
GOTC 2023
全球开源技术峰会
THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE
OPEN SOURCE, INTO THE FUTURE
「基础设施与软件架构」专场
本期议题：从 OpenCloudOS 的演进之路，看开源操作系统的突围与演进
陈海武 2023年5月28日

OpenCloudOS 发展历程及定位

OpenCloudOS：十年积累，千万节点

GOTC

2010 开始研发，10+年积累　　**1000** 规模超千万，行业领先　　**99.999%** 可用性满足企业级要求

自主研发时代
自主研发运营、持续打磨

2010年
开始自主研发

代替外购SUSE

2011年
发布第一个版本TS1

精简内核

稳定性/性能提升

新硬件支持

新技术引进

功能定制

持续运营打磨

2016年
发布TS2

自研覆盖99%

支撑微信、QQ、游戏等核心业务/tlinux

创新研发时代
向外生长、社区生态、技术引领

2019年
发布TS3，升级品牌

输出到公有云客户

小红书、作业帮、拼多多

TCE 专有云平台基于TS构建，落地私有云客户

中行、建行、微众、富融、中银国际、深证通、富途

2020年
私有化场景拓展

私有云客户：

招行、上海银行、贵州银行、富融、内蒙古农信；

华夏、人保、民生人寿；

银联、中金、中信建投、方正证券

2021年
OpenCloudOS社区成立

打造生态、引领核心技术

私有云客户：

央行、农行、厦门国际、东莞银行、广州农商；

太平洋保险、山西证券、越秀金融云

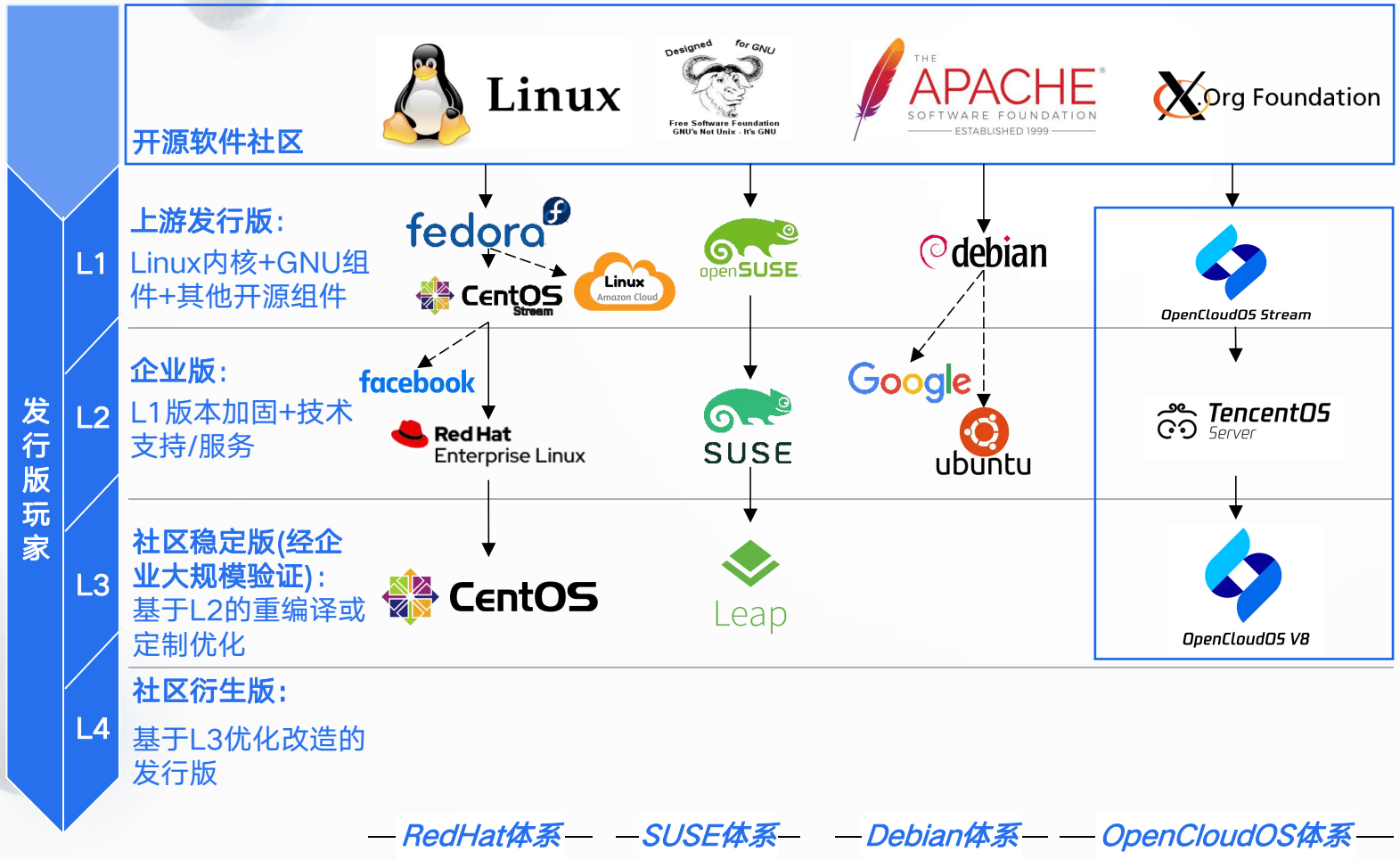
2022年
规模1000万

私有云客户：

人行、华夏银行、长江银行、上海农商、红塔银行

东北证券、中原消金、中体彩

Linux行业现状与OpenCloudOS社区定位：L1/L2/L3



行业问题:开源供应链安全风险

红帽不再维护CentOS8；国产OS对其强依赖，影响较大；**开源软件供应链存在安全风险**

L1国产发行版不足

L1上游发行版需聚焦多方协同开发，投入大，社区版本未经过大规模生产环境验证，**非稳定版本，无法直接用于生产环境**

L2国产商业版不足

L2国产商业版本稀缺。主要原因是上游社区维护**能力与投入不足**

L3/L4国产发行版不足

L3/L4社区聚焦版本的稳定和生产价值，但需要依赖可靠上游版本（商业版本）；

腾讯是OS上游的核心贡献者，是国内OS的核心技术力量

Linux KVM社区连续6年国内贡献第一
贡献20% Features
内核整体核心贡献保持国内前列

OpenJDK社区连续4年国内贡献第一
(JDK15-18)
国内唯一Notable贡献厂商

OpenCloudOS 创办自主开源OS社区

Top 8 全球开源影响力

自主内核演进

- TK1 2.6
- TK2 3.10
- TK3 4.14
- TK4 5.4
  - 3年打磨，完全稳定
  - 兼顾稳定性和新场景(云原生)

OpenCloudOS 9 版本介绍—产品特色

紧跟社区上游，创新先进
Linux最新Kernel 6.1

2000+用户态软件
独立编译维护
不依赖第三方发行版

高效 经济
性能全方位升级

开放开源稳定易用
社区持续技术支持

下载地址	http://mirrors.opencloudos.org/opencloudos/9.0/isos/
代码仓库	https://gitee.com/src-opencloudos-rpms
缺陷跟踪	http://bugs.opencloudos.tech/

获取并体验 OpenCloudOS 9.0

OpenCloudOS 云原生实践

nettrace – 网络包全生命周期跟踪&诊断&监控工具集

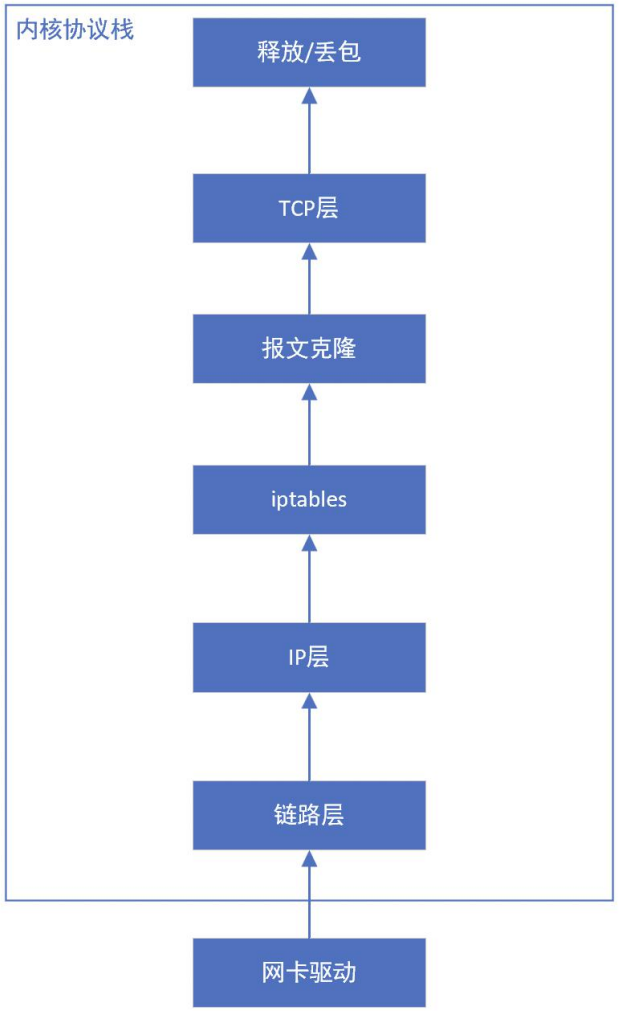
问题： 网络丢包/节点ping不通？

云原生场景下，网络环境越来越复杂，报文在节点上的处理路径也越来越长。当发生网络故障时，传统的网络定位工具（tcpdump、 ftrace、 kprobe、 dropwatch等）存在一定的短板，无法深入内核进行分析，使得问题定位困难部署、周期长。

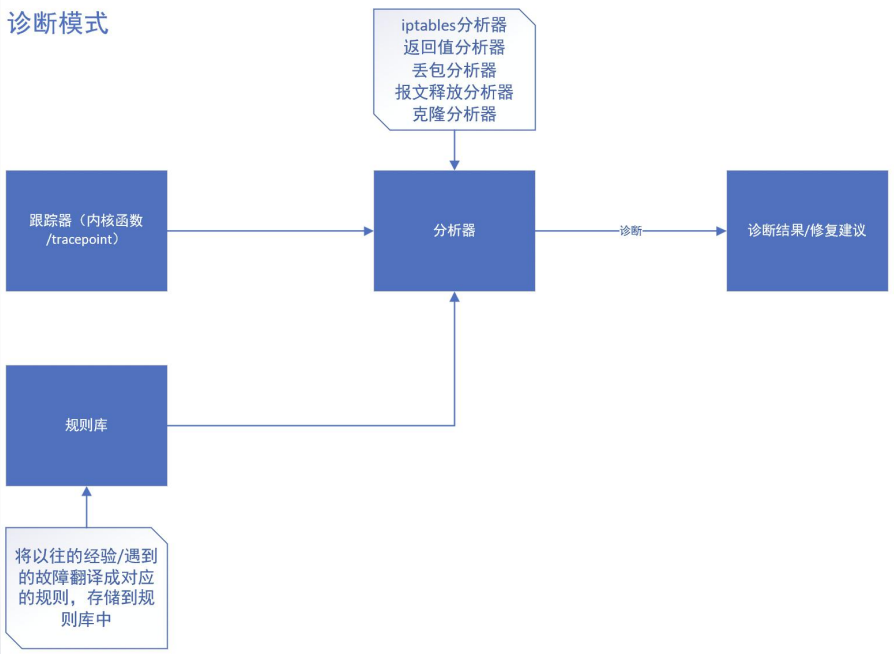
nettrace（自研）：基于eBPF实现的用于云原生场景的网络定位、诊断和监控工具：

- ✓ 报文生命周期跟踪，快速定位问题
- ✓ 主动故障诊断，一键解决网络故障
- ✓ 常态化网络异常监控，实时发现网络问题
- ✓ **集群报文染色，解决容器集群范围网络包跟踪难题**
- ✓ **新增丢包原因Feature，在Kernel社区提交近100个patch，上了lwn新闻**

生命周期跟踪示意图



nettrace – 网络包全生命周期跟踪&诊断&监控工具集



现象：节点ping不通

诊断命令：nettrace –p icmp --diag --saddr 192.168.122.8

诊断结果：iptables在filter表INPUT链中的防火墙规则导致

```
*************** ffff889fb6af8c00 **************
[474.482815] [__netif_receive_skb_core] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0
[474.482831] [br_nf_pre_routing   ] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0
[474.482836] [nf_hook_slow        ] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0 *iptables in HOOK: PRE_ROUTING*
[474.482840] [ipv4_conntrack_in   ] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0
[474.482848] [__nf_ct_refresh_acct] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0
[474.482877] [nf_hook_slow        ] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0 *ebtable in HOOK: PRE_ROUTING*
[474.482884] [__netif_receive_skb_core] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0
[474.482892] [ip_rcv              ] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0
[474.482895] [ip_rcv_core         ] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0
[474.482902] [nf_hook_slow        ] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0 *iptables in HOOK: PRE_ROUTING*
[474.482906] [ip_rcv_finish       ] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0
[474.482911] [ip_route_input_slow ] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0
[474.482916] [fib_validate_source ] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0
[474.482936] [ip_local_deliver    ] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0
[474.482939] [nf_hook_slow        ] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0 *iptables in HOOK: INPUT* *packet is dropped by netfilter (
[474.482942] [ipt_do_table        ] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0 *iptables table:filter, chain:INPUT* *packet is dropped*
[474.482952] [kfree_skb           ] ICMP: 192.168.122.8 -> 192.168.122.1 ping request, seq: 0 *packet is dropped by kernel*
---------------- ANALYSIS RESULT --------------------
[1] ERROR happens in nf_hook_slow(netfilter):
        packet is dropped by netfilter (NF_DROP)
    fix advice:
        check your netfilter rule

[2] ERROR happens in ipt_do_table(netfilter):
        packet is dropped
    fix advice:
        check your netfilter rule

[3] ERROR happens in kfree_skb(life):
        packet is dropped by kernel
    location:
        nf_hook_slow+0x96
    drop reason:
        NETFILTER_DROP
```

适用场景：

- 出现网络故障时，启动命令跟踪特定流（报文），进行诊断分析
- 使用情况：普通用户/专家/运维都在使用中

收益：提升工作效率

云原生SLI-Sevice Level Indicator

问题： 公有云顺*科技在TKE环境中业务偶现卡顿

- 容器内5-8s内日志无输出，业务莫名卡顿
- 随机出现，无复现规律，持续时间短(秒级)
- 资源无瓶颈，cpu/内存/io/网络 都正常无突高
- 监控无异常，秒级监控也束手无策

SLI的要求：

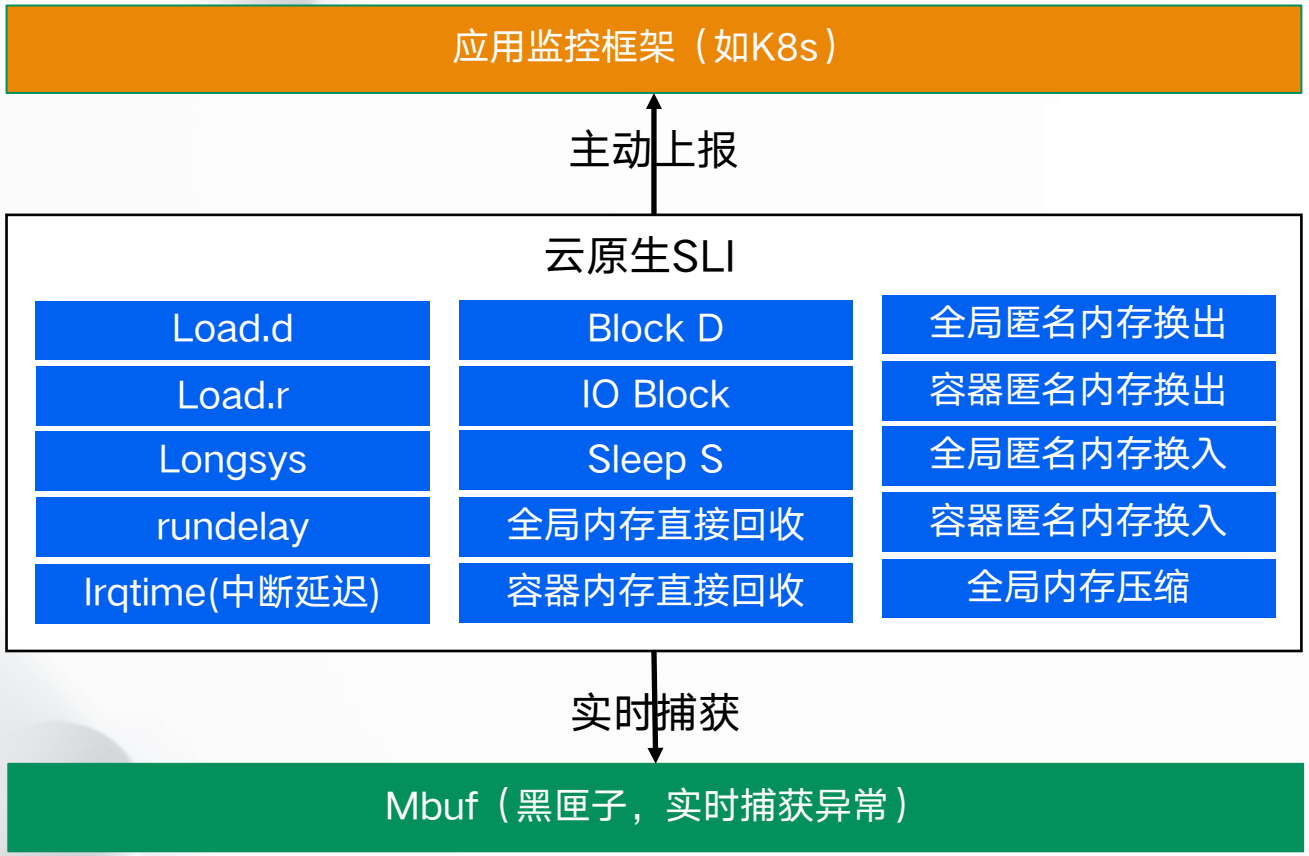
- 容器级别、专业、专用
- 常态化部署(低开销)
- 主动监控异常、主动记录
- 毫秒级响应

典型的业务随机抖动问题

行业难题！

SLI指标	分类	说明
Load.r	CPU	容器内处于R态的进程平均数量，判断overload
Load.d	CPU	容器内处于D态的进程平均数量，判断锁竞争或IO阻塞
longsys	CPU	内核态长延迟。系统调用、缺页、中断等原因可能导致
rundelay	CPU	调度延迟。判断CPU争抢
irqtime	CPU	中断延迟。判断中断带来的影响
Block D	CPU	D状态阻塞延迟。判断锁竞争或IO阻塞
IO Block	IO	IO阻塞延迟。判断IO问题
内存指标	内存	内存关键延迟。判断各种细致的内存问题

云原生SLI-Sevice Level Indicator



方案：

- SLI+Mbuf，常态化部署，捕获随机抖动

业务卡顿解决效果：

- 抓到关键第一现场，定位问题，优化内核后解决

收益：

- 提升解决问题效率 & 提供监控/调度关键数据

```
schedlat_wait_thr = 18446744073709551615 ms
schedlat_block_thr = 18446744073709551615 ms
schedlat_ioblock_thr = 16 ms
schedlat_sleep_thr = 18446744073709551615 ms
schedlat_rundelay_thr = 18446744073709551615 ms
schedlat_longsys_thr = 18446744073709551615 ms
```

```
891764923095:record reason:schedlat_ioblock comm:cc1 pid:17113 duration=20595081
891764924011:[<0>] rq_qos_wait+0xc2/0x160
891764925288:[<0>] wbt_wait+0x213/0x2c0
891764926136:[<0>] __rq_qos_throttle+0x25/0x40
891764927231:[<0>] blk_mq_make_request+0x102/0x5e0
891764928296:[<0>] generic_make_request+0x17e/0x340
891764929226:[<0>] submit_bio+0xaf/0x1a0
891764930524:[<0>] ext4_io_submit+0x4d/0x60
891764931770:[<0>] ext4_bio_write_page+0x192/0x510
891764932513:[<0>] mpage_submit_page+0x57/0x70
```

如意：在离线容器集群混部架构

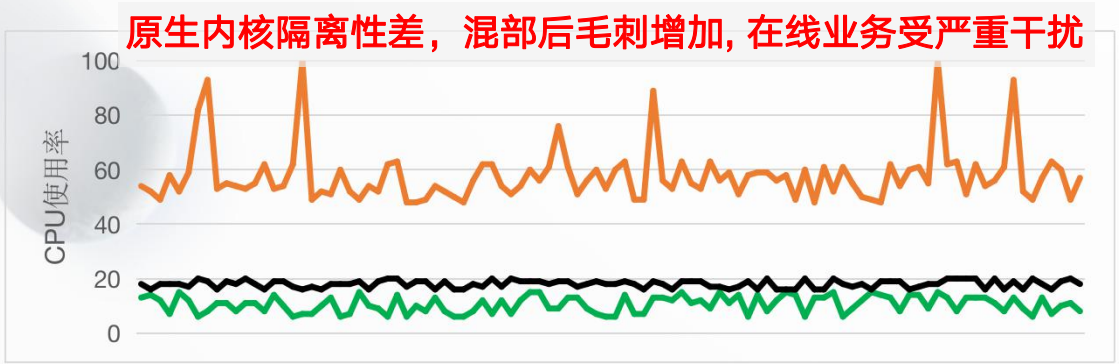
在离线业务混部背景

业界难题：IDC整体自研利用率低，CPU利用率：<15%

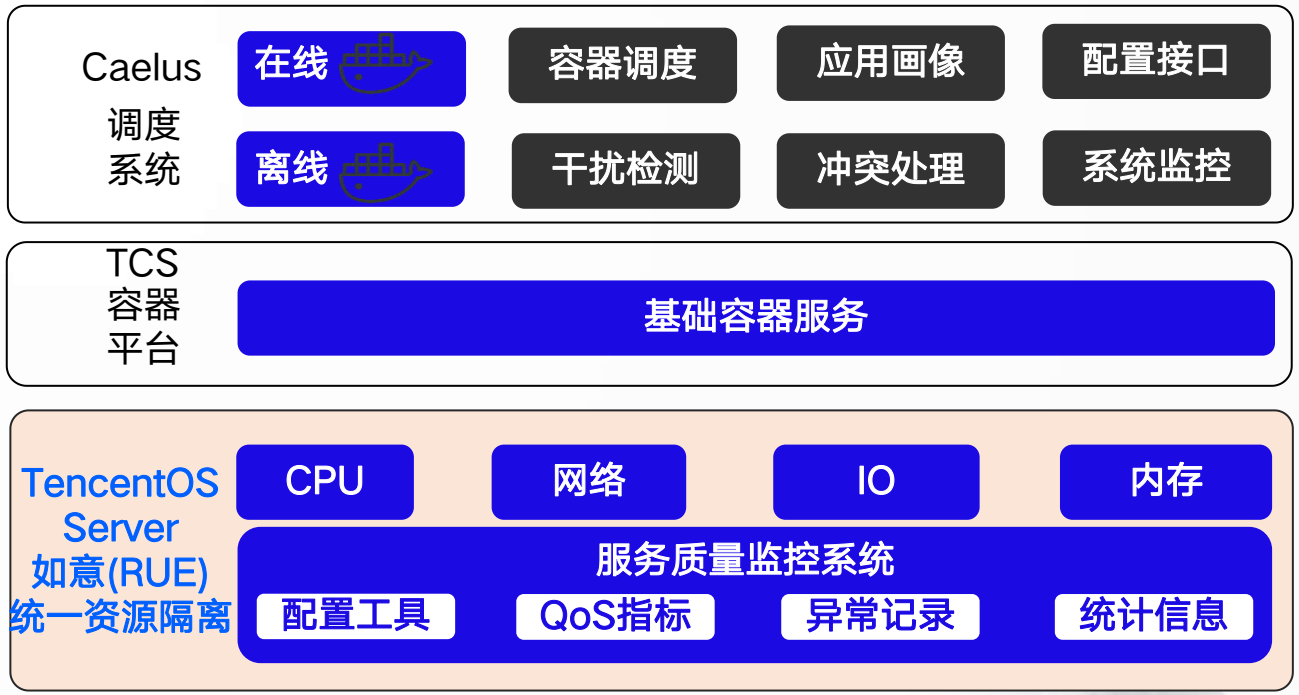
在线业务：常规业务

离线业务：TDW大数据(压缩和常规)、AI训练、广告转码

关键问题



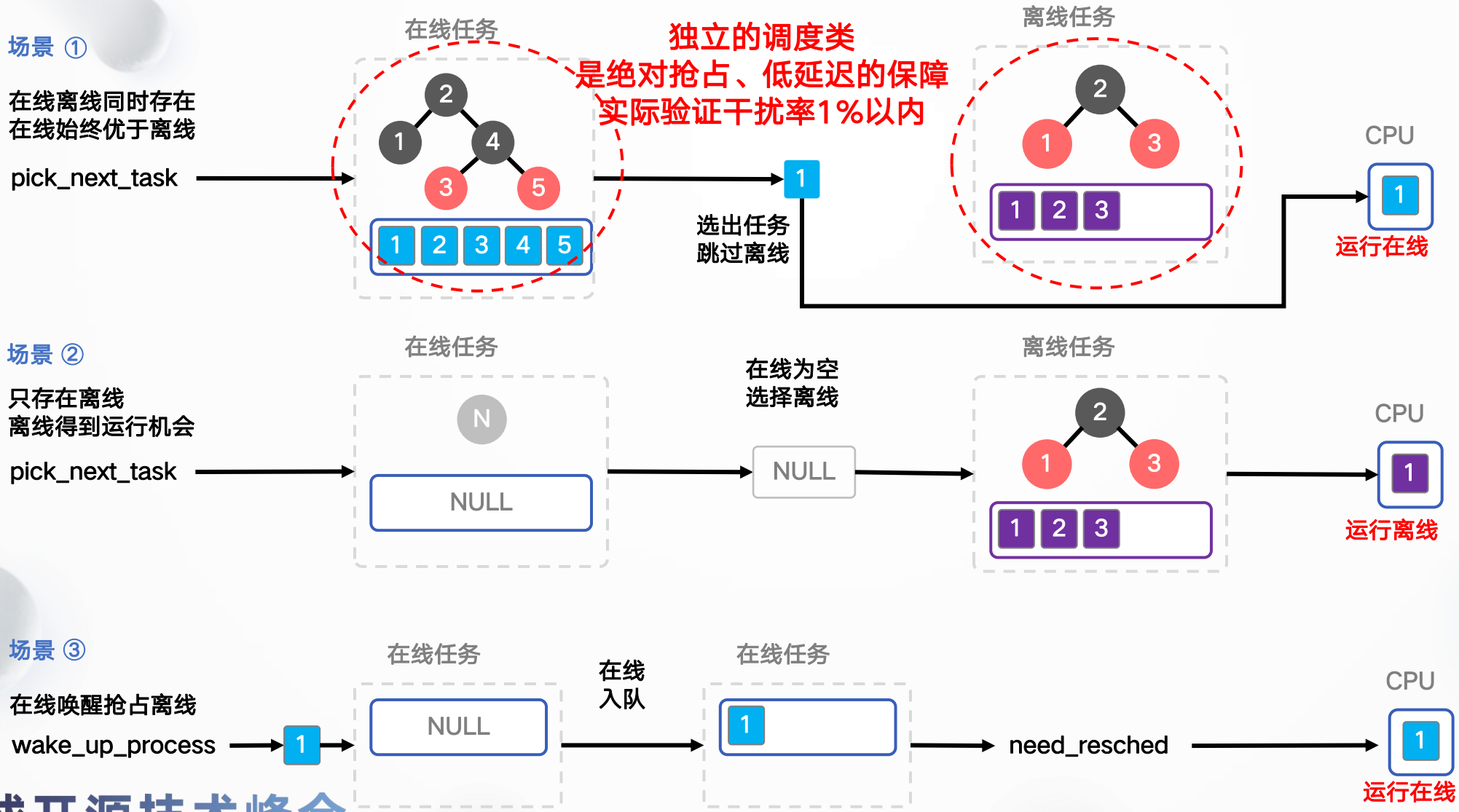
全场景混部技术(容器+物理机)



整体方案：底层资源隔离(RUE)+上层容器编排

- 上限高，适用性广，不挑业务
- OS内核层自动的容器隔离，冲突处理
- 相比社区，内核隔离能力全面增强

如意CPU QoS：绝对抢占(BT调度器)

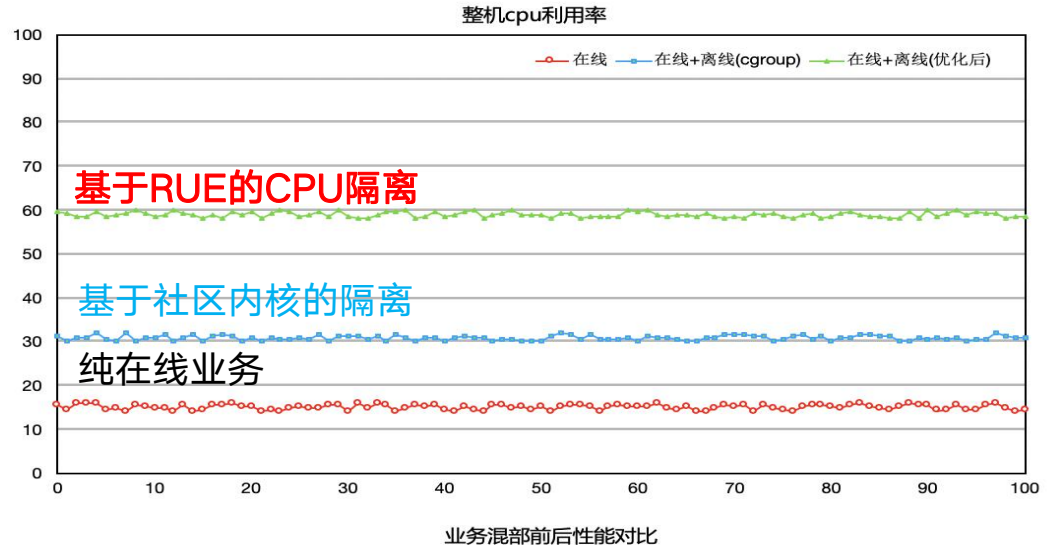


如意效果：资源利用率翻倍，成本降低50%

混部关键指标

- 混部大盘CPU利用率:<15%->30%，成本降低50%
- 干扰率：离线对在线的影响小，干扰率<1%
- 抖动：在线业务的io、网络带宽稳定，波动率<5%
- 覆盖规模超2000万核
- 样板集群CPU占用率:65%，行业标杆

CPU隔离效果对比



收益

自研业务

WXG、PCG、CDG大量业务

腾讯TencentOS荣获“OSCAR尖峰奖”，创新突破节省上亿度电资源

某头部互联网厂商

资源利用率提升100%

成本节省数百万/年

其它厂商

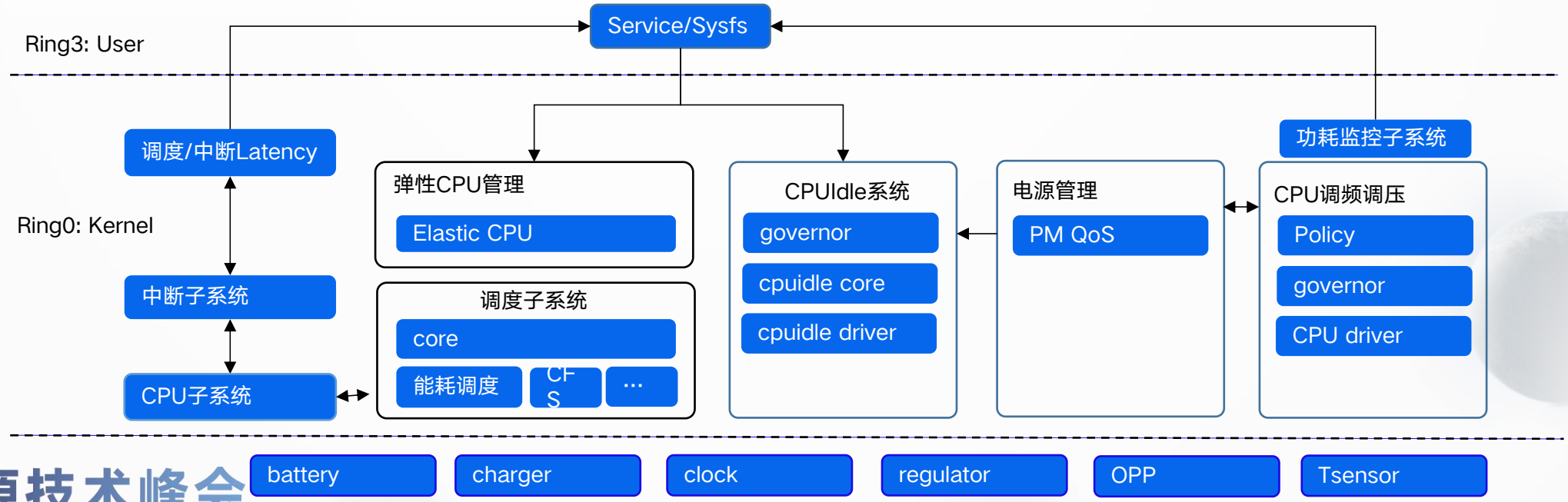
成本降低43%

绿色低碳的下一代云原生操作系统

响应国家“碳中和”战略：OpenCloudOS的“悟能”技术系统，在操作系统级别CPU弹性运行与深度睡眠，结合功耗监控和细粒度的电源管理，软硬件结合达到节能目的。

实测结果：减少5%-30%的能源消耗、小于0.1%的性能损耗

- 腾讯应用：服务器整机功耗省5-30%，腾讯数据中心整体能耗省6亿KWH/年，减少碳排放24万吨/年
- 宏观应用：可推广到所有IDC，减少所有IDC 5-30%碳排放，对国家整体碳中和战略有长远的重要意义。



全面优化：性能全面提升

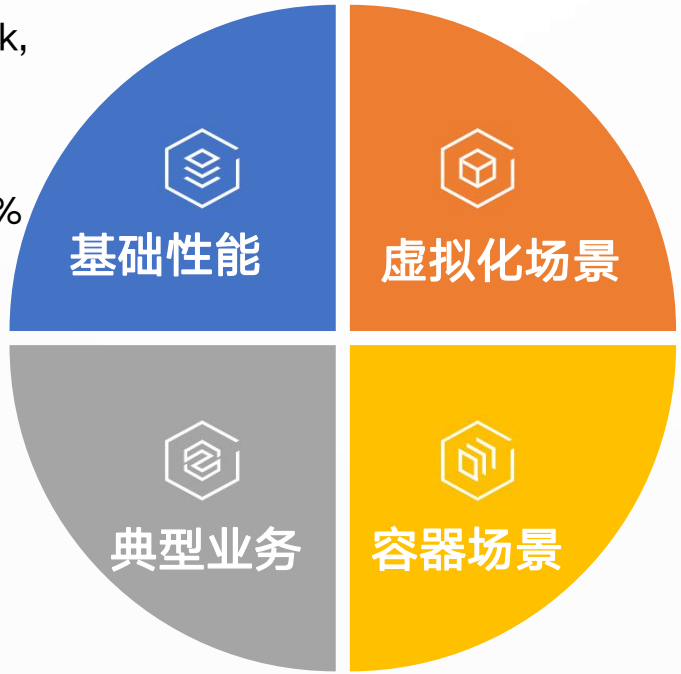
OpenCloudOS 针对海量业务场景进行深度的性能优化，相比Upstream原生内核有明显的提升

基础性能提升10-200%

- Spinlock优化：实现LLC-aware Spinlock，文件锁测试性能提升200%
- Unixbench：上下文切换提升10%
- Stress-ng：并发读取随机数性能提升28%
- 网络PPS：优于CentOS8 59%-65%
- 网络带宽：优于CentOS8 29%-38%

典型业务性能提升达150%

- Nginx：性能最大提升150%
- Redis：性能提升11%



虚拟化性能提升达150%

- PV IPI/PV TLB Shootdown：降低虚拟化开销提升达130%-150%
- Fastpath IPI：IPI和timer性能提升30%和16.5%
- 并行化的vMMU方案：虚拟机缺页处理时内存分配性能提升100%
- 8项技术获评KVM社区年度核心突破

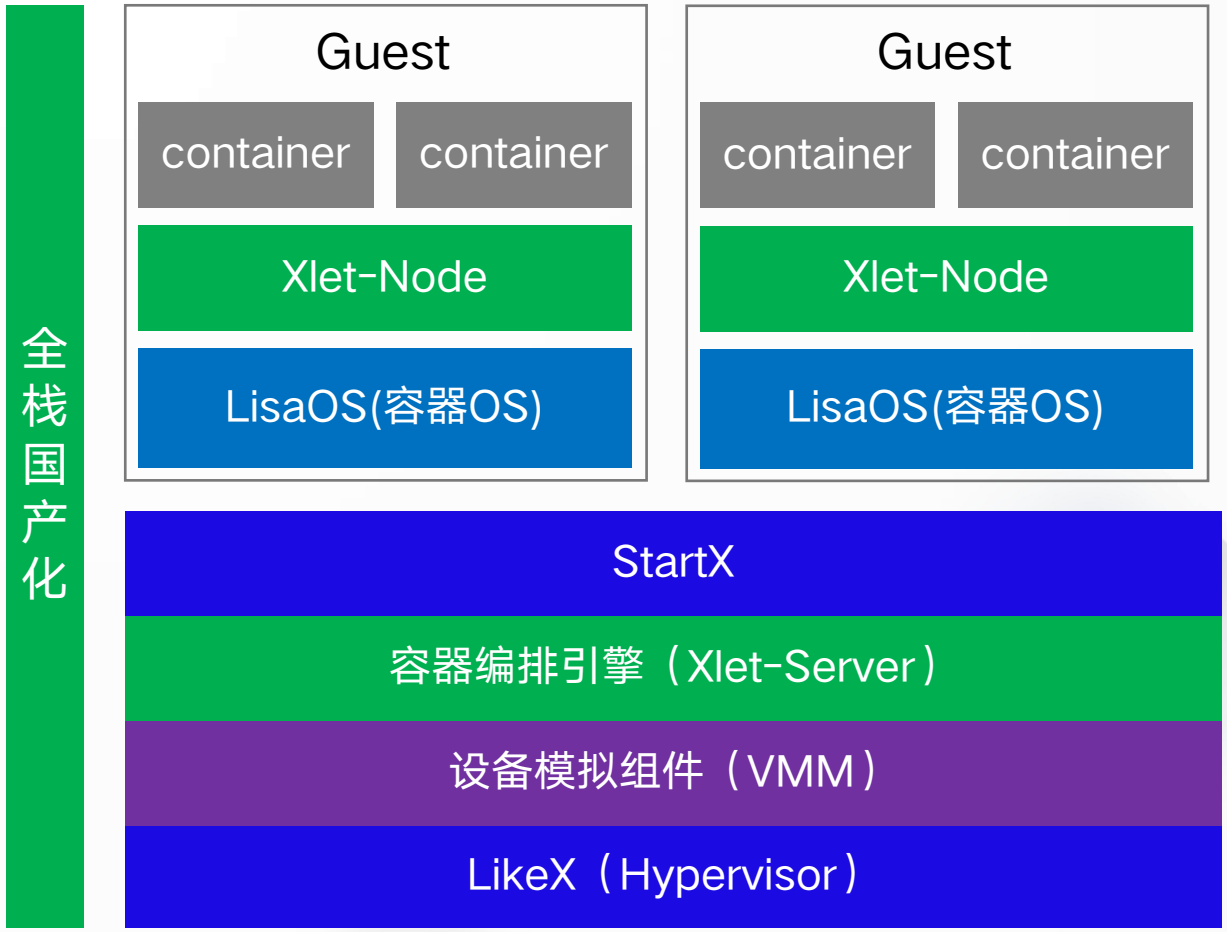
容器场景性能提升达187%

- LRU lock优化：多容器并发访问page cache性能提升187%
- Ebpf能力增强：完整支持Cillium，容器网络快速转发，性能提升50%+

自主可控的下一代云原生操作系统

下一代云原生操作系统以容器为中心，虚拟化技术主要用于容器隔离，能更好地支持Serverless/FaaS场景。
通过参与贡献国际顶级开源项目获得原创技术能力，从而实现全栈国产化、自主可控的下一代云原生操作系统。

特征	轻量化
	Safety（可靠性）
	Security（安全性）
解决痛点	解决版本碎片化的问题
	容器运行时与内核结合问题
	多租户安全隔离问题
优势	支撑业务敏捷发布
	支撑业务长时间稳定运行
	支撑业务快速弹性伸缩



OpenCloudOS的价值和可行性探讨
——欢迎加入 OpenCloudOS 社区

OpenCloudOS“千百双扶”计划

OpenCloudOS “千百双扶”计划由 OpenCloudOS 社区联合生态合作伙伴、国内 10 余家优质联盟与孵化器等单位、机构共同发起，旨在 CentOS 停服局势下扶持千百家中小企业，快速实现操作系统的平滑升级和迁移；同时也帮助其在关键技术领域获得趋势洞察、实现生态圈层内的资源合作与交流。

申请并加入的企业，可获得如下资源支持:

技术咨询和迁移支持服务

- **建立专属技术服务群**
- 1v1 专项迁移咨询及指导
- 提供迁移全流程技术支持
- “兼容性测试”评估认证

生态资源扶持

- 中小企业支持政策定期培训与解读
- “OpenCloudOS核心合作伙伴”认证
- 专项宣推（联动主流媒体宣发扩圈）
- 标杆或灯塔项目联合营销

联盟共赢体系

- OpenCloudOS 500+生态企业高端私享会
- OpenCloudOS 全年市场活动展示机会
- 100+ SIG（Special Interest Group）加入/发起权
- 营销/技术赋能 Workshop 优先参与机会

投资孵化加持

- 优先进入腾讯产品生态投资通道
- 优质产品有机会推荐上架腾讯云市场
- 优先享有与腾讯其他产品对接集成合作
- 优先获得全球优质孵化器、知名VC和高校合作

加入OpenCloudOS社区您可以获得：

- 自如应对CentOS停服风险
- 下一代云原生操作系统的社区技术保障
- 与11大行业500+生态伙伴合作的机会
- 与社区100个SIG一起发展开源操作系统生态
- 有机会参与“千百双扶”计划
- 触达10万+开发者

同时您有机会：

- 成为理事会成员，制定OpenCloudOS 社区的战略方向
- 成为TOC成员，对OpenCloudOS社区进行技术决策
- 成立SIG，在特定领域探索及贡献OpenCloudOS社区

加入社区官方交流群

添加社区小助手

GOTC

THANKS

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE