从混部到 Serverless 化，腾讯自研业务云原生成本及稳定性优化实践

吕祥坤

腾讯云容器高级开发工程师

InfoQ

精彩继续！
更多一线大厂前沿技术案例

北京站

GMTC
全球大前端技术大会

时间：9月15-16日
地点：北京·国际会议中心

扫码查看大会详情>>

北京站

QCon
全球软件开发大会

时间：9月17-19日
地点：北京·富力万丽酒店

扫码查看大会详情>>

杭州站

ArchSummit
全球架构师峰会

时间：9月26-27日
地点：杭州·和达希尔顿逸林酒店

扫码查看大会详情>>

个人介绍

负责腾讯超千万核自研业务容器上云平台 TKEx 的研发设计工作，支持包括 QQ、腾讯会议、腾讯文档在内的海量自研业务实现云原生架构升级。在有状态服务应用改造、云原生应用发布与流量管理、大规模集群的运营效率、稳定性提升等方向有较多的经验与积累。

大纲

腾讯自研业务容器化上云历程及主要问题

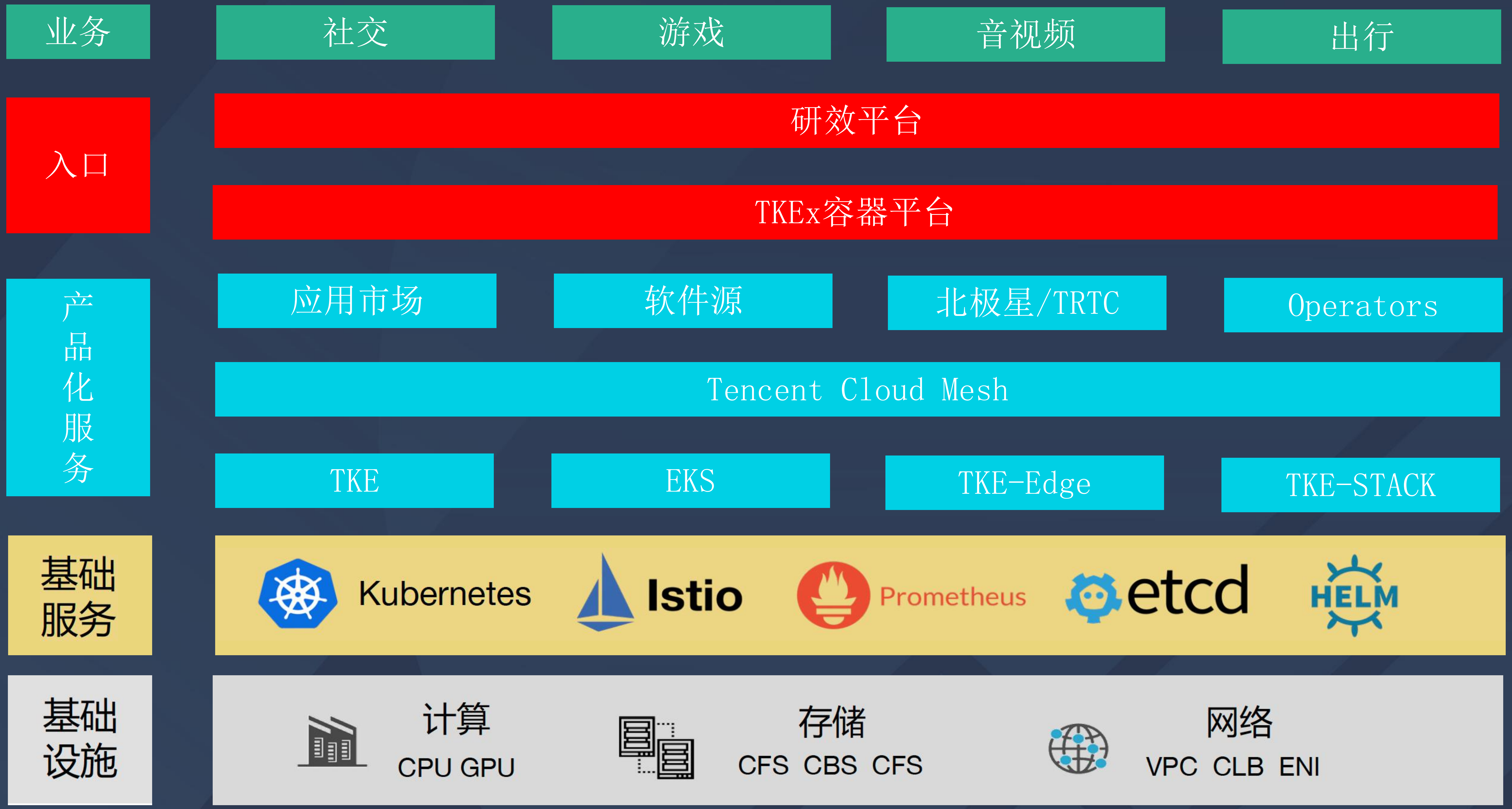
在线混部集群的资源利用率提升方案

拥抱腾讯云弹性容器服务 EKS 价值所在

存量 K8s 集群应用平滑迁移弹性容器服务 EKS 的落地实践

腾讯自研业务容器化上云历程及主要问题

自研业务容器化上云简略架构



业务：社交 | 游戏 | 音视频 | 出行

入口：研效平台 | TKEx容器平台

产品化服务：应用市场 | 软件源 | 北极星/TRTC | Operators
Tencent Cloud Mesh
TKE | EKS | TKE-Edge | TKE-STACK

基础服务：Kubernetes | Istio | Prometheus | etcd | HELM

基础设施：
计算 CPU GPU
存储 CFS CBS CFS
网络 VPC CLB ENI

独立集群架构

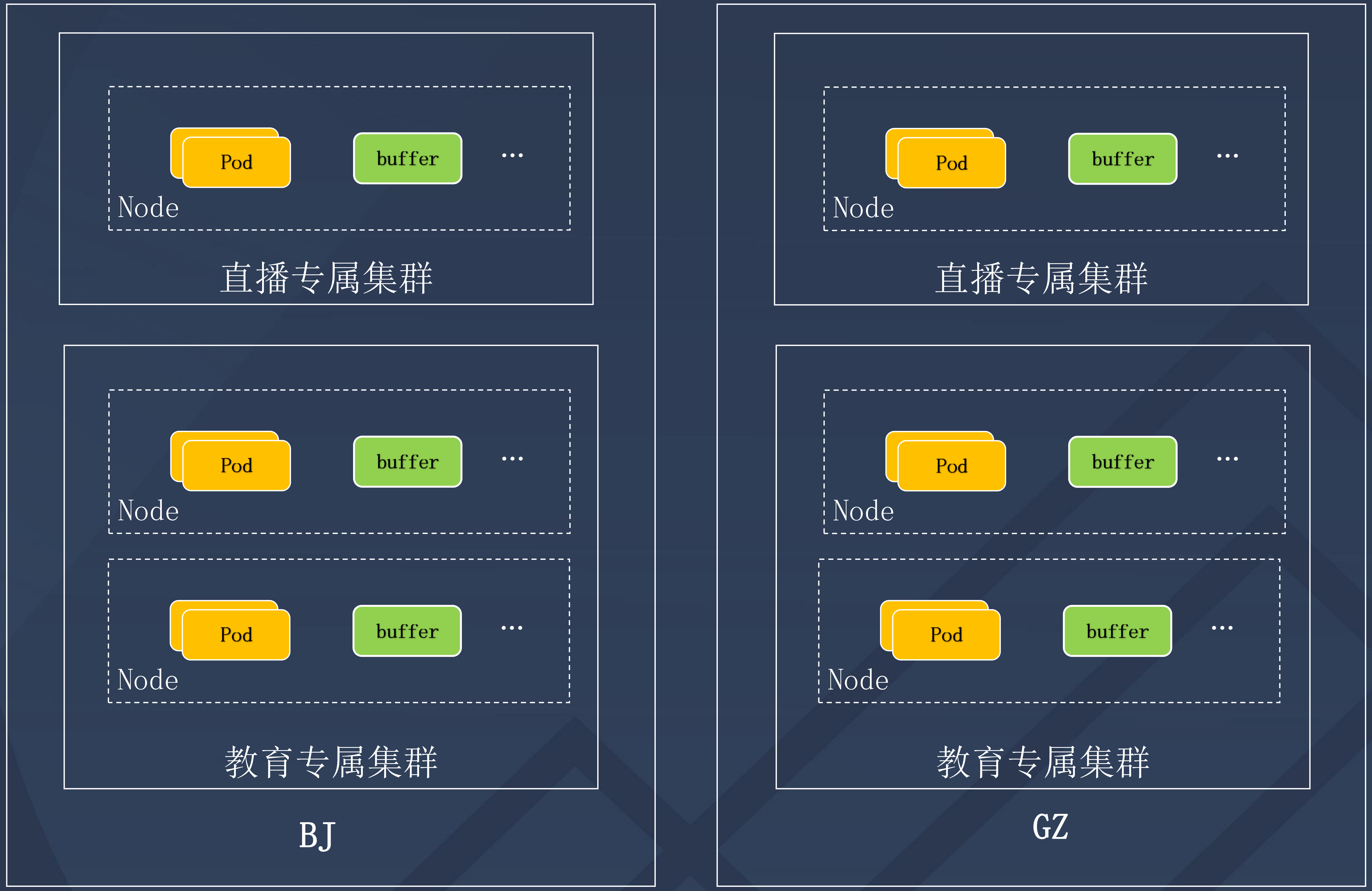
问题：

节点装箱率差

资源利用率低

运营成本高

海量节点运维成本高

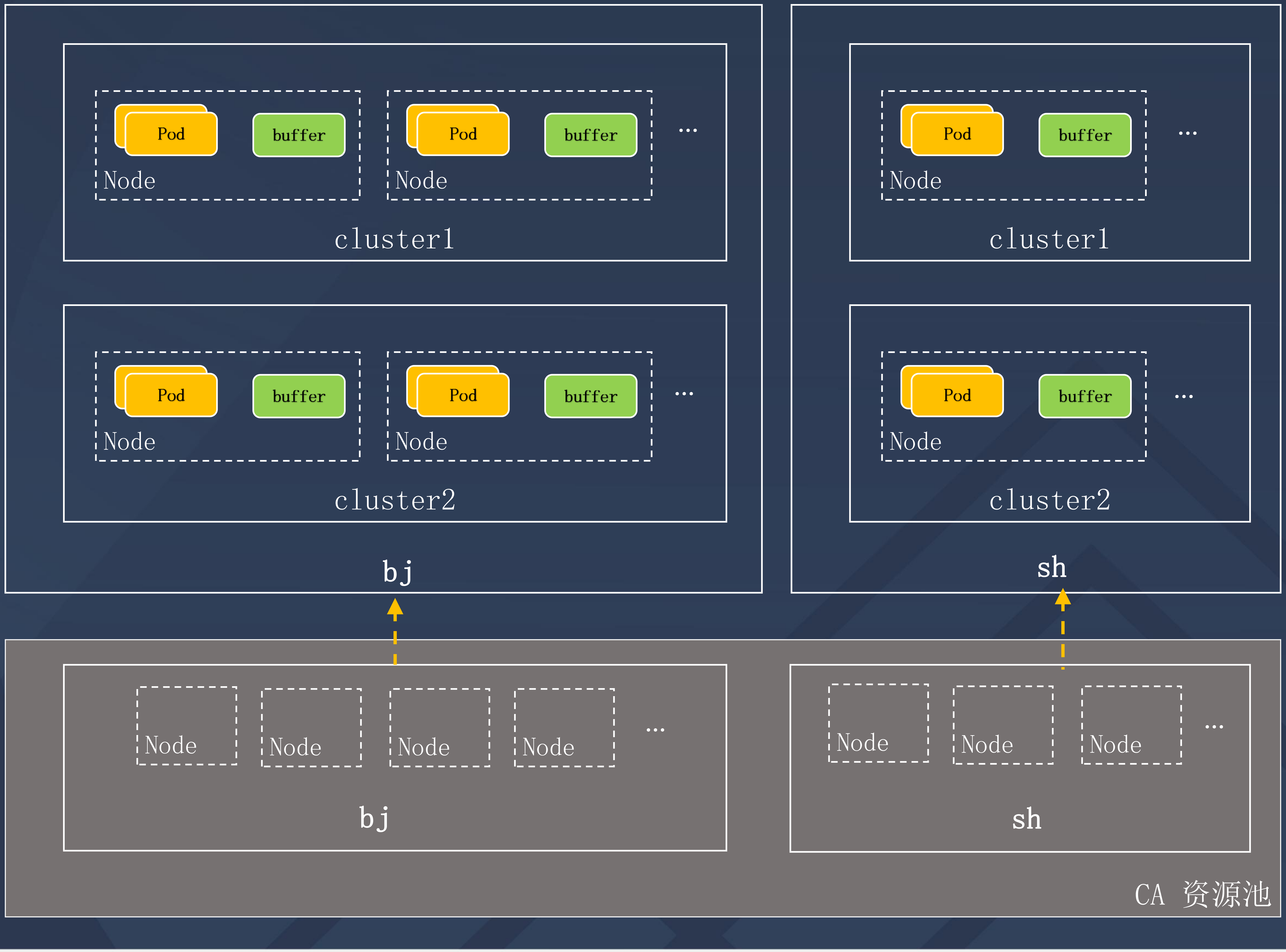


公共集群架构

期望：

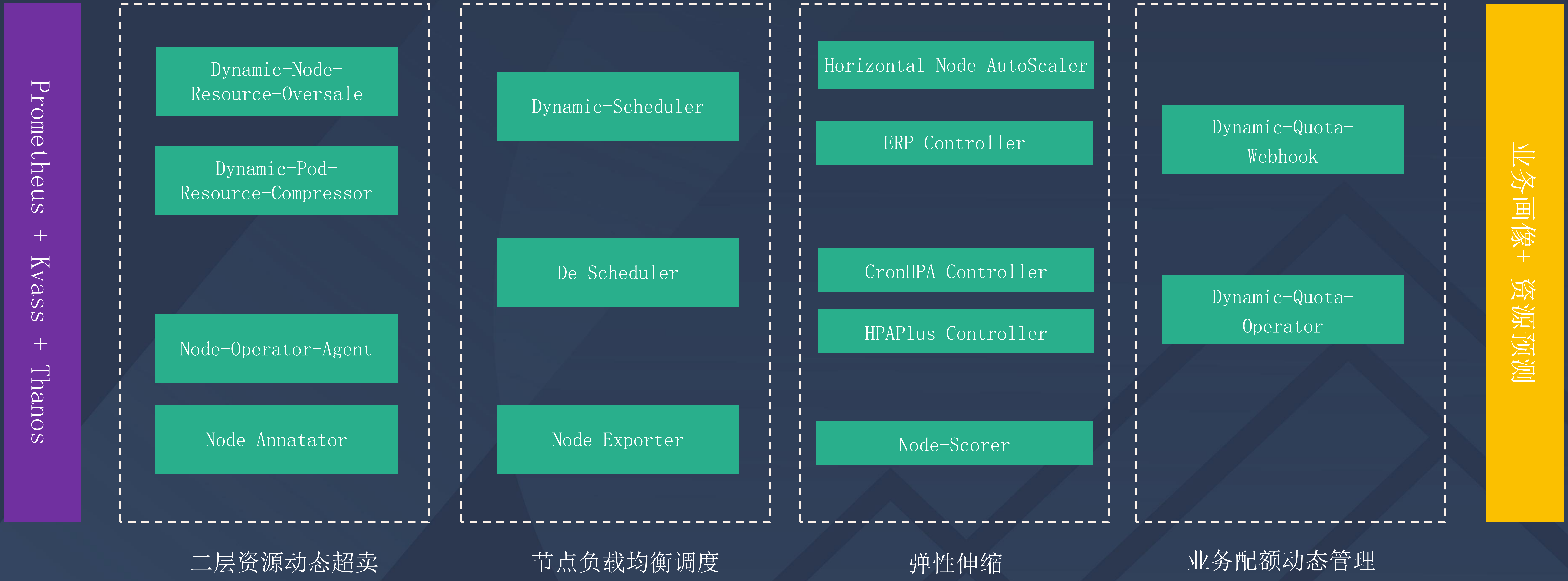
通过混部提升节点资源利用率

减少机器数量，降低成本



在线混部集群的资源利用率提升方案

在线混部集群利用率提升方案



Prometheus + Kvass + Thanos

Dynamic-Node-Resource-Oversale

Dynamic-Pod-Resource-Compressor

Node-Operator-Agent

Node Annatator

二层资源动态超卖

Dynamic-Scheduler

De-Scheduler

Node-Exporter

节点负载均衡调度

Horizontal Node AutoScaler

ERP Controller

CronHPA Controller

HPAPlus Controller

Node-Scorer

弹性伸缩

Dynamic-Quota-Webhook

Dynamic-Quota-Operator

业务配额动态管理

业务画像+ 资源预测

动态调度器

痛点：

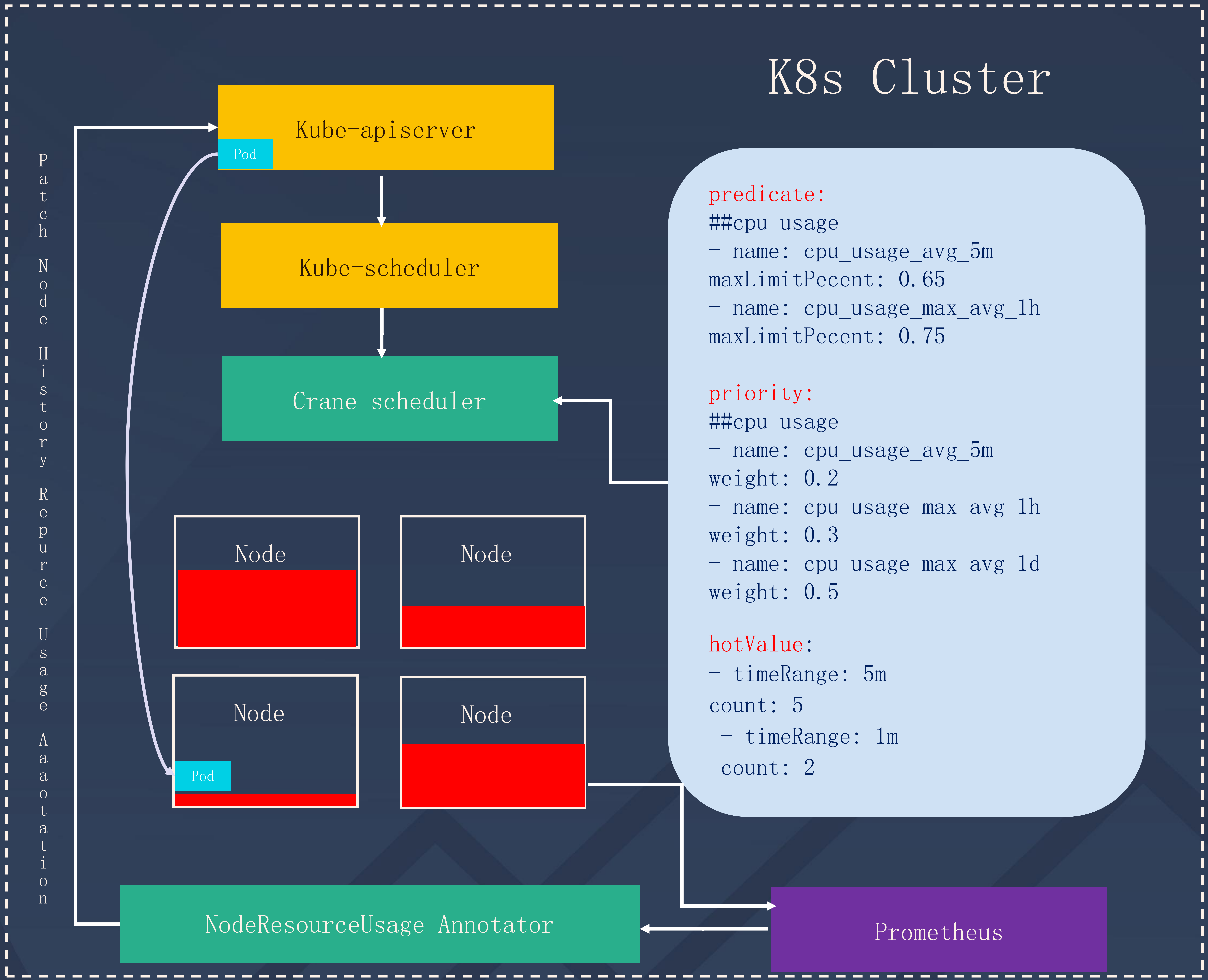
Kubernetes原生调度器属于静态调度，当大量业务混部在一个集群时，必然出现节点负载不均衡，Pod调度时仍可能往高负载节点上调度，造成业务服务质量下降。

方案：

自研动态调度器：让节点的关键资源在集群节点中均衡分布

Cpu/ Memory / Disk usage/ Network io / System load / Iowait / softirq

自研热点动态补偿算法解决调度热点问题



```
predicate:
##cpu usage
- name: cpu_usage_avg_5m
maxLimitPecent: 0.65
- name: cpu_usage_max_avg_1h
maxLimitPecent: 0.75

priority:
##cpu usage
- name: cpu_usage_avg_5m
weight: 0.2
- name: cpu_usage_max_avg_1h
weight: 0.3
- name: cpu_usage_max_avg_1d
weight: 0.5

hotValue:
- timeRange: 5m
count: 5
 - timeRange: 1m
 count: 2
```

二层动态资源超卖

技术挑战：

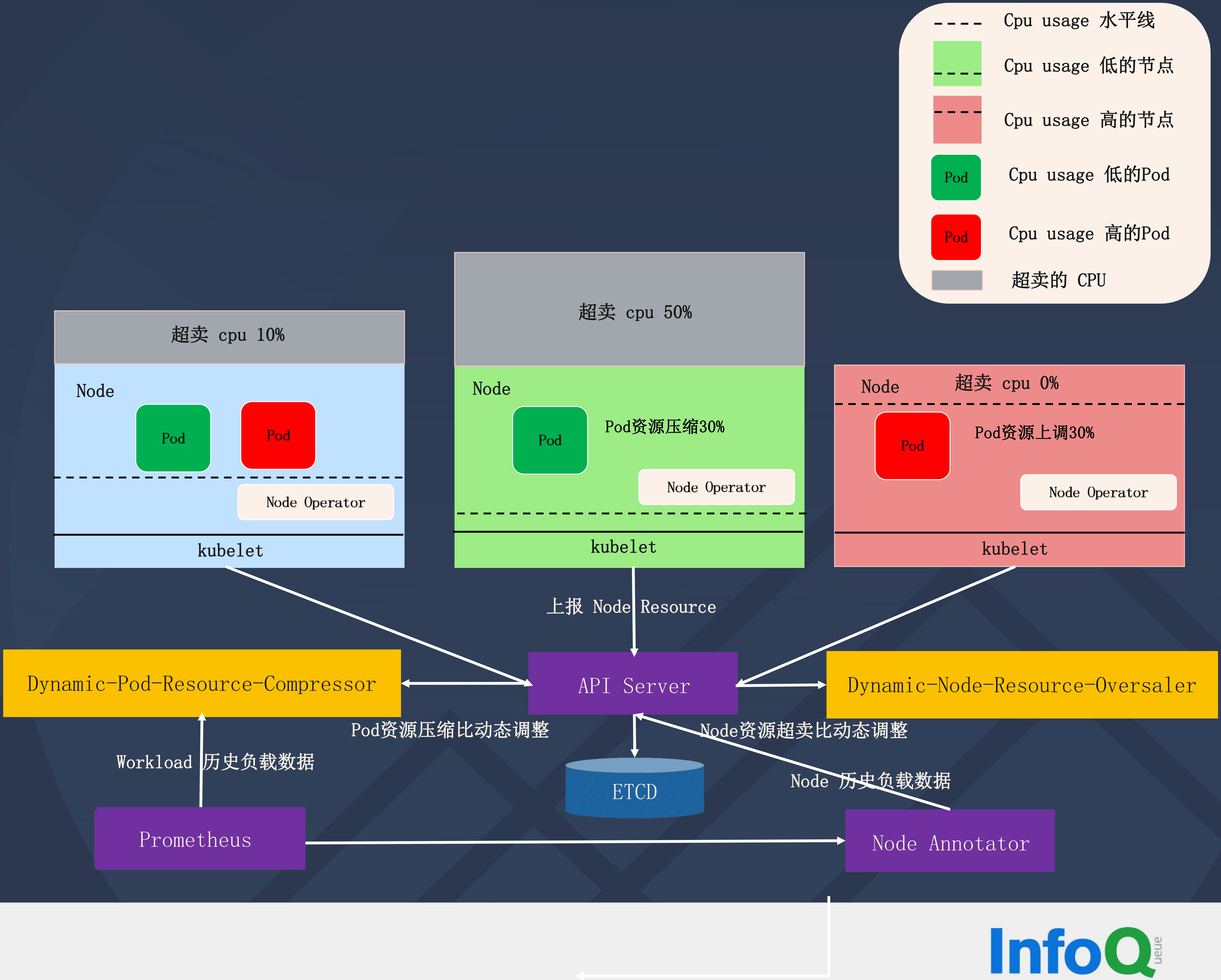
节点超卖比的安全控制，尽量不影响业务稳定性。

节点资源超卖对Kubernetes驱逐机制和资源预留机制的影响。超卖比变化需要动态调整kubelet对应的配置。

超卖比需要根据节点实时的负载数据进行动态调整，防止造成节点负载过高。

如果出现预料外的节点高负载，通过de-scheduler及时降低节点负载。

极端情况如果出现大面积节点高负载，通过HNA进行秒级扩容。



弹性伸缩-集群

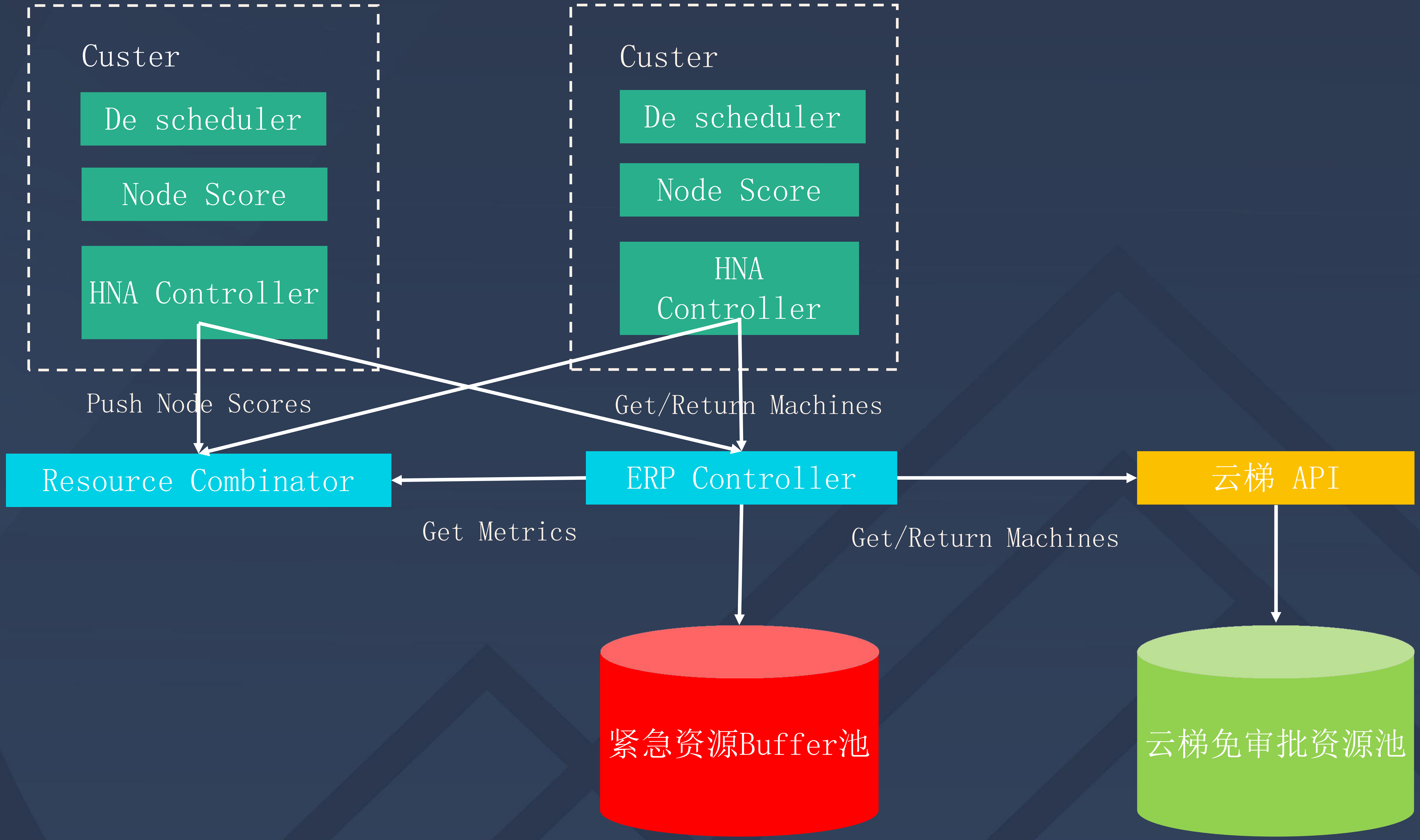
目标：

自研多集群资源协调器，将多集群的闲置资源构建统一的平台级Buffer池，让资源在多集群高效流转。

节点上下线实现自动化和标准化。

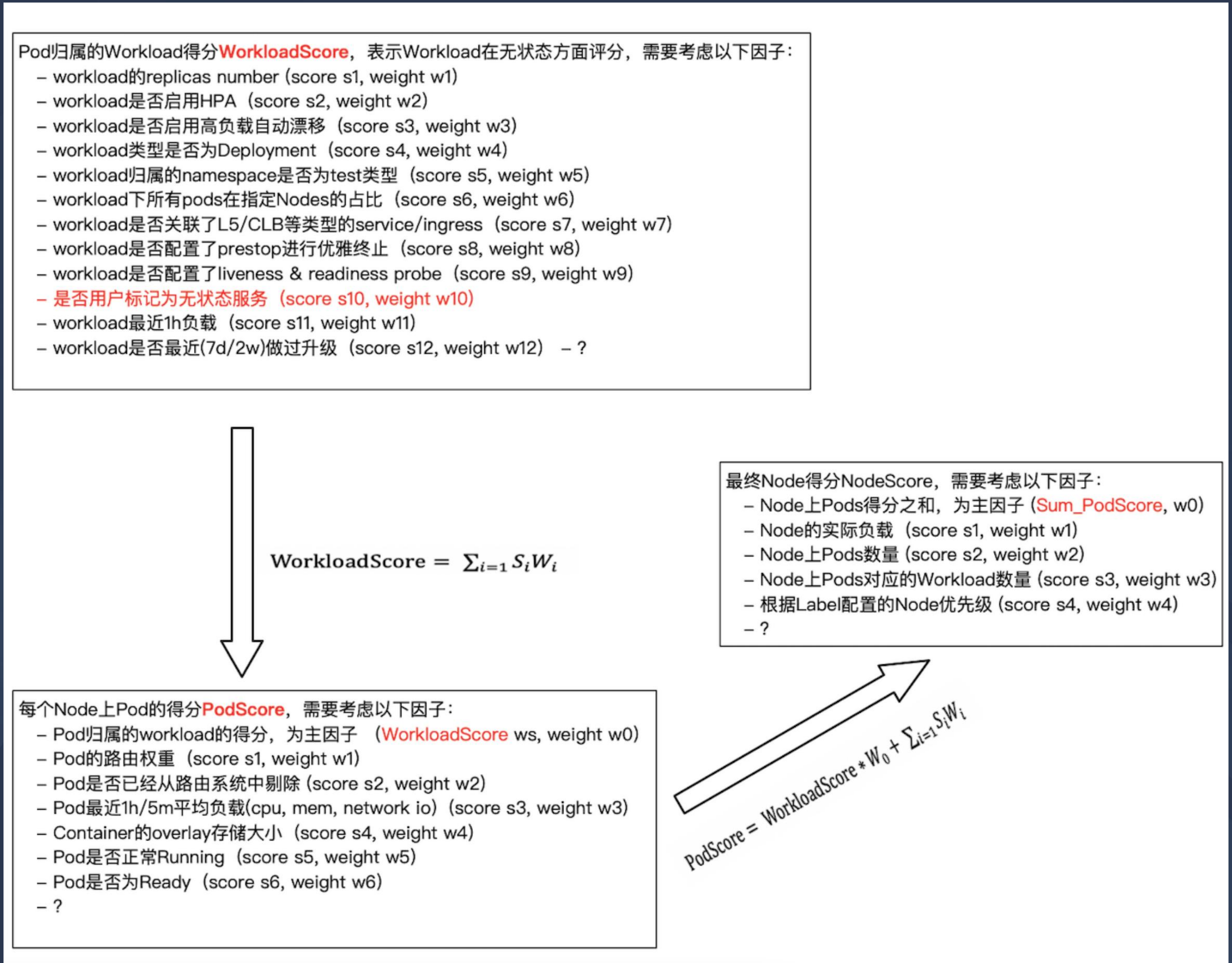
二级弹性资源池方案，支持常规扩缩容和紧急扩缩容2种场景。

集群负载高或者资源不足时，最快可实现小于1分钟的扩容速度，这样对降低集群负载有极大的益处。



集群缩容策略

node scorer 会周期性地对所有节点进行打分，我们认为某个阈值下得分的 node 都是可以缩容的



Pod归属的Workload得分WorkloadScore，表示Workload在无状态方面评分，需要考虑以下因子：

- workload的replicas number (score s1, weight w1)
- workload是否启用HPA（score s2, weight w2)
- workload是否启用高负载自动漂移（score s3, weight w3)
- workload类型是否为Deployment（score s4, weight w4)
- workload归属的namespace是否为test类型（score s5, weight w5)
- workload下所有pods在指定Nodes的占比（score s6, weight w6)
- workload是否关联了L5/CLB等类型的service/ingress（score s7, weight w7)
- workload是否配置了prestop进行优雅终止（score s8, weight w8)
- workload是否配置了liveness & readiness probe（score s9, weight w9)
- 是否用户标记为无状态服务（score s10, weight w10)
- workload最近1h负载（score s11, weight w11)
- workload是否最近(7d/2w)做过升级（score s12, weight w12） – ?

$\text{WorkloadScore} = \sum_{i=1} S_i W_i$

每个Node上Pod的得分PodScore，需要考虑以下因子：

- Pod归属的workload的得分，为主因子 （WorkloadScore ws, weight w0)
- Pod的路由权重（score s1, weight w1)
- Pod是否已经从路由系统中剔除 (score s2, weight w2)
- Pod最近1h/5m平均负载(cpu, mem, network io)（score s3, weight w3)
- Container的overlay存储大小（score s4, weight w4)
- Pod是否正常Running（score s5, weight w5)
- Pod是否为Ready（score s6, weight w6)
- ?

$\text{PodScore} = \text{WorkloadScore} * W_0 + \sum_{i=1} S_i W_i$

最终Node得分NodeScore，需要考虑以下因子：

- Node上Pods得分之和，为主因子 (Sum_PodScore, w0)
- Node的实际负载（score s1, weight w1)
- Node上Pods数量 (score s2, weight w2)
- Node上Pods对应的Workload数量 (score s3, weight w3)
- 根据Label配置的Node优先级 (score s4, weight w4)
- ?

弹性伸缩-业务

HPAPlus-Controller：

支持业务常规弹性伸缩场景。

支持HPA对象自定义关键配置：扩缩容速率/计算周期/指标容忍度等。

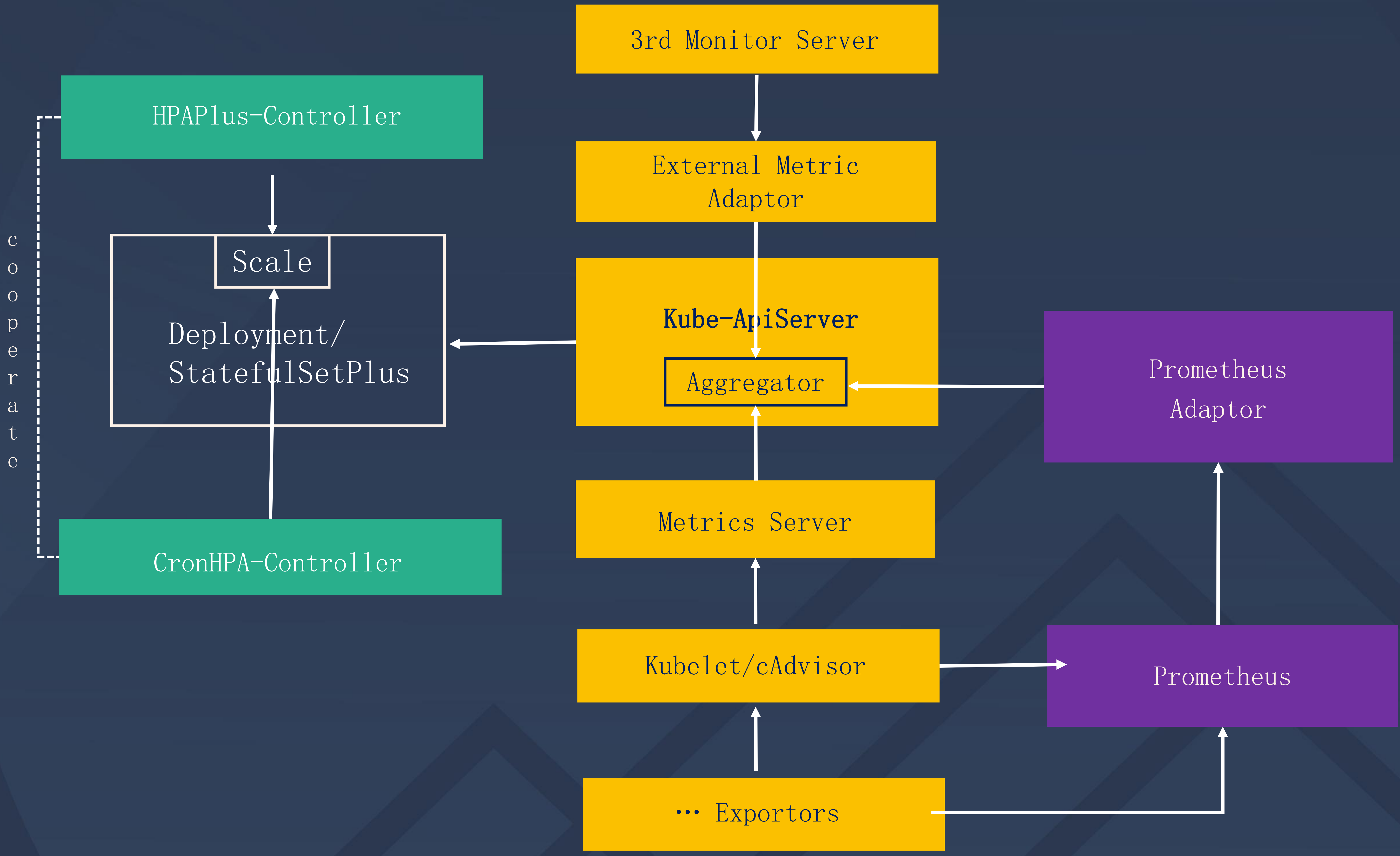
支持弹性的maxReplicas策略，避免超出预期的流量受限于maxReplicas配置太低，导致业务雪崩。

性能优化：支持几千个业务HPA对象并行弹性伸缩计算逻辑。

CronHPA-Controller：

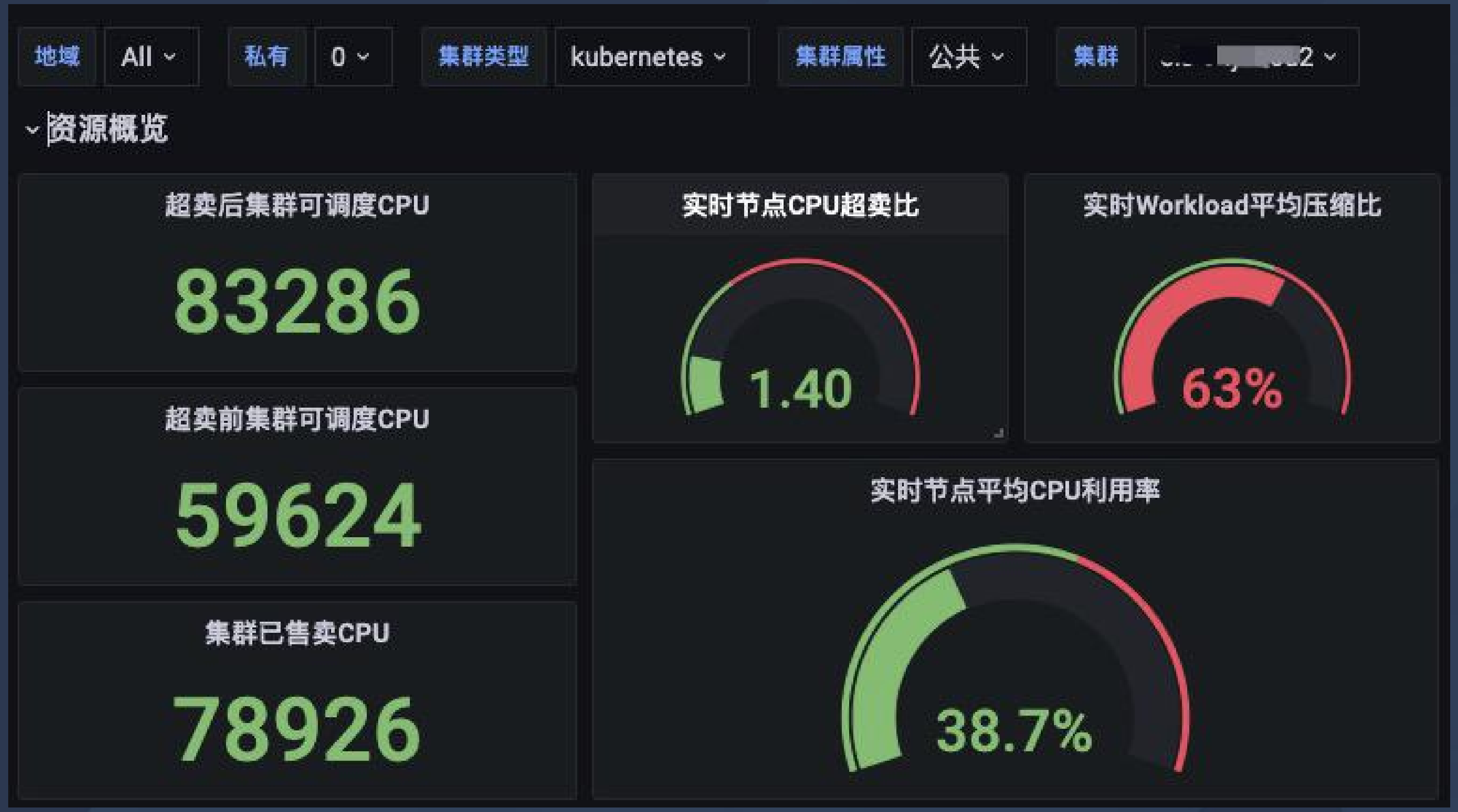
支持业务周期性弹性伸缩场景。

HPA与CronHPA联动决策：支持业务计划内的定时扩容策略，如果业务实际流量超过预估流量，仍能自动扩容。



在线混部利用率提升方案效果

通过在线业务混部超卖方案，
集群CPU平均利用率提升到30%~40%



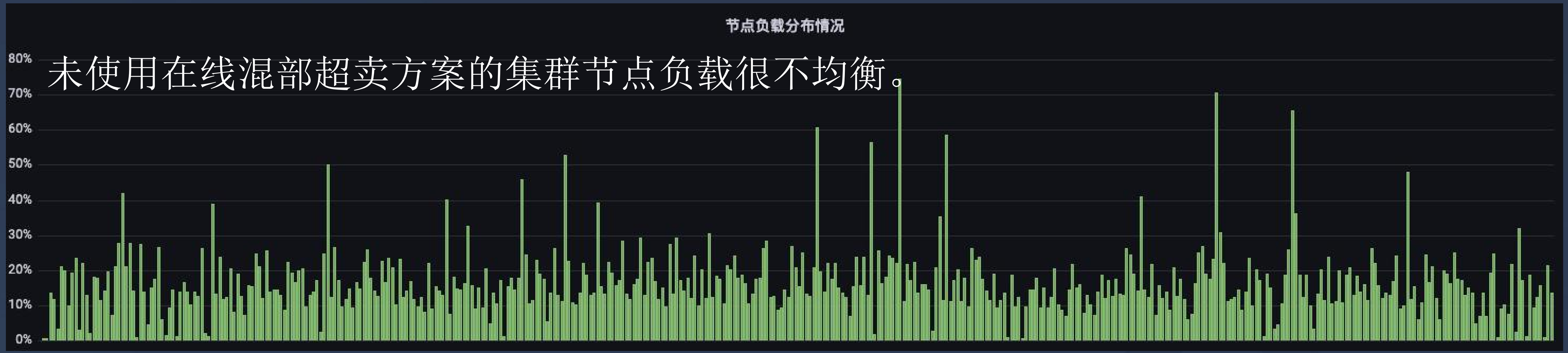
地域 All ˅ 私有 0 ˅ 集群类型 kubernetes ˅ 集群属性 公共 ˅ 集群 [redacted] ˅

˅ 资源概览

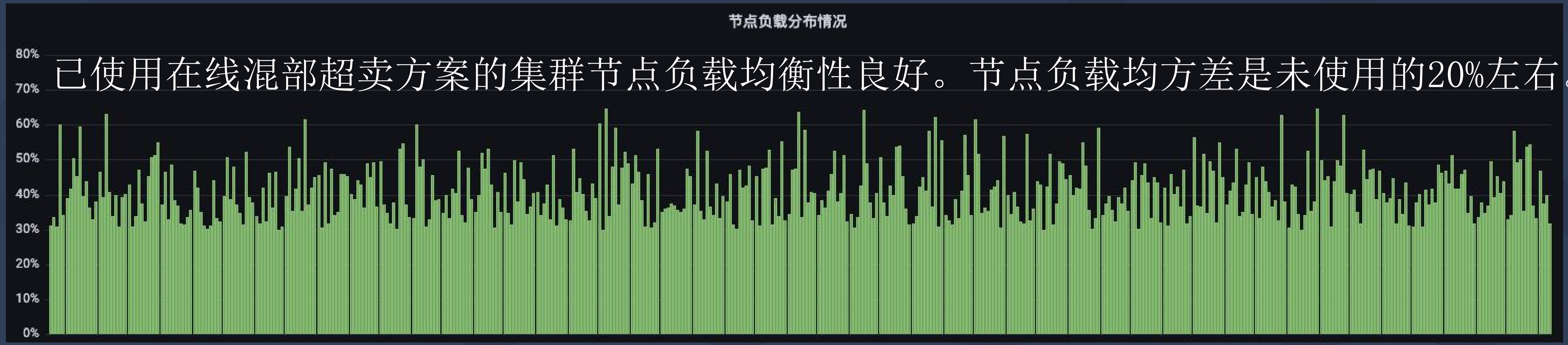
超卖后集群可调度CPU	实时节点CPU超卖比	实时Workload平均压缩比
83286	1.40	63%

超卖前集群可调度CPU	实时节点平均CPU利用率
59624	38.7%

集群已售卖CPU
78926



未使用在线混部超卖方案的集群节点负载很不均衡。



已使用在线混部超卖方案的集群节点负载均衡性良好。节点负载均方差是未使用的20%左右。

通过Crane开源全套技术
https://github.com/gocrane/crane

稳定性提升方案

业务经常因节点关键资源抢占导致业务服务质量下降。深入内核，从内核层面提供更丰富的节点及容器级的稳定性指标。在节点层面进行自愈，在容器层面进行协同调度编排。

Dockerd/Containerd/Kubelet状态和异常日志分析

Umount failed | Shim残留 | Syncloop hung住 | 进程D状态 | Cgroup 泄露/残留 | Container残留

Pod restart frequency

Pod Load.r/load.d

OS稳定性指标检测

Pod Long sys

Memory usage | 数据盘 usage | PID Pressure | D状态进程数 | Iowait | System load | FD Pressure |节点网络异常检测

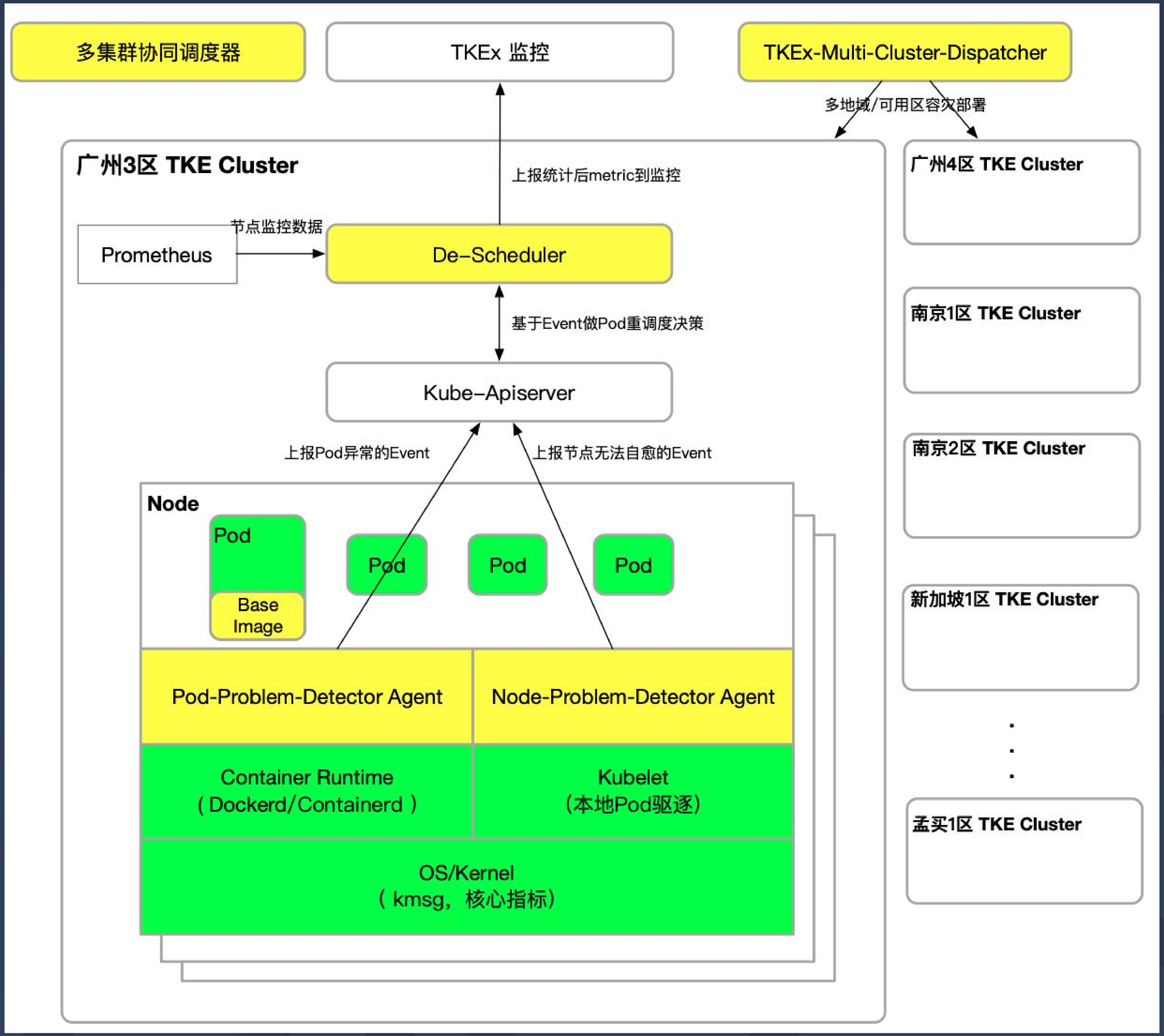
Pod Cpu调度延时

Pod Iowait

内核稳定性事件检测 （云原生TencentOS）

Pod 内存分配延时

Kernel死锁 | Softlockup | Hungtask | RCU Stall | Kernel Panic



拥抱腾讯云弹性容器服务 EKS 价值所在

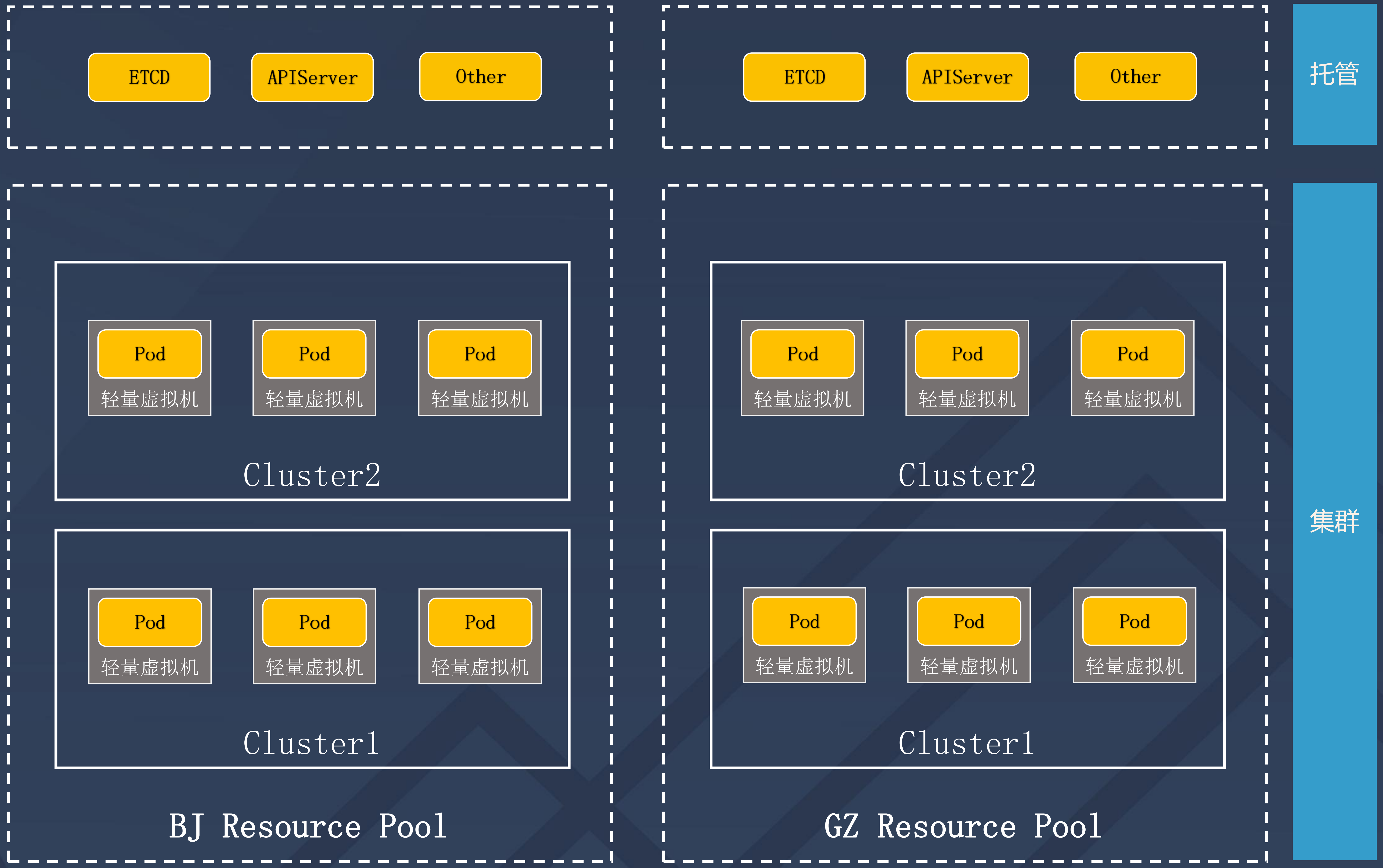
腾讯云弹性容器服务 EKS – 架构演进最终选择

介绍：

采用 Serverless 架构，以 Pod 为交付资源 —— 无须在集群添置节点，即可部署工作负载。容器弹性不受固定资源池限制，理论上可以无限制扩容。

采用Pod间虚拟化隔离技术，每个 Pod 独享虚拟机，不会收到集群其他异常 Pod 干扰。

通过 Pod 而非节点计费 —— 根据 Pod 的资源配置及运行时间计费，容器运行结束自动停止计费，无须为 buffer 资源付费。



产品优势

提供云原生标准协议

- 支持 k8s 编排，完全兼容社区 k8s api
- 支持 k8s 扩展性
- 支持社区生态

高性能

- 计算、网络性能媲美云服务器
- 定制内核
- 负载均衡流量直通容器

高可用

- 主备、多副本管控组件
- 可自动/指定跨 zone 部署应用
- 容器支持热迁移

支持异构算力

- 丰富的 Intel 型号
- 腾讯云自研 AMD
- 多种主流 GPU 型号
- 虚拟化 GPU

弹性效率

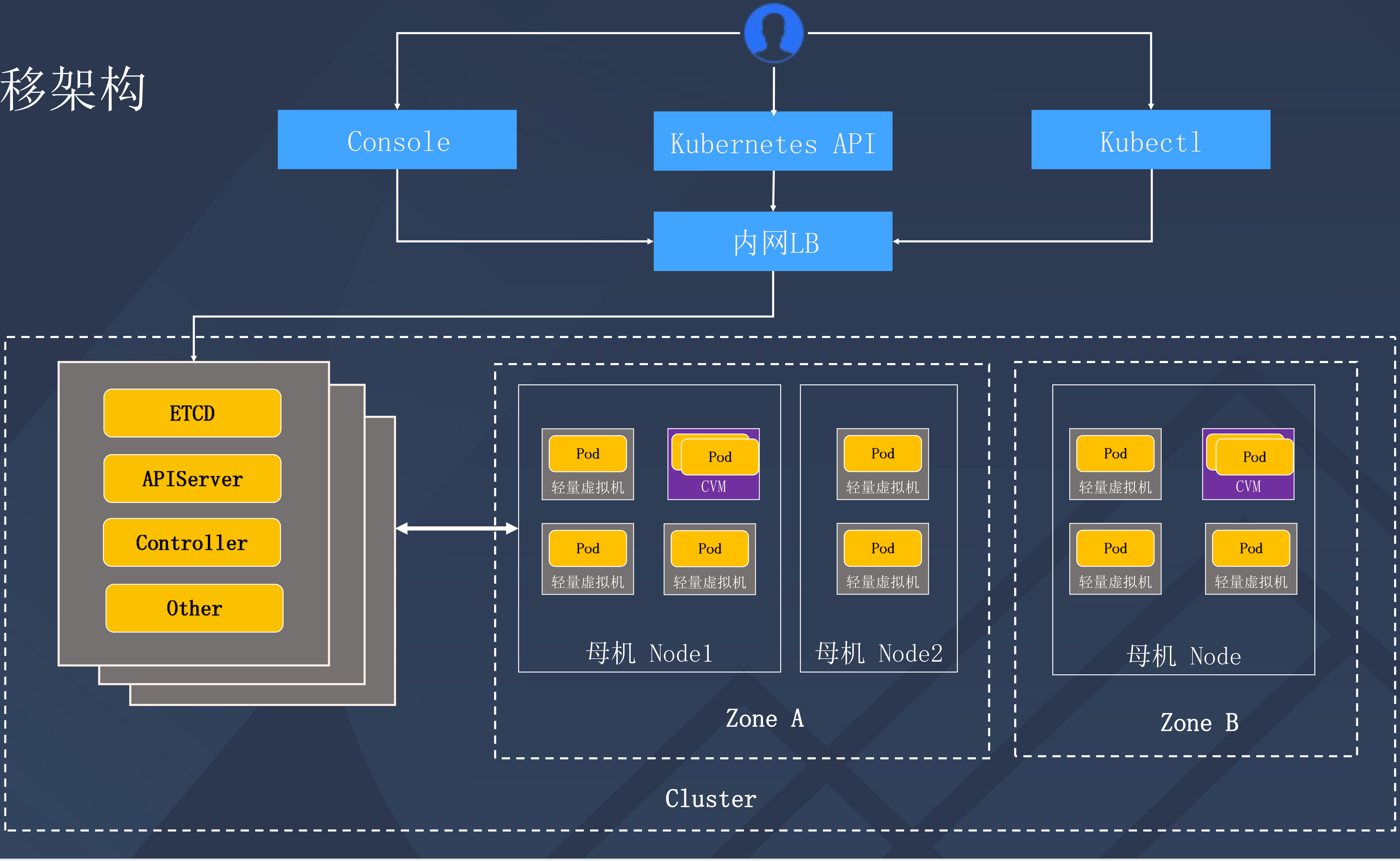
- 秒级冷启动
- 支持镜像复用技术
- 支持数万容器并发创建
- 支持敏感扩容、定时扩容

安全性

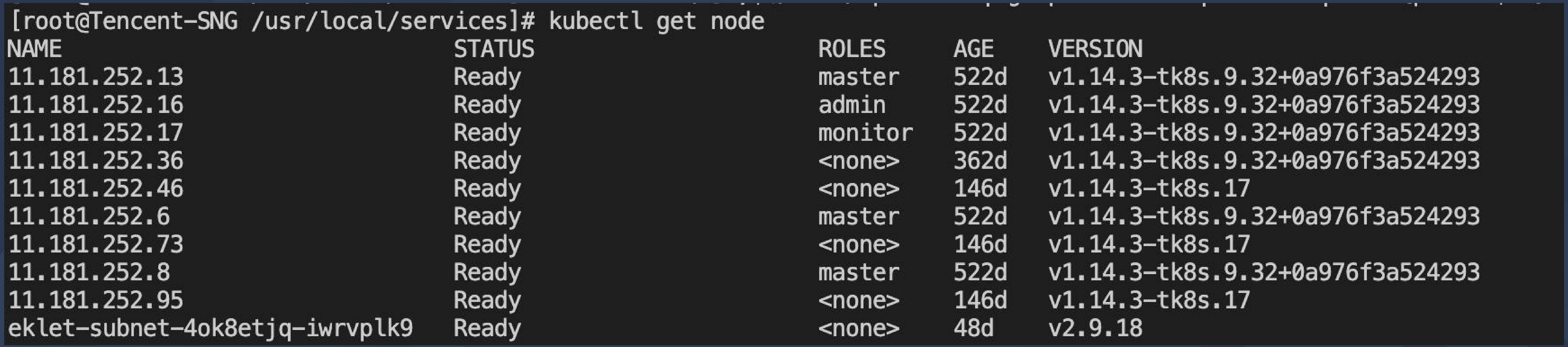
- 容器间虚拟化隔离
- 集群间网络、管控隔离
- 租户间绝对隔离
- 管控与数据面隔离

存量 K8s 集群应用平滑迁移弹性容器服务 EKS 的落地实践

平滑迁移架构

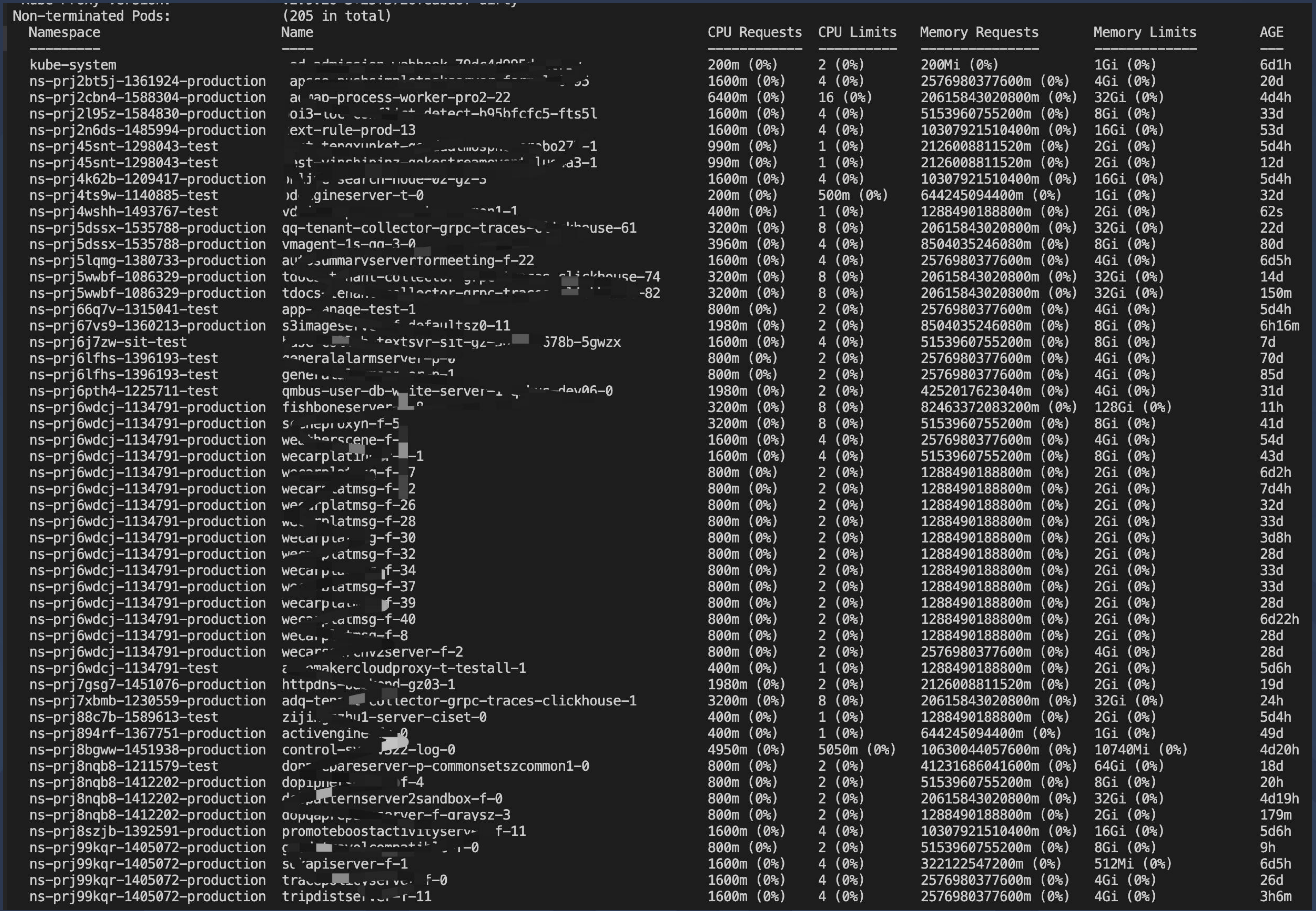

Console
Kubernetes API
Kubectl
内网LB
ETCD
APIServer
Controller
Other
Pod
轻量虚拟机
Pod
CVM
Pod
轻量虚拟机
Pod
轻量虚拟机
Pod
轻量虚拟机
Pod
轻量虚拟机
母机 Node1
母机 Node2
Zone A
Pod
轻量虚拟机
Pod
CVM
Pod
轻量虚拟机
Pod
轻量虚拟机
母机 Node
Zone B
Cluster

命令行展示



```
[root@Tencent-SNG /usr/local/services]# kubectl get node
NAME                            STATUS                      ROLES     AGE    VERSION
11.181.252.13                   Ready                       master    522d   v1.14.3-tk8s.9.32+0a976f3a524293
11.181.252.16                   Ready                       admin     522d   v1.14.3-tk8s.9.32+0a976f3a524293
11.181.252.17                   Ready                       monitor   522d   v1.14.3-tk8s.9.32+0a976f3a524293
11.181.252.36                   Ready                       <none>    362d   v1.14.3-tk8s.9.32+0a976f3a524293
11.181.252.46                   Ready                       <none>    146d   v1.14.3-tk8s.17
11.181.252.6                    Ready                       master    522d   v1.14.3-tk8s.9.32+0a976f3a524293
11.181.252.73                   Ready                       <none>    146d   v1.14.3-tk8s.17
11.181.252.8                    Ready                       master    522d   v1.14.3-tk8s.9.32+0a976f3a524293
11.181.252.95                   Ready                       <none>    146d   v1.14.3-tk8s.17
eklet-subnet-4ok8etjq-iwrvplk9  Ready                       <none>    48d    v2.9.18
```

命令行展示



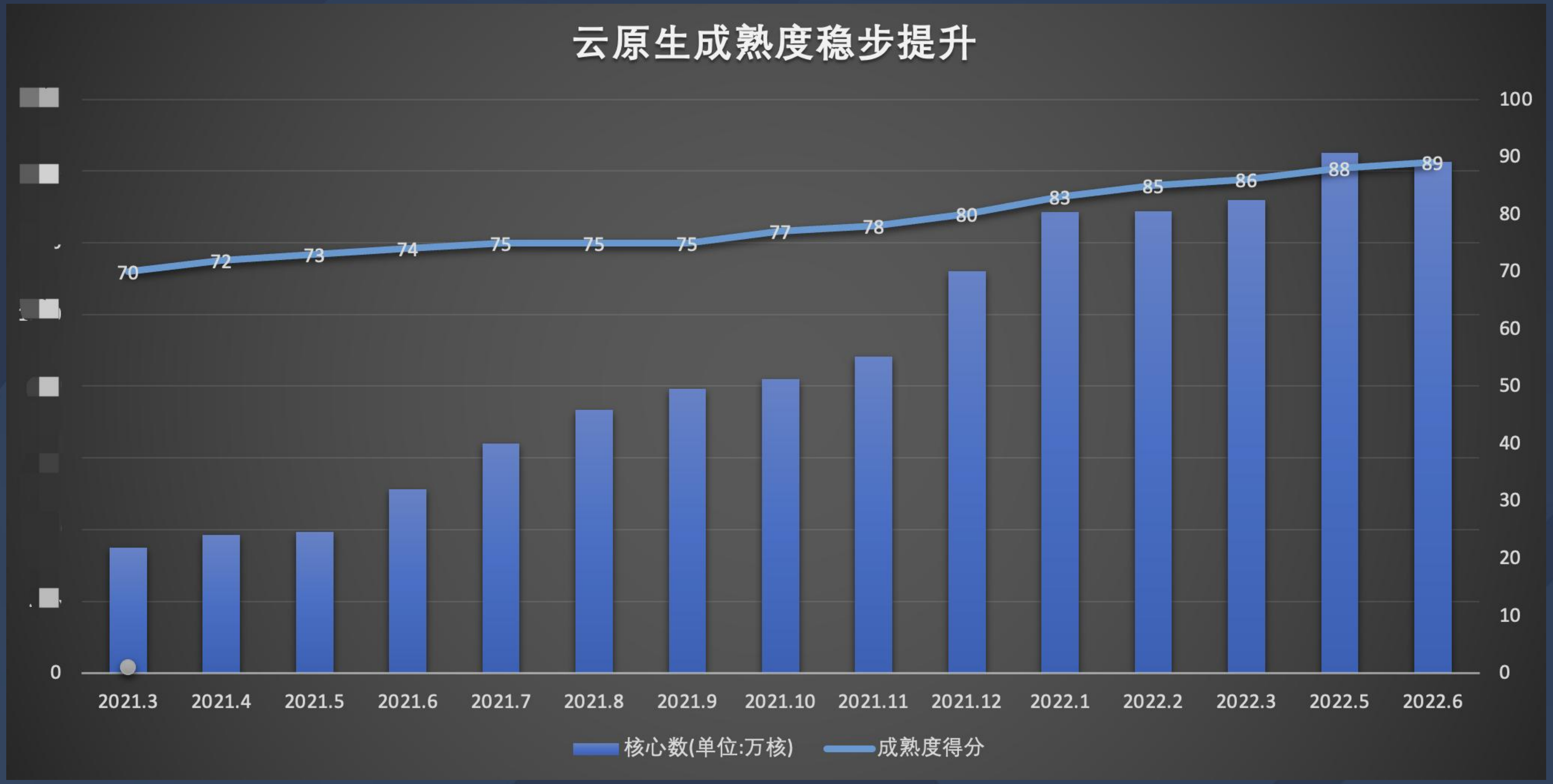
```
Non-terminated Pods:                 (205 in total)
  Namespace                          Name                                                CPU Requests  CPU Limits   Memory Requests       Memory Limits    AGE
  ---------                          ----                                                ------------  ----------   ---------------       -------------    ---
  kube-system                        [redacted]                                          200m (0%)     2 (0%)       200Mi (0%)            1Gi (0%)         6d1h
  ns-prj2bt5j-1361924-production     ap[redacted]                                        1600m (0%)    4 (0%)       2576980377600m (0%)   4Gi (0%)         20d
  ns-prj2cbn4-1588304-production     a[redacted]ap-process-worker-pro2-22                6400m (0%)    16 (0%)      20615843020800m (0%)  32Gi (0%)        4d4h
  ns-prj2l95z-1584830-production     oi3-[redacted]-detect-b95bfcfc5-fts5l               1600m (0%)    4 (0%)       5153960755200m (0%)   8Gi (0%)         33d
  ns-prj2n6ds-1485994-production     [redacted]ext-rule-prod-13                          1600m (0%)    4 (0%)       10307921510400m (0%)  16Gi (0%)        53d
  ns-prj45snt-1298043-test           [redacted]bo27[redacted]-1                          990m (0%)     1 (0%)       2126008811520m (0%)   2Gi (0%)         5d4h
  ns-prj45snt-1298043-test           [redacted]st[redacted]a3-1                          990m (0%)     1 (0%)       2126008811520m (0%)   2Gi (0%)         12d
  ns-prj4k62b-1209417-production     [redacted]search-node-02-gz-3                       1600m (0%)    4 (0%)       10307921510400m (0%)  16Gi (0%)        5d4h
  ns-prj4ts9w-1140885-test           [redacted]gineserver-t-0                            200m (0%)     500m (0%)    644245094400m (0%)    1Gi (0%)         32d
  ns-prj4wshh-1493767-test           vd[redacted]1-1                                     400m (0%)     1 (0%)       1288490188800m (0%)   2Gi (0%)         62s
  ns-prj5dssx-1535788-production     qq-tenant-collector-grpc-traces-[redacted]use-61    3200m (0%)    8 (0%)       20615843020800m (0%)  32Gi (0%)        22d
  ns-prj5dssx-1535788-production     vmagent-1s-qq-3-0                                   3960m (0%)    4 (0%)       8504035246080m (0%)   8Gi (0%)         80d
  ns-prj5lqmg-1380733-production     au[redacted]summaryserverformeeting-f-22            1600m (0%)    4 (0%)       2576980377600m (0%)   4Gi (0%)         6d5h
  ns-prj5wwbf-1086329-production     tdo[redacted]clickhouse-74                          3200m (0%)    8 (0%)       20615843020800m (0%)  32Gi (0%)        14d
  ns-prj5wwbf-1086329-production     tdocs-[redacted]-82                                 3200m (0%)    8 (0%)       20615843020800m (0%)  32Gi (0%)        150m
  ns-prj66q7v-1315041-test           app-[redacted]anage-test-1                          800m (0%)     2 (0%)       2576980377600m (0%)   4Gi (0%)         5d4h
  ns-prj67vs9-1360213-production     s3imagese[redacted]defaultsz0-11                    1980m (0%)    2 (0%)       8504035246080m (0%)   8Gi (0%)         6h16m
  ns-prj6j7zw-sit-test               [redacted]textsvr-sit-gz-[redacted]578b-5gwzx       1600m (0%)    4 (0%)       5153960755200m (0%)   8Gi (0%)         7d
  ns-prj6lfhs-1396193-test           [redacted]eralalarmserve[redacted]                  800m (0%)     2 (0%)       2576980377600m (0%)   4Gi (0%)         70d
  ns-prj6lfhs-1396193-test           genera[redacted]-1                                  800m (0%)     2 (0%)       2576980377600m (0%)   4Gi (0%)         85d
  ns-prj6pth4-1225711-test           qmbus-user-db-w[redacted]ite-server-[redacted]dev06-0  1980m (0%)  2 (0%)       4252017623040m (0%)   4Gi (0%)         31d
  ns-prj6wdcj-1134791-production     fishboneserver-[redacted]                           3200m (0%)    8 (0%)       82463372083200m (0%)  128Gi (0%)       11h
  ns-prj6wdcj-1134791-production     s[redacted]eproxyn-f-5                              3200m (0%)    8 (0%)       5153960755200m (0%)   8Gi (0%)         41d
  ns-prj6wdcj-1134791-production     wea[redacted]erscene-f-[redacted]                   1600m (0%)    4 (0%)       2576980377600m (0%)   4Gi (0%)         54d
  ns-prj6wdcj-1134791-production     wecarplati[redacted]-1                              1600m (0%)    4 (0%)       5153960755200m (0%)   8Gi (0%)         43d
  ns-prj6wdcj-1134791-production     w[redacted]g-f-[redacted]7                          800m (0%)     2 (0%)       1288490188800m (0%)   2Gi (0%)         6d2h
  ns-prj6wdcj-1134791-production     wecar[redacted]atmsg-f-[redacted]2                  800m (0%)     2 (0%)       1288490188800m (0%)   2Gi (0%)         7d4h
  ns-prj6wdcj-1134791-production     we[redacted]rplatmsg-f-26                           800m (0%)     2 (0%)       1288490188800m (0%)   2Gi (0%)         32d
  ns-prj6wdcj-1134791-production     w[redacted]platmsg-f-28                             800m (0%)     2 (0%)       1288490188800m (0%)   2Gi (0%)         33d
  ns-prj6wdcj-1134791-production     wecarpla[redacted]g-f-30                            800m (0%)     2 (0%)       1288490188800m (0%)   2Gi (0%)         3d8h
  ns-prj6wdcj-1134791-production     we[redacted]atmsg-f-32                              800m (0%)     2 (0%)       1288490188800m (0%)   2Gi (0%)         28d
  ns-prj6wdcj-1134791-production     wecarp[redacted]f-34                                800m (0%)     2 (0%)       1288490188800m (0%)   2Gi (0%)         33d
  ns-prj6wdcj-1134791-production     w[redacted]atmsg-f-37                               800m (0%)     2 (0%)       1288490188800m (0%)   2Gi (0%)         33d
  ns-prj6wdcj-1134791-production     wecarpla[redacted]f-39                              800m (0%)     2 (0%)       1288490188800m (0%)   2Gi (0%)         28d
  ns-prj6wdcj-1134791-production     we[redacted]atmsg-f-40                              800m (0%)     2 (0%)       1288490188800m (0%)   2Gi (0%)         6d22h
  ns-prj6wdcj-1134791-production     weca[redacted]-f-8                                  800m (0%)     2 (0%)       1288490188800m (0%)   2Gi (0%)         28d
  ns-prj6wdcj-1134791-production     wecar[redacted]env2server-f-2                       800m (0%)     2 (0%)       2576980377600m (0%)   4Gi (0%)         28d
  ns-prj6wdcj-1134791-test           a[redacted]makercloudproxy-t-testall-1              400m (0%)     1 (0%)       1288490188800m (0%)   2Gi (0%)         5d6h
  ns-prj7gsg7-1451076-production     httpdns-[redacted]nd-gz03-1                         1980m (0%)    2 (0%)       2126008811520m (0%)   2Gi (0%)         19d
  ns-prj7xbmb-1230559-production     adq-ten[redacted]collector-grpc-traces-clickhouse-1  3200m (0%)   8 (0%)       20615843020800m (0%)  32Gi (0%)        24h
  ns-prj88c7b-1589613-test           ziji[redacted]hu1-server-ciset-0                    400m (0%)     1 (0%)       1288490188800m (0%)   2Gi (0%)         5d4h
  ns-prj894rf-1367751-production     activengine[redacted]0                              400m (0%)     1 (0%)       644245094400m (0%)    1Gi (0%)         49d
  ns-prj8bgww-1451938-production     control-[redacted]322-log-0                         4950m (0%)    5050m (0%)   10630044057600m (0%)  10740Mi (0%)     4d20h
  ns-prj8nqb8-1211579-test           dor[redacted]pareserver-p-commonsetszcommon1-0      800m (0%)     2 (0%)       41231686041600m (0%)  64Gi (0%)        18d
  ns-prj8nqb8-1412202-production     dopip[redacted]f-4                                  800m (0%)     2 (0%)       5153960755200m (0%)   8Gi (0%)         20h
  ns-prj8nqb8-1412202-production     d[redacted]patternserver2sandbox-f-0                800m (0%)     2 (0%)       20615843020800m (0%)  32Gi (0%)        4d19h
  ns-prj8nqb8-1412202-production     dopqap[redacted]erver-f-graysz-3                    800m (0%)     2 (0%)       1288490188800m (0%)   2Gi (0%)         179m
  ns-prj8szjb-1392591-production     promoteboostactivityserv[redacted]f-11              1600m (0%)    4 (0%)       10307921510400m (0%)  16Gi (0%)        5d6h
  ns-prj99kqr-1405072-production     g[redacted]-0                                       800m (0%)     2 (0%)       5153960755200m (0%)   8Gi (0%)         9h
  ns-prj99kqr-1405072-production     s[redacted]apiserver-f-1                            1600m (0%)    4 (0%)       322122547200m (0%)    512Mi (0%)       6d5h
  ns-prj99kqr-1405072-production     tra[redacted]f-0                                    1600m (0%)    4 (0%)       2576980377600m (0%)   4Gi (0%)         26d
  ns-prj99kqr-1405072-production     tripdistse[redacted]-11                             1600m (0%)    4 (0%)       2576980377600m (0%)   4Gi (0%)         3h6m
```

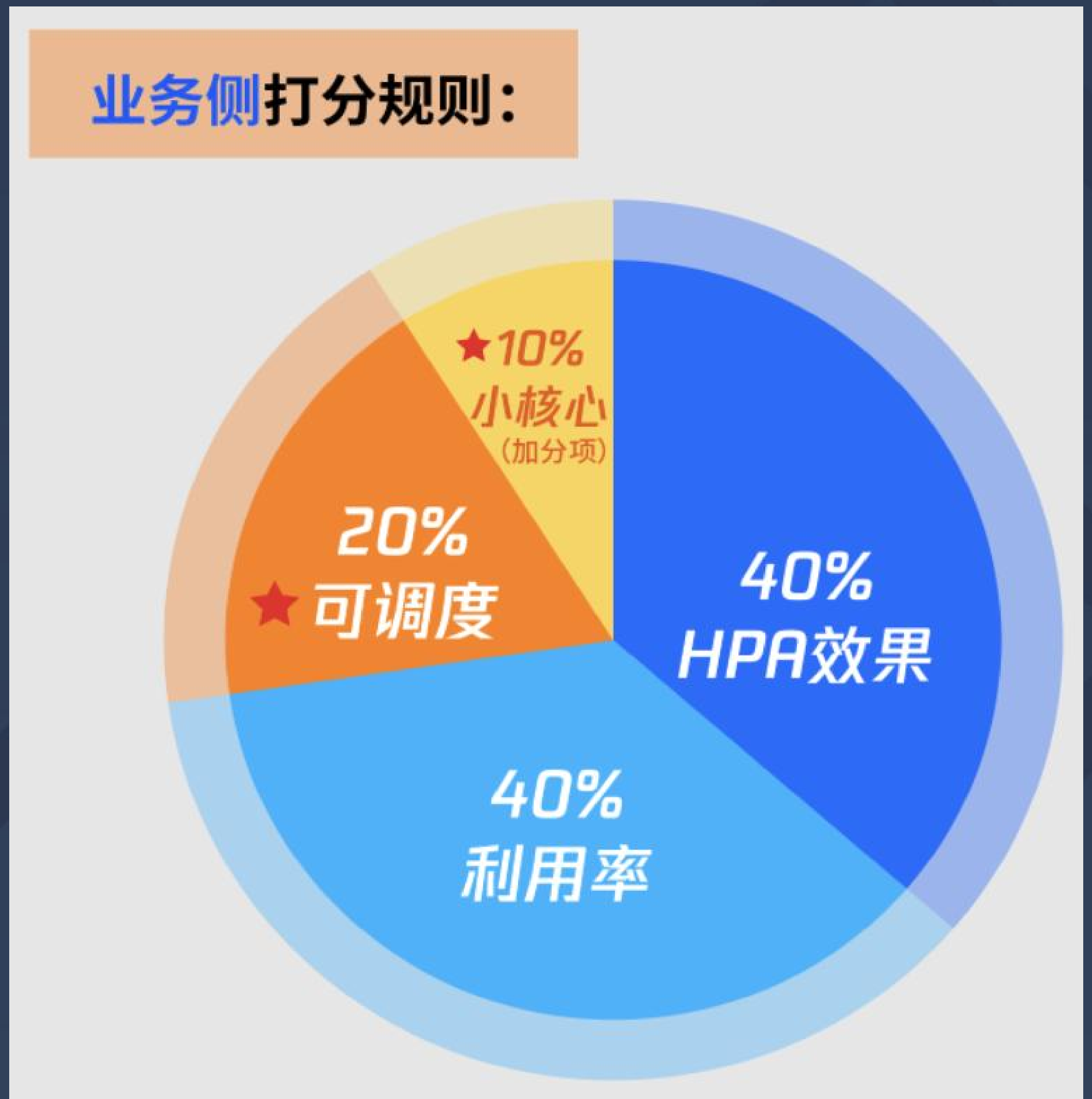
正在做的事情

引入EHPA，预测传入的峰值流量并提前扩展其副本数，提升用户使用 HPA 的信心。

通过服务负载历史数据，给用户进行资源配置推荐、以此提升资源利用率。

针对需要固定副本数的工作负载，进行副本数推荐；可以开启HPA 的工作负载，进行上下限副本数推荐





InfoQ 传媒和整合营销服务

对技术人群极具影响力的新闻网站 / 技术社区

InfoQ 是一家全球性的在线新闻 / 社区网站，创立于 2006 年，创始人是 Floyd Marinescu。目前全球拥有英、法、中、日共五种语言的站点。InfoQ 中国于 2007 年由极客邦科技创始人兼 CEO 霍太稳引入中国。

十五年来，InfoQ 致力于促进软件开发及相关领域知识与创新的传播，凭借在技术服务领域的深耕。

300W+	150W+	100W+	300W+	1600+
InfoQ 网站 月访问量	积累公众号 粉丝	微博 粉丝	覆盖中高端 技术开发者	CTO、 技术高管

THANKS

Global

Architect Summit