KYLINSOFT
麒麟软件

银河麒麟云底座操作系统 V10

产品白皮书

麒麟软件有限公司

2024 年 7 月

商标声明

KYLINSOFT 麒麟软件和其他麒麟商标均为麒麟软件有限公司的商标。本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受麒麟软件有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，麒麟软件有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容有可能变更，麒麟软件有限公司保留在没有任何通知或提示的情况下对内容进行修改的权利。除非另有约定，本文档仅作为使用指导，并不确保手册内容完全没有错误。本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

目 录

1. 概述

1.1. 背景

操作系统(Operating System，简称 OS)是承载各种信息设备和软件应用运行的基础平台，是配置在计算机硬件上的第一层软件。它是一组控制和管理计算机硬件和软件资源，合理地对各类作业进行调度以及方便用户的程序集合。操作系统是用来对整个计算机系统的硬件和软件资源进行配置和管理，控制所有应用程序运行，提供人机交互的平台，是计算机工作的灵魂，CPU、数据库、办公软件、中间件、应用软件等需要与操作系统深度适配。现如今，操作系统发展迅速，逐步进入了社会生活的各个方面，涉及大型计算机、个人计算机、移动便携设备、其他自动化设备等各个层次的应用领域。

1.2. 关于麒麟

麒麟软件有限公司（简称“麒麟软件”）是中国电子信息产业集团有限公司（CEC）旗下科技企业，2019 年 12 月由天津麒麟信息技术有限公司和中标软件有限公司强强整合而成。

麒麟软件以安全可信操作系统技术为核心，面向通用和专用领域打造安全创新操作系统产品，现已形成桌面操作系统、服务器操作系统、万物智联操作系统、工业操作系统、智算操作系统产品等为代表的产品线，达到国内最高的安全等级，全面支持飞腾、鲲鹏、龙芯等国产主流 CPU，在系统安全、稳定可靠、好用易用和整体性能等方面具有领先优势，并为党政、行业信息化及国家重大工程建设提供安全可信的操作系统支撑。根据赛迪顾问统计，麒麟软件旗下操作系统产品连续 13 年位列中国 Linux 市场占有率第一名。

麒麟软件注重核心技术创新，2018 年荣获“国家科技进步一等奖”，2020 年发布的银河麒麟操作系统 V10 被国资委评为“2020 年度央企十大国之重器”，相关新闻入选中央广播电视总台“2020 年度国内十大科技新闻”，2021 年麒麟操作系统入选央视《信物百年》纪录片，2022 年入选工信部“2022 年国家技术创新示范企业”，2023 年发布的“开放麒麟 1.0”被国资委评为“2023 年度央企十大国之重器”，麒麟软件有限公司技术中心被多部委共同认定为“国家企业技术中心分中心”，入选国资委“创建世界一流专精特新示范企业”,2024 年麒麟操作系统被中国国家博物馆收藏，这是中国国家博物馆收藏的第一款国产操作系统。麒麟软件荣获“中国电力科学技术进步奖一等奖”、“水力发电科学技术奖一等奖”、“中国版权金奖·推广运用奖”等国家级、省部级和行业奖项 600 余个，并被授予“国家规划布局内重点软件企业”、“国家高技术产业化示范工程”、“科改示范行动企业”、“国有重点企业管理标杆创建行动标杆企业”等称号。通过 CMMI 5 级评估，现有博士后工作站、省部级企业技术中心、省部级基础软件工程中心等，先后申请专利 891 项，其中授权专利 408 项，登记软件著作权 647 项，主持和参与起草国家、行业、联盟技术标准 70 余项，被国家知识产权局成功认定为“国家知识产权优势企业”。

麒麟软件在北京、天津、上海、长沙、广州、深圳、太原、郑州、武汉、南京、南昌、济南、南宁、成都、沈阳、厦门等地设有分支机构，服务网点遍布全国 31 个省会城市和 2 个计划单列市。

麒麟软件高度重视生态体系建设，与众多软硬件厂商、集成商建立长期合作伙伴关系,建设完整的自主创新生态链,为国家网信领域安全创新提供有力支撑。截至 2024 年 4 月 30 日，麒麟软件已与 23100 多家厂商建立合作，硬件适配数超 71 万项，软

件适配数超 378 万项，总量超过 449 万项，生态适配官网累计注册用户数超 6.6 万人。

麒麟软件积极贯彻人才是第一资源的理念，以麒麟软件教育发展中心为组织平台，联合政产学研各方力量，探索中国特色的网信人才培养模式,目前已形成了源自麒麟操作系统的“5 序”课程体系、教材体系、认证体系、师资体系、平台体系，并与工信部教育与考试中心联合推出“百城百万”操作系统培训专项行动，持续为我国培养各类操作系统专业人才。

在开源建设方面，成立桌面操作系统根社区 openKylin，旨在以“共创”为核心、以“开源聚力、共创未来”为社区理念，在开源、自愿、平等、协作的基础上，通过开源、开放的方式与企业构建合作伙伴生态体系，共同打造桌面操作系统顶级社区，推动 Linux 开源技术及其软硬件生态繁荣发展。截至 2024 年 4 月 30 日，openKylin 社区用户数量超过 110 万，社区会员突破 450 家，开发者数量超 6200 人，创建 103 个 SIG 组。从 2022 年开始，openKylin 连续两年获评中国信通院“先进级可信开源社区”。此外，麒麟软件正式成为开放原子开源基金会白金捐赠人；作为 openEuler 开源社区发起者，以 Maintainer 身份承担 80 个项目，除华为公司外贡献第一；在 OpenStack 社区贡献位列国内第一、全球第三。

2. 产品介绍

2.1. 产品简介

银河麒麟云底座操作系统 V10 是麒麟软件面向中小云厂商和央国企等云计算场景，针对云宿主机关键诉求，依托服务器操作系统产品重要成果，整合云业务专业能力，将计算底座与上层业务解耦，融合最新的云和容器的开源技术，打造的开放创新、能力先进、自主合规、安全无忧、技术兜底的云底座操作系统。

银河麒麟云底座操作系统 V10 版本提供长达 6 年维护支持，包括 2 年标准服务期、2 年扩展服务期和 2 年最终服务期。支持操作系统的小版本升级，提供软件包在线升级服务。

标准服务期：提供对软硬件新功能和新特性的支持，缺陷和安全漏洞修复，以及技术支持。

扩展服务期：缺陷和安全漏洞修复，以及技术支持。

最终服务期：关键安全漏洞修复，以及最后阶段技术支持。

2.2. 产品特性与优势

银河麒麟云底座操作系统 V10 面向云计算 IaaS 下虚拟机、容器计算节点为主的操作系统技术形态，其主要的产品特性和优势包括：

- 可完全替换 CentOS：

基于 openEuler 生态，南北向主流软硬件支持，并针对云厂商对源码深度定制的需求而完全开源，不被锁定。

- 云底座优选体验：

面对云原生场景和虚拟化场景，在 5.10 内核基础上，针对混部技术进行了深度优化，充分发挥包括绑核动态亲和性控制、抢占多优先级控制、SMT 优先级防反转、IO QoS 控制等功能，优化动态优先级切换能力，使能 eBPF 能力，提供面对软硬一体化场景的 DPU 无感卸载功能，支持鲲鹏 CPU 优化特性，以及通用 vDPA 驱动支持统一智能网卡 virtio 设备接口规范。

新版本内核在特定场景下的系统性能、故障切换和资源利用率上获得了显著提升。通过并行化 iSCSI 链路切换，大幅缩短故障切换时间，减少业务中断；内核大块 IO 限速优化，实现了更加精细的 I/O 速率管理，确保了数据流动的平顺高效；内存动态隔离与释放策略，提供了灵活性更强的资源管控方案；支持安全迁移内存，增强系统稳定性；核隔离增强减少噪声干扰，提升业务性能。这些优化不仅提高系统性能，还增强稳定性和安全性，为开发者提供高效开发环境。

除此之外，该版本支持高性能用户态网络协议栈，基于 dpdk 和 lwip 提供一套兼顾高性能与通用性的 4 层网络协议栈运行库，提供兼容 POSIX API 的接口，确保应用程序无需进行改动即可直接享受到更高层次的网络性能体验。

- 虚拟化和容器技术增强：

云底座操作系统的核心功能组件现已支持多租户混部能力、可信引导能力、国密虚拟化应用、主流容器管理技术。具体而言，产品囊括了主流虚拟化 hypervisor qemu 和 libvirt 管理引擎、虚拟化镜像工具、虚拟设备 vTPM、edk2 引导引擎、docker/containerd/podman/crio 容器引擎和运行时、 kata container 3.0、虚拟交换 openvswitch、在离线混合部署控制引擎等，支持精准辨识虚拟机内存故障，针对大内存虚拟机的热迁移性能进行了深度优化，从而形成了全方位的云场景服务框架。

- 容器云&虚拟化编排优化：

基础仓库中默认提供 kubernetes 和 openstack 主流使用版本，额外提供更多的版本选择。k8s 包括 1.24、1.25、1.26 版本。openstack 包括 Train 和 Walley 版本。同时，面对国产 CPU 和当前主流资源调度场景，提供基于 S2500 的绑核调度优化，以及虚拟机动态绑核策略。

- 硬件生态扩展：

支持 CPU 类型包括第四代英特尔® 至强® 可扩展处理器、第五代英特尔® 至强® 处理器、英特尔® 至强® 6 处理器、海光 C86-3G、海光 C86-4G、鲲鹏 920、鲲鹏 920 V200、飞腾 S2500、飞腾 2000+等。支持主流 Nvidia CX 系列云场景网络卸载能力的智能网卡、BlueField 2 DPU 板卡等。

- 基础软件优化：

继承通用 OS 和 OpenEuler 2203 LTS 的基础功能，裁剪非云场景的功能。在 ISO 镜像和仓库上做区分。达成云底座操作系统的软件功能最小集，强化云场景特点，分离不必要的场景依赖和功能，减轻运维成本，降低安全漏洞潜在风险。保留基础功能包括：服务管理、网络管理、基础工具、文件和存储管理、安装引导组件、PAM 和 openssl 管理、通用安全机制、开发者编程语言和库、包管理工具、监控管理、日志管理等。此外，在调优配置上新增分布式存储 Ceph OSD 节点典型优化方案。

- 安全可信：

统一技术方案，系统提供安全套件基础设施，为安全套件厂商提供 SDK，提供统一的使用规范，保障系统的稳定运行。系统提供三员分立，按职能分割和最小授权原则，分别授予它们各自为完成所承担任务所需的最小权限，并形成相互制约关系。提供执行管控、进程防杀、模块防卸载功能，保障系统的安全可信运行。

➢ 特选运维和系统工具：

除去基础系统的运维工具外，扩展系统探测 eBPF 工具，网络探针 nettrace 等便于对在线业务分析的手段，引入故障分析助手和一键式性能分析调优工具，旨在大幅简化操作系统运维过程中的故障排查与性能瓶颈定位工作，优化工作流程，进而提升系统稳定性和性能水平，确保用户专注于核心业务。

除此之外，产品对用户提供二次系统集成工具，并引入了 NestOS 不可变操作系统概念。NestOS 不可变操作系统采用 rpm-ostree 封装技术，确保系统核心组件处于只读状态，从根本上增强系统稳定性与安全性；通过优化构建工具 nestos-assembler 与配置文件 nestos-config，为云底座提供了一套完整且易于管理的不可变操作系统解决方案，确保每次系统更新或部署都能保持一致性和可靠性，致力于为用户提供更加安全、高效、稳定的云平台体验。

2.3. 主要功能及特性

2.3.1. 在离线资源混部功能

为了解决大规模集群中资源利用率低效问题，在线业务和离线业务混合部署逐步成为提升资源利用率的主要尝试方式。但在离线混合部署之间会有天然的性能干扰，这使得性能折损变成是一种对提升资源利用率的妥协。

在 K8S 场景下，根据内核提供的接口，rubik 解决方案提供了完整的 Qos 控制机制来实现在线服务 QOS 保障。

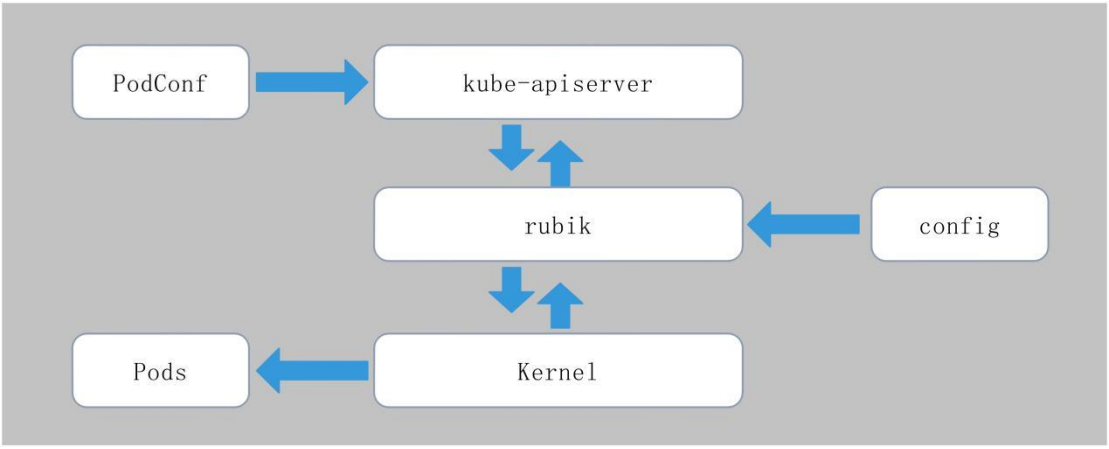


图 1 在离线混布功能技术方案

- rubik 通过监听 kube-apiserver 创建和更新 Pod 的事件来触发资源调配行为；
- rubik 读取用户创建的 Pod 配置，从 annotation 中读取相应的 Pod 优先级信息；
- rubik 读取全局配置来实现部分功能的开关以及参数管理；
- rubik 操作 Pod 的 cgroup 接口来对 Pod 的 Qos 进行配置；
- rubik 动态监控机器状态，可动态控制离线任务的内存水位线。

2.3.2. 腾云 S2500 绑核调度策略

openstack-nova-scheduler 和 kubelet 均有自身的绑核调度和决策方法，已经针对 CPU 硬件拓扑特点实现了，NUMA 亲和，超线程，绑核或不绑核等调度方式。通过策略的决策，使虚拟机的计算性能能够达到更优的效果。

腾云 S2500 CPU 采用双 socket，多 NUMA 的拓扑设计。由于每 4 个 Core 共享 1 个 L2。这种硬件特征导致 openstack-nova-scheduler 和 kubelet 在默认调度行为进行绑核分配时，无法达到原有性能提升目的。

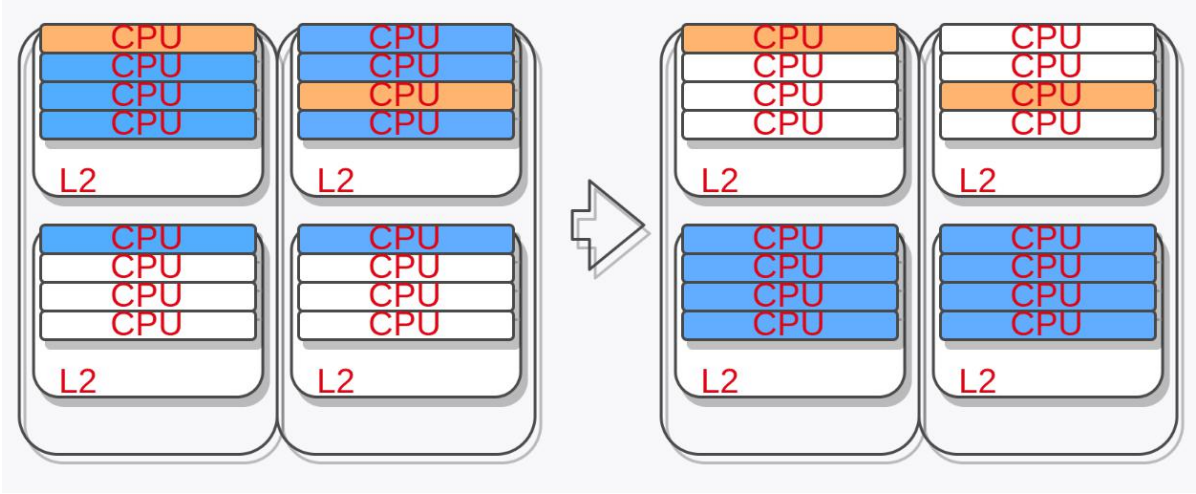


图 2 优化前后绑核对比示意

对于 kubelet，如果 NUMA 上已经分配 CPU 时(图 2 黄色部分)，新分配的 CPU(上图蓝色部分)大概率会和原有 CPU 形成 L2 争抢，从而无法达到最优性能，因此对该绑核逻辑改造以保障 L2 独占。

对于 openstack-nova-scheduler，面对（如图 3 S2500 CPU 拓扑示意）S2500 的拓扑特征会导致在默认调度行为进行绑核分配时，虚拟机 CPU 大概率会形成 L2 争抢，而无法达到默认调度分配预期。

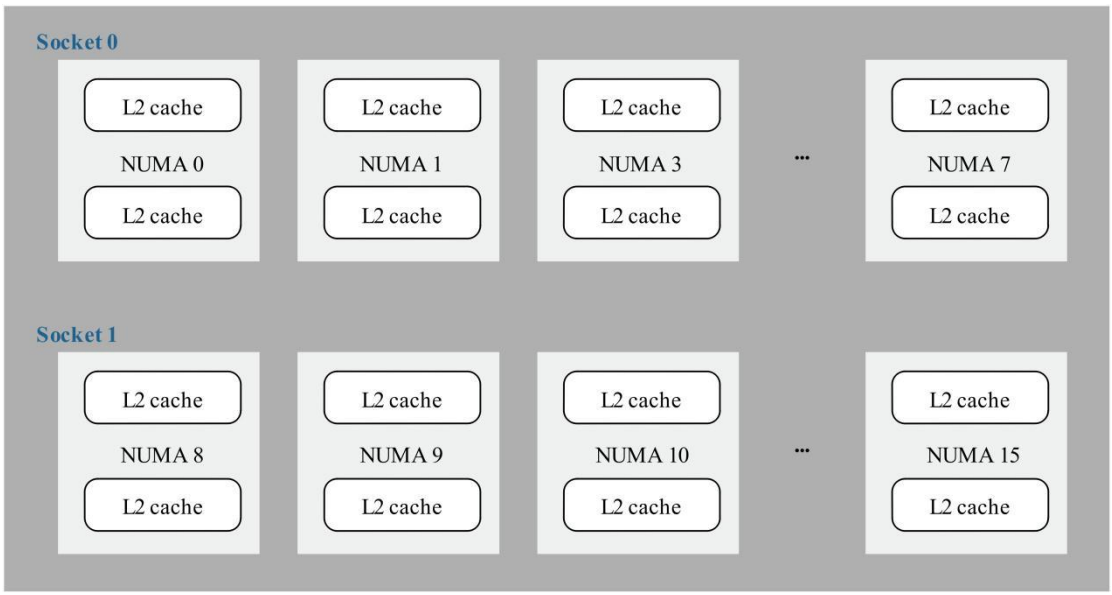


图 3 S2500 CPU 拓扑示意

为了在腾云 S2500 上满足云场景多租户高性能需求，基于腾云 S2500 双 socket，多 numa 以及 L2 缓存的特征，在 nova-scheuler 绑核调度策略上采用尽量不跨路，不抢占 L2，分散 numa 的方案。在 kubelet 上与 nova-scheduler 类似，不同的是从 NUMA 节点上选取 CPU 进行绑定时，先筛选出 L2 空闲的 CPU 优先进行选择，若未分配够，则继续从其他 CPU 进行选择。

实现效果

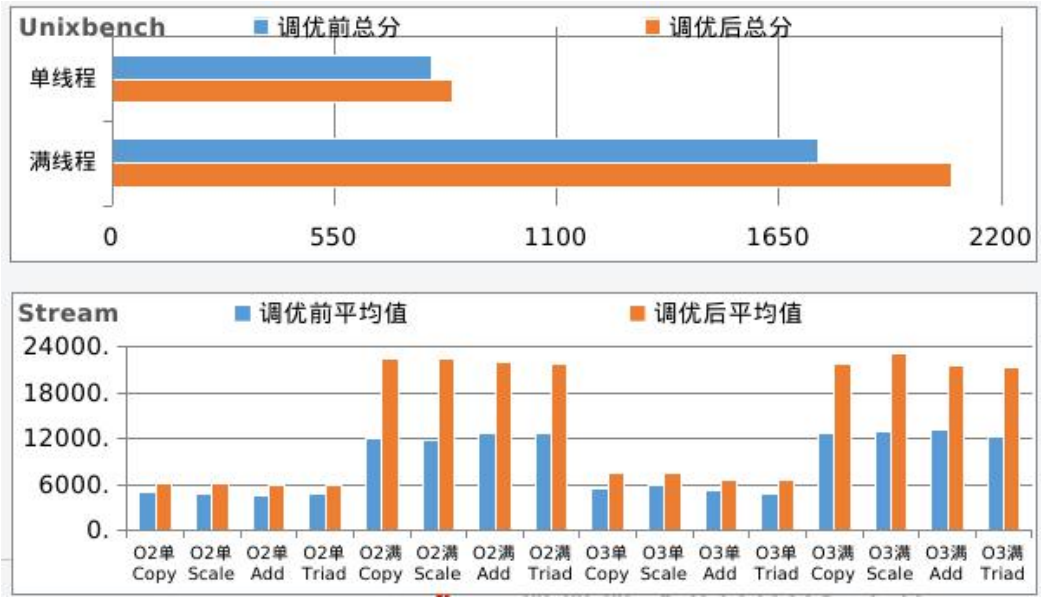


图 4 腾云 S2500 绑核调度调优效果

以 OpenStack（T 版），4C8G 虚拟机为例。使用 stream 工具测试优化后的被调度实例的内存带宽效果，可以看出，内存带宽在单线程和满线程情况下提升均较为明显。在单线程测试下，可以提升 20%-35%，在满线程测试下，可以提升 70%-90%。使用 unixbench 测试综合性能，调优后和调优前对比（如图 5），整体性能更优。单线程提升 6%，多线程提升 18%。

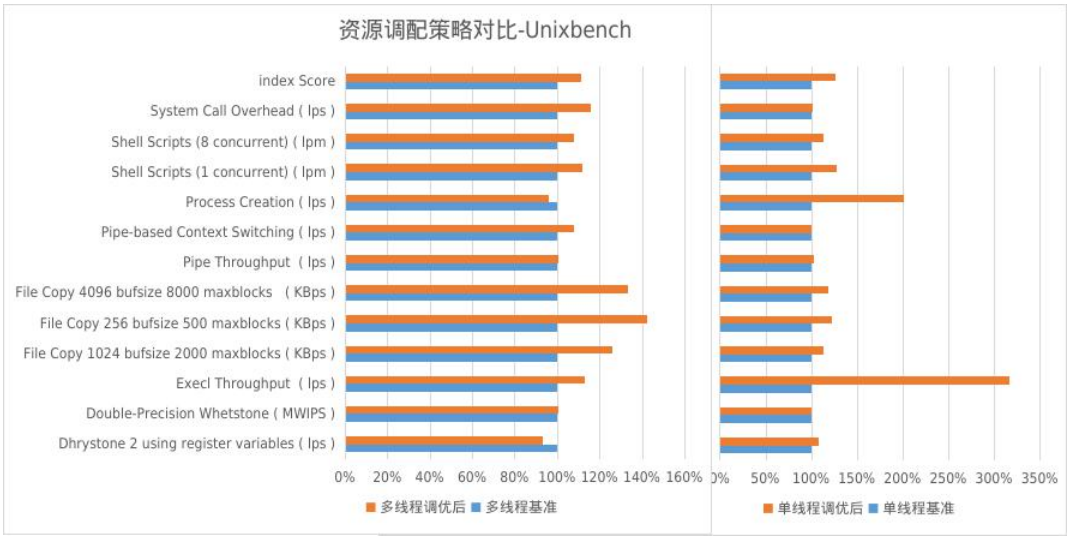


图 5 测试 Unixbench 指标性能对比图

容器场景下,以 4C4G 为例,在已有 CPU 资源被其他 POD 占用的情况下部署新 Pod,测试 Unixbench 指标性能单线程提升 26%，多线程提升 11%；

2.3.3. 鲲鹏 CPU Cluster-Aware 调度感知优化

鲲鹏处理器微架构拓扑包含 Cluster 结构层级,每个 Cluster 含有 4-8 个处理器核心，Cluster 内的处理器核心直连到同一个 L3 缓存片段。通过在内核中支持对于 Cluster 拓扑层级的感知，在 NUMA 调度域的之上构建 Cluster 调度域，优先将数据相关进程调度在同一 Cluster 的处理器核心上，从而提高数据相关进程之间的数据交换速度。

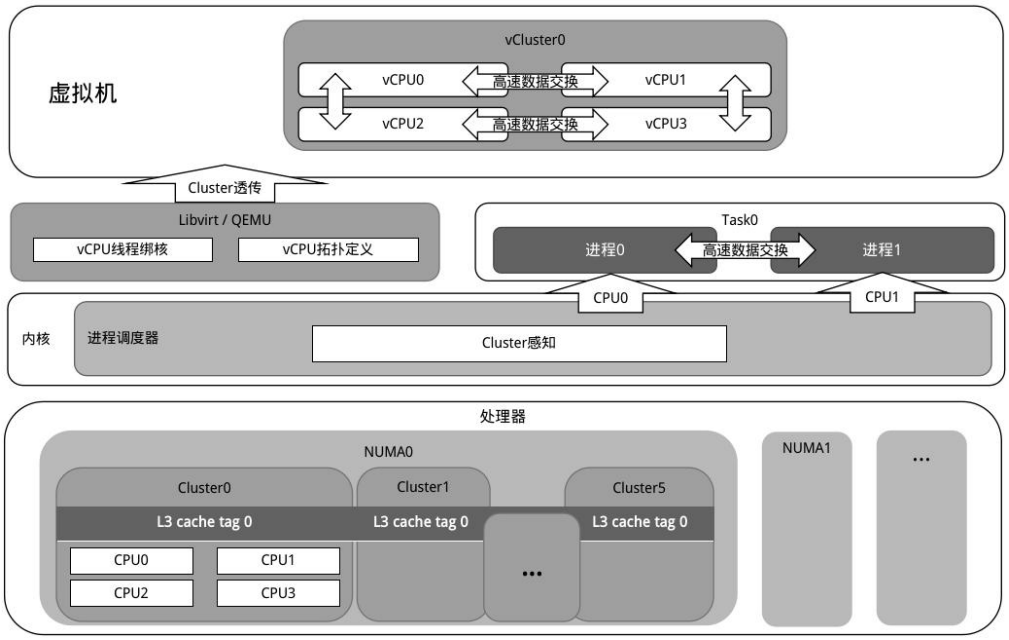


图 6 鲲鹏处理器微架构拓扑

对于多线程协作型任务，支持 Cluster-Aware 感知功能能够将共享数据的线程优先调度到同一 cluster 的核心上，提高线程之间的数据交换速度，从而优化任务执行速度。

在虚拟化场景下，支持将处理器 Cluster 拓扑结构透传到虚拟机中，在虚拟机 GuestOS 上实现 Cluster-Aware 调度（GuestOS 内核需支持 Cluster-Aware 调度）。

实现 Cluster-Aware 调度后，数据相关的不同进程会被优先调度到同一个 Cluster 内的处理器核心上，由于相同 Cluster 的核心共享同一个 L3 缓存片段，进程间可以达到更快的数据交换速度。

根据基准测试结果，启用 Cluster-Aware 调度功能后，多线程 Stream 测试结果提升明显，提升幅度约为 10%-15%；UnixBench 的进程管道相关测试项测试结果提升明显，提升幅度约为 10%-25%。

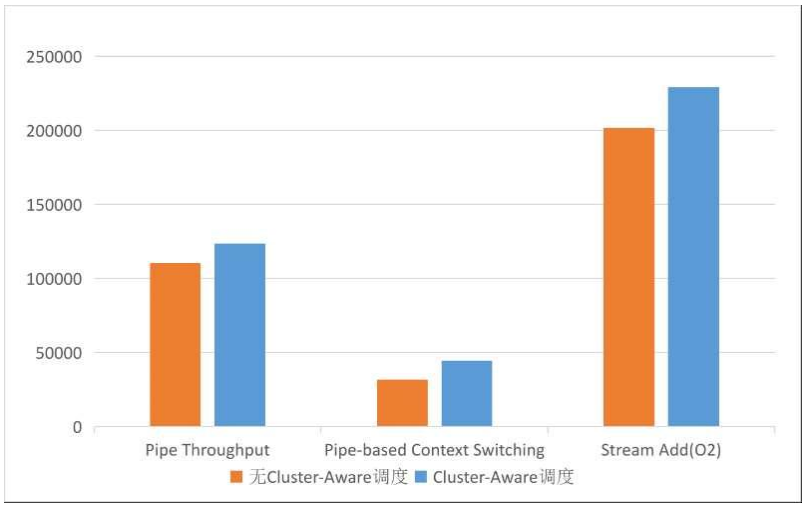


图 7 Cluster-Aware 调度功能优化对比图

2.3.4. 全面支持原生 Cilium 网络增强特性

Cilium 是一种应对现代微服务架构中动态性和安全挑战的解决方案，利用 LinuxBPF 技术，利用 LinuxBPF，Cilium 保留了透明地插入安全可视性+强制执行的能力，但这种方式基于 service/pod/container 标识（与传统系统中的 IP 地址识别相反），并且可以在应用层（例如 HTTP）进行过滤。因此，通过将安全性与寻址分离，Cilium 不仅可以在高度动态的环境中应用安全策略，而且除了提供传统的第 3 层和第 4 层分割之外，还可以通过在 HTTP 层操作来提供更强的安全隔离。这种方法克服了传统网络安全方法在微服务环境中的限制，如 IP 地址变动和端口复用，同时保持高度可扩展性，即使在大规模环境下也能够有效运行。

Cilium 依赖于内核的 eBPF 功能以及与 eBPF 集成的不同子系统。较高版本的内核提供更完整的 eBPF 功能支持，银河麒麟高级操作系统 V10(Host 版)内核版本为 5.10x，可以支持 Cilium 的多数高级功能：

- Bandwidth Manager：通过使用 EDT 和 eBPF 配置和优化 TCP 和 UDP 工作负载，实现对 Pod 的有效速率限制，以及支持 BBR 拥塞控制。

- Egress Gateway：出口网关是管理和控制内部网络到外部网络的出口流量的网络架构模式。
- Vxlan Tunnel Endpoint (VTEP) Integration：允许第三方 VTEP 设备通过 VXLAN 与 Cilium 管理的 Pod 直接进行通信。
- Wireguard Transparent Encryption：使用 WireGuard 实现端到端流量的透明加密；
- Full Support for Session Affinity：来自同一 Pod 或主机的连接始终选择相同的服务终端，以实现流量的负载均衡和会话保持；
- BPF-Based Proxy Redirection：通过重定向网络流量到代理（如 Envoy）来实现高性能的流量转发和处理；
- Socket-Level LB bypass in pod netns：允许在 Pod 命名空间内绕过套接字级别负载均衡器，使用原始的 ClusterIP 地址进行自定义重定向和操作，以适应特定的网络环境和需求；
- BPF-based host routing：基于 eBPF 技术实现的主机路由功能，能够绕过 iptables 和上层主机堆栈，提供更快的网络命名空间切换和较低的开销。

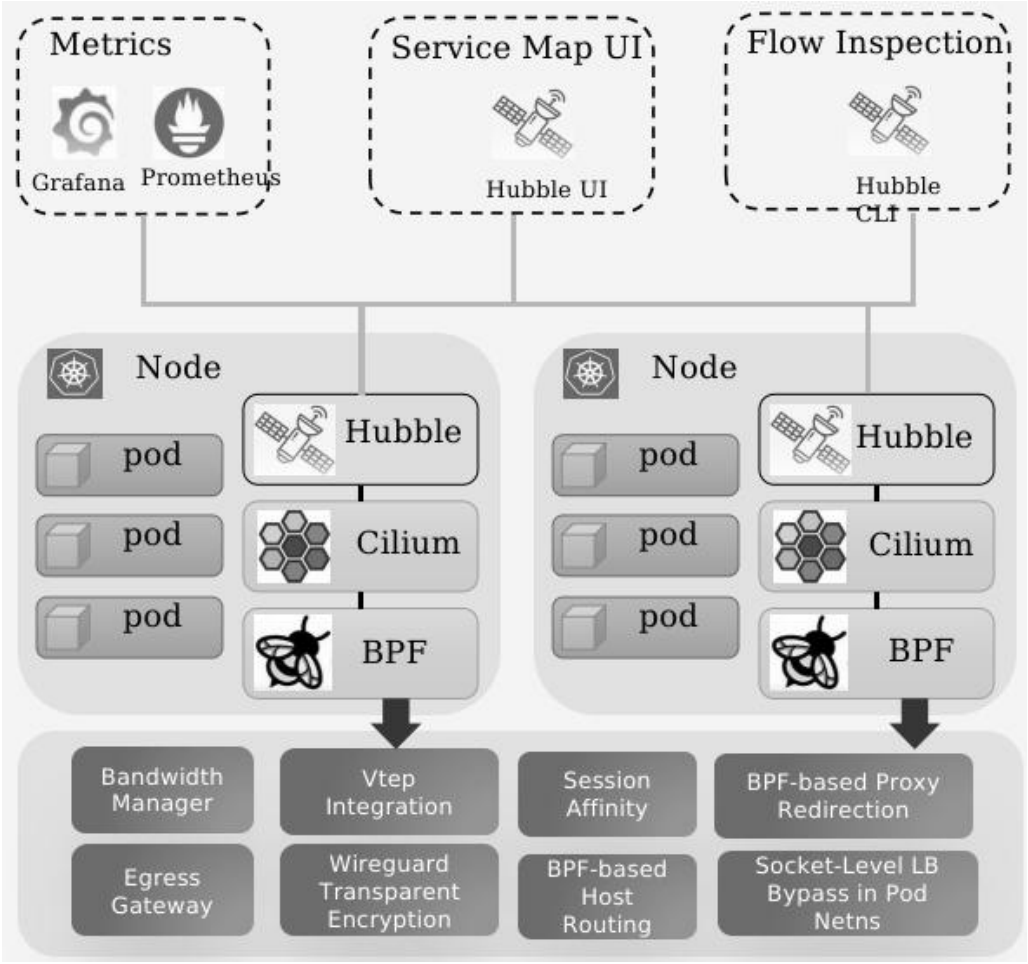


图 9　cilium 高级特性技术架构

Cilium 主要使用场景是在 Kubernetes 中，除了作为 KubernetesCNI 之外，还有很多其他功能，包括不限于：多集群网络、负载均衡（完全替代 Kube-proxy）和 ServiceMesh。

2.3.5. 支持低延迟高性能网络协议 Gazelle

Gazelle 是一款高性能用户态协议栈，兼顾高性能与通用性。它提供标准 POSIX API，应用程序可通过 LD_RELOAD 方式加载 gazelle，从而在无需代码修改的情况下，享受显著提升的网络性能。

Gazelle 基于 DPDK 在用户态直接读写网卡报文，共享大页内存传递报文，并轮询

模式收报，有效规避了中断与上下文切换带来的开销，确保了数据处理的连续性和高速度。

Gazelle 使用轻量级 LwIP 协议栈,将从 DPDK 接收到的报文直接在用户态进行拆解组装转发，以配合 DPDK 将报文收发处理全部在用户态完成，实现了数据包收发处理的全用户态闭环，显著提升了应用网络 I/O 的吞吐能力。

Gazelle 不仅减少了传统内核空间与用户空间数据交换的延迟，还大幅提高了数据处理的并发性，为构建高吞吐、低延迟的云原生应用和服务奠定了坚实的技术基石。

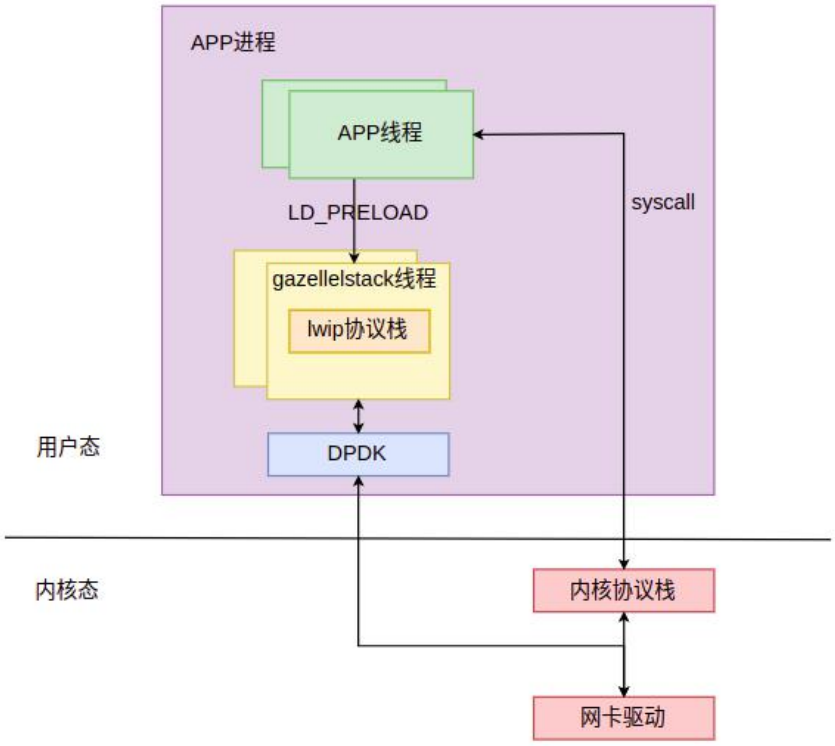


图 10 gazelle 技术架构图

2.3.6. ARM64 下 OSQ 互斥锁优化

多线程锁争抢一直是常见问题，在 arm64 下引入 MCS 锁机制。主要解决的问题是：在多 CPU 系统中，每当一个 spinlock 的值出现变化时，所有试图获取这个 spinlock 的

CPU 都需要读取内存，刷新自己对应的 cache line，而最终只有一个 CPU 可以获得锁，也只有它的刷新才是有意义的。锁的争抢越激烈（试图获取锁的 CPU 数目越多），无谓的开销也就越大。

MCS 锁机制的核心思想：每一个 CPU 都分配一个自旋锁结构体，自旋锁的申请者（per-CPU）在 local-CPU 变量上自旋，这些结构体组建成一个链表，申请者自旋等待前驱节点释放该锁；OSQ(optimistci spinning queue)是基于 MCS 算法的一个具体实现，并通过了迭代优化。

通过对比优化前后我们发现，以 sysbench mutex 为例，在并发多线程锁竞争的场景下有不错的性能开销降低。

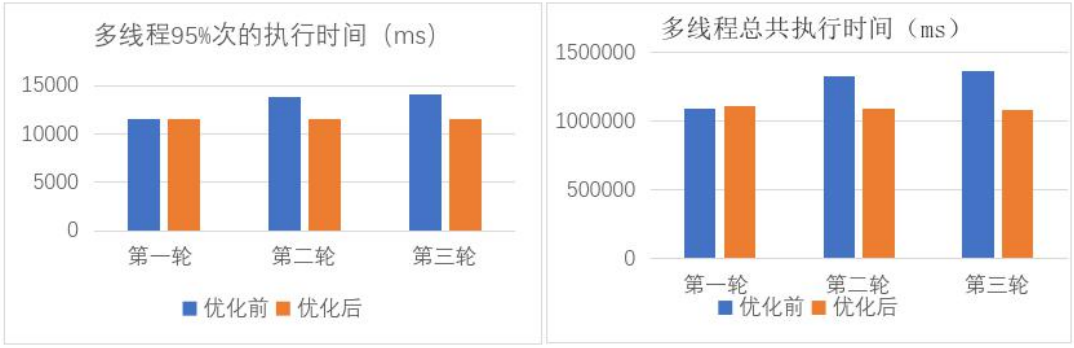


图 12 OSQ 锁优化效果

2.3.7. DPUOS 无感卸载

云 IaaS 平台虚拟化的技术演进逐步向软硬一体化发展，并在 DPU/IPU 卡上得以发挥。借助 DPU/IPU，不仅为高性能数据和 IO 场景提供更大技术施展空间。然后在软硬件不断发展的过程中，HostOS 所承载的资源开销越来越大和复杂。IaaS 的成本优势正在被大部分系统管理进程和云平台控制平面组件进程压缩。这对拥挤的计算节点提供更多的利润空间形成阻碍：

- openstack，kuberentes，libvirtd，dockerd，监控等 Node 节点的 agent 类组件繁多复杂，资源开销大。可以借助 DPU/IPU 进行卸载部署。
- 对第 1 点中提到的组件拆分要求开发者对卸载组件具备代码层级的了解。云厂商维护组件较多，相关组件卸载至 DPU 带来的拆分工作量巨大。
- 拆分工作在组件升级时很难继承，组件升级维护的成本较高，需要移植适配相关拆分代码。

通过在 HostOS 和 DPUOS 之间建立一个集 VFS，进程间通信，异步 IO 框架，Mount 等为一体的增强网络网络文件系统，为卸载到 DPU 侧的管理面进程和 HOST 侧的业务进程提供一致的运行时视图，达到应用对卸载低感知或零感知的效果。控制平台和系统的服务只需要少量适配管理面业务代码，保证了业务的软件兼容性和演进性，降低组件维护成本。从实现层面，本方案提供的机制可以结合定制策略实现不同场景下的进程无感卸载目标。

2.3.8. Ceph 存储节点性能优化

Ceph 是一个专注于分布式的、弹性可扩展的、高可靠的、性能优异的存储系统平台，可以同时支持块设备、文件系统和对象网关三种类型的存储接口。

tuned 是 Linux 系统中的一种调优服务，tuned 允许用户创建和自定义自己的调优配置文件，以适应不同的系统和应用场景。

面对云场景常用的分布式存储 Ceph 场景云底座系统提供典型的优化方案。通过 tuned 预置参数对 ceph 分布式存储节点进行性能优化。利用 tuned 预置经验参数，如内核参数、磁盘参数等，对 ceph 分布式的性能进行优化。

在 tuned 脚本中，首先筛选出 ceph 相对应的 osd 磁盘，对 osd 磁盘进行参数修改。通过 Tuned 的场景化优化，可以为 Ceph 存储节点基础性能提升 10%左右性能效果，如下图：

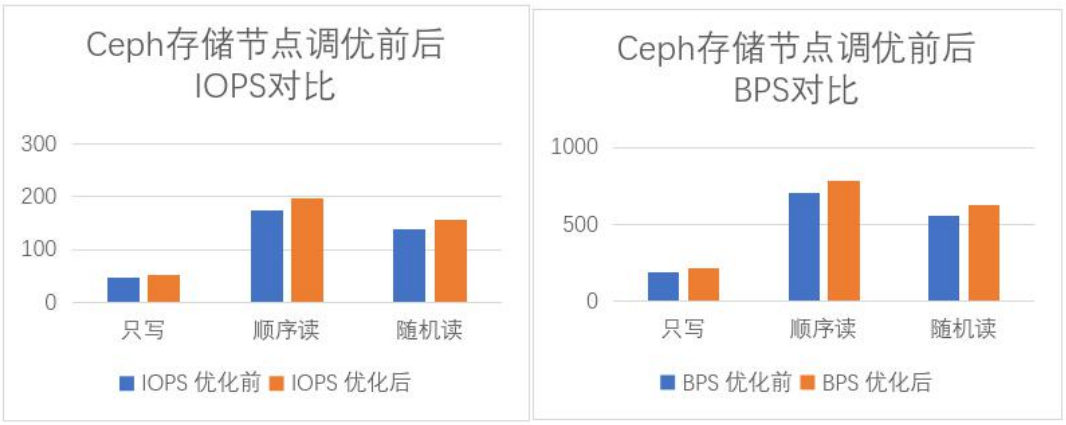


图 13 Ceph 调优方案效果

2.3.9. Generic-vDPA

基于 SR-IOV 的数据卸载方案使得虚拟机比肩物理机的网络性能，但各硬件产商在通信流程层面如驱动告知设备、驱动感知等协议都各不相同，没有统一透明的接口，OS 厂商实现这类设备时复杂度较高，且对虚拟机热迁移也不够友好。银河麒麟云底座操作系统为 virtio 驱动程序提供数据通路加速功能，隐藏 vDPA hardware 实现的复杂性，并为内核和用户空间提供一个安全统一的接口来使用。vDPA 可以由多种不同类型的设备实现，并预留支持热迁移 KABI。

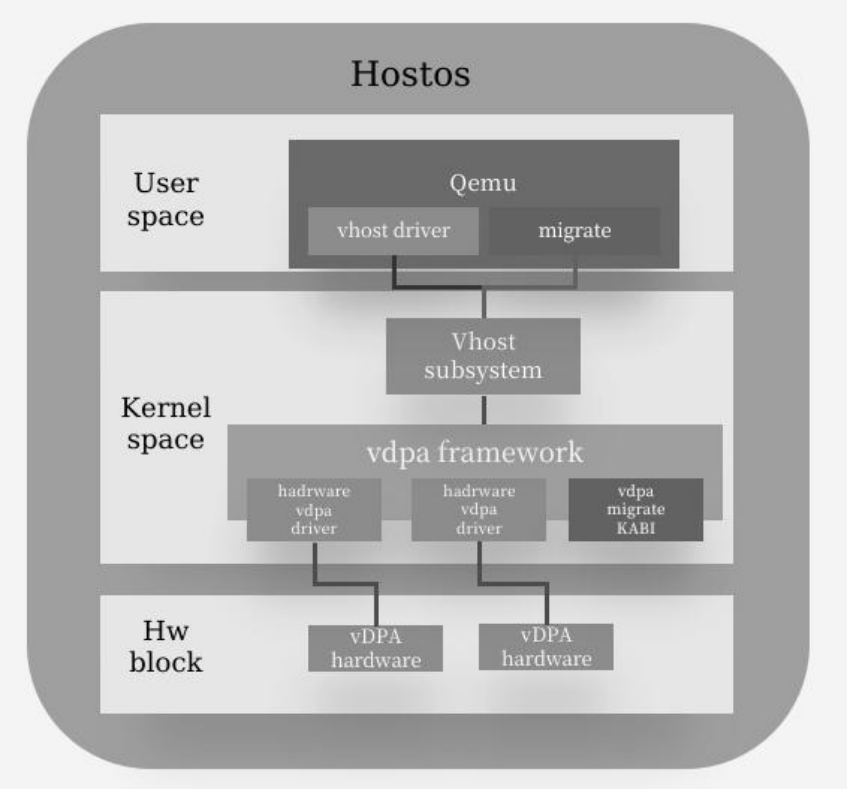


图 14 Generic-vDPA 架构图

vDPA 设备的数据路径是遵循 Virtio 协议规范的。因此，可以直接被宿主机上的 virtio 驱动直接访问。同时，该技术框架对 vhost 内核子系统进行了扩展，赋予了类似 VFIO 技术框架的功能，允许将 vDPA 设备用来进行数据通信的硬件资源透传给虚拟机使用。因此，虚拟机的 virtio 驱动进行数据通信时，直接访问硬件资源，提升虚拟机设备性能。由于虚拟机驱动是原本的 virtio 驱动，当需要支持热迁移时，Qemu 可以灵活切换会软件模拟的方式，来保证热迁移的顺利进行。

2.3.10. 系统级故障分析工具

鉴于现代系统架构的日益复杂性，故障排查与调试已成为一项极具挑战性的任务。面对诸如已知问题重现、配置错误频发、日志记录不足、监控机制缺失等常见难题，不仅对技术人员的专业技能提出了更高要求，同时也直接影响到服务的稳定性和用户体验。为满足市场高要求及内部 SLA/OLA 标准，我们提出引入“系统体检”机制。该

机制旨在通过主动扫描与问题检查，结合强化的主动监控体系，实现故障的早期预警与关键信息的即时收集，以此降低故障冲击，加速问题解决流程。整体架构如下图所示：

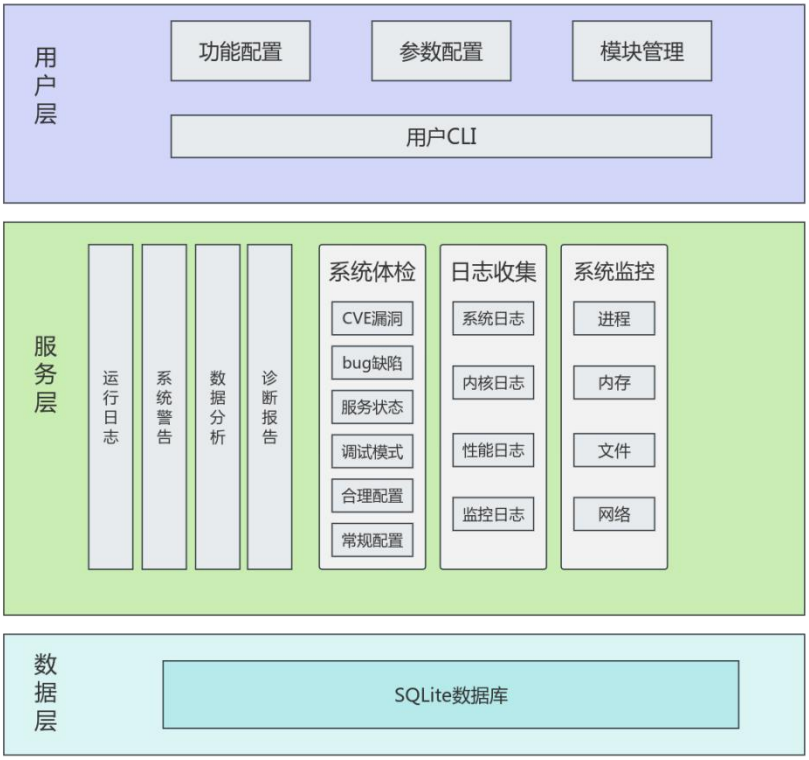


图 15 系统级故障分析工具技术架构图

系统体检功能覆盖多个维度，包含系统上安全漏洞的检测、软件包以及内核缺陷的检查、系统组件服务状态检查、调试项的状态检查、对系统上的特定配置进行合理性检查以及对系统的必要性配置进行状态检查。

日志收集包含一键收集全量日志的功能、选择性的收集和问题相关的日志以及在监控到系统异常时收集异常点相关的日志信息。

系统监控包含指定目录或文件的元数据和数据操作监控、对系统进程和系统负载的监控、对内存回收状态、内存泄露的监控以及对网络包在各协议栈流转的监控。

“系统体检”机制及其配套的日志收集与系统监控体系，构成了一个全面、主动的故障预防与响应框架，不仅有助于降低故障发生概率，还能在故障出现时迅速采取

行动，最大化减少服务中断时间，确保业务连续性和用户满意度。

2.3.11. 一键式性能收集工具

一键式性能调优工具，作为优化应用性能的强大助手，其设计初衷在于深入剖析系统性能指标，精准定位性能瓶颈，并以直观的 HTML 报告形式呈现调优建议，助力用户实现应用性能的显著提升。

该工具由三大核心功能模块构成，分别为数据采集、性能分析与静态调优，共同构成了一个闭环的性能优化流程。

数据采集模块负责全面捕捉系统层面的性能数据，涵盖 CPU 利用率、内存占用、网络带宽、输入输出（I/O）操作、Java 虚拟机（JVM）状态、系统日志记录，以及 perf 工具捕获的热点数据信息。这一阶段，旨在建立全面的性能基线，为后续的分析环节提供丰富的数据支撑。

性能分析模块则是工具的核心，通过对采集到的数据进行深度挖掘，高亮展示异常指标，识别导致性能下降的关键因素，并基于算法模型给出针对性的调优建议。这一过程，不仅依赖于先进的数据分析技术，还融入了丰富的性能调优经验，确保建议的实用性和有效性。

静态调优功能主要是针对典型应用场景下发优化配置参数。

图 16 性能收集工具架构图

用户可根据实际需求灵活配置功能参数配置，例如在读写 I/O 密集型工作负载下，适当延长磁盘信息采集周期，以获取更详尽的性能数据，进而得出更精确的分析结果，指导更合理的调优决策，显著提升性能优化的效率和效果。

2.3.12. 生成不可变操作系统镜像

随着云计算领域的迅猛发展，操作系统作为其基石，其稳定性和安全性正逐渐成

为行业关注的焦点。在这一背景下，NestOS 作为一款遵循不可变/原子化原则设计的操作系统，凭借其独特的架构与显著优势，成为容器云场景下的理想选择。

本产品提供基于 NestOS 技术路线的不可变操作系统镜像定制能力。通过本产品的软件源，用户可以便捷地访问一整套构建工具、配置工具以及安装部署工具软件包，从而实现自定义的 NestOS 镜像构建。同时，为简化用户使用操作，本产品将提供构建容器镜像和使用脚本的封装，以及 NestOS 配置规范的校验和转换为点火文件的能力。NestOS 整体架构如下：

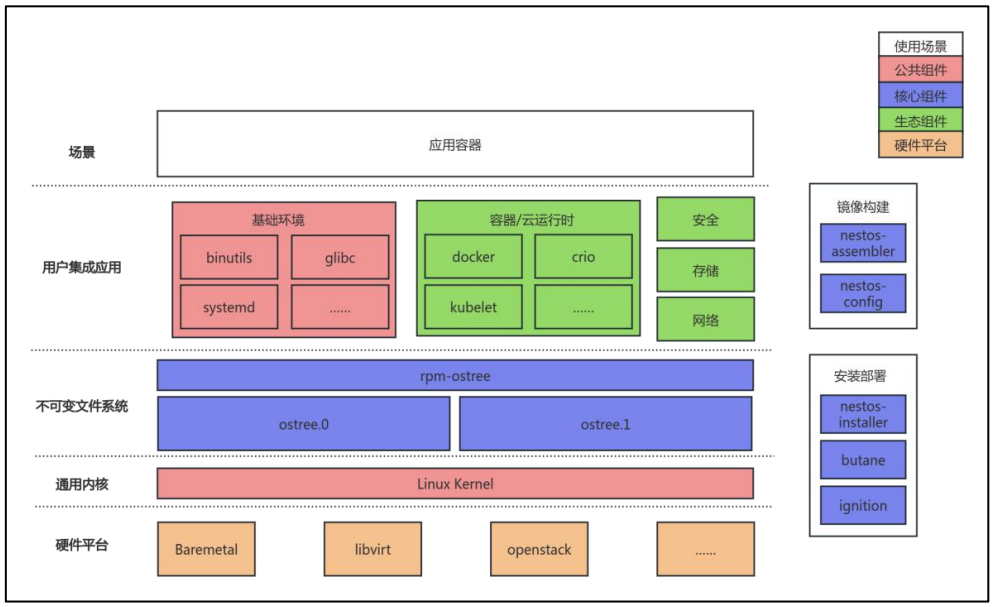


图 17 NestOS 技术架构图

其能力主要涉及以下几方面：

- NestOS 采用基于 ostree 和 rpm-ostree 技术的操作系统封装方案，将关键目录设置为只读状态，保障核心系统文件和配置不被恶意修改。

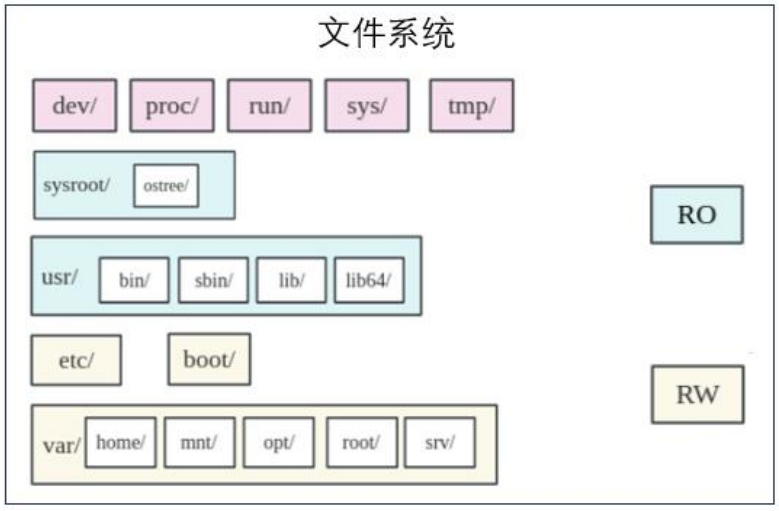


图 18 NestOS 目录结构

- 支持通过 RPM 包的形式分发构建工具 nestos-assembler 和配置文件 nestos-config，帮助用户便捷搭建构建环境，快速实现镜像构建；
- 支持配置规范版本 V1.0.0 与变体 nestos 设置，与 ignition 配置规范 v3.3.0 对应，支持灵活多样的用户自定义配置。
- NestOS 支持通过 OCI 格式镜像更新，实现以镜像为最小粒度进行操作系统版本的切换，满足快速部署和升级的需求。

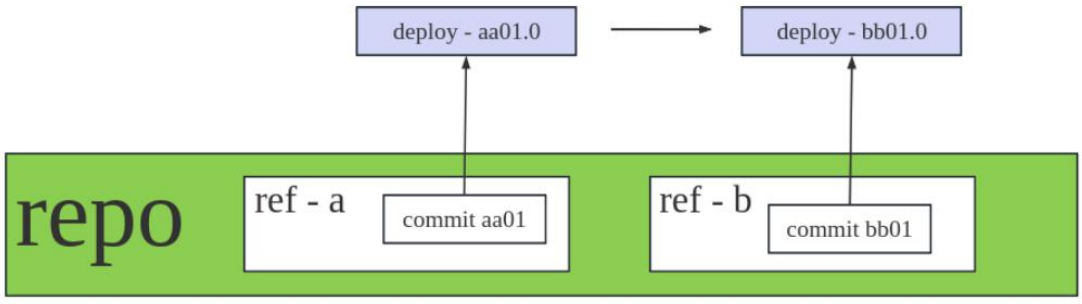


图 19 Ostree 部署切换示意图

2.3.13. 国密应用和可信计算

云计算、云桌面等虚拟化技术在党政、金融等行业的迅猛发展，虚拟加密磁盘，热迁移加密、vTPM 等常用功能长期使用 AES、SHA、RSA 等国际通用密码算法体系的现状亟需改变。银河麒麟云底座操作系统通过拓展支持国密算法，实现对虚拟化技术加密算法的自主创新。

通过对在 nettle 加密算法库上实现 SM2，SM3，SM4 国密算法、在 gnutls 证书上嵌入可提供解析 SM2 认证证书的功能，实现在 libvirt、qemu 上拓展支持虚拟化磁盘国密算法加密、热迁移国密算法加密功能。

2.3.14. 安全启动

安全启动（Secure Boot）是一项关键的安全技术，它运用公私钥加密原理，对系统启动部件进行签名和验证，以确保只有经过认证的代码才能在系统中执行。这一机制通过层层递进的验证流程，构建起一道坚固的防线，有效地抵御了未授权或恶意代码的加载，从而保护了系统的完整性和用户数据的安全。

在安全启动的过程中，各验证组件按照预定顺序逐一进行数字签名的验证与加载，具体流程如下：

- BIOS（Basic Input/Output System）验证：作为系统启动的第一步，BIOS 负责验证下一个启动部件——shim 的数字签名。若 shim 的签名正确无误，BIOS 将加载并执行 shim。
- shim 验证：shim 作为一个过渡层，其主要职责是验证接下来的启动部件——grub 引导加载程序的数字签名。通过这一验证步骤，shim 确保了 grub 引导加载程序的

合法性和完整性。

- grub 验证：grub 引导加载程序在被 shim 验证通过后，将继续验证内核镜像 vmlinuz 的数字签名。这一验证过程确保了内核本身未被篡改，保证了系统核心的纯净与安全。
- vmlinuz（内核镜像）验证与加载：一旦 vmlinuz 的数字签名被 grub 验证通过，内核将被加载至内存中并开始执行，至此，安全启动流程顺利完成，系统进入正常运行状态。

通过这一系列严谨的验证流程，安全启动机制有效地阻止了未经认证的、可能携带恶意代码的启动部件在系统中运行，从而显著降低了系统遭受攻击的风险，保护了用户数据免受侵害。这一机制在现代操作系统中扮演着至关重要的角色，尤其是对于云环境和企业级应用而言，安全启动成为了构建安全可信计算环境的基石。

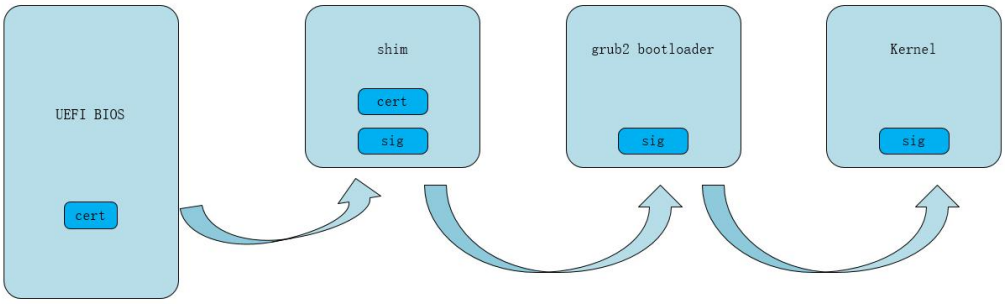


图 21 安全启动验签流程图

2.3.15. 动态度量

动态度量通过周期性地检查特定运行时程序的完整性，确保其未被恶意篡改。这一机制的核心目标是持续监视系统中的特定软件，一旦检测到篡改行为，立即记录详细的告警信息至审计日志，为后续的分析和响应提供依据。

动态度量的检查对象可以根据安全策略灵活配置，通常包括进程代码段、内核代码段和只读数据段等关键区域。为了在性能与实效性之间取得平衡，度量周期的频度可依据具体应用场景和资源条件进行调整，确保既不影响系统正常运行，又能有效捕捉安全事件。

在用户层面上，动态度量技术不仅能够监测进程是否遭到篡改，还具备通过策略配置实现对被篡改进程的主动干预能力，即在检测到篡改行为后，可根据预设策略自动终止受影响进程，防止其进一步危害系统安全。这种主动防御机制，大大增强了系统的自我保护能力，降低了安全事件对业务连续性的影响。

此外，为了便于用户管理和操作，动态度量机制还提供了命令行接口，支持策略查询与配置功能，使得用户能够根据自身需求，灵活调整度量规则和响应策略，确保系统安全策略的实施与优化始终与业务需求保持同步。

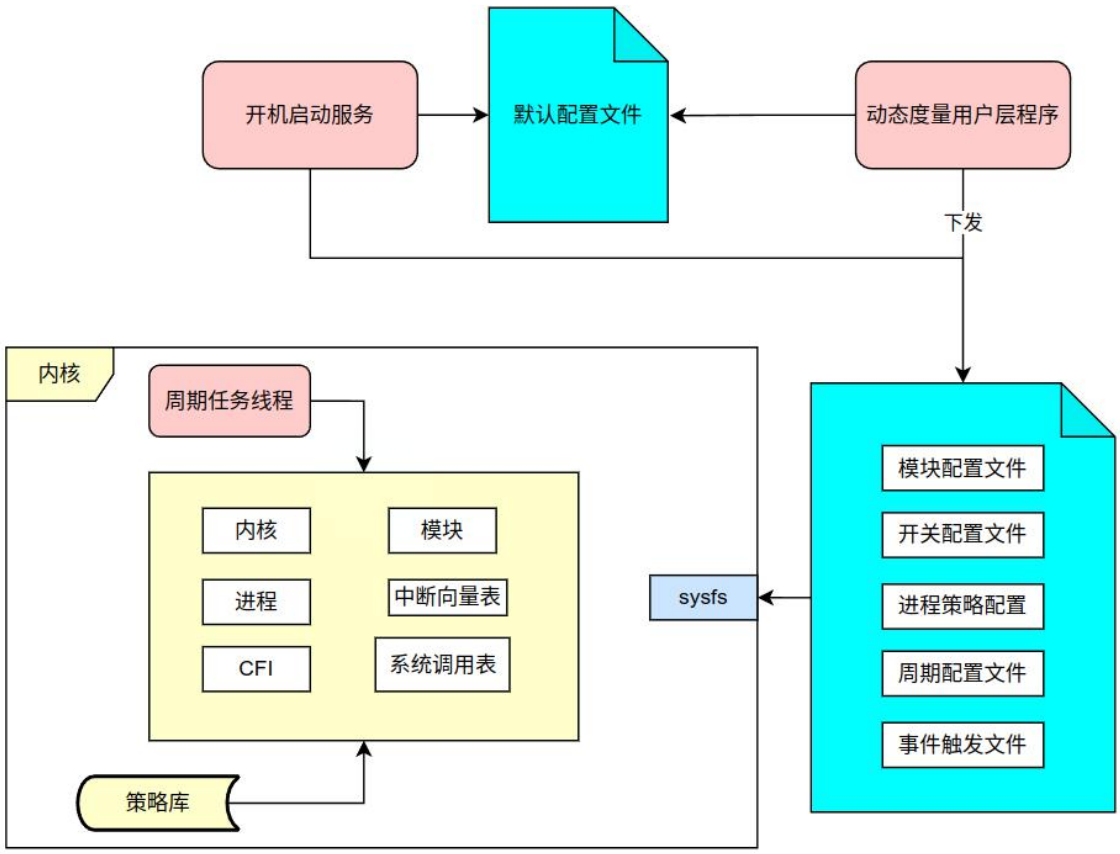


图 22 动态度量结构框图

动态度量作为一种主动的安全防护技术，通过周期性的对运行时程序检查，有效监测并阻止了恶意篡改行为，为系统提供了持续的安全保障。其灵活的配置选项和主动干预能力，使其成为构建多层次安全防御体系的重要组成部分，尤其在云环境和复杂网络架构中，动态度量技术的应用显得尤为重要。

2.4. 产品技术指标

类别	技术参数
基础核心	Kernel: 5.10
标准符合度	符合 POSIX 标准
	产品研发符合 CMMI5 标准
	符合 CGL 5.0
	符合 GB18030-2022 标准
架构支持	飞腾：FT-2000+/64、腾云 S2500 鲲鹏：920、920 V200 海光：海光 C86-3G、海光 C86-4G Intel/AMD 等服务器平台
最大支持的逻辑核数	X86：8192 ARM：1024
最大的物理内存	X86：16TB[64T] ARM：16TB[64T]
页大小支持	ARM：64K X86：4K
文件系统支持	默认使用 XFS，支持 EXT3、EXT4、GFS、GFS2 等。
安装引导	支持 GRUB2 引导，支持 MBR 及 GPT(GUID 分区表)分区，支持 NTFS 文件系统。
计算支持	支持双核及多核处理器； 支持并优化 NUMA 体系架构，支持在运行时分配 1GB hugetlbfs。

存储支持	内置支持快速块设备作为慢速块设备缓存以加速 IO； 支持 swap 压缩以减少 IO 并提高性能； 支持 FCoE、iSCSI，支持将 Ceph 块设备视为常规磁盘设备条目，挂载到某个目录并使用标准文件系统格式化，比如 XFS 或者 EXT4。	
网络支持	支持网卡 bonding。	
虚拟化支持	支持 KVM 虚拟化； 内置单机虚拟化管理程序； 支持作为 KVM、Xen、Hyper-V、ESXi 虚拟机。	
应用开发运行环境	开发工具	集成 Qt 等开发框架；
		支持 GCC 包含的 C、C++、Objective C、Objective C++ 和 Fortran 等； 相应支持库（libstdc++、libgcj...等）；
		支持 Python， Perl，Shell，Ruby，PHP 等脚本语言。
	运行环境	支持 JDK 1.8、11、17 等。
常用应用支持	默认提供 qemu、libvirt、Openvswith、docker、dpdk 等应用；	
易用性	安装	提供全中文化的图形操作界面及帮助；
	中文处理	采用 i18n（国际化）技术和标准； 支持最新国家标准字符集（如：GB18030-2022）。
	提供常用的系统服务支持并提供常用服务的中文帮助文档。	
	提供 OpenLMI 以简化任务配置及服务器管理； 提供图形化的远程桌面查看工具，支持 SSH、SPICE、VNC、RDP 协议 支持按需启动守护进程。	
高可用性	支持负载均衡；	
	支持多种网卡 Bonding，提高可用性；	
	支持存储多路径并提供国际标准 multipath 驱动。	
可扩展性	支持至少 10000 个“sd”设备；	
可维护性	提供在线升级服务、提供 HostOS 自动化软件包安装回滚工具；	
	支持动态内核补丁，支持在不重启的情况下为内核打补丁；	

	提供 Kdump 用于系统崩溃时的信息收集，支持最大 3TB 内存； 支持固件辅助转储 fadump； 支持系统 crash 时对系统崩溃信息进行分析。	
	支持 sosreport 收集系统配置和运行主机上的诊断信息，协助排查故障。	
	提供程序错误自动报告工具，统一不同源的出错数据集合，捕获、处理并记录所有来自内核追踪架构的可靠性、可用性及可服务性（RAS）出错事件。	
	提供强制访问控制故障排除工具。	
	提供 oProfile、papi 、elfutils 等内核性能分析工具； 提供一键式性能调优工具； 提供系统级故障分析工具； 提供 Performance Co-Pilot 对系统级性能测定进行采集、归档和分析的工具、服务及库套件。其轻加权、分布式架构的特点使其特别适合复杂系统的集中分析； 提供 SystemTap 在整个非特权用户空间运行的基于 DynInst 检测，同时也支持基于 Byteman 的 Java 应用程序精确探测。	
安全性	系统安全性	自研内核统一访问控制安全框架 KYSEC； 支持 LSM 统一访问控制安全框架； 可信计算 TCM/TPCM、TPM2.0（内核标准支持）； 内置国密算法，支持基于国密算法的加解密应用； 支持 PAM 认证机制，支持密码及秘钥管理； 支持安全启动； 支持动态度量； 支持内核模块黑名单。
	文件安全性	支持文件系统加密； 支持文件完整性检查。
	网络安全性	支持 Firewalld（IPtables），支持动态管理的防火墙，并支持网络“区域”以便为网络及其相关链接和接口分

		配可信度，支持 IPv4 和 IPv6 防火墙设置，支持以太网桥接并有独立的运行时和持久配置选项，提供可直接添加防火墙规则的服务或者应用程序接口。
	强制访问控制	支持强制访问控制，内置式一体化安全体系，支持多策略融合的强制访问控制机制。
	审计	提供系统审计日志。
软件兼容性	私有云：OpenStack 等； 云原生：kubernetes、fluentd、containerd、crio、podman、docker、 nginx 等； 分布式存储：Ceph 等。	
硬件兼容性	兼容国内外主流的服务器、存储等硬件产品，包括联想、曙光、浪潮、方正、华为等； 智能 DUP 卡：Mellanox 等。	

3. 生态适配

服务器整机、虚拟化、云原生、云厂商适配请访问麒麟软件官网的软件和硬件兼容适配列表页面进行查看。（https://eco.kylinos.cn）

4. 应用场景

银河麒麟云底座操作系统 V10 提供中文化的操作系统环境和常用管理工具。支持多种安装方式，提供完善的系统服务和网络服务；集成多种易用的编译器并支持众多开发语言，全面兼容国内外的软硬件厂商。

基于银河麒麟云底座操作系统 V10，用户可以轻松构建大型数据中心、并支持云原生、虚拟化、云平台、分布式存储等场景，在国产化平台上，可以替换 CentOS，满足云厂商对 Host OS 的需求；支持海量的层出不穷的云组件，支持不断创新的国产芯片和整机，满足国家安全等级要求的云场景通用操作系统。

5. 开发环境与工具

5.1. 系统开发环境

以 GCC 为核心并集成了 Eclipse 强大的开发环境，几乎覆盖了集成开发环境（IDE）的每个方面，其中 C/C++(CDT)和 Java(JDT)是 Eclipse 两个主要的开发工具包。银河麒麟云底座操作系统 V10 同时支持跨平台应用 Qt 开发框架。

银河麒麟云底座操作系统 V10 对 GCC 进行了漏洞修复和功能增强，支持旧版本的更新移植。目前的 GCC 包含 C、C++、Objective C、Chill、Fortran 和 java 的前端，并包括这些语言的支持库（libstdc++、libgcj...）。GCC 的开发是 GNU 计划的一部分，旨在增强包括 GNU/Linux 在内的 GNU 系统的编译器。GCC 的开发完全是在开放的环境中进行的，并支持其他的平台。

银河麒麟云底座操作系统 V10 还可支持诸如 Python，Perl，Shell，Ruby，PHP 等脚本语言。

5.2. 构造工具

开发大型的软件程序是一个复杂的过程。构造工具通过实现构造过程中某些步骤自动化达到简化过程的目的。make 是 Linux 系统的主要构造工具，它可以使你很容易描述如何编译程序，通常的构造工具包括：

- make：自动地确定一个大程序的哪一部分需要编译，并启动命令重新编译它们；
- Antoconf：一个可以自动配置源代码包的工具；
- Automake：一个为 autoconf 生成 Makefile.ini 文件的工具；
- RPM/DNF：包管理工具。

5.3. 调试器

调试器可以使程序员观察到另一程序执行的内部情况，或查看另一程序在崩溃时

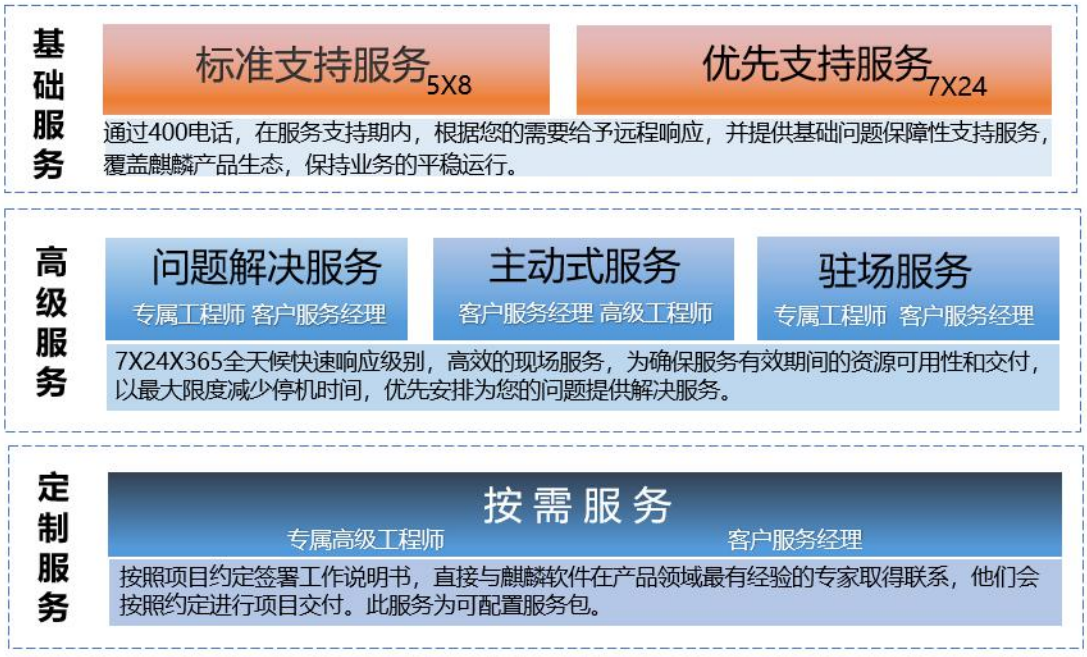
正在做些什么。 GNU 的调试器 GDB 可以帮助程序员做以下 5 类工作：

- 启动程序，规定任何对程序有影响的参数；
- 在进程中设置断点，暂停程序的执行；
- 当进程处于停止或暂停状态时，检查程序的状态；
- 修改进程的内部参数；
- GDB 目前可用于调试用 C 或 C++编写的程序。

6. 技术服务

麒麟软件有限公司拥有完善的技术服务体系和一流的服务团队。服务遵循 ISO27001、ISO20000、ITSS 等体系标准要求，为客户提供专业的厂商级服务。

可提供多种服务模式，包括基础服务、高级服务、定制服务等，服务产品如下：



为了满足不同用户、不同场景的需求，让用户能够享受周到、专业的服务，麒麟软件依据地域情况形成覆盖全国的技术服务团队，服务网点遍布全国 31 个省会城市+2 个计划单列市，主要区域均可快速响应客户服务请求。

7. 结束语

多年来，麒麟软件通过坚持自主创新、并持续融入国际开源社区的方式，形成了一只具有强大凝聚力和雄厚科研能力的核心技术团队，在国内外操作系统行业中独树一帜。

麒麟操作系统及相关软硬件产品和解决方案已经在政府、电力、电信、金融、能源、交通、邮政、教育等行业以及国家援外项目中得到了成功应用，并将进一步联合芯片、整机、数据库、中间件、应用软件和系统集成等上下游产业伙伴企业，持续共建网信产业生态环境。

未来，麒麟软件将继往开来，展现国企担当，立足对标世界、中国最好的操作系统产品目标，为用户提供统一、高品质的操作系统产品、方案和技术服务；为国产计算机提供安全智能可靠的“中国大脑”，让中国的软件基础设施不再受制于人。