SYDNEY

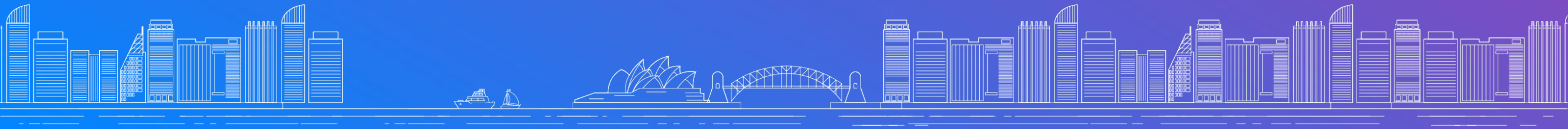
KCD

GROWING CLOUD NATIVE TOGETHER

5 & 6 SEPTEMBER

From Bootstrap to Scale: Optimising Kubernetes Operators

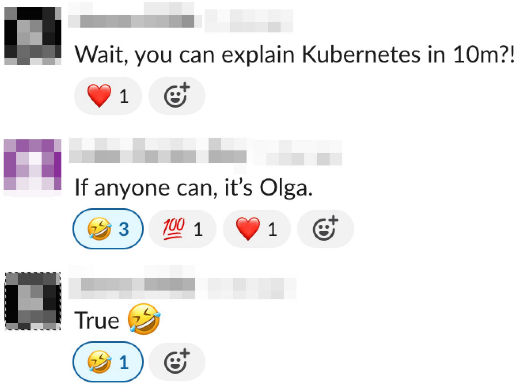
Olga Mirensky - Platform Engineer - ANZ Plus



Introduction and Agenda

I am Olga. My teammates believe that I can explain Kubernetes in 10 minutes.

Can I explain operators in 20 minutes?



[redacted]

Wait, you can explain Kubernetes in 10m?!

❤️ 1

[redacted]

If anyone can, it's Olga.

🤣 3 💯 1 ❤️ 1

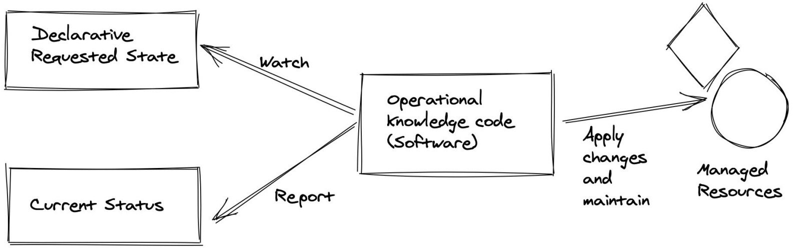
[redacted]

True 🤣

🤣 1

Agenda: What is operator and why build one? Operators in ANZ. Scaling challenges. Frameworks, optimisations and what's new.

What is an Operator and Why Build One


Declarative Requested State
Watch
Operational knowledge code (Software)
Apply changes and maintain
Managed Resources
Report
Current Status

- KRM (Kubernetes Resource Model) + GitOps.
- Continuous reconciliation.
- Codify knowledge, best practices, controls.

Image source github.com/cncf/tag-app-delivery: Operator-WhitePaper_v1-0.md

Operator Stack (Go)

kubebuilder

High-level framework for building Kubernetes operators, provides scaffolding, utilities and patterns

controller-runtime

Abstraction for building controllers.

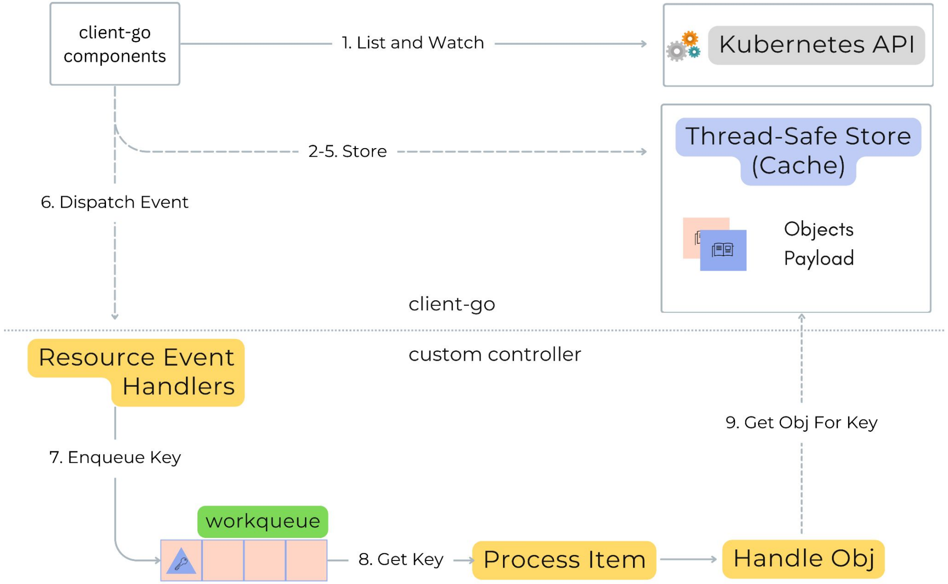
Cache, Manager, Builder, Handler

Client-go

Official Go client library for interacting with the Kubernetes API. Clientset, Informer, Rate Limiter

Controllers 101

Simplified version of github.com/kubernetes/sample-controller/blob/master/docs/controller-client-go.md



Challenges Scaling Operators

Horizontal Scaling

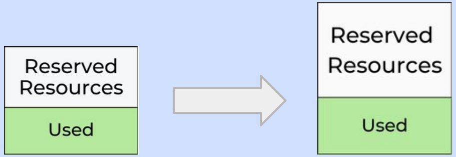
pod-1
pod-2
pod-n

Leader Election

<u>Solution Attempts</u>

FluxCD Sharding

Vertical Scaling



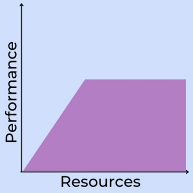
Reserved Resources

Used

Reserved Resources

Used

Little to no gain in performance beyond some point.



Concurrency and Rate Limiting

MaxConcurrentReconciles

GroupKindConcurrency

Internal Concurrency

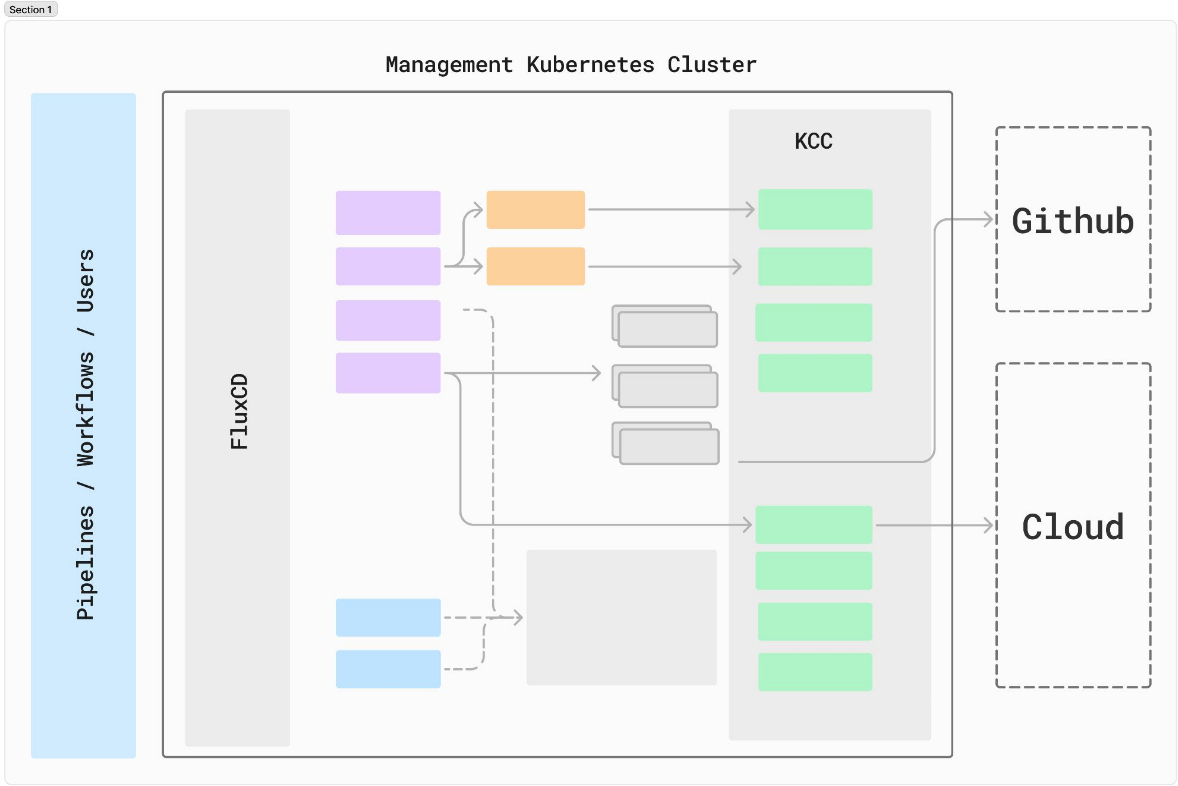
<u>Bottlenecks - rate limiting</u>

Client QPS / Burst

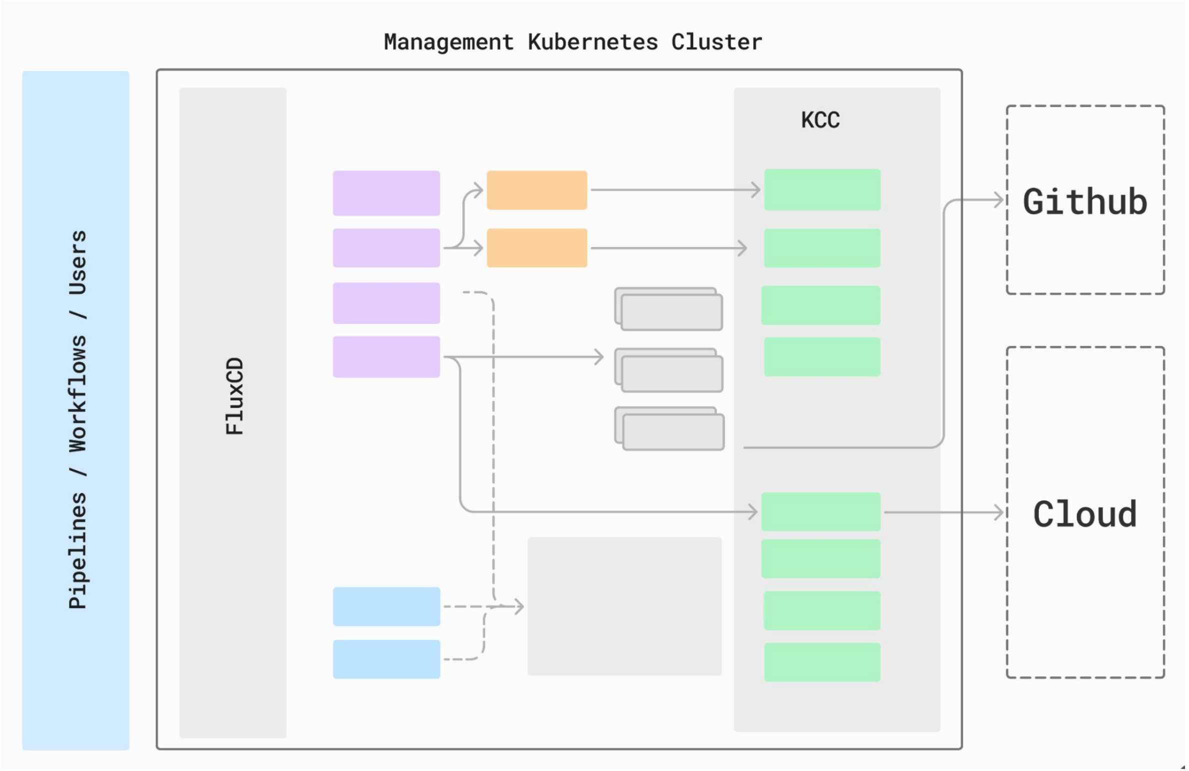
API Server QPS / Burst

External dependencies

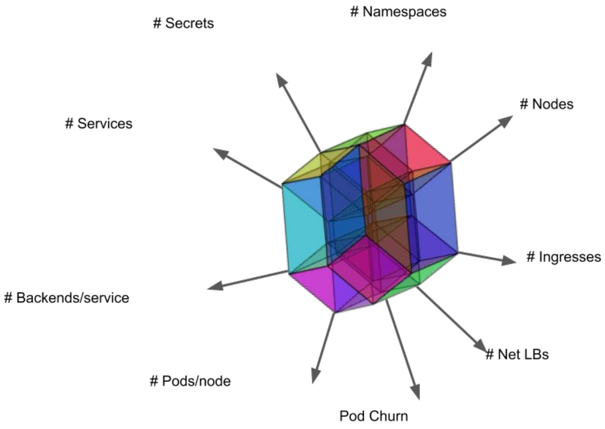
App in 60min - Powered by Operators



Kubernetes as a Platform Control Plane



Thresholds and Limits to Inform SLOs



Source of hypercube image: http://www.gregegan.net/APPLETS/29/29.html

SIG Scalability - Scalability Envelope

Scaling along one axis too far limits other dimensions.

Interdependencies: e.g. number of Services and number of backends in each Service.

Notice the churn as a dimension.

Thresholds Dimensions

Dimension	Kubebuilder Abstraction
Custom Resources	`For()`
Managed Resources	`Owns()`
Watched Resources AND / OR	`Watches()`
Enqueued Resources	`Enqueue…() called from Watches`
External Events	`N/A`

Payload Size

2000 x

```yaml
apiVersion: mygroup/v1
kind: mykind
spec:
- name: item1
  properties: obj
```

OR 1 x

```yaml
apiVersion: mygroup/v1
kind: mykind
spec:
- name: item1
  properties: obj
...
- name: item2000
  properties: obj
```

Example - URLMaps limits

2,000 host rules, path matchers

1,000 path rules per matcher

1,000 Hosts per host rule

...

```
pathMatcher[].defaultRouteAction.weightedBackendServices[].headerAction.requestHeadersToAdd[]
```

Capacity Planning - Load Testing

Testing Frameworks	Chainsaw	Kyverno	Declarative e2e test. Templating. No support for load test (yet?)
	Cluster Loader 2	kubernetes/perf-tests SIG Scalability	K8s load testing framework. Supports CR, measurements, chaos and more.

Emulation

Generic resource emulation

client-gen + fake client

Simulate fake/hollow nodes and pods

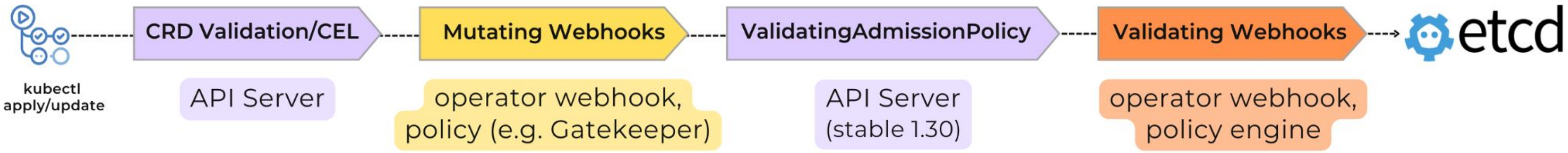
KWOK

Kubernetes Without Kubelet
(KWOK. SIG Testing)

Kubemark
(SIG Scalability)

Optimisation Opportunities

Fail Fast And Explicitly

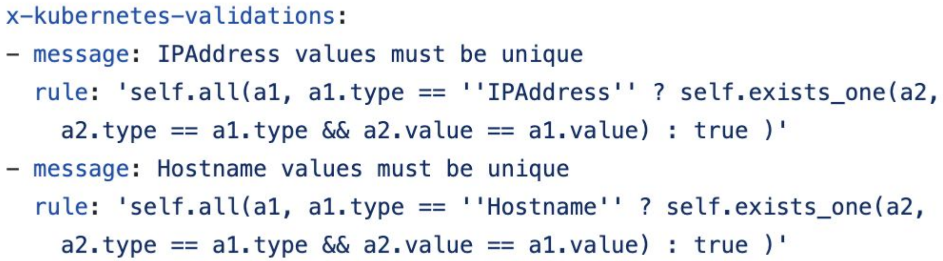


Offload work from the operator

Immediate feedback on failure

Cluster hygiene

CEL - Common Expression Language

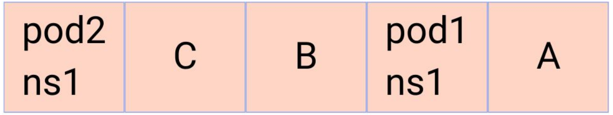


```
x-kubernetes-validations:
- message: IPAddress values must be unique
  rule: 'self.all(a1, a1.type == ''IPAddress'' ? self.exists_one(a2,
    a2.type == a1.type && a2.value == a1.value) : true )'
- message: Hostname values must be unique
  rule: 'self.all(a1, a1.type == ''Hostname'' ? self.exists_one(a2,
    a2.type == a1.type && a2.value == a1.value) : true )'
```

Example `XValidation` markers translated into CRD (gateway-api)

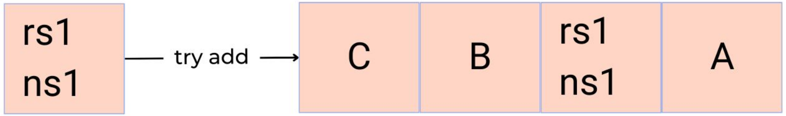
Workqueue Deduplication

Scenario 1: Enqueue each object



pod2 ns1	C	B	pod1 ns1	A

Scenario 2: Enqueue common key



rs1 ns1	try add →	C	B	rs1 ns1	A

Duplicate keys are not added

Enqueue common key

If reconciliation is exactly the same for different object

Typed Reconciler (experimental)

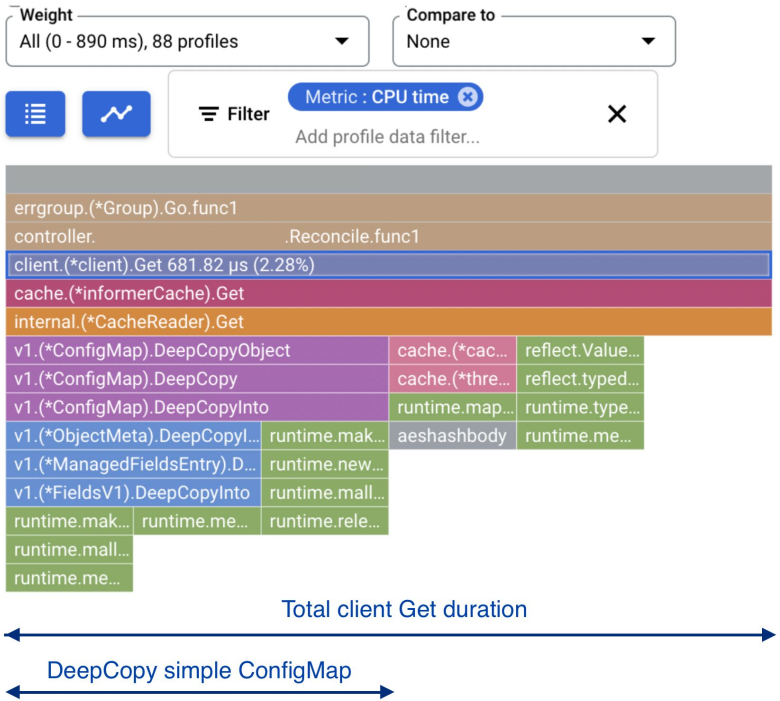
Controller-runtime -

examples/typed/main.go

Convert For to Watch

```
Named("myCR").
Watches(v1.myCR,EnqueueCustom).
```

Cache



Weight: All (0 - 890 ms), 88 profiles

Compare to: None

Filter: Metric : CPU time

Add profile data filter...
errgroup.(*Group).Go.func1
controller. .Reconcile.func1
client.(*client).Get 681.82 μs (2.28%)
cache.(*informerCache).Get
internal.(*CacheReader).Get
v1.(*ConfigMap).DeepCopyObject
cache.(*cac...
reflect.Value...
v1.(*ConfigMap).DeepCopy
cache.(*thre...
reflect.typed...
v1.(*ConfigMap).DeepCopyInto
runtime.map...
runtime.type...
v1.(*ObjectMeta).DeepCopyI...
runtime.mak...
aeshashbody
runtime.me...
v1.(*ManagedFieldsEntry).D...
runtime.new...
v1.(*FieldsV1).DeepCopyInto
runtime.mall...
runtime.mak...
runtime.me...
runtime.rele...
runtime.mall...
runtime.me...

Total client Get duration

DeepCopy simple ConfigMap

Less Payload ⇒ less time/CPU in deep copy

- `PartialMetadata`
- `TransformStripManagedFields`

`UnsafeDisableDeepCopy` (When safe)

- Entire Cache
- Per Group
- Per Object

Avoid Writes

- Dirty flag
- `CreateOrUpdate`

Field Indexers

speed up retrieval from large cache

What’s next

New Features

Typed Reconciler (experimental, v0.19)

NewQueue

NewCache

Advanced Settings

Controller-runtime source and pkg.go.dev:

Cache Options (cache.go)

Cache Options Design (cache_options.md)

Controller Options (controller.go)

SYDNEY

KCD

GROWING CLOUD NATIVE TOGETHER

5 & 6 SEPTEMBER

Questions & Feedback