CHINA LINUX KERNEL
中国Linux内核开发者大会

vivo

字节大规模内核版本迁移实践

字节跳动
贺中坤

迁移背景介绍

Why

内核版本在社区支持时间有限
硬件迭代
软件新feature带来优化

What

保证稳定性的无感升级

How

怎么才能做到

通算集群的挑战

- 业务场景复杂
- 混部后单机压力大，触发边界问题
- 内核版本RC后的测试难以覆盖全
- 迁移中途遇见问题回退版本

稳定性要求高，迭代慢

新增智算集群挑战

- 智算逐步超过通算
- 围绕者高速的GPU逐步重构硬件和基础架构
  - 更强大单节点
  - 更高速的互联
  - 硬件迭代速度快

对稳定性和迭代速度有更高的要求

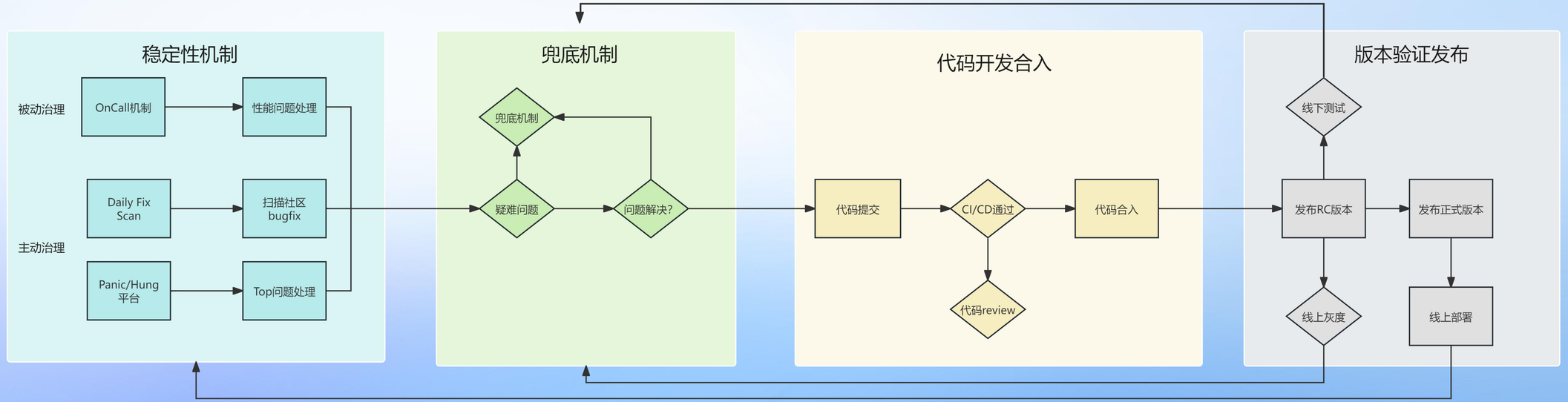
CHINA LINUX KERNEL
中国Linux内核开发者大会

vivo

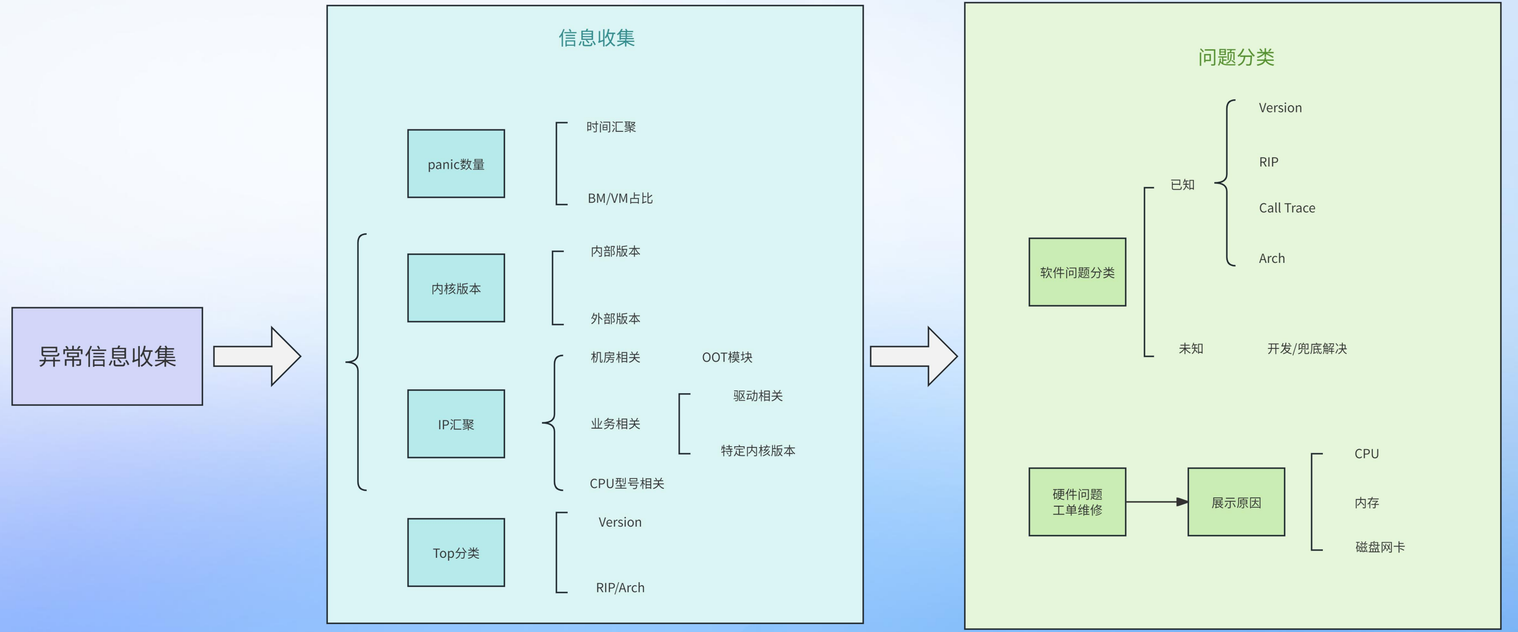
目录

CONTENTS

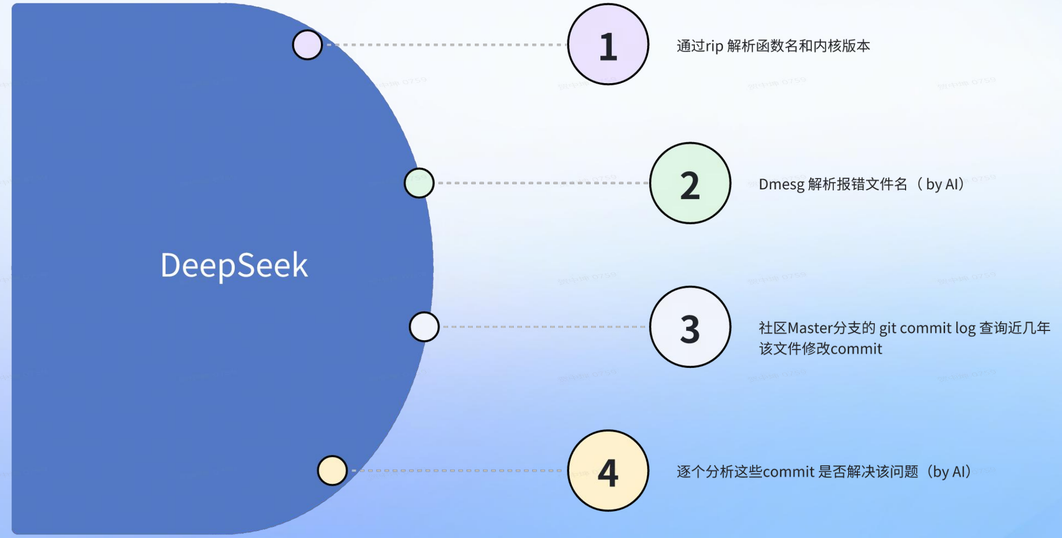
系统概览

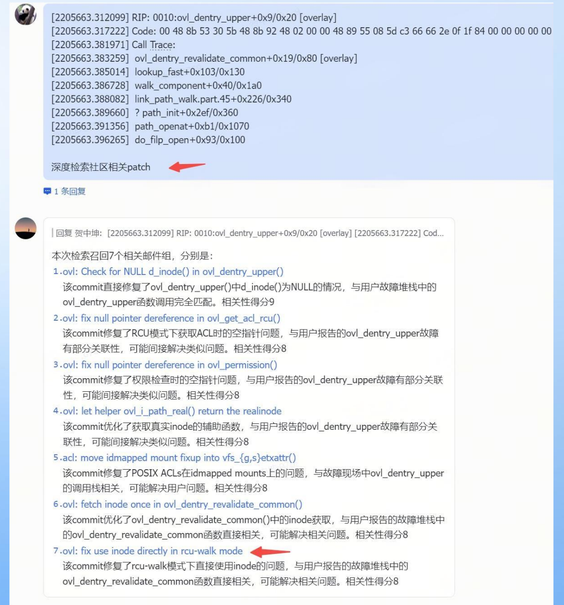

稳定性机制
被动治理
OnCall机制
性能问题处理
主动治理
Daily Fix Scan
扫描社区bugfix
Panic/Hung平台
Top问题处理
兜底机制
兜底机制
疑难问题
问题解决?
代码开发合入
代码提交
CI/CD通过
代码合入
代码review
版本验证发布
线下测试
发布RC版本
发布正式版本
线上灰度
线上部署

稳定性建设-平台

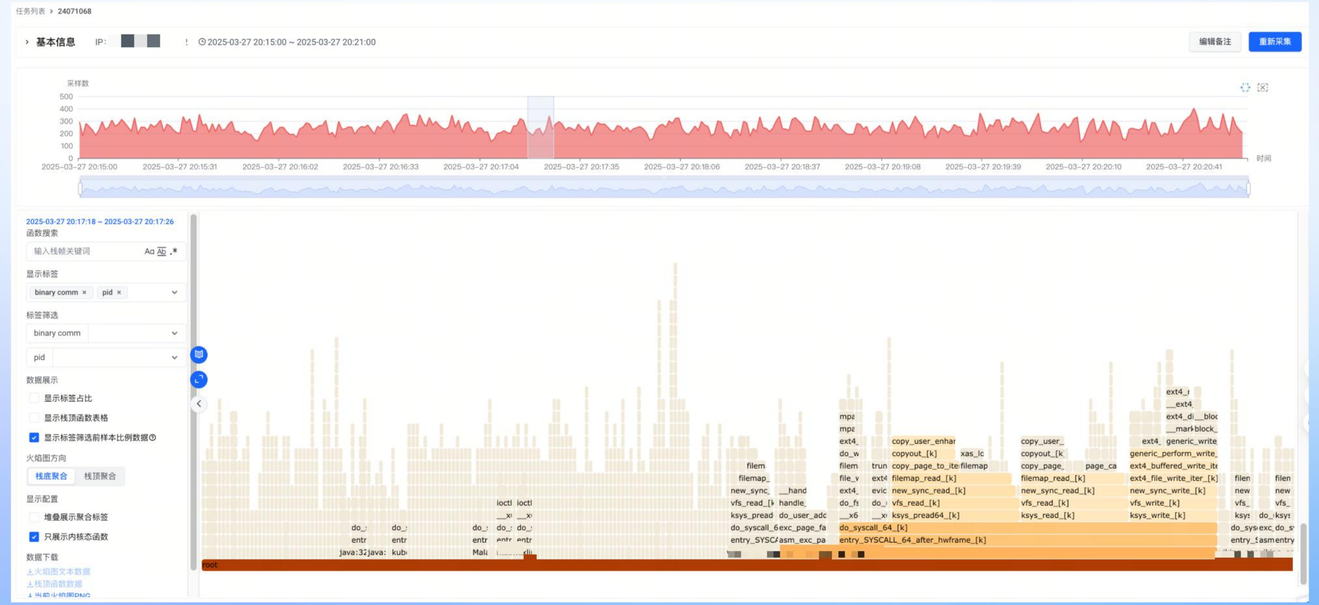

异常信息收集
信息收集
panic数量
时间汇聚
BM/VM占比
内核版本
内部版本
外部版本
IP汇聚
机房相关
OOT模块
业务相关
驱动相关
特定内核版本
CPU型号相关
Top分类
Version
RIP/Arch
问题分类
软件问题分类
已知
Version
RIP
Call Trace
Arch
未知
开发/兜底解决
硬件问题
工单维修
展示原因
CPU
内存
磁盘网卡

稳定性建设-AI检索问题





稳定性建设-debug



perf的缺点是什么？

1：无法常态化采集

2：数据量大

3：无法采集用户态栈

如何debug 偶现的性能抖动问题？

稳定性建设-debug

典型业务场景	机器规格	活跃CPU使用量 (关联CPU占用)	活跃elf数量 (关联内存占用)	CPU占用	内存占用	存储占用
clickhouse独占机器	128C / 1.5T	67C	253	0.13C	2.5G	100M / 天
mysql独占机器	128C / 1.5T	67C	253	0.13C	2.5G	100M / 天
容器场景(单机30 Pod)	128C / 1.5T	67C	253	0.13C	2.5G	100M / 天
容器场景(单机1000 Pod)	128C / 1.5T	67C	253	0.13C	2.5G	100M / 天

持续的profiling

1：开销低

2：包含内核用户态

3：存储占用低

兜底机制

内核技术方向

- 追随社区，回馈社区
- 方案推开源，尽量少做闭源代码
  - 长期技术债减少
- 构建调度/内存/存储/网络等方向专家团队

兜底还得依靠人

- ByteDance 在Linux社区贡献patches 960+
- 多个重要的feature合入Linux社区
- Maintainer/Reviewer 3位

代码开发合入机制


发现问题/需求
社区方案
已解决 — backport
未解决 — 内外开发并行
内部方案
已经解决 — 升级更新内核/hotfix
未解决 — 开发代码
上层问题
优化业务代码 — review业务逻辑
用户态辅助 — 做出demo演示
Fetch-CI
Autoconfig-Build-for-Each-arm64
Autoconfig-Build-for-Each-x86-64
aarch64-buster-kasan-build
X86_64-buster-kasan-build
Release-Note-Generation
Archive-Business-Distributions
1 Prepare
2 Check Patch
3 Pre-Build
4 Build
5 Build Debug
6 Test LTP
7 Release
8 Deploy
Upstream-Fix-Scanner
Upstream-Variance-Viewer
Upstream-Format-Validator
Upstream-Necessity-Validator
Downstream-Format-Validator
MR-Behind-Validator
Compatibility-Validator
Pachset-Integrity-Viewer
X86_64-buster-build
aarch-64k-build
X86_64-buster-debug-test
aarch64-buster-debug-test

代码开发合入机制

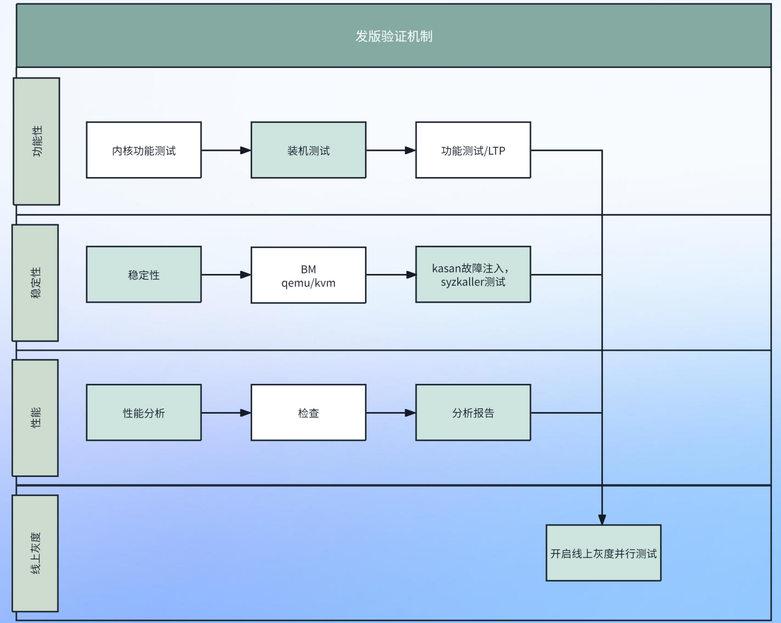
Upstream-Fix-Scanner	扫描MR中的patch是否有上游的相关修复
Upstream-Variance-Viewer	呈现上下游提交的差异，方便审查
Upstream-Necessity-Validator	检验回合bug而没有bug引入patch的情况
Up/downstream-Format-Validator	验证作者日期格式等，防止意外修改
MR-Behind-Validator	验证源分支是否没有明显落后于目标分支

Daily Fix San

- 扫描fix标题的patch
- 使用AI处理不带fix标签的patch扫描，解决patch识别不准确的问题

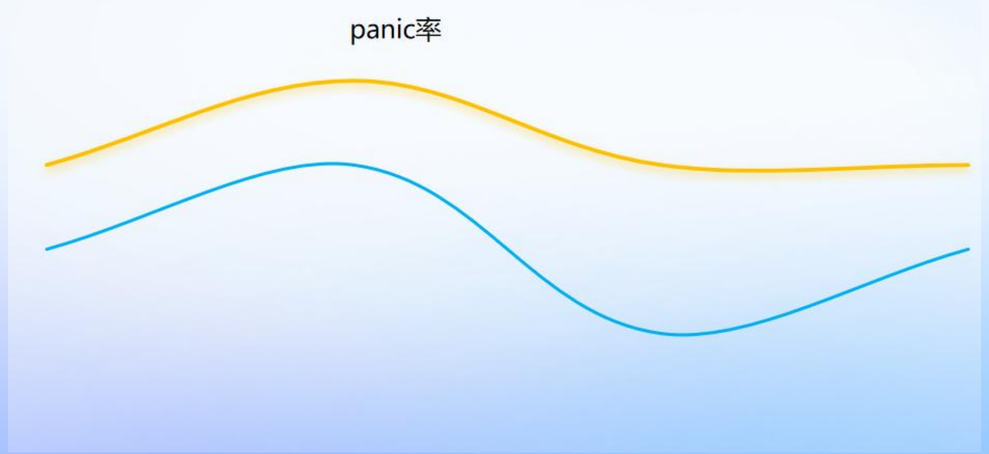
识别出待合入的bufix，再相关领域进行筛选

发版验证机制


发版验证机制
功能性
内核功能测试
装机测试
功能测试/LTP
稳定性
稳定性
BM
qemu/kvm
kasan故障注入，
syzkaller测试
性能
性能分析
检查
分析报告
线上灰度
开启线上灰度并行测试


稳定性建设
hung
panic
性能监控
内存
cpu
irqoff

6 实践成果展示



1：panic降低

2：hungtask降低

3：问题定位效率大幅提升

实践成果-cpu优先级反转

问题现象

- 混部场景
- 低优业务获得cpu时间较少
- 持有全局锁(cgroup)
- 得不到运行后整机卡顿

解决方案

- 限流时标记cfs_rq throttled但并不将它从rq上摘除
- task在返回用户态之前将自己从rq上摘除，这时并不持有内核资源
- 方案已经合入6.18

https://lore.kernel.org/lkml/20250829081120.806-1-ziqianlu@bytedance.com/

实践成果-内存优先级反转

问题现象

- 内存申请路径
- 持有全局资源(jbd2 handle)
- 全局或者cgroup内存不足
- loop in memory reclaiming

解决方案

- add task flag ACCOUNTFORCE
- cgroup内存不足时force charge
- 强制charge部分在退出到用户态之前回收，不持有锁
- 准备推动方案force charge,回退到用户态时recalim和OOM
- 全局OOM场景没有较好的办法

https://lore.kernel.org/all/71a4bbc284048ceb38eaac53dfa1031f92ac52b7.1750234270.git.hezhongkun.hzk@bytedance.com/

问题现象

- 管控程序访问cpu.stat频率高
- cgroup_rstat_flush_locked
- 5.4 rstat延时 250us
- 5.15 rstat延时 7ms?

- cgroup v2关中断时间太长,影响网络收发包

解决方案

社区进行多年优化

- global->sigle memcg flush
- flush context sleep able
- 全局rstat锁拆分到sub sys

- 5.15方案是 mutex global flush, 性能与准确度平衡

问题现象

- 用户态分配Jemalloc/TCMalloc
- 依赖 madvise系统调用释放物理内存
- 用户地址空间大规模膨胀
- PTE 页表占用约为虚拟地址空间大小的 2‰

- 线上发现数十GB的页表

解决方案

- page table相关的基础设施进行了进一步改造[1]
- 实现PTE页表的同步回收[2]
- 为修复其产生的UAF问题，再次对基础设施进行了改造[3]

[1] https://lore.kernel.org/all/cover.1727332572.git.zhengqi.arch@bytedance.com/#r
[2] https://lore.kernel.org/all/cover.1733305182.git.zhengqi.arch@bytedance.com/
[3] https://lore.kernel.org/all/cover.1736317725.git.zhengqi.arch@bytedance.com/#r

CHINA LINUX KERNEL
中国Linux内核开发者大会

vivo

第二十届中国
Linux 内核开发者大会

2025年11月1日　中国·深圳

THANKS

CHINA LINUX KERNEL
中国Linux内核开发者大会

vivo

第二十届中国 Linux 内核开发者大会

2025年11月1日　中国·深圳

赞助单位

vivo　阿里云　HUAWEI　火山引擎

龙芯中科　蚂蚁开源　oppo　腾讯云

支持媒体及社区

51CTO　CSDN　开发者茶思屋 CHASPARK　InfoQ

极星会　思否　阅码场

支持单位

迪捷软件　OpenAnolis 龙蜥社区　OpenCloudOS　清华大学