UNIVERSITY OF CAMBRIDGE

Themelios: a model-checked reimplementation of Kubernetes

Andrew Jeffery, Richard Mortier

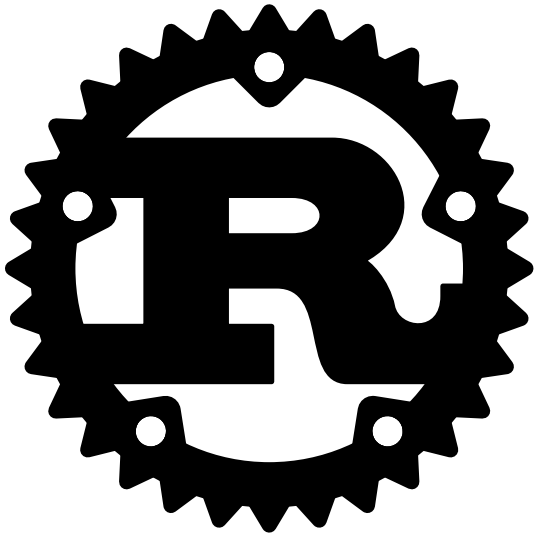
18th April 2024 @UKSys

Themelios: Breaking it down

A model-checked reimplementation of Kubernetes

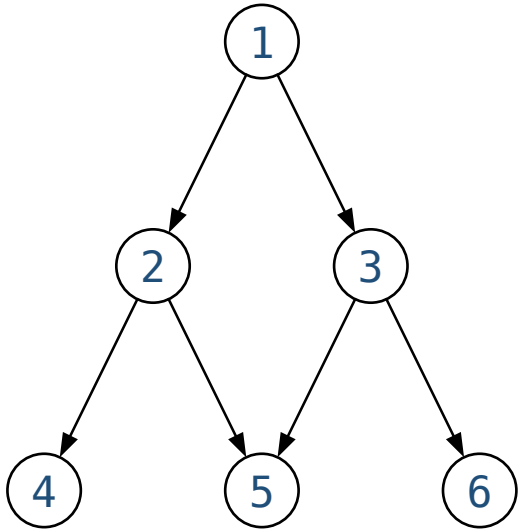
Themelios: Breaking it down

A model-checked reimplementation of Kubernetes



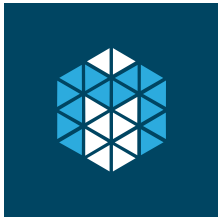
Themelios: Breaking it down

A model-checked reimplementation of Kubernetes



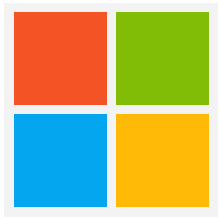
Kubernetes isn't the *only* one

Kubernetes



Mesos

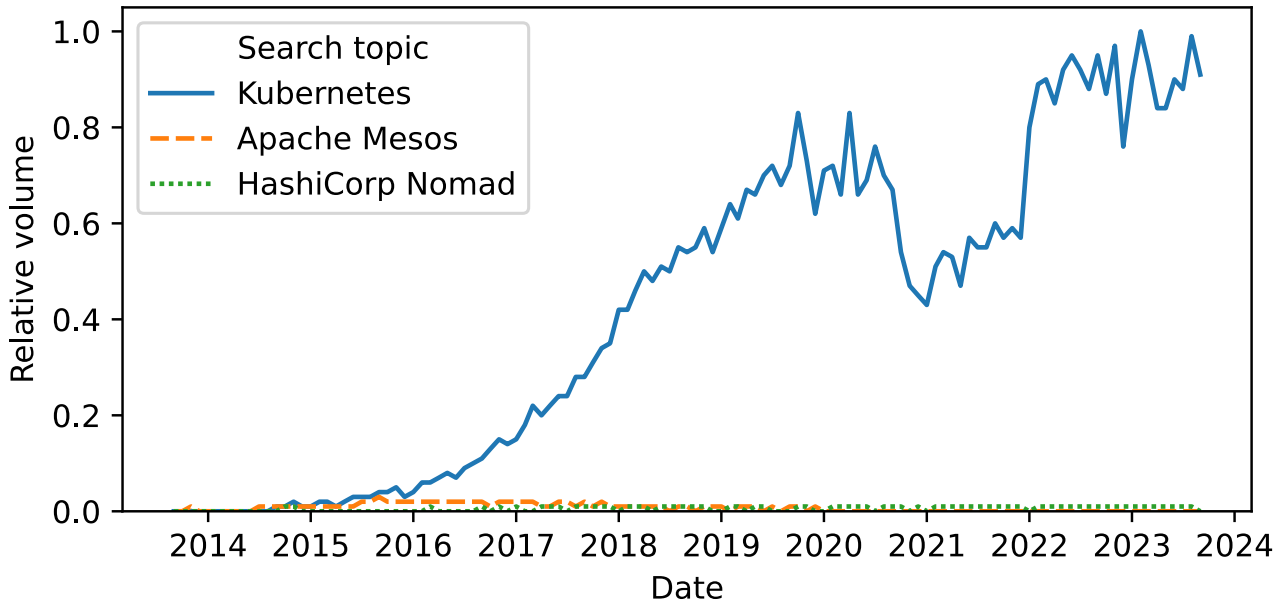
Twine



Autopilot

Borg

But Kubernetes is *the* one



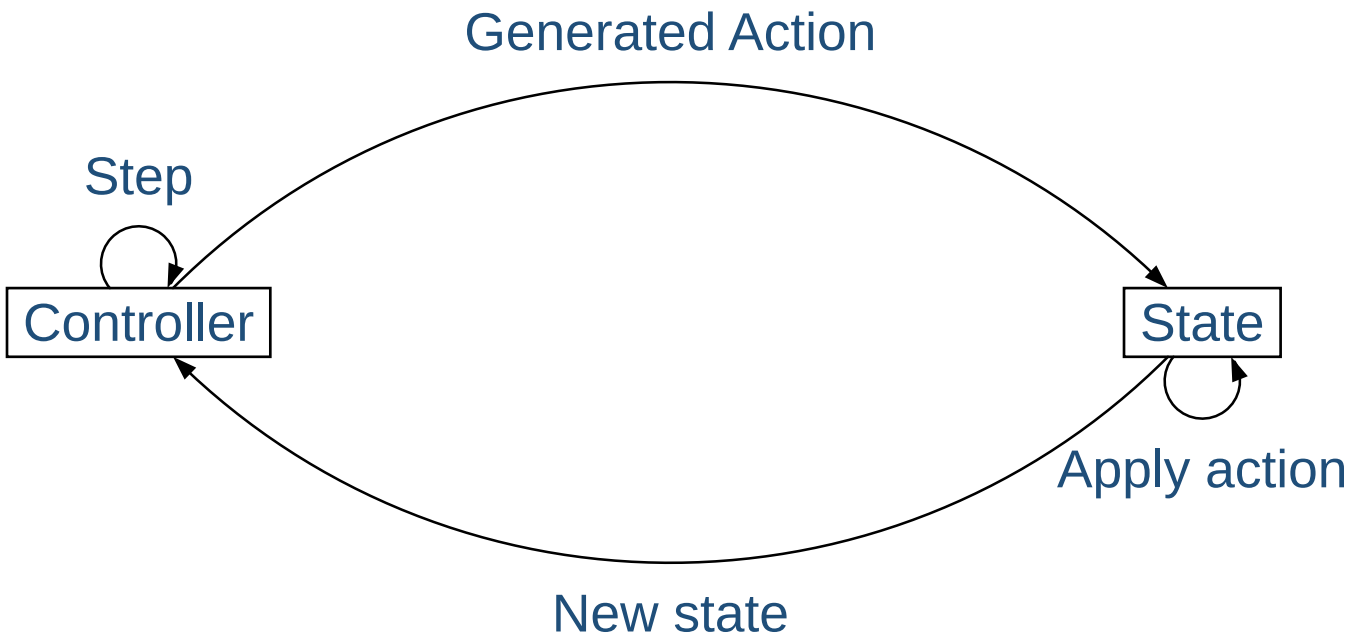
But what is going on?

Kubernetes is vulnerable to stale reads, violating critical pod safety guarantees #59848

Open

smarterclayton opened this issue on Feb 14, 2018 · 89 comments

The model


Generated Action
Step
Controller
State
Apply action
New state

The model — controllers

```
trait Controller {
  fn step(&mut self, // local state
          gs: &GlobalState) // etcd state
          -> Option<Action>; // what to do
}

// typically managed by an API server
impl GlobalState {
  fn apply(&mut self, action: Action) {
    // ...
  }
}
```

The model — state

State

Actions

```
{
  "pods": {},
  "nodes": {
    "node-0": []
  },
}
```

The model — state

State

```
{
  "pods": {
    "llama2": {
      "image": "hf/llama:2"
      "node": null,
    }
  },
  "nodes": {
    "node-0": []
  },
}
```

Actions

- `CreatePod("llama2")`

The model — state

State

```
{
  "pods": {
    "llama2": {
      "image": "hf/llama:2"
      "node": "node-0",
    }
  },
  "nodes": {
    "node-0": []
  },
}
```

Actions

- `CreatePod("llama2")`
- `SchedulePod("llama2", "node-0")`

The model — state

State

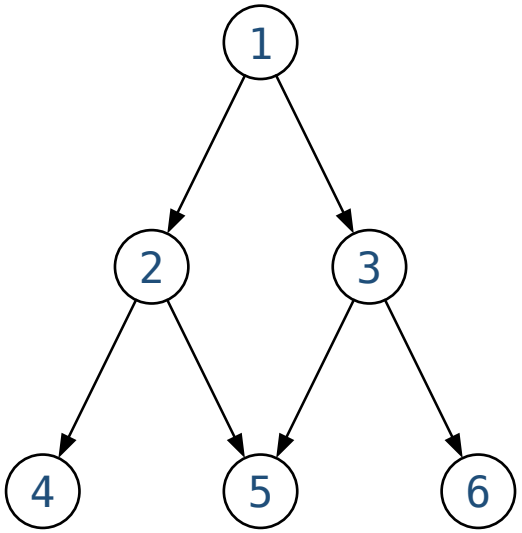
```
{
  "pods": {
    "llama2": {
      "image": "hf/llama:2"
      "node": "node-0",
    }
  },
  "nodes": {
    "node-0": ["llama2"]
  },
}
```

Actions

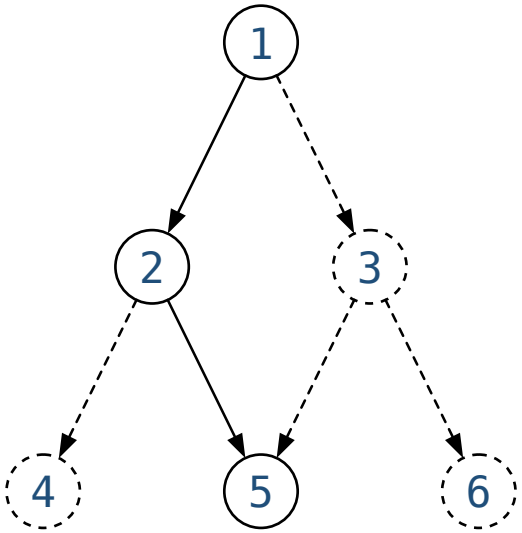
- `CreatePod("llama2")`
- `SchedulePod("llama2", "node-0")`
- `RunPod("node-0", "llama2")`

What can we do with it? Check it!

Exhaustive



Simulation

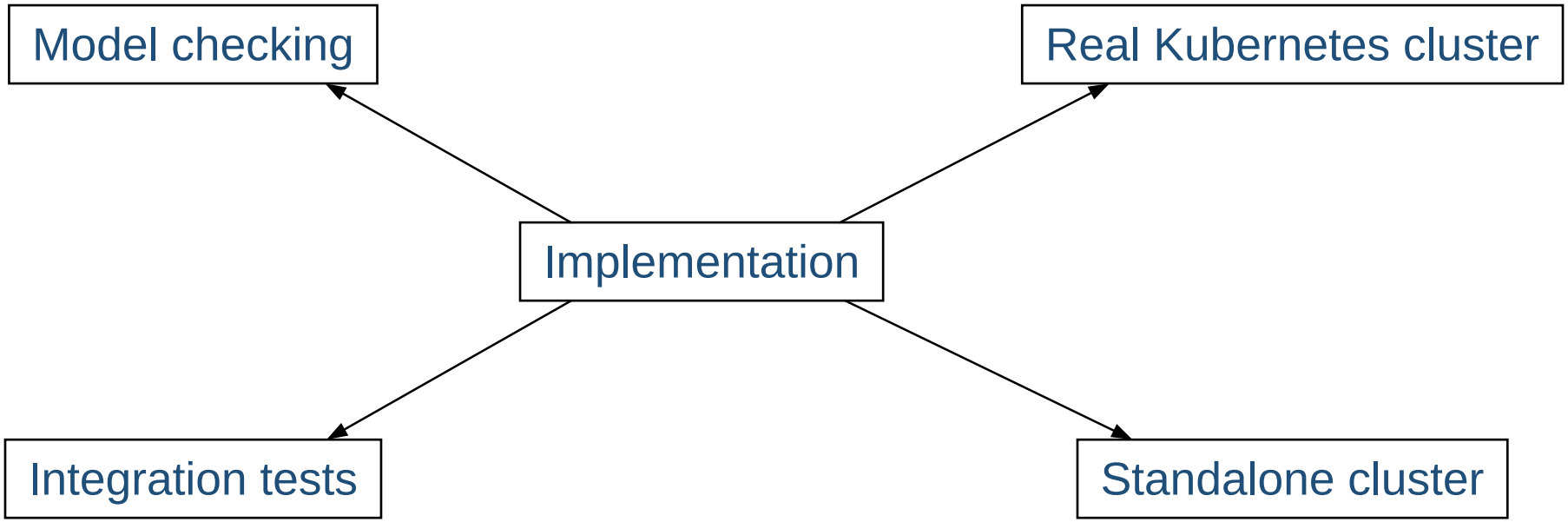


For the stale reads property:

Pods running on Nodes **always** have unique names across the cluster

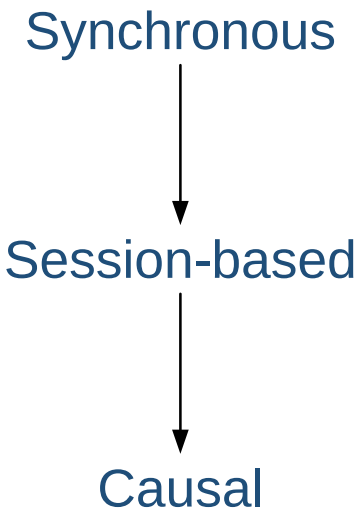
What can we do with it? Run it!

Single, checked implementation is used everywhere



What can we do with it? Adapt it!

Weaken the consistency model
(swap out etcd)



And more!

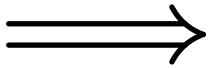
- Change the architecture
- Add more controllers
- Change the existing controller behaviour

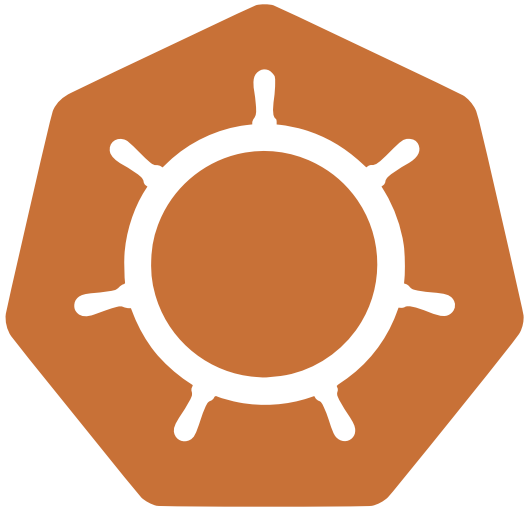
And how fast does it check?

Consistency	Total states explored	μs per state
Synchronous	24,758,530	2.464
Monotonic Session	23,819,028	2.561
Resettable Session	23,461,999	2.600
Causal	10,100,172	6.040

Best results of checking runs of 60s, max depth of 100

Let's start building strong foundations





Kubernetes

Themelios