白皮书

英特尔 MCA+MFP 助力京东构建稳定、高效的云服务

JDT 京东科技

“云主机质量的提升对客户至关重要，可以极大的降低业务中断，为了达成云主机 99.975% 的 SLA 承诺，京东云和英特尔深度合作，利用了英特尔的 MCA 和 MFP 技术，缩小故障域执行故障隔离或提前在线迁移，降低了内存 CE 和 UE 故障对云主机的影响。”

—— 龚义成
京东科技京东云事业群 IaaS 产品研发负责人

“一直以来系统可靠性和稳定性都是京东云服务的最基本的服务宗旨，京东云在基础技术突破上也从未停止探索，这次京东云联合英特尔对 CPU 的 Advanced RAS 功能进行定制优化，同时围绕英特尔 MFP 故障预测模型结合使用场景进行联合研发,实现内存故障预测和修复。通过这一技术将有效提升京东云服务的可靠性和稳定性。我们也期待通过双方努力持续提升硬件系统的可靠性和稳定性。”

—— 陈国峰
京东科技京东云事业群硬件研发负责人

目录

引言

“英特尔 MCA Recovery 与 MFP 技术的成功部署，帮助京东云主机因内存故障导致的宕机概率降低了 40%，内存故障条件下的热迁移成功率提升了 50%，大大提高了主机的可靠性和稳定性，改善了用户体验，提升了 SLA，使京东云在激烈的市场竞争中，占有技术优势。”

数字经济高速发展的今天，下面的生活场景已经逐渐被人们所熟知：

“一位货运司机，不再通过线下寻客，在手机上安装一个应用，通过货运平台接收订单，这样可以减少货车的空驶率，货运回款有所保障，既降低成本又增加收入。”在货运平台的背后是每天几十万货运司机与货主的交流信息，过亿的运输货值，几百万的货车轨迹坐标，数亿的行驶数据，PB 级的数据量。

“后疫情时代，在线学习成为常态，一名小学生在线上平台上网课，与教师无延迟交流，使用各种教学工具学习知识。”在线教育系统的提供商每天需要面对日活千万的用户，提供一百余种教学功能和教具，覆盖全球六大洲，1080P 的高清直播，延迟需要低于 200ms。

“又到一年 618，京东的商家们正在厉兵秣马、积极备战。一位商家正在通过京东的平台更新产品、制作广告、设定折扣；还可以定制智能客服应对海量咨询，联系主播为活动热身，只待促销大卖。”然而大促期间，京东平台也会迎来数亿次的攻击，千亿元的订单，万亿级流量的考验。

“互联网+”改变了人们的生活方式，颠覆了传统行业的经营模式，但这些都需要有安全、稳定、高效的基础技术架构做保障，京东云就为这些场景提供了完美的解决方案。

背景

京东云是京东科技集团旗下领先的云计算品牌，依托于京东科技集团在人工智能、大数据、云计算、物联网领域的前沿科技能力，提供包含公有云、专有云、混合云在内的多云、安全、可信赖的基础云服务，为全球互联网、金融、城市、交通、能源等客户提供领先的云计算服务与行业解决方案。2016 年 4 月京东云正式商用，进军中国云计算市场；2017 年 6 月，京东业务全部上云；2021 年 4 月，京东云 IaaS 市场占有率升至中国第五，跻身国内云计算第一梯队。

作为全球容器化最彻底的云平台之一，京东云拥有全球最大规模的 Docker 集群、全球最大规模的 Kubernetes 集群，支撑万亿级电商交易，实现京东 618 购物节订单 100% 云上完成、以及京东物流、京东健康全量上云。历经京东 618、11.11、春晚等万亿级流量洪峰考验，京东云服务多个视频、媒体、在线教育、游戏等客户，服务最高可用性保证达 99.995%。

痛点：业务激增对京东云稳定与可靠性提出更高要求

如今京东云覆盖各个行业领域超过 2500 家的合作伙伴，随着用户规模不断增大，特定行业与云原生类用户对应用开发和运营模式提出许多新的要求，传统用户也正在将更多复杂业务迁移上云，这些持续变化的技术需求对京东云服务提出新的挑战。

作为云服务的核心资源云主机，它的可靠性、可用性、可维护性直接决定了云服务的质量和水平。如今硬件故障的发生是造成主机宕机的重要因素，传统方式下，一组服务停止工作只会影响到自己的业务和用户，但是在云环境下，服务终止将会导致云服务提供商违反服务水平协议（SLA）并造成巨大的经济损失。在众多的硬件故障中，内存错误是当今数据中心中所面对的最严重故障之一。目前京东云数据中心内存错误在整体硬件故障中的占比达到 37%，为此京东云建立了完善的云主机故障预测和恢复系统，希望通过对内存错误的发现与预测，通过在线快速迁移恢复技术，减少内存错误对云主机造成的影响。

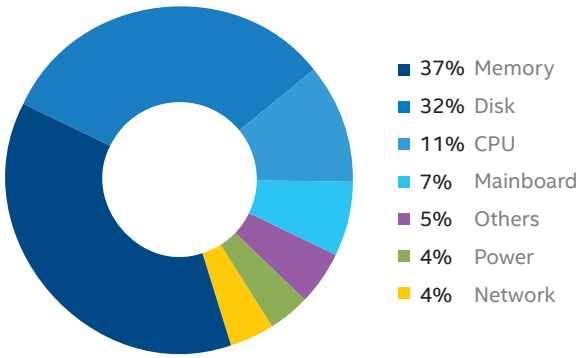


图 1　京东云硬件故障分布

但由于目前云主机中各软硬件系统兼容性的影响，恢复系统仍然无法对许多内存故障引起的宕机，进行快速恢复。例如恢复系统无法对存储优化型云主机进行热迁移；恢复系统在日常巡检时及时发现内存错误，在热迁移过程中出现系统宕机等问题，增加了云主机的故障率。

如果能建立一套实时洞察云主机内存状况、预测潜在的内存故障并对内存错误进行有效的回复的解决方案，对提高京东云服务的稳定性与可靠性，提高终端用户的 SLA，降低京东云数据中心的总体拥有成本都有极大的帮助。

英特尔 MCA Recovery+MFP，助力京东云提供高效稳定的服务

京东云与英特尔在云计算领域一直保持着紧密而广泛的合作，为终端用户提供专业而且高性价比的云服务是双方合作的初衷。为了解决内存错误的困扰，双方再次携手，通过引入英特尔 MCA Recovery 与 Memory Failure Prediction（MFP）技术，结合京东云的故障恢复系统，用来降低内存错误对京东云主机稳定性的影响。

内存错误

目前主机出现的内存错误主要分为可纠正错误（Corrected Error，简称 CE）和不可纠正错误（Uncorrected Error，简称 UE）。可纠正错误目前最为普遍的解决方案是通过纠错码（ECC）克服双列直插式内存模块（DIMM）的一些可纠正错误。

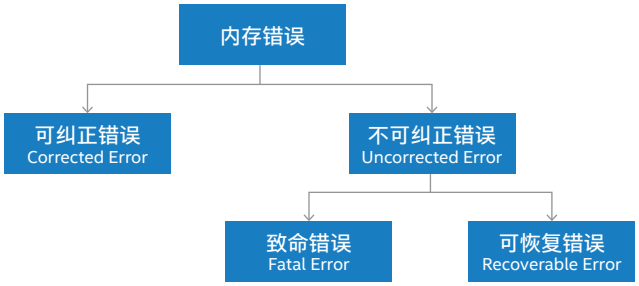


图 2　内存错误分类 1

不可纠正错误（UE）通常会造成比较严重的灾难性后果，如主机操作系统挂起，系统崩溃、宕机。UE 错误也可以分为 Fatal Error、SRAR、SRAO 以及 UCNA。

1. Fatal Error：非常严重的 UE 错误。此类错误系统无法对其修复，该错误会导致处理器内部处于混乱或者不稳定的状态，只能通过复位系统进行恢复。出现这种 UE 错误目前暂无好的恢复手段。

2. RAR（Software Recoverable Action Required）：发生这种错误后，操作系统 / 应用程序需要执行某种操作（例如隔离 / 终止失败线程）来恢复此无法纠正的错误。此类错误是恢复技术可以重点恢复的错误类型。

3. SRAO（Software Recoverable Action Optional）：出现这种错误后，操作系统 / 应用程序根据用户设定的策略选择执行某种操作（例如隔离 / 终止失败线程），用以恢复此类错误。

4. UCNA（Uncorrectable Error No Action required）：出现的错误不是位于关键路径上，该错误没有触发 MCE，通常不需要采取任何操作。

基于内存错误的分析和了解，可以判断出，制定一套针对 SRAR 与 SRAO 两种 UE 错误的预测+恢复的技术解决方案，可以有效降低内存故障对主机的影响。在经过双方技术专家的反复测试与权衡，最终选择英特尔 MCA Recovery 与 MFP 技术解决此类问题。

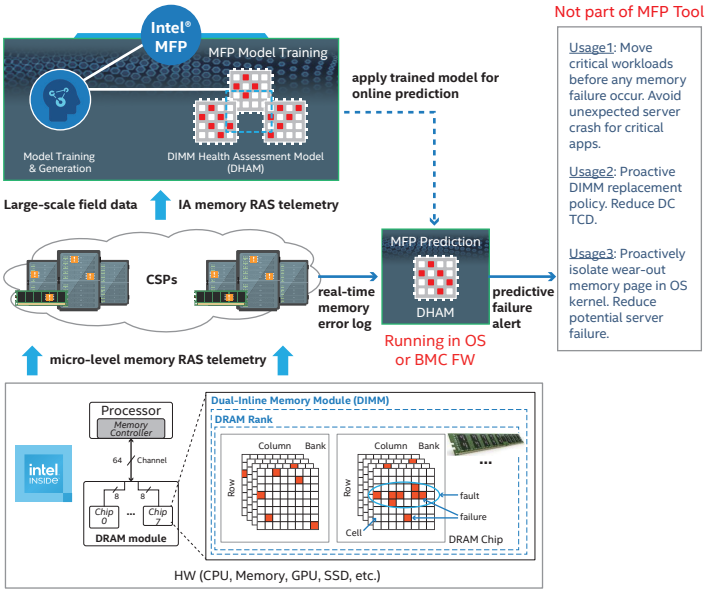
	Mci_STATUS							MCG_STATUS						
								Errored Thread		Other Threads				
	Val ID	UC	PCC	Service	AR (Action Required)	ADDRV	MISCV	RIPV	EIPV	RIPV	EIPV	Signaling	ADDR in Kernel Space	SW Action
Uncorrected Errors	1	1	1	×	×	×	×	×	×	×	×	MCERR	×	System Crash
SRAR - Instruction	1	1	0	1	1	1	1	0	0	1	0	MCERR	NO	Take Specific Recovery Action
SRAR - Instruction	1	1	0	1	1	1	1	0	0	1	0	MCERR	YES	Kerne Panic
SRAR - Data Load	1	1	0	1	1	1	1	1	1	1	0	MCERR	NO	Take Specific Recovery Action
SRAR - Data Load	1	1	0	1	1	1	1	1	1	1	0	MCERR	YES	May Kernel Panic
SRAO	1	1	0	1	0	1	1	1	0	1	0	MCERR	×	Optional for Recovery Action
UCNA	1	1	0	1	0	1	1	×	×	×	×	CMCI	×	Log the Error and Optional for Recovery Action
CE	1	0	0	1	0	1	1	×	×	×	×	CMCI	×	Log the Error and No Corrective Action Required

表 1　内存错误分类 2

MFP

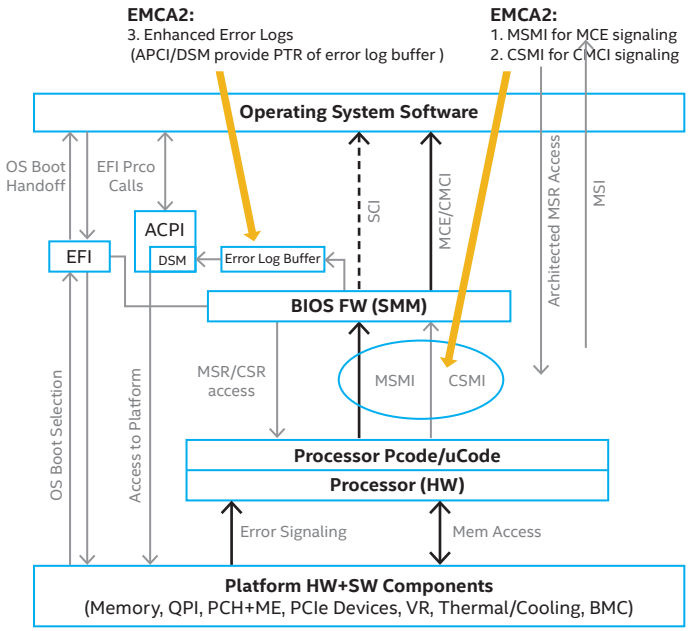
英特尔® MFP[1] 是一种通过主动内存故障管理提高主机可靠性的数据驱动技术。它通过对历史故障数据的学习，可以自主的对主机内存故障做出预测，并在发生灾难性结果前通知系统管理员。

英特尔® MFP 通过对成千上万的 EDAC 日志对内存微观层面故障数据进行学习和数据挖掘，以此训练和建立 DIMM 健康评估模型（DHAM）。MFP 部署后，会实时监控主机内存运行状况，分析主机不同层面的内存错误，包括 DIMM、rank、bank、column、row 和 cell等，将主机内存状况与 DIMM 健康评估模型进行对比，以预测发生内存故障的可能。



MCA Recovery

MCA Recovery[2] 是“英特尔高级 RAS”功能，利用 CPU 的 MCA 架构体系，结合固件（比如 UEFI 固件）对发现的不可纠正的硬件错误（UE）进行隔离，从而使系统从这类错误中恢复出来的一种技术。



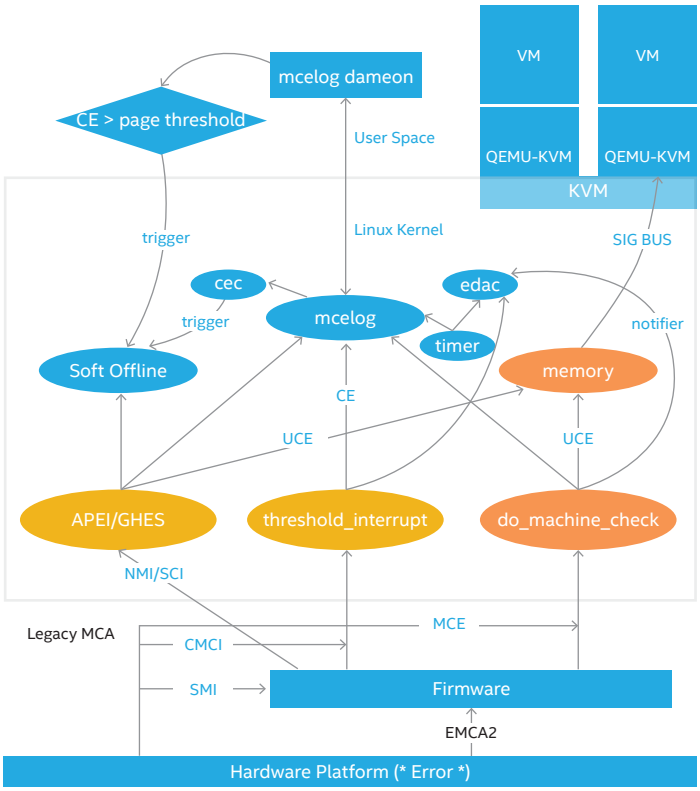
Enhanced MCA Gen2 (EMCA2) based FW First Model of Error Handling in Intel® Xeon Processor E5/E7 Family

图 3　MCA recovery 技术原理图

[1] Intel® Memory Failure Prediction: https://www.intel.com/content/www/us/en/software/intel-memory-failure-prediction.html
[2] Intel MCA Recovery 技术介绍：https://partneruniversity-prc.intel.cn/diweb/catalog/launch/package/4/eid/777016

使用 MCA Recovery 技术进行恢复有几个先决条件需要满足：

1. 内存 UE 是非致命错误。
2. 内存故障地址不在内核空间中。
3. 访问错误地址的应用或进程可以被 kill。



UCE: Uncorrected Error
CE: Corrected Error

图 4　MCA 内存错误恢复流程图

如图 4 所示，当一个内存错误发生后，CPU 会通过 MSMI 通知 BIOS，同时 CPU 会进入 SMM Mode 模式（错误处理模式）；BIOS 接到指令后，会对错误的数据进行收集整理并把它存入 API TABLE 或者 BMC 中，同时发送指令给 OS，告知发生的错误及错误类别；当 OS 接收到指令后，会把发生错误的内存地址进行隔离，同时把仍然使用错误地址数据的应用删掉，从而保证主机不会发生宕机，实现错误恢复。

MCA Recovery 与 MFP 的部署，配合京东云的故障恢复系统，大幅降低京东云主机因内存故障造成的系统崩溃。当某一个节点出现潜在 UE 可能时，京东云可用性分析系统，通过英特尔® MFP 技术可以实时 UE 错误预警，从而触发第一层系统保护 —— 基于内存故障预测的热迁移，避免潜在内故障导致的云主机宕机。如果 UE 故障发生在 MFP 预测范围外，触发第二层系统保护 —— 通过 MCA Recovery 进行恢复，故障恢复系统在 MCA Recovery 的帮助下隔离受影响的内存页，防止页面被其它应用/进程再次使用，如果内核能够成功执行恢复，只要没有故障，系统就可以保持在线。同时数据中心维护将收集故障日志以确定哪个 DIMM 发生故障，将出现 UE 的内存页面所在的云主机在线迁移到一个健康的节点，并将故障节点上报给服务器管理系统进行内存更换。

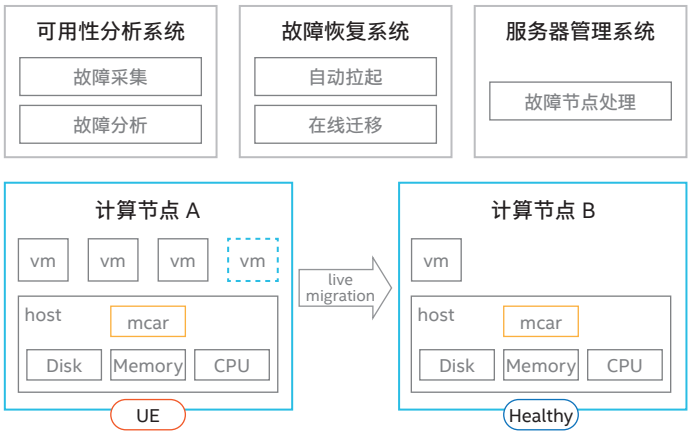


图 5　MCA Recovery 在京东云故障恢复系统中的架构

Tip：部署经验

1. 部署 MCA Recovery 要打开 CPU Data Poisoning 功能。

   CPU Data Poisoning 的作用是，一旦出现内存错误，CPU 会对这种错误进行标签处理，为后续的操作提供一个标识。

   其打开的方式有两种：通过 BIOS 的 System Memory Poison 选项（表2）或在 OS 里设置 MSR 寄存器 0×178 bit0设置为1。

(178h) MSR_MCG_CONTAIN				
Machine Check Containment Mode Register. This register is used to configure Containment Mode capabilities of the machine check architecture of the processor. Before attempting to access this register, the SW must test the Software Error Recovery Support [24] in the IA32 MCG_CAP register (MSR 179h). If the Software Error Recovery Support Bit is not set to 1, a #GP exception may be raised on access to this register.	Core	0h	RW	**POISON_ENABLE** - Enable Poison Mode. When set to 1, Enables Poison. **Mode** - Erroneous data coming from memory will be Poisoned. Errors may be reported in several places. When set to 0 (default), indicates Legacy Mode - No poisoning available.

表 2　BIOS 的 System Memory Poison 选项

2. 操作系统支持

主机内核需要加上英特尔 MCA Recovery 相关 patch。并在内核配置上检查如下的配置：

```
CONFIG _ X86 _ MCE=y
CONFIG _ ACPI _ APEI=y
CONFIG _ ACPI _ APEI _ GHES=y
CONFIG _ ACPI _ APEI _ MEMORY _ FAILURE=y
COFIG _ ARCH _ SUPPORTS _ MEMORY _ FAILURE=y
CONFIG _ MEMORY _ FAILURE=y
CONFIG _ X86 _ MCE _ INTEL=m
CONFIG _ ACPI _ APEI _ EINJ=m
CONFIG _ HWPOISON _ INJECT=m
```

在部署过程中发现，部分机型的 BIOS 设置项目找不到或者隐藏 CPU Data Poisoning，只能通过操作系统对 MSR 进行设置。

性能验证

实际部署前，京东云通过 Ras-Tools 模拟不同类型的内存故障，对部署了 MCA+MFP 的服务器进行了压力测试，测试环境以及机器配置如下：

CPU	Intel® Xeon® Gold 6148
内存	32G DDR4 *12
操作系统	centos 7.4 + Intel patch

表 3　机器配置

在整个测试过程中，使用 Ras-Tools 工具模拟注入 Ue Single、Ue Double、Ue THP、Ue Store、Ue Instr、Ue Patrol、Ue Llc、Ue Mlock、Cmcistorm 等九种类型的故障。整个测试过程中，CE 与 UE 错误都可以被正常巡检出来，并触发恢复流程，故障降级与内存页隔离，保证主机的稳定。

测试主机的宕机频率由部署前 UE 注入 10 ~ 20 次，部署后主机 UE 注入 1500 ~ 6800 次后才产生宕机，稳定性、可靠性大幅提升。

结论

MCA Recovery+MFP 的成功部署，使得京东云数据服务中心可以实时监控各结点云主机的内存使用状况，及时发现出现主机出现的内存故障并加以恢复，使得计算节点主机的宕机率减少 40%，内存故障条件下的热迁移成功率提高了 50%。极大改善了由于内存故障造成主机宕机的稳定性，为保证云主机 99.975% 的可用性提供了强有力的技术支撑。新技术的运用，有效提升了云主机 SLA，提高了终端用户的服务质量，降低了京东云数据中心的总拥有成本；在激烈的云市场竞争中，占据了技术优势。

在未来，京东云仍将与英特尔开展广泛的技术合作，无论是开发和运维的平台级优化，还是云计算趋势性产品的研发，英特尔与京东云的合作必将为中国云计算产业的发展提供助力。