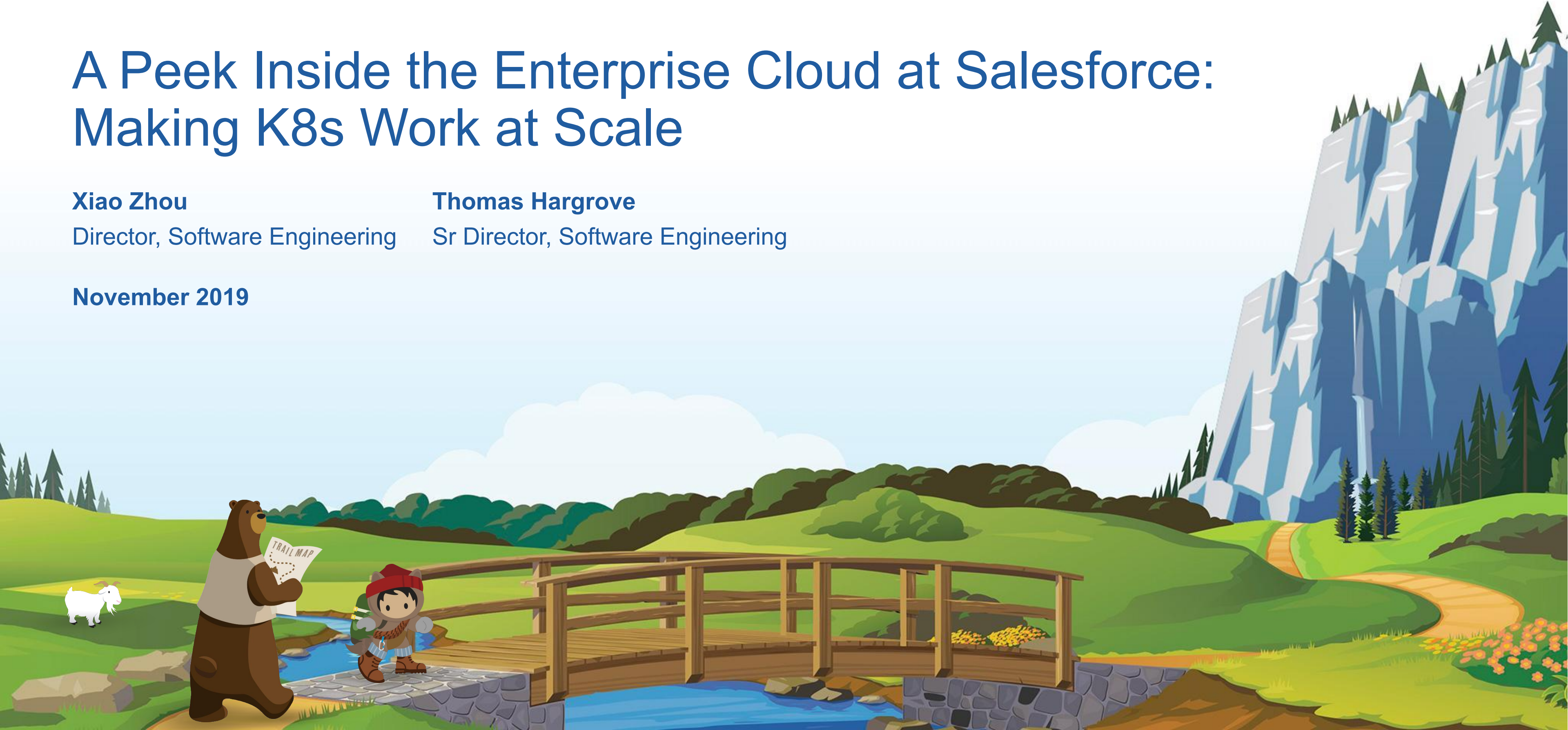
A Peek Inside the Enterprise Cloud at Salesforce: Making K8s Work at Scale

Xiao Zhou
Director, Software Engineering

Thomas Hargrove
Sr Director, Software Engineering

November 2019



Agenda

Unique Challenges for Enterprise Cloud at Salesforce

Solutions

Multitenancy & Security
Deployment Management
Orchestrated Production Visibility
Testing/Monitoring/Alerting

Future Projects

Unique Challenges for the Enterprise Cloud with K8s

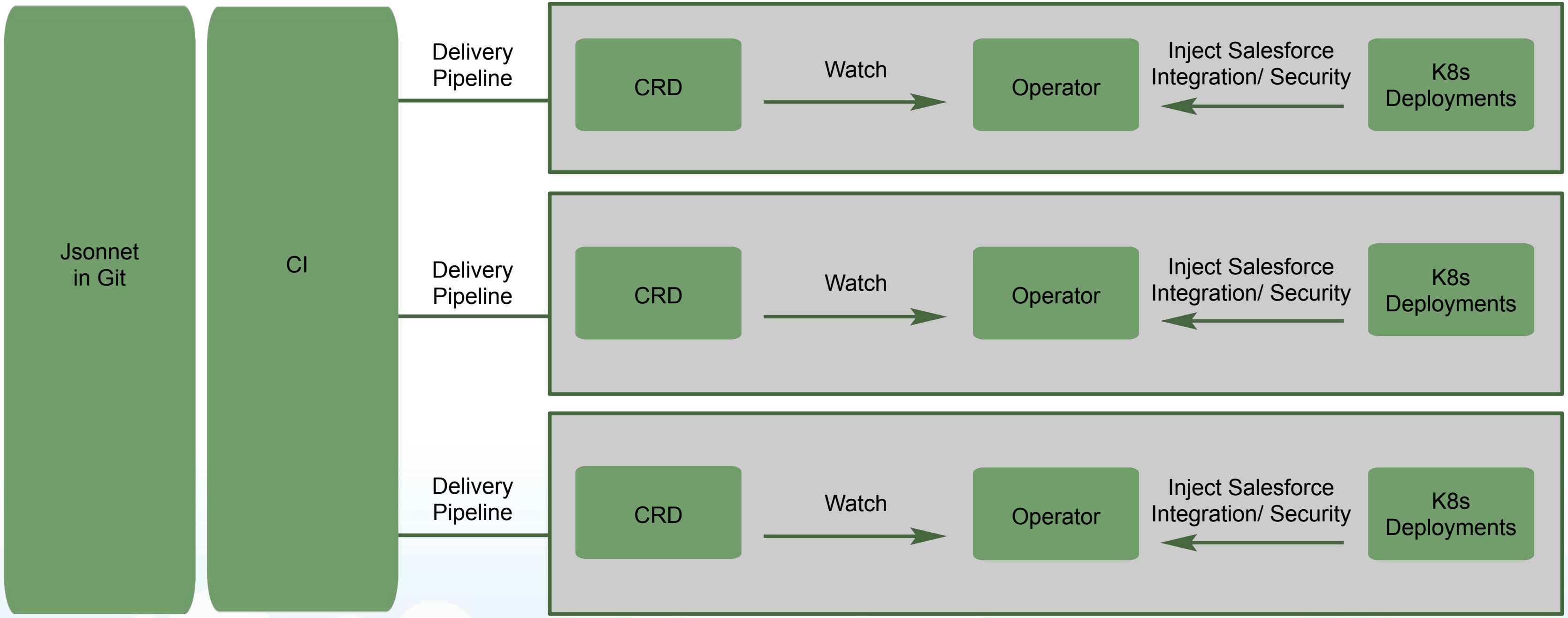
- 1st Party, many Prod data centers, thousands of hosts
- Bare metal
- Restrictive privileges
- Internal integration
- Internal customers: hundreds of namespaces and apps
- Requirement for advanced tools
- Network isolation
- Container security

Multitenancy & Security

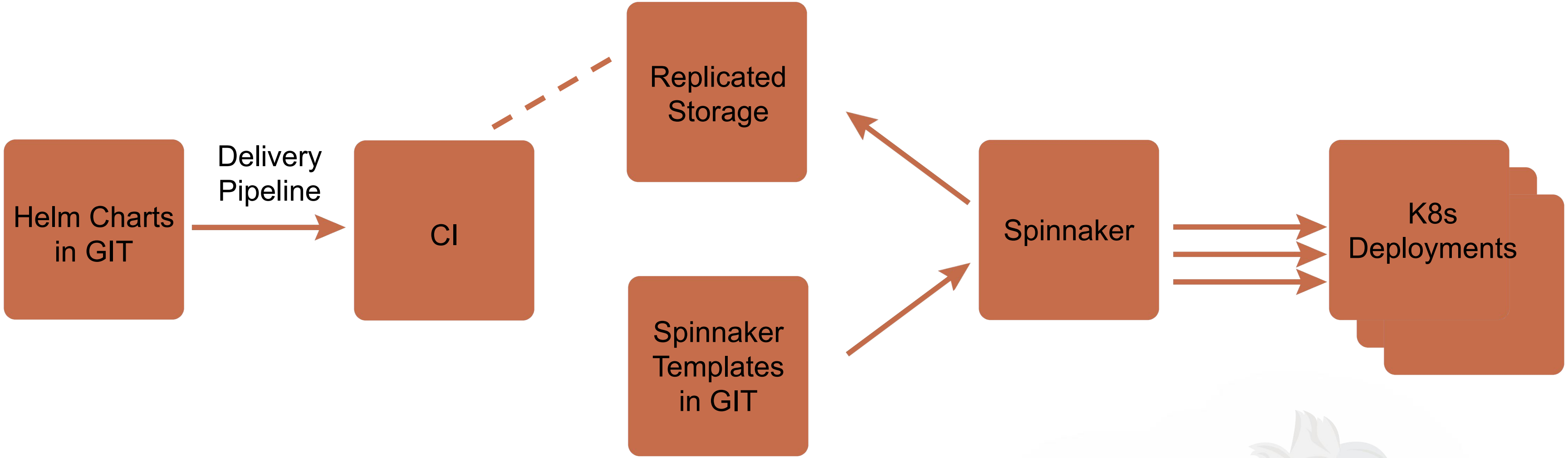
- Shared clusters with tenant protections
    - mTLS communication with internal PKI service
    - OPA (Open Policy Agent)
    - RBAC (Role-Based Access Control) to limit access to namespaces
- Use internal secret management instead of k8s secrets
- Container scanning for forensic analysis and security concerns
- Code Signing
- Automated patching w/ health checks, change tracking and sequencing
- Service mesh
- Detections for malicious behavior

Internal Deployment Management - Version 1

Goal-seeking system using GIT for source of truth



Internal Deployment Management - Version 2



Helm Charts in GIT

Delivery Pipeline

CI

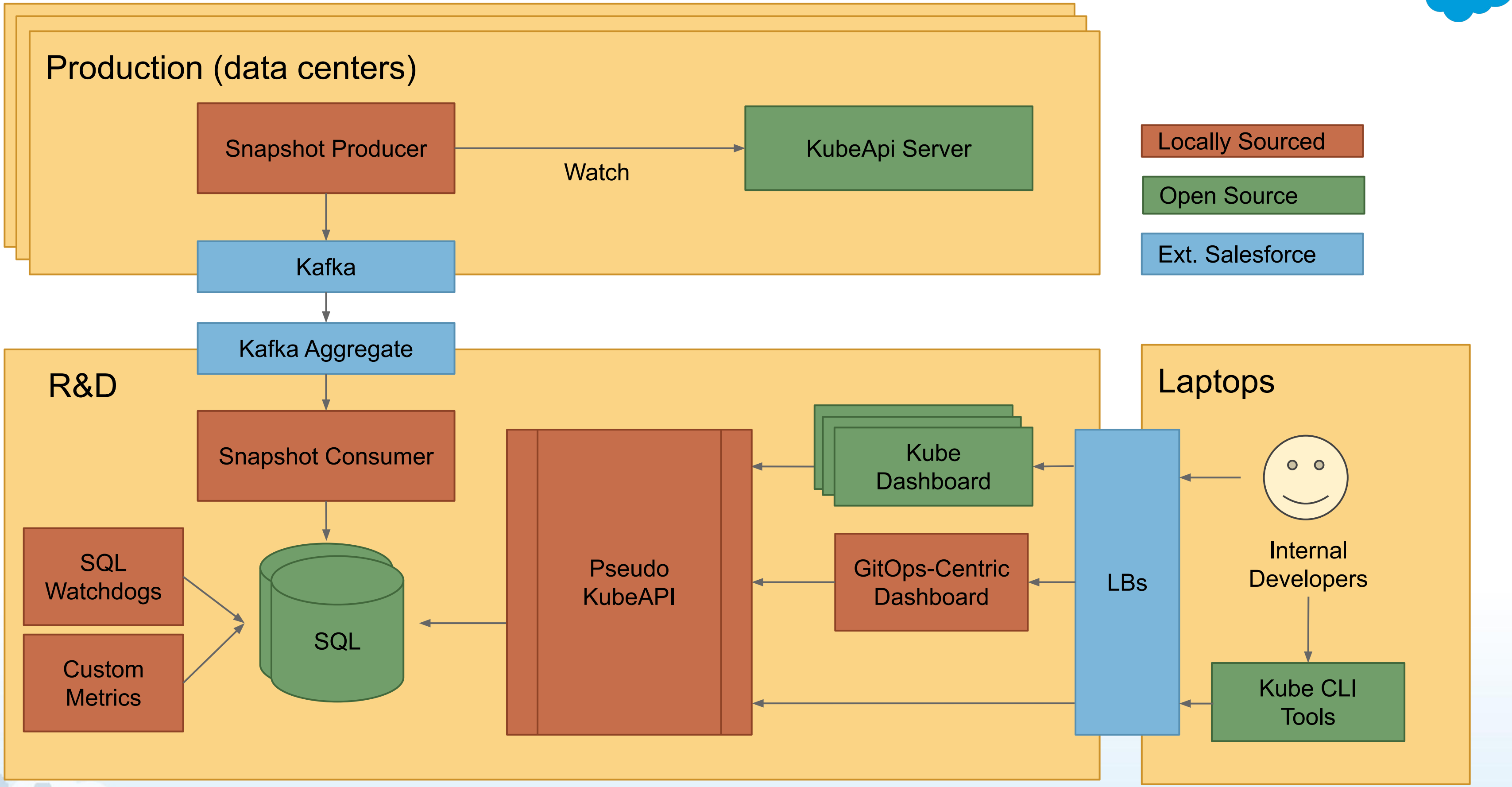
Replicated Storage

Spinnaker Templates in GIT

Spinnaker

K8s Deployments

Orchestrated Production Visibility



Example SQL Query - Pods with failing init containers

```
select
  cluster,
  namespace,
  name as podName,
  Payload->>'$.spec.nodeName' as nodeName,
  Payload->>'$.status.phase' as phase,
  Payload->>'$.status.initContainerStatuses[*].restartCount' as initContainerRestartCount,
  Payload->>'$.status.initContainerStatuses[*].state.*.message' as initContainerMessage,
  Payload->>'$.status.message' as message
from k8s_resource
where kind = 'Pod' and
  Payload->>'$.status.phase' = 'Pending' and
  Payload->>'$.status.initContainerStatuses[*].state.*.message' is not null
```

Example SQL Query - Pods with failing init containers

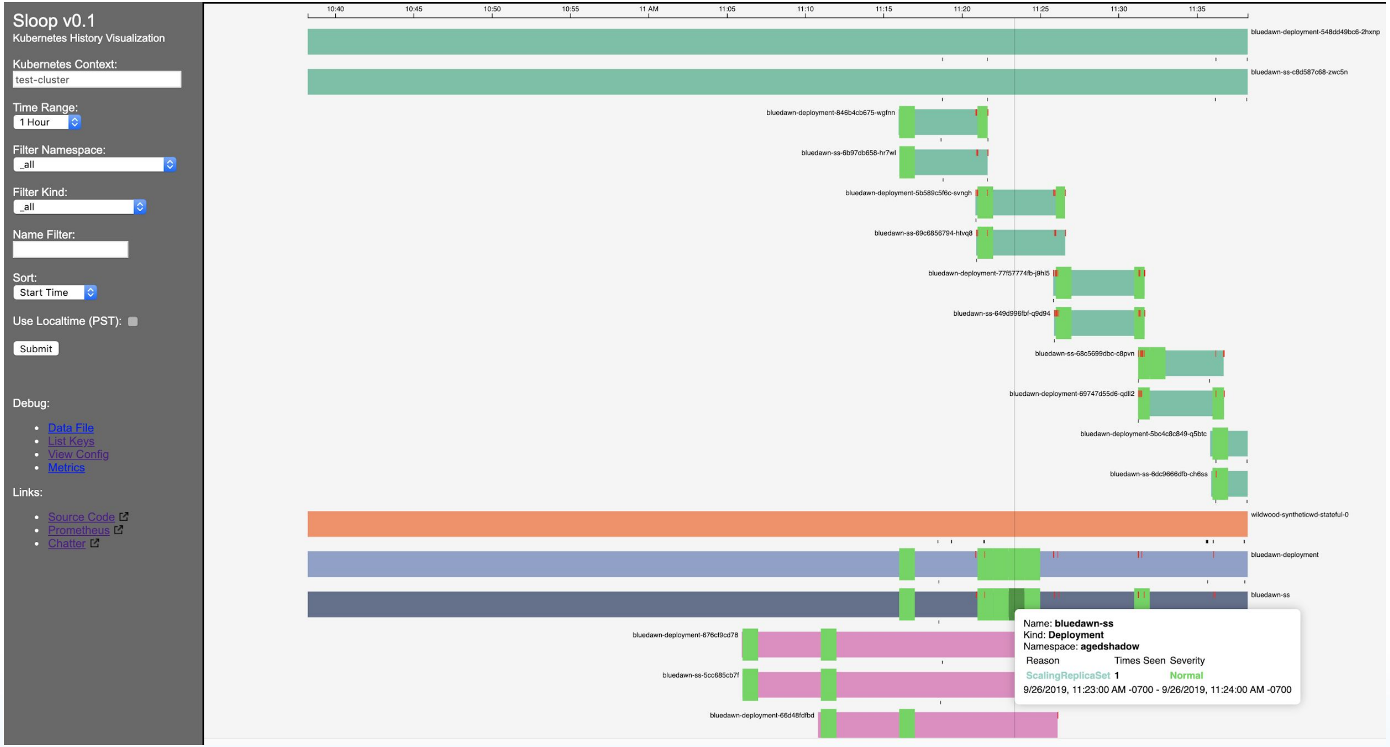
```
select
  cluster,
  namespace,
  name as podName,
  Payload->>'$.spec.nodeName' as nodeName,
  Payload->>'$.status.phase' as phase,
  Payload->>'$.status.initContainerStatuses[*].restartCount' as initContainerRestartCount,
  Payload->>'$.status.initContainerStatuses[*].state.*.message' as initContainerMessage,
  Payload->>'$.status.message' as message
from k8s_resource
where kind = 'Pod' and
  Payload->>'$.status.phase' = 'Pending' and
  Payload->>'$.status.initContainerStatuses[*].state.*.message' is not null
```

cluster	namespace	podName	nodeName	phase	initContainerRestartCount	initContainerMessage
cluster07	dense-stardust	potassium-7578655876-78jnd	cluster07-node014	Pending	[1652, 0, 0]	["Back-off 5m0s restarting failed container=cert-init-container..."]
cluster07	dense-stardust	potassium-7578655876-sx5qm	cluster07-node014	Pending	[9346, 0, 0]	["Back-off 5m0s restarting failed container=cert-init-container..."]
cluster21	quick-sun	halibut-67b8955978-wx9qs	cluster21-node004	Pending	[705, 0]	["Back-off 5m0s restarting failed container=cert-init-container..."]
cluster21	quick-sun	halibut-99c7f886b-p9rp7	cluster21-node010	Pending	[861, 0]	["Back-off 5m0s restarting failed container=cert-init-container..."]

Testing/Monitoring/Alerting

- Big investment in unit tests and watch dogs
- Watchdog detect issues and page on call
- Phase releases for all changes
  - Phase 1 - Test beds
  - Phase 2 - RnD
  - Phase 3 - Prod canary
  - Phase 4 - Rest of production
- Disabled by default feature flags
- Repair automation based on watchdog alerting
- Strive for consistency

Sloop - Kubernetes History Visualization



Sloop v0.1
Kubernetes History Visualization

Kubernetes Context: test-cluster

Time Range: 1 Hour

Filter Namespace: _all

Filter Kind: _all

Name Filter:

Sort: Start Time

Use Localtime (PST): □

Submit

Debug:

- Data File
- List Keys
- View Config
- Metrics

Links:

- Source Code
- Prometheus
- Chatter

Name: **bluedawn-ss**
Kind: **Deployment**
Namespace: **agedshadow**

Reason	Times Seen	Severity
ScalingReplicaSet	**1**	**Normal**

9/26/2019, 11:23:00 AM -0700 - 9/26/2019, 11:24:00 AM -0700

10:40
10:45
10:50
10:55
11 AM
11:05
11:10
11:15
11:20
11:25
11:30
11:35
bluedawn-deployment-548dd49bc6-2hxnp
bluedawn-ss-c8d587c68-zwc5n
bluedawn-deployment-846b4cb675-wgfnn
bluedawn-ss-6b97db658-hr7wl
bluedawn-deployment-5b589c5f6c-svngh
bluedawn-ss-69c6856794-htvq8
bluedawn-deployment-77f57774fb-j9hl5
bluedawn-ss-649d996fbf-q9d94
bluedawn-ss-68c5699dbc-c8pvn
bluedawn-deployment-69747d55d6-qdll2
bluedawn-deployment-5bc4c8c849-q5btc
bluedawn-ss-6dc9666dfb-ch6ss
wildwood-syntheticwd-stateful-0
bluedawn-deployment
bluedawn-ss
bluedawn-deployment-676cf9cd78
bluedawn-ss-5cc685cb7f
bluedawn-deployment-66d48fdfbd

https://github.com/salesforce/sloop

Demos: Wednesday 1-3pm, Thursday 3-5pm as the Salesforce booth

Future Projects

- Robust mutating webhook validation and rollout sequencing
- Cost-to-serve reporting and alerting
- One-step onboarding: repo, build, publish, Helm chart, Spinnaker pipeline
- Better Helm validations at PR time
- Open source version of our visibility pipeline

thank you

Q & A