



KubeCon



CloudNativeCon

North America 2023





KubeCon



CloudNativeCon

North America 2023

SIG Scalability: Intro + DeepDive

Wojciech Tyczynski, Google (SIG Scalability TL)

Marcel Zieba, Isovalent (Sig Scalability Chair)

What do we do?



KubeCon



CloudNativeCon

North America 2023

- Define & Drive
 - Coordinate & Contribute
 - Monitor & Measure
 - Preserve & Protect
 - Consult & Coach
- scalability definition & goals
 - performance improvements
 - performance of the system
 - from scalability regressions
 - community about scalability

Not to confuse with SIG Autoscaling!

What is Kubernetes Scalability?



KubeCon



CloudNativeCon

North America 2023

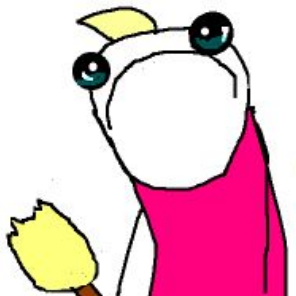
WHAT DO WE WANT?



SCALABLE CLUSTERS!



WHAT DOES IT MEAN?



What is Kubernetes Scalability?



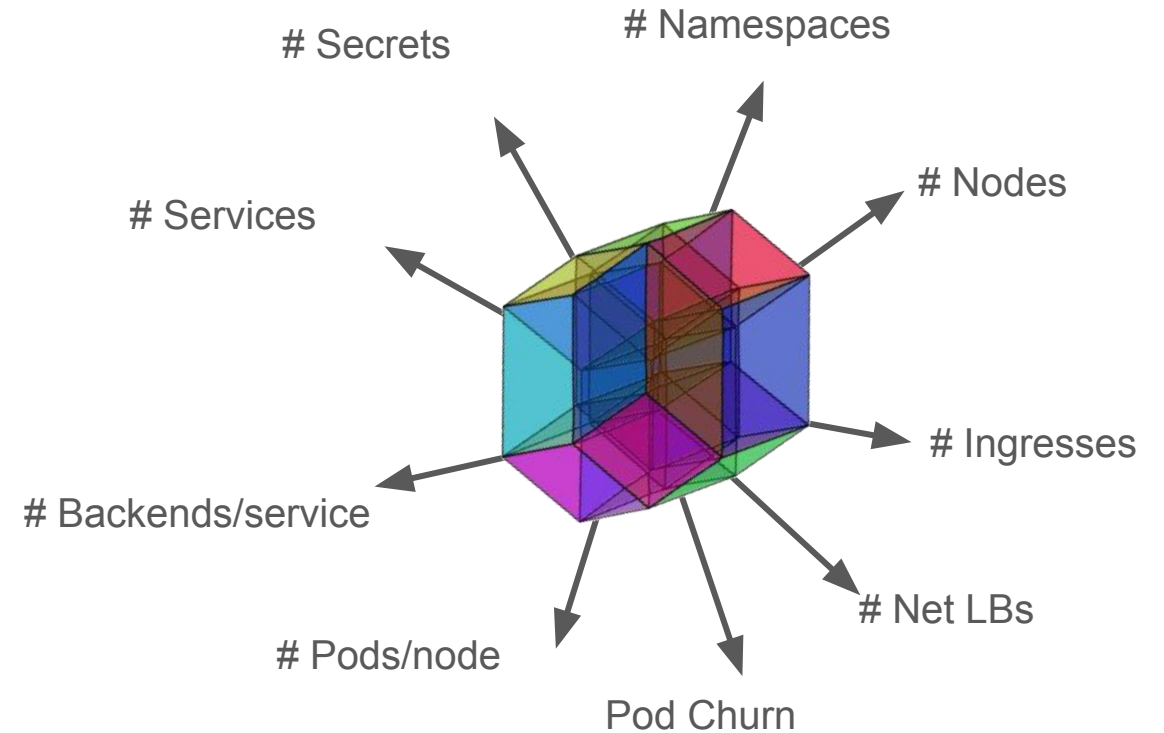
KubeCon



CloudNativeCon

North America 2023

~~Scalability = # nodes~~



Scalability Envelope

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

What is Kubernetes Scalability?



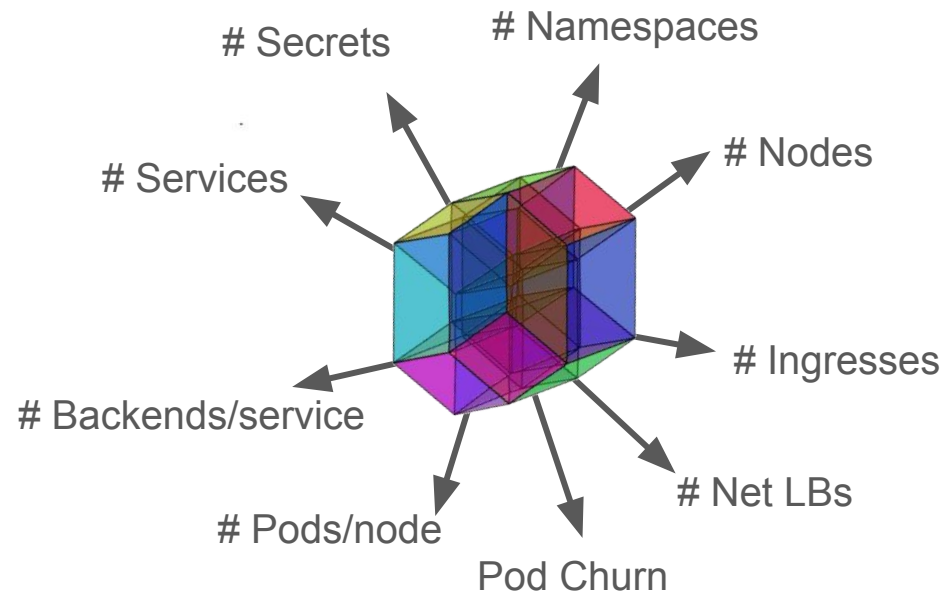
KubeCon



CloudNativeCon

North America 2023

Scalability Envelope - a *safe zone*, within which your cluster is *happy*.



**ALL (scalability)
SLOs are satisfied**

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

Kubernetes Scalability SLIs/SLOs



KubeCon



CloudNativeCon

North America 2023

SLI - Service Level Indicator

SLO - Service Level Objective

1. API Call Latency
2. Pod Startup Latency
3. In-Cluster Network Programming Latency
4. DNS Programming Latency
5. In-Cluster Network Latency
6. DNS Latency

More at github.com/kubernetes/community/tree/master/sig-scalability/slos

SLIs/SLOs - case study



KubeCon



CloudNativeCon

North America 2023

2015: ([blog post](#))

SLO: “99% of all our API calls return in less than 1 second”

2023: ([definition](#))

SLI: Latency of processing mutating API calls for single objects for every (resource, verb) pair, measured as 99th percentile over last 5 minutes

SLO: In default Kubernetes installation, for every (resource, verb) pair, excluding virtual and aggregated resources, 99th percentile per cluster-day ≤ 1 s

Scalability limits



KubeCon



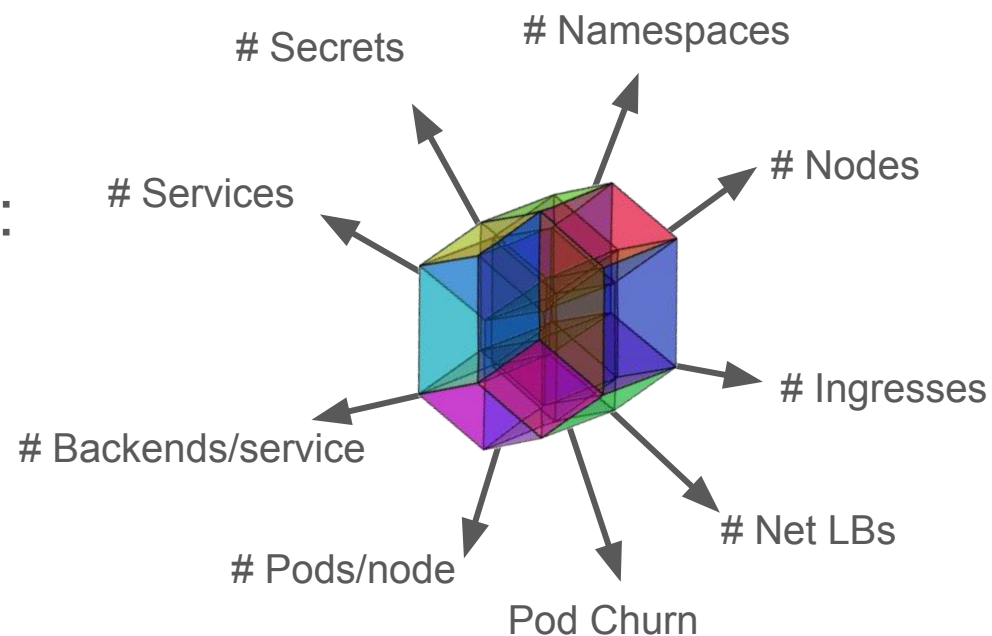
CloudNativeCon

North America 2023

Precise definition of **scalability envelope** is ~impossible.

Fortunately, we have reasonable approximation:

- #Nodes $\leq 5,000$
- #Services $\leq 10,000$
- ...



More at sig-scalability/configs-and-limits/thresholds.md

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



KubeCon



CloudNativeCon

North America 2023

Scalability Testing Infrastructure

Test framework - ClusterLoader2



KubeCon



CloudNativeCon

North America 2023

- **A “bring your own yaml” test framework**
 - User (semi) declaratively describes desired state of the cluster
 - ClusterLoader brings the cluster to state
 - At the same time verifying Scalability SLOs
- **Designed for easy extensibility**
 - Provides extra observability
 - And a bunch of extra features
 - See kubernetes/perf-tests/clusterloader2 for more details

Cluster simulation - Kubemark

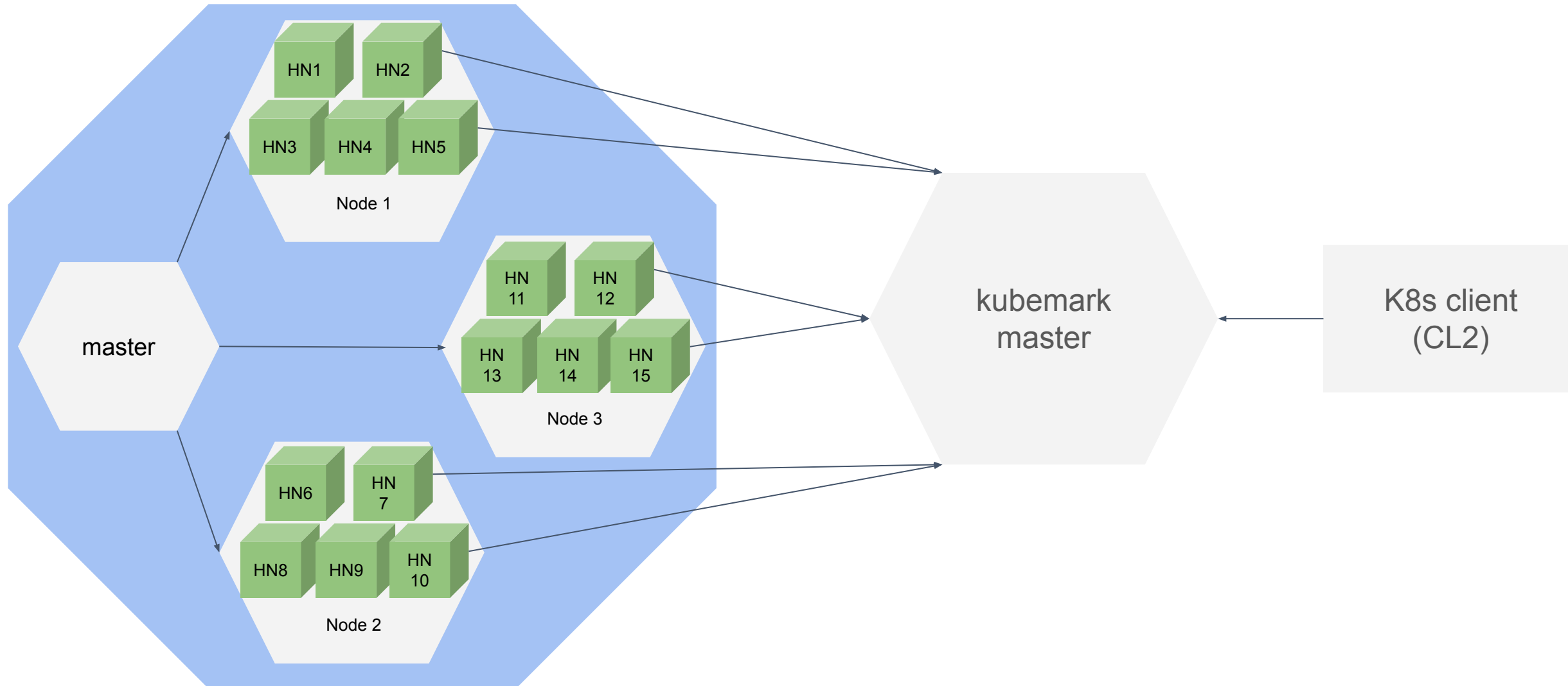


KubeCon



CloudNativeCon

North America 2023



Observability & debuggability - Perfdash



KubeCon



CloudNativeCon

North America 2023

Performance Dashboard

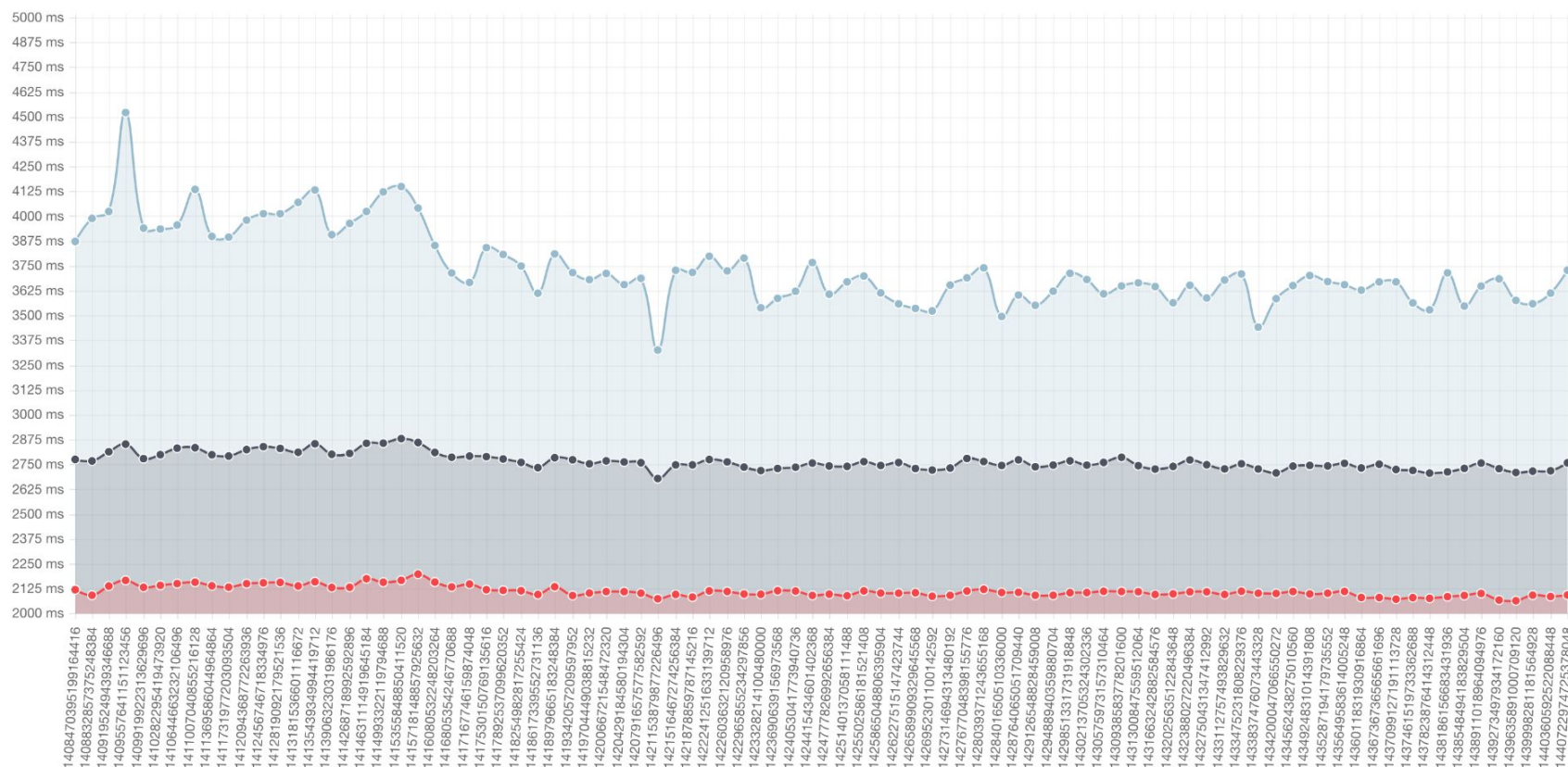


gce-5000Nodes

▼ E2E

▼ LoadPodStartup

▼ pod_startup



Runs over time

Observability & debuggability - Grafana

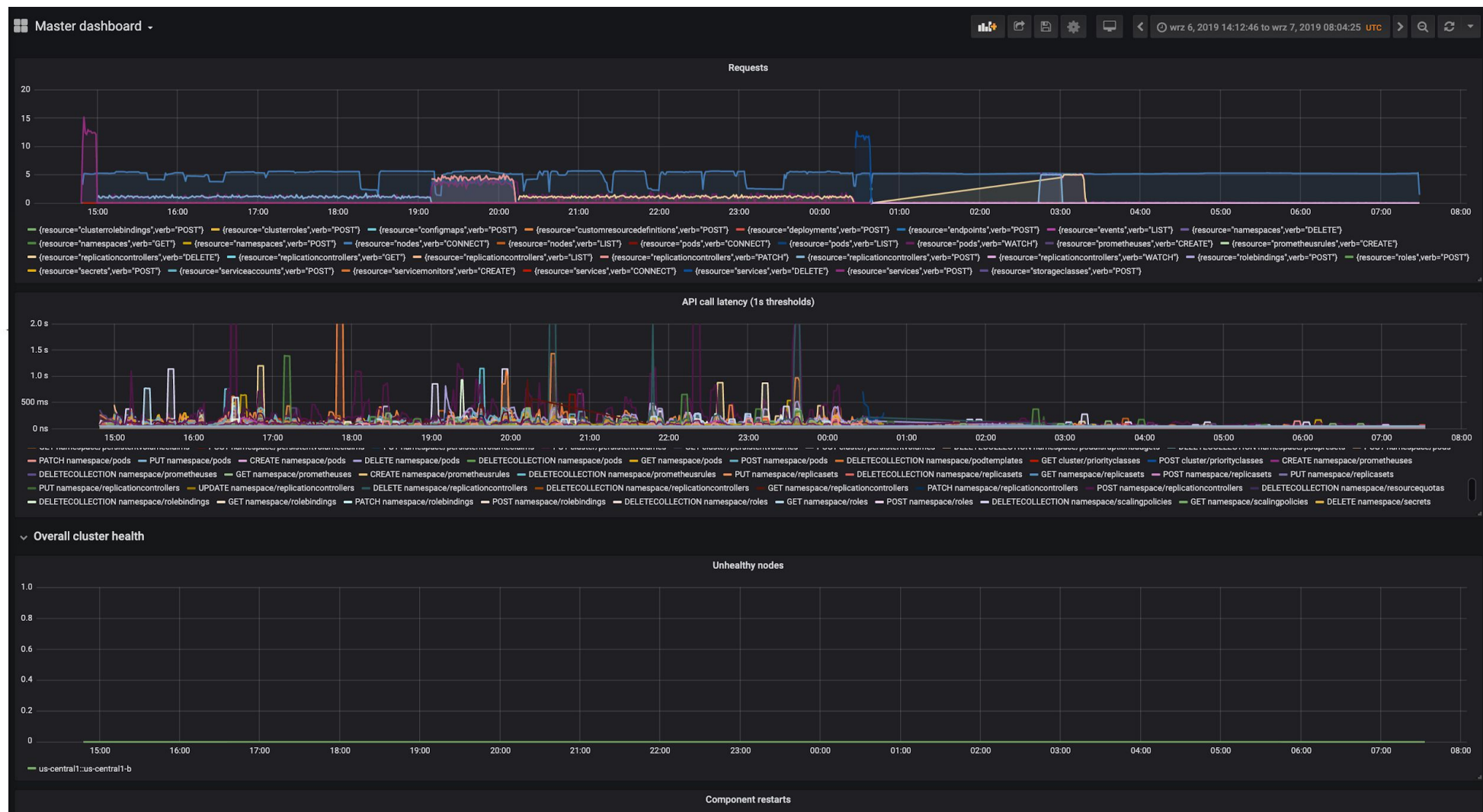


KubeCon



CloudNativeCon

North America 2023



Observability & debuggability - Profiling



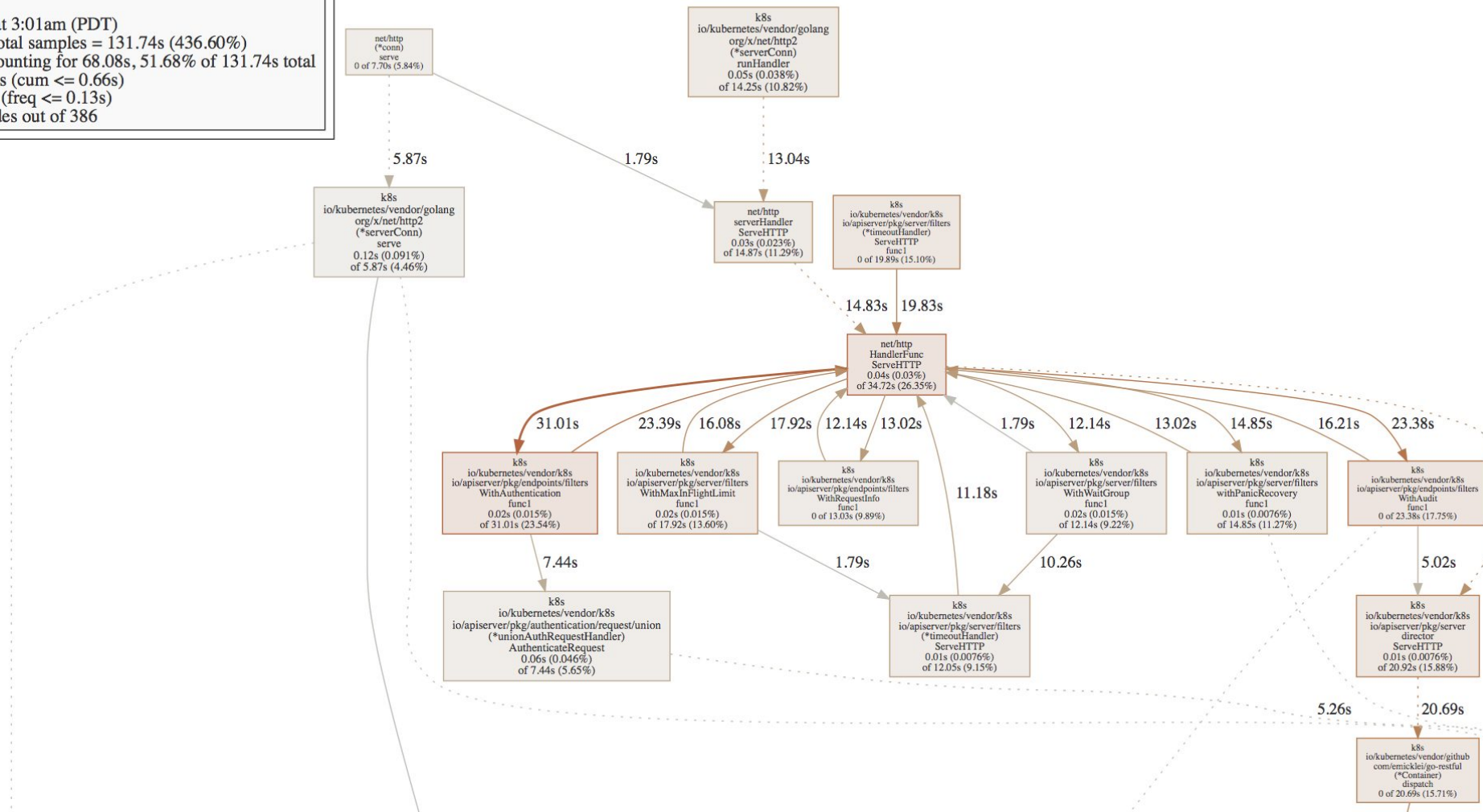
KubeCon



CloudNativeCon

North America 2023

File: kube-apiserver
Type: cpu
Time: Apr 8, 2019 at 3:01am (PDT)
Duration: 30.17s, Total samples = 131.74s (436.60%)
Showing nodes accounting for 68.08s, 51.68% of 131.74s total
Dropped 1663 nodes (cum <= 0.66s)
Dropped 217 edges (freq <= 0.13s)
Showing top 80 nodes out of 386





KubeCon



CloudNativeCon

North America 2023

Protecting Kubernetes Scalability

Scalability tests



KubeCon



CloudNativeCon

North America 2023

Periodic tests:

- Release blocking tests:
 - Performance 100 nodes
 - Performance 5000 nodes
 - Correctness 5000 nodes
- Non-release blocking
 - Kubemark
 - Storage
 - Benchmarks
 - ...

Optional Presubmit tests:

- Performance 100 nodes
 - kubernetes
 - Perf-tests
- Performance 5000 nodes

Scalability Tests Infrastructure



KubeCon



CloudNativeCon

North America 2023

	sig-scalability	sig-scalability-gce	sig-scalability-node	sig-scalability-kubemark	sig-scalability-perf-tests	sig-scalability-benchmarks	sig-scalability-experiments	sig-scalability-golang	sig-scalability-network	
Summary	gce-master-scale-correctness	gce-master-scale-performance	gce-cos-master-scalability-100	gce-cos-1.20-scalability-100	gce-cos-1.21-scalability-100	gce-cos-1.22-scalability-100	gce-cos-1.19-scalability-100			
										<div>Show All Alerts</div> <div>Hide All Alerts</div> <div>Sort by Status</div>
	gce-master-scale-correctness: FLAKY 8 of 10 (80.0%) recent columns passed (38319 of 38326 or 100.0% cells)									Last update: 09-23 21:45 CEST Tests last ran: 09-23 14:01 CEST Last green run: 9462ca231
	gce-master-scale-performance: PASSING 9 of 9 (100.0%) recent columns passed (504 of 504 or 100.0% cells)									Last update: 09-23 22:20 CEST Tests last ran: 09-23 19:03 CEST Last green run: 5b489e284
	gce-cos-master-scalability-100: FLAKY 8 of 9 (88.9%) recent columns passed (545 of 549 or 99.3% cells)									Last update: 09-23 22:40 CEST Tests last ran: 09-23 22:35 CEST Last green run: 6c2f64448
	gce-cos-1.20-scalability-100: PASSING 10 of 10 (100.0%) recent columns passed (580 of 580 or 100.0% cells)									Last update: 09-23 22:18 CEST Tests last ran: 09-23 18:02 CEST Last green run: 2624cc613
	gce-cos-1.21-scalability-100: PASSING 10 of 10 (100.0%) recent columns passed (580 of 580 or 100.0% cells)									Last update: 09-23 22:18 CEST Tests last ran: 09-23 14:02 CEST Last green run: 401153d9d
	gce-cos-1.22-scalability-100: PASSING 10 of 10 (100.0%) recent columns passed (620 of 620 or 100.0% cells)									Last update: 09-23 21:59 CEST Tests last ran: 09-23 20:02 CEST Last green run: 2c0e4a232
	gce-cos-1.19-scalability-100: PASSING 9 of 9 (100.0%) recent columns passed (522 of 522 or 100.0% cells)									Last update: 09-23 22:19 CEST Tests last ran: 09-23 22:02 CEST Last green run: 7b343ec8f

Scalability regressions



KubeCon



CloudNativeCon

North America 2023

- Scalability is *sensitive*
- We've seen regressions come from pretty much everywhere:
 - Golang
 - Operating System
 - Controllers
 - API machinery
 - Scheduler
 - Etcd
 - Kubelet
 - ...
- We often debug/fix them ourselves, or triage to relevant SIGs



KubeCon



CloudNativeCon

North America 2023

Driving Scalability Improvements

Scalability improvements



KubeCon



CloudNativeCon

North America 2023

Two categories of work:

- Improving reliability at scale
 - Scalability can be thought of as reliability at scale
- Pushing the limits

[Most joined with SIG API-machinery]

Scalability improvements



KubeCon



CloudNativeCon

North America 2023

Improving reliability [at scale]:

- API Priority & Fairness - [#1040](#)
 - GA-ing in 1.29
- Improved upgrades experience
 - Graceful shutdown - e.g. [#114925](#)
- API streaming lists - [#3157](#)
 - Second Alpha in 1.29

Scalability improvements



KubeCon



CloudNativeCon

North America 2023

Justification is critical for complexity vs ROI tradeoff

Pushing the limits

- Improved CRD scalability:
 - Binary encoding for CRDs - [#4222](#)
 - Pre-alpha in 1.29
 - Efficient event serialization - [#120300](#)
- Localized kube-apiserver improvements
 - e.g. faster compression - [#112296](#)
 - e.g. page size progressing - [#108569](#)
- Improvements towards higher throughput

How to get involved?



KubeCon



CloudNativeCon

North America 2023

Where to find us?

- Home page: [README](#)
- Public Meetings: Thursdays 17.30 UTC (bi-weekly)
- Slack channel: [#sig-scalability](#)
- Mailing List: [#kubernetes-sig-scale](#)

How to get involved?

- [kubernetes/perf-tests help-wanted](#)
- [kubernetes/kubernetes help-wanted](#)



PromCon
North America 2021



**Please scan the QR Code above
to leave feedback on this session**