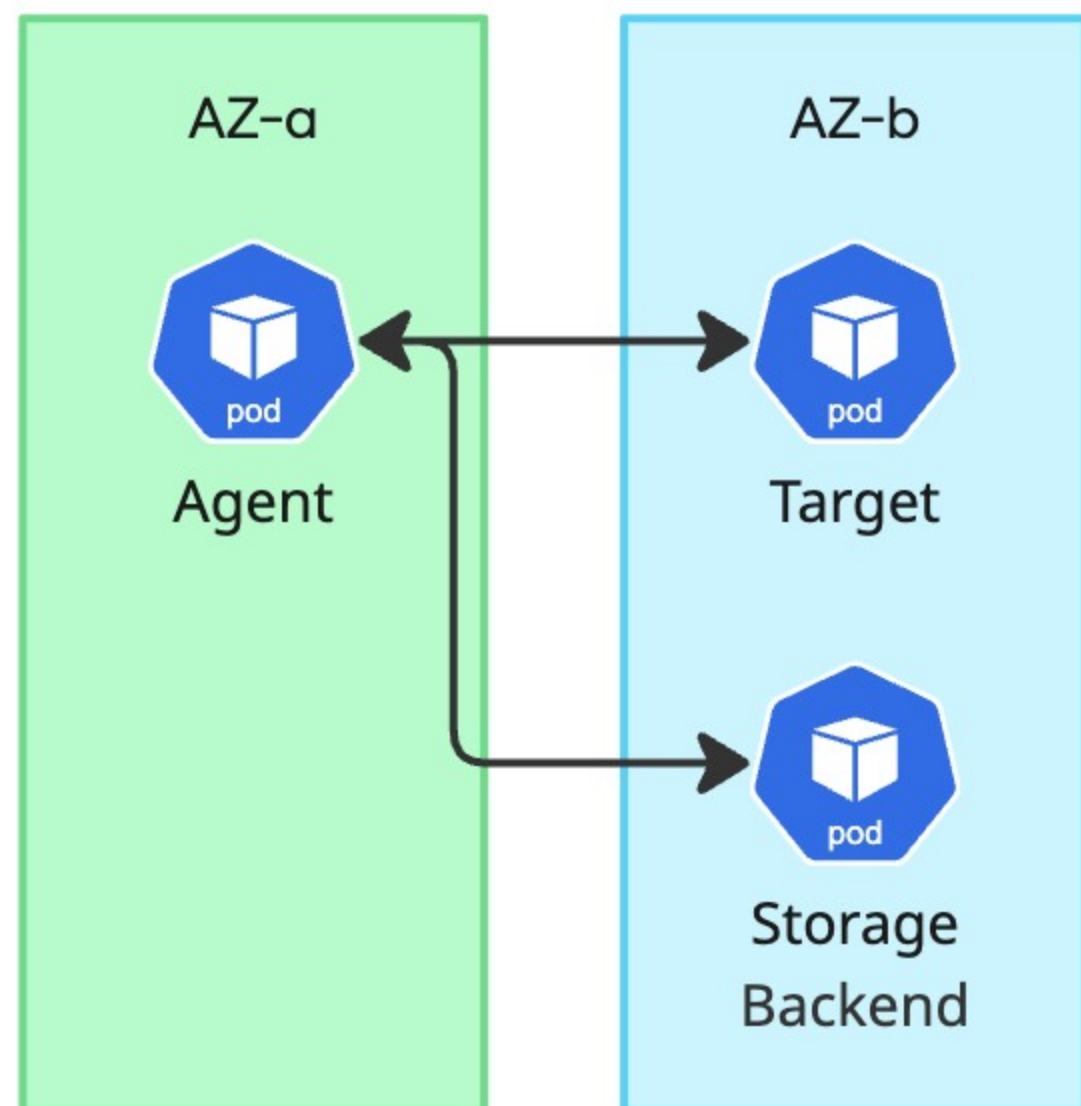
KubeCon



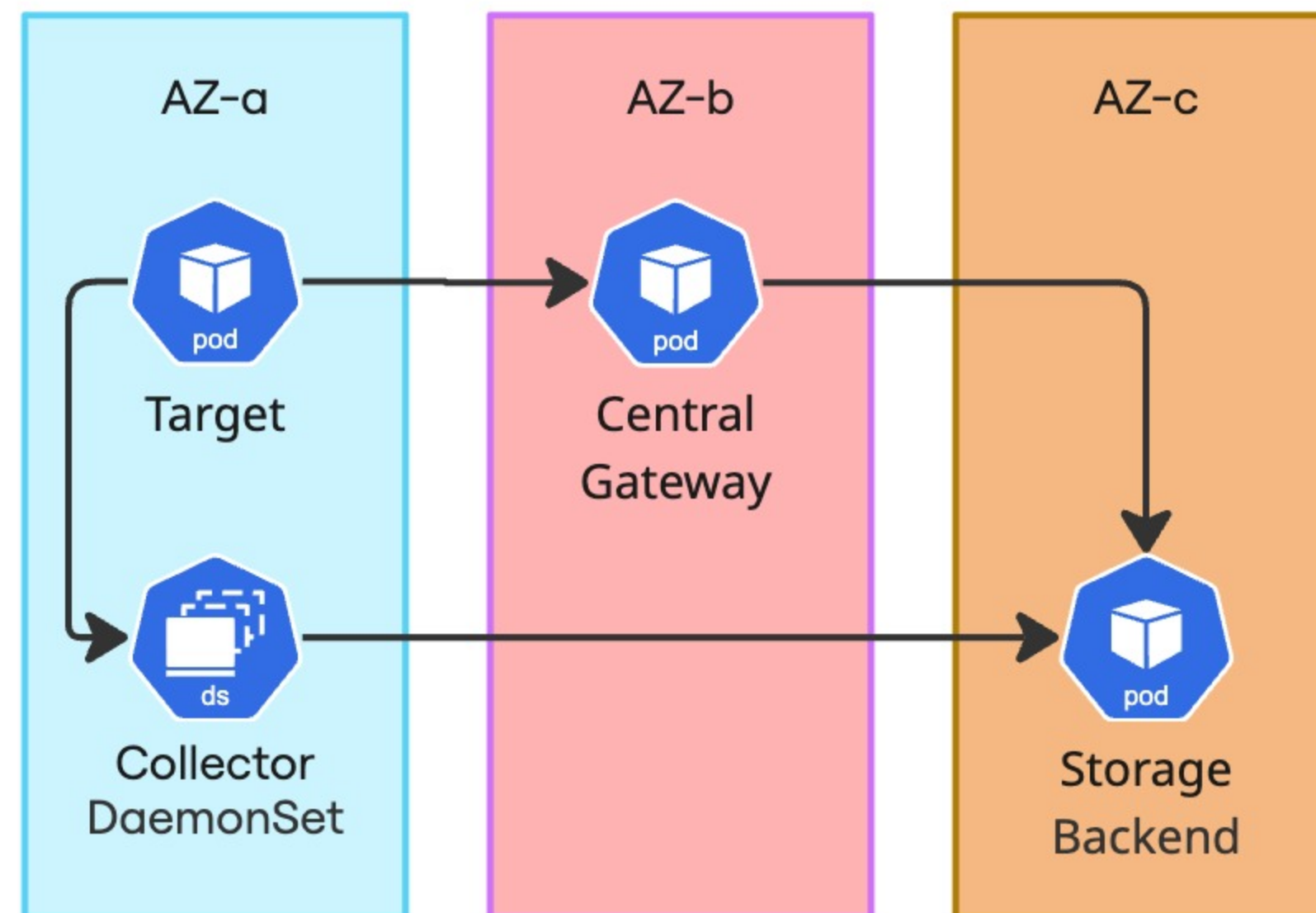
CloudNativeCon

Europe 2026

PULL (Prometheus-style)



PUSH (OTLP)



Anatomy of a Scandal



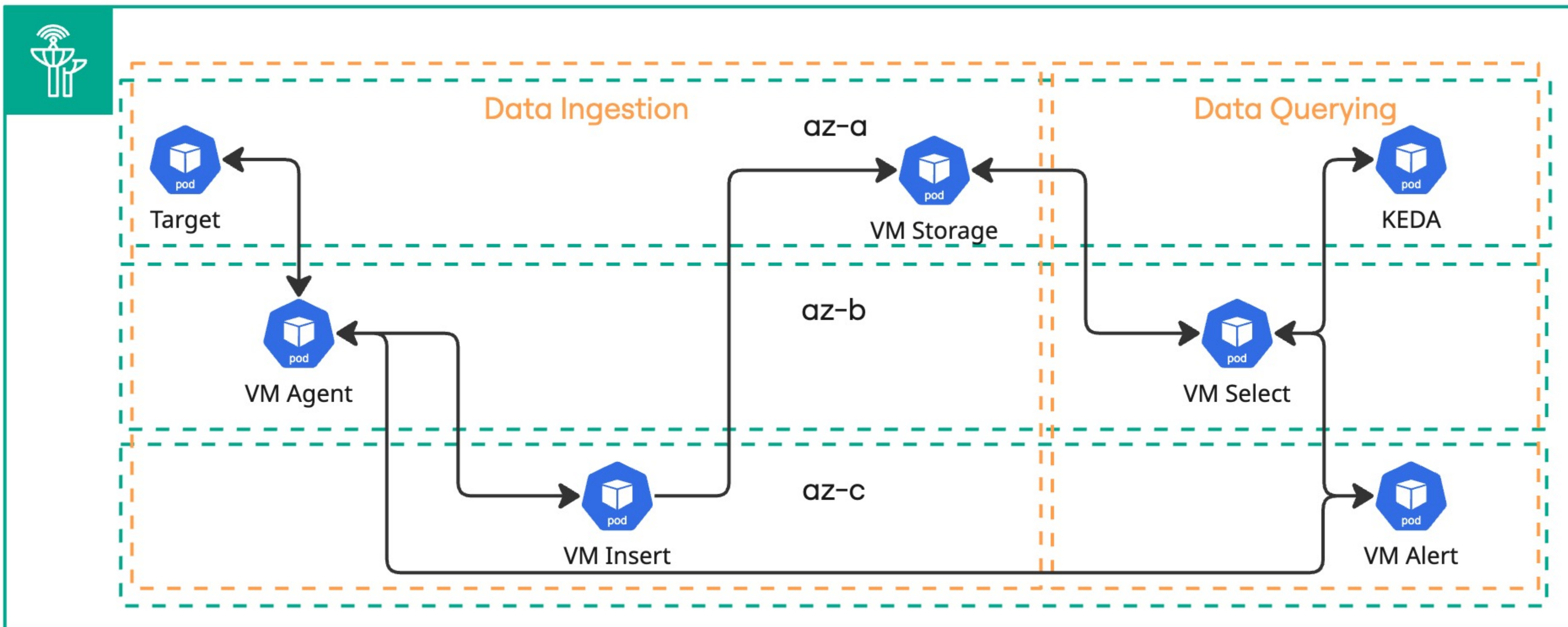
KubeCon



CloudNativeCon

Europe 2026

In a VictoriaMetrics, pull based land....



Anatomy of a Scandal



KubeCon



CloudNativeCon

Europe 2026

 Platform	 Collection	 Write Path	 Storage	 Read Path
VictoriaMetrics (Cluster)	vmagent	vminsert (stateless router)	vmstorage	vmselect
Prometheus (Agent Mode)	Prometheus agent	None	Prometheus TSDB	Prometheus Remote Read
Thanos (Receiver Mode)	Prometheus or any remote_write sender	Thanos Receive (accepts remote_write, replicates via hashing)	Object storage Thanos Store Gateway	Thanos Query Query Frontend
Grafana Mimir	Prometheus/Grafana Alloy/any remote_write sender	Distributor Ingester	Object storage (S3, GCS, Azure Blob) via Ingester flush	Querier  Query Frontend
Cortex	Prometheus or any remote_write sender	Distributor Ingester	Object storage or DynamoDB/Cassandra/BigTable (via Ingester flush)	Querier  Query Frontend



Impacts

- Higher architectural complexity
- Cost savings justified the time investment
- Increased cognitive load to the Observability team
- Zero impact on end-user experience
- Let's not chase a zero network bill



Non-Negotiables

- ⊙ Maintain current compute resource levels
- ⊙ No additional spending on new architecture
- ⊙ Preserve availability, reliability, and observability metrics
- ⊙ Scalable architecture
- ⊙ Centralized querying and alerting

Our context: Miro's metrics platform at scale



KubeCon



CloudNativeCon

Europe 2026

These are numbers that make this problem real, in one of our largest environments:



~140M

active time series



4.8M/s

peak ingestion rate



13,400+

active scrape targets



35 × 2

agent shards × HA replicas

- Pull-based VictoriaMetrics stack, scrape interval: 15–60 seconds
- 12–19 GB of metrics data per scrape interval (uncompressed)
- At peak: >25 TB/day of raw scrape volume flowing through the system



At this scale, \$0.02/GB cross-AZ doesn't sound like much — until it does.

Miro's usecase: Optimization in two phases



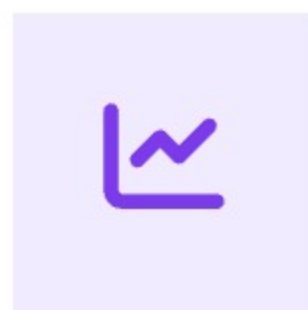
KubeCon



CloudNativeCon

Europe 2026

Due to the challenging number of architecture changes, we followed a two phased approach



Phase 1

Metrics collection

- ✓ Make vmagents AZ-aware
- ✓ Scrape only targets in the same AZ
- ✓ Quick win, low risk
- ⚠ Ingestion still crosses AZs -> Phase 2



Phase 2

Metrics ingestion,
storage and querying

- ✓ Full per-AZ VictoriaMetrics Stack
- ✓ Ingestion and query fully AZ-local
- ✓ Federated multi-level query
- ⚠ Bigger architectural change

Miro's usecase

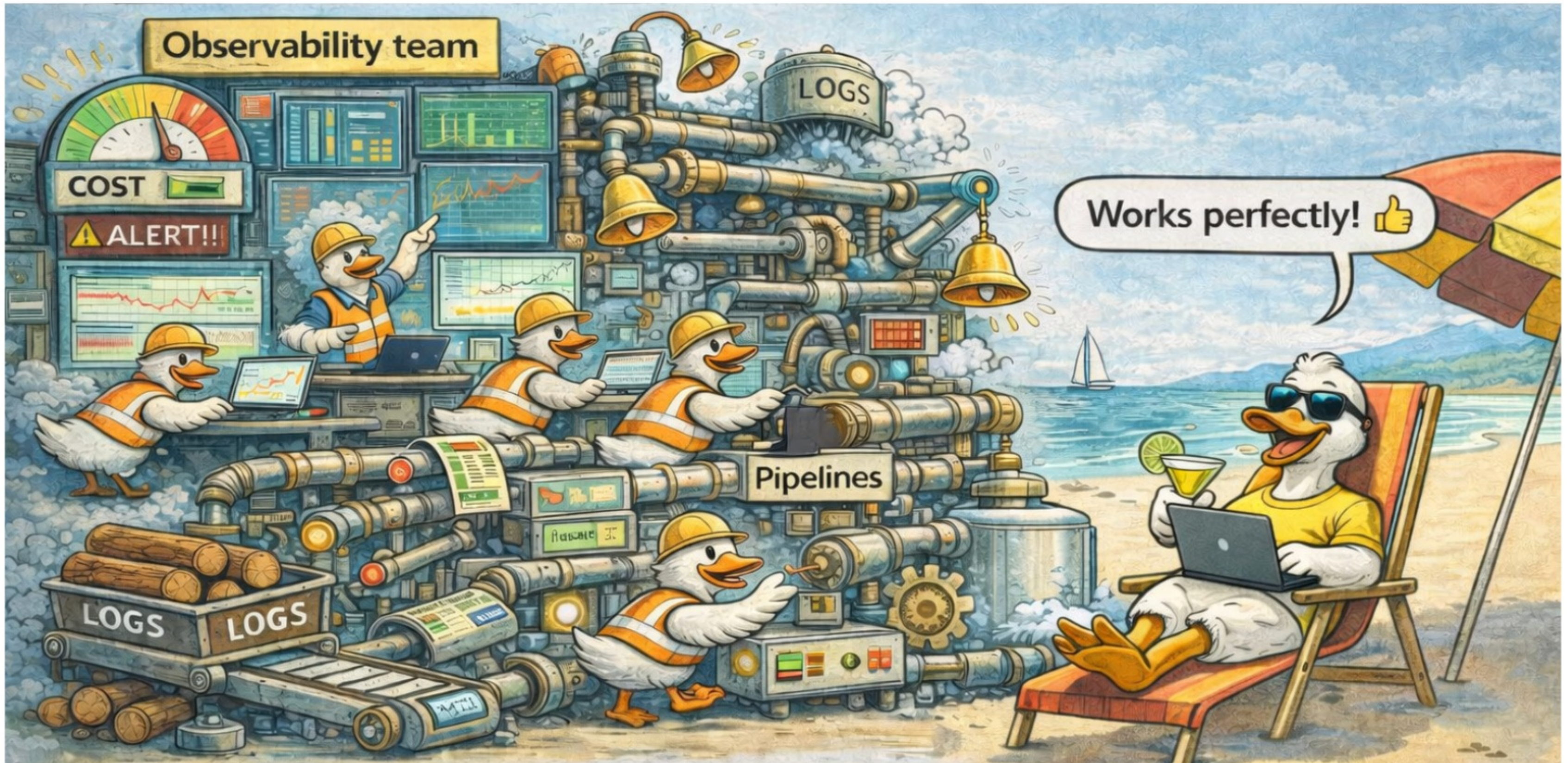


KubeCon



CloudNativeCon

Europe 2026



Metrics Collection – Pre-requisites



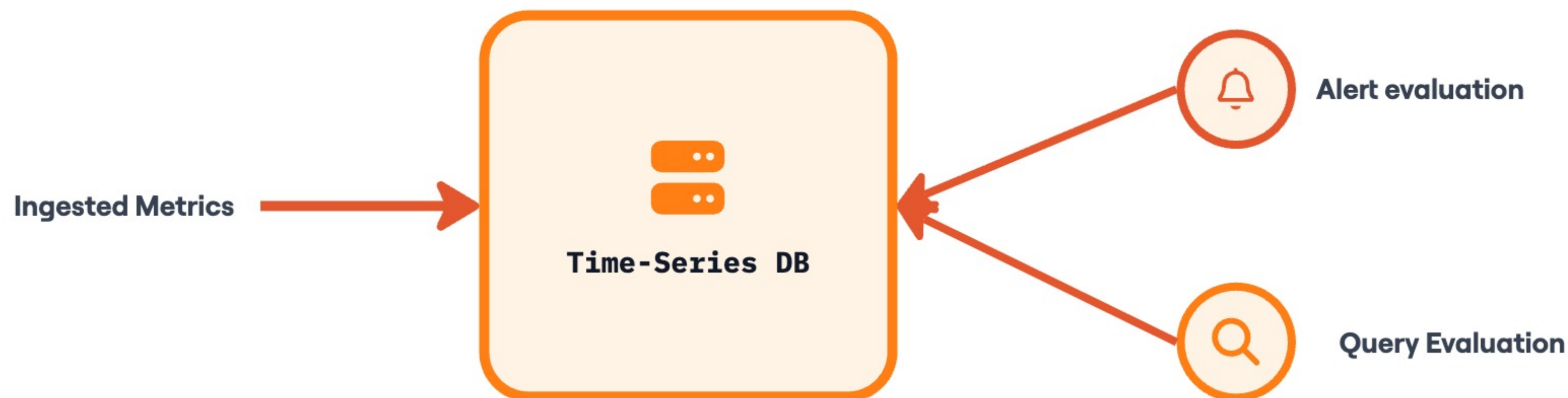
KubeCon



CloudNativeCon

Europe 2026

Central storage for alerting and querying purposes



Metrics Collection – Pull Metrics



KubeCon

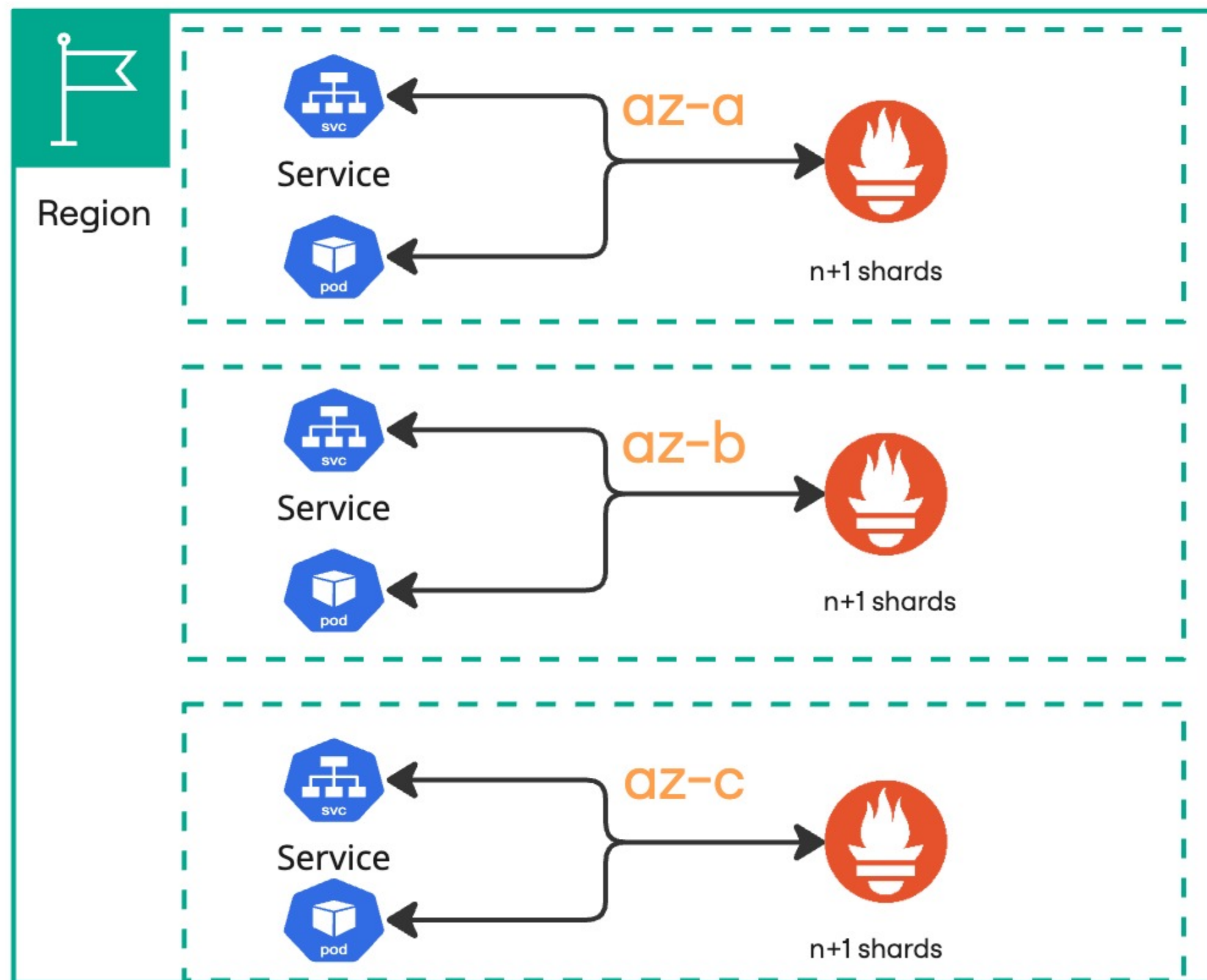


CloudNativeCon

Europe 2026

Key Architecture Points

- **Zone-Isolated Deployment:** One Prometheus/vmagent per availability zone
- **Intelligent Target Scraping:** Metric relabeling ensures zone-local scraping only
- **Horizontal Scalability:** Full sharding support (n+1 shards per zone)



Metrics Collection – Target Topology awareness



KubeCon



CloudNativeCon

Europe 2026

Topology metadata has to be available in the Pods

```
1 k -n kube-system describe po coredns-7877dd48f7-tqdrd
2 Name: coredns-7877dd48f7-tqdrd
3 Namespace: kube-system
4 Labels: k8s-app=kube-dns
5         pod-template-hash=7877dd48f7
6         topology.kubernetes.io/region=eu-west-1
7         topology.kubernetes.io/zone=eu-west-1b
```

📶 Thanks to KEP 4742 - <https://github.com/kubernetes/enhancements/issues/4742>

✅ Enabled by default in Kubernetes 1.35

```
10 > k -n kube-system describe po coredns-69b4756d98-fmqmx
11 Name: coredns-69b4756d98-fmqmx
12 Namespace: kube-system
13 Annotations: topology.kubernetes.io/region: eu-west-1
14              topology.kubernetes.io/zone: eu-west-1a
```

📶 You will have to mutate pods during node binding

✅ <https://kyverno.io/blog/2024/02/19/assigning-node-metadata-to-pods/>



```
1 scrape_configs:
2 - job_name: some-scraping-
3   kubernetes_sd_configs:
4   - role: endpoints
5     attach_metadata:
6       node: true
```

📶 Prometheus Native, supported by platforms

⚠️ Increased API Server Pressure & Memory Consumption

Metrics Collection – Intelligent Target Scraping



KubeCon



CloudNativeCon

Europe 2026

Target relabeling and filtering for the win...

Kubernetes SD

```
1  scrape_configs:
2  - job_name: some-scraping-job
3    kubernetes_sd_configs:
4      - role: endpoints
5    relabel_configs:
6      - ...
7      - source_labels:
8        - __meta_kubernetes_pod_label_topology_kubernetes_io_zone
9        regex: "eu-west-1a"
10     action: keep
```

EC2 SD

```
1  scrape_configs:
2  - job_name: system_metrics
3    ec2_sd_configs:
4      - region: eu-west-1
5        filters:
6          - name: instance-state-name
7            values: [running]
8          - name: vpc-id
9            values: [vpc-12345678]
10         - name: availability-zone
11           values: ["eu-west-1a"]
12    relabel_configs:
13      - ...
```

Metrics Collection – Global configs



KubeCon



CloudNativeCon

Europe 2026

PrometheusAgent

```
1  apiVersion: monitoring.coreos.com/v1alpha1
2  kind: PrometheusAgent
3  metadata:
4    name: prometheus-agent
5  spec:
6    scrapeClasses:
7      - name: az-filter
8        default: true
9        relabelings:
10         - sourceLabels:
11             - __meta_kubernetes_pod_annotation_topology_kubernet
12               es_io_zone
13             - __meta_kubernetes_pod_label_topology_kubernetes_io
14               _zone
15             - __meta_ec2_availability_zone
16         regex: ".*eu-west-1a.*"
17         action: keep
```

VMAgent

```
1  apiVersion: operator.victoriametrics.com/v1beta1
2  kind: VMAgent
3  metadata:
4    name: vmagent
5  spec:
6    extraEnvs:
7      - name: ZONE_NAME
8        valueFrom:
9          fieldRef:
10            fieldPath: metadata.labels['topology.kubernetes.io
11                      /zone']
12      - name: DEDICATED_AZ_PLACEMENT
13        value: ""
14    serviceScrapeRelabelTemplate:
15      - action: keep
16        regex: (^(%{ZONE_NAME}%{DEDICATED_AZ_PLACEMENT}))$)
17        sourceLabels:
18          - __meta_kubernetes_pod_label_topology_kubernetes_io_zo
19            ne
20    probeScrapeRelabelTemplate:
21      - action: keep
22        regex: (^(%{ZONE_NAME}%{DEDICATED_AZ_PLACEMENT}))$)
23        sourceLabels:
24          - __meta_kubernetes_pod_label_topology_kubernetes_io_zo
25            ne
```

Metrics Collection – The Hidden topology problem



KubeCon



CloudNativeCon

Europe 2026

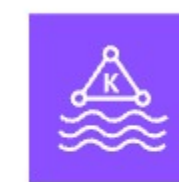
PROBLEM



What about targets from managed services or opaque URLs that are often not associated with one single zone?



```
1  kind: VMAgent
2  spec:
3    extraEnvs:
4      - name: ZONE_NAME
5        valueFrom:
6          fieldRef:
7            fieldPath: metadata.labels['topology.kubernetes.io/zone']
8      - name: DEDICATED_AZ_PLACEMENT
9        value: "|"
10   serviceScrapeRelabelTemplate:
11     - action: keep
12       regex: (^({ZONE_NAME}%{DEDICATED_AZ_PLACEMENT}))$ # example "^(eu-west-1a|)$"
13     sourceLabels:
14       - __meta_kubernetes_pod_label_topology_kubernetes_io_zone
```



Amazon
MSK



Amazon
EKS



blackbox-exporter

Fast forwarding to the final solution

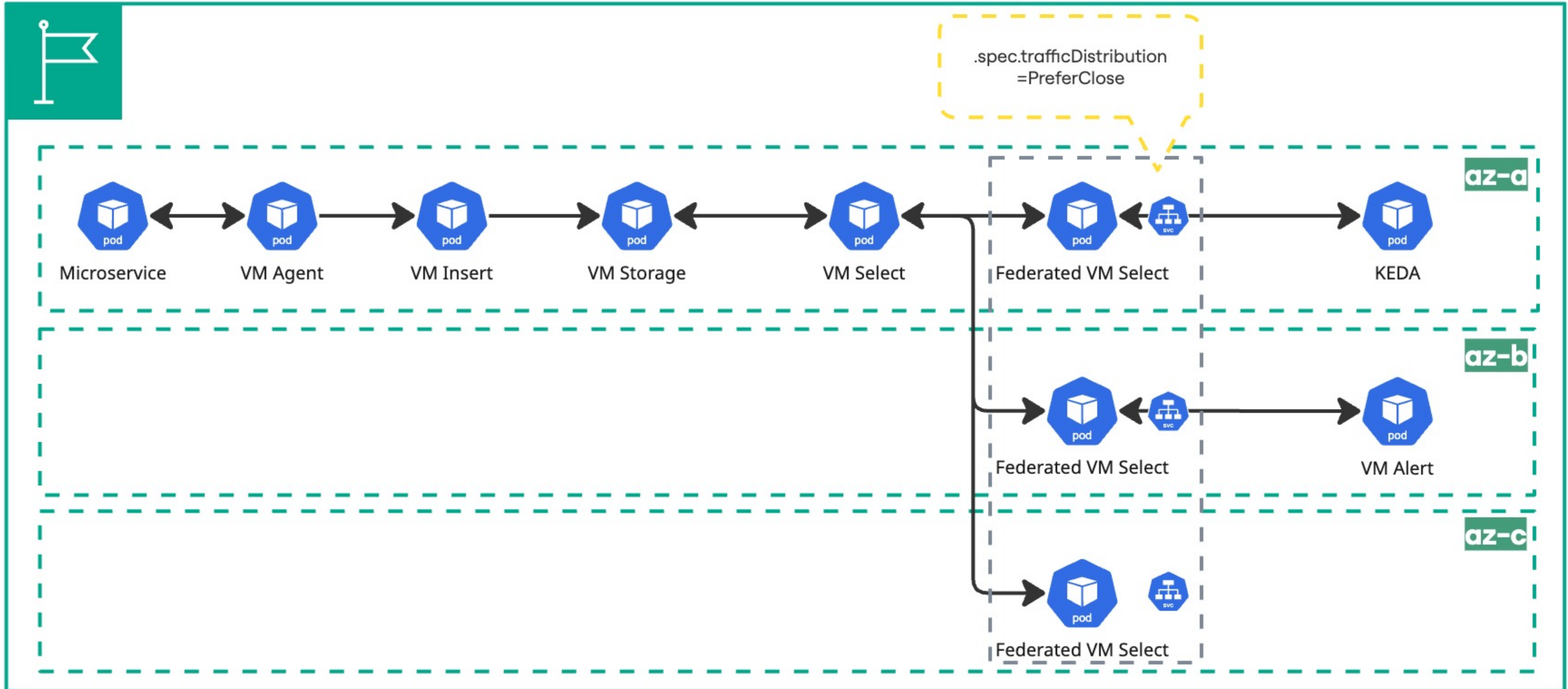


KubeCon



CloudNativeCon

Europe 2026



Show case



KubeCon



CloudNativeCon

Europe 2026

<https://github.com/rodrigorfk/kubecon-eu-2026-az-observability>



And if you were wondering, how much did we save so far?



FYDISNEYMISFITS | TK-THE-TIGER

A LOT

GIF





KubeCon



CloudNativeCon

Europe 2026



THANK YOU!

***Let us know
your thoughts***

