中国Linux内核开发者大会

MPAM特性，arm64资源隔离技术的应用和upstream现状

汪少博 华为OS内核实验室
邮箱：bobo.shaobowang@huawei.com

openEuler公众号

openEuler

目录

1 背景及MPAM介绍

2 MPAM软硬实现

3 MPAM和RDT

4 MPAM开发验证

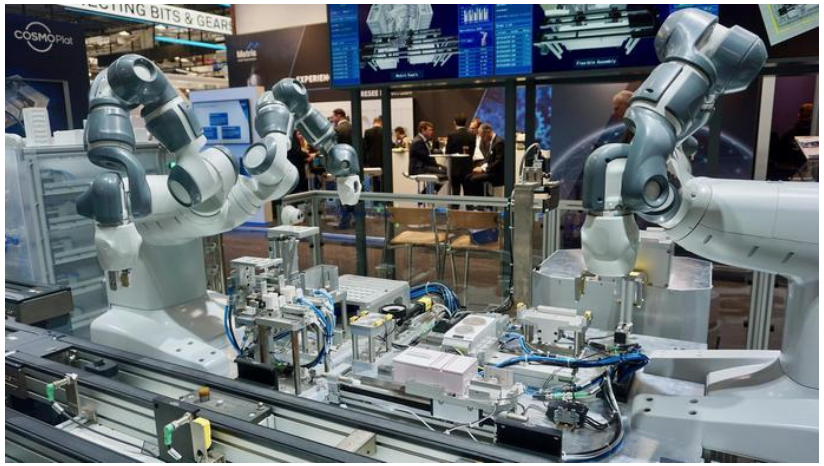
5 资源隔离应用场景

6 MPAM应用扩展

openEuler

背景

各种数据报告显示目前数据中心的机器利用率平均在10%左右，通过混部（Co-location，包括资源隔离等手段），可有效降低30%成本，极大提升资源利用率。[1]



工业机器人以及医学自动化器械等对实时性要求极高，如何保障工业软件的时延是重中之重。

5G工业时代的到来将引发巨大的数据风暴，如何保障数据传输时延是基础；

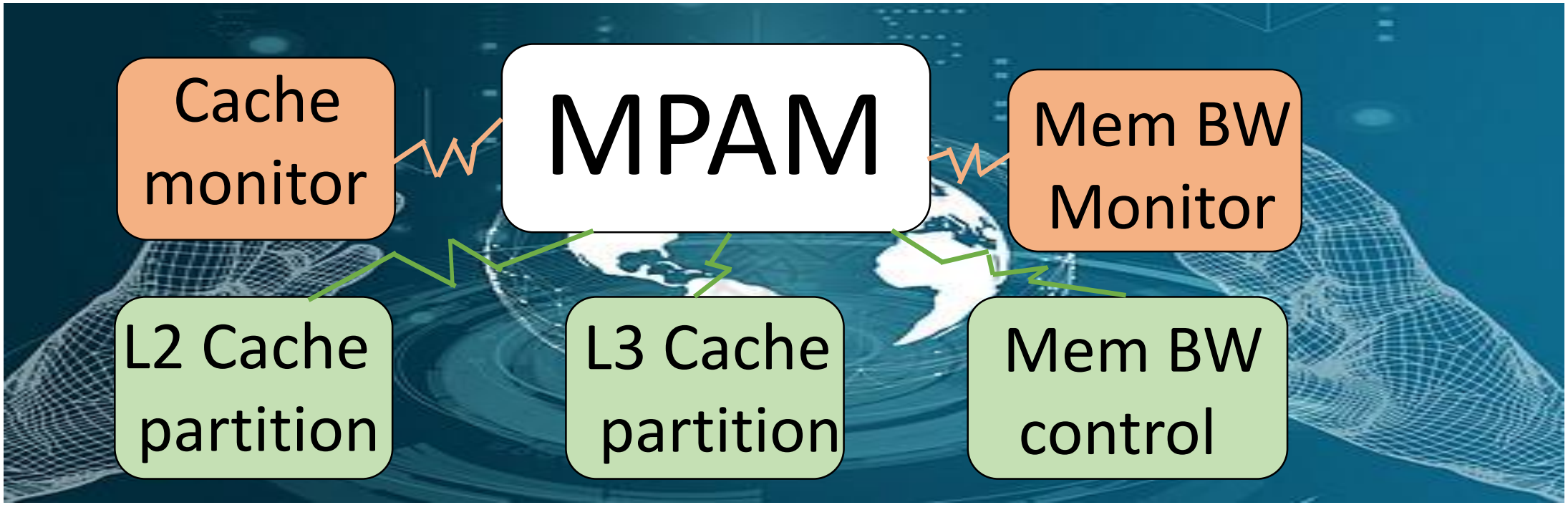
以智能城市理念建造的城市大脑将处理海量数据，如何合理利用有限存储资源业务的资源可以有效降低社会运作成本。

[1] https://myslide.cn/slides/6192

openEuler

MPAM介绍

MPAM（Memory System Resource Partioning and Monitoring）是继Intel x86 RDT特性的另一个在arm64架构下的访存资源隔离和监控特性。

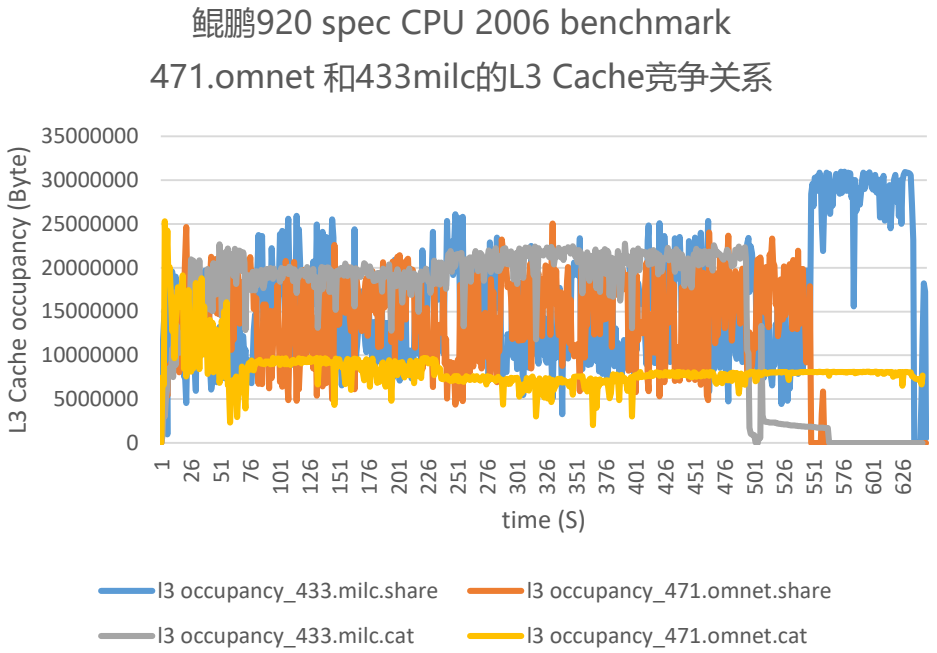


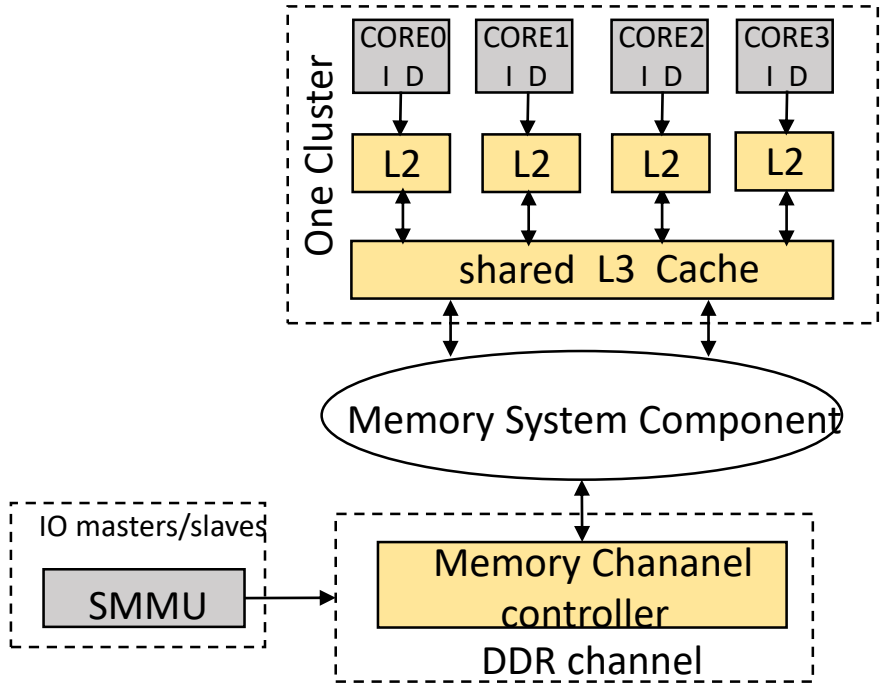
[1] https://myslide.cn/slides/6192

MPAM介绍

现代服务器中某些共享资源（例如L3 Cache）中的干扰，可能导致关键业务性能的急剧退化或系统性能整体下降[1][2][3][4]。

MPAM涉及共享资源包括：L2 Cache，L3 Cache，DMC带宽，通过给CPU/SMMU源头标记partID，使得业务流对用户可见，用户通过配置MSCs完成对业务流的控制。





Memory System Component (MSC)是资源控制的基本单位

[1] Oh M, Choi J, Cho S, et al. Analyzing and modeling the impact of memory latency and bandwidth on application performance[C]//Proceedings of the 33rd Annual ACM Symposium on Applied Computing. 2018: 1095-1101.

[2] Nikas K, Papadopoulou N, Giantsidi D, et al. DICER: Diligent Cache Partitioning for Efficient Workload Consolidation[C]//Proceedings of the 48th International Conference on Parallel Processing. 2019: 1-10.

[3] Tootoonchian A, Panda A, Lan C, et al. Resq: Enabling slos in network function virtualization[C]//15th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 18). 2018: 283-297.

[4] Mancuso R, Dudko R, Betti E, et al. Real-time cache management framework for multi-core architectures[C]//2013 IEEE 19th Real-Time and Embedded Technology and Applications Symposium (RTAS). IEEE, 2013: 45-54.

openEuler

目录

1 背景及MPAM介绍

2 MPAM软硬实现

3 MPAM和RDT

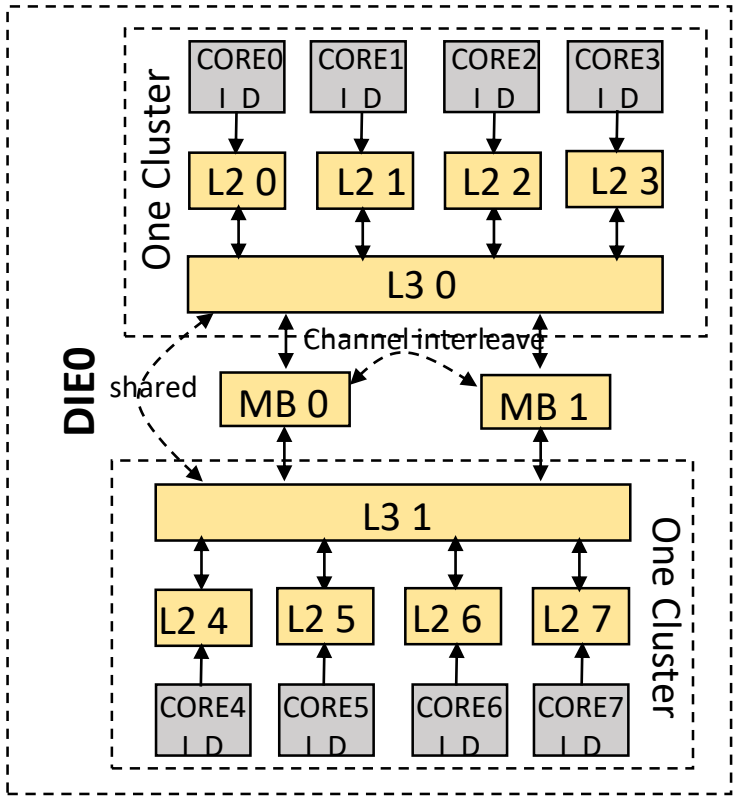
4 MPAM开发验证

5 资源隔离应用场景

6 MPAM应用扩展

MPAM软硬实现

MPAM驱动管理一组MSCs为一个域，每种资源对应的域不尽相同。



L2: 0=xx;1=xx;2=xx;3=xx;4=xx;5=xx;6=xx;7=xx
L3: 0=xx
MB: 0=xx

User interface

e.g. Domains list:
L2: 0 -> 1 -> 2 -> 3 -> 4 -> 5 -> 6 -> 7
L3: 0
MB: 0

Software resource export

Software internal

L2 0	L2 1	L2 2	L2 3	L2 4	L2 5	L2 6	L2 7
L3 0				L3 1			
MB 0				MB 1			
DIE0 MSCs ioremap							

每种资源可能支持的功能不一样
鲲鹏920服务器支持了：
L3的cpbm；MB的max，hardlimit
下个版本将会对Capabilities做全面升级

Capabilities probe

cpbm	cmax	priority
L2 Cap		

cpbm	cmax	priority
L3 Cap		

pbm	max	priority	hardlimit
MB Cap			

MPAM ACPI规定了如何管理组织各个资源的MSCs。

MPAM arm官方手册：https://developer.arm.com/documentation/ddi0598/latest

openEuler

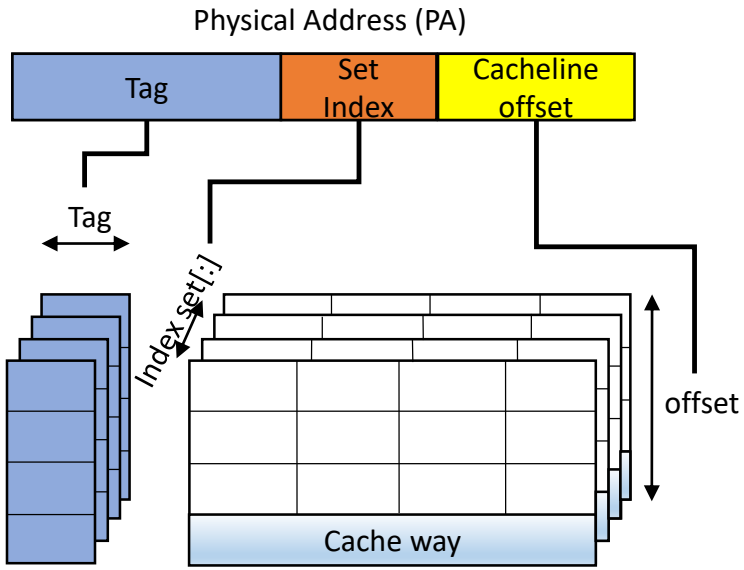
MPAM软硬实现

- 上游通过CORE/SMMU为访存请求标记PARTID和PMG
- 下游配置PARTID和PMG完成资源配置及监控
- vPARTID到PARTID的映射在虚拟化hypervisor中进行管理

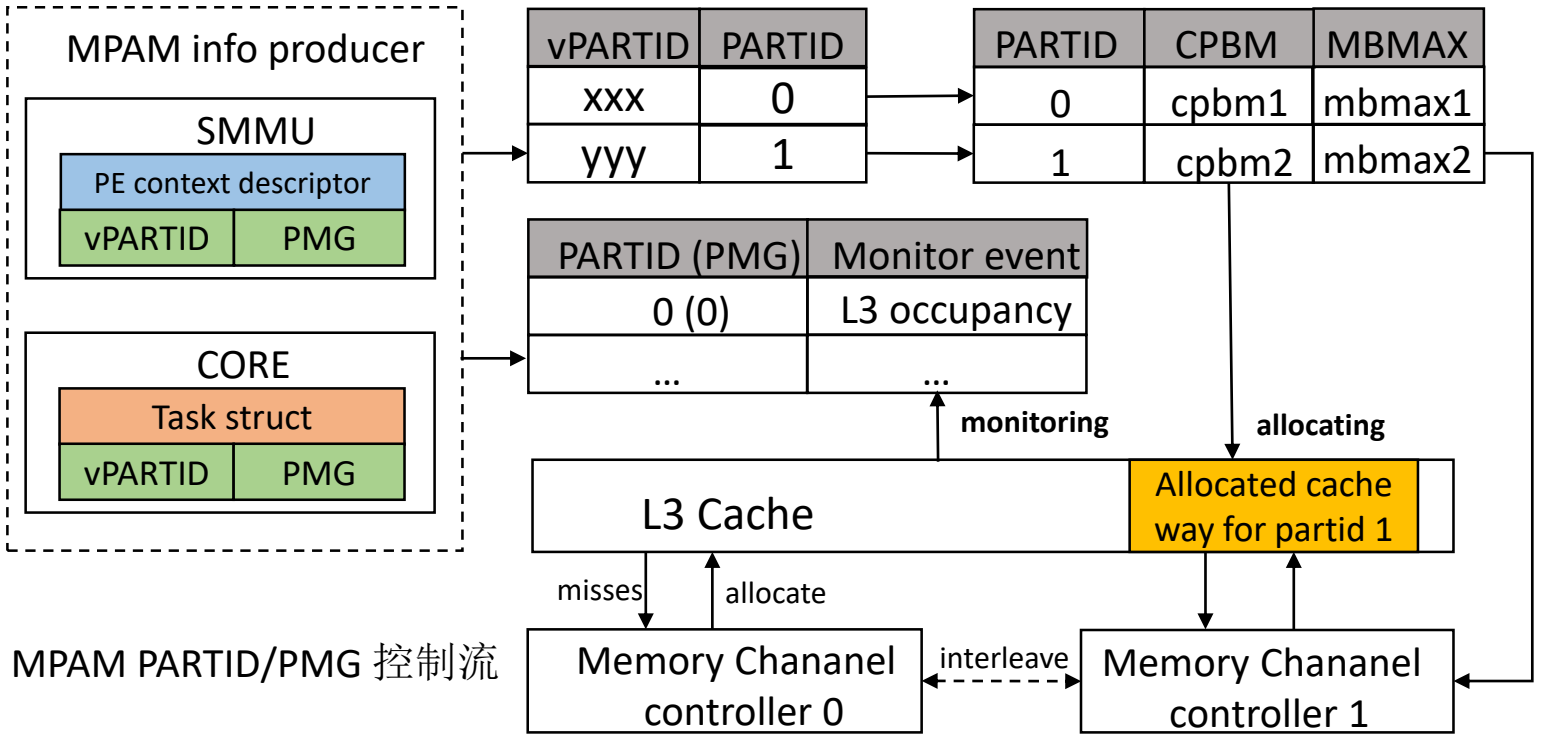
配置示例：

- CPBM将Cache way映射成为bitmask
- MBMAX将业务流带宽限制到当前通道所能达到最大带宽的百分比

物理结构上一条Cache way跨越多个Cache set，1条Cache way对应MPAM Cache Portion Bit Mask（CPBM）的一个bit



L3 Tag Structure



MPAM PARTID/PMG 控制流

openEuler

目录

1 背景及MPAM介绍

2 MPAM软硬实现

3 MPAM和RDT

4 MPAM开发验证

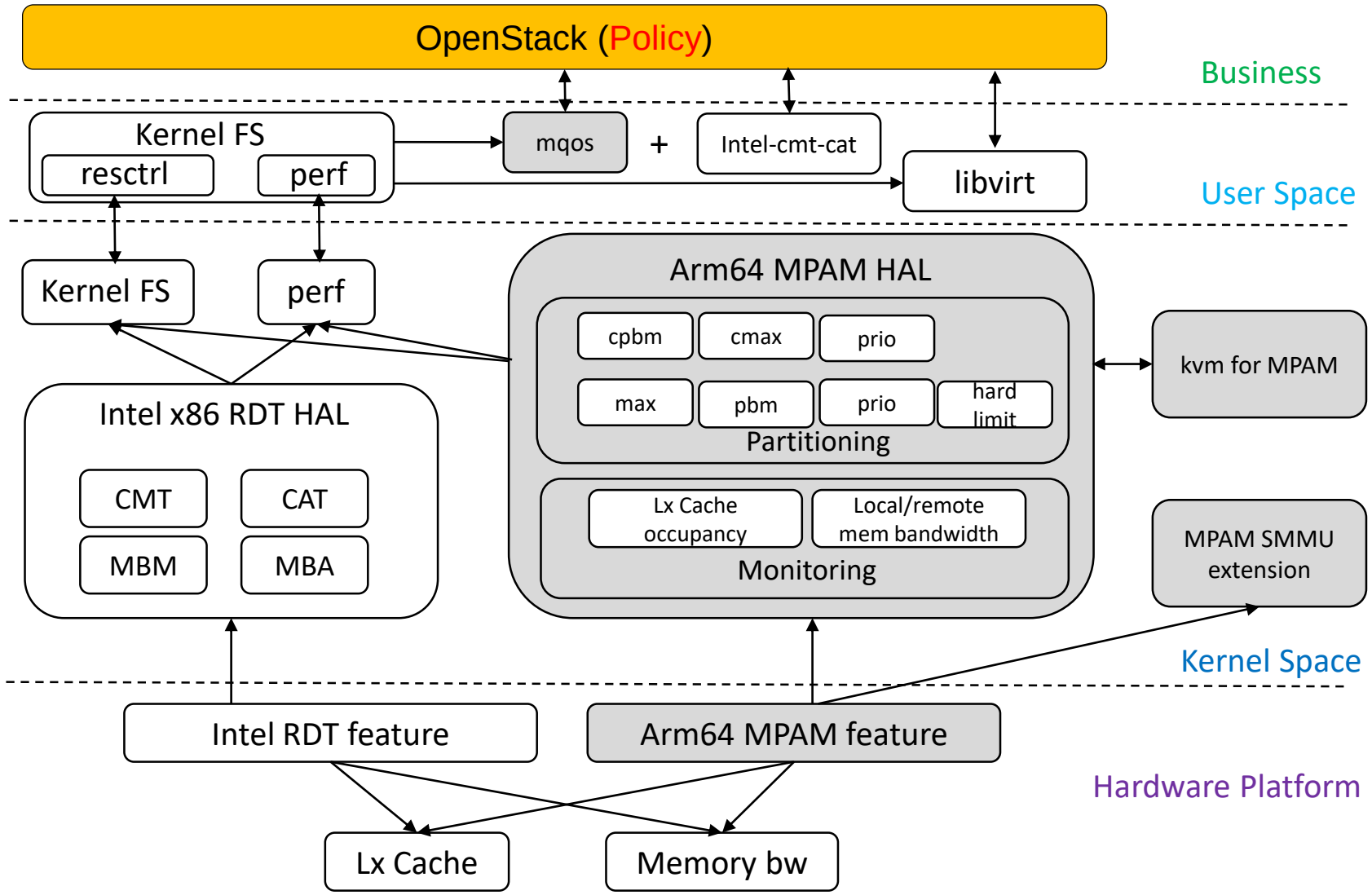
5 资源隔离应用场景

6 MPAM应用扩展

openEuler

MPAM和RDT

Hardware， Kernel， User space上MPAM和RDT的异同



- RDT已有多种部署框架，例如RMD和AppForMix
- Libvirt使用resctrl完成虚拟机的部署核静态配置；
- mqos为用户提供一**套管理资源池的的用户态，以及一套性能分析工具**
- Intel-cmt-cat为RDT的用户态工具[1]

- RDT kernel: Linux 4.10+
  MPAM kernel:openeuler 4.19.36+
- MPAM使用kvm管理vPARTID到PARTID的映射，可实现**虚拟机内部划分资源**，而RDT不支持，RDT需要通过类似vCat[2]的机制完成虚拟机内部资源划分。

- RDT 通过读写每个CPU的MSR寄存器完成配置，MPAM寄存器按照资源划分，需要ACPI/DT指定基地址；RDT通过Cacheinfo上报资源，MPAM需要ACPI上报。

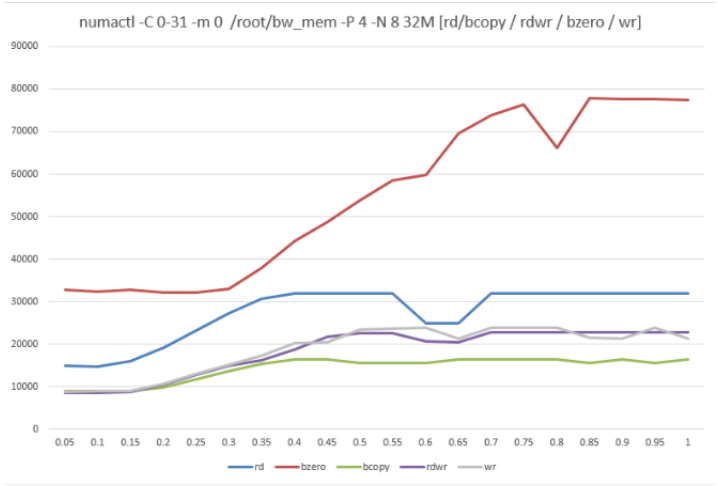
[2] Xu M, Thi L, Phan X, et al. vCAT: Dynamic cache management using CAT virtualization[C]//2017 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS). IEEE, 2017: 211-222.

[1] https://github.com/intel/intel-cmt-cat

MPAM和RDT

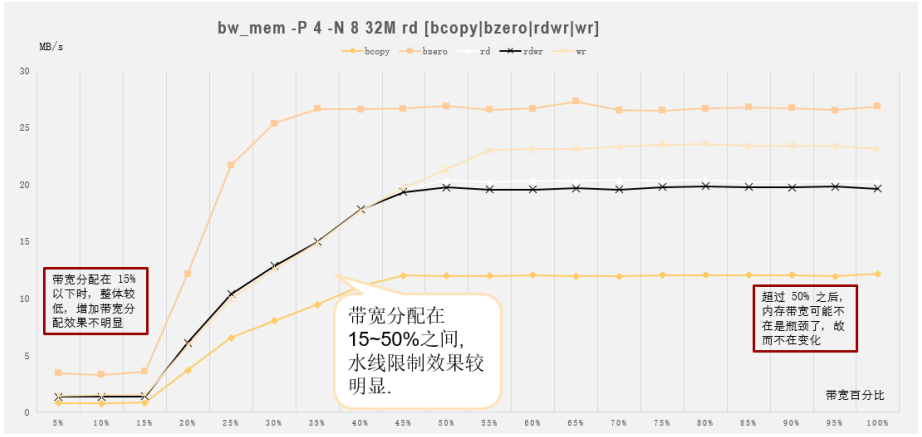
MPAM和RDT在测试指标上的异同

鲲鹏920 MPAM带宽max stride



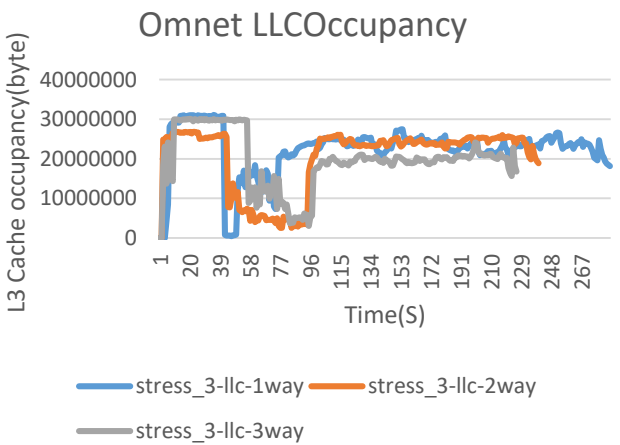
MPAM带宽max stride限制模式

RH2288-v5 CPU: Skylake Intel(R) Xeon(R) Gold 6140 CPU

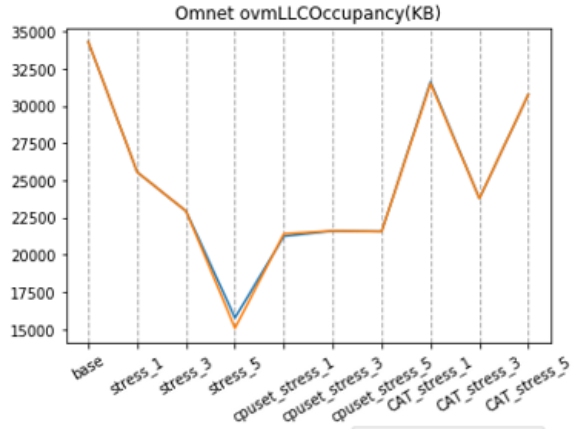


RDT的带宽限制特点在于流量较大时相对有效，且有效区间集中在百分比下半段。

鲲鹏920，MPAM配置stress干扰的Cache way个数越多，Omnet的Cache容量越小



Intel RDT CAT效果



openEuler

目录

1 背景及MPAM介绍

2 MPAM软硬实现

3 MPAM和RDT

4 MPAM开发验证

5 资源隔离应用场景

6 MPAM应用扩展

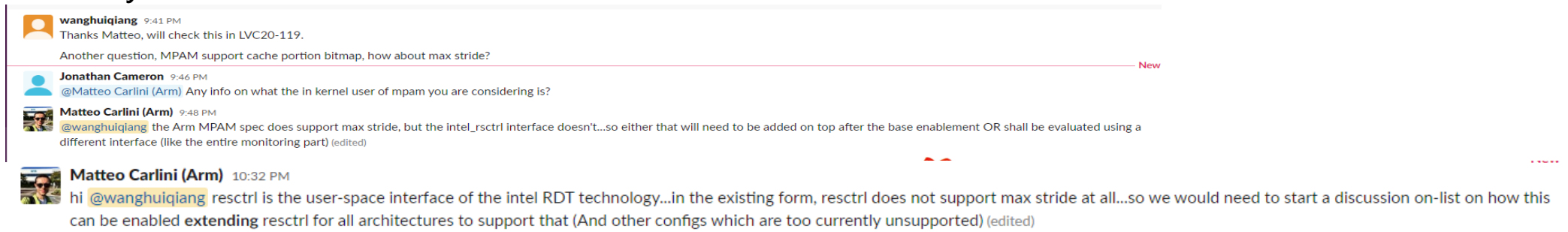
MPAM开发验证

1. Linux社区upstream当前状态：

未支持MPAM驱动，arm在linaro connect大会上表示将会在2021年底推到社区，当前已有自己本地分支[1]。

2. 社区兼容性问题：

- Arm 决定将 MPAM 和 RDT使用同一套用户态接口（resctrl），但MPAM专有的priority，hardlimit，SMMU Memory Max stride等功能无法满足，Arm期望社区接口能做增强。



[1] http://www.linux-arm.org/git?p=linux-jm.git;a=summary

openEuler

MPAM开发验证

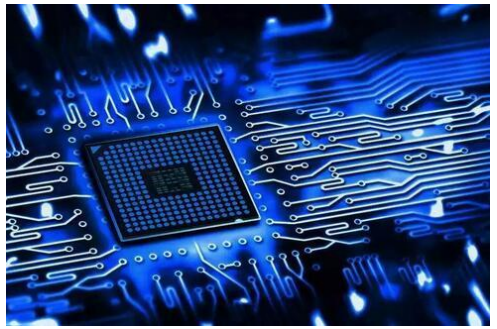
3. openEuler支持

- openEuler20.03版本已初步支持MPAM驱动。
- mqos工具后续在openeuler开源。
- 基于最新arm分支版本[1]，完成开发一个**全量支持MPAM的驱动版本**，该版本将会后续在openeuler开源。

官方网站

4. 芯片功能验证

- 特定场景进行实验
  1）重点测干扰强度，干扰层级，分析关键业务对Cache和带宽的敏感程度。
- 平台部署方式实验
  1）DIE交织，访存Channel，L3 Cache partition/shared对访存隔离的影响。
- 边界和压力实验
  1）实时线程吃低带宽时，压低低带宽的隔离效果；
  2）大带宽下的限制效果以及不同水线下的延时测试。



[1] http://www.linux-arm.org/git?p=linux-jm.git;a=summary

openEuler

目录

1 背景及MPAM介绍

2 MPAM软硬实现

3 MPAM和RDT

4 MPAM开发验证

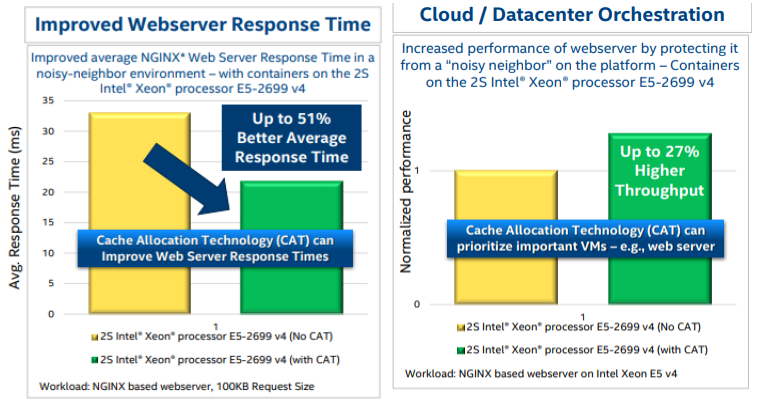
5 资源隔离应用场景

6 MPAM应用扩展

openEuler

资源隔离应用场景

RDT CAT隔离场景

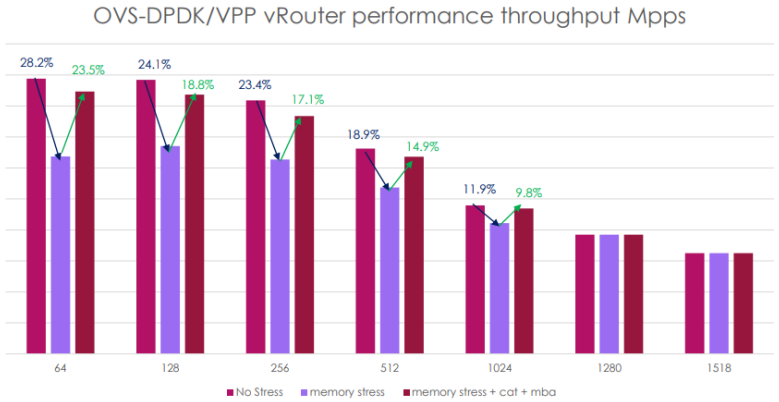


Latency Results. Using the AppFormix* software suite and Intel's Cache Allocation Technology (CAT) up to a 51%[1] reduction in average response time (latency) can be achieved, improving the experience for end users.

Throughput Results. Using the AppFormix* software suite and Intel's Cache Allocation Technology (CAT) to contain a "noisy neighbor" and prioritize the NGINX* web server yields up to a 27%[1] performance increase for the web server.

使用CAT提升web服务器性能[1]

RDT CAT+MBA隔离场景



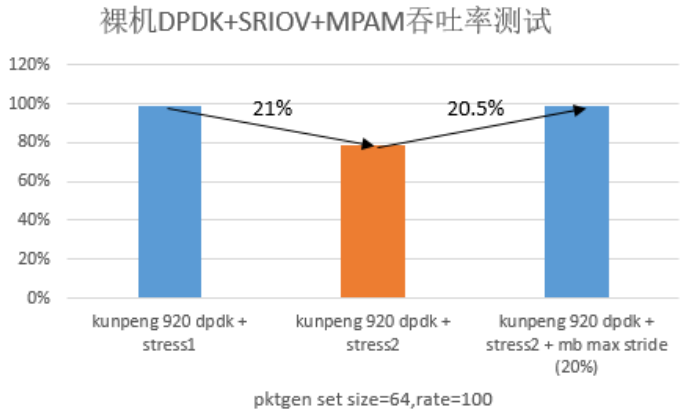
dpdk应用使用RDT提升性能[2]

DPDK + MPAM带宽限制场景(裸机上跑DPDK+SRIOV，用Pktgen做测试)

加压1：taskset 0-11 ./stress-ng --cache 12

加压2：--cache 9 –cache-flush –cache-prefetch –aggressive –cpu 2 –cpu-method matrixprod [3]

	solo	加压1	加压2	加压2 + MPAM max stride
每秒平均指令数	2E+09 + 5E+07	2E+09 + 3E+07	1.7E+09	2E+09 + 4E+07
指令数下降比（和solo模式相比）	NA	0.6%	22%	0.7%
dpdk性能下降（和solo模式相比）	NA	≈0	≈21%	≈0



实验表明使用MPAM带宽限制后，Pkts Rx和TX有明显提升。

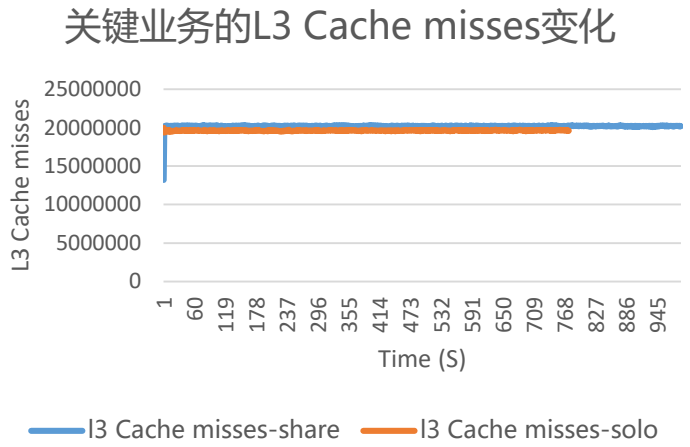
[1] https://builders.intel.com/docs/cloudbuilders/Intel_AppFormix_SolutionBrief_Final.pdf
[2] https://www.dpdk.org/wp-content/uploads/sites/35/2019/07/01-DPDK_on_multicore_CPU_100.pdf
[3] https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/increasing-platform-determinism-pqos-dpdk-paper.pdf

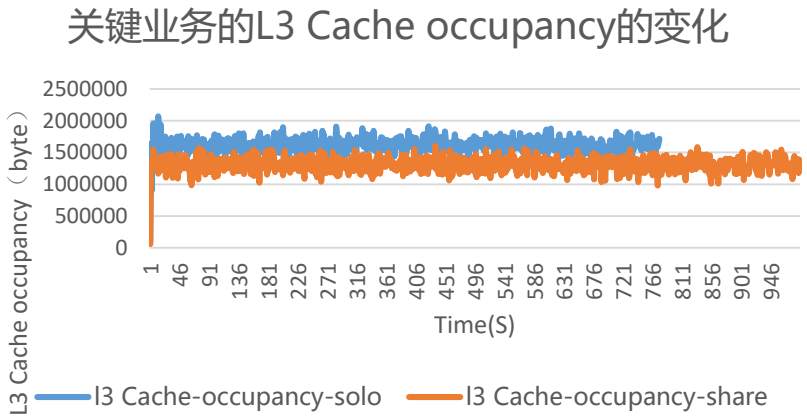
openEuler

资源隔离应用场景

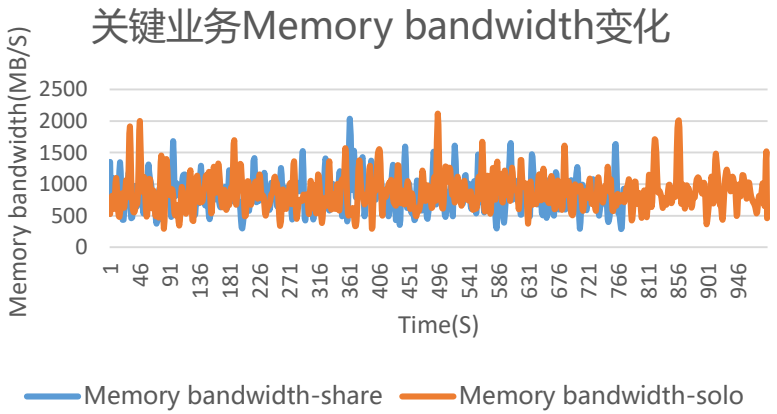
MPAM L3 Cache隔离场景

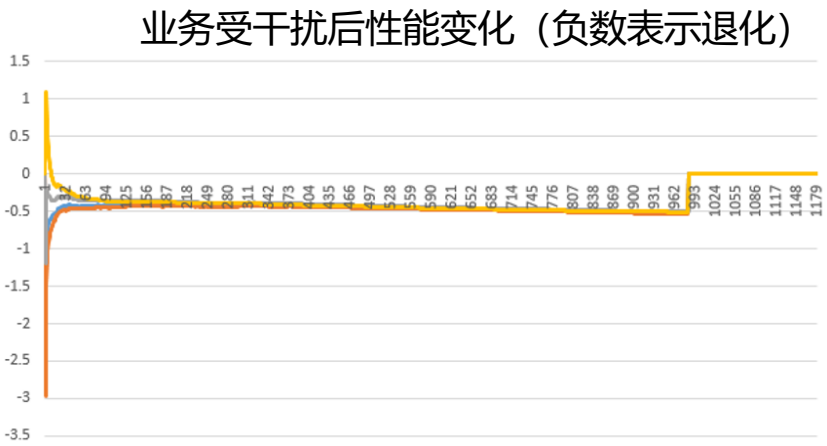
一般这种场景带宽竞争较小，竞争层级主要发生在L3 Cache





MPAM monitor统计关键业务L3 Cache Occupancy有所下降，带宽无明显变化。





L3 Cache misses导致关键业务访存延时变大，造成关键业务执行速度变慢5%，使用Cache隔离后业务性能回复正常。

openEuler

目录

1 背景及MPAM介绍

2 MPAM软硬实现

3 MPAM和RDT

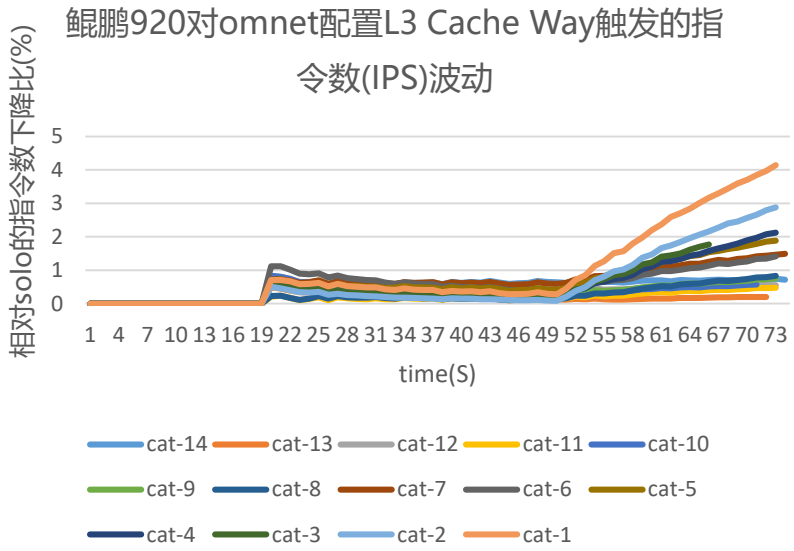
4 MPAM开发验证

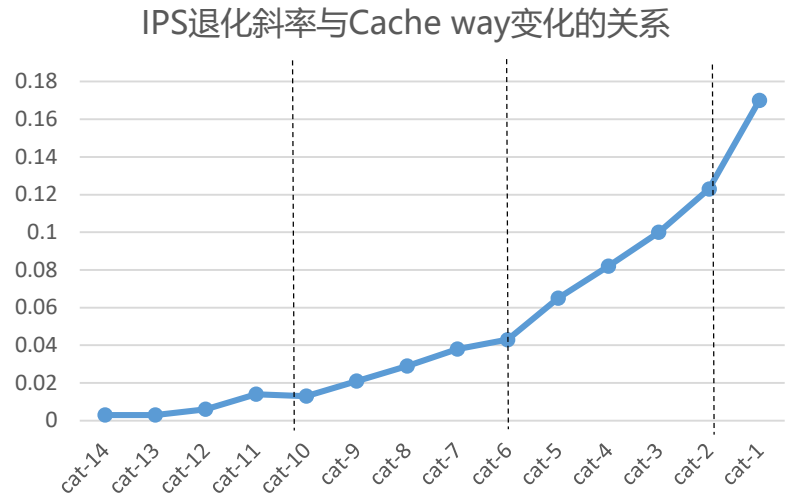
5 资源隔离应用场景

6 MPAM应用扩展

MPAM应用扩展

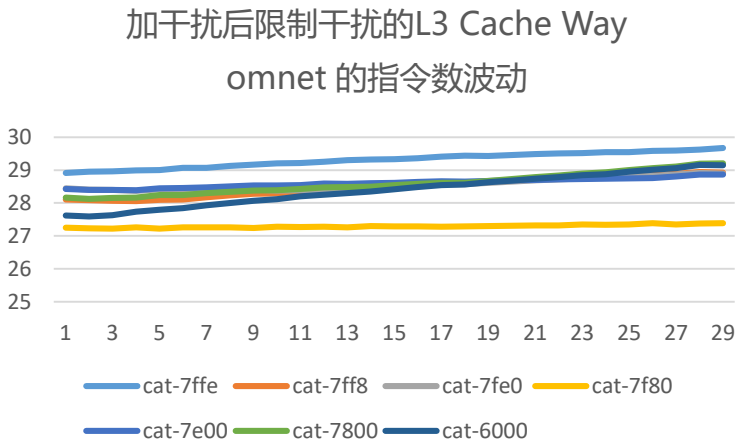
后续工作扩展

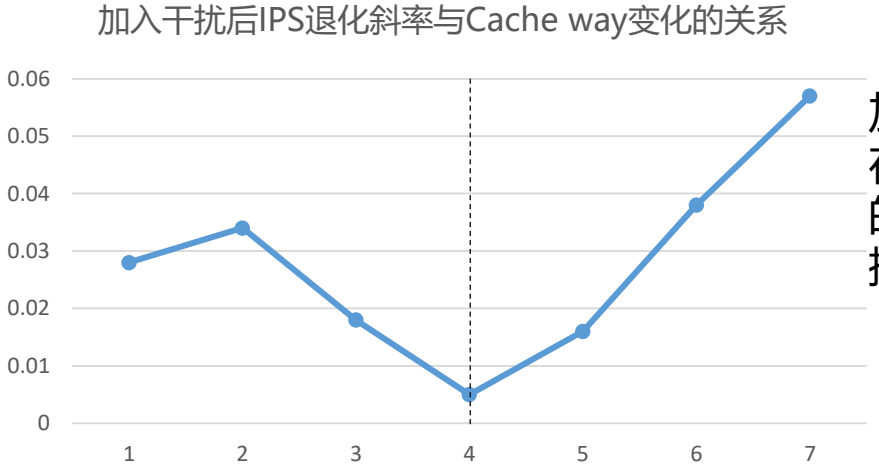




IPS退化斜率呈现指数上升，说明L3 Cache的减小和业务性能退化呈现指数关系。

$$f(c) = \begin{cases} slope_0 * c + offset_j & 0 < j < m \\ slope_1 * c + offset_k & m \le k < n \\ \dots & n \le l < o \end{cases}$$





加入干扰，系统压力变化，L3 Cache的干扰在某个压力区间内不是主导业务性能退化的主要原因，寻求最优比可以节省共享资源，提升整体性能。

openEuler

openEuler