Linux内核在百万级服务器云计算环境下面临的可靠性挑战

浙江大学计算机学院

阿里云智能事业群

陈义全

2018.12.16

云计算面临的可靠性挑战

November 22, 2018 AWS Network Failure Takes Down Crypto Trading in South Korea

https://cryptochronicle.com/aws-network-failure-takes-down-crypto-trading-in-south-korea/

September 4, 2018 Microsoft Azure and Office 365 Services Go Down in Texas Service Area

https://redmondmag.com/articles/2018/09/04/azure-office-365-down-in-texas.aspx

云计算面临的可靠性挑战

阿里云627故障

2018年6月27日下午，我们在运维上的一个操作失误，导致一些客户访问阿里云官网控制台和使用部分产品功能出现问题，引发了大量吐槽。

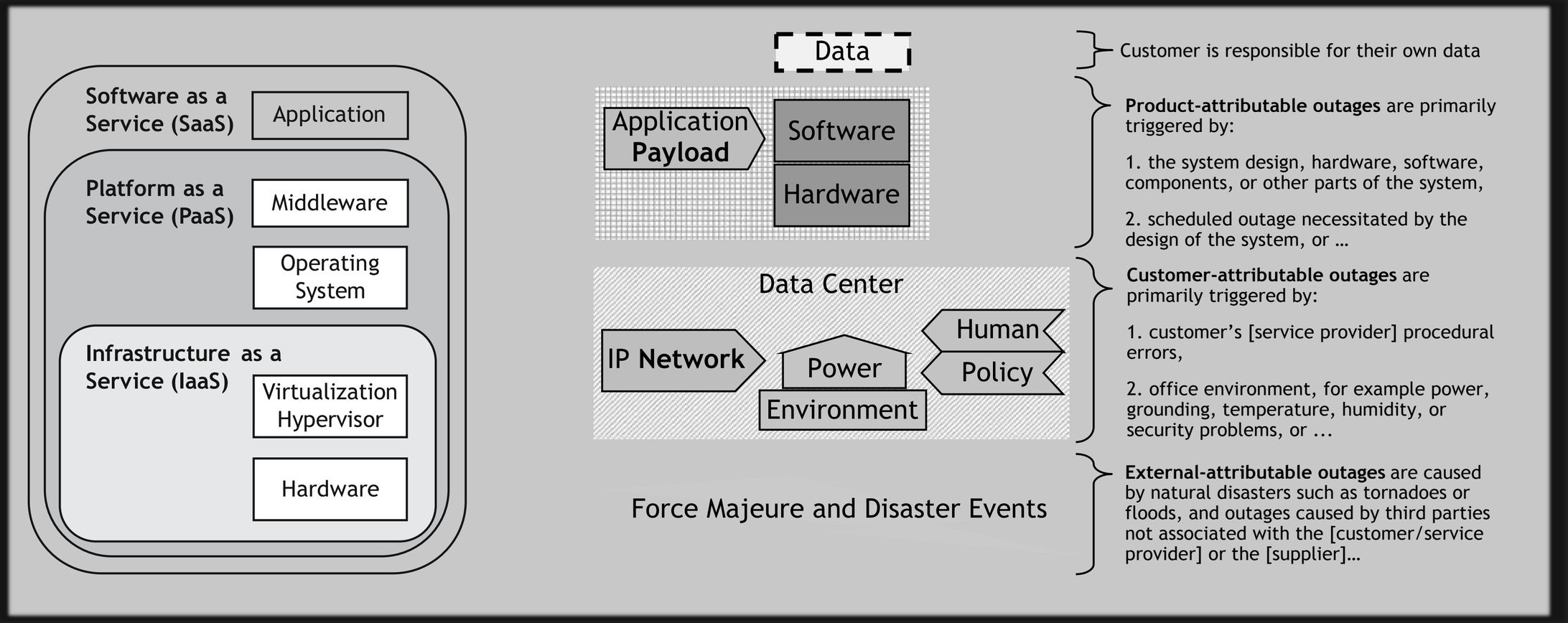
https://yq.aliyun.com/articles/603866

腾讯云“数据丢失事件”

2018年7月20日，北京清博数控科技有限公司所属“前沿数控”平台一块操作系统云盘，因受腾讯云北京三区部分物理硬盘固件版本bug导致的静默错误（写入数据和读取出来的不一致）影响，文件系统元数据损坏。

http://finance.sina.com.cn/roll/2018-08-06/doc-ihhhczfc6817406.shtml

云计算可靠性框架



Eric Bauer, Randee Adams, *Reliability and Availability of Cloud Computing,* Wiley-IEEE Press,2012

IaaS面临的可靠性挑战

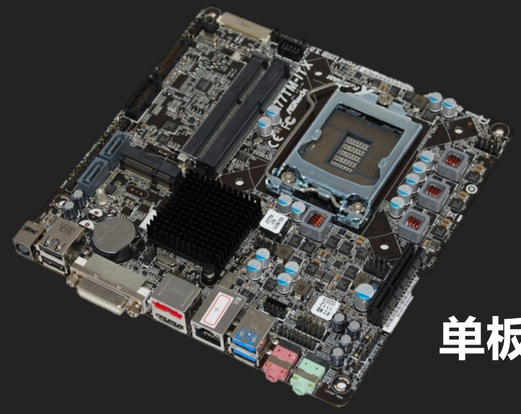
- ✓ 若日宕机率3%%，100万台服务器，每天就有**300**台服务器宕机
- ✓ Linux内核自身稳定性，非常关键
- ✓ 不宕机，同时也要保障QoS及安全性

服务器硬件面临的可靠性挑战

- 硬件电子系统（芯片、单板、服务器系统）日益复杂

芯片

200m transistors



单板

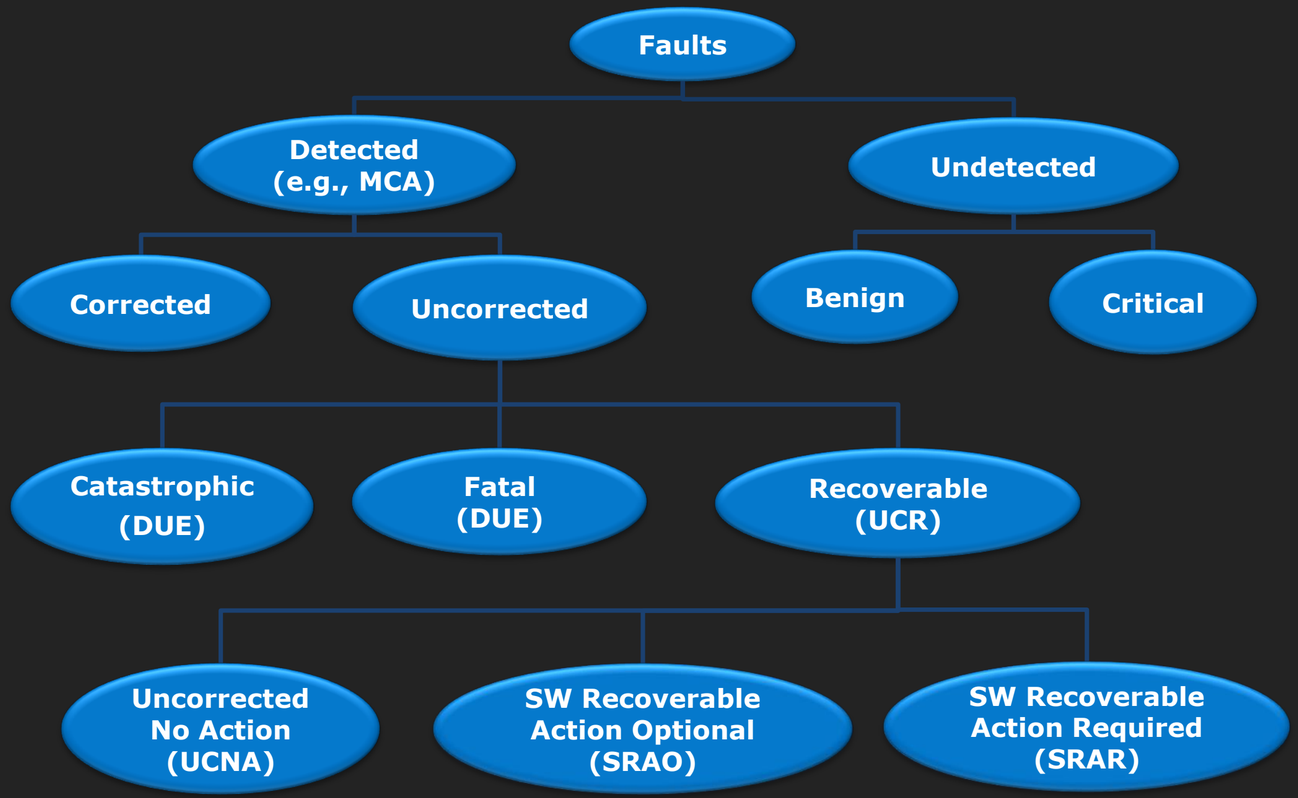
Tens/hundreds of chips

服务器系统

Many PCBs

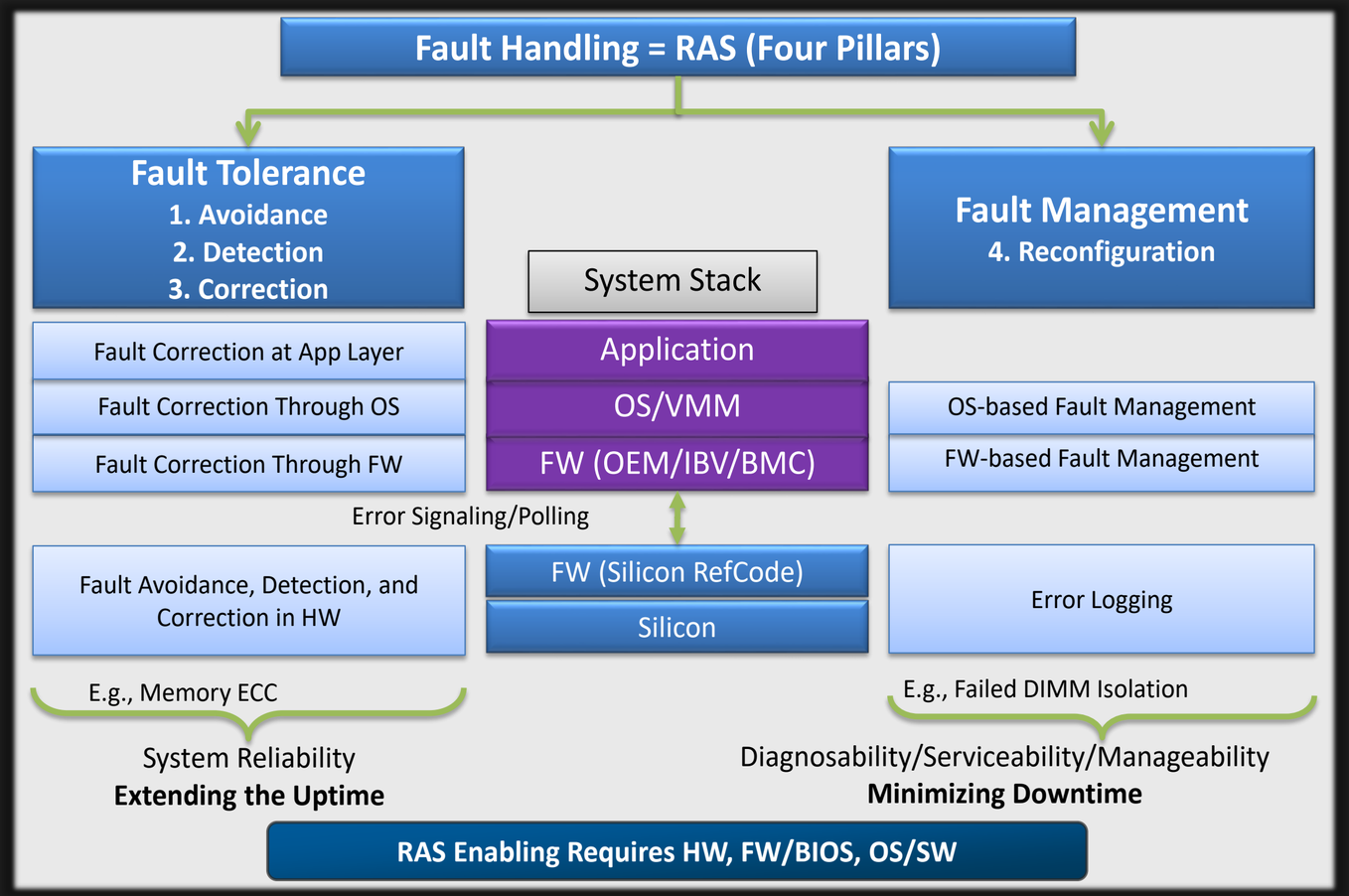
- 硬件器件有寿命的，不可避免**老化**，引起故障或宕机
- 测试不完美，缺陷芯片可能被集成进单板，在特定运行条件和环境影响下造成服务器系统宕机

x86架构错误分类



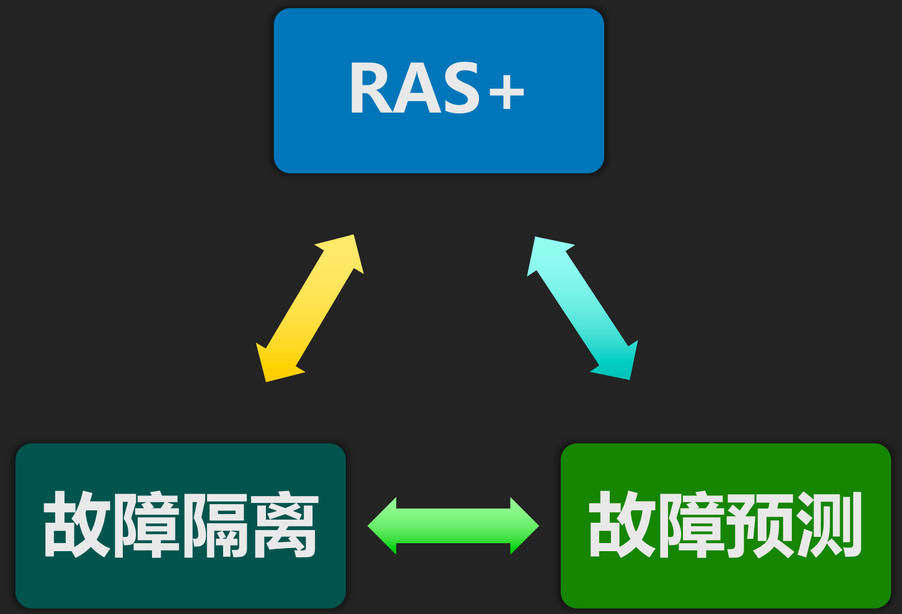
MCA: *Machine Check Architecture*
DUE: *Detectable but Uncorrected Error*
UCR: *Uncorrected Recoverable*

x86服务器RAS框架



Lucia Mao，Spike Yuan，*Server RAS and UEFI CPER*，March 27 - 31, 2017

提高硬件可靠性措施



RAS+

在硬件RAS基础上做增强改进

故障隔离

对故障单元的有效隔离，减少故障的扩散半径

故障预测

基于大数据和机器学习算法，预测系统故障概率，降低故障影响业务

Linux内核面临的可靠性挑战

挑战#1：差异化

挑战#2：精细化

挑战#3：unknown问题

挑战#4：稳定性

挑战#5：故障域

挑战#6：QoS

挑战#1：差异化

1、xxx业务陆续有机器宕机，宕机日志提示是 Fatal hardware error导致内核panic；

2、发生故障的设备，section type: unknown, eb5e4685-ca66-4769-b6a2-26068b001326，提示是一个pci 设备，内核没有打印详细的 BDF 信息。

```
[Hardware Error]: Hardware error from APEI Generic Hardware Error Source: 1
[Hardware Error]: APEI generic hardware error status
[Hardware Error]: severity: 1, fatal
[Hardware Error]: section: 0, severity: 1, fatal
[Hardware Error]: flags: 0x01
[Hardware Error]: primary
[Hardware Error]: section type: unknown, eb5e4685-ca66-4769-b6a2-26068b001326
Kernel panic - not syncing: Fatal hardware error!
```

N.2.9 PCI/PCI-X Component Error Section

Type: {0xEB5E4685, 0xCA66, 0x4769, {0xB6, 0xA2, 0x26, 0x06, 0x8B, 0x00, 0x13, 0x26}}

Table 279. PCI/PCI-X Component Error Section

Mnemonic	Byte Offset	Byte Length	Description
Validation Bits	0	8	Indicate which fields are valid: Bit 0 – Error Status Valid Bit 1 – Id Info Valid Bit 2 – Memory Number Valid Bit 3 – IO Number Valid Bit 4 – Register Data Pair Valid Bit 5-63 Reserved

故障没有区分，直接Panic

差异化建议

硬件故障，按照严重级别区别处理，减少直接panic

- FATAL
- ERROR
- WARN
- INFO
- DEBUG

挑战#2：精细化

```
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.702113] [Hardware Error]: Hardware error from APEI Generic Hardware Error Source: 1
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.710499] [Hardware Error]: APEI generic hardware error status
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.716831] [Hardware Error]: severity: 1, fatal
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.721770] [Hardware Error]: section: 0, severity: 1, fatal
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727751] [Hardware Error]: flags: 0x01
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727753] [Hardware Error]: primary
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727754] [Hardware Error]: section type: unknown, eb5e4685-ca66-4769-b6a2-26068b001326
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727756] Kernel panic - not syncing: Fatal hardware error!
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727758] Pid: 0, comm: swapper Tainted: G        W  ----------------   2.6.32-220.23.2.al…
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727760] Call Trace:
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727761]    [] ? panic+0x78/0x145
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727774]  [] ? ghes_notify_nmi+0x17c/0x180
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727779]  [] ? notifier_call_chain+0x55/0x80
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727782]  [] ? atomic_notifier_call_chain+0x1a/0x20
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727787]  [] ? notify_die+0x2e/0x30
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727790]  [] ? do_nmi+0x1a1/0x2b0
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727792]  [] ? nmi+0x20/0x30
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727797]  [] ? native_write_msr_safe+0xa/0x10
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727798]  <>    [] ? intel_pmu_disable_all+0x3f/0x110
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727805]  [] ? x86_pmu_disable+0x52/0x60
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727811]  [] ? perf_pmu_disable+0x2b/0x40
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727814]  [] ? perf_event_task_tick+0x2a5/0x2f0
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727820]  [] ? scheduler_tick+0xd3/0x270
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727824]  [] ? tick_sched_timer+0x0/0xc0
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727829]  [] ? update_process_times+0x52/0x70
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727832]  [] ? tick_sched_timer+0x66/0xc0
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727835]  [] ? __run_hrtimer+0x8e/0x1a0
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727839]  [] ? read_tsc+0x9/0x20
2018-08-28 03:08:32 rdbtantalumna62a011… conman [35517899.727841]  [] ? hrtimer_interrupt+0xe6/0x250
2018-08-28 03:08:33 rdbtantalumna62a011… conman [35517899.727845]  [] ? smp_apic_timer_interrupt+0x6b/0x9b
```

宕机打印信息较为笼统，对问题定位帮助有限

精细化建议

Panic时，打印更多有价值信息（自动解析）

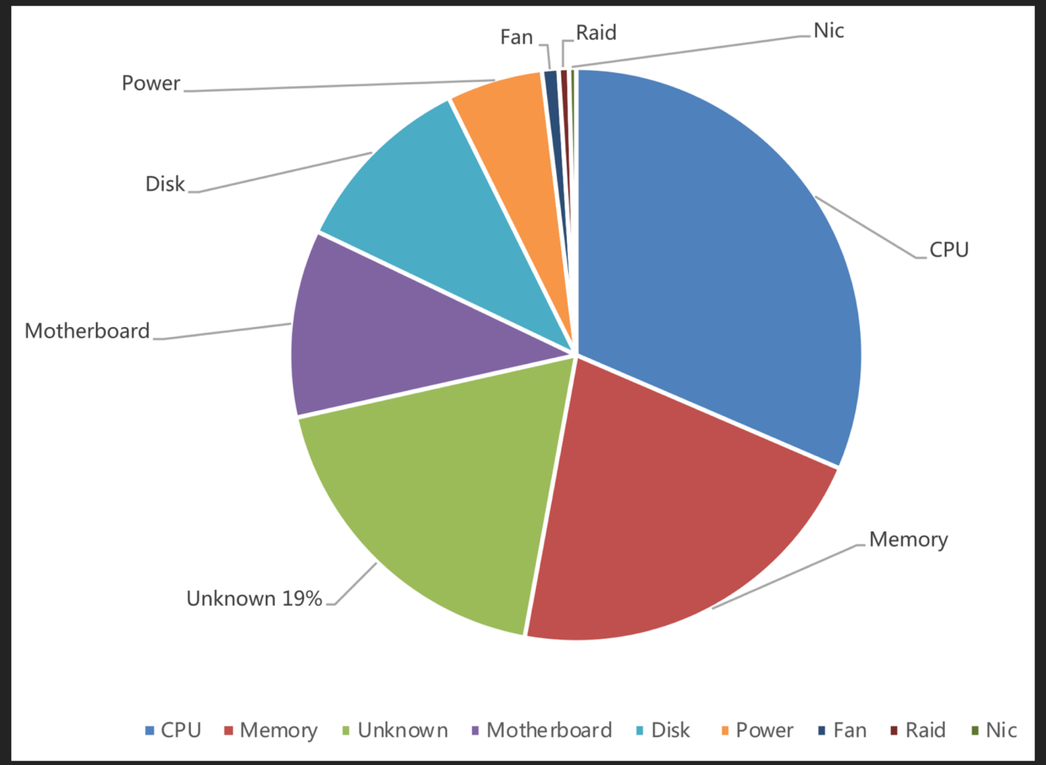
- 硬件问题
  - *部件名称*
  - *部件位置*
  - *故障原因*
- 内核Bug

挑战#3： unknown问题

```
2018-11-29 18:02:26 i46f15219.cloud.sgl… conman    Alibaba Group Enterprise Linux Server release 7.2 (Paladin)
2018-11-29 18:02:26 i46f15219.cloud.sgl… conman    Kernel 3.10.0-327.ali2016.alios7.x86_64 on an x86_64
2018-11-29 18:02:26 i46f15219.cloud.sgl… conman    i46f15219 login:
2018-11-29 18:02:26 i46f15219.cloud.sgl… conman    Alibaba Group Enterprise Linux Server release 7.2 (Paladin)
2018-11-29 18:02:26 i46f15219.cloud.sgl… conman    Kernel 3.10.0-327.ali2016.alios7.x86_64 on an x86_64
2018-11-29 18:02:26 i46f15219.cloud.sgl… conman    i46f15219 login:
2018-11-29 18:03:55 i46f15219.cloud.sgl… conman    i46f15219 login:
2018-11-29 18:03:55 i46f15219.cloud.sgl… conman    Alibaba Group Enterprise Linux Server release 7.2 (Paladin)
2018-11-29 18:03:55 i46f15219.cloud.sgl… conman    Kernel 3.10.0-327.ali2016.alios7.x86_64 on an x86_64
2018-11-29 18:03:55 i46f15219.cloud.sgl… conman    i46f15219 login:
2018-11-29 18:03:55 i46f15219.cloud.sgl… conman    Alibaba Group Enterprise Linux Server release 7.2 (Paladin)
2018-11-29 18:03:55 i46f15219.cloud.sgl… conman    Kernel 3.10.0-327.ali2016.alios7.x86_64 on an x86_64
2018-11-29 19:08:05 i46f15219.cloud.sgl… conman      [0m[1m[01;00H[0m[2;37;40m                [2J [1;1H [2J [1;1H [0;37;40m [25;60H [2J [1;1H [1;1H [1;1HBIOS Date: 08/…
2018-11-29 19:08:42 i46f15219.cloud.sgl… conman    [    0.000000] Initializing cgroup subsys cpu
2018-11-29 19:08:42 i46f15219.cloud.sgl… conman    [    0.000000] Initializing cgroup subsys cpuacct
2018-11-29 19:08:42 i46f15219.cloud.sgl… conman    [    0.000000] Linux version 3.10.0-327.ali2016.alios7.x86_64 (admin@e81408e7a752) (gcc version 4.8.5 20150623 (…
2018-11-29 19:08:42 i46f15219.cloud.sgl… conman    [    0.000000] Command line: BOOT_IMAGE=/vmlinuz-3.10.0-327.ali2016.alios7.x86_64 root=UUID=ae533e49-1f28-4740-9…
2018-11-29 19:08:42 i46f15219.cloud.sgl… conman    [    0.000000] e820: BIOS-provided physical RAM map:
2018-11-29 19:08:42 i46f15219.cloud.sgl… conman    [    0.000000] BIOS-e820: [mem 0x0000000000000000-0x00000000000997ff] usable
2018-11-29 19:08:42 i46f15219.cloud.sgl… conman    [    0.000000] BIOS-e820: [mem 0x0000000000099800-0x000000000009ffff] reserved
2018-11-29 19:08:42 i46f15219.cloud.sgl… conman    [    0.000000] BIOS-e820: [mem 0x00000000000e0000-0x00000000000fffff] reserved
2018-11-29 19:08:42 i46f15219.cloud.sgl… conman    [    0.000000] BIOS-e820: [mem 0x0000000000100000-0x0000000069edbfff] usable
2018-11-29 19:08:42 i46f15219.cloud.sgl… conman    [    0.000000] BIOS-e820: [mem 0x0000000069edc000-0x000000006c86cfff] reserved
2018-11-29 19:08:42 i46f15219.cloud.sgl… conman    [    0.000000] BIOS-e820: [mem 0x000000006c86d000-0x000000006ca11fff] usable
2018-11-29 19:08:42 i46f15219.cloud.sgl… conman    [    0.000000] BIOS-e820: [mem 0x000000006ca12000-0x000000006d4d4fff] ACPI NVS
2018-11-29 19:08:42 i46f15219.cloud.sgl… conman    [    0.000000] BIOS-e820: [mem 0x000000006d4d5000-0x000000006f345fff] reserved
2018-11-29 19:08:42 i46f15219.cloud.sgl… conman    [    0.000000] BIOS-e820: [mem 0x000000006f346000-0x000000006f7fffff] usable
2018-11-29 19:08:42 i46f15219.cloud.sgl… conman    [    0.000000] BIOS-e820: [mem 0x000000006f800000-0x00000000ffffffff] reserved
```

突然宕机，无内核日志输出

服务器宕机原因分布


Fan
Raid
Nic
Power
Disk
CPU
Motherboard
Memory
Unknown 19%
CPU
Memory
Unknown
Motherboard
Disk
Power
Fan
Raid
Nic

unknown问题建议

不能一死了之，要想办法打印信息

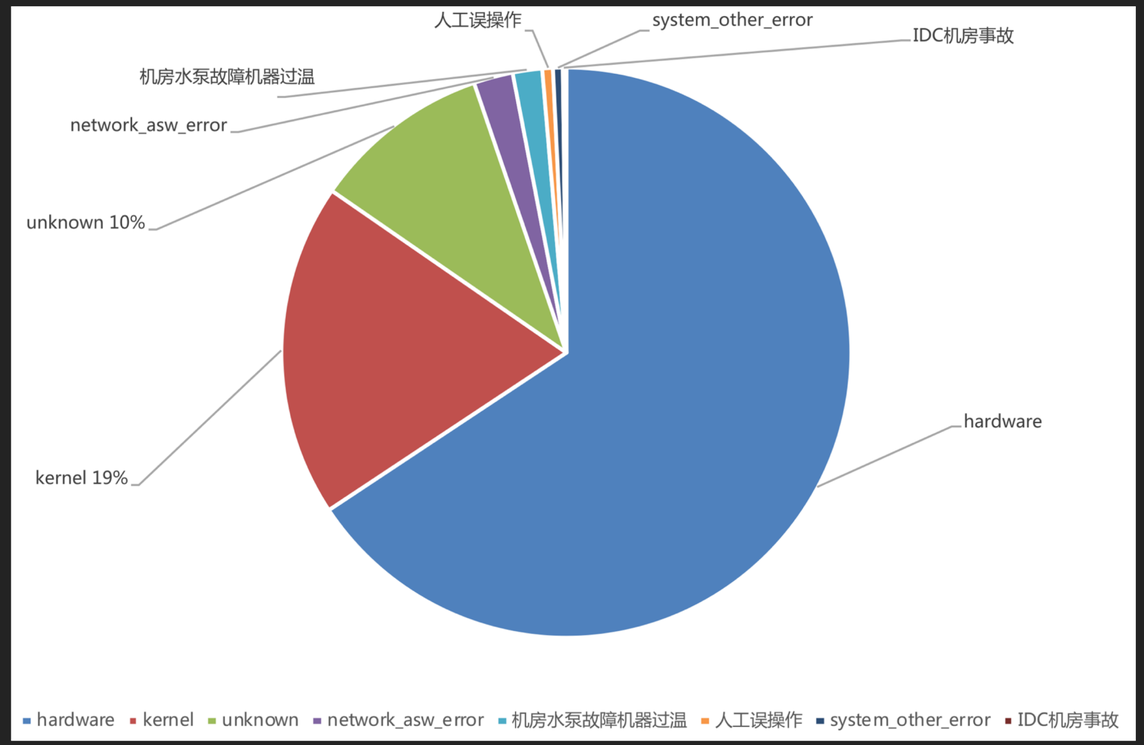
[unknown | Huoshui]温度超标 raw: reboot by fht	G42	BDW	AY10F	960607	2018/8/27 6:56
[unknown idc]Error_Time:2018-09-06 01:53:47 Error_Creator:cruiser Error_Level:严重 Error_Code:Bus F	G41G2	BDW	AY216L	1549255	2018/9/6 1:54
[unknown | discovery]驻场启动失败，主板故障，报修单W18091009052927466	S10-3S	IVB	AY11I	972703	2018/9/8 12:49
[unknown idc]Error_Time:2018-10-11 02:42:58 Error_Creator:cruiser Error_Level:致命 Error_Code:PCIe	G41G2	BDW	AY216L	1563135	2018/10/11 2:44
IPMI日志不可用，无其他异常，真实宕机原因不详	S10-3S	IVB	AY66F	1539415	2018/8/29 0:38
[unknown idc]Error_Time:2018-09-24 22:55:40 Error_Creator:cruiser Error_Level:严重 Error_Code:Bus F	G41G2	BDW	AY216L	1557634	2018/9/24 22:56
[unknown | Huoshui]UCE raw:	G41G1	BDW	AY211I	1437451	2018/6/26 14:15
[unknown | discovery]驻场重启失败,维修单	N32	HSW	AY74H	764204	2018/10/23 5:46

挑战#4： 稳定性

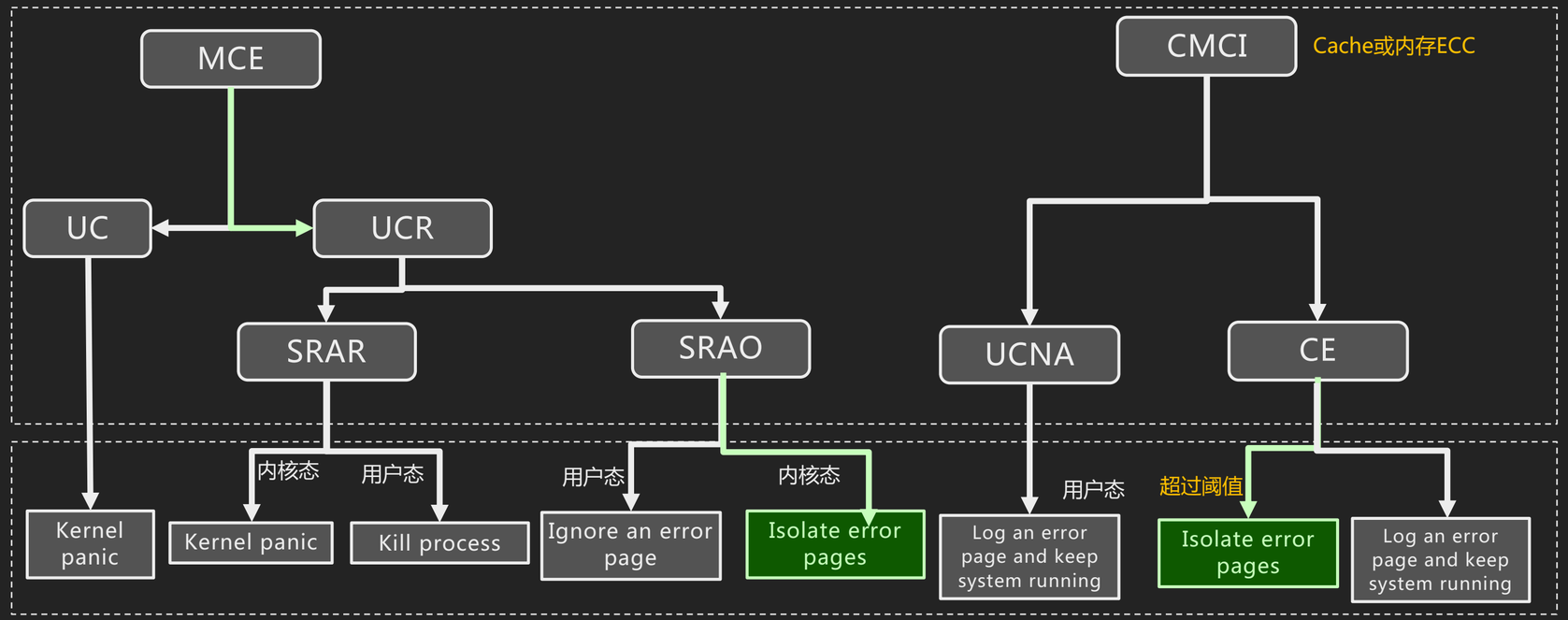
```
[12424406.711946] BUG: unable to handle kernel NULL pointer dereference at           (null)
[12424406.720486] IP: [<ffffffff810e6029>] clockevents_program_event+0x99/0xf0
[12424406.727903] PGD 0
[12424406.730606] Oops: 0002 [#1] SMP
[12424406.734562] Modules linked in: kpatch_D559221(OE) kpatch_D602147(OE) kpatch_D537536(OE) kpatch_D723518(OE) kpatch_D644168(OE) kpatch_D656712(OE) kpatch_D601425(OE) kpatch_D528317(OE) fuse btrfs zlib_deflate raid6_pq xor
d dm_mod kpatch_D549556(OE) kpatch_D543129(OE) bcache(OE) vrbd(OE) nbd ip6_tables iptable_filter ebtable_nat ebtables tcp_diag inet_diag binfmt_misc nf_conntrack_ipv4(F) nf_defrag_ipv4(F) nf_conntrack(F) kpatch_D579677(OE) kp
isert(OE) iscsi_target_mod target_core_mod ib_iser(OE) libiscsi scsi_transport_iscsi ib_srpt(OE) ib_srp(OE) scsi_transport_srp(OE) ib_ipoib(OE) rdma_ucm(OE) ib_ucm(OE) ib_uverbs(OE) ib_umad(OE) rdma_cm(OE) ib_cm(OE) iw_cm(OE)
ding intel_powerclamp coretemp
[12424406.810629]  intel_rapl crc32_pclmul ghash_clmulni_intel aesni_intel lrw gf128mul glue_helper ablk_helper cryptd iTCO_wdt iTCO_vendor_support ipmi_devintf pcspkr sg i2c_i801 lpc_ich shpchp mfd_core wmi ipmi_si ipmi_msgh
t_net(OE) tun macvtap macvlan vfio_pci(OFE) vfio_iommu_type1(OE) vfio(OFE) ip_tables mlx5_ib(OE) ib_core(OE) ext4 kvm_intel_3(OE) mbcache kvm_intel_0(OE) jbd2 kvm_intel_1(OE) ast syscopyarea sysfillrect sysimgblt kvm_intel_2(
 drm_kms_helper nvme ttm mlx5_core(OE) mlx_compat(OE) ptp pps_core drm vxlan ip6_udp_tunnel udp_tunnel i2c_core sd_mod crc_t10dif crct10dif_generic crct10dif_pclmul crct10dif_common ahci libahci libata
[12424406.879024] CPU: 85 PID: 0 Comm: swapper/85 Tainted: GF       W  OE K------------   3.10.0-327.ali2016.alios7.x86_64 #1
[12424406.891085] Hardware name: Inspur AliServer Thor02-2U/YZMB-00824-101, BIOS 3.0.23 03/05/2018
[12424406.900826] task: ffff882f62609a00 ti: ffff882f62628000 task.ti: ffff882f62628000
[12424406.909638] RIP: 0010:[<ffffffff810e6029>]  [<ffffffff810e6029>] clockevents_program_event+0x99/0xf0
[12424406.920167] RSP: 0018:ffff882f6262bd48  EFLAGS: 00010046
[12424406.926854] RAX: 0000000000000000 RBX: ffff882fadb52080 RCX: 00000000000006e0
[12424406.935388] RDX: 0000000000719a6f RSI: 00000000fa9b8b6c RDI: 00000000000006e0
[12424406.943934] RBP: ffff882f6262bd60 R08: 0000000000000000 R09: 0000000000000000
[12424406.952491] R10: 00000000000006e43 R11: 0000000000000000 R12: 000000000325625e
[12424406.961059] R13: 0000000000000001 R14: ffff882fadb52aa0 R15: ffff882fadb53160
[12424406.969636] FS:  0000000000000000(0000) GS:ffff882fadb40000(0000) knlGS:0000000000000000
[12424406.979193] CS:  0010 DS: 0000 ES: 0000 CR0: 0000000080050033
[12424406.986424] CR2: 0000000000000000 CR3: 000000011b986000 CR4: 00000000003627e0
[12424406.995070] DR0: 0000000000000000 DR1: 0000000000000000 DR2: 0000000000000000
[12424407.003727] DR3: 0000000000000000 DR6: 00000000fffe0ff0 DR7: 0000000000000400
[12424407.012390] Call Trace:
[12424407.016384]  [<ffffffff810e7b54>] tick_program_event+0x24/0x30
[12424407.023793]  [<ffffffff810a9c02>] __remove_hrtimer+0xd2/0xe0
[12424407.031038]  [<ffffffff810a9f10>] hrtimer_start_range_ns+0x300/0x3d0
[12424407.038984]  [<ffffffff810a9ff2>] hrtimer_start+0x12/0x20
[12424407.045970]  [<ffffffff810e85a8>] tick_nohz_stop_sched_tick+0x2a8/0x2e0
[12424407.054173]  [<ffffffff8101e4a5>] ? native_sched_clock+0x35/0x80
[12424407.061764]  [<ffffffff810e868a>] __tick_nohz_idle_enter+0xaa/0x180
[12424407.069620]  [<ffffffff810e8bcd>] tick_nohz_idle_enter+0x3d/0x70
[12424407.077225]  [<ffffffff810ddfee>] cpu_startup_entry+0x9e/0x290
[12424407.084654]  [<ffffffff8104ac6a>] start_secondary+0x1da/0x250
[12424407.091997] Code: 2a 45 84 ed 74 25 0f 1f 40 00 48 89 df e8 d0 f9 ff ff 5b 41 5c 41 5d 5d c3 66 0f 1f 84 00 00 00 00 00 48 89 fe 4c 89 e7 ff 53 10 <5b> 41 5c 41 5d 5d c3 45 84 ed 75 d3 5b 41 5c 41 5d b8 c2 ff ff
[12424407.115511] RIP  [<ffffffff810e6029>] clockevents_program_event+0x99/0xf0
[12424407.124022]  RSP <ffff882f6262bd48>
[12424407.129193] CR2: 0000000000000000
```

Linux内核自身稳定性

IaaS宕机原因分布


人工误操作
system_other_error
IDC机房事故
机房水泵故障机器过温
network_asw_error
unknown 10%
kernel 19%
hardware
hardware
kernel
unknown
network_asw_error
机房水泵故障机器过温
人工误操作
system_other_error
IDC机房事故

稳定性案例：*内存故障隔离*



稳定性案例：*内存故障隔离*

1	当页面不在被使用，或者migration成功时，隔离物理页面。	支持
2	当隔离失败时，在页面被释放的时候，重新隔离页面。	不支持
3	将错误页面数据信息保持在磁盘或者非易失性内存中，直到硬件被替换	不支持
4	重启系统后，直接隔离已经保持在磁盘或者非易失性内存中的物理页面	不支持
5	当发生内存的MCA错误(CE、UC、UCR)时，需要显示Location信 息，方便后期维修。	不支持
6	mcelog中无法直观显示每个socket、每个channel、及每个DIMM上 发生CE、UC、UCR错误的次数，需要功能增强，方便管理员查看。	不支持

增强

RETRY
当页面正在被使用，且migration失败的时候，增加retry机制

Continuity after isolation
引入一个磁盘数据库，用于保存已经被隔离的错误页面信息; 在服务器启动的时候，直接读取磁盘数据库中的错误页面信息，并主动隔离这 些错误页面

ToolsImprovement
在mcelog工具中，增加一个summary功能，可以直观的显示每个socket上，每个 channel，每个DIMM上，发送CE、UC、及UCR的次数

https://lore.kernel.org/lkml/1531452366-11661-2-git-send-email-n-horiguchi@ah.jp.nec.com/#r

挑战#5： 故障域

局部故障导致系统整体宕机

故障域建议

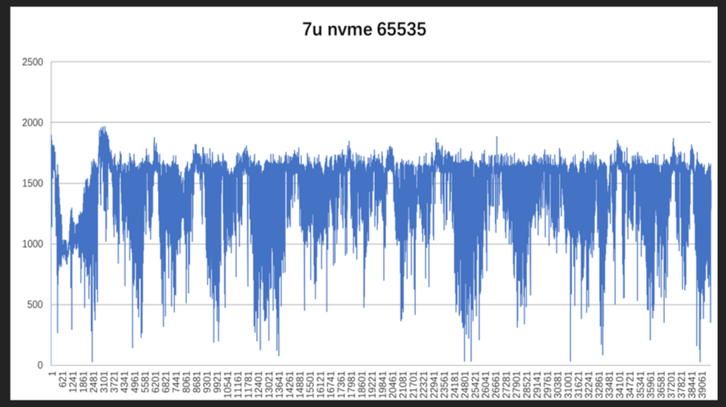
- **内核功能分区**

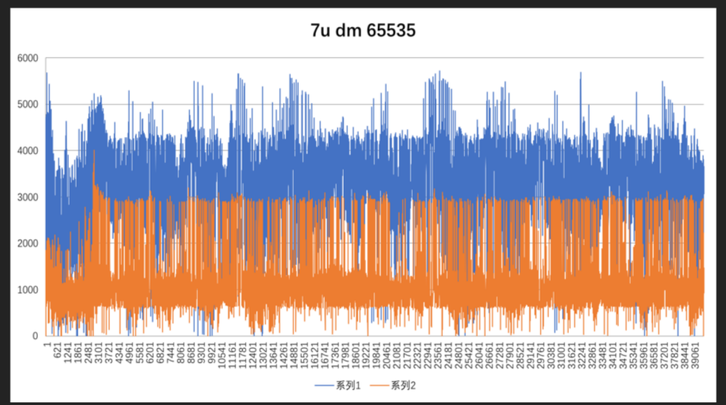
  容器技术往下走，把内核按照功能进行分区

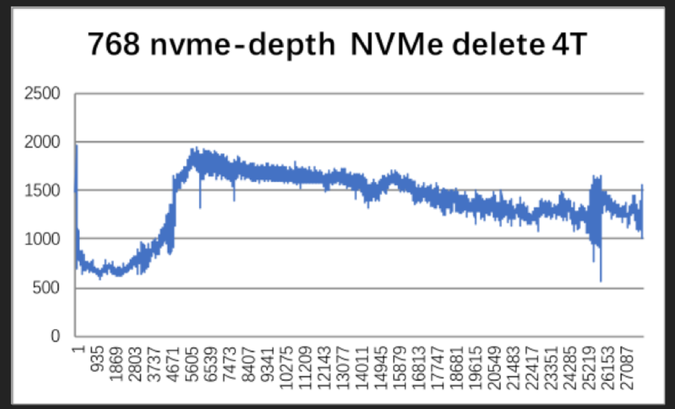
  在内核功能panic以后，让整个内核有规划的宕掉或者不宕掉

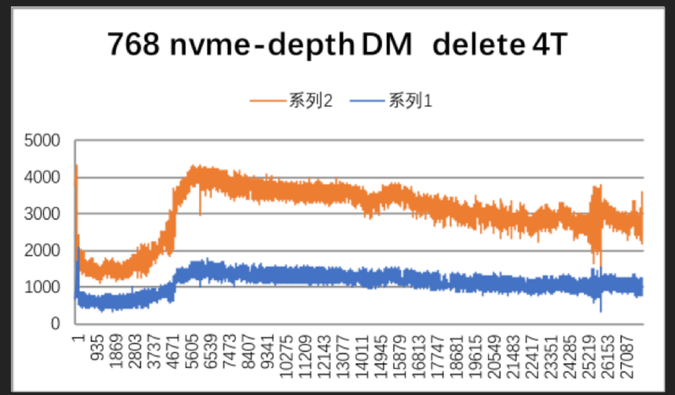
- **用户态驱动**

挑战#6： QoS


7u nvme 65535

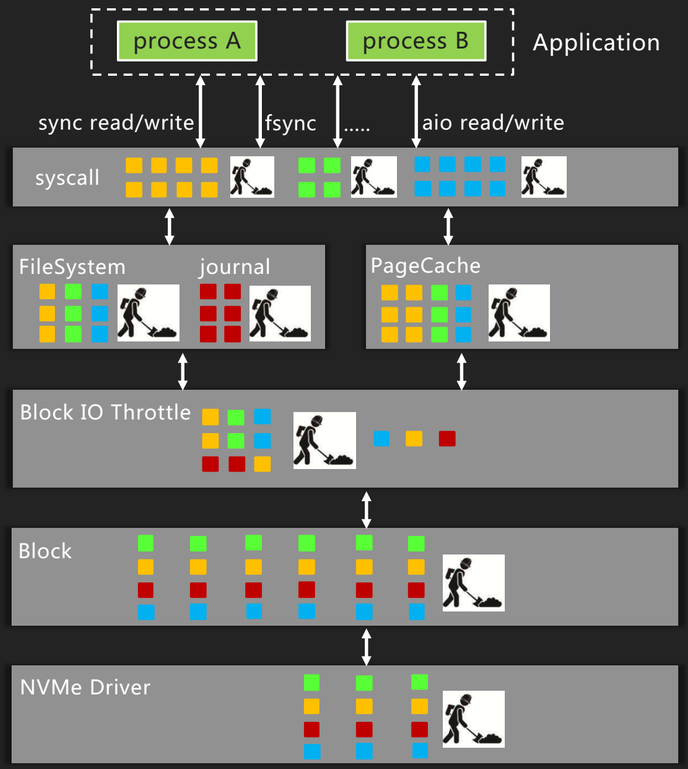

7u dm 65535
系列1
系列2


768 nvme-depth NVMe delete 4T

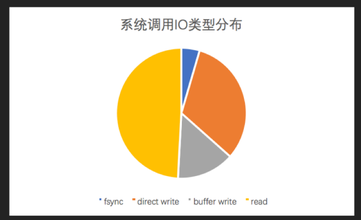

768 nvme-depth DM delete 4T
系列2
系列1

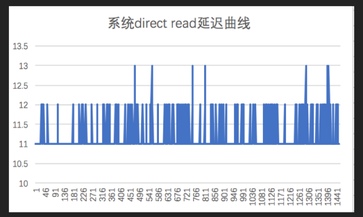
AliIPF *Alibaba IO Profiler*

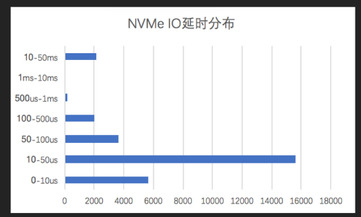
线上IO抖动自动检测、分析，精确IO画像

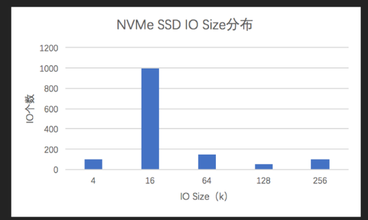


- **IO性能诊断**
- **IO画像**



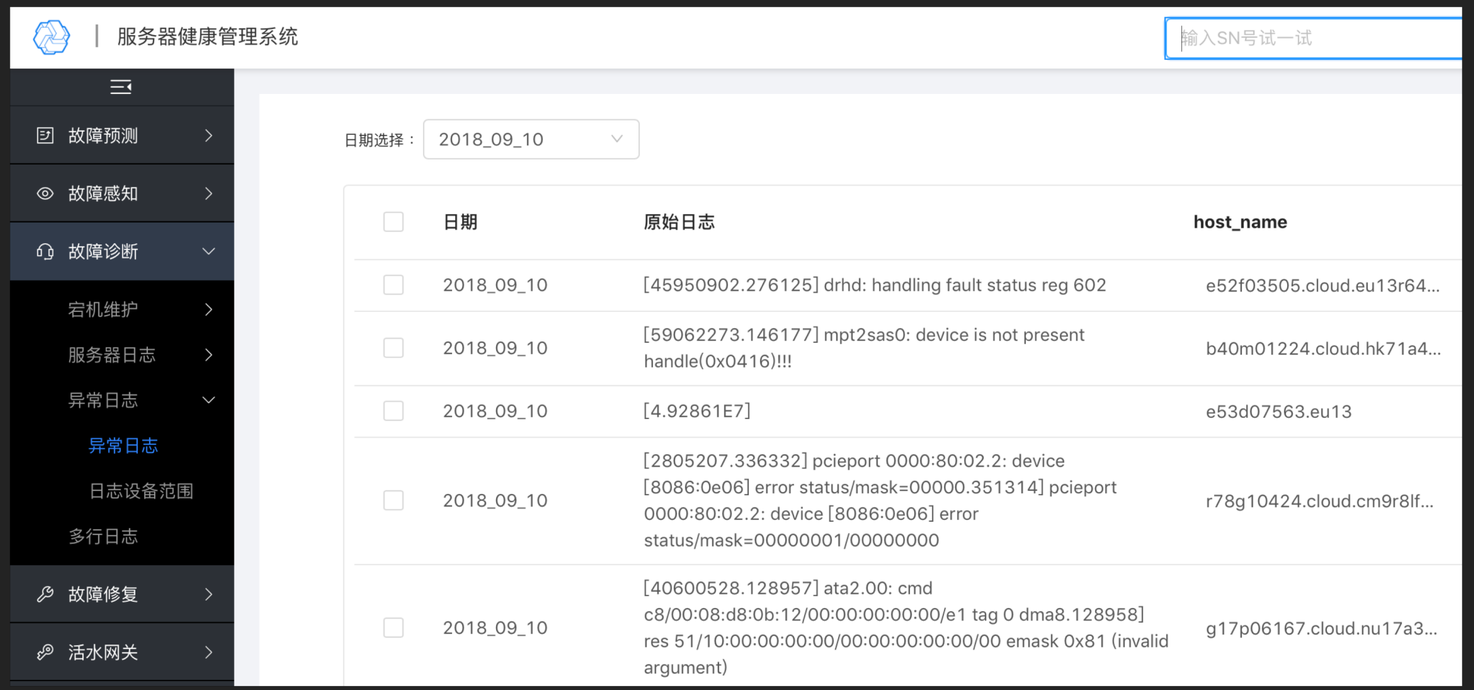






自动化、智能化的Linux内核？

- ✓ 故障诊断
- ✓ 故障预测
- ✓ 故障修复



Alibaba Group
阿里巴巴集团

浙江大学
1897
ZHEJIANG UNIVERSITY

谢谢！