Kunpeng-V 双层调度技术

范恒龙 华为技术有限公司

openEuler

HUAWEI

作者简介

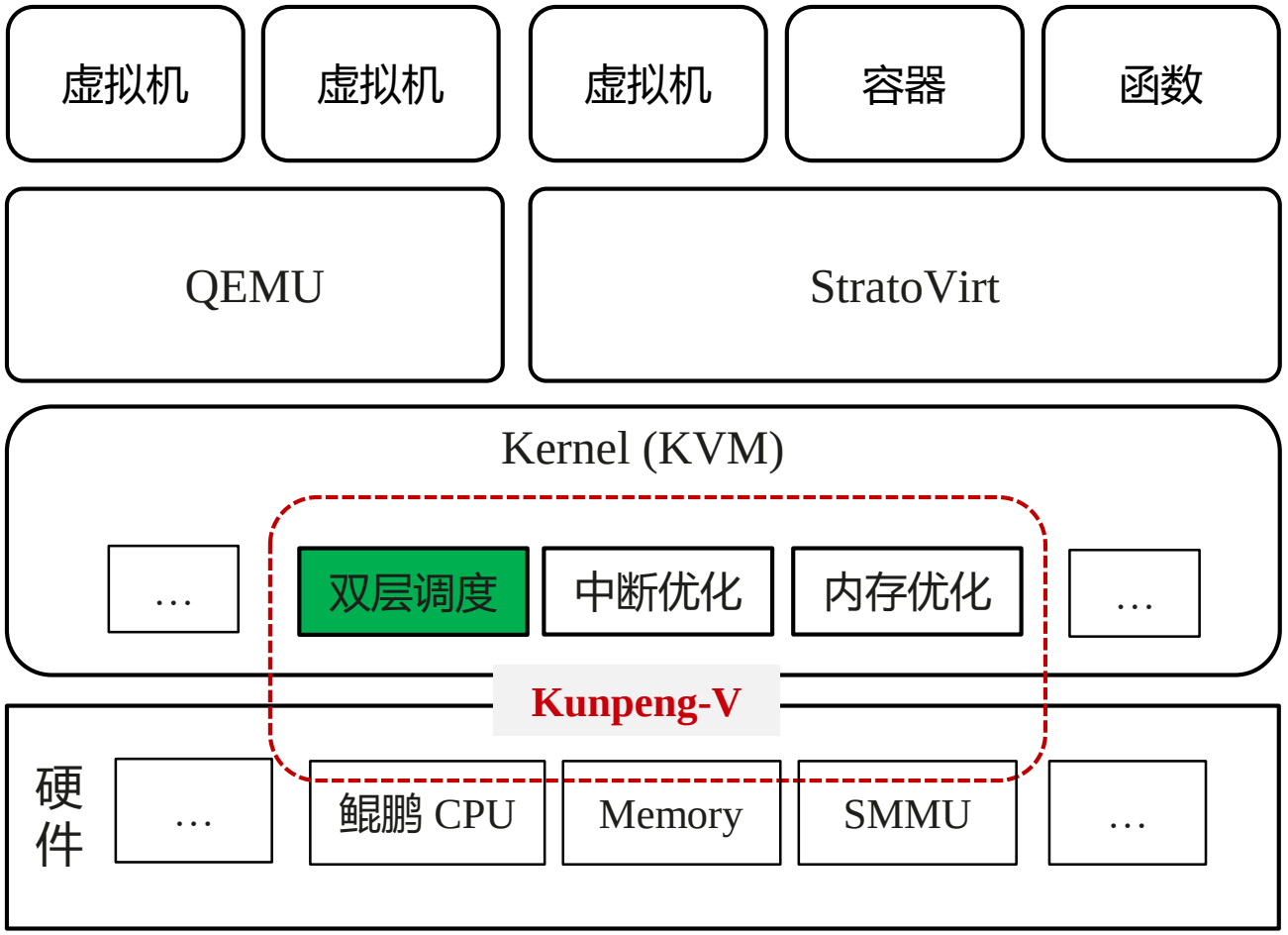
范恒龙：

华为技术有限公司虚拟化技术专家，华为第一代虚拟化版本核心开发工程师，虚拟化运维专家，曾连续负责华为公有云虚拟化多个版本系统设计工作。

从事虚拟化和操作系统技术研究10年，对XEN/KVM软件架构，X86/ARM芯片架构有深入研究，同时对虚拟化各类问题域有丰富的分析和处理经验。

目前投入Kunpeng-V 双层调度技术研究和突破。

Kunpeng-V：基于鲲鹏平台的虚拟化技术体系



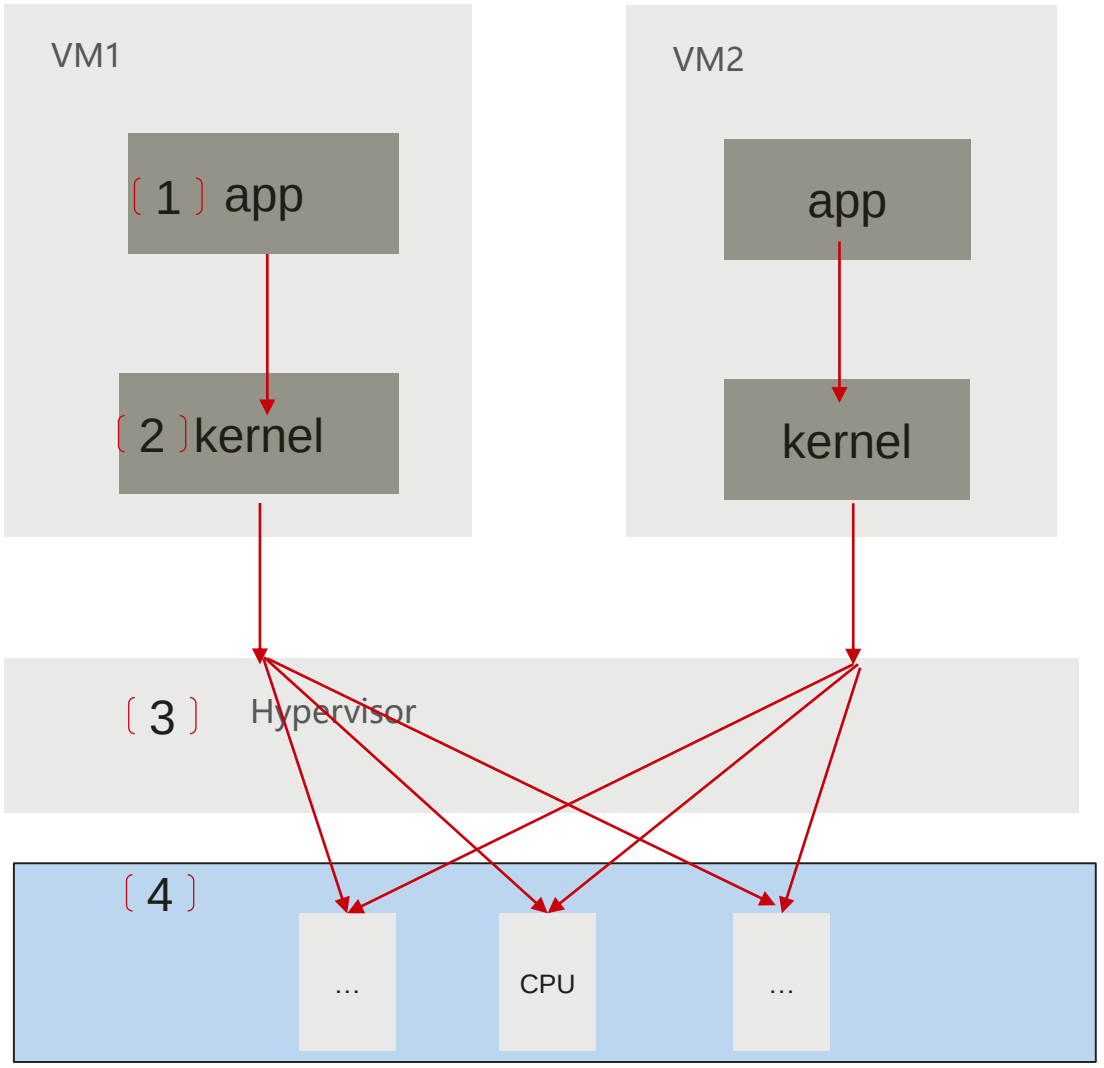
openEuler

HUAWEI

Kunpeng-V双层调度问题背景介绍

典型问题：
超分混部16U、8U、4U虚拟机，通过给虚拟机加压，构造主机cpu80%压力场景，16U虚拟机unixbench跑分低于8U虚拟机

学术界总结问题：
Scaling Guest OS Critical Sections with eCS
Sanidhya Kashyap Changwoo Min†
Taesoo Kim Georgia Institute of Technology
Virginia Tech†
临界区语义GAP：spinlock、rwlock、rcu、interrupts、mutex、rwsem

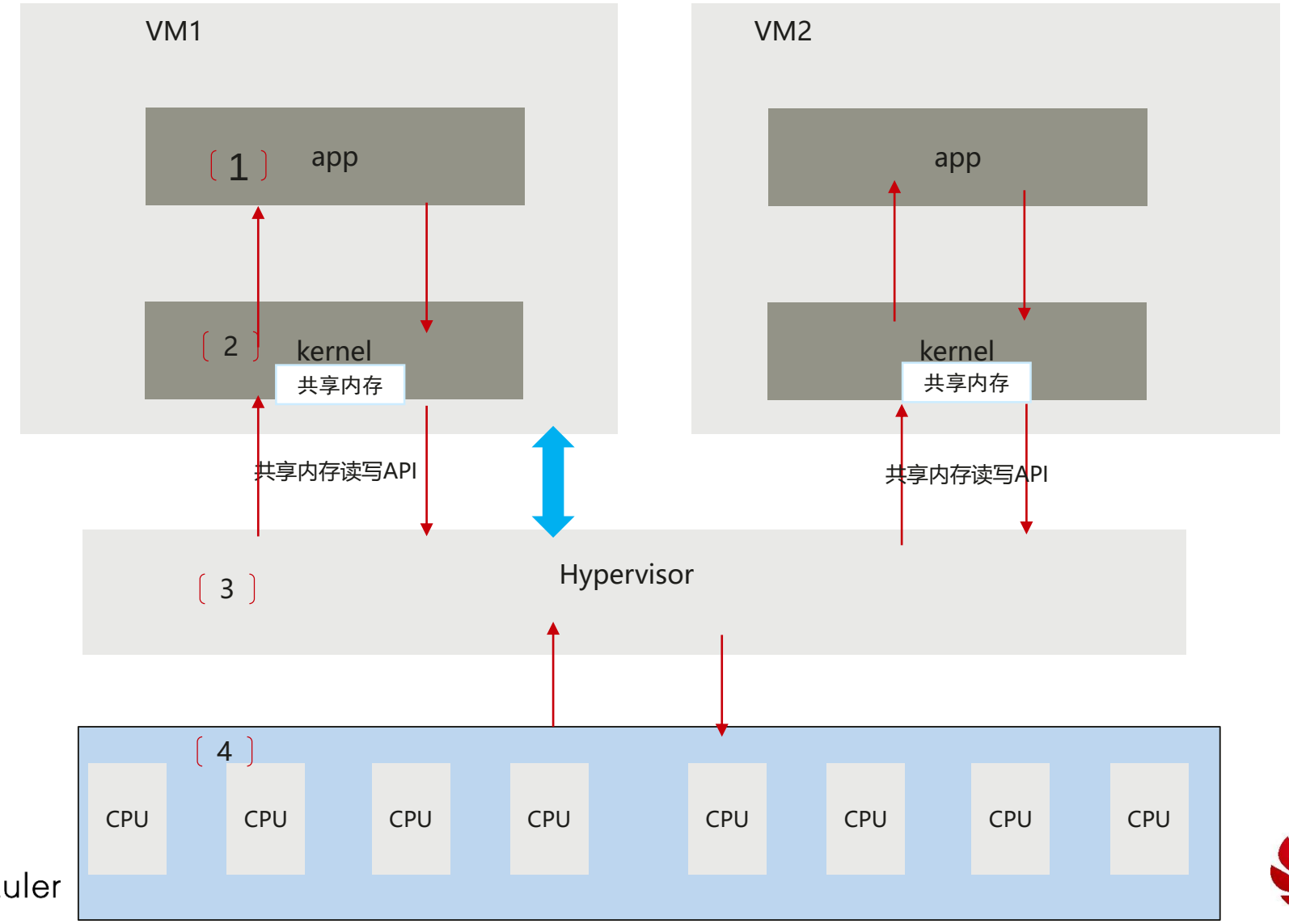


openEuler

HUAWEI

Kunpeng-V双层调度框架

核心思路：不同调度器之间通过共享内存建立通信接口，运行时将所在层产生的敏感信息通过接口传递到下一个调度层
架构组成：**信息传递框架+调度策略**



openEuler

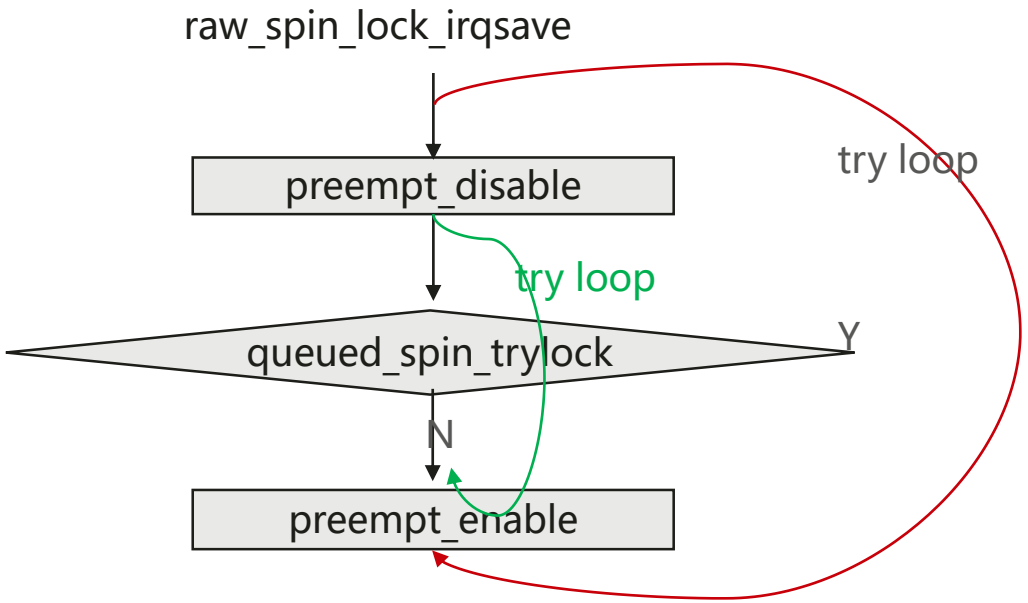
HUAWEI

Kunpeng-V双层调度语义信息

类型ID	信息传递方向	信息传递方案	传递的关键信息
1	hypervisor->vm kernel	共享内存	pcpu负载信息
2	hypervisor->vm kernel	共享内存	smt拓扑信息
3	hypervisor->vm kernel	共享内存	vcpu被抢占信息
4	vm kernel->hypervisor	共享内存	vm kernel对临界区访问次数

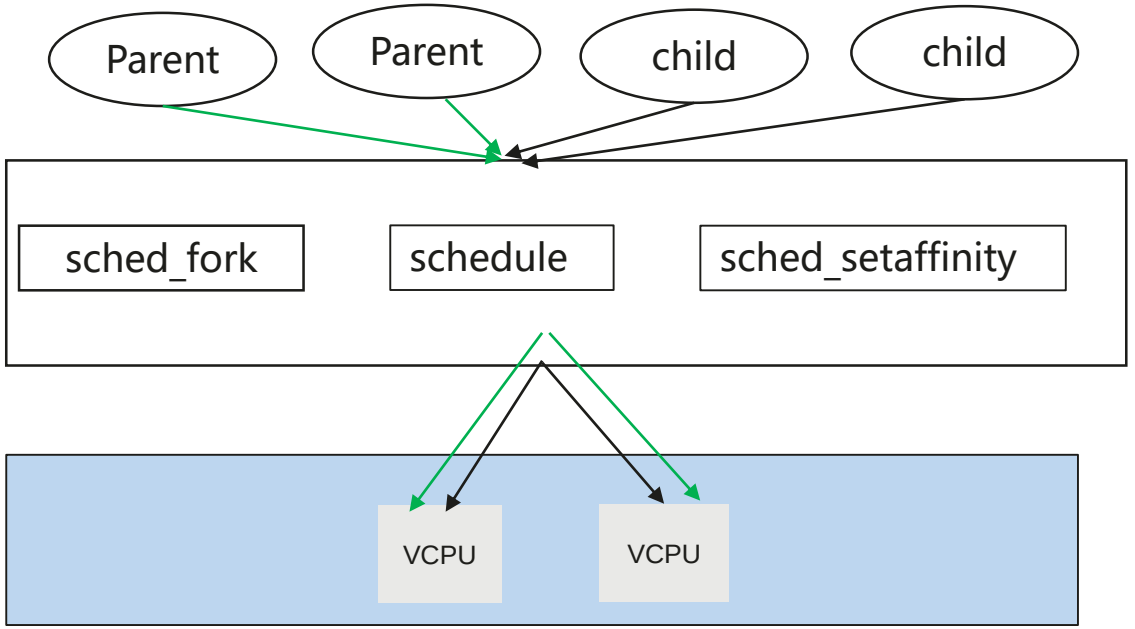
Kunpeng-V双层调度：PCPU负载信息使用策略（1/2）

锁loop机制优化，充分利用CPU资源：try loop优化



mutex_lock slowpath路径,
rwsem_lock block schedule路径均有方案做优化

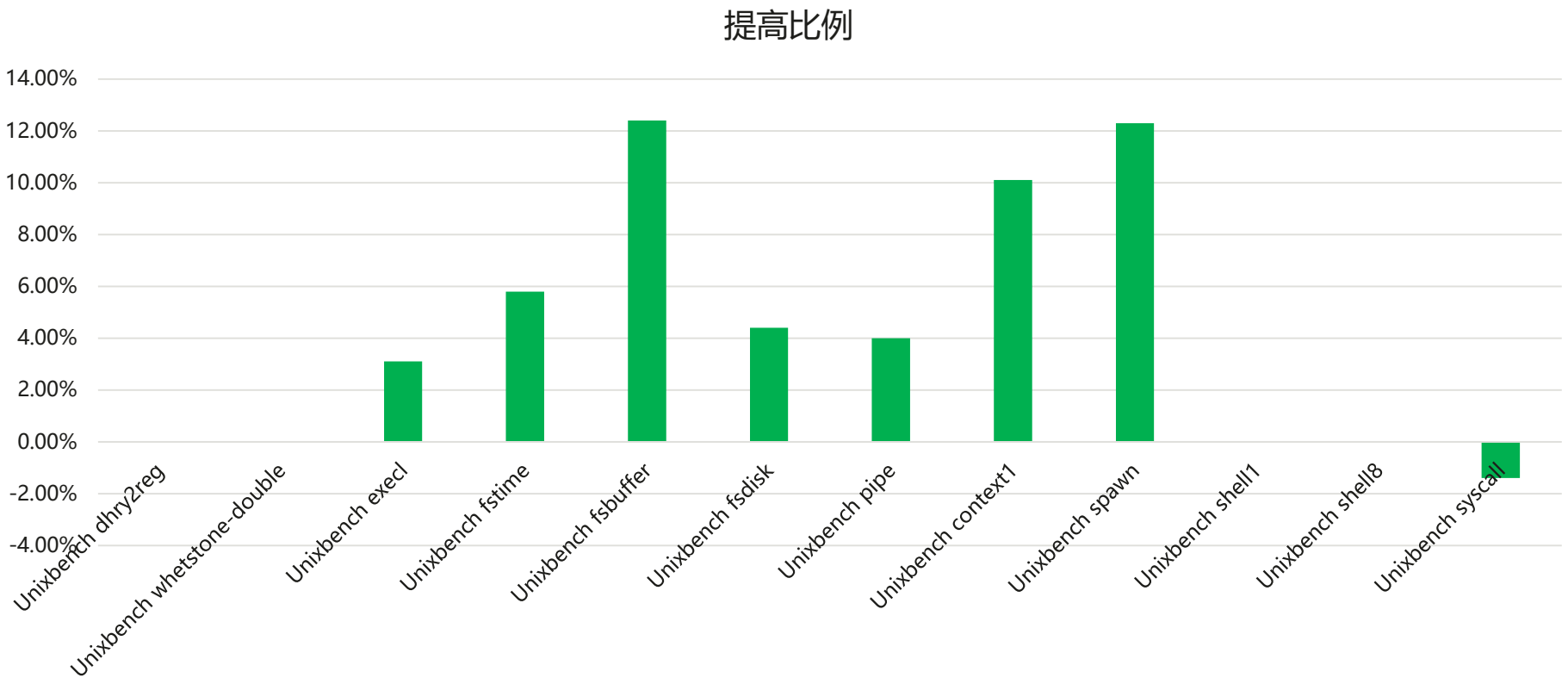
同名持锁父子进程均衡分布



openEuler

HUAWEI

Kunpeng-V双层调度：PCPU负载信息使用策略（2/2）



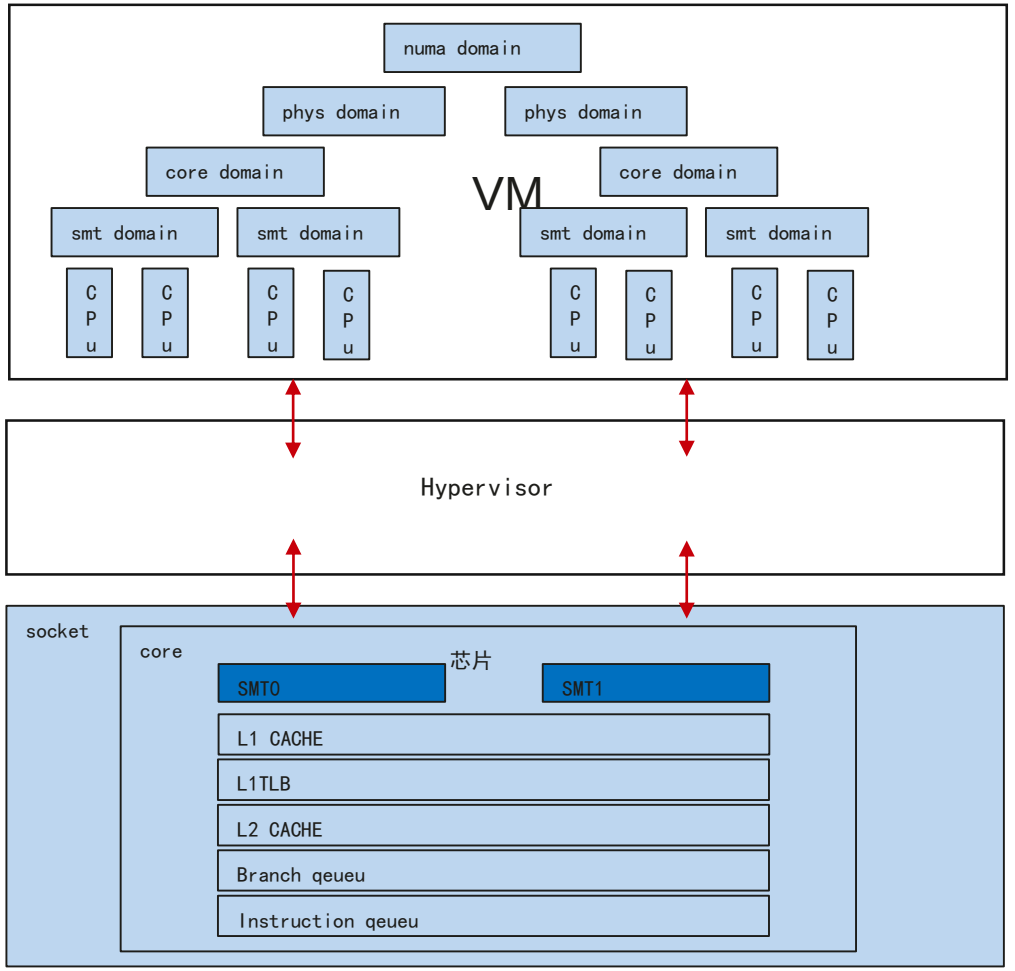
openEuler

HUAWEI

Kunpeng-V双层调度：SMT语义信息使用策略（1/2）

1、通过QEMU 模拟PPTT表，已经给虚拟机呈现超线程拓扑结构，充分使用OS基于SMT的调度策略。

2、通过PMU监控虚拟机的指令特征，做SMT上的业务混合部署。



openEuler

HUAWEI

Kunpeng-V双层调度：SMT拓扑信息策略（2/2）

测试项名称	基线数据-保守值	基线数据-理想值	(spawn,context1)	提升比例
spawn	176.5		206.2	16.83%
context1	93.6		98.3	5.02%

			(spawn,dhry2reg)	提升比例
spawn	176.5		210.7	19.38%
dhry2reg	722.9		774.5	7.14%

			(spawn,execl)	提升比例
spawn	176.5		187.5	6.23%
execl	165.7		217.4	31.20%

			(spawn,fstime)	提升比例
spawn	176.5		194.4	10.14%
fstime	316.7		375.5	18.57%

			(whetstone-double,execl)	提升比例
whetstone-double	215.4		272.2	26.37%
execl	165.7		184.2	11.16%

			(whetstone-double,fstime)	提升比例
whetstone-double	215.4		274.8	27.58%
fstime	316.7		400.3	26.40%

			(context1,execl)	提升比例
context1	93.6		102.1	9.08%
execl	165.7		203.3	22.69%

			(context1,fstime)	提升比例
context1	93.6		135.8	45.09%
fstime	316.7		435	37.35%

			(dhry2reg,execl)	提升比例
dhry2reg	722.9		784.4	8.51%
execl	165.7		205.2	23.84%

			(dhry2reg,fstime)	提升比例
dhry2reg	722.9		1004.3	38.93%
fstime	316.7		419.5	32.46%

			(execl,fstime)	提升比例
execl	165.7		232.3	40.19%
fstime	316.7		372.2	17.52%

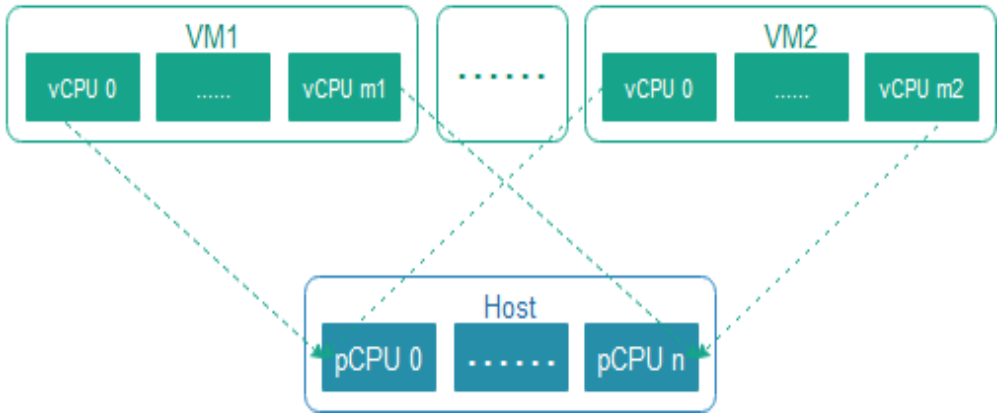
			(fstime,pipe)	提升比例
fstime	316.7		485.7	53.36%
pipe	277.3		278.5	0.43%

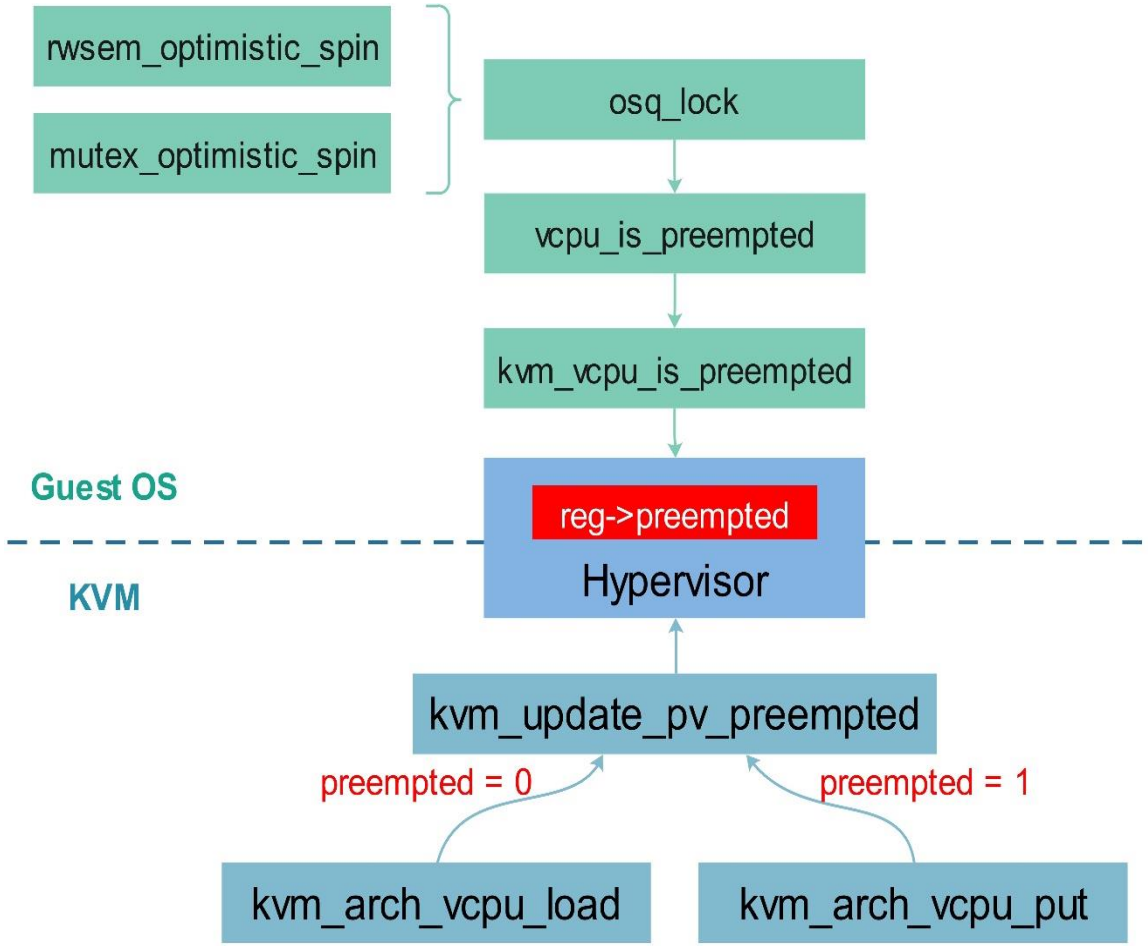
			(fstime,syscall)	提升比例
fstime	316.7		462.3	45.97%
syscall	194.1		204.3	5.26%

openEuler

HUAWEI

Kunpeng-V双层调度：vcpu抢占信息使用策略（1/2）



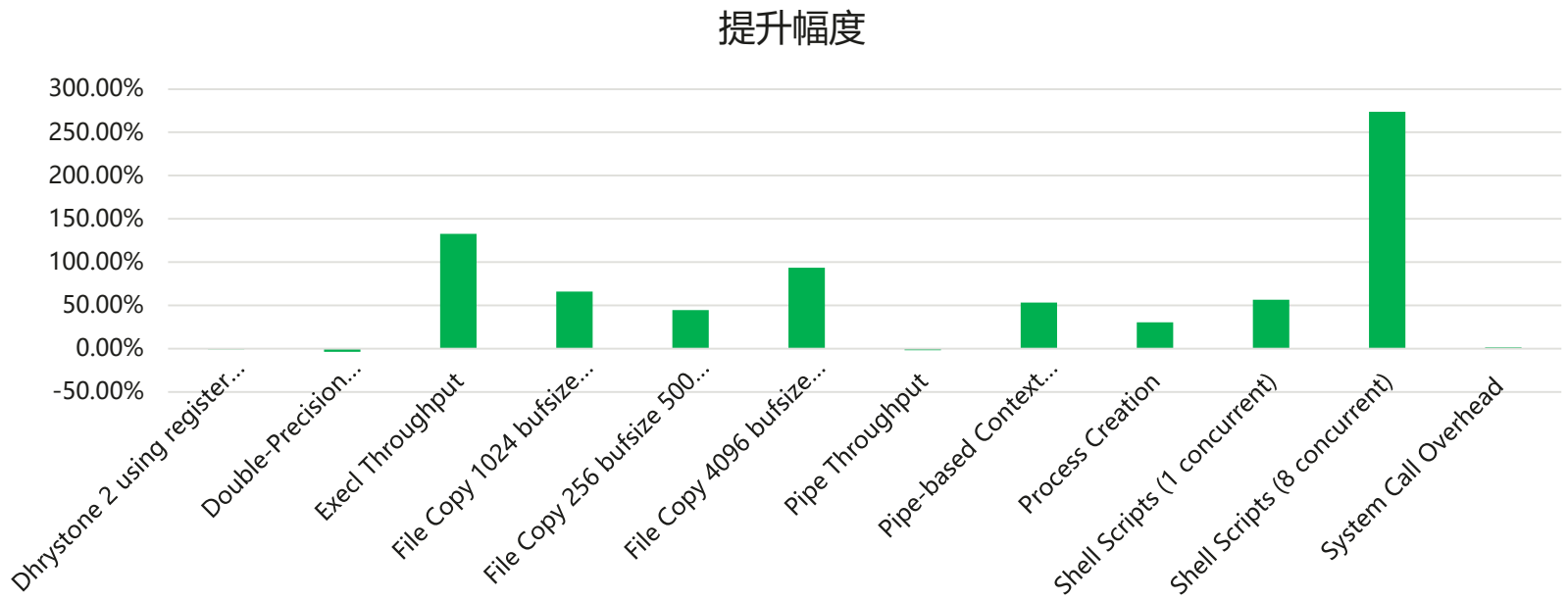


openEuler

HUAWEI

Kunpeng-V双层调度：vcpu抢占信息使用策略（2/2）

测试环境配置：
kunpeng-4826环境上，node1起3台虚拟机，配置24u8g，每个node的上虚拟机cpu范围绑核到该node上全部cpu，性能工具unixbench只在一虚拟机上跑，其余虚拟机全部加压60%压力，压力构造使用stress加压，cpulimit限制加压进程cpu使用率。



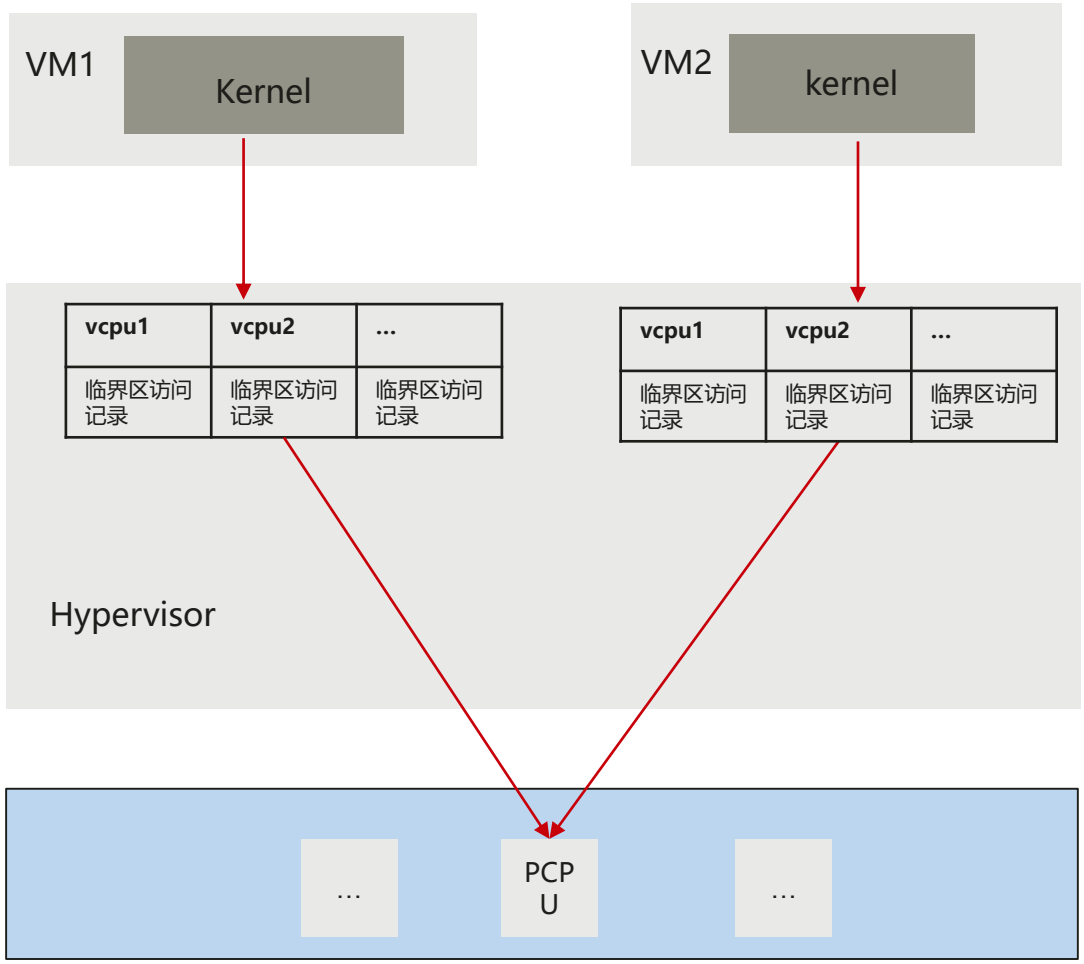
openEuler

HUAWEI

Kunpeng-V双层调度：临界区访问信息使用策略（1/2）

初始化：
1、Hypervisor层每个VCPU都有一个结构体struct pvsched_vcpu_state记录临界区访问的信息。
2、VM内部percpu变量pvsched_vcpu_state在虚拟机内申请的GPN号记录为base
3、vm kernel在启动时初始化
4、hypercall 把虚拟机内部的临界区通信的结构体PFN传递到hypervisor的vcpu结构体成员变量中，用于hypervisor访问临界区通信信息。

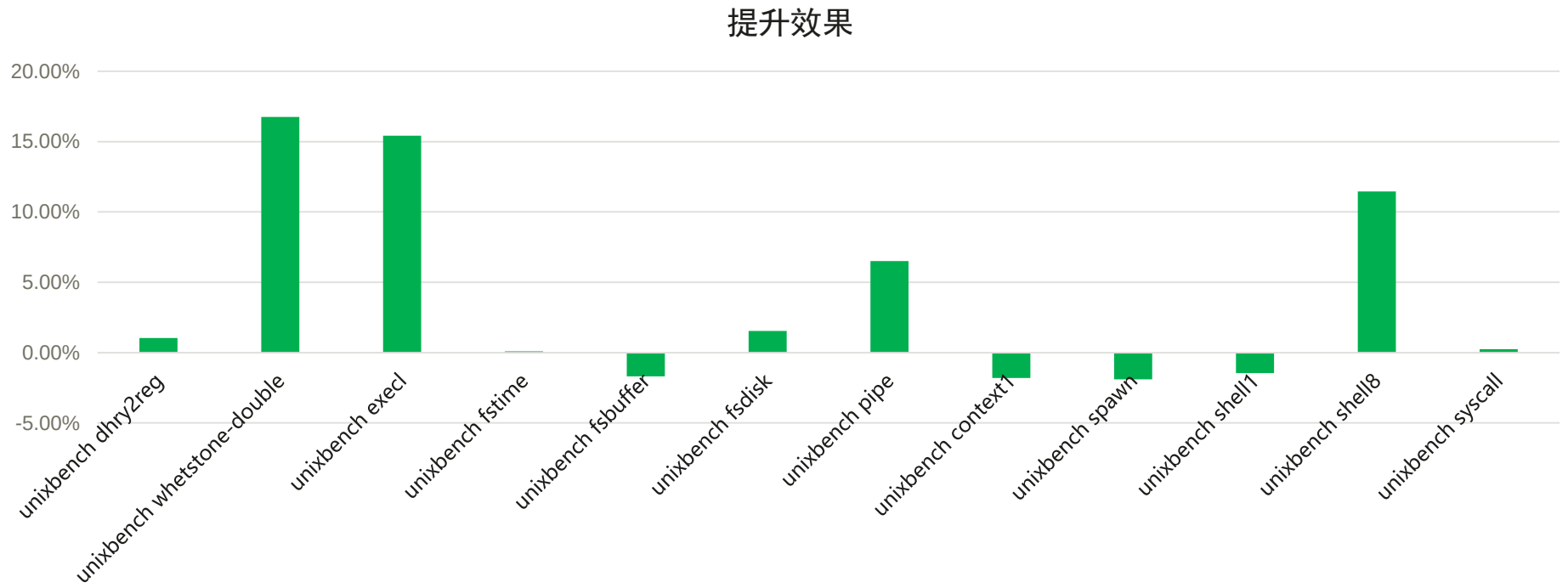
调度：
1、schedule_tick:临界区访问记录大于0的，增加额外时间片
2、wfe trap: 选择临界区访问记录最大值的vcpu，优先调度



openEuler

HUAWEI

Kunpeng-V双层调度：临界区访问信息使用策略（2/2）



openEuler

HUAWEI

Kunpeng-V双层调度：开源计划说明

1、Kunpeng-V 已开源部分代码：
https://openeuler.org/
qemu,kernel/kvm模块

2、正在进行的工作，开源计划在openeuler官网会及时刷新

欢迎关注

社区网站

openEuler

https://openeuler.org

代码托管平台

gitee

https://gitee.com/openeuler

微信交流群

添加小助手微信号“openeuler123”拉你进群

Q & A

openEuler

HUAWEI