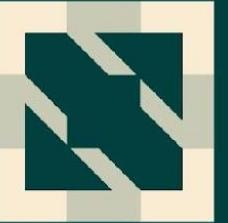




KubeCon



CloudNativeCon

S OPEN SOURCE SUMMIT

China 2023



KubeCon



CloudNativeCon



OPEN SOURCE SUMMIT

China 2023

Scaling up and Scaling Out Networking, Observability, and Security with Cilium

Bill Mulligan, Isovalent & Jaff Cheng, Trip.com Group

Agenda

- **Cilium at Trip.com**
- **Cilium on Alibaba Cloud**
- **Challenges & Solutions at Scale**
- **Takeaways**



eBPF-based:

- Networking
- Security
- Observability
- Service Mesh & Ingress

Foundation

CLOUD NATIVE COMPUTING FOUNDATION

Technology

eBPF envoy

Over 100 USERS.md entries

Adobe What Makes a Good Multi-tenant Kubernetes Solution VIDEO 1 - VIDEO 2	Alibaba Cloud Building High-Performance Cloud Native Pod Networks READ BLOG	aws AWS picks Cilium for Networking & Security on EKS Anywhere READ BLOG	Bell Bell uses Cilium and eBPF for telco networking VIDEO 1 - VIDEO 2	AccuKnox AccuKnox uses Cilium for network visibility and network policy enforcement READ BLOG	ACOSS ACOSS uses Cilium as their main CNI plugin for self hosted Kubernetes WATCH VIDEO	ArangoDB ArangoDB Oasis uses Cilium to separate database deployments in a multi-tenant cloud environment WATCH VIDEO	ayedo Ayedo builds and operates cloud native platforms using Cilium WATCH VIDEO
CapitalOne Building a Secure and Maintainable PaaS WATCH VIDEO	cengn Cloud Native Networking with eBPF WATCH VIDEO	datadog Datadog is using Cilium in AWS (self-hosted k8s) WATCH VIDEO	DigitalOcean Managed Kubernetes: 1.5 Years of Cilium Usage at DigitalOcean WATCH VIDEO	ByteDance ByteDance uses Cilium as their CNI for self-hosted Kubernetes clusters WATCH VIDEO	canonical Canonical's Kubernetes distribution microk8s uses Cilium as CNI plugin WATCH VIDEO	CIVO Civo is offering Cilium as the CNI option for Civo users to choose it for their Civo Kubernetes clusters WATCH VIDEO	COGNITE Cognite uses Cilium as the CNI plugin for industrial DataOps WATCH VIDEO
E-CAPITAL TRANSFER ect888 uses Cilium as their CNI and for load balancing READ BLOG	GitLab Kubernetes Network Policies in Action with Cilium VIDEO	Google Google chooses Cilium for Google Kubernetes Engine (GKE) networking READ BLOG	IKEA IKEA uses Cilium for their self-hosted bare-metal READ BLOG	elasticpath Elastic Path uses Cilium in their production CloudOps READ BLOG	F5 F5 uses Cilium VXLAN tunnel integration with iMAGINE WATCH VIDEO	finleap connect finleap connect uses Cilium on a bare metal private cloud WATCH VIDEO	FORM3 Form3 is using Cilium in their production clusters (self-hosted, bare-metal, private cloud) WATCH VIDEO
MÁSMÓVIL Scaling a Multi-Tenant Kubernetes Clusters in a Telco WATCH VIDEO	Meltwater Meltwater is using Cilium in AWS on self-hosted multi-tenant k8s clusters as the CNI plugin WATCH VIDEO	MOBILA Mobila uses Cilium as their CNI for their internal network READ BLOG	Microsoft Microsoft is using Cilium as their CNI plugin on EKS for its IoT SaaS READ BLOG	monitrix Monitrix uses Cilium in self-hosted clusters on bare-metal and Openstack READ BLOG	inmobi inmobi uses Cilium to run their customer's infrastructure READ BLOG	ISOVALENT Cilium is the platform that powers Isovalent's enterprise networking, observability, and security solutions WATCH VIDEO	JUMO JUMO uses Cilium as the CNI plugin for all of their AWS-hosted EKS clusters WATCH VIDEO
PostFinance PostFinance is using Cilium as their CNI for all mission critical, on-premise k8s clusters READ CASE STUDY	sky eBPF & Cilium at Sky WATCH VIDEO	sky BET SkyBet uses Cilium as their CNI READ BLOG	Trip.com Trip.com uses Cilium both on-premise and in AWS BLOG 1 - BLOG 2	KRYPTOS LOGIC Kryptos uses Cilium as the CNI for their on-prem Kubernetes clusters READ BLOG	Kube-OVN Kube-OVN uses Cilium to enhance the CNI service performance, security and monitoring READ BLOG	KUBERMATIC Kubernetic uses Cilium as the CNI for its Kubernetes installer and platform WATCH VIDEO	KUBESPHERE Kubesphere is an open-source lightweight tool for deploying Kubernetes clusters and addons WATCH VIDEO
Northflank Northflank uses Cilium as its CNI plugin across GCP, Azure, AWS and bare metal READ BLOG	overstock.com Overstock uses Cilium as their CNI for self-hosted bare metal clusters READ BLOG	Palantir Palantir is using Cilium as their main CNI plugin in AWS (self-hosted k8s) READ BLOG	PLAID Plaid uses Cilium as the CNI for its serverless database platform READ BLOG	REPLY LIQUID Liquid Reply is a consulting firm that uses Cilium in client projects READ BLOG	Melenion Inc Melenion uses Cilium as the CNI for its on-premise production clusters READ BLOG	MUX Mux uses Cilium on self-hosted clusters in GCP and AWS to run its video streaming/analytics platforms READ BLOG	myfitnesspal MyFitnessPal trusts Cilium with high volume user traffic on AWS and GKE READ BLOG
PlanetScale PlanetScale uses Cilium as their CNI plugin in self-hosted Kubernetes on AWS READ BLOG	radiofrance Radio France uses Cilium in their self-hosted clusters on AWS READ BLOG	rapyuta robotics Rapyuta Robotics uses Cilium as their main CNI plugin for host-based clusters READ BLOG	SAP SAP uses Cilium for projects across AWS, Azure, GCP, and OpenStack READ BLOG	sproutfi Sproutfi uses Cilium as the CNI on its GKE based clusters READ BLOG	SUPERORBITAL SuperOrbital uses Cilium in their customer engagements READ BLOG	TAILOR BRANDS Tailor Brands uses Cilium in their EKS clusters READ BLOG	The New York Times The New York Times uses Cilium on EKS to build multi-region multi-tenant shared clusters READ BLOG
Scaleway Scaleway uses Cilium as the default CNI for Kubernetes Capsule READ BLOG	SCHUBERG PHILIS Schuberg Philis uses Cilium as the CNI for mission critical Kubernetes clusters they run for their customers READ BLOG	SIMPLE Simple uses Cilium as default CNI for EKS READ BLOG	smile SmileDirectClub uses Cilium in self-hosted clusters vSphere and EC2 for manufacturing READ BLOG	T Systems TSI uses Cilium for its Open Sovereign Cloud product READ BLOG	yahoo/ Yahoo is using Cilium for L4 North-South Load Balancing for Kubernetes Services READ BLOG		



eBPF-based:

- Networking
- Security
- Observability
- Service Mesh & Ingress

Foundation



Technology



Deploy on your preferred cloud



Use your favorite Kubernetes distribution



Major  cloud providers have now
picked  for Networking & Security
in their Kubernetes platforms

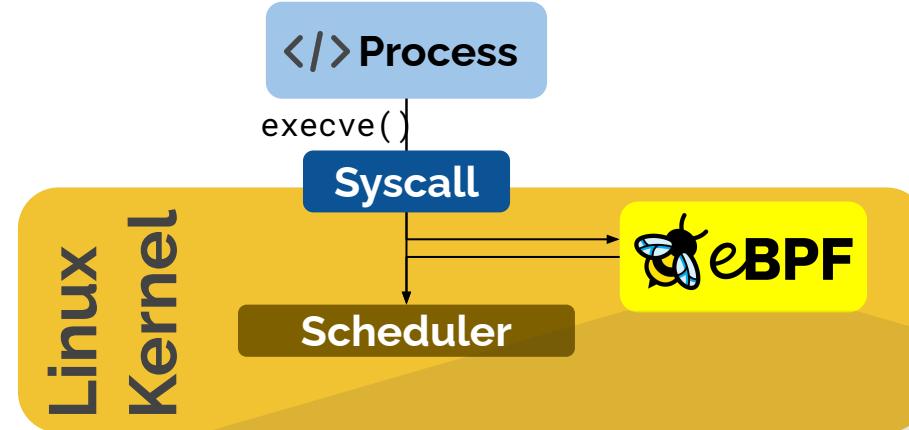


Google Cloud



Makes the Linux kernel programmable in a secure and efficient way.

“What JavaScript is to the browser, eBPF is to the Linux Kernel”



```
int syscall__ret_execve(struct pt_regs *ctx)
{
    struct comm_event event = {
        .pid = bpf_get_current_pid_tgid() >> 32,
        .type = TYPE_RETURN,
    };

    bpf_get_current_comm(&event.comm, sizeof(event.comm));
    comm_events.perf_submit(ctx, &event, sizeof(event));

    return 0;
}
```



Cilium CNI

Scalable, Secure,
High Performance
CNI Plugin





Cilium CNI

Scalable, Secure,
High Performance
CNI Plugin



Cilium Service Mesh

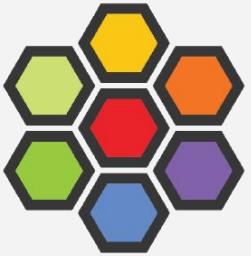
Sidecar-free Mesh &
Ingress





Cilium CNI

Scalable, Secure,
High Performance
CNI Plugin



Cilium Service Mesh

Sidecar-free Mesh &
Ingress



Hubble

Network
Observability





Cilium CNI

Scalable, Secure,
High Performance
CNI Plugin



Cilium Service Mesh

Sidecar-free Mesh &
Ingress



Hubble

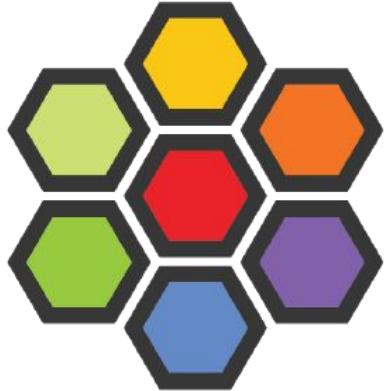
Network
Observability



Tetragon

Security Observability &
Runtime Enforcement

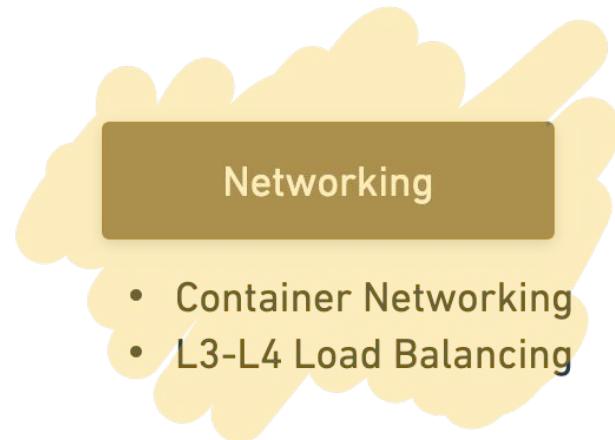
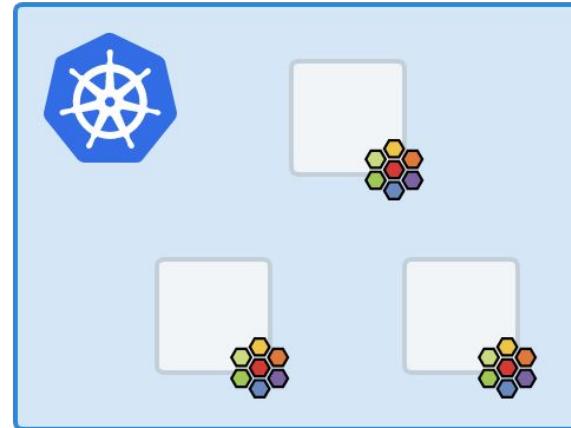




cilium

The Origins

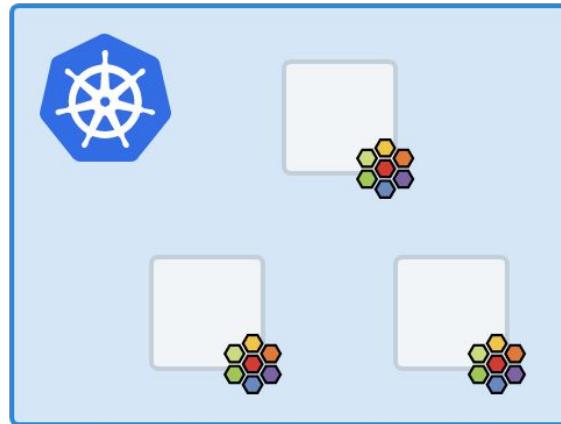
The Beginning



Vision

Inent & identity-based, high performance
container networking platform built using eBPF

Network Security 1.0



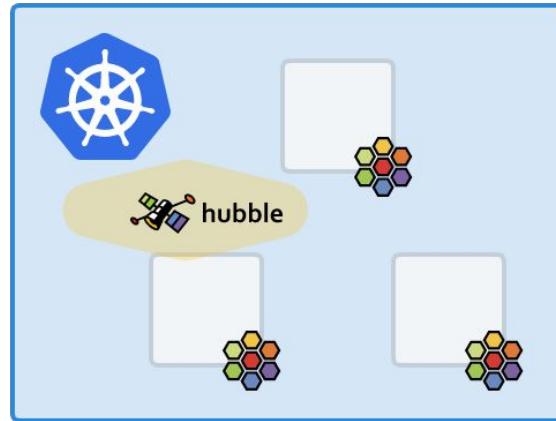
Networking

- Container Networking
- L3-L4 Load Balancing

Network Security

- Network Policy L3-L7
- Encryption

Hubble



Networking

- Container Networking
- L3-L4 Load Balancing

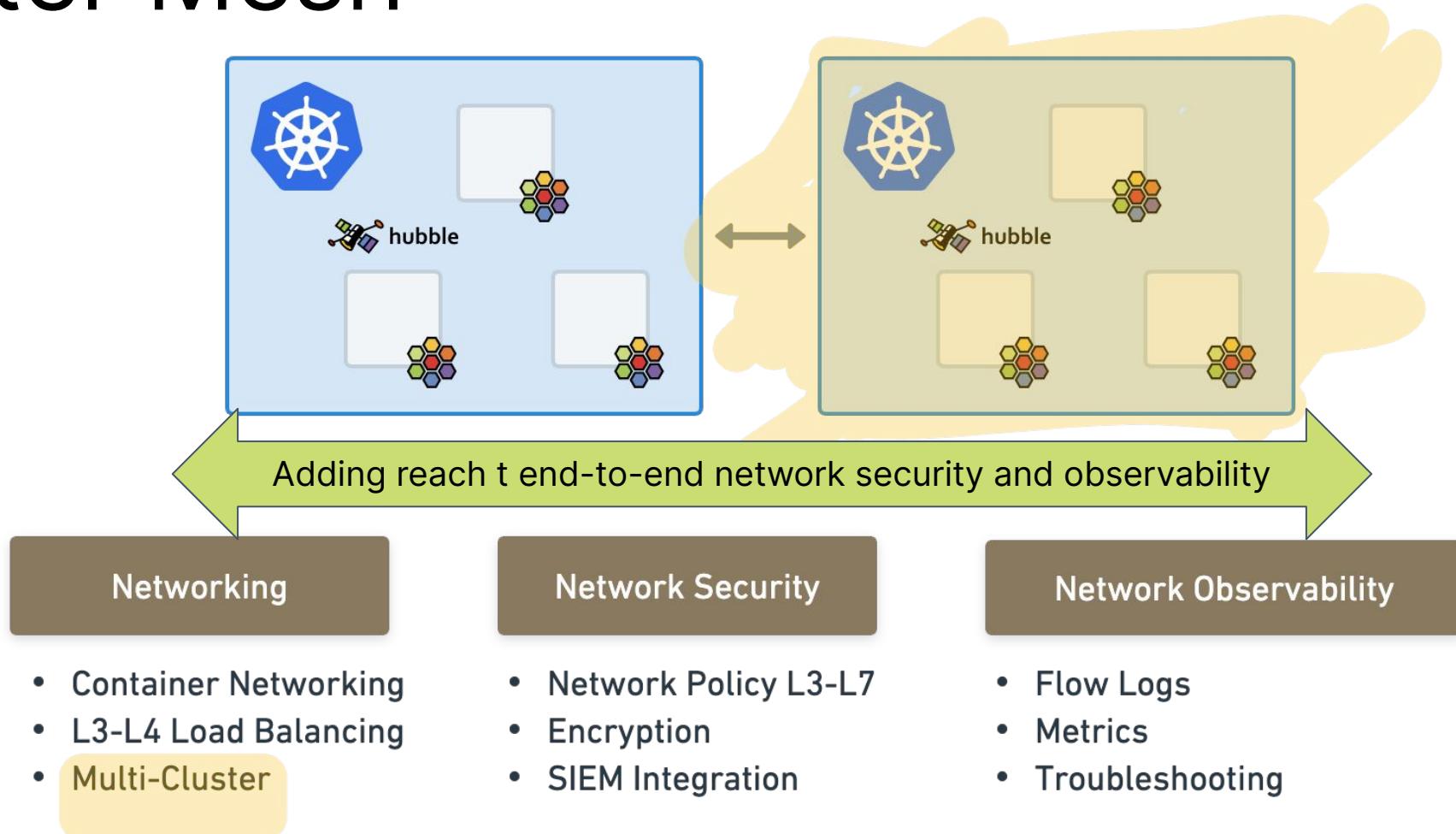
Network Security

- Network Policy L3-L7
- Encryption
- SIEM Integration

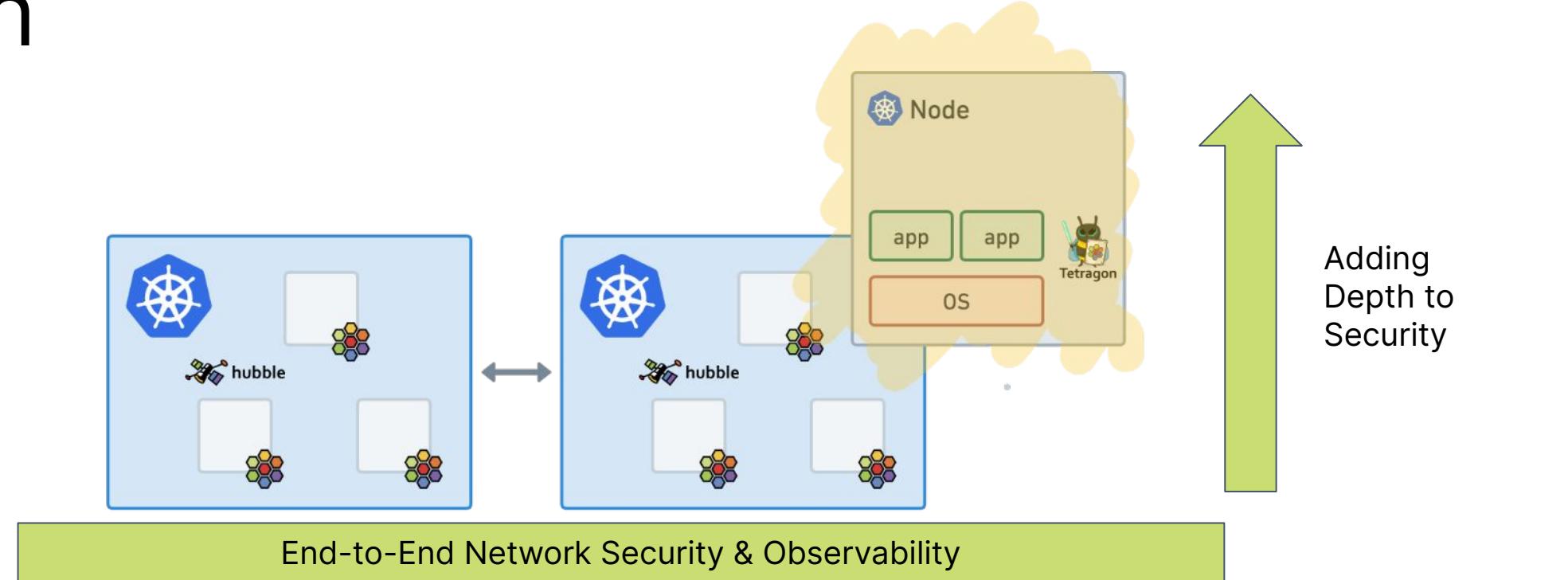
Network Observability

- Flow Logs
- Metrics
- Troubleshooting

Cluster Mesh



Tetragon



Networking

- Container Networking
- L3-L4 Load Balancing
- Multi-Cluster

Network Security

- Network Policy L3-L7
- Encryption
- SIEM Integration

Network Observability

- Flow Logs
- Metrics
- Troubleshooting

Runtime Security

- Runtime Enforcement

Runtime Observability

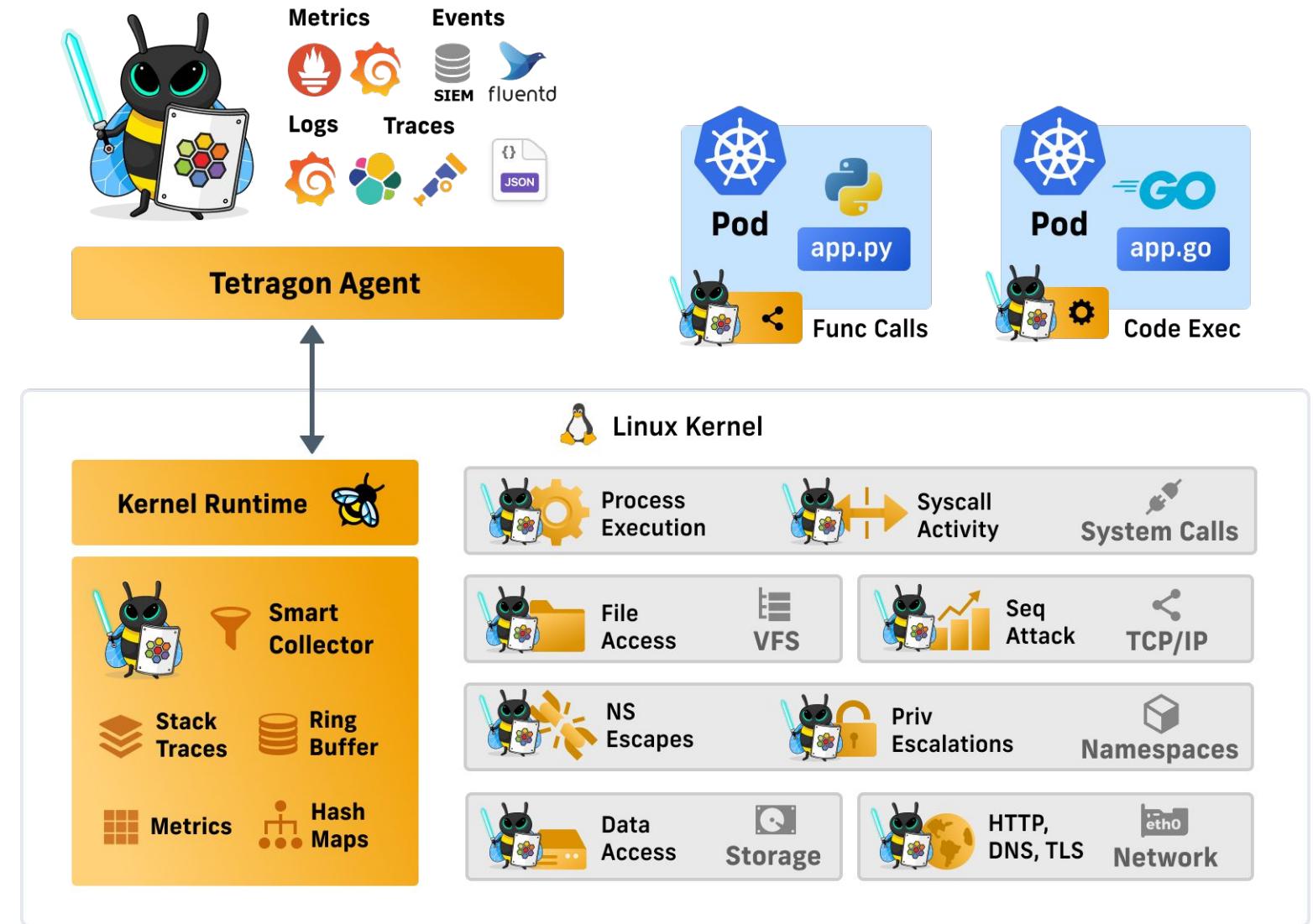
- Syscall, File, Privilege, and Network Observability
- SIEM Integration



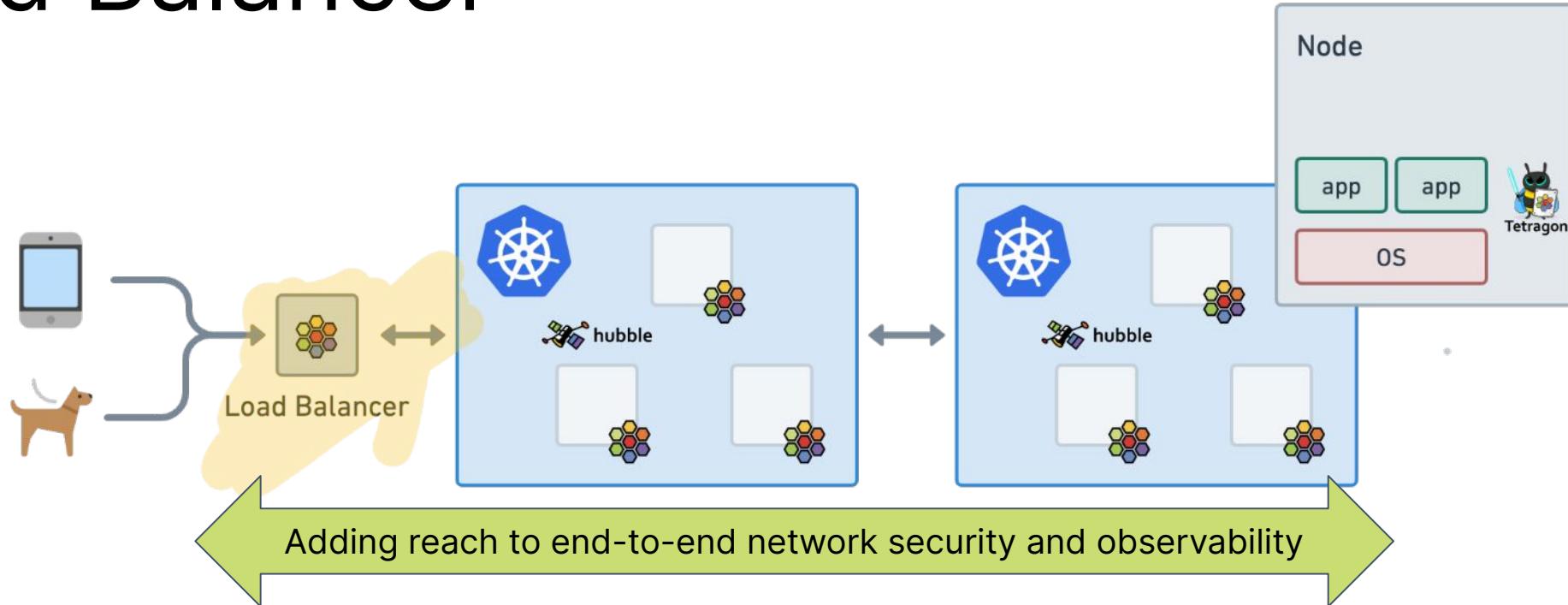
Tetragon

Security Observability & Runtime Enforcement

 CLOUD NATIVE COMPUTING FOUNDATION



Load Balancer



Networking

- Container Networking
- L3-L4 Load Balancing
- Standalone LB
- Multi-Cluster

Network Security

- Network Policy L3-L7
- Encryption
- SIEM Integration

Network Observability

- Flow Logs
- Metrics
- Troubleshooting

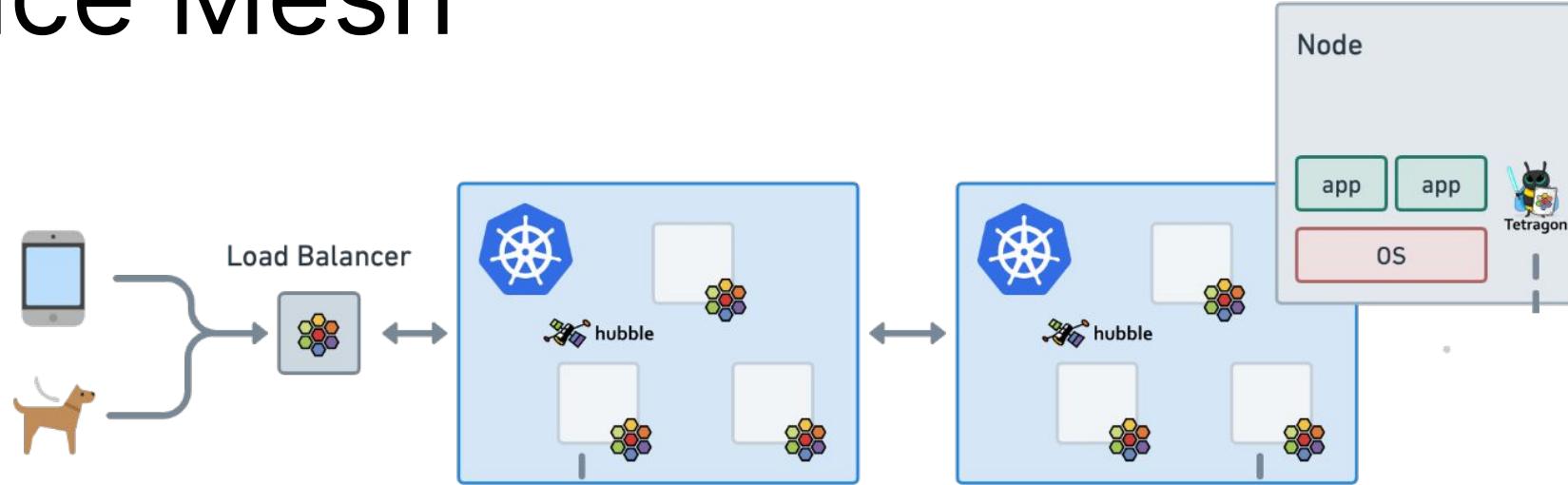
Runtime Security

- Runtime Enforcement

Runtime Observability

- Syscall, File, Privilege, and Network Observability
- SIEM Integration

Service Mesh



Networking

- Container Networking
- L3-L4 Load Balancing
- Standalone LB
- Multi-Cluster

Network Security

- Network Policy L3-L7
- Encryption
- SIEM Integration

Network Observability

- Flow Logs
- Metrics
- Troubleshooting
- Network Time Machine

Service Mesh

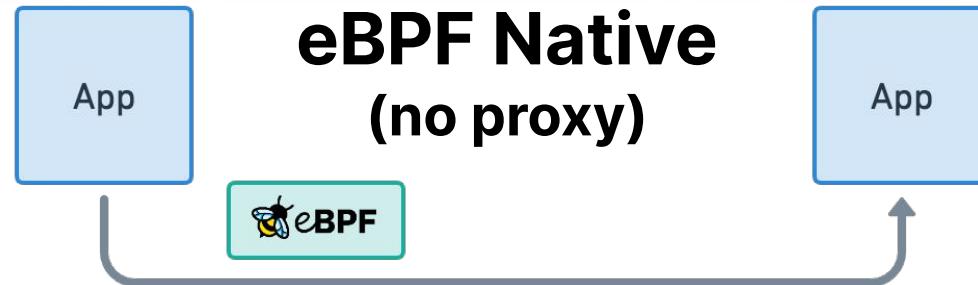
- L7 Load Balancing
- Tracing

Runtime Security

- Runtime Enforcement

Runtime Observability

- Syscall, File, Privilege, and Network Observability
- SIEM Integration
- Analytics



Whenever possible

Traffic Management

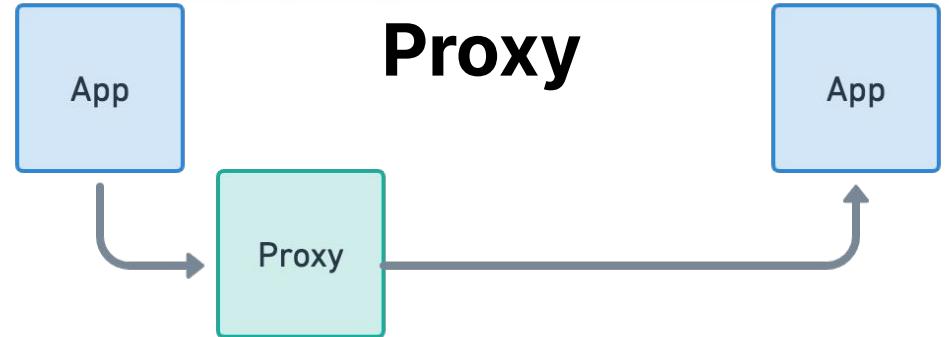
- L3/L4 forwarding & Load-balancing
- Canary, Topology Aware Routing
- Multi-cluster

Security

- Network Policy
- mTLS

Observability

- Tracing, OpenTelemetry, & Metrics
- HTTP, TLS, DNS, TCP, UDP, ...



When eBPF cannot do it

Traffic Management

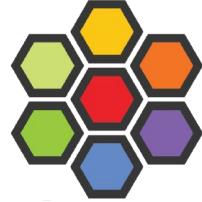
- L7 Load-balancing & Ingress

Resilience

- Retries, L7 Rate Limiting

Security

- TLS Termination & Origination
- L7 Network Policy*



Cilium 1.14 and beyond

Roadmap Highlights

- Mutual Authentication for NetworkPolicy
- SPIFFE Integration
- Day 2 Operations Enhancements
- Grafana Dashboards in Hubble UI
- Istio Ambient Mesh & zTunnel integration
- Did we mention more Grafana already?
- Cilium Mesh

Roadmap Highlights

Mutual Authentication

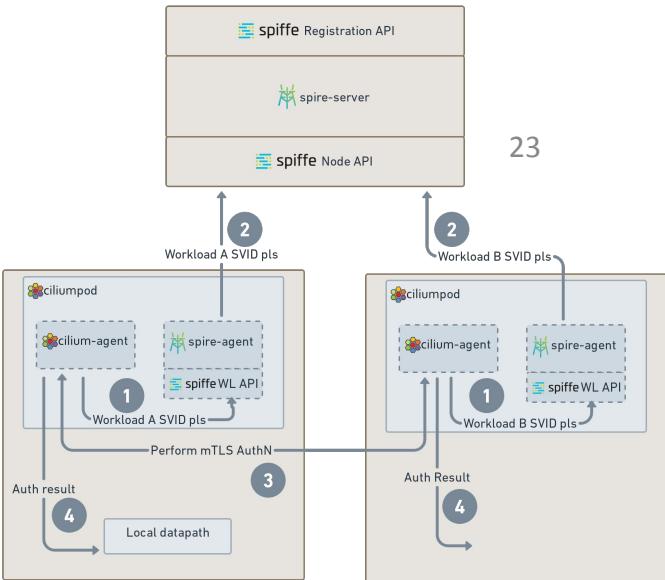
Mutual Auth via just Network Policy

```
apiVersion: cilium.io/v2
kind: CiliumNetworkPolicy
metadata:
  name: frontend-backend
spec:
  endpointSelector:
    matchLabels:
      role: backend
  ingress:
    - fromEndpoints:
        - matchLabels:
            role: frontend
  auth:
    required: strict
```



SPIFFE Integration

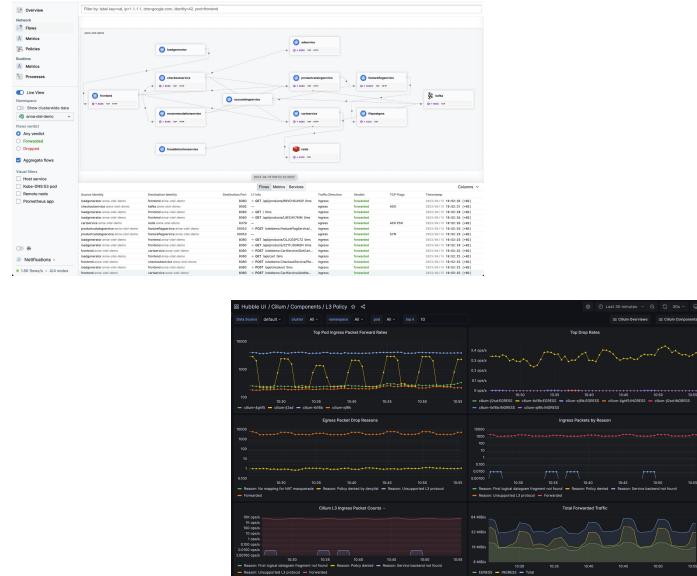
Certificate management via SPIFFE/SPIRE + SPIFFE ID selector matching



Day 2 Ops

cilium

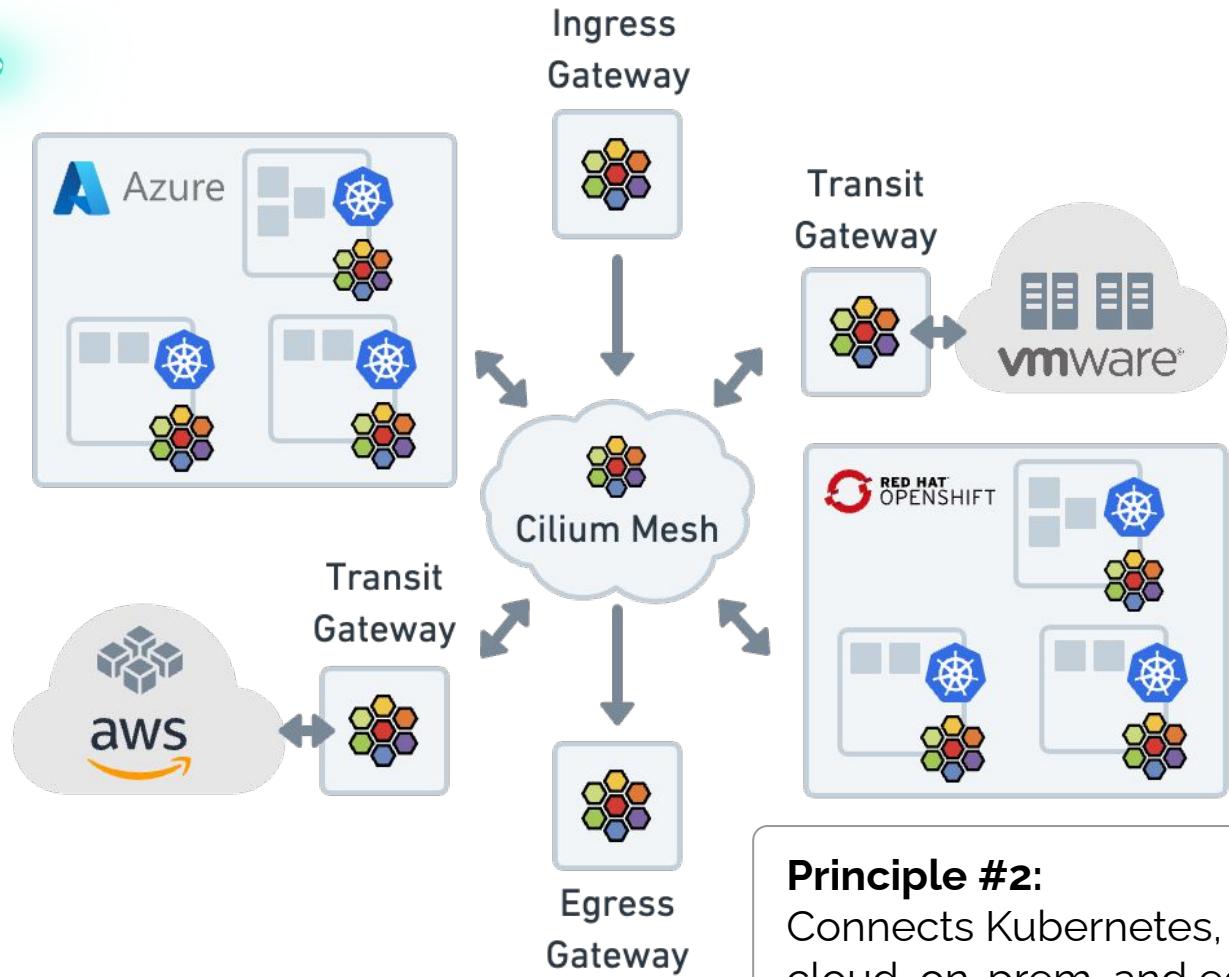
Assisted monitoring, proactive troubleshooting, simplified day 2 ops



Cilium Mesh



One Mesh to Connect Them All



Principle #1:

Combines all Cilium components into a single mesh:

- Kubernetes Networking (CNI)
- Cluster Mesh (Multi-Cluster)
- Ingress & Egress Gateway
- Load Balancer
- Service Mesh

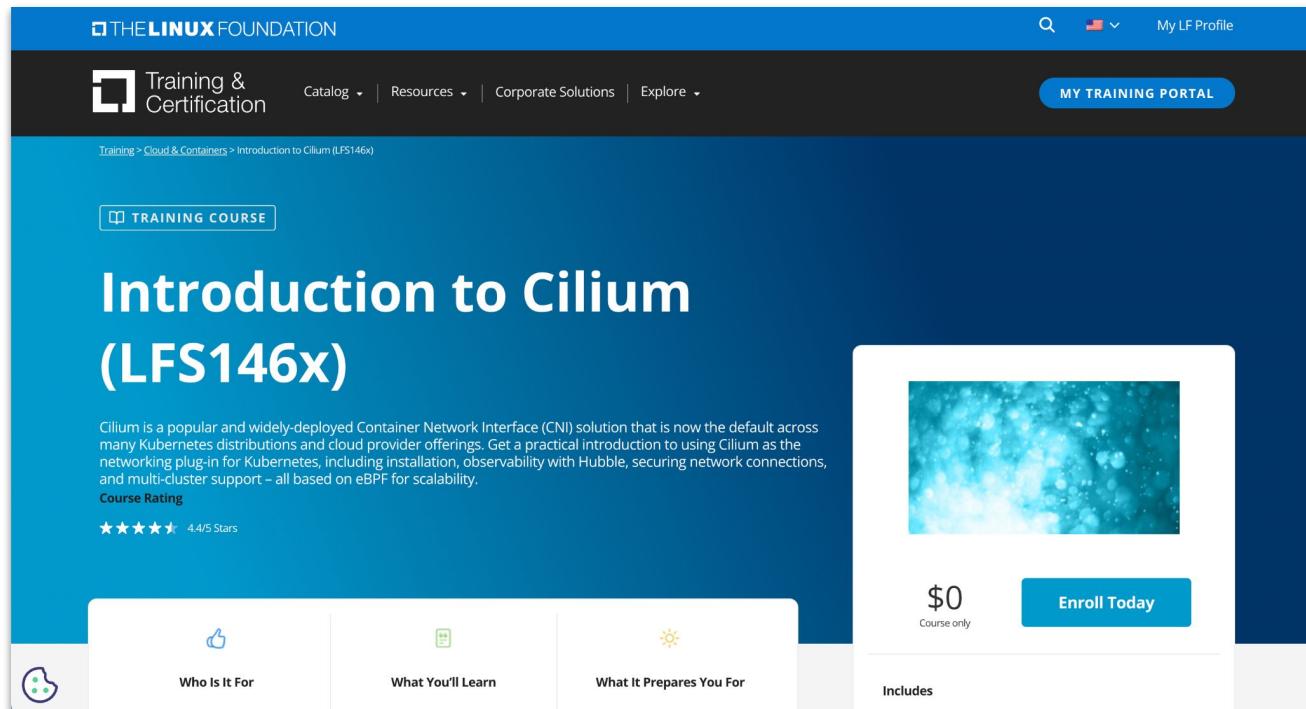
Principle #2:

Connects Kubernetes, VMs, and Servers across cloud, on-prem, and edge.



LF Intro to Cilium Course

training.linuxfoundation.org - LFS146x



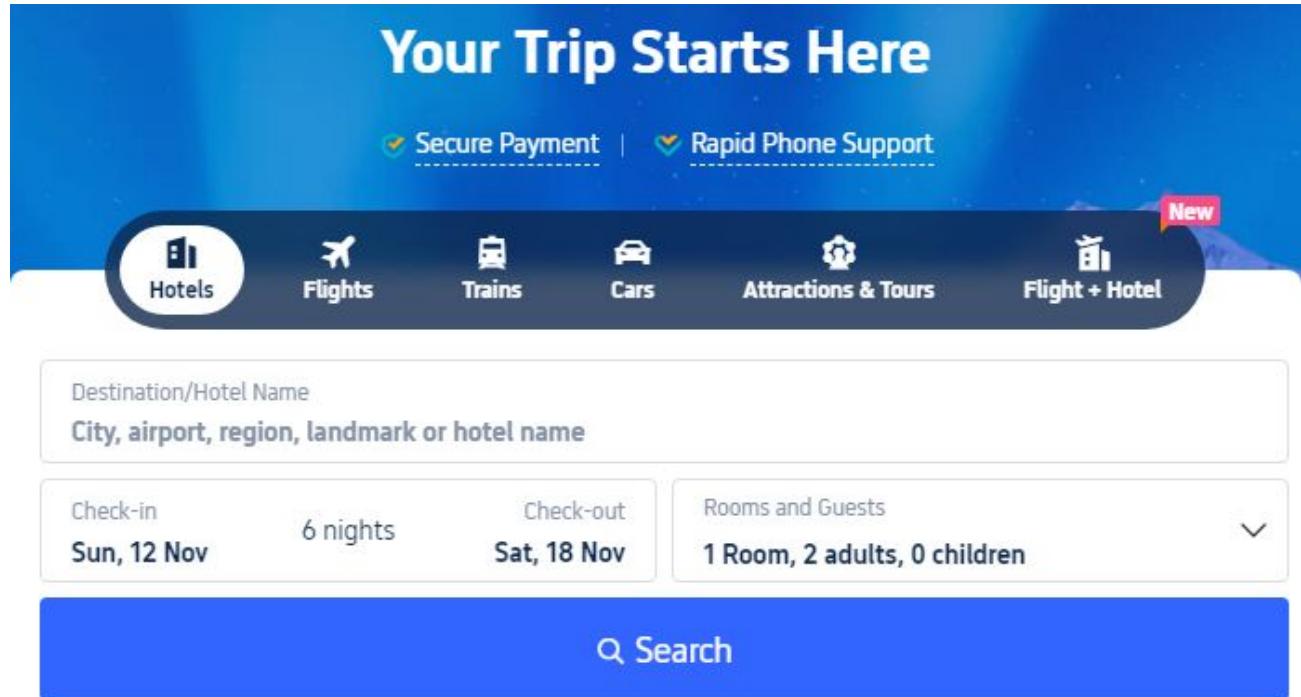
The screenshot shows the course landing page for "Introduction to Cilium (LFS146x)" on the Linux Foundation's training portal. The page features a large blue header with the course title. Below the title, a brief description states: "Cilium is a popular and widely-deployed Container Network Interface (CNI) solution that is now the default across many Kubernetes distributions and cloud provider offerings. Get a practical introduction to using Cilium as the networking plug-in for Kubernetes, including installation, observability with Hubble, securing network connections, and multi-cluster support – all based on eBPF for scalability." A "Course Rating" section indicates 4.4/5 Stars. At the bottom, there are three sections: "Who Is It For" (with a person icon), "What You'll Learn" (with a book icon), and "What It Prepares You For" (with a sun icon). A prominent call-to-action button on the right says "Enroll Today".

About Trip.com Group

- Leading online travel platform
- Flights, hotels, tours and more
- 400 million users worldwide

Cloud team @Trip.com Group

- R&D of cloud infra over the globe
- Virtualization, networking, storage, security



Cilium at Trip.com

20K Nodes

350K Pods

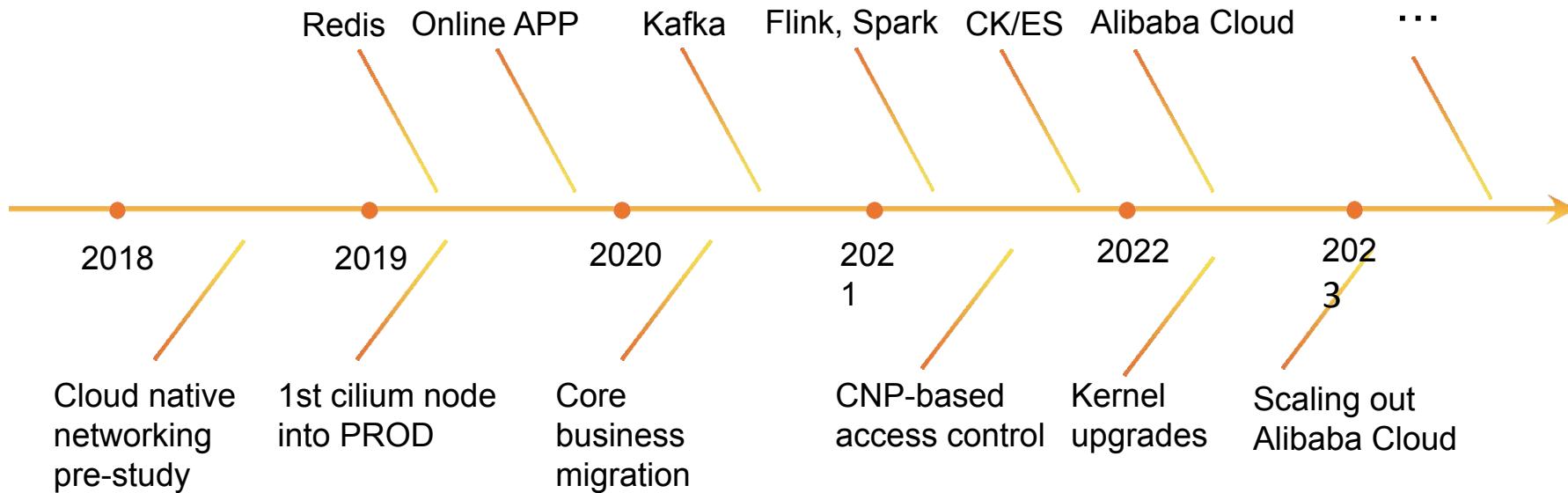
3K Network Policy

200K Hubble events/s

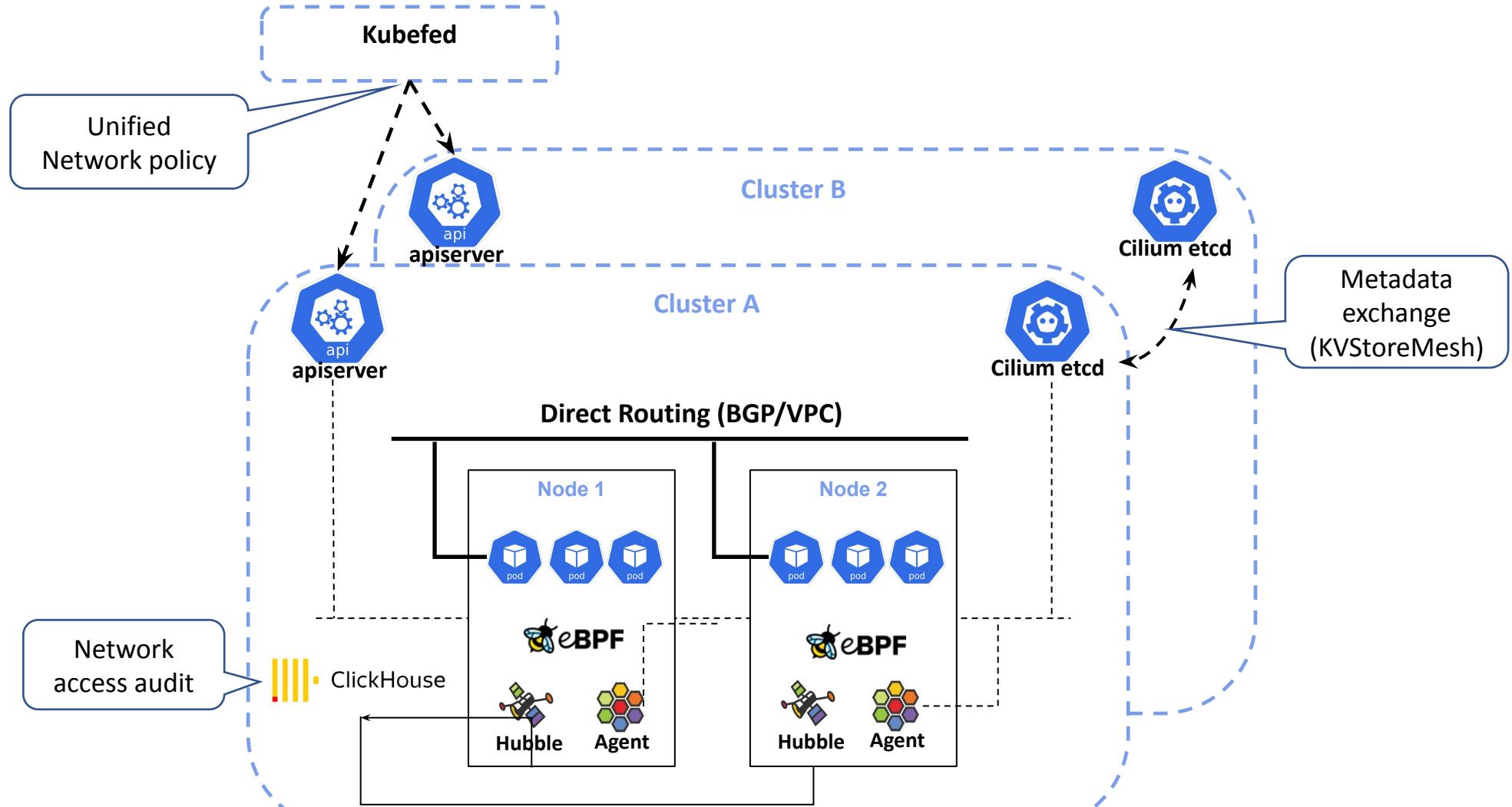
On Prem
Data Centers



History



Deployment overview

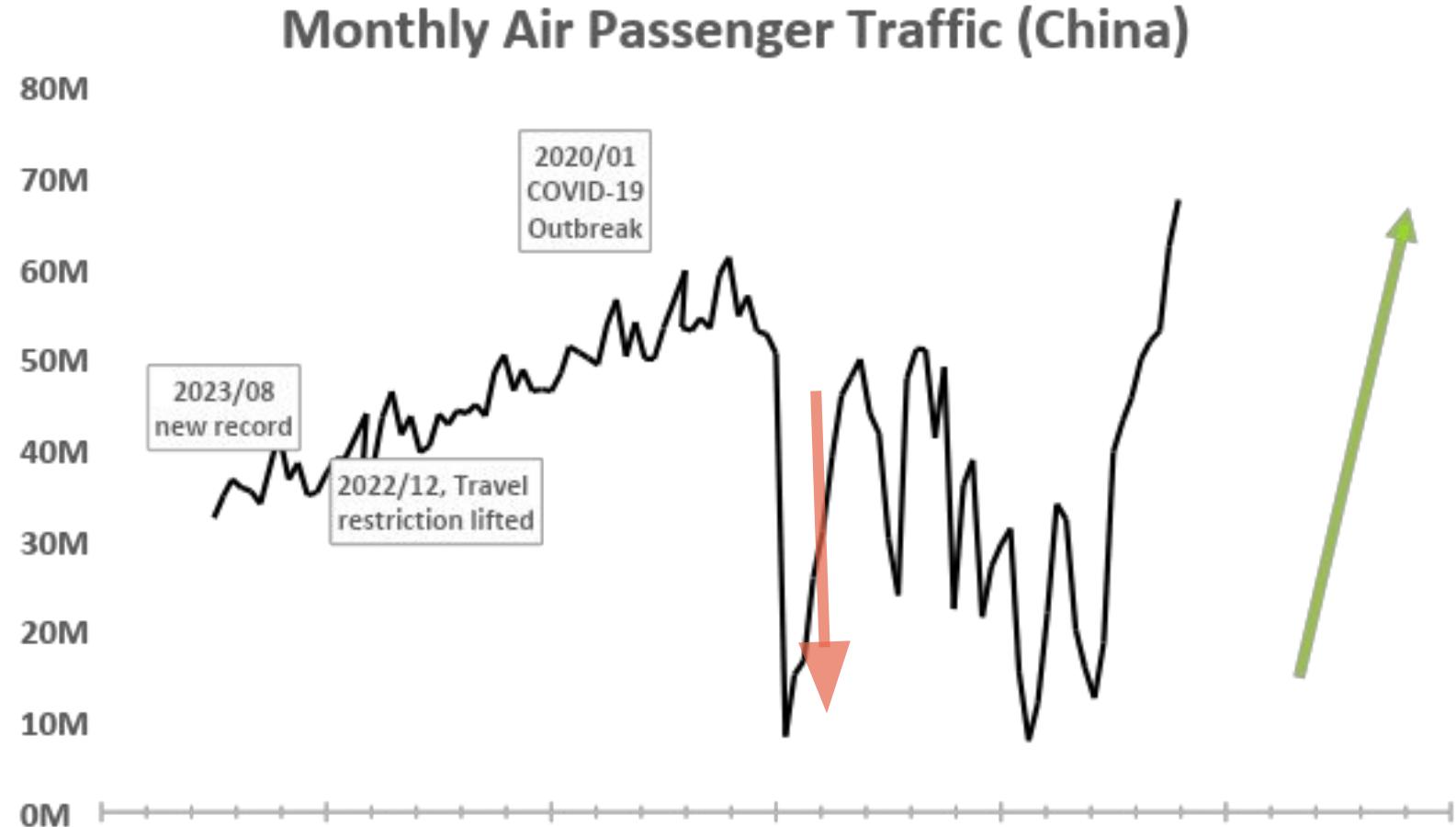


Cilium brings us

- **Cluster scalability (up 6000+ nodes/cluster)**
- **Data plane stability (4+ years in PROD)**
- **K8s-Native features (network policy, LB)**
- **Network observability**
- **Consistent experience (user & admin, tech stack, CNI)**

Embracing change

- Cost Efficiency
- Auto-scaling
- Capacity



Data source: Civil Aviation Administration of China

Cilium on Alibaba Cloud

3K ~ 5K

Auto-scaling nodes

10K~40K

Pods

100 Nodes

concurrent creation supported

<1 mln

From pod creation to running on a triggered node

Challenges

- **Node/pod scaling performance**
 - Cilium operator: slow ENI resync during IP allocation
 - Cilium agent: delayed IP deficit status report
 - Scheduling: IP availability not considered by kube-scheduler
- **Control plane resiliency**
 - Potential thundering herd
 - Network policy degradation plan
- **Early user of *--ipam=alibabacloud***
 - Encountered panic on the first run
 - Bug fixes

ENI IPAM Architecture

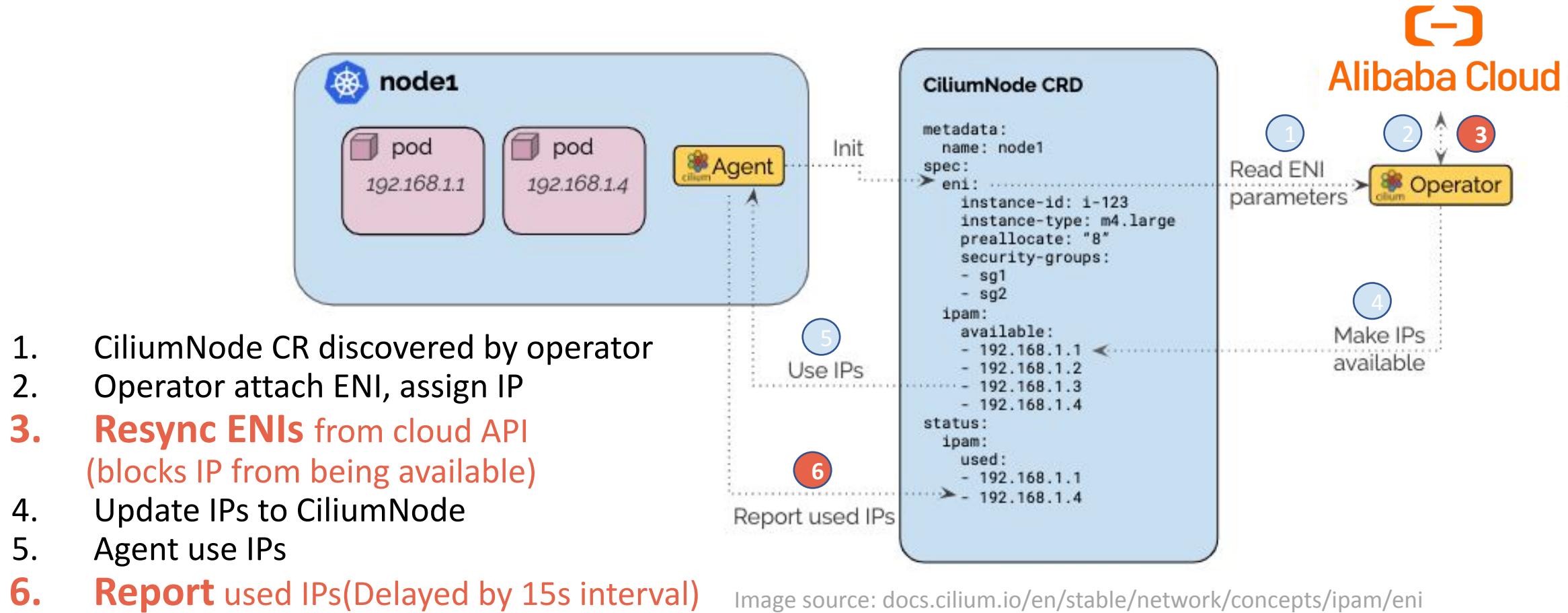


Image source: docs.cilium.io/en/stable/network/concepts/ipam/eni

Operator improvements

1. Drop unused subnet filter in ENI resync

```
for subnet in subnets {  
    // List ENIs from individual subnets  
    DescribeNetworkInterfaces(subnet)  
}
```



```
// List all ENIs  
DescribeNetworkInterfaces()
```

Resync Time significantly reduced



API requests reduced



Data from a cluster with 1500 ENIs, 40+ subnets

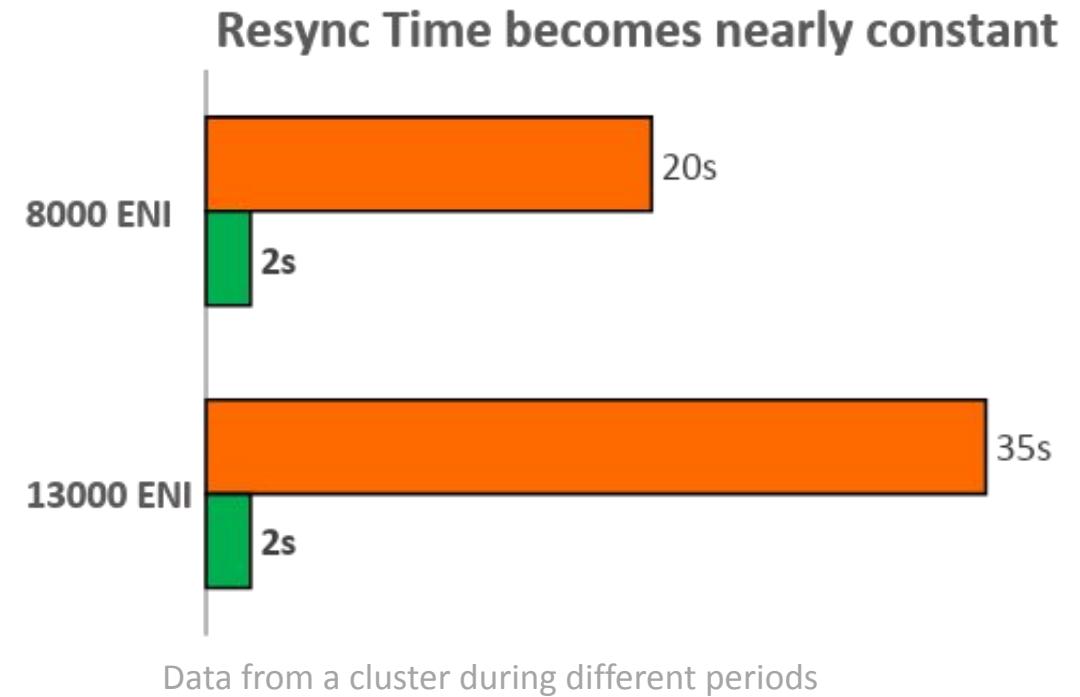
Operator improvements

2. Resync ENIs selectively on relevant instance

```
// List all ENIs  
DescribeNetworkInterfaces()
```



```
// List ENIs on instance  
DescribeNetworkInterfaces(instance)
```



Agent improvements

Accelerate agent startup and node bootstrap

- Add a new flag `--ipam-cilium-node-update-rate` to configure the delay of IP usage report (hardcoded as 15s previously)
- Fix multiple issues that cause agent fatal on initialization
 - Race condition in CiliumNode update
 - Enlarge cloud API quotas to avoid API rate limit

Scheduling Problems

- **IP address availability is not considered by kube-scheduler**
 - Pod could stuck in *ContainerCreating* while other nodes are available
<https://github.com/kubernetes/enhancements/issues/611>
- **Pod IP pre-warming is limited by IP address space**
- **Different node types have different IP capacities**
 - PodCIDR /24, /25, /26; Cloud provider flavors...

IP capacity-aware Scheduling

Solution: *Extended Resources + Device Plugin*

1. Device plugin read IPAM status from cilium-agent

```
$ curl --unix-socket /path/to/cilium.sock http://localhost/v1/healthz | jq -r .ipam.status
IPv4: 2/10 allocated,
```

2. Calculate & Advertise IP capacity to apiserver

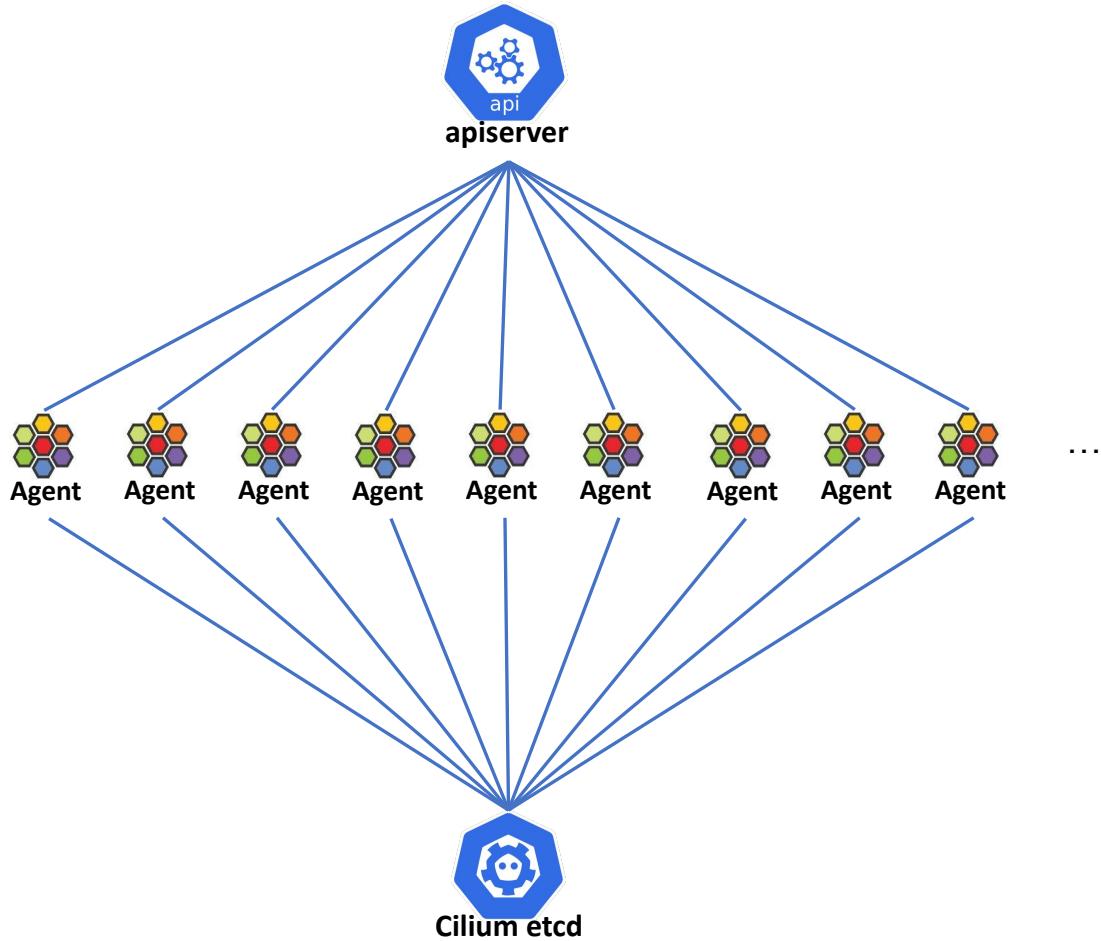
```
kind: Node
...
status:
  allocatable:
    example.com/ip: "8"
    cpu: "40"
  ...

```

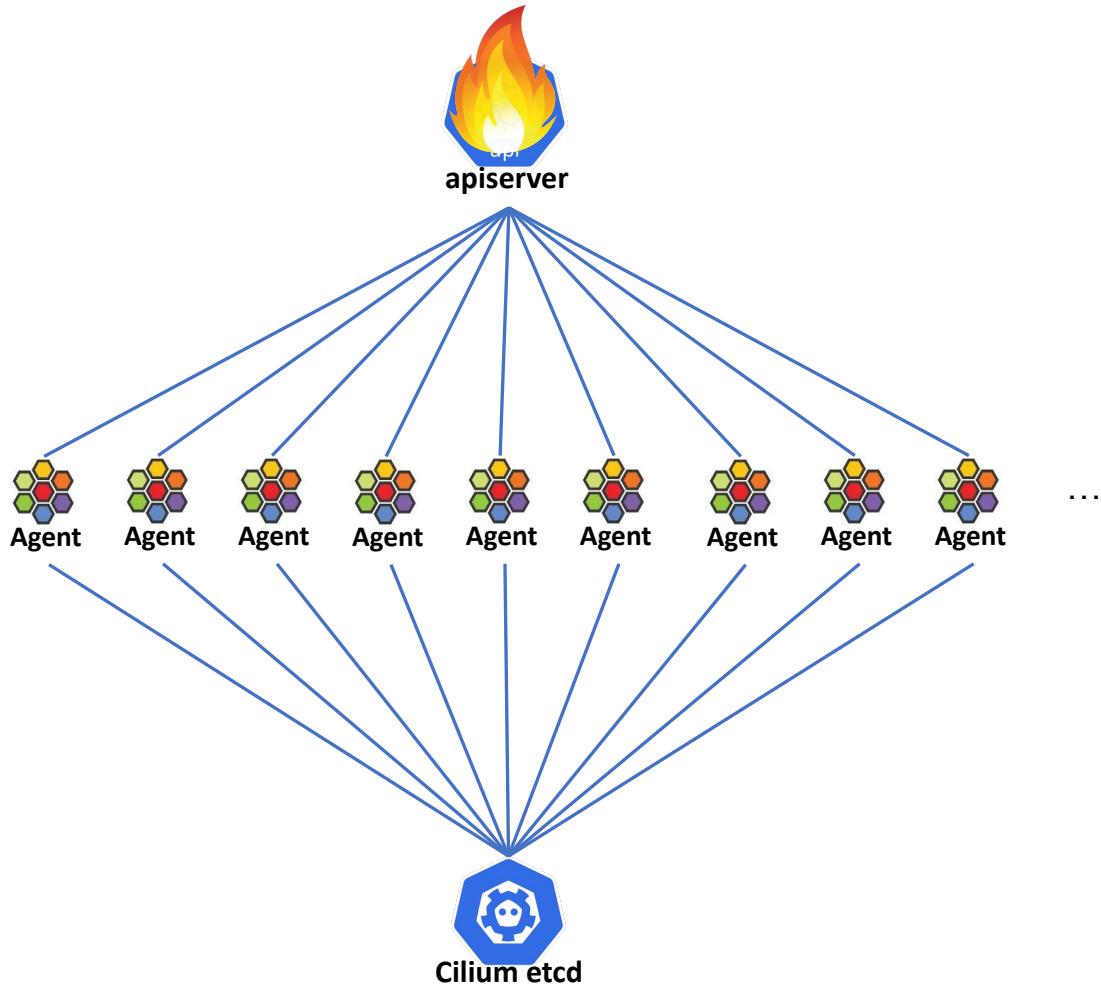
3. Webhook injector adds IP requests to non-hostNetwork pod spec

```
resources:
  limits:
    example.com/ip: "1"
    cpu: "4"
    memory: 8Gi
  requests:
    example.com/ip: "1"
    cpu: "4"
    memory: 8Gi
```

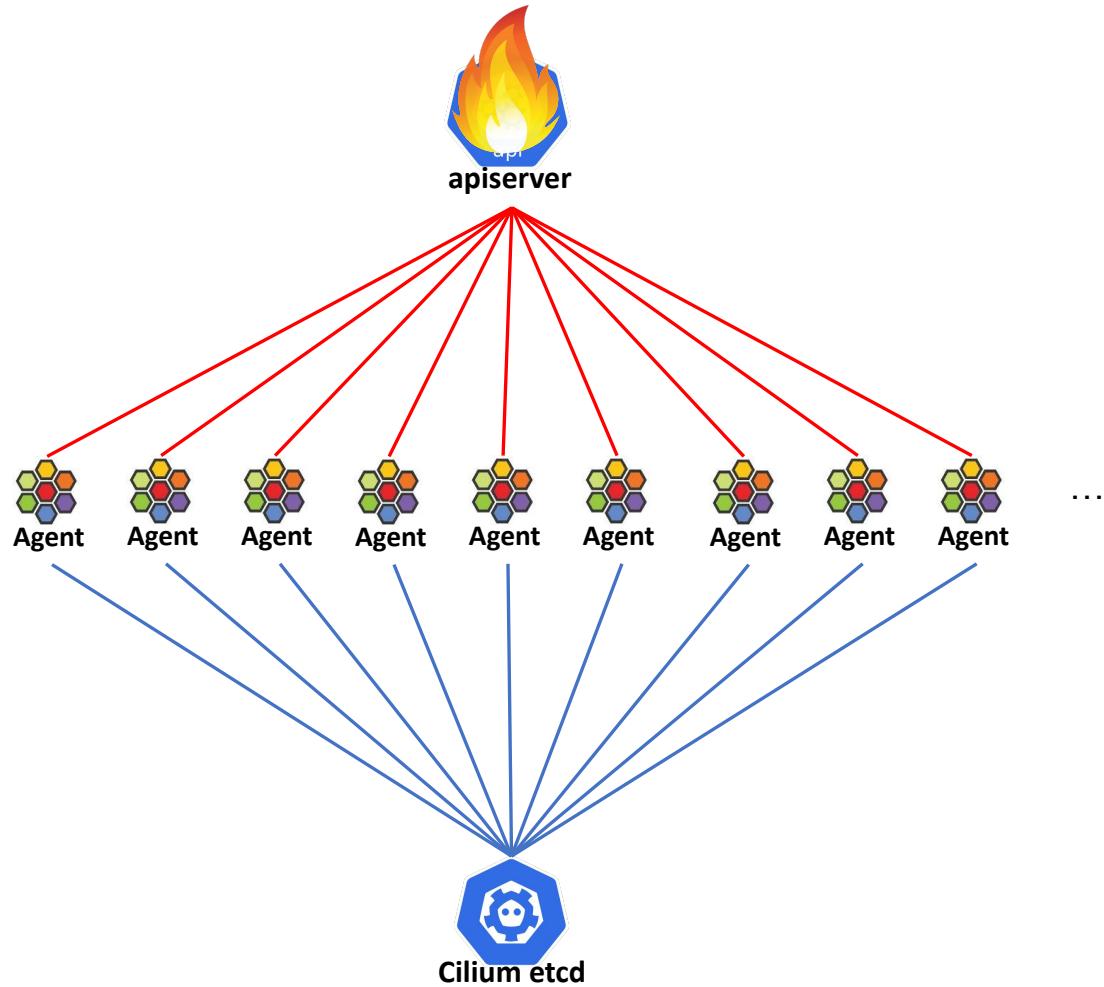
Thundering herd



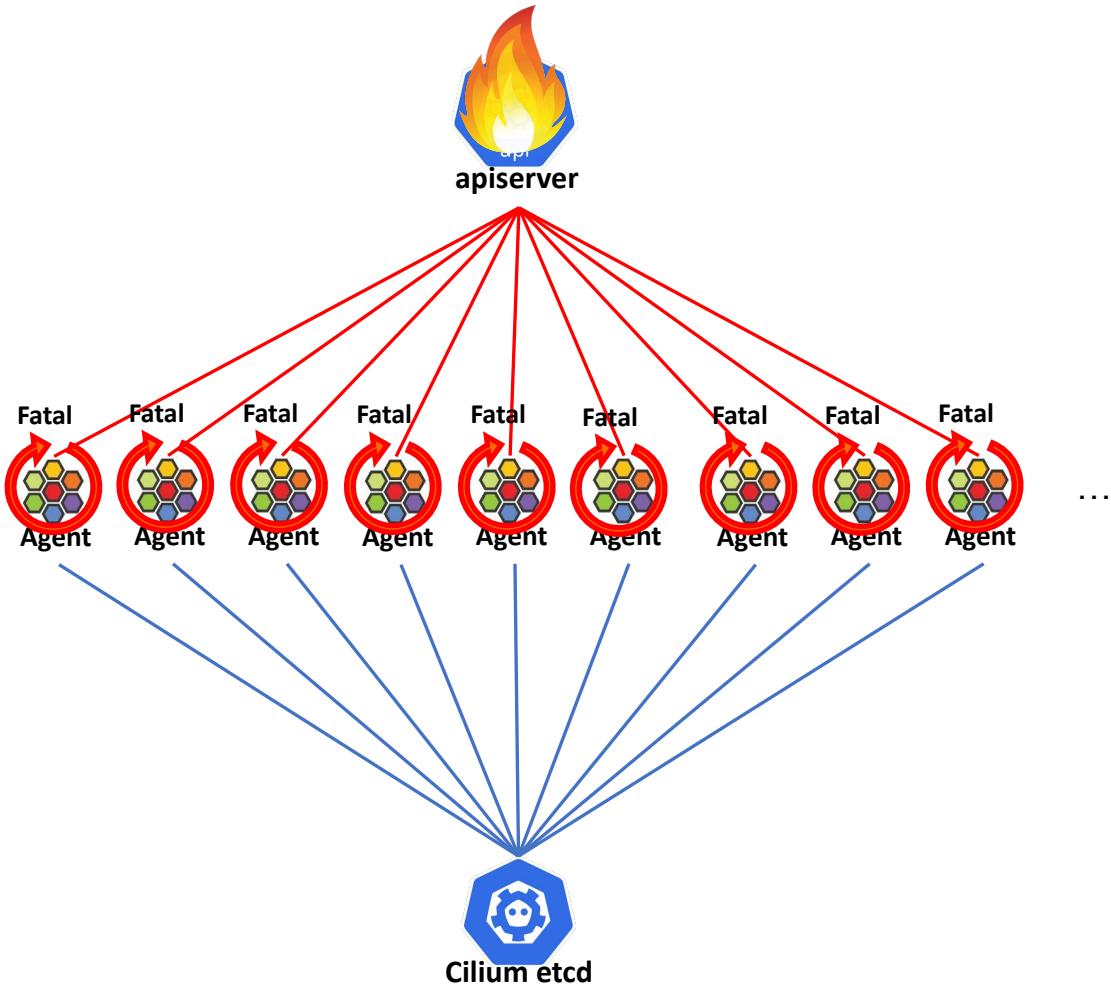
Thundering herd



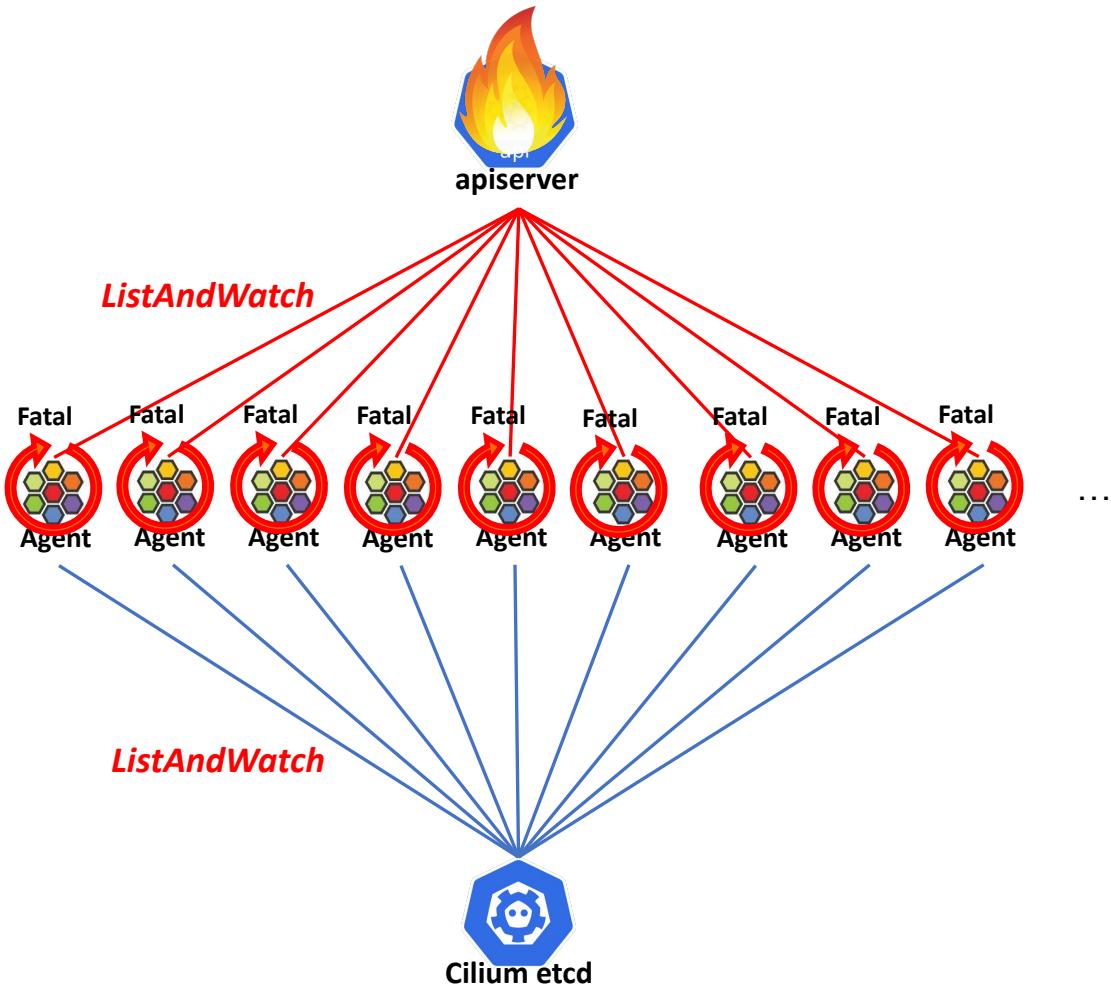
Thundering herd



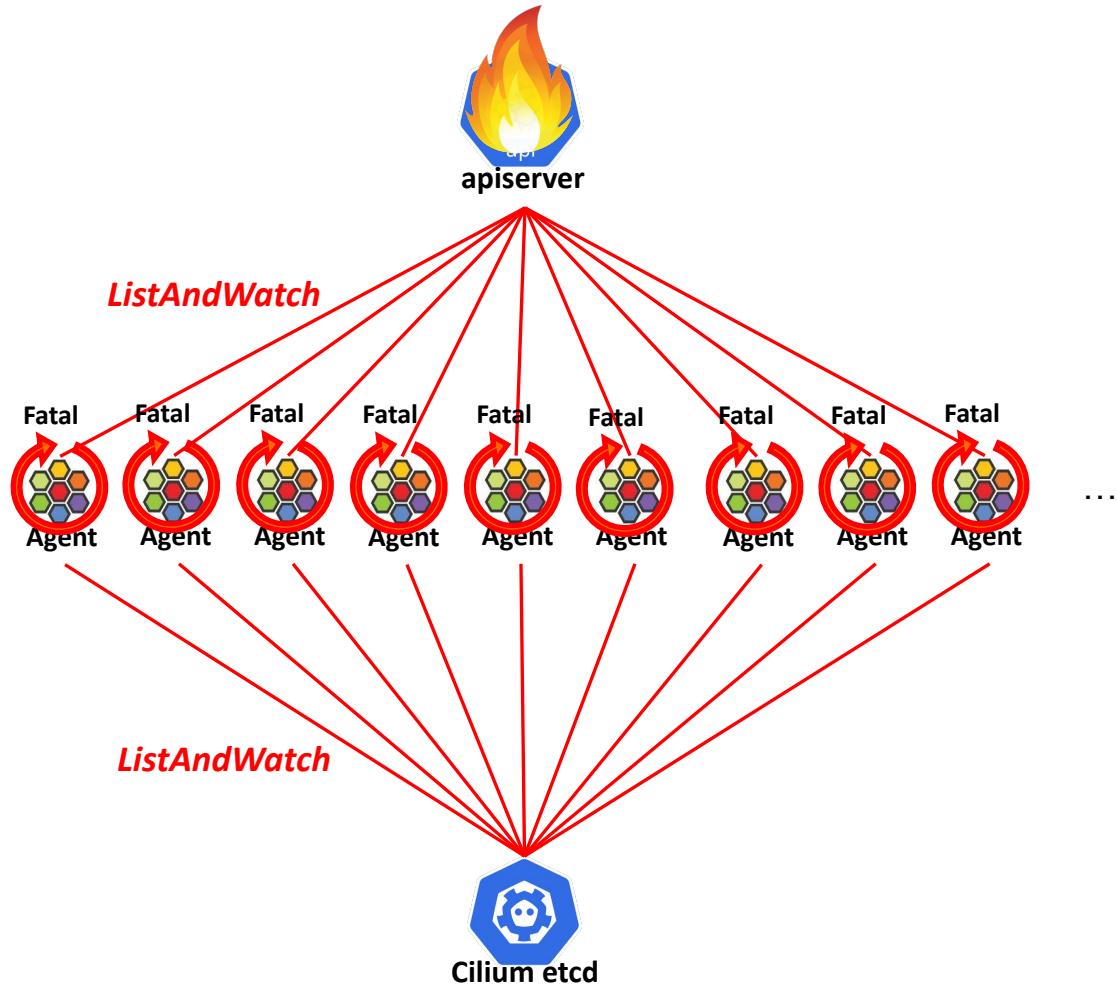
Thundering herd



Thundering herd



Thundering herd



Thundering herd

Limitations of Pod container

- **Exponential backoff**
 - Backoff step too small for a cluster of agents: 10s, 20s, 40s, ... capped at 5m
- **Randomness**
 - Restart policy doesn't support jitter

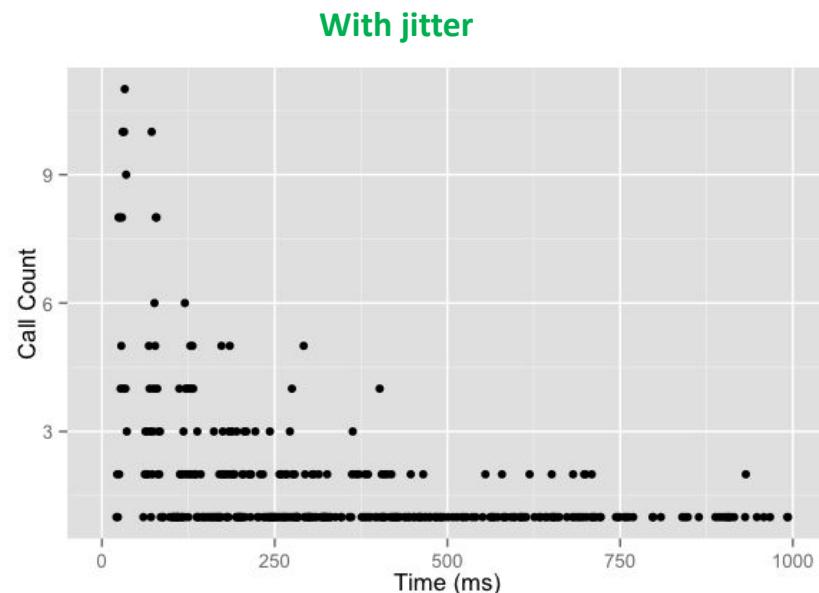
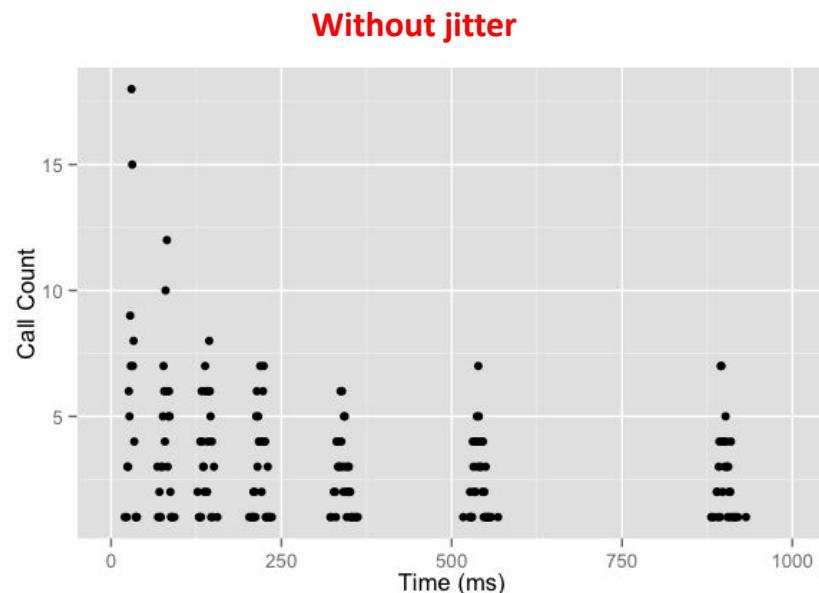
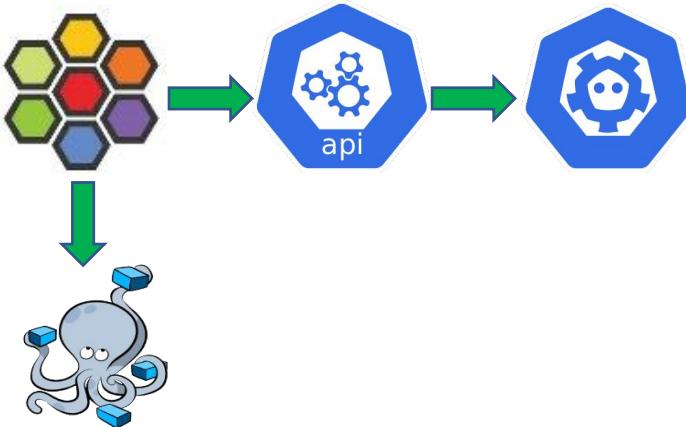
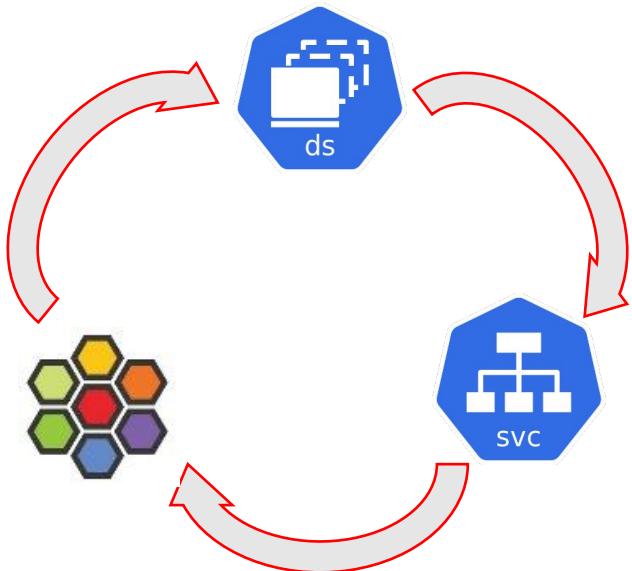


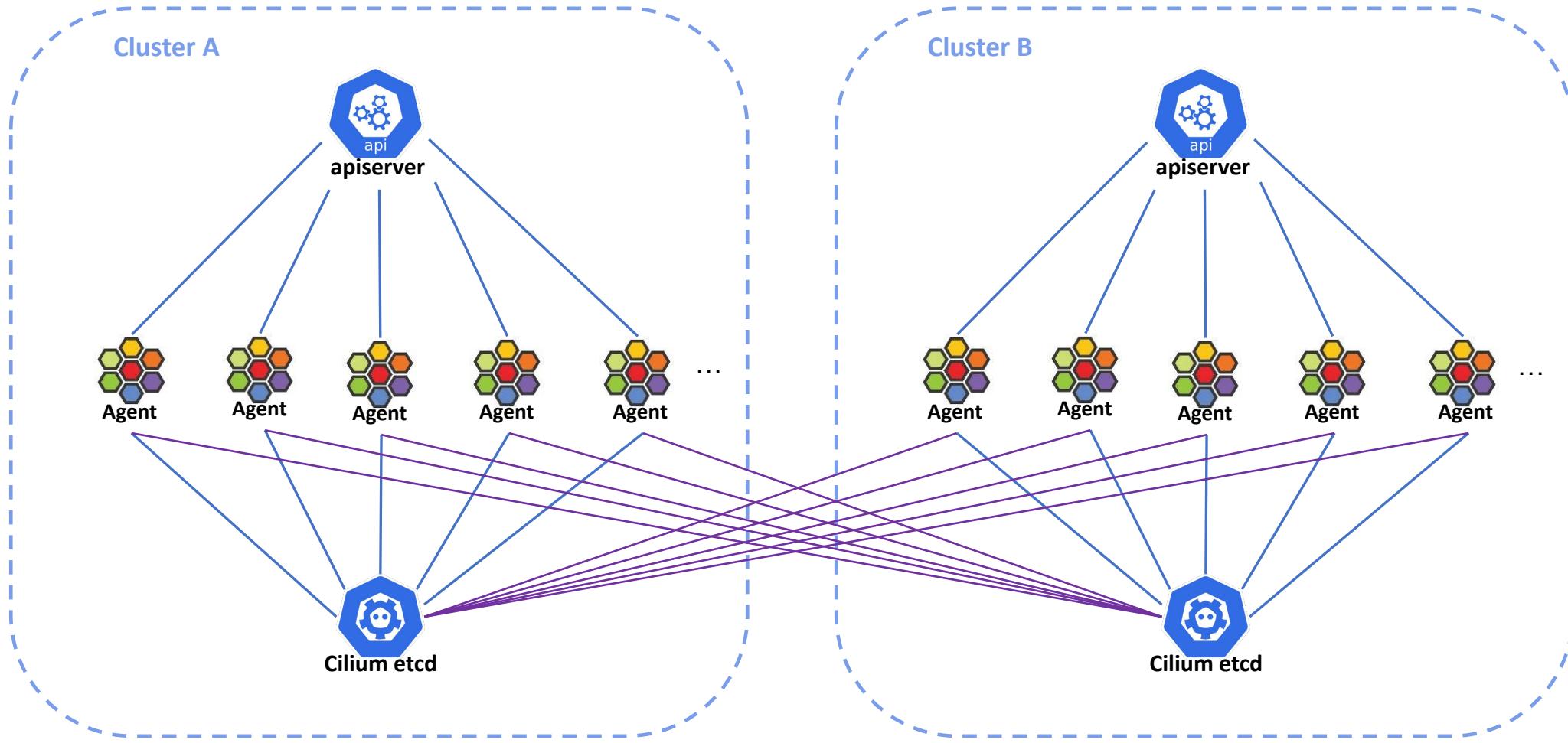
Image source: aws.amazon.com/cn/blogs/architecture/exponential-backoff-and-jitter/

Docker Compose

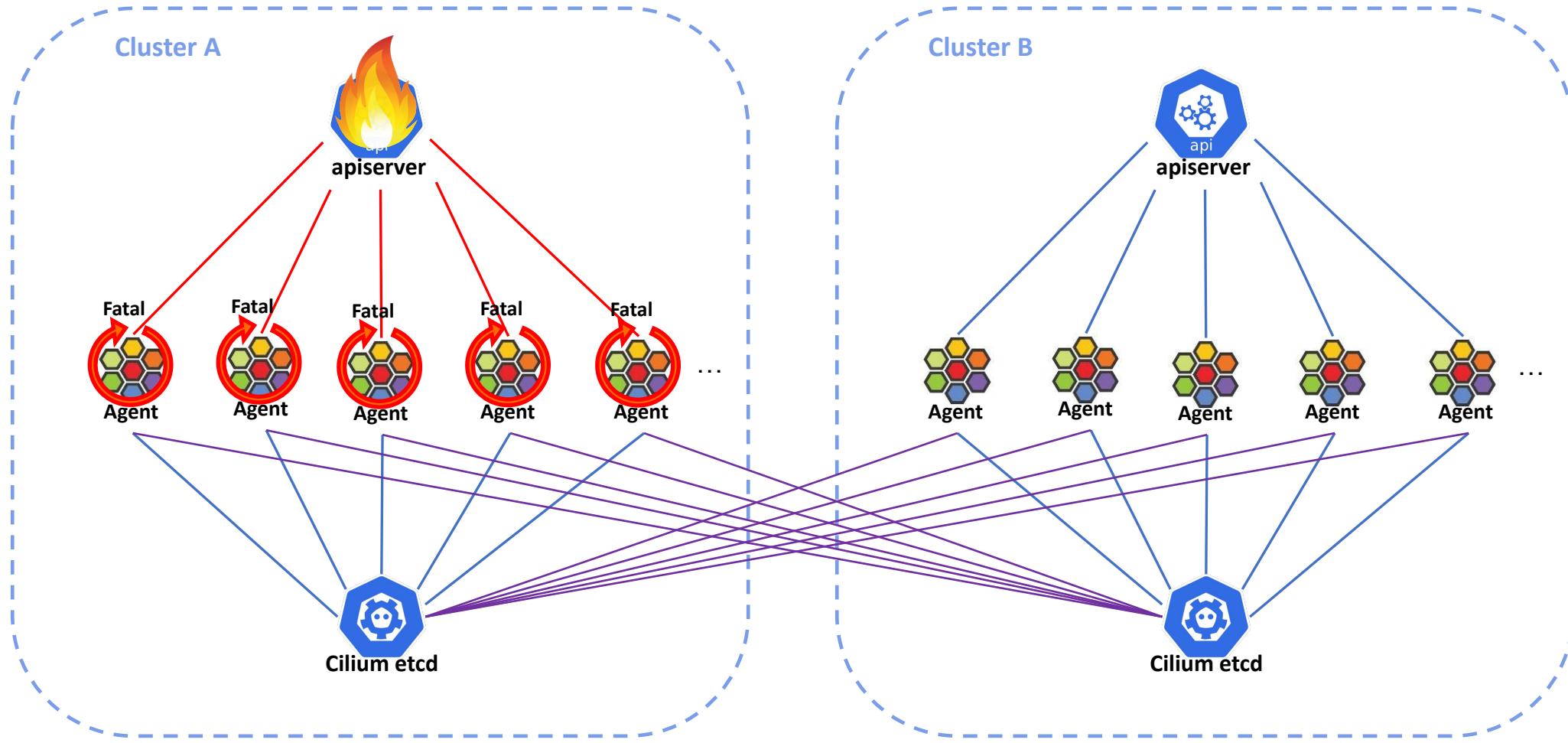
- Full control of custom backoff step and jitter
- Better control over canary release
- Decouple deployment, avoid chicken-egg problems
 - imagine creating an agent pod depends on a Webhook service (rare but possible)



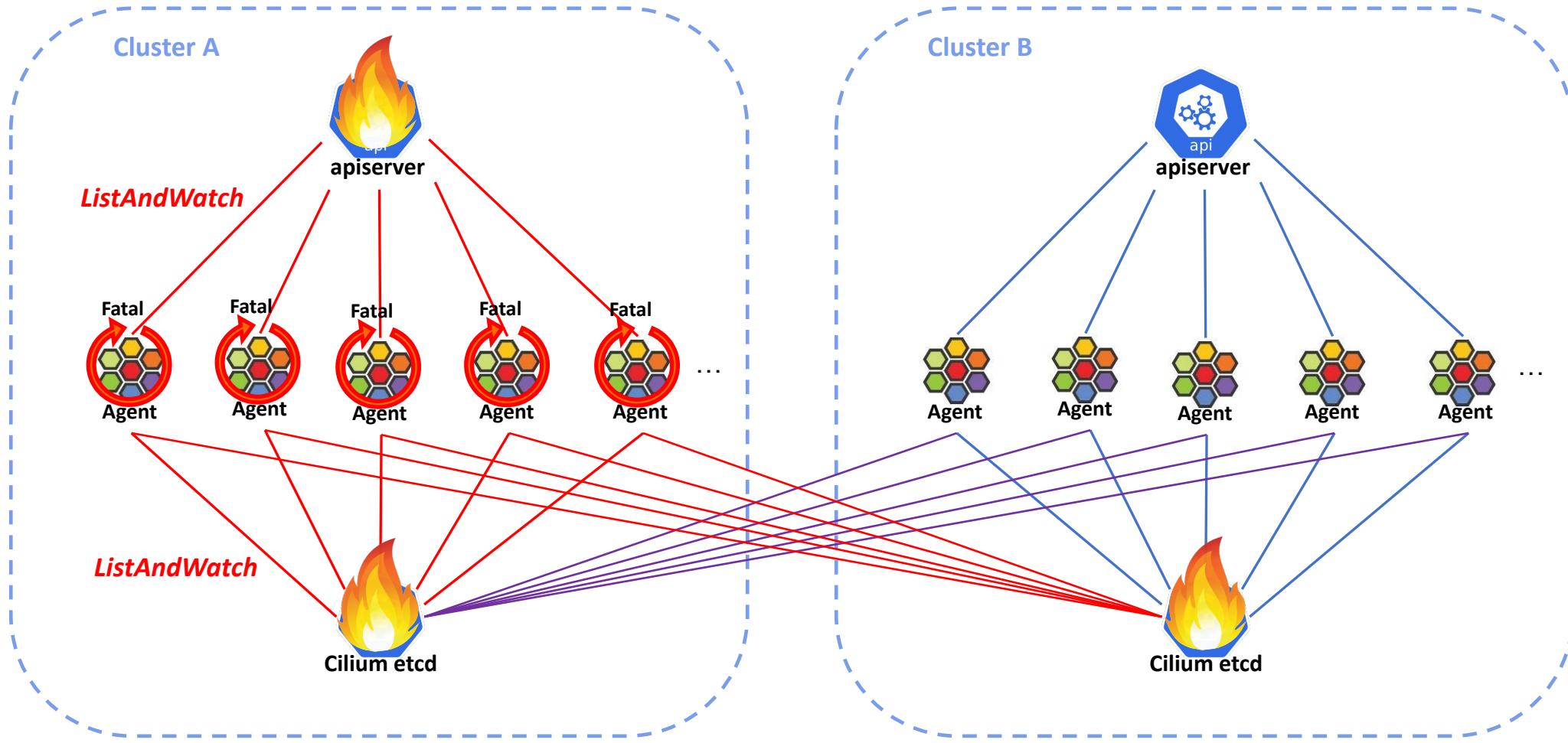
Fault propagation



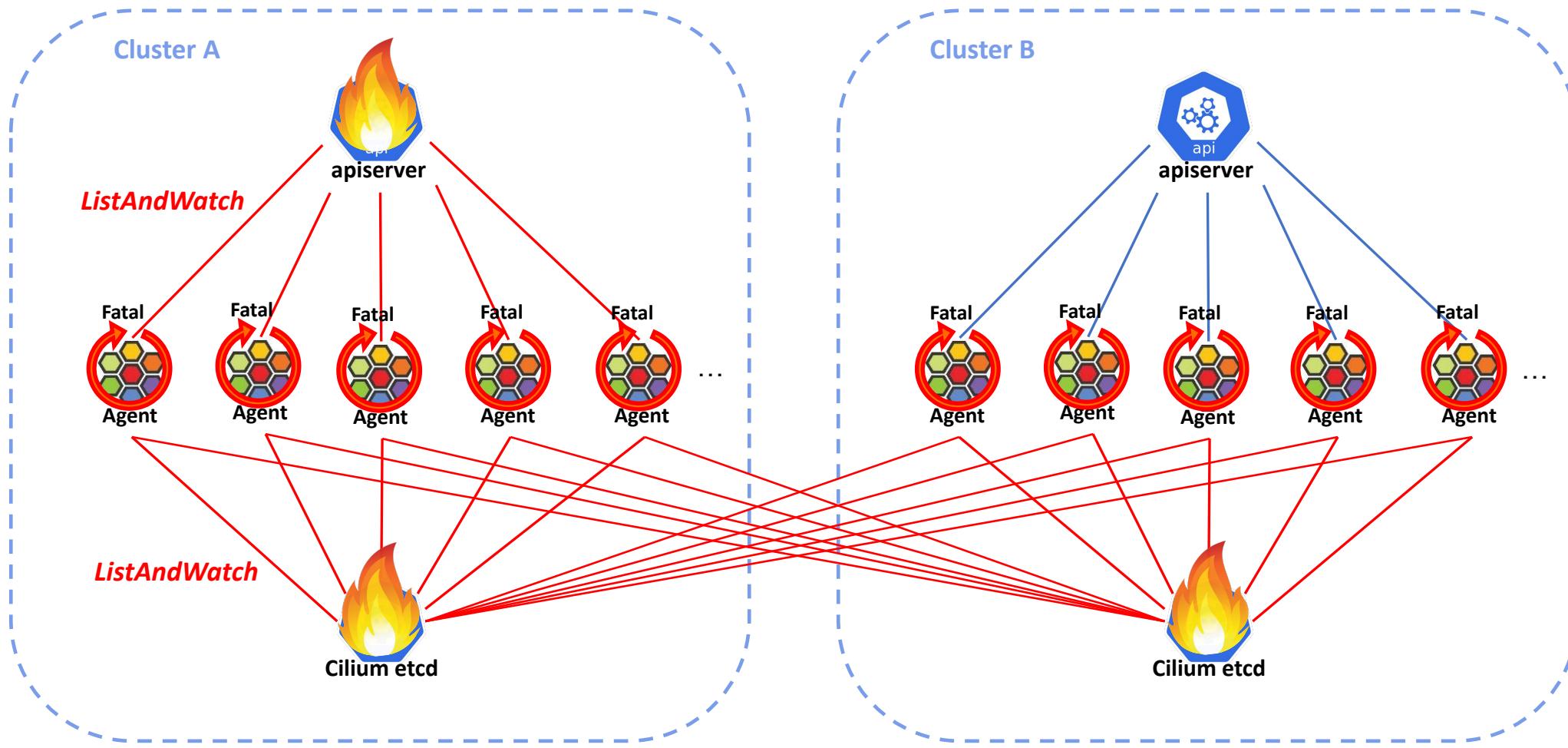
Fault propagation



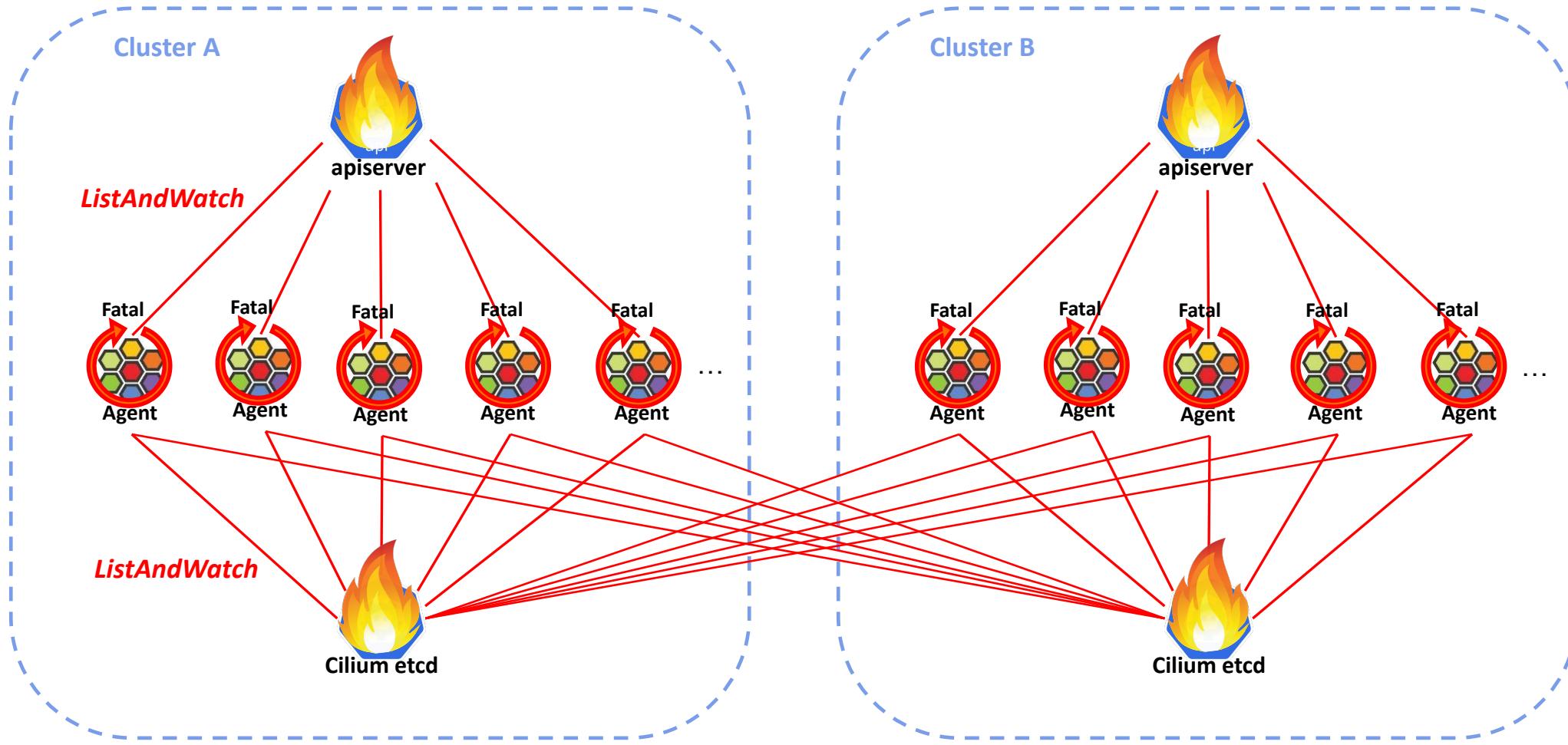
Fault propagation



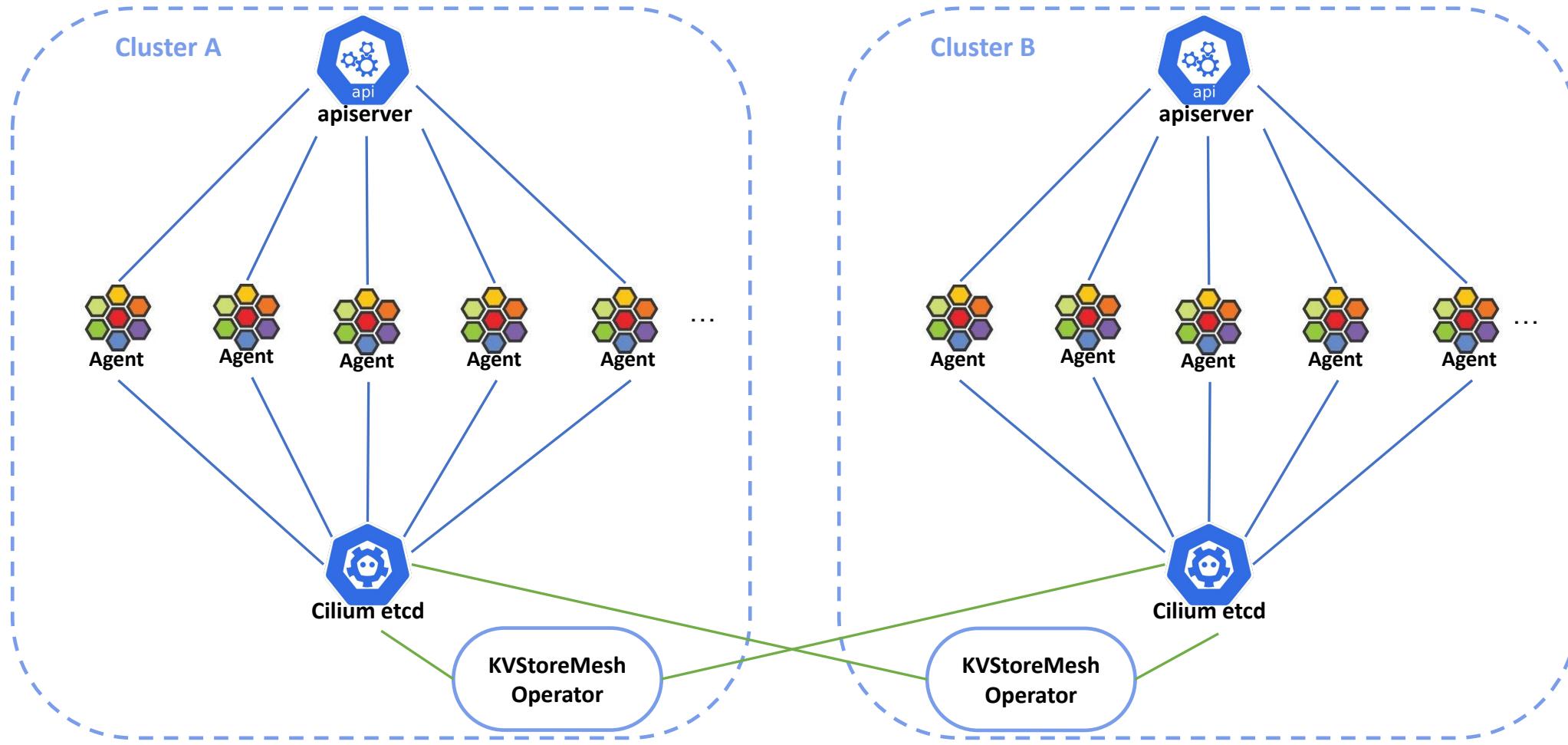
Fault propagation



Fault propagation



Solution: KVStoreMesh



KVStoreMesh has also been implemented in upstream version 1.14, currently in beta.

Network policy degradation



Last resort for policy degradation when control plane and agent are down

Key takeaways

- Prefer Kernel patch version > 100
- Pay attention to cloud API quotas
- For big subnets, pay attention to ARP table capacity
net.ipv4.neigh.default.gc_thresh{1,2,3}
- Control retry behaviors
- Incident action plan

Thank you!