RAS-MCA Recovery

由 yaoaili [么爱利]创建于九月 01, 2020

1、MCA Recovery 简述以及收益

1.1 什么是 MCA Recovey:

MCA Recovery 是 Intel Advance CPU 的一个 RAS 特性（purley），是 Machine Check Recovery 的简称；这个特性需要硬件、BIOS、Kernel、运维相配合，目的是降低内存故障带来的系统影响。

MCA Recovery 又简称内存隔离技术，其主要工作方式和步骤如下:

- a) 系统 CPU 检测到内存出现了故障，触发系统处理流程;
- b) 系统判断故障是否可恢复，如果不可恢复，系统宕机;
- c) 系统可恢复，系统将故障内存也踢出系统的内存管理系统，同时进行标记;

1. 进行其他可能的操作，比如 kill 相关进程;
2. 通过系统 log 或者 BMC sel log 触发虚机迁移策略，以及后续的运维。

1.2 可能收益

对于服务器来讲，不具备此 Feature 的机器在内存发生故障、产生 Uncorrectable Error（后用 UE 指代）时候，都会导致一个必然的结果，服务器宕机；随着服务器内存的逐渐增大，由于内存 UE 导致的宕机情况大概占了云数据中心硬件宕机 50%的情况。在此 feature 使能的情况下，根据内存的区域区分为两种 case：一种是内核代码区域，另一种非内核代码区域，当内

存 UE 发生在内核代码区域时，系统会宕机，当内存 UE 发生在非内核代码区域时，系统不会宕机，最大的可能是 kill 相关进程。目前尚无内核内存与非内核内存的比例真是数据，我们假定内核与非内核内存的比例为 **1:4**，**内核与非内核内存 UE 发生的概率相同**，那么我们会降低内存宕机可能性的 **80%**。

MCA Recovery Enable	No	Enabled
系统行为及收益	100%宕机	小概率宕机，大概率不宕机，不宕机机器可触发迁移

2、MCA Recovery 原理

2.1 Reliability, Availability, and Serviceability

1.Reliability 可靠性，正确运行计算，不出错误结果；

2.Availability 机器运行的连续性，长时间运行不宕机；

3.Serviceability 运维特性，出现了问题快速恢复的能力

正常情况下，一台 server 在软件不出问题的情况下，理想情况是永远的运行下去，但是硬件上由于设计的原因，老化的原因，环境的原因，硬件会出现问题，导致 server 不能继续正确运行，或者系统软件或者其他软件能处理这些错误，或者需要宕机，更换硬件，修好会服务才能继续运行。对于这些错误系统需要感知并记录这些错误的信息，通过对错误信息的分析提供下一步的处理策略。如果系统不能感知硬件出现的问题，那么这种错误会被消费，可能导致错误的结果，也可能导致系统不明原因的崩溃或者异常。这种情况对于数据中心的服务器来讲不可接受。

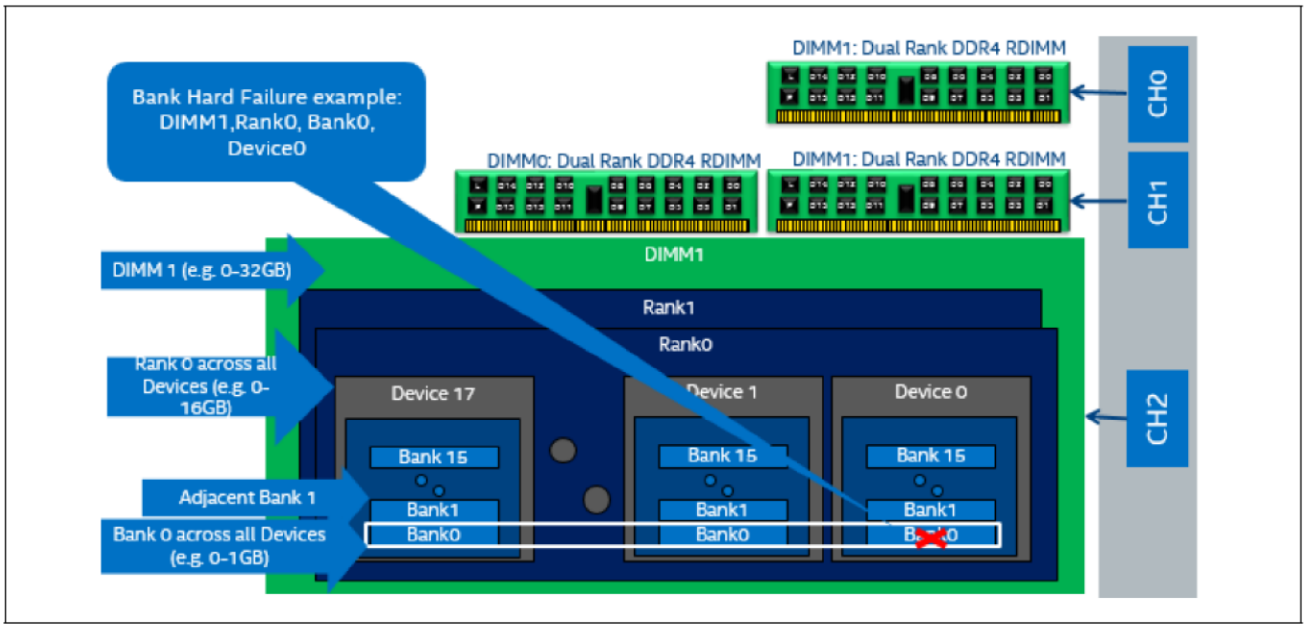
Intel 在其 CPU 内部设计了 MCA 架构模块，来记录不同硬件的产生的错误，并在硬件层面对该硬件错误的级别进行定义，然后通过中断请求 OS，BMC，BIOS 来进行处理；

MCA 记录故障信息的部分由多组 bank 寄存器沟通，每个 bank 记录某一部件相关的硬件错误信息。

在云数据中心宕机的统计中，其中内存故障占了很大的比例，主要是因为目前的内存容量越来越来，频率越来越高，芯片密度越来越高，内存出错的可能性随着这些因素以及时间的因素逐渐加大。

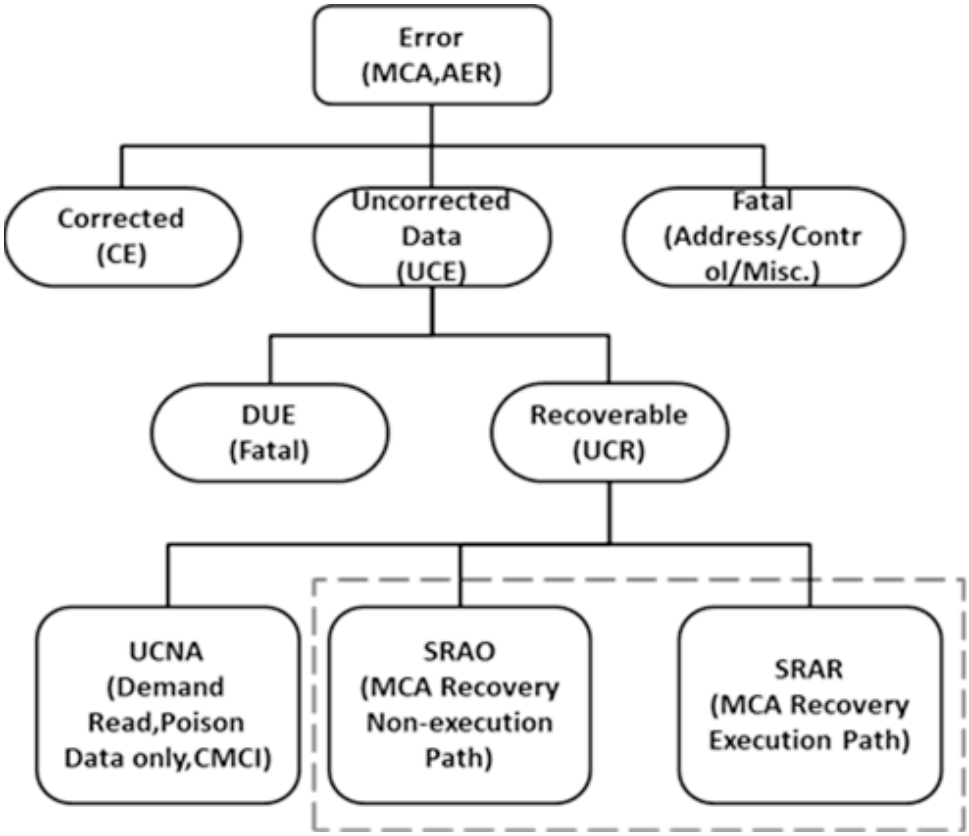
2.2 ECC内存

Figure 8-2. Definitions for Memory Regions: DIMM Devices (Chips), Banks, Ranks and DIMMs



2.3 内存故障：

2.3.1 错误的分类：



2.3.2 内存故障的来源：

Fault Type	Possible Causes (Examples only)	Fault Coverage (RAS Features)
Bit (Cell) error	High energy particle strike -Soft Error (SE). Transient error.	SDDC, Demand Scrub, Patrol Scrub
	Persistent fault	PCLS
Row error	Persistent fault	SDDC, PPR
Bank failure	Persistent fault	ADC (SR), ADDDC (MR)+1[1]
Rank/Device failure	Persistent fault	ADDDC (MR)+1[1], SDDC+1[1], Rank Sparing
Addr/Cmd error	Transient/ Persistent fault	DDR4 CMD/ADDR Parity error check and retry
Multi-device error (UCE)	Persistent fault + SE	MCA-recovery[1], Address Range Mirroring[1]
Connector failure	Electrical noise. Transient error.	Transaction retry, DDR4 Write Data CRC
	Wear-out or manufacturing defect. Persistent error.	Memory disable/map-out for FRB
Channel failure	Board defect	

2.4 MCA Recovery 的一些处理流程：

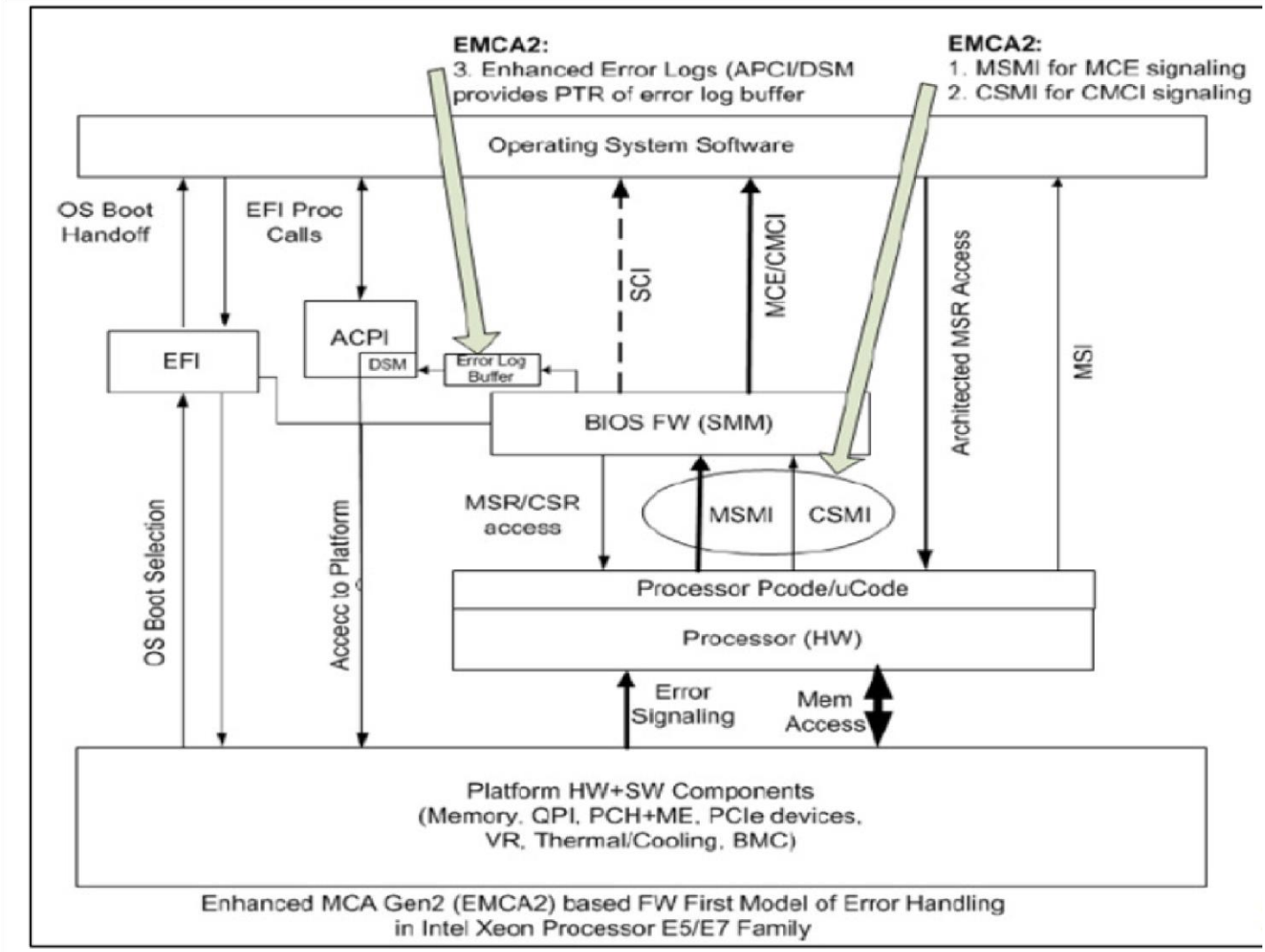


Figure 4-1. Firmware First Model EMCA2 Error Handling

2.5 MCA recovery 包含如下 5 个特性：

1 R_SYS02 MCA Recovery – Execution Path Advanced 执行路径

这种错误是由内核或者业务代码访问触发的 UCE，此种错误，对于位于内核段的错误，硬件通告级别为 panic，系统会宕机，对于处于非内核段的数据，

系统会选择访问 UCE 对应物理页的进程（可能是多个）kill，然后对相关页进行 unmap，标记为 poision

2 R_SYS02 MCA Recovery – Non-execution Path Advanced 非执行路径

此特性针对两种情况，一种是 patrol scrub 发现的 UCE，另一种是 LLC Explicit Write Back transaction 触发的 UCE，此两种情况的故障对于包含

MCA Recovery 特性的系统来讲为 SRAO 类型错误；

3 R_SYS03 MCA 2.0 Recovery (as per EMCA Gen2 architecture) Advanced ---执行路径

EMCA Gen2 is a capability that allows firmware to intercept errors triggered via Machine

Check Architecture (corrected and uncorrected errors) enabling a Firmware First Model (FFM) of error handling and possible recovery.

BIOS 协助处理的故障恢复，FFM 模式；

4 R_CPU10 Complex Instruction Recovery Improvements (New in Purley Platform)

CPU feature，了解即可，软件无法干预；

5 R_SYS11 Local Machine Check (LMCE) based Recovery (New in Purley Platform)

通常情况吓，发生 UE 的情况，MCA 架构会广播 MCE 中断到所有 core，所有的 core 进入 do_machine_check,然后各 CPU 按照

MCE 进入顺序依次进行故障扫描，

由选出的 monarch 来处理，其他 core 等待并同步，这个特性打开后，发生 UE 后，MCA 架构会将中断触发给受影响的单个 core，由单个 core 完成处理；

2.6 MCA 故障处理的流程：

1.**系统访问内存对应的地址(SRAR)**，ECC 校验失败，触发内存错误，进入 MCE 中断；

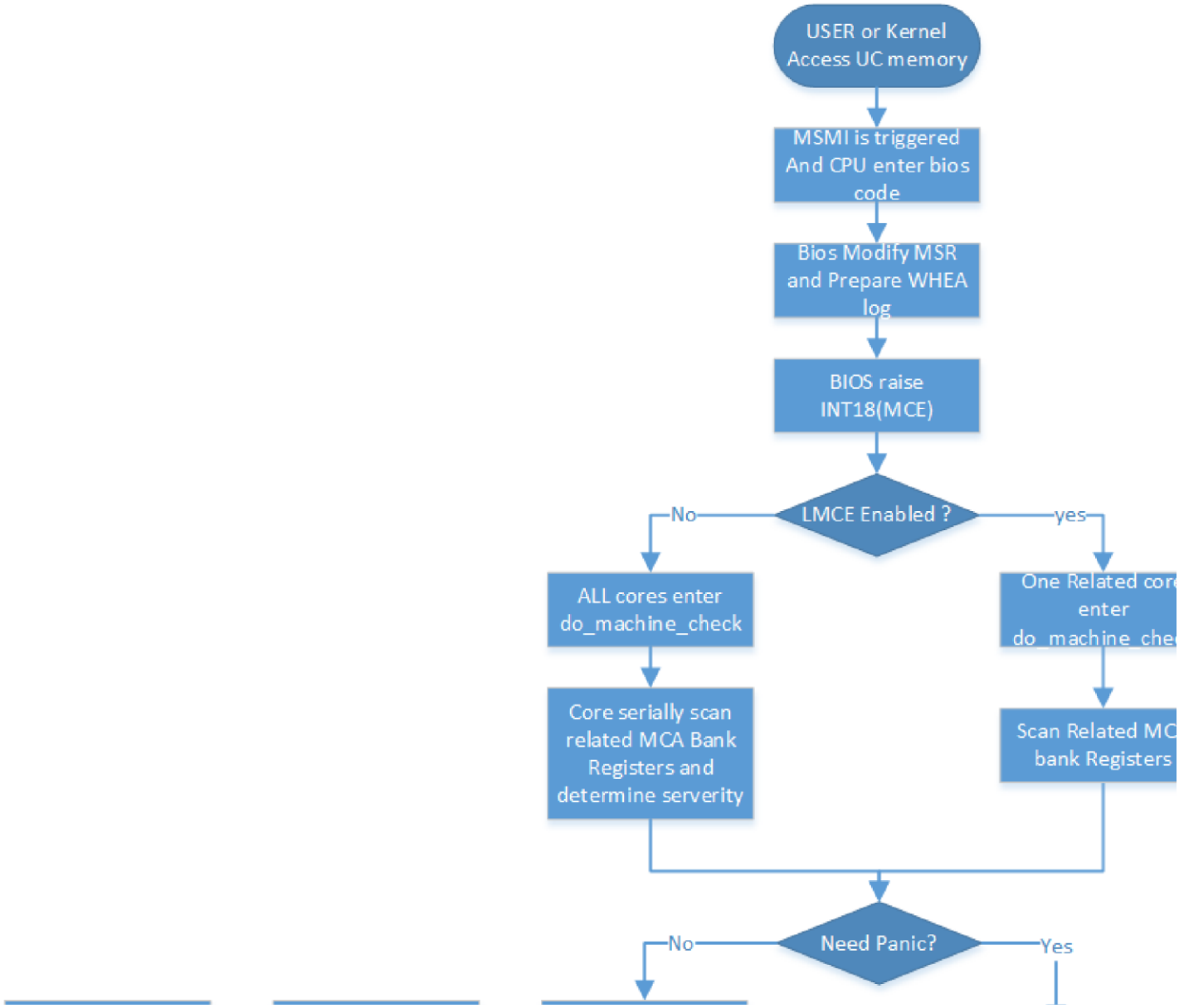
2.在中断函数中，根据 MCA MSR 确定需要采取的操作；如果需要 panic；

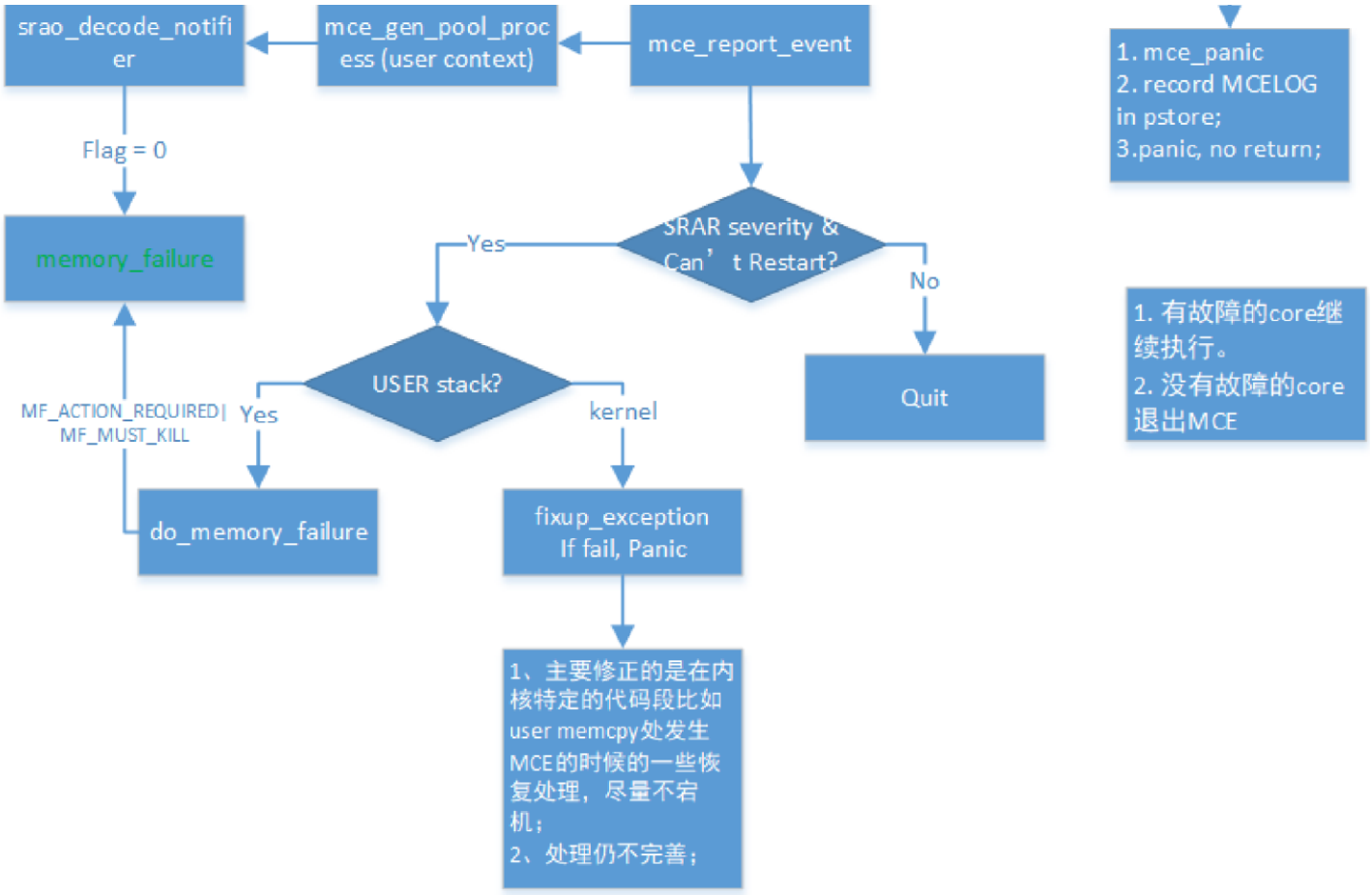
3.对于不需要的 panic 的情况，对相关内存对应的 page 进行隔离，都需要打上 poision 标记，有两种情况：

a) 此 page 正在被使用，那么需要进行相关处理，使得 page 相关的资源全部释放，比如 ummap，kill 等，由 MCA 保留一个 refcount，page 既不在 Buddy 系统里，也不再各种 LRU 链表里，内核失去了对该 page 的 track；

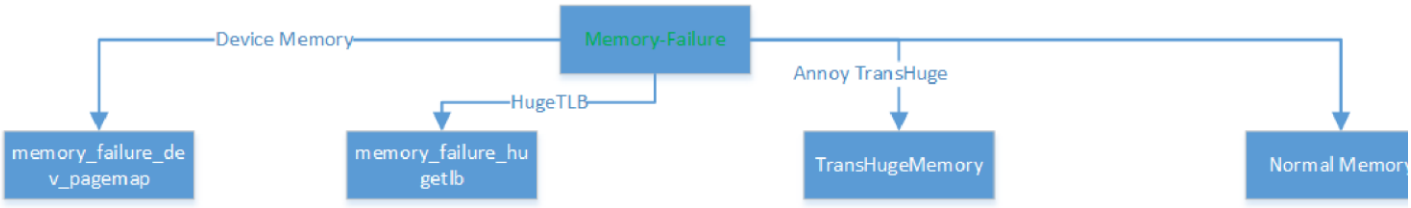
b) 此 page 没有被使用，那么会将 page 返回 buddy 系统，在系统继续运行，可能会将此 page 分配，在将 page 从 buddy 链表提出后，会对页面进行 check，如果看到 poision，则会重新申请页面，内核在这个时候也失去了对此 page 的 track；

2.6.1 Do_machine_check:

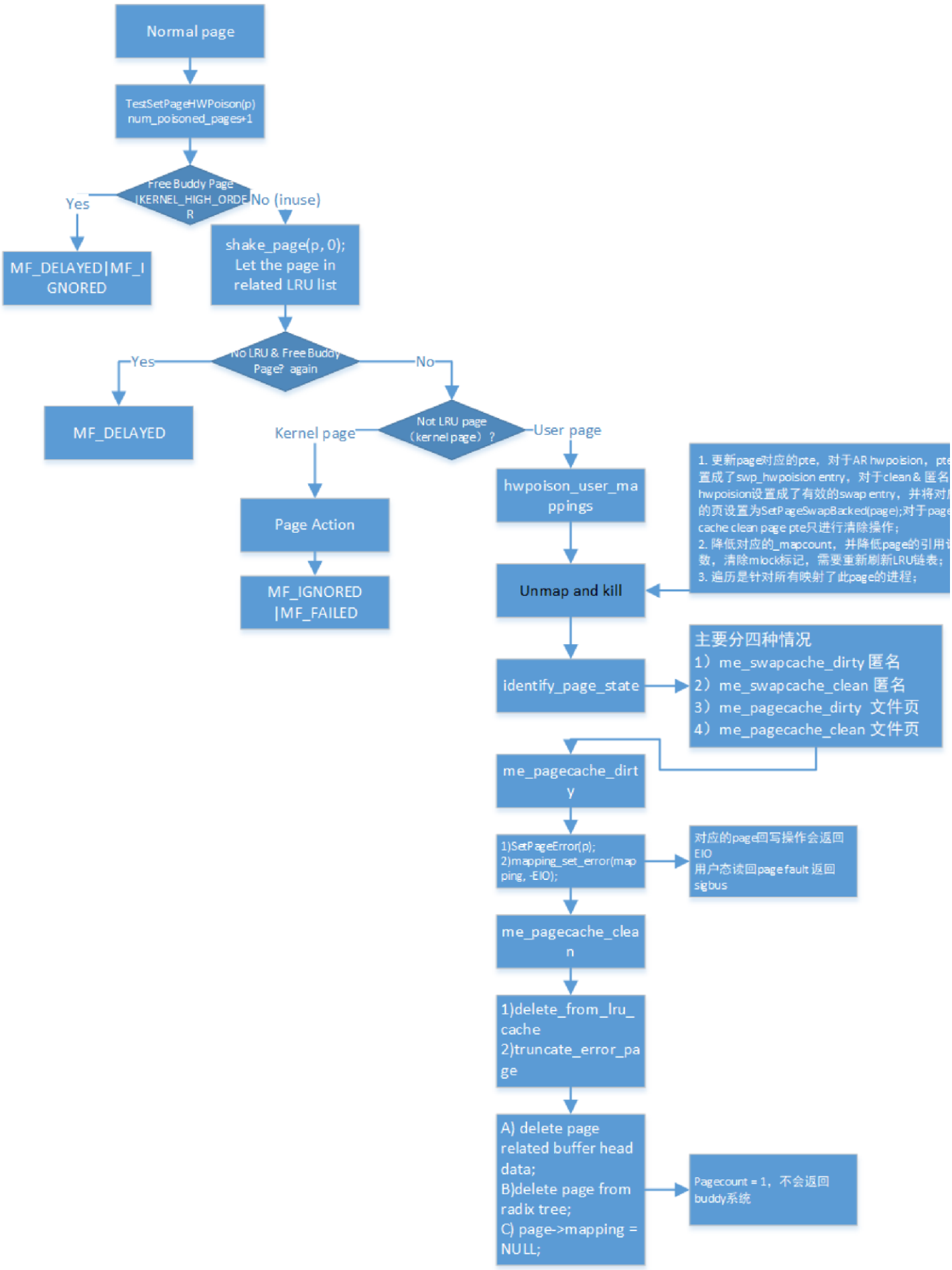

USER or Kernel Access UC memory
MSMI is triggered And CPU enter bios code
Bios Modify MSR and Prepare WHEA log
BIOS raise INT18(MCE)
LMCE Enabled ?
No
yes
ALL cores enter do_machine_check
Core serially scan related MCA Bank Registers and determine serverity
Scan Related MC bank Registers
Need Panic?
No
Yes



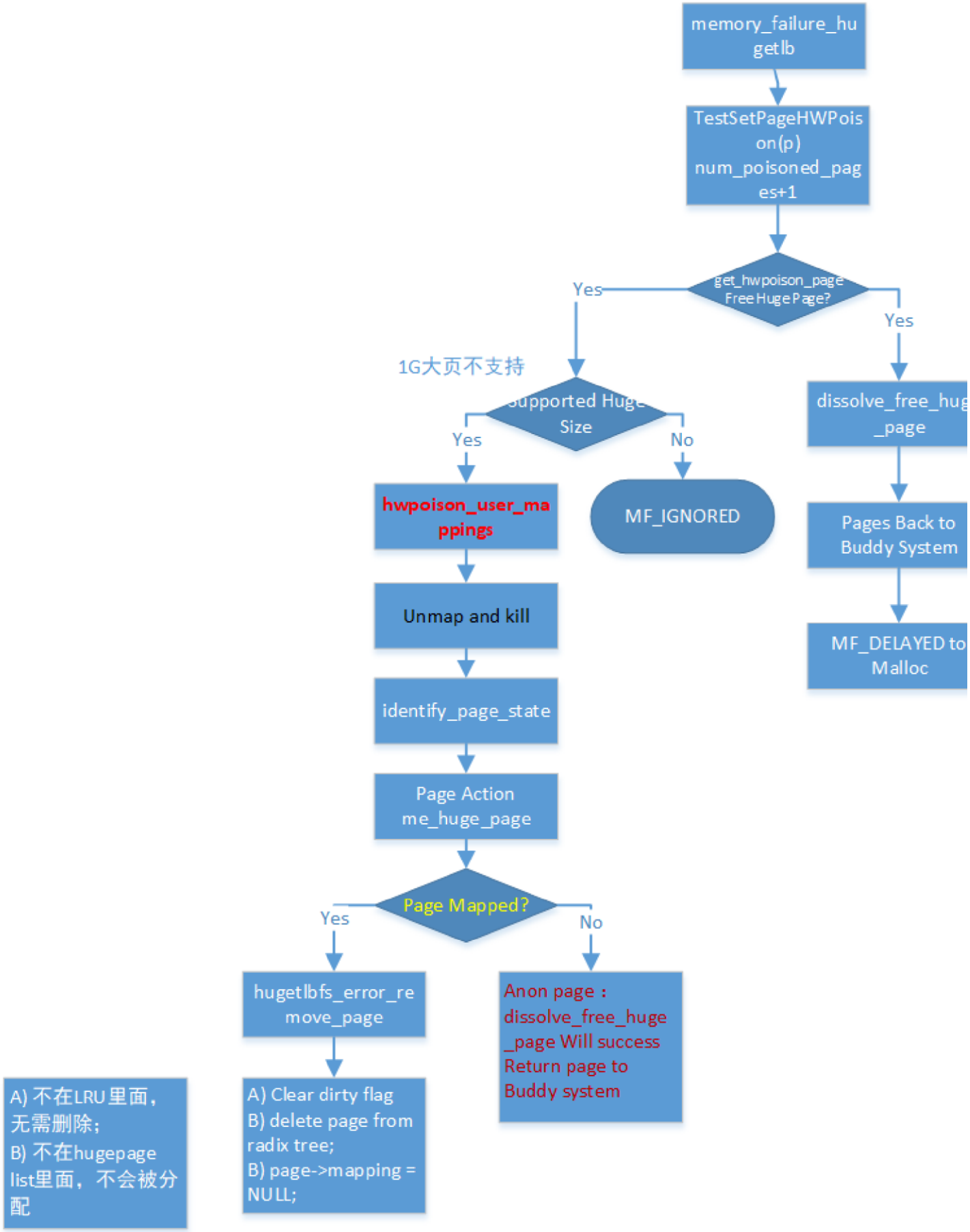
2.6.2 Memory_failure



2.6.3 Normal_page_process



2.6.4 memory_failure_hugetlb



3、运维策略（待更新）

1. 运维工具需要检测 BMC SEL log 和 kernel 系统日志、mcelog 日志，并根据不同的情况，进行区分处理
2. 宕机的 case，需要在系统重启后，自动收集如下日志：1、BMC SEL LOG 2、VM Core 、vmcore-dmesg.txt、3、Message、mcelog 4、宕机的时间 上述日志合并录入 bugzilla；运维根据宕机的情况，通知客户，启用备用机等等； 3. 对于不宕机的 case：系统需要监控系统日志，收集 1、BMC SEL LOG 2、Message、mcelog，并分析系统日志里的恢复情况，无论是否有 kill 的情况，都需要触发在线迁移的流程，将业务虚机转至无故障的机器上；
4. ODM 介入，故障机器检修；

4、Test Related:

4.1 kernel config: (default is open)

- CONFIG_X86_MCE=y
- CONFIG_DEBUG_FS=y
- CONFIG_X86_MCE_INTEL=y
- CONFIG_ACPI_APEI=y
- CONFIG_ACPI_APEI_GHES=y
- CONFIG_ACPI_APEI_MEMORY_FAILURE=y

CONFIG_ACPI_APEI_EINJ=m or CONFIG_ACPI_APEI_EINJ=y

4.2 BIOS config:

- — System Errors: enable
- — System Poison: enable
- — S/W Error Injection Support (Test need)

4.3 Test tool

- git://git.kernel.org/pub/scm/utils/cpu/mce/mce-inject.git
- git://git.kernel.org/pub/scm/utils/cpu/mce/mce-test.git
- make & make install

4.4 Test process

- cd mce-test/work
- touch function & vi function

```
APEI-INJ cases/function/apei-inj/runtest.sh
SRAR-DCU cases/function/core_recovery/runtest_dcu.sh
SRAR-IFU cases/function/core_recovery/runtest_ifu.sh
#HWpoison
```

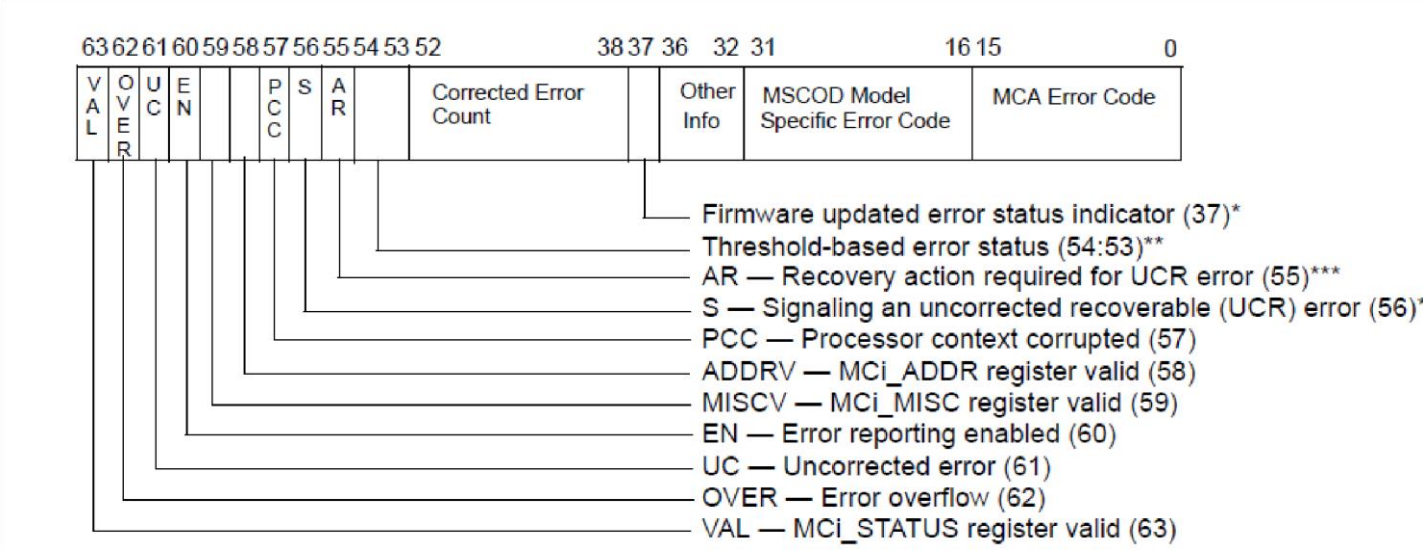
```
HWPOISON-SOFT cases/function/hwpoison/run_soft.sh
HWPOISON-HARD cases/function/hwpoison/run_hard.sh
HWPOISON-HUGEPAGE cases/function/hwpoison/run_hugepage.sh
HWPOISON-HUGEPAGE-OVERCOMMIT cases/function/hwpoison/run_hugepage_overcommit.sh
HWPOISON-THP cases/function/hwpoison/run_thp.sh
```

- save & exit
- cd ../mce-test
- ./runmcetest

5. 其他:

PCC (processor context corrupt)：(Intel 目前未反馈 CPU 设置此标记的逻辑)

Indicates (when set) that the state of the processor might have been corrupted by the error condition detected and that reliable restarting of the processor may not be possible. When clear, this flag indicates that the error did not affect the processor's state, and software may be able to restart.



无标签