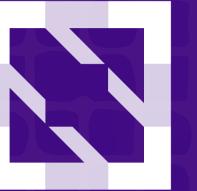




KubeCon



CloudNativeCon



China 2019



1-5-10: How to Fast Recover Container Failure at Large Scale

Huan Xiong



KubeCon



CloudNativeCon



OPEN SOURCE SUMMIT

China 2019



Agenda



KubeCon



CloudNativeCon



OPEN SOURCE SUMMIT

China 2019

1

Definition of 1-5-10

2

An incident replay

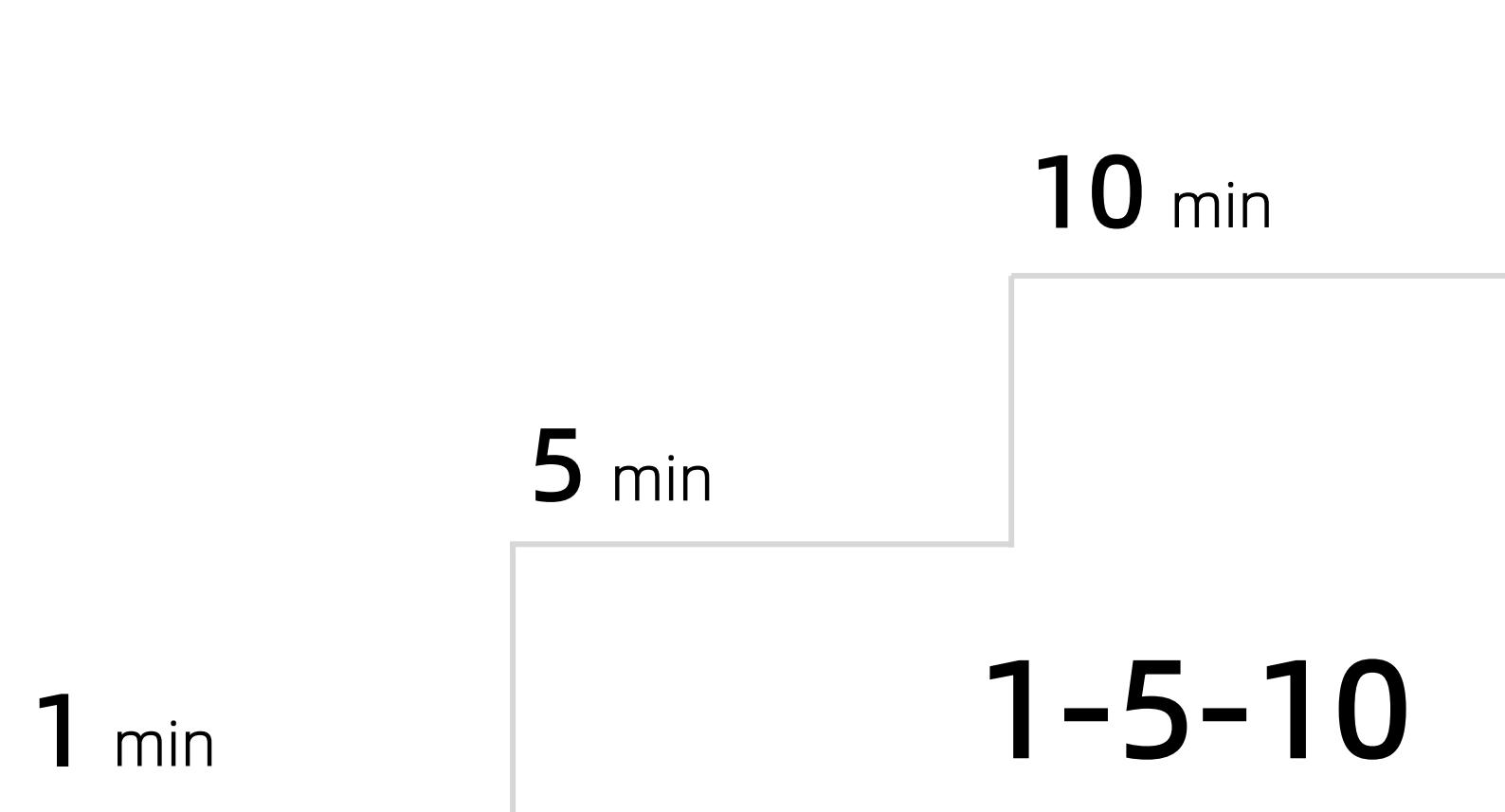
3

How to 1-5-10--offline

4

How to 1-5-10--online

Definition of 1-5-10





Agenda



KubeCon



CloudNativeCon



OPEN SOURCE SUMMIT

China 2019

①

Definition of 1-5-10

②

An incident replay

③

How to 1-5-10--offline

④

How to 1-5-10--online

Incident Timeline

- 15:17 Starts agent upgrade, 3 batchs, first batch x nodes.
- 16:30 Starts to upgrade second batch xx nodes.
- 16:55 The new feature of agent is activated.
- 17:30 Internal monitoring systems began generating alerts.
- 17:39 Starts to upgrade last batch xxx nodes.
- 17:47 xxxx pods are impacted by agent.
- 17:50 Stops upgrade job of agent.
- 18:15 Difficult to identify the root cause.
- 18:19 changes load balancing to increase traffic to other clusters.
- 18:28 Roll back agent.
- 19:03 Finish to roll back agent, the status was updated to green.

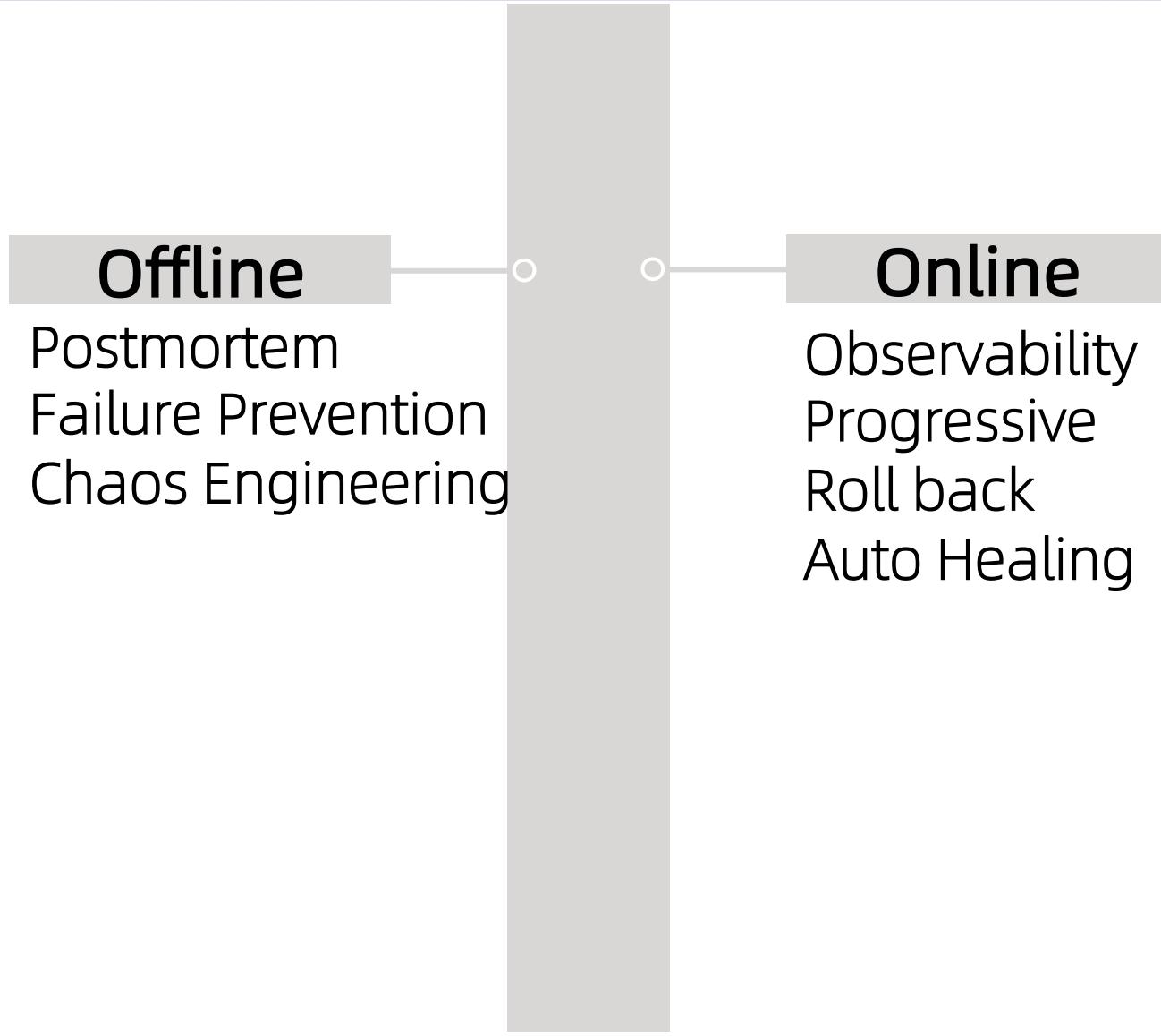
1-5-10?



What we learned

- 1 Not sensitive to potential failures; alerts received but not handled in time.
- 2 Lack of risk awareness; new features are not fully validated.
- 3 Progressive deployments are not perfect.
- 4 Lack of observability.
- 5 Lack of solutions to stopping loss in time.

How to improve





Agenda



KubeCon



CloudNativeCon



OPEN SOURCE SUMMIT

China 2019

①

Definition of 1-5-10

②

An incident replay

③

How to 1-5-10--offline

④

How to 1-5-10--online



Postmortem

阶段Phase	关键时间点Time	要点KeyPoint
发生Occur	故障注入的时间点 故障发生的时间点	故障的根因? 以后能不能避免同类问题?
发现Detect	故障发现的时间点	发现途径是什么? 自动还是人工? 是谁发现的? 业务方还是责任方? 能不能快速发现? 能不能提前发现?
定位Identify	故障相关人接手的时间点 定位出根因的时间点 找出止血方法的时间点	相关人是否及时接手? 定位根因的方法? 看监控, 查日志? 是否通过人工排查才找出止血方法?
恢复Recover	系统恢复的时间点 业务恢复的时间点	恢复的方式是什么? 回滚或是其他? 执行恢复操作花了多长时间? 未来能不能快速恢复?



Failure Prevention

Design > Develop > Test > 准入

Stability is vital in the process of design, development, validation and deployment.

Chaos Engineering

Prepare > Inject > Test > Analysis & Optimization

By introducing faults to the system while it's running and interrupting its normal workloads, we can discover some potential bugs and improve the stability and resiliency of our system.



Agenda

1

Definition of 1-5-10

2

An incident replay

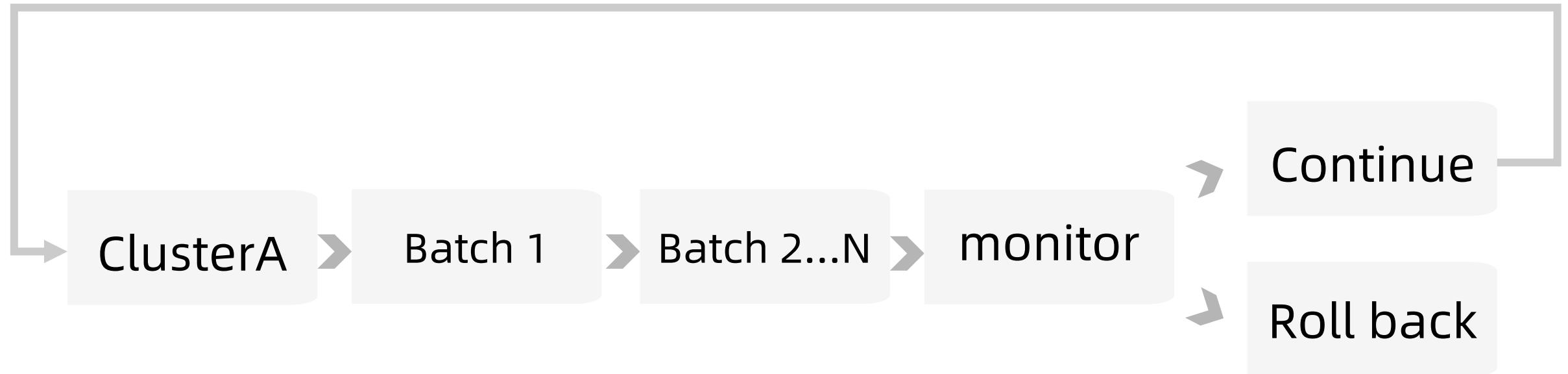
3

How to 1-5-10--offline

4

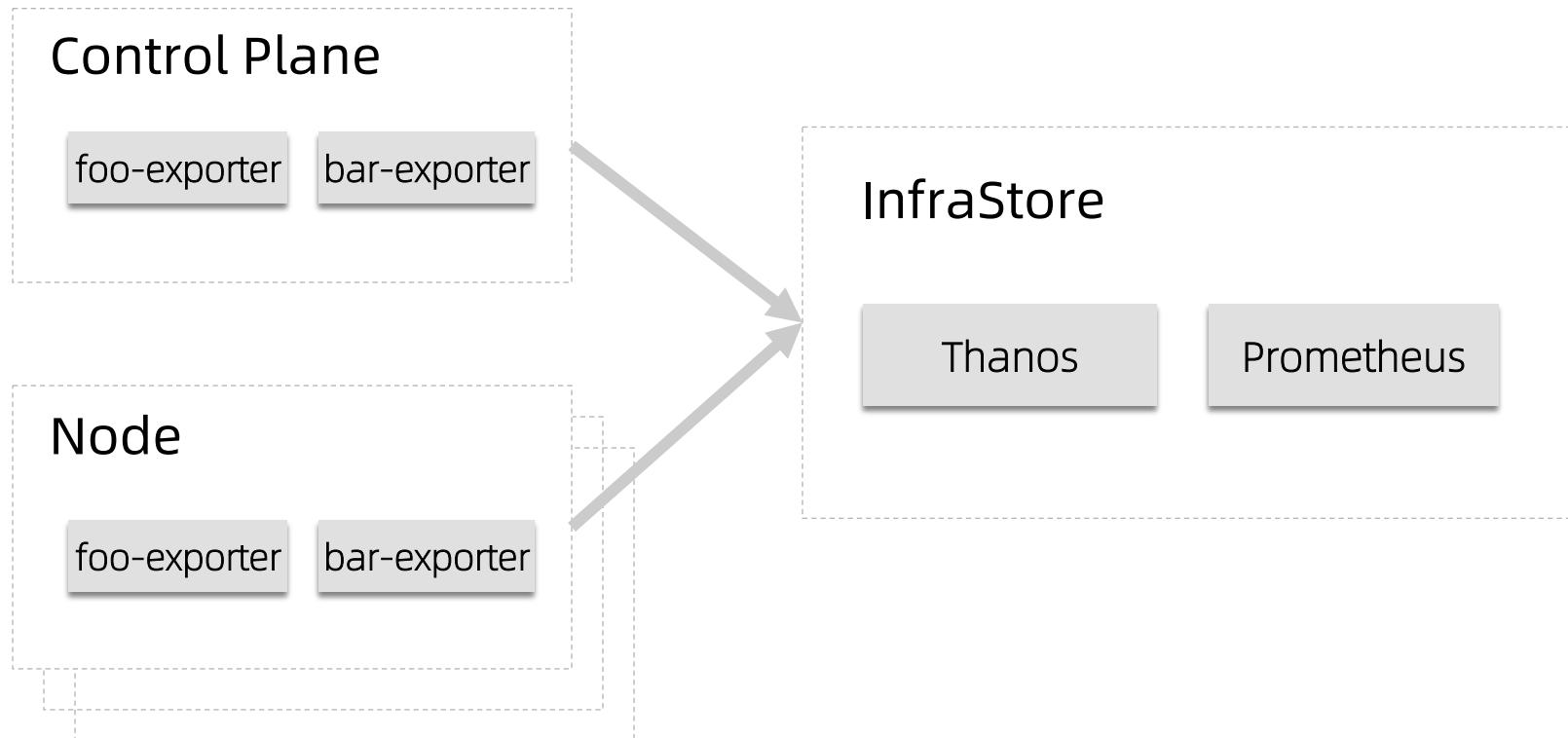
How to 1-5-10--online

Progressive Deployment



- ✗ Make sure the strategy of stopping loss is ready before enforcing any changes.

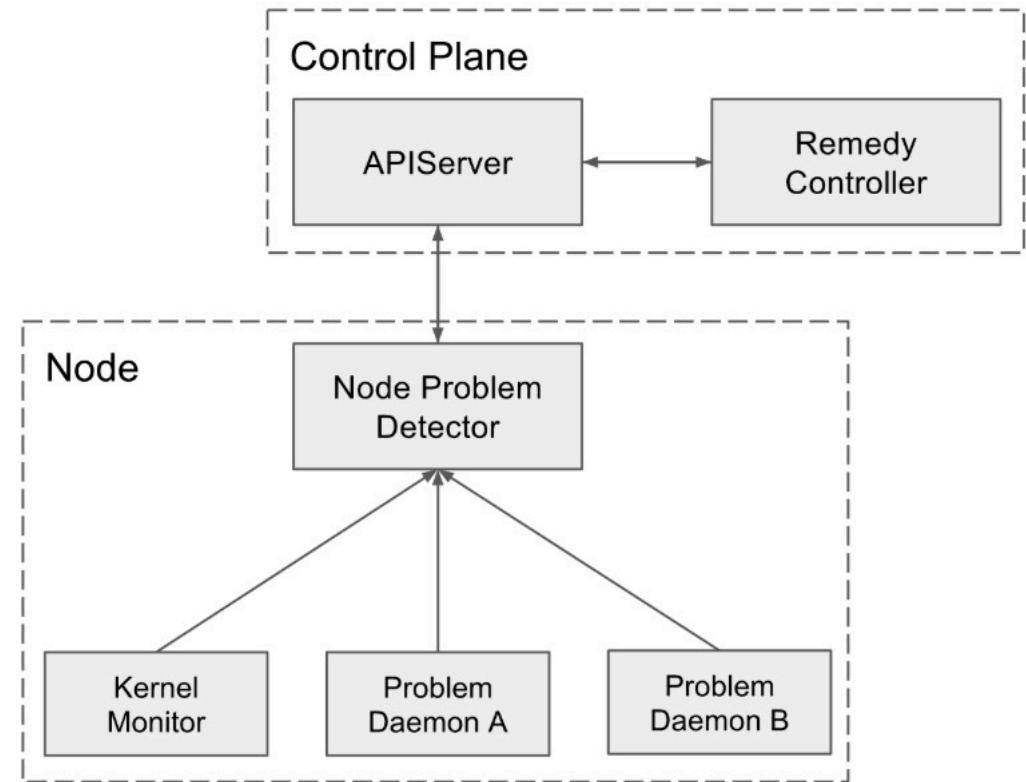
Observability

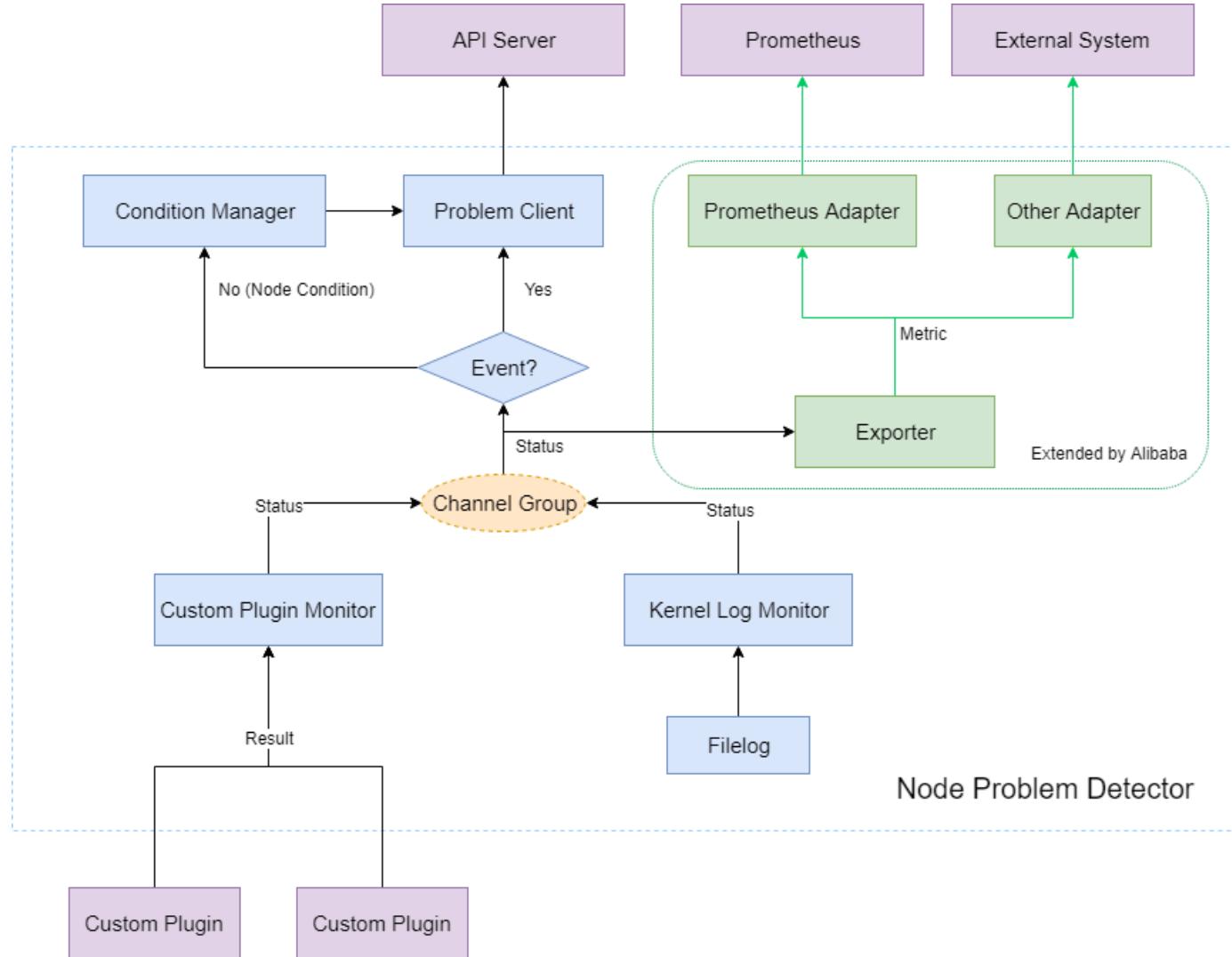


Node Problem Detector: A DaemonSet detects node problems and reports them to APIServer.

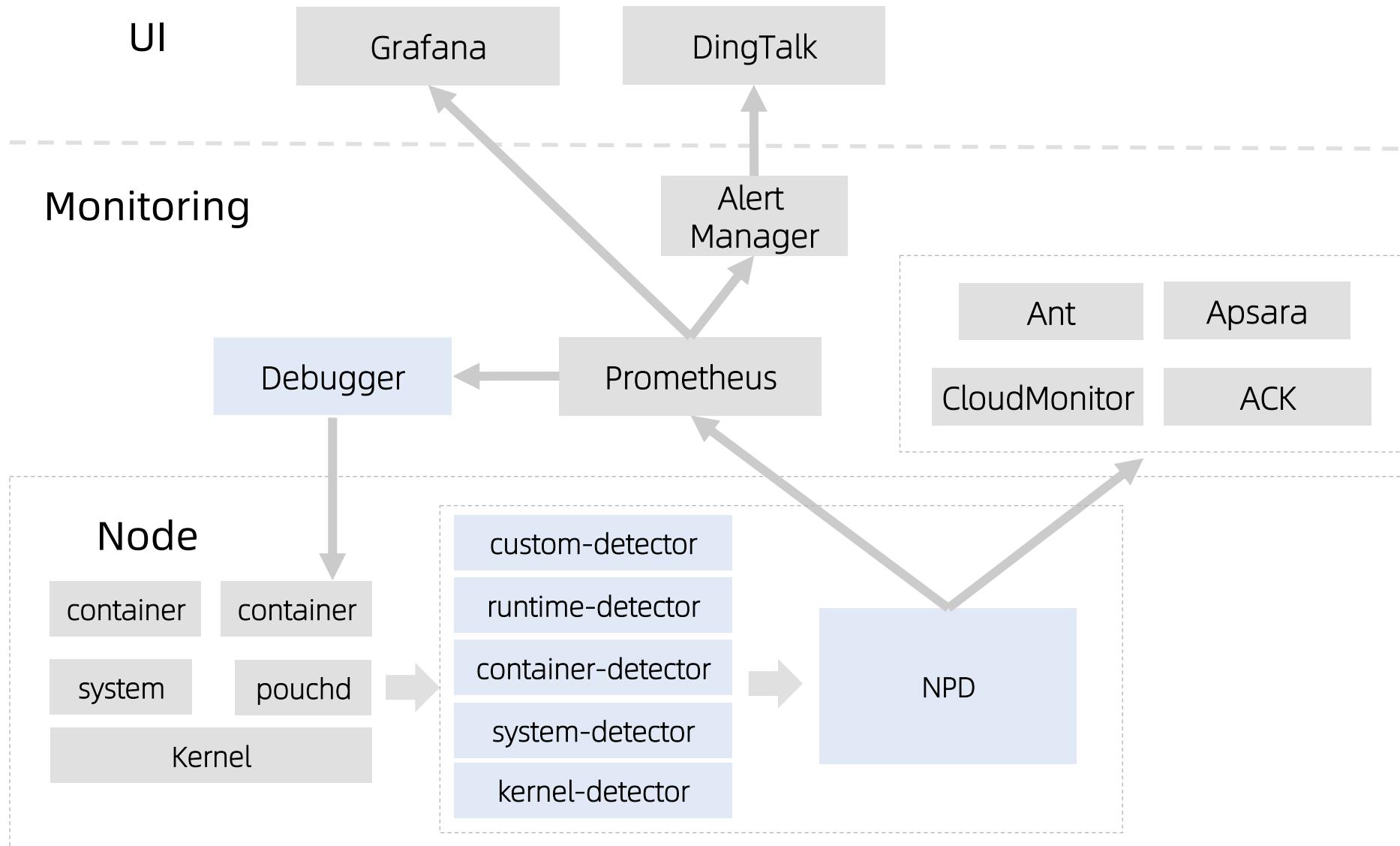
Enhancement of NPD

- Adapter of Prometheus
- External monitoring system

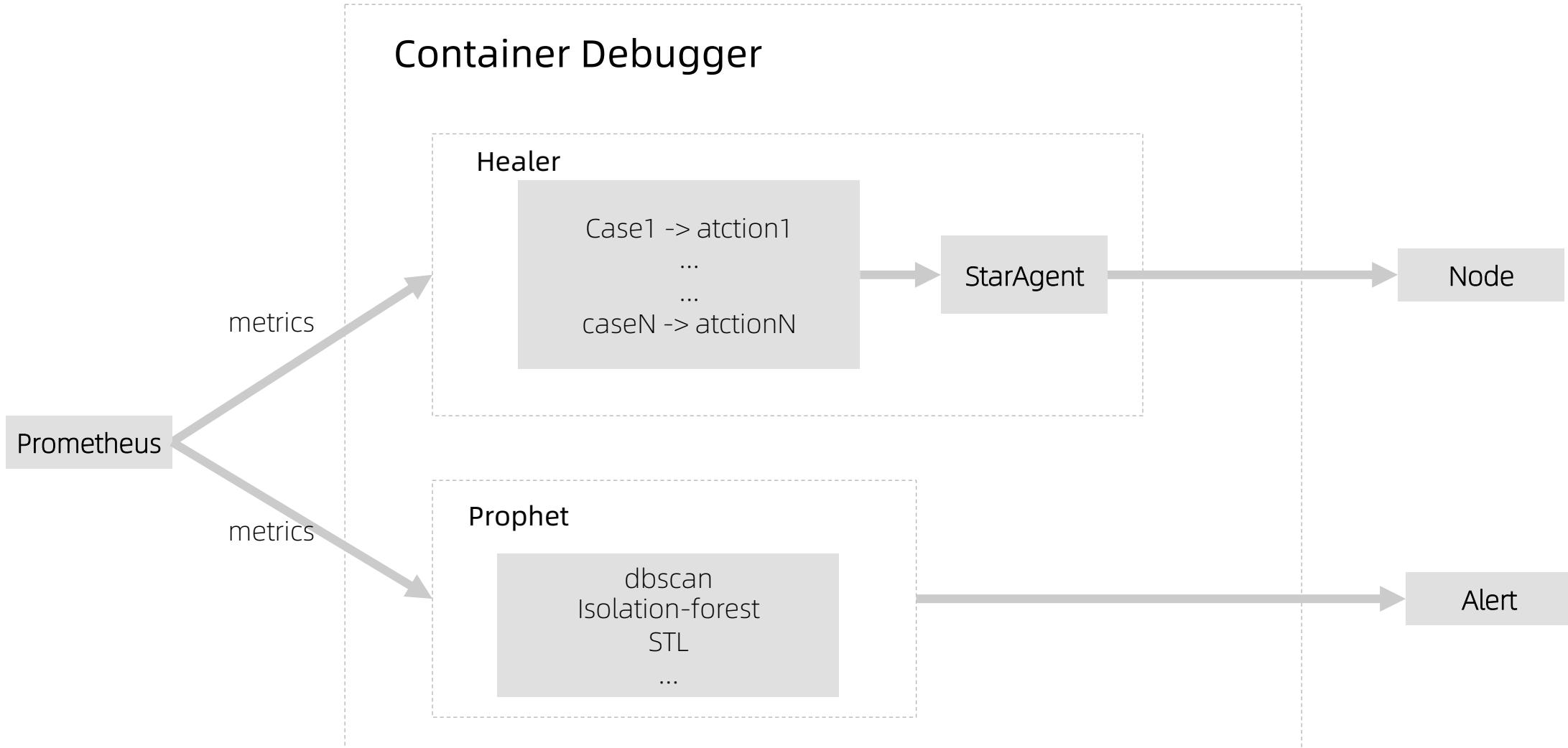




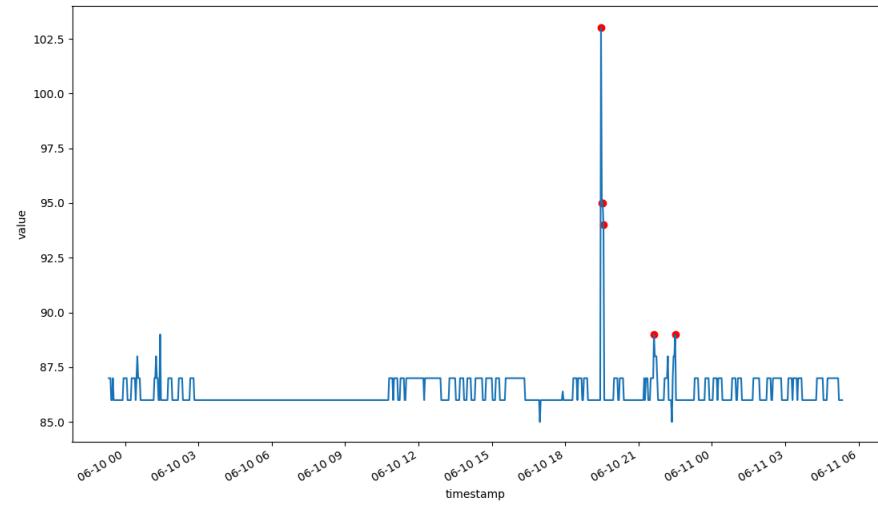
Auto Healing



Container Debugger



Prophet



KubeCon



CloudNativeCon



OPEN SOURCE SUMMIT

China 2019

Q & A



免登陆听课

动手实践课后自测

CKA课程内容同步

阿里云原生最佳实践



立即听课