KYLINSOFT
麒麟软件

银河麒麟高级服务器操作系统 V10

技术白皮书

麒麟软件有限公司

2024 年 03 月

目录

1 概述

银河麒麟高级服务器操作系统 V10 是吸收开源社区技术成果，融合 openEuler 社区各个 LTS/SP 版本的特性和功能，基于自身超 20 年 Linux 产业技术实践积累，安全深度优化的产品。为满足客户针对作为业务系统基座的操作系统产品长期维护的需要，V10 的内核主版本（4.19）在整个产品生命周期内保持不变，并融入包括 openEuler 社区在内的开源社区新版本的各类新特性，以保证操作系统技术领先。

麒麟软件一直坚持回馈社区、在欧拉社区的贡献排名第二。麒麟 V10 充分融合了欧拉社区的创新特性和麒麟软件深厚的产品化研发能力。

安全增强、Linux 内核是麒麟独立研发维护的。内核源码基础来自 kernel.org 官方社区。

麒麟 V10 已经发布了四个版本：V10、V10 SP1、V10 SP2、V10 SP3 2303。去年启动了麒麟 V10 SP3 2403 的研发工作。经过 18 轮的版本迭代，各项产品指标均已符合麒麟产品标准要求。在此，我们对麒麟 V10 SP3 2403 的技术特性进行介绍。

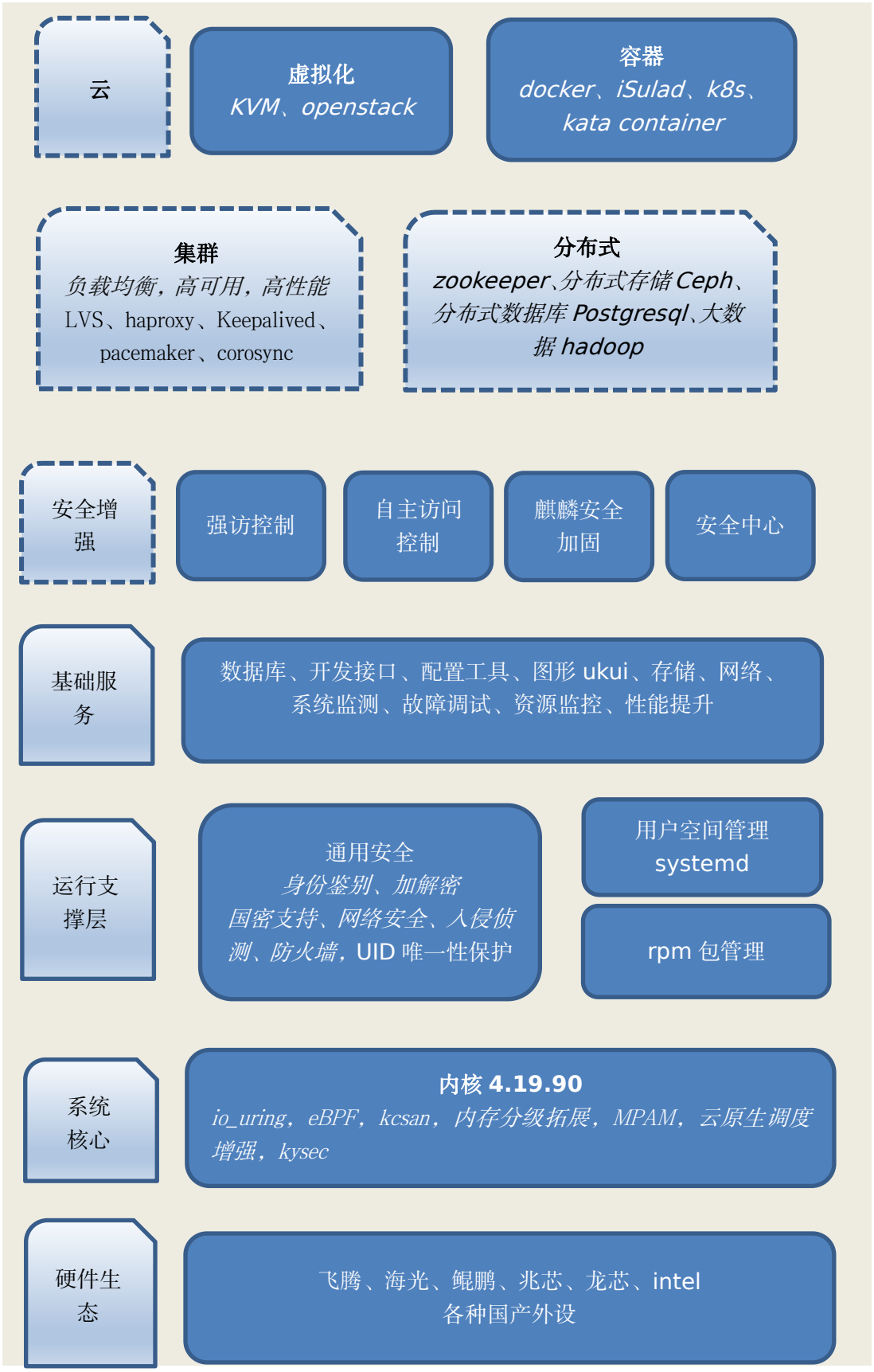
麒麟 V10 SP3 2403 是一种 Linux 操作系统。Linux 也叫 GNU Linux，是开源组织 GNU 推动的产物。Linux 主要技术特征是：宏内核、开源组件众多、内部包含核内空间和核外两部分。硬件的价值必须在软件运行中体现，软件无法脱离硬件独立运行。驱动程序是软硬件协同的基础。麒麟 V10 SP3 2403 包含了大量的驱动程序，是麒麟与友商广泛深入合作的成果，对硬件生态体系形成了坚实的基础保障。服务器软件生态是最能体现软件价值的地方，麒麟研发与测试团队联合友商或社区协力完成了各种重要的生态软件版本适配工作。

功能、安全、稳定与性能，一直是麒麟服务器研发团队不懈努力的坐标点。麒麟 V10 SP3 2403 的内核功能特性、性能特性、硬件支持等方面，有很大的提升和拓展，特别是对新国产硬件的支持。麒麟 V10 SP3 2403 的安全特性也进一步增强，国密支持、安全加固配置、安全漏洞修复等方面研发成果丰富。面向云原生的容器、虚拟化技术一直麒麟 V10 的重点

研发方向，SP3 2403 的虚拟显示、虚拟盘加密、热迁移、容器安全与性能等方面提升明显。针对集群和分布式技术，以及大数据应用，麒麟 V10 SP3 2403 集成了丰富的功能组件和工具，可以适应各种需求场景。构成操作系统的系统组件之间关系复杂，但是这些关系不是来自组件之外，任何一个关系都是组件内部的逻辑衍生出来的。所以，麒麟研发团队一直追求深入理解每个组件的内部逻辑，为系统的稳定和性能在逻辑层面寻找问题原因和待优化点。

下面是麒麟 V10 SP3 2403 的架构特性图。相关内容在后续章节展开介绍。

系统架构特性图


云
虚拟化
KVM、openstack
容器
docker、iSulad、k8s、
kata container
集群
负载均衡，高可用，高性能
LVS、haproxy、Keepalived、
pacemaker、corosync
分布式
zookeeper、分布式存储 Ceph、
分布式数据库 Postgresql、大数
据 hadoop
安全增
强
强访控制
自主访问
控制
麒麟安全
加固
安全中心
基础服
务
数据库、开发接口、配置工具、图形 ukui、存储、网络、
系统监测、故障调试、资源监控、性能提升
运行支
撑层
通用安全
身份鉴别、加解密
国密支持、网络安全、入侵侦
测、防火墙，UID 唯一性保护
用户空间管理
systemd
rpm 包管理
系统
核心
内核 4.19.90
io_uring，eBPF，kcsan，内存分级拓展，MPAM，云原生调度
增强，kysec
硬件生
态
飞腾、海光、鲲鹏、兆芯、龙芯、intel
各种国产外设

2 内核

V10.SP3 2403 内核继续沿用 V10 整个生命周期的 4.19 内核, 在功能特性、性能优化、安全加固、生态适配等多方位持续发力，累计合入 IO 网络、容器虚拟化等重要功能特性 68 项，优化 Unixbench、容器网络等场景性能 15 处，适配新发布 cpu 平台 6 个，升级阵列卡、光纤卡、网卡等重要外设驱动 23 个，适配 LSI、Intel、mallox 等新型板卡 53 款，累计更新 patch 20000 余个,修复内核问题数 950 余个，修复 CVE 漏洞 320 余个。

2.1 I/O 网络

- multipath 多路径支持 Historical Service Time (HST) 和 I/O Affinity 路径选择器
- EXT4 文件系统新增加密特性支持
- block 层全面升级为多队列通道
- overlayfs 文件系统新增 userxattr 属性支持
- BFQ IO 调度器调度优先级策略 QOS 支持
- Bonding balance-alb/balance-tlb 两种模式支持 IPv6 ns/na 特性
- Ipv6 支持 DNR 双转发功能
- NVMe 新增 host-auth 和 target-auth 认证特性支持
- NVME 新增 Nvme-Over-Tcp 特性支持
- tcp 支持 tw_timeout 功能特性

2.2 容器虚拟化

- vcpu 新增 stall detector 检查器
- cgroup 新增 freezer controller 特性支持
- arm64 平台新增 PTP_KVM 支持
- virtio 支持到 V1.1 版本特性
- memcg 新增 memory.events 接口属性
- 新增 CONFIG_VHOST_SCSI 特性支持
- 新增混部 SMT 驱离防止优先级反转特性支持
- 新增 CPU 调度负载均衡混合部署特性功能
- 新增 MemCG 异步水位线混合部署特性功能

2.3 进程调度

- ARM64 新增 CLUSTER 调度域
- 新增 IO 亲和调度器支持
- 调整 ARM64 架构的默认抢占策略为 PREEMPT_VOLUNTARY

ebpf 增强

- eBPF 新增 bpf_for_each_map_elem helper 接口支持
- eBPF 新增 ringbuf 支持
- eBPF skb_verdict 新增 SK_PASS 特性支持
- eBPF TRACING 模式支持 bpf_get_socket_cookie helpers 特性
- 新增 eBPF BTF_KIND_FLOAT 支持
- 新增 bpf_redirect_peer 接口
- 新增 bpf_bpf_redirect_neigh 接口
- 新增 bpf_ktime_get_boot_ns 接口
- 新增 skb 丢包 Reason 类型支持
- 新增 available_filter_functions_addrs 接口

安全增强

- 新增 x86 arm64 平台硬件加速算法库
- 新增 内核签名国密算法支持
- 新增 内核模块签名国密算法支持
- 新增 ktsl 国密算法支持
- [kysec] 联网管控功能增加黑白名单机制切换
- [kysec] 新增 静态度量 pcr 扩展特性

其他特性增强

- ACPI 新增 HMAT 特性支持
- page_owner 新增 timestamp 和 pid 以及进程名称的支持
- perf 新增 ARMv8.3-SPE 支持
- 内核启动参数新增 hostname=参数来设置默认的主机名
- RCU 新增 stall diagnosis information 特性支持
- 新增 UKFEF 统一内核故障框架支持
- 内核新增 ARM64 高性能特性 LSE 支持
- X86 新增 GOV_SCHEDUTIL 调频支持
- USB 新增 raw-gadget 特性支持
- 新增 KT0206 音频卡驱动支持
- PCIe 5.0 特性增强
- 华为 Kunpeng 平台 RAS 特性增强

2.7 国内外主流 CPU 等芯片支持

银河麒麟服务器 V10.SP3 2403 完成的硬件生态适配重点是一些新的 cpu：飞腾 S5000C、鲲鹏 920+ 、龙芯 3D5000、兆芯 40000、海光 4 号。并且对飞腾 S2500 等已支持的 cpu 进行

了优化。另外还增加了一些针对厂家全型号的特性支持。具体如下：

- 飞腾 S5000C PCIe/DDR PMU 驱动支持；
- 飞腾 S5000C 网络亲和性支持
- 新增飞腾 CEU 算法驱动支持
- 新增飞腾 E2000 BMC 内置显示驱动支持
- Optee 新增飞腾设备树描述信息支持
- 新增飞腾 S2500 RAS 基础功能支持
- 新增 X100 相关 SOC 外设驱动
- ARM64 架构新增 AMU 扩展属性支持
- 新增 Phytium DDR/C2C/PCIE PMU 支持
- 鲲鹏 920+ 新增 HISI PTT 驱动支持
- 鲲鹏 920+ HISI SPI/SFC 驱动支持
- 鲲鹏 920+ HISI 架构新增 ll_cache_miss_rd 和 ll_cache_rd 通用事件支持
- 鲲鹏 920+ HISI ras 功能增强
- LoongArch64 架构新增 BTF 属性支持
- LoongArch64 架构新增 ETMEM 支持
- LoongArch64 架构 ls2k500sfb 内置显示驱动增强
- LoongArch64 架构新增 SRIOV 支持
- LoongArch64 架构使能 rubik 支持
- 新增 LoongArch64 平台 ls2k500sfb BMC 支持
- 新增 Arm64 架构 Inspur BMC 显示驱动支持
- 兆芯 40000 新增 SM2 加密算法驱动支持，添加对 ZXPAUSE 指令的支持
- 兆芯平台新增 SM3/SM4 硬件算法支持
- 兆芯平台增强 PMU 支持
- 兆芯架构新增 MWAIT C-state 支持
- 新增 Hygon 4 号支持
- Hygon 平台 虚拟化增强

另外，银河麒麟服务器 V10.SP3 2403 内核针对服务器存储和网络相关的重要驱动进行了升级，如 vrcraid 阵列卡驱动、sssraid 阵列卡驱动、SSSNIC/3s9xx 网卡、楠菲以太网卡 PS1600 驱动、沐创网卡驱动 rnpgbe、Intel QAT 驱动、Mucse RNP/RNPVF/RNPM 网卡驱动、网讯网卡驱动、光润通网卡驱动、等。

3 系统组件

Linux 作为最成功的服务器操作系统，应用广泛，社区中技术积累深厚。特别是基础组件方面的稳定性和可靠性非常高。一方面这些组件应用广泛，经受了各种环境的考验；另

一方面，新增补丁的合入一般都非常谨慎，逻辑论证和验证工作比较充分。 所以，银河麒麟服务器操作系统也秉承这种思路，充分吸收了开源组件的稳定成果；并且在引入补丁的过程中，严格评审、充分验证，排除不稳定的因素，为银河麒麟服务器操作的稳定奠定坚实的基础。

3.1 系统组件介绍

内核以及之上的系统基础部分，是麒麟服务器操作系统架构组成中的核心。基本特征是：系统启动后，支持用户字符环境登录，进入命令行环境，进行各种操作配置。最小化安装一般就是系统基础部分，不过一般情况下，最小化安装后仍然可以裁剪一些服务或者工具，裁剪后留下的组件的基础性更强。一个基础组件被上层组件依赖的程度越高，比如 C 库被依赖的程度非常高，或者依赖次数很多，比如 python 被很多包依赖，我们就可以理解为这个组件相对更加基础。基础组件的功能主要是：

- 系统运行——涉及内核、initramfs 镜像、systemd、login、shell、pam 等。
- 配置管理——服务管理、文本工具、进程管理、网络管理、存储管理等。

基础组件之上，麒麟操作系统还包括存储、网络、虚拟化、容器、集群、数据库、安全、性能、开发、图形、等很多方面的组件。每一个方面一般会提供多个优选后的组件供用户选择。

下面是一些系统组件举例：

组件名称	功能说明
glibc	GNU 发布的 libc 库，即 c 运行库。glibc 是 linux 系统中最底层的 api，几乎其它任何运行库都会依赖于 glibc。
glib2	GLib 是构成项目基础的底层核心库比如 GTK+和 GNOME。它为 C 提供数据结构处理，可移植性包装器，以及用于此类运行时功能的接口作为事件循环、线程、动态加载和对象系统。

xlib	Xlib 是一个用 c 语言编写的 X Window System 协议的客户端库，它包含有与 x 服务器进行通信的函数，编程者可以在不了解 x 底层协议的情况下直接使用它进行编程。
gtk	是一套源码以 LGPL 许可协议分发、跨平台的图形工具包。已成为一个功能强大、设计灵活的一个通用图形库，支持创建基于 GUI 的应用程序。
coreutils	Coreutils 软件包包括一整套基本的 shell 工具。是 GNU 提供了一整套比较基本的工具软件包，是这些工具的集合。
bash	Bash 是一个命令处理器，通常运行于文本窗口中，并能执行用户直接输入的命令。
openssh	OpenSSH 是 SSH 协议的免费开源实现。SSH 协议族可以用来进行远程控制，或在计算机之间传送文件。
audit	审计服务，专门用来记录安全信息，用于对系统安全事件的追溯。
systemd	提供更优秀的框架以表示系统服务间的依赖关系，并依此实现系统初始化时服务的并行启动，同时达到降低 Shell 的系统开销的效果。
firewalld	firewalld 是一个防火墙服务守护程序，提供动态可自定义具有 D-Bus 接口的防火墙。
libX11	基于 X11 协议的客户端；X Client 最重要的工作就是处理来自 X Server 的动作，将该动作处理成为绘图数据，再将这些绘图数据传回给 X Server。
libdrm	直接渲染管理器运行库。DRM 是 Linux 内核层的显示驱动框架，它把显示功能封装成 open/close/ioctl 等标准接口，用户空间的程序调用这些接口，驱动设备，显示数据。
xorg-x11-server	基于 X11 协议的服务端。管理硬件设备（驱动），键盘鼠标显示器等。
dbus	D-Bus 最主要的用途是在 Linux 桌面环境为进程提供通信，同时能将 Linux 桌面环境和 Linux 内核事件作为消息传递到进程。
udev	udev 的功能是管理/dev 目录底下的设备节点。它同时也用来接替 devfs 及热插拔的功能。
qt	Qt 是一个跨平台 C++图形用户界面应用程序开发框架。它既可以开发 GUI 程序，也可用于开发非 GUI 程序。
Python	Python 是一种面向对象、解释型、弱类型的跨平台脚本语言，它也是一种功能强大而完善的通用型语言。
Perl	Perl 一种功能丰富的计算机程序语言，用于各种任务，包括系统管理，Web 开发，网络编程，GUI 开发等。
grub2	是一个多重操作系统启动管理器。用来引导不同系统。
anaconda	是一个安装程序管理器。
dnf	rpm 是用来管理软件包的程序。dnf 基于 RPM 包仓库进行管理，从指定的服务器自动下载 RPM 包并且安装，可以自动处理依赖性关系，并且一次安装所有依赖的软件包。
networkmanager	让用户更轻松的处理网络需求，尤其是无线网络，能够自动发现网卡并配置 IP 地址。

gdisk	用于 GPT 磁盘的类似 fdisk 的分区工具。gdisk 的特点是命令行界面，相当直接地操作分区表结构、恢复工具，帮助您处理损坏的分区表，以及将 MBR 磁盘转换为 GPT 格式的能力。
upower	提供了一个守护进程、API 和命令用于管理连接到系统的电源设备的线路工具。
httpd	httpd 是 Apache 超文本传输协议(HTTP)服务器的主程序。被设计为一个独立运行的后台进程，它会建立一个处理请求的子进程或线程的池。
rsyslog	负责收集 syslog 的程序。
nftables	网络包过滤框架与工具
firefox	图形环境中的浏览器
cracklib	密码检查工具与 API
cryptsetup	磁盘加密工具
openssl	SSL/TLS 协议，数据加密传输
pcp	一系列可用于监控和管理系统性能的服务
cockpit	系统管理员可以通过 cockpit 执行网络配置、检查日志、服务管理、内核转储等任务。
UKUI	桌面环境

3.2 系统组件的主要功能增强

3.2.1 功能增强——新增软件

麒麟 V10 SP3 2403 新增一百多个 rpm 包，进一步增强了系统功能。比如：

- ceph-deploy， 分布式存储 Ceph 的管理和部署工具
- code-oss， Visual Studio Code 开发编码工具
- drbd90-utils， DRBD 网络存储工具，支持 HA 集群
- kmod-drbd90，DRBD 网络存储工具，支持 HA 集群
- entr, 文件变动监控工具
- etmem, 内存分级扩展
- hadoop， 分布式大数据框架
- hardlink， 硬链接工具

- hct， 海光加解密加速库
- iodump， io 细节获取工具
- iowatcher， blktrace 结果可视化工具
- kunpengsecl，鲲鹏专用远程认证安全组件
- kylin-activation-ukey-driver， 麒麟认证 ukey 支持
- kysdk-module-authorize，kysdk 授权模块
- nettrace，内核层网络包跟踪调试
- LZMA-SDK， 压缩工具
- memstrack， 内存分配使用情况分析工具
- p7zip，压缩工具
- zx_gmi，兆芯 gmi 密码库
- extuner， 性能调优工具

Extuner 是基于麒麟操作系统的一键式场景性能调优工具，通过执行一条命令，收集服务器的全量性能数据，包含 CPU、内存、网络、IO、系统参数等，形成可视化数据报告，并直观展示异常指标、调优建议、调优指导等，使用者可以通过报告了解系统当前系统状态及瓶颈，从而根据指导内容，有针对性的对系统进行调优操作。同时该工具预置多款典型应用场景(如数据库、web、中间件等)的经验优化参数，用户可通过一键式设置，即可提升应用性能。

- kylin-sysassist， 系统助手

支持系统状态体检，能扫描出系统中存在的已知安全漏洞和系统缺陷，并提供相应的解决方案；能识别系统运行中的异常服务、不合理配置以及未开启的调试核心配置，并提供相应的建议和操作说明。该功能可提供详细的分析报告供用户查看。支持系统日志一键收

集，能对系统日志进行快速地自定义或者全量收集。该功能可提供统一的日志集合文件。支持系统监控，能对系统进程的资源进行周期性监控，监控期间可根据预定的指征，动态调整监控频率并开启进程跟踪，记录进程运行前后的关键信息；能对指定的文件或者目录进行监控并捕获操作者的进程名、操作时间、父进程等详细信息；能对指定网络参数类型的数据包进行实时跟踪并捕获数据包流向路径、时间和状态等信息；能对系统内存回收状态进行监控并捕获触发内存回收的进程相关信息；能对内存的使用进行周期性监控并评估出可能存在内存泄漏；能对内存的碎片化程度进行监控，并当指定大小的连续内存页趋于碎片化时给与告警。（loongarch64 架构暂不支持该功能）

3.2.2 功能增强——组件升级优化

还有大量的组件升级优化，特别是 C 库、账户秘钥等基础重要组件的功能增强。下面是一些重要系统组件的功能方面的变动举例：

glibc： 系统基础库

在 init_cacheinfo 中增加对兆芯的虚拟机检测支持；

在 glibc-compat-2.17 包中添加 libpthread_nonshared.a；

添加兆芯补丁，提升 memcpy 等接口的性能；

systemd： 用户空间管理

移除可执行文件的 rpath/runpath 属性值、提升安全；

添加一些内存 overflow 溢出检查；

journalctl 命令新增功能参数--facility=kern；

优化 udev 规则文件的匹配逻辑中的文件权限设定；

shadow： 秘钥支持

UID 唯一性增加 flag， off 状态 -u 参数默认 getpwuid，on 状态，-u 使用 uid_used。

增加 zh_HK 的支持。

xorg-x11-server：图形服务

增加 multi-gl sietium driver 支持；

增加芯动科技风华显卡支持；

增加支持凌久 GP201 显卡驱动。

gcc：C 编译器

优化 int128 的长整型的使用，在运行时进行更准确的判断。

新增对 cpu 核心 tsv110 的支持，增加 tsv110 调度支持

优化部分进行了重构，比如冷热点的优化，循环的优化内容。

新增 bolt 优化功能。

audit：系统审计服务

在 auditd 停止时释放异步刷新锁；

audit 设备管控需求合入通用,添加程序黑名单类型；

添加 kysec_ppro 和 kysec_netctl 日志类型。

cryptsetup：存储加密

添加了对新的 no_read/write_wrokqueue dm-crypt 选项的支持；

增加了对 dm-verity 设备的 panic_on_corruption 选项的支持；

支持用于在线 LUKS2 重新加密的 --master-key-file 选项；

integritysetup 支持新的 dm-integrity HMAC 重新计算选项

integritysetup 在 dump 命令中显示重新计算扇区。

java-11-openjdk：java 语言

TLS, 1.2 的默认 Diffie-Hellman 密钥大小从 1024 位增加到 2048 位

JDK 现在接受 PKCS#1 格式的 RSA 密钥。这意味着 RSA 密钥可以在更多格式下使用

支持 GB18030-2022 标准，与 Unicode 11.0 同步；

引入了系统属性 jdk.jar.maxSignatureFileSize 来配置 JAR 文件验证期间签名相关文件的最大字节数；

更新了 PKCS#12 密钥库中默认的 MAC 算法，基于 SHA-256；

libabigail：ABI 兼容性

提升了 Linux Kernel 二进制文件的类型比较优化；

添加了 abidiff 的--debug-tc 调试选项。

multipath-tools：多路径存储

新增命令 del maps 刷新映射表并删除不识别设备；

配置文件新增 protocol 子部分，实现针对不同存储协议进行配置和属性设置，从而为不同存储协议提供最佳配置；

新增 remove_local_disk 开关，当开启时，只在 FC 或 iSCSI 设备上创建多路径设备 。

mesa：图形加速

新增特性 virtio-gpu 支持硬件编解码（H264、H265）

openssl：安全传输

支持内核模块国密签名功能，openssl 上层包需要合入相关补丁

增加兆芯国密支持 SM2、SM3、SM4

增加 TLCP 的支持。

支持 sm2utl，返回值与其他模块保持一致，支持默认 ID

pam：用户认证

pam_exec 实现了 quiet_log 选项。

pam_mkhomedir 在/etc/login.defs 中增加了对 HOME_MODE 和 UMASK 的支持。

为提供的库增加 pkgconfig 文件。

增加了--with-systemdunitdir 配置选项来指定 systemd 单元目录。

增加了--with-misc- conf-bufsize 配置选项来指定 libpam_misc 的 misc_conv()函数中的缓冲区大小，将该参数的默认值从 512 提高到 4096。

pam_faillock:增加了不设置 pam_fail_delay 的 nodelay 选项

扩展 libpam API 与 pam_modutil_check_user_in_passwd 功能。

configure 添加了--disable-unix 选项来禁用 pam_unix 模块的构建。

pam_faillock 将/run/faillock/$USER 权限从 0600 修改为 0660。

pam_limits 增加了对 nonewprivs 项的支持。

pam_pwhistory 添加 SELinux helper。

pam_unix 和 pam_usertype:避免某些定时攻击。

pam_wheel 在 getlogin 失败的情况下实现 PAM_RUSER 回退。

删除了 pam_cracklib 模块，使用 pam_passwdqc(来自 passwdqc 项目)或者 pam_pwquality(来自 libpwquality 项目)。

移除已弃用的 pam_tally 和 pam_tally2 模块，使用 pam_faillock 代替。

pam_env 不赞成读取用户环境，将被删除。

pam_motd 按照用户和组筛选 motd。

extuner：一键式性能调优工具

通过执行一条命令，收集服务器的全量性能数据，包含 CPU、内存、网络、IO、系统参数等，形成可视化数据报告，并直观展示异常指标及调优建议。同时该工具预置多款典型应用场景(如数据库、web、中间件等)的经验优化参数，用户可通过一键式设置，即可提升应用性能。

authd：RFC 1413 ident 协议的守护进程

将 MD5 加密算法更改为 SHA256 算法，增强工具安全性。

authselect：管理系统授权

使用新 popt 库方法，优化内存处理模块，增强工具安全性。

nvme-cli：存储驱动程序工具库

新增两个 RPC 命令 bdev_nvme_cuse_register 与 bdev_nvme_cuse_unregister，用于创建和注销 CUSE 设备。

nvmetcli：存储设备工具

增加 modprobe 对 kmod lib 的支持；

修复保存报告时名称错误问题；

修复 Test_invalid_input 在 py3 失败问题。

hikptool：鲲鹏芯片带内信息收集工具

用于鲲鹏芯片上增强带内信息收集和提升问题定位能力的 linux 用户态工具，支持 SAS、SERDES、CXL、SATA、RoH、RoCE、SoCIP、NIC、PCIE 模块信息查询功能。

Nagios：网络监控工具

支持通过 systemd 并行启动多个 Nagios 服务。

3.2.3 组件依赖关系优化

操作系统之所以成为系统，最主要的技术基础就是依赖关系。在每一个 rpm 包中都有依赖关系的配置信息。我们可以使用 rpm 命令查询出来。根据常见包的依赖关系情况推断，在系统中依赖配置的数量至少应该是五倍的软件包数量。依赖关系可以分为安装依赖和编译依赖。 依赖关系不是线性的，是有交叉、有相互依赖的。 对于一个确定的包，它的多个依赖配置对于它的重要程度是不同的，有的是必须的，有的是可配置的。上面这些因素决定了，操作系统的依赖关系维护和优化是一项复杂艰巨的工程。

Linux 开源软件的依赖关系网的形成是长期演化的结果。银河麒麟高级服务器操作系统 V10 总体上继承了社区形成的比较完备的依赖关系。并在如下方面的做了优化：

3.2.3.1 过依赖问题解决

软件编译时，为了增加某一项功能，经常需要增加需要的依赖。这些社区增加的依赖项配置，有些不符合麒麟的产品定位要求，所以我们要移除这种依赖。比如，

- ceph:将依赖 selinux-policy-base 改为 selinux-policy-minimum
- git:分离 git-core 子包，将 git 依赖最小化
- gnulib：更改编译依赖 java 的版本，从 1.3 改为 1.8
- initial-setup：删除对 python3-libreport 的过时依赖
- lightdm：删除对 pam_console.so 的依赖。

3.2.3.2 依赖配置不当导致功能问题或风险

这里主要介绍，依赖配置项设置不合理导致问题的情况。 比如

- sendmail：添加 procmail 安装依赖，修复发送邮件时缺少 procmail 命令。
- dovecot：添加安装依赖 tar，解决执行 dovecot-sysreport 命令报错

- libpfm：修复 libpfm-devel 包安装没有依赖 libpfm 的问题
- openmpi：修改编译依赖 java-devel 改为 java-1.8.0-openjdk-devel 解决编译问题。
- lightdm：修复缺少 systemd-pam 依赖导致 lightdm.service 启动有报错，且导致 lightdm 无法启动 user 级别的 dbus.service，最终导致用户登录桌面后出现 xfce-polkit 启动报错。

3.3 系统问题修复情况说明

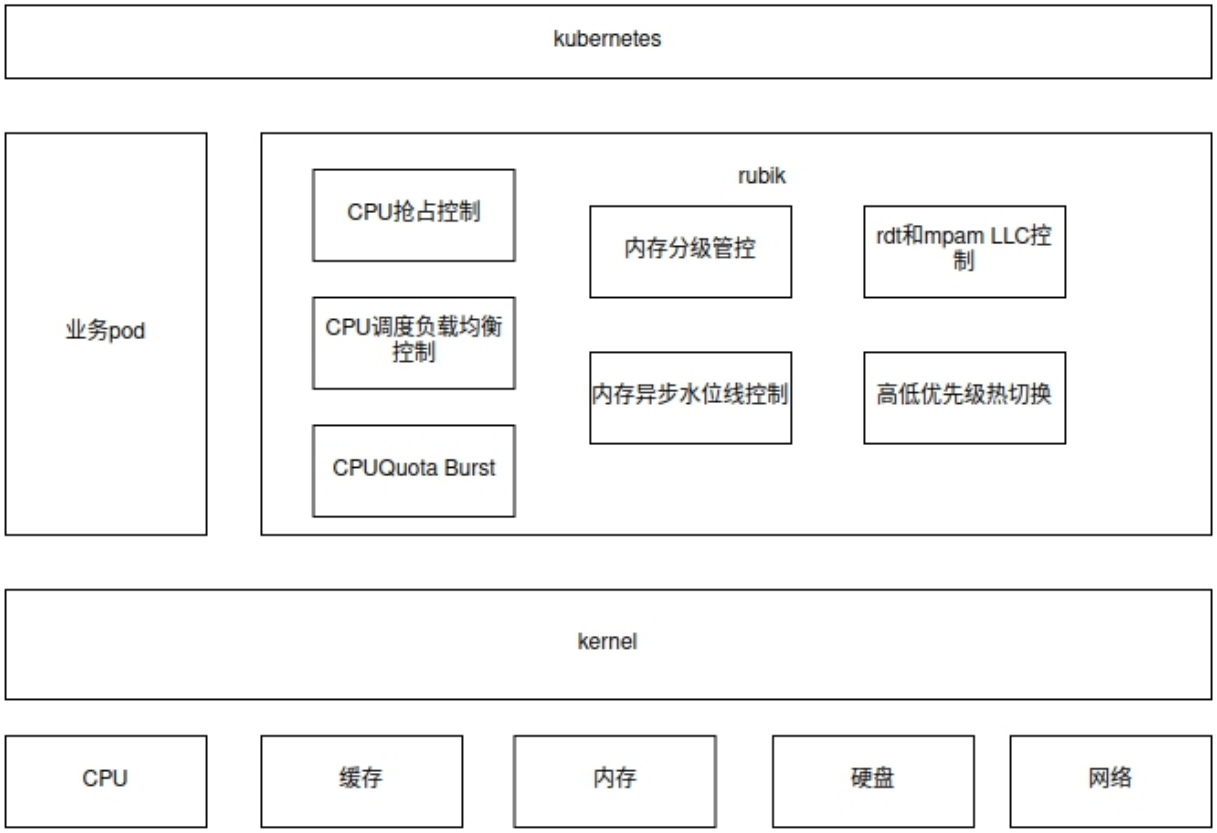
版本迭代过程中保障系统的稳定是非常重要的。源码变动对版本稳定有不利的一方面。配套的测试验证工作是复杂的、成本高昂的过程，任何的疏漏都会埋下问题隐患。并且，概率性问题经常发生，可能是因为线程执行条件不满足，也可能是问题对资源动态敏感，导致验证测试对概率性问题的覆盖不全。麒麟 V10 SP3 2403 的研发迭代遵循了最小变动的原则，没有必要的补丁是不会合入的，对于合入的改动要分析清楚。

问题修复在麒麟 V10 SP3 2403 的研发工作中占很大一部分。这也是操作系统产品研发的一个特点。其中一部分是社区中报告并修复的，如果相关修复对麒麟 V10 的产品特性有帮助，我们会安排引入。引入前会做评审，引入后安排测试验证工作。其他问题是麒麟测试研发迭代中发现并修复的问题。问题数量多、内容杂，具体请见相关发布说明文档，在此不赘述。

4 容器

4.1 在离线混合部署

在离线混合部署是提高资源利用率的有效手段，但是混合部署会导致互相干扰，影响业务 Qos，银河麒麟高级服务器操作系统 V10SP3 2403 提供了针对容器场景在离线混合部署场景的解决方案，架构如下：



支持的功能包括：

- 支持低优先级调整为高优先级；

基于原生多优先级控制逻辑，支持低优先级调整为高优先级的热切换功能，直接通过对应 cgroup 接口 cpu.qos_level 进行修改，无需重建 cgroup，无需重启任务。在 Pod 的优先级不确定时，通过热切换可以快速切换优先级同时避免业务重启导致数据丢失，减少资源浪费。

- 支持 memcg 异步水位线控制;

该特性通过限制混部时离线应用使用的总内存，通过内核提供的 memcg 级的水位线接口，动态调整离线应用内存上限，实现对离线业务的内存压制，从而保障在线业务的服务质量。

- 支持内存 qos 分级管控；

该特性允许高优先级 pod 和低优先级 pod 在形成内存资源抢占时，高优先级 Pod 绝对抢占低优先级 Pod 的资源，低优先级业务优先 OOM kill，从而保障高优先级业务运行。

- 支持 rdt 和 mpam llc 控制能力；

该功能基于 arm64 平台的 MPAM 和 Intel 平台的 RDT 功能,提供针对访存带宽和 LLC(最后一级缓存) 的细粒度隔离，在 Rubik 组件启动时，配置高中低三个组和一个 max 组，每一个组对应一个 LLC 百分比限制和访存带宽百分比限制。离线任务配置对应的分组，该任务对应的 LLC 使用率和访存带宽使用率将被限制在对应比例范围内。

- 支持 CPU Quota burst；

该特性允许 pod 在低于 CPU quota 时累积 burst time，在有突发负载到来时，允许使用容器累积的 burst time，并使 cpu 利用率短暂的突破 quota 限制，保障业务响应。当业务平时使用率较低，但是业务突发负载时又需要保障及时响应时，可使用该特性。

- 支持 CPU 抢占控制；

该特性允许高优先级 pod 和低优先级 pod 在形成 cpu 资源抢占时，高优先级 Pod 绝对抢占低优先级 Pod 的资源，保障高优先级业务运行。当资源抢占较少时，低优先级 Pod 可恢复 cpu 资源。

- 支持 CPU 调度负载均衡控制;

该特性保障资源负载较高的时候，能均衡 CPU 使用率，保障所有 CPU 都能得到均衡调度，不出现部分 CPU 负载过高阻塞任务调度的情况。

4.2 容器根镜像支持

随着产业数字化建设不断推进，云原生需求不断增长，容器作为云原生的关键技术，企业各类应用场景对容器的需求显著提升。相比于虚拟化部署，容器部署的优势在于轻量、高

度自动化和降低运维成本和资源占用。当前云原生场景下的容器镜像生态中，每一个容器镜像都是在一个基础镜像上进行再制作而来，比如 pytorch 镜像，是由 centos 或 ubuntu 等镜像，增加一系列的包和脚本制作而成，而业务镜像又从 pytorch 镜像制作而成。所谓根镜像，指的是一个镜像所依赖的最基础的镜像。在一个集群中，应尽量使用来自同一个根镜像制作而来的容器镜像，这样能尽量减少镜像维护成本和磁盘资源占用。同时为了满足不同厂商的不同业务场景需求，麒麟推出四类根镜像，这四类根镜像遵循标准 OCI 规范，可无缝分发，且足够精简并保障镜像安全性，这四类根镜像分别是：

根镜像名称	platform(基础镜像)	init	minimal	micro
业务场景	传统业务场景。	systemd 等系统管理服务有着强依赖需求的场景。	完成微服务化，客户组件对部署环境有一定要求，需要部署环境提供软件包下载工具的场景。	高度微服务化，对部署环境无要求的场景。
特点	稳定，适用于大多数场景，功能完整	提供完整的系统管理服务	轻量，使用 microdnf 替代传统的包管理器	比 minimal 更轻量，安全性高
大小	基础	大于基础镜像	小于基础镜像	微小，小于 100m
包管理器	具备基础包管理器 rpm，yum，dnf 等	具备基础包管理器 rpm，yum，dnf 等	microdnf	无
基础系统工具	包含大多数基础系统工具	包含大多数基础系统工具和 systemd 调试工具	只包含最基础的系统工具，不包括初始化和 systemd 相关工具，不包含 python 环境	仅由 bash、glibc-common、coreutils 和 kylin 必备软件包如 kylin-release
镜像启动参数	/bin/bash	/sbin/init	/bin/bash	/bin/bash

根据不同的使用场景可以使用不同类型的根镜像来进行二次构建业务容器镜像，这样可以使镜像快速分发，减小底层资源开销，便于快速集成，加快 CICD 效率；裁剪不必要的工具链，减少镜像的攻击面，按周期更新镜像，持续满足 CVE 保障服务，来增强镜像安全性。

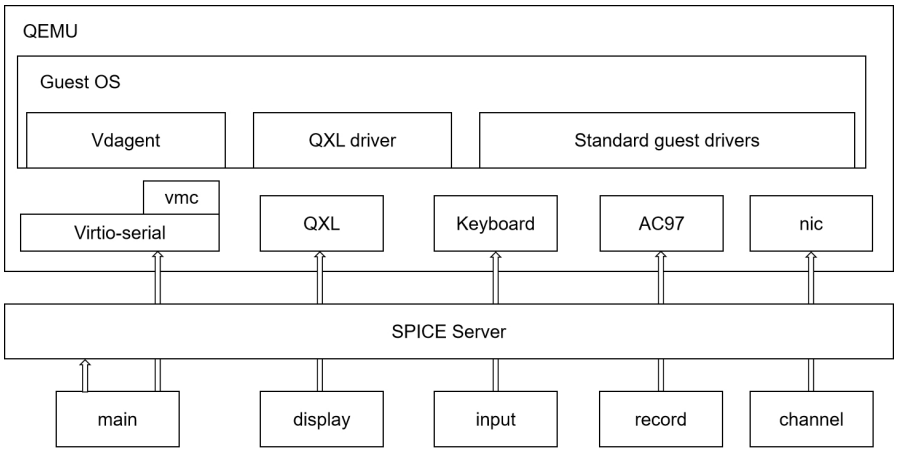
5 虚拟化

5.1 虚拟化图像显示增强

5.1.1 spice、qxl 以及 h264 编码：

Spice 协议是开源的桌面虚拟化数据传输协议,提供客户端访问远程机器显示和设备(如键盘、鼠标、音频等...)。Spice 实现了类似于与本地机器交互的用户体验，同时尝试将大部分密集型 CPU 和 GPU 任务转移给客户端。

如下图所示，Spice 协议是构建起整个虚拟桌面的核心。借助于 Spice 协议，虚拟云桌面各个组件之间才能够顺利交互。



Spice Server 支持的 QXL 设备，负责为虚拟机提供虚拟显卡，包括初始化显卡的 ROM 与 RAM 等的地址映射、IO 端口映射、显存区域更新、光标位置通知、设备 IRQ 请求、显卡模式设置及重置等显卡基本功能。同时，QXL 设备后端，负责与 Spice Server 交互，从而实现在终端实时进行虚拟桌面的显示。

此次版本中，spice 协议在 x86_64 及 aarch64 平台全面回归，满足用户对 spice 显示协议的使用需求。Spice 协议中还新增支持 h264 高级视频编解码的配置，提高在网络传输时的图

像显示质量。另外，在麒麟 aarch64 版本虚拟机中也已支持添加 QXL 显示设备。

5.1.2 virtio gpu 3D 渲染优化：

优化了 qemu 在 virtio-gpu 显卡配置下的客户机图像输出，同时确保在 spice server 启用视频流处理机制时，在客户机中播放视频可以正确触发视频流处理机制，不会出现画面卡顿、抖动、撕裂等问题。

减少不必要的屏幕刷新操作，并优化屏幕刷新及 glReadPixels 处理，提升 glReadPixels 操作效率，提升客户机中显示性能。

5.2 运维和稳定性增强

5.2.1 热迁移使用 multifd 技术

热迁移：在 Linux 系统中，热迁移是允许在不中断虚拟机服务的情况下将运行中的虚拟机从一台物理主机迁移到另一台物理主机

迁移流：在热迁移中，迁移流是指在源主机和目标主机之间传输虚拟机状态的数据流。这个数据流通常包括虚拟机的内存内容、CPU 状态、网络连接状态以及存储状态等信息。迁移流需要在源主机和目标主机之间建立一个可靠的通信通道，以确保数据能够以高速、低延迟地传输。

当只有一个单一的迁移流时，会导致几个问题：

- 处理接收的 CPU，在 10g 和更快的情况下是一个瓶颈；
- 复制所有的迁移页面，即使我们可以直接发送；
- 使透明大页更难以使用。

在 qemu 原有的热迁移 TLS 加密传输模式中，增加了 multifd 多迁移流支持，对迁移过程进行优化。将迁移流分成两个部分，一个用于处理控制信息的发送，其他的则用于 RAM 页面内容的发送。避免了不要的拷贝操作。

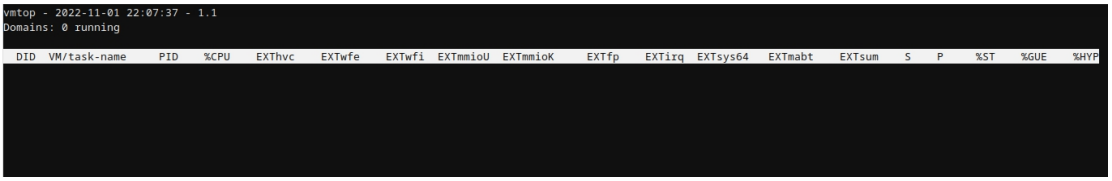
2. 优化 IO 悬挂功能

IO 悬挂功能，当 IO 发生错误时默认自动重试，超时会上报警告。使能该功能后，能够减少因网络抖动导致的网络存储 IO 异常报错问题。

这里增强了 IO 悬挂的参数检查处理，在参数设置异常时，虚拟机将无法正常启动。

3. vmtop 功能优化

vmtop 是运行于主机端的虚拟机监控工具，如下图所示，其能够以动态方式实时查看虚拟机资源使用量，包括 CPU 和内存使用量，以及 vCPU KVM 退出事件。vmtop 为定位虚拟化问题和性能优化提供了极大的便利，是一种能集成多方信息以便监控虚拟机的实用工具。



5.3 飞腾 CPU 虚拟化支持增强

在虚拟化中，对虚拟机的 CPU 模型有特定需求，常常需要对它的特性和拓扑结构进行指定。其中 mode 属性可以用来更容易地配置 guest CPU，使其尽可能接近 host CPU。mode 属性可指定为一下几种：

- custom 模式；
- host-model 模式；
- host-passthrough 模式。

Custom 模式，在这种模式中，cpu 元素描述应该呈现给客户机的 cpu。当没有指定 mode 属性时，这是默认值。这种模式使得持久客户机无论在哪个主机上启动，都将看到相同的硬件。

host-model 模式，本质上是将主机 CPU 定义从功能 XML 复制到域 XML 的捷径。由于

CPU 定义是在启动域之前复制的，因此可以在不同的主机上使用完全相同的 XML，同时仍然提供每个主机支持的最佳 guest CPU。

host-passthrough 模式，在这种模式下，guest 可见的 CPU 应该与 host CPU 完全相同。

这里主要补全 vCPU 模型及拓扑：vCPU 的 custom 模式中新增支持飞腾 2000+、腾云 S2500 、飞腾 S5000C CPU 的定义，并支持其对应的 host-model 模式，方便用户在飞腾平台进行 Guest OS 的 vCPU 配置。

6 安全

6.1 银河麒麟高级服务器操作系统 v10 安全体系介绍

麒麟操作系统通过多方面有效保障系统的安全性，包括增强的身份认证、细粒度的自主访问控制、多策略融合的强制访问控制、主动标记的执行控制机制、管理员分权机制、可信启动、可信度量、安全加固、安全审计、网络安全、数据加密、数据备份、客体重用、沙箱保护机制和国密算法支持等安全机制，提供全新升级的安全中心、文件保护箱和日志查看器等安全套件，可有效防止病毒、木马入侵，加强了数据的安全性保护，提高了系统运行环境的安全性以及保证了违规操作的不可抵赖性。

6.1.1 身份认证

身份认证是银河麒麟操作系统的主要安全机制之一，身份认证必须准确鉴别用户的身份，以便为访问控制和安全审计打下坚实的基础。银河麒麟操作系统提供用户 UID 唯一性保护功能，确保系统遗留用户的文件的安全隔离；提供用户密码强度的检查方案、多种密码算法支持，为系统账户安全提供安全保障。

银河麒麟操作系统实现了用户 UID 唯一性保护功能,即一个 UID 只能被一个用户使用，

即使该用户被删除后，该 UID 也不能再次被其他用户使用，可以有效避免该用户的遗留文件被非法访问，确保用户数据隔离和禁止客体重用。

银河麒麟操作系统提供账户安全增强检查方案，防止用户账户信息被暴力破解。提供如下方面的账号和密码检查：

- 账户有效期
- 账户到期前提醒
- 密码使用期限
- 密码重复限制
- 密码错误次数
- 密码强度配置
- 账户锁定配置

银河麒麟操作系统提供了多种算法，为用户密码保密方式提供多种选择，支持 SM3 国家商用密码算法，确保在不同的场景支持不同的算法保密方式。提供有如下算法：

- SHA256 算法
- SHA512 算法
- SM3 算法

6.1.2 自主访问控制

自主访问控制是一种由客体的拥有者管理的“谁能访问什么”的控制机构。客体拥有者能够有选择地授予其他用户某些访问权限。传统的 UNIX 文件权限模式是一种控制粒度很粗的自主访问控制（Discretionary Access Control，简称 DAC）机制。它将用户分为文件属主、同组用户和其它三类，分别指定读、写、执行权限。文件的拥有者无法指定文件只被某个

具体的用户或某几个具体的用户所组成的组所访问。

银河麒麟操作系统遵循“Posix 1003.1e”标准，设计实现了基于访问控制列表的细粒度自主访问控制策略。访问控制列表（Access Control List，简称 ACL）是客体的一个自主访问控制实体系统设计，可以决定主体能否访问该客体。操作系统设计并支持二种 ACL 类型，分别是访问 ACL（access ACL）和缺省 ACL（default access ACL）。访问 ACL 可以和文件、目录、连接等文件系统客体关联，控制对它们的访问许可权。缺省 ACL 仅仅和目录关联，在具有缺省 ACL 的目录下创建的客体将把父目录的缺省 ACL 继承为自己的访问 ACL。

银河麒麟操作系统中的 ACL 安全功能主要由内核功能模块、用户开发接口和实用工具三个部分组成。其中，内核功能模块运行于银河麒麟操作系统的内核态，为自主访问控制提供底层支持；用户开发接口为一个静态库或者动态库，为开发人员编写基于 ACL 机制的安全应用提供支持；实用工具是一个命令行工具集合，合法用户可以通过命令行工具设置、获取、验证和删除客体的 ACL 安全属性。

6.1.3 强制访问控制

为了保证系统安全，信息系统必须能够实施满足机密性和完整性需求的信息安全隔离，操作系统安全机制是提供这类安全隔离的基础。然而，目前大多数操作系统缺乏实施安全隔离的强有力的安全措施，导致应用中的安全机制容易被篡改或旁路，使系统安全受到威胁。

为了解决这一问题，银河麒麟服务器操作系统设计实现基于类型增强（TE）的访问控制策略。该策略在强制访问控制框架实现，对系统主客体根据不同的应用目的划分为不同的类型域，并定义类型之间的转换规则和访问控制规则，保障系统数据的安全隔离。

TE 策略为系统中的主客体定义了类型属性，并且在类型定义的基础上定义访问控制规则和类型转换规则。主体访问客体时，策略将根据访问控制规则判断是否允许此次访问发

生；进程改变执行映像时，将根据 TE 策略的类型转换规则，判断进程类型是否发生转换。

6.1.4 KYSEC 安全机制

病毒、木马等非法或恶意代码是通过篡改或替换系统应用程序而达到对系统攻击进而试图进入系统以获取其非法目的，因此利用操作系统安全机制确保系统应用程序和重要数据的完整性可有效防止这种非法或恶意代码给系统安全带来的可能危害。

KYSEC 安全体系是麒麟软件自主研发的一款操作系统安全防护机制，该体系在白名单控制和强制执行控制策略框架下，实现了应用执行控制、应用联网控制、内核模块防卸载保护、进程防杀死保护、文件只读保护等安全防护策略。

6.1.4.1 执行控制机制

银河麒麟操作系统提出了基于标记的执行控制机制；该机制通过对系统应用程序标记，实现对执行程序的识别和行为约束，确保应用程序来源的可靠性和程序本身的完整性。执行控制机制具有程序执行权限控制，提供进程标记和客体标记功能，实现标记继承和转换机制，确保只有完整可靠的进程才允许执行。

银河麒麟操作系统通过软件执行控制和管理模块来确保程序的合法性和完整性。系统通过对应用软件实施签名验证，确保软件来源的合法性和软件包的完整性；对系统文件及合法安装的文件进行主动标记，只有来源可靠且完整的应用程序可执行，确保应用在运行时的完整性。

6.1.4.2 应用联网控制

应用联网控制是对应用程序访问网络权限进行控制。在应用进行联网操作时，控制模块会根据内核应用联网许可来检查该应用程序是否允许联网，若允许则该应用程序可以正常进行联网操作，否则会阻止该应用程序的联网操作。并且会将该应用程序联网操作权限

结果同应用程序对应的进程信息进行记录，在该应用程序（当前进程下）联网时会直接根据策略决定本次是否允许联网。

6.1.4.3 管理员分权机制

在传统的 Unix/Linux 操作系统上,都存在一个超级用户 root,超级用户拥有最高的特权，负责管理操作系统所有的资源，既是各类政策的制定者，又是这些政策的执行者；既是各类资源的管理者，又是这些资源的使用者。集各种权力与一身，虽然能够提高系统的整体运行效率，但是如果被恶意利用的话，系统的权力就容易被窃取并滥用，整个系统的信息就可能泄露，系统也可能遭到破坏。

(1) 最小特权机制

目前主流的计算机操作系统都存在一个无所不能的超级用户。这个超级用户虽然为使用系统带来了极大的方便，但也是一个巨大的安全隐患。为此，人们提出了最小特权的概念，即应用或进程仅具有完成其功能所必需的权限。草案标准 POSIX1003.1e 所定义的权能机制正是实施最小特权的一种有效措施。

银河麒麟操作系统将超级用户的特权抽象为各种各样的权能（CAP），不同的权能代表不同的特权。进程权能控制技术基本遵循 POSIX1003.1e 草案标准，又有所扩展，不但支持赋予进程和文件权能，还引入角色权能，使最小特权管理和操作系统的基于角色定权的强制访问控制框架融为一体，极大提高了系统的可配置性和易用性。另外，利用角色权能的概念，可以方便地将操作系统原来的超级用户划分成多个相互制约的管理员，如系统管理员、安全管理员、审计管理员等，实现管理员分权，从而大大降低超级用户所带来的安全威胁。

银河麒麟操作系统的主体是进程，进程只有具有了权能，才能够代表用户进行特权操

作。权能机制作为强制访问控制的一部分，用户所拥有的权能由用户所关联角色的权能决定，用户与代表用户的进程所拥有的权限由其拥有的权能决定，当用户或进程欲执行某项特权操作时，系统将检查主体是否具有相应的权能，如果具有相应的权能，则允许执行该项操作，否则拒绝。缺省情况下，普通用户和进程不具有任何权能，因此，其权限也是最低的。进程权能控制技术与角色定权技术结合，银河麒麟操作系统方便地实现了管理员分权功能。

利用角色权能、进程权能和文件权能，使得不同用户执行相同的文件可能有不同的权限，同一个用户执行不同的文件也有不同的权限，从而可以明确各个进程运行时的权限，使其仅具有完成其功能所必需的能力，实现最小特权。

银河麒麟操作系统针对系统用户权限进行重新设计，使其既能满足正常系统管理使用，又符合“最小特权”原则，实现了三权分立的管理员分权功能。分别设立了安全管理员、系统管理员、审计管理员。安全管理员统一管理操作系统中与强制访问控制等安全机制有关的事件和信息；系统管理员负责系统资源管理与操作；审计管理员负责系统审计规则制订、审计日志分析等审计相关工作。按职能分割和最小授权原则分别授予它们各自为完成自己所承担任务所需的最小权限，并形成相互制约关系。

6.1.4.4 系统文件防护

银河麒麟操作系统提供三种内置文件保护功能，实现对系统内核模块、应用进程和关键文件的保护，可确保内核模块防卸载保护、进程防杀死保护、系统关键文件的只读保护。

(1) 内核模块防卸载保护

KYSEC 文件防护机制在内核加载启动过程中，由内核初始化防卸载链表，然后由系统服务将核外的防卸载模块列表消息发送到内核中，完成防卸载保护策略的载入，从而在模

块卸载时，禁止关键核心内核模块被非法卸载，实现对内核模块的防卸载保护。

(2) 进程防杀死保护

银河麒麟操作系统进程防杀死保护是系统将禁止该受保护程序进程被杀死。根据系统进程防护策略，实现对受保护进程的放杀死保护，同时也提供对受保护进程意外的保护措施。

(3) 文件防篡改保护

银河麒麟操作系统基于安全标记技术，实现对文件的防篡改保护，可有效防止受保护文件被删除、修改、移动等操作。

6.1.5 系统安全加固

银河麒麟服务器操作系统 V10 在系统安全的基础上，实现对系统配置的安全检查和应用服务加固保护功能。

系统配置加固包括安全服务、内核参数、安全网络、系统命令、系统审计、系统设置、潜在危险等 16 类安全加固检查项，提供等保三级、麒麟安全默认模板，等保三级共 46 个加固项；麒麟安全为麒麟推荐加固模板，共 70 个加固项；提供自定义模板，用户可针对 139 个加固项快速细粒度的自定义并加固，用户也可以根据大类（15 个大类）进行粗粒度的自定义和加固。提供简单易用的图形和命令行交互式工具，方便用户进行操作。

6.1.6 国密算法支持

银河麒麟操作系统提供全栈国密支持，包括 Linux 内核、OpenSSL、libgcrypt、gnulib、gnutls、nettle、rustcrypto、TPM/TCM、OpenSSH、KYSEC 等基础组件。银河麒麟操系统内核模块签名支持国密算法，提供一套通用的密码算法框架，对算法进行统一管理，并提供算法接口供其他模块使用，支持 SM2、SM3 和 SM4 算法，KTLS 也支持国密 SM4 对数据进行

加密；OpenSSL 作为系统最基础的密码学算法库，支持 SM2 密钥对生成、SM2 加解密、SM2 签名验签、SM2 密钥交换、SM3 和 SM4 算法，支持 TLCP 协议，并且通过 engine 支持海光 CPU 国密算法、兆芯 CPU 国密算法、鲲鹏 CPU 国密算法的调用；libgcrypt、nettle 和 rustcrypto 加密算法库，支持国密算法 SM2、SM3、SM4；gnulib 支持 SM3；gnutls 传输层安全协议，支持 SM2、SM3、SM4；PAM 向系统提供认证服务，支持 SM3 算法，对用户口令进行摘要计算，并以摘要值的形式完成用户口令的存储和校验；OpenSSH 是系统常用的远程登录工具，支持国密算法 SM2、SM3、SM4，对版本号协商、密钥和算法协商、请求认证、会话认证、通信过程采用国密算法，同时提供国密和非国密切换命令，一键达到国密配置；KYSEC 主要实现应用来安装控制、应用执行控制、内核模块保护、文件保护、分区加密、文件系统加密和安全通信等安全功能，支持 SM3 算法，对全盘文件进行摘要计算，以摘要值的形式完成文件特征存储及文件完整性的校验。

银河麒麟操作系统还支持可信芯片（如：TPM、TCM）的 SM2、SM3 和 SM4 算法，提供 TPM 和 TCM 软件栈国密开发接口；支持 x86/arm 指令集国密算法加速，提升 50%。

6.1.7 可信计算

银河麒麟操作系统基于 TCM1.0/TCM2.0/TPM2.0/KYEE 提供可信计算的能力，基于飞腾芯片提供双体系架构麒麟可信执行环境(KYEE)与麒麟通用计算环境，并且针对 CPU 的多核架构建立资源隔离和交互机制，计算部件无法访问安全部件的内容，通过把加密密钥等数据存放在安全部件，实现了隐私数据的隔离，只有安全部件和对应的安全应用能够访问，从而进异步保障数据的安全性。

可信计算的“信任链建立”，银河麒麟操作系统基于 TCM1.0/TCM2.0/TPM2.0/KYEE 实现可信启动功能，保护系统在启动过程中关键组件的完整性，建完整可信信任链。并提供简单易用的管理工具，方便用户对关键组件进行配置以及查看信任链状态。

银河麒麟操作系统还提供基于可信根的静态度量功能，可以对系统程序运行时进行完整性度量。提供动态度量功能，根据周期或事件机制对内核、内核模块、进程度量，并根据度量策略和 TPM 状态进行度量扩展，及时发现系统存在的风险；提供命令行交互式工具，方便用户查看和配置。

银河麒麟操作系统还支持鲲鹏、海光等平台上 TPCM 可信 3.0。支持机密计算，适配了海光 4 号 CSV3、鲲鹏 itrustee_tzdriver。

6.1.8 安全启动

银河麒麟操作系统支持 X86、arm、loongarch 平台上安全启动，实现国密和非国密双签名，打通从固件到 shim、grub、内核的签名验证，完成主流固件厂商的适配。

6.1.9 安全中心

安全中心是一款系统安全图形管理工具，包含账户安全、安全体检、病毒防护、网络保护、应用控制与保护等功能。

- 账户安全提供了密码强度配置和账户锁定配置功能，提高系统账户安全性；
- 安全加固提供了系统安全扫描和安全加固，保障系统安全；
- 网络保护提供了应用程序联网控制功能，提高网络访问安全性；
- 应用控制与保护提供了对执行控制、防护的安全配置功能，保障系统运行环境的安全和稳定。
- 安全内存通过安全内存模组上的安全芯片内的安全区，使芯片在处理内存数据流的同时控制使用，排除恶意指令，加强第二重防控。
- 可信度量，BIOS 固件可通过启动度量方式，在 CPU 启动时联动获取最安全的初始状态，保证系统启动阶段安全可信，有效防止固件被篡改，防止受外设 ROM 或第

三方 UEFI 可执行程序实施入侵式安装后门攻击。

- 指令流安全预检测摆脱了对文件、流量、数据、行为等特征的依赖，采用了内存指令控制流检测技术，并与机器学习与人工智能技术深度结合，可从系统的更底层发现漏洞攻击代码的执行，提供了不依赖漏洞及攻击代码的特征的漏洞检测防护能力。

文件保护箱是基于内核级数据隔离机制的保护工具，提供用户间数据隔离和加密保护功能，支持国密算法，实现一箱一密、一文一密的细粒度控制，保障用户数据安全。

6.1.10 文件保护箱

麒麟文件保护箱是基于内核级数据隔离机制的保护工具，提供用户间数据隔离和加密保护功能，支持国密算法，实现一箱一密、一文一密的细粒度控制，保障用户数据安全。

具有如下特点：

多重防护：支持用户间数据隔离以及细粒度的权限控制，保障数据安全。

安全加密：支持一箱一密、一文一密的透明加密机制，且对密钥进行安全管理，能够满足政企和金融级客户的核心安全诉求。

丰富算法：支持标准国际算法、国密算法和硬件级加密算法，能够满足不同安全等级的加密应用场景。

高兼容性：支持保护箱版本兼容机制，用户升级适配无感知，保证用户数据安全存储、永不丢失。

简单易用：支持内置文件管理器，实现统一管理，操作简单、易于上手。

6.2 安全提升

安全体系中，麒麟 V10 SP3 2403 中的新增特性有：

- 安全加固支持通过模板文件对 139 个加固项进行自定义参数设置；

- 国密算法支持，增加 gnutls、nettle、rustcrypto、openssh、tpm/tcm 软件栈
- 账户保护——UID 唯一性保护，增加配置文件，提供增强级保护
- 新增安全启动，并在 grub2 加载验证器等方面做了优化
- 新增基于 TPM2.0/TCM1.0/TCM2.0 的可信启动
- 新增静态度量和动态度量
- 海光安全特性支持
- 新增基于硬件的国密支持，并提供开发接口

6.3 安全漏洞修复

麒麟 V10 SP3 2403 研发阶段共修复内核安全漏洞 278 个、核外安全漏洞 1010 个，其中包含 338 个重要和严重高危级别的漏洞。比如，apache 的允许选择加载主机密钥 SSH 服务器的 CVE-2022-45047， DNS 协议方面的 CVE-2022-2906 等共 5 个安全漏洞、objdump 工具中的 compare_symbols 函数可以导致拒绝服务 CVE-2022-47696，具体请见发布说明文档的重要的安全漏洞修复表格。

7 银河麒麟高可用集群

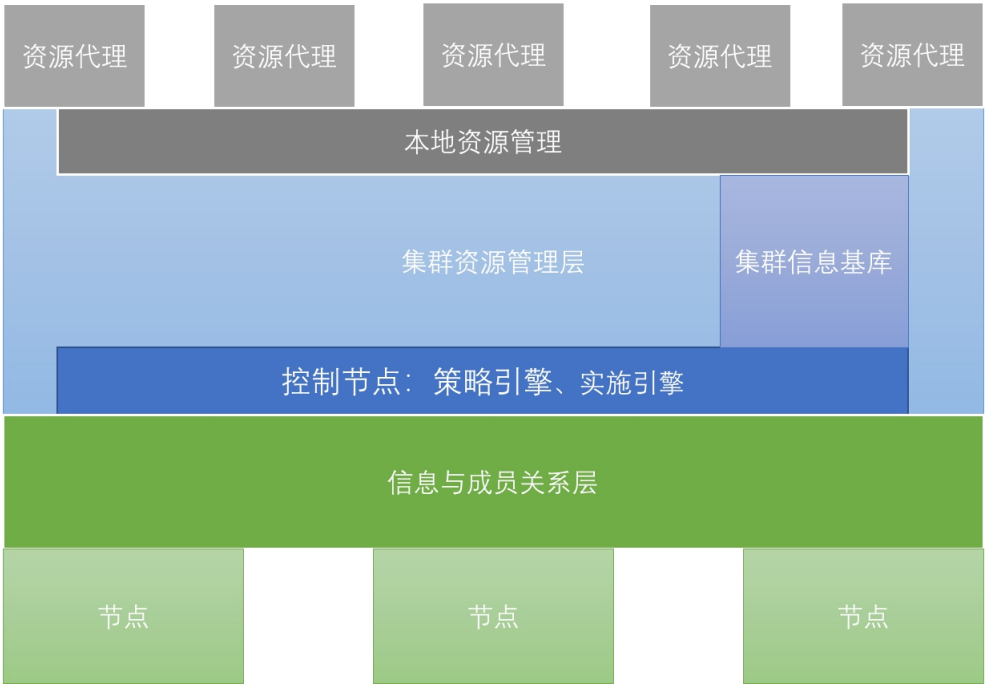
银河麒麟高可用集群软件介绍

银河麒麟高可用集群软件是基于国产银河麒麟高级服务器操作系统开发的高可用性产品，能够给用户提供多种灵活的高可用组合解决方案，保护用户的服务器集群对外提供不间断的运行环境。

7.1.1 软件架构

银河麒麟高可用集群的软件架构分为三层:信息与成员关系(Messaging and Membership)层，集群资源管理(Cluster Resource Manager (CRM))层与资源代理(Resource Agent (RA))

层组成。架构图如下所示：



其中每个层次的详细说明如下：

- 信息与成员关系层

本层又可细分为信息传递层和成员关系层。其中信息传递层提供传递集群信息的机制，可通过单播、组播、广播的方式，实时快速传递信息，传递的内容为高可用集群的集群事务，信息传递层只负责传递信息，不负责信息的计算和比较。成员关系层的作用是控制节点（DC）通过 Cluster Consensus Menbership Service（CCM 或者 CCS）服务，根据 Messaging 层提供的信息产生完整的成员关系。CCM 服务可以承上启下监听底层接受的心跳信息，当监听不到心跳信息的时候会重新计算整个集群的票数和收敛状态信息，并将结果转递给上层，让上层做决定采取怎样的措施。CCM 还可以以本节点做为视角生成各节点状态的拓扑结构概览图，保证该节点在特殊情况下能够采取对应的动作。

- 集群资源管理层

本层用于实现集群服务，集群资源管理层收集信息传递层传递的节点信息，对信息进行计算和比较，并做出相应的动作，如服务的启动、停止和资源转移、资源的定义和资源分配等。它需要借助 Messaging Layer 来实现工作，因此工作在 Messaging Layer 上层。本层包含集群资源管理器 (CRM, cluster Resource Manager)，集群信息基库(CIB，Cluster Infonation Base)，策略引擎（ PE，Policy Engine ），实施引擎（ TE, Transition Engine ），本地资源管理（ LRM，Local Resource Manager ）等组件。下面分别对每个组件进行详细介绍：

集群资源管理器：核心组件，实现资源的分配和管理。它需要借助信息传递层来实现工作，因此工作在信息传递层上层。每个节点上的 CRM 都维护一个 CIB 用来定义资源特定的属性,哪些资源定义在同一个节点上。主节点上的 CRM 被选举为 DC(Designated Coordinator 控制节点，主节点挂掉会选出新的 DC)，成为管理者，它的工作是决策和管理集群中的所有资源。DC 上会额外运行两个进程 PE 和 TE。CRM 会推选出一个用于计算和比较的节点，作为控制节点，其中计算由 PE 实现，计算出结果后的动作控制由 TE 实现。

集群信息基库：XML 格式的配置文件，工作的时候常驻内存。集群的所有信息都会反馈在 CIB 中。在每一个节点上都包含一个 CRM，且每个 CRM 都维护这一个 CIB，只有在主节点上的 CIB 是可以修改的，其他节点上的 CIB 都是从主节点那里复制得到的。

策略引擎：定义资源转移的一整套转移方式，但只做策略，并不亲自来参加资源转移的过程，而是让实施引擎来执行自己的策略。

实施引擎：执行策略引擎做出的策略的并且只有 DC 上才运行 PE 和 TE。

本地资源管理：在每个节点上都有一个 LRM，这是 CRM 的一个子功能，接收 TE 传递过来的事务，在节点上采取相应动作，如运行 RA 脚本等。

- 资源代理层

资源代理层提供能够对集群资源进行管理的脚本，对集群资源的管理操作包括但不限

于启动，停止、重启和查询状态信息等。脚本由 LRM 本地资源管理器负责运行。资源代理分为 OCF、service、systemd、stonith 几类。

7.1.2 工作模式

集群的工作模式支持：

- 主备模式/双机互备模式

主备模式组建的高可用集群通常包含 2 个及以上节点，如果只有两个节点也称之为双机热备，其中一台作为主节点（active），另一台作为备节点（standby）。在这种模式下，一台服务器作为主服务器，正常情况下其承担所有的服务；另外一台服务器作为待机服务器，正常情况下除了监控主服务器的状态，不进行其他的操作。一旦主服务器宕机，待机服务器就接手工作，成为新的主服务器，客户仍然可以使用同样的服务器 IP 地址、服务、数据库及其它内容，保证业务使用不受影响。

- 主主模式

主主模式下，故障节点请求会自动的转移到另外一个正常的节点上或通过负载均衡器在剩余的正常的节点上进行负载均衡，这种模式下集群中的节点通常部署了相同的软件并具有相同的参数配置，同时各服务在这些节点上并行运行。只有两个节点的集群也称之为双机互备，两个节点在正常情况下各自独立运行自己的应用，同时又都作为对方的待机服务器，通过心跳监控对方的状态。一旦某一服务器宕机，另一台服务器就承担所有的服务，这是一种互为冗余的模式。

- N+1/N+M 模式

通过支持多个节点，pacemaker 可以通过允许多个 Active/Passive 集群合并且共享公共备份节点，从而显著降低硬件成本。N+1 模式就是多准备一个额外的备机节点，当集群中某一节点故障后该备机节点会被激活从而接管故障节点的服务。在不同节点安装和配置有不同软件的集群中，即集群中运行多个服务的情况下，该备机节点具备接管任何故障服务的

能力，而如果整个集群中只运行一个服务，则 N+1 模式退变为 Active/Passive 模式。在单个集群运行多种服务的情况下，N+1 模式下仅有一个备机节点可能无法提供充分的冗余，因此集群需要提供 M 个备节点以保证集群在多个服务同时发生故障的情况下仍然具备高可用性，正常情况下除了监控主服务器的状态，不进行其他的操作。一旦运行业务的主机出现故障，备机服务器依据设定的顺序接管业务，成为新的主机，继续对外提供服务。

- N to N 模式

该模式是 Active/Active 模式和 N+M 模式的结合，集群将故障节点的服务和访问请求分散到集群其余的正常节点中，在该模式下的集群中并不需要有 standby 节点的存在，但是需要所有的 Active 节点均有额外的剩余可用资源。当共享存储可用时，每个节点都有可能用于故障转移。Pacemaker 可以运行多个服务副本来分散工作负载。

7.1.3 资源类型及脚本适配

在 Pacemaker 中，资源管理器支持不同种类的资源代理，这些受支持的资源代理包括 OCF、LSB、Upstart 、Systemd、Service、Fencing、Nagios Plugins。最为常用的有 OCF(Open Cluster Framework)资源代理 LSB(Linux Standard Base)资源代理、Systemd 和 Service 资源代理。

银河麒麟高可用集群软件除支持系统常用软件及软件外， 适配国产数据库及中间件，开源软件，构成丰富的生态。

此外支持 bundle/docker，满足 docker 用户的使用需求。

主要功能增强

银河麒麟高可用集群软件 V10SP3 2403 重点新增如下功能：

- 新增 docker-compose 支持
- 新增虚拟机资源迁移支持

- 新增 booth 支持
- 新增 remote、guest 类型节点配置和使用支持
- 新增 ukey 激活、ukey 续保支持
- 新增集群利用率功能支持
- 新增 TAG 标记功能支持
- 新增资源自启动功能支持
- 新增解组解克隆功能支持

8 分布式技术

分布式技术对服务器应用场景非常重要，是资源优化配置的重要手段，是大型项目中存储和算力保障的技术基础。由此诞生了大家熟悉的分布式存储和分布式计算，比较典型的场景就是大数据。微服务也是依赖分布式技术的先进的程序架构设计思想。

分布式系统的主要特点是不同的节点做不同的事情，所以节点中任务进程的管理是分布式技术的核心。zookeeper 是 Apache 研发的分布式服务协调程序，麒麟 V10 升级 zookeeper 至高版本，修复编译报错、修复相关 BUG、SOURCES 中新增配置、优化软件流程等。

分布式存储是重要的应用场景，ceph 分布式服务组件在麒麟 V10 上支持良好。麒麟 V10 SP3 2403 在修复安全漏洞、解决安装问题、调整依赖包、允许自动更新 RADOS 对象、性能优化等面，对 CEPH 进行提升。

etcd 是一个分布式的、可靠的 key-value 存储系统，用于存储分布式系统中的关键数据。内部采用 raft 协议作为一致性算法，基于 Go 语言实现。麒麟 V10 对 etcd 升级解决安全编译问题。

分布式数据库是分布式存储的一个重要方面，麒麟 V10 对 Mariadb、PostgreSQL 均可支

持。其中对 Mariadb 进行了升级，增强了安全性、修复了大量的 CVE 漏洞、解决了空指针使用问题等。对 PostgreSQL，解决了在龙芯架构上的编译问题，保证龙芯架构上 Mariadb 的兼容性，修复任意查询 SQL 注入 CVE 等。