White Paper

Intel MCA+MFP Helps JD Stable and Efficient Cloud Services

"Improving the quality of the cloud host is very important to customers, which can greatly reduce business interruption. In order to achieve the 99.975% SLA commitment of the cloud host, JD and Intel has cooperated deeply, using Intel's MCA and MFP technologies to narrow the fault domain and perform fault isolation or online migration in advance, reducing the impact of memory CE and UE failures on the cloud host."

—— Gong Yicheng

Head of IaaS Product R&D of JD Cloud Business Group of JD Technology

"System reliability and stability have always been the most basic service tenet of the services of JD Cloud, and JD Cloud has never stopped exploring basic technological breakthroughs. This time, JD Cloud and Intel have customized and optimized the Advanced RAS function of CPU, and jointly researched and developed the fault prediction model of Intel MFP in combination with usage scenarios to realize memory failure prediction and repair. This technology will effectively improve the reliability and stability of JD Cloud's services. We also expect to continuously improve the reliability and stability of the hardware system through the efforts of both parties. "

—— Chen Guofeng

Head of Hardware R&D of JD Cloud Business Group of JD Technology

Contents

Foreword

"With the successful deployment of Intel's MCA Recovery and MFP technologies, the down probability of JD Cloud's host due to memory failures has been reduced by 40%, and the success rate of hot migration under memory failures has been increased by 50%, which has greatly enhanced the reliability and stability of the host, improved the user experience, and promoted the SLA, so that JD Cloud has a technological advantage in the fierce market competition."

Today, with the rapid development of digital economy, the following life scenarios have gradually become familiar to people:

"A freight driver no longer seeks customers offline, but installs an application on his mobile phone, and receives orders through the freight platform, which can reduce the unloaded ratio of his truck and ensure the payment of transportation expenses, which not only reduces costs but also increases income." Behind the freight platform is the exchange of information between hundreds of thousands of freight drivers and shippers every day, the transportation value of over 100 million, the track coordinates of millions of trucks, hundreds of millions of driving data, and PB-level data volume.

"In the post-COVID-19 epidemic era, online learning has become a normal state. A primary school student takes online classes online, communicates with teachers without delay, and uses various teaching aids to learn knowledge." The online education system providers need to face millions of daily active users every day, provide more than 100 kinds of teaching functions and teaching aids, covering six continents around the world, and provide 1080P high-definition live broadcast, and the delay needs to be less than 200ms.

"On June 18 of the year, JD's merchants are making great efforts to actively prepare for sales promotion. A merchant is updating products, making advertisements and setting discounts through the JD platform. A merchant can also customize the intelligent customer service to deal with massive consultations, and contact an anchor to warm up for the event to only wait for the promotion to sell. " However, during the promotion period, the JD platform will also face the tests of hundreds of millions of attacks, hundreds of billions of orders and trillions of traffic.

"Internet +" has changed people's life style and subverted the business models of traditional industries, but all these need to be guaranteed by a safe, stable and efficient basic technical architecture, and JD Cloud has provided perfect solutions for these scenarios.

Background

JD Cloud is a leading cloud computing brand of JD Technology Group. Relying on the cutting-edge scientific and technological capabilities of JD Technology Group in the fields of artificial intelligence, big data, cloud computing and Internet of Things, JD Cloud provides multi-cloud, safe and reliable basic cloud services including public cloud, proprietary cloud and hybrid cloud, and provides leading cloud computing services and industry solutions for customers in the fields of the Internet, finance, city, transportation, energy and other fields around the world. In April 2016, JD Cloud was officially commercialized and entered the Chinese cloud computing market. In June 2017, all of JD's businesses were on the cloud. In April 2021, the market share of the IaaS of JD Cloud rose to the fifth place in China, ranking among the first echelon of cloud computing in China.

As one of the most thoroughly containerized cloud platforms in the world, JD Cloud has the world's largest Docker cluster and the world's largest Kubernetes cluster, supporting trillion-level e-commerce transactions, realizing 100% completion of orders for the JD 618 Shopping Festival on the cloud, and JD Logistics and JD Health all going to the cloud. JD Cloud has stood the test of trillions of traffic peaks at JD 618, JD 11.11, Spring Festival Evening, etc. JD Cloud serves many customers in the fields of video, media, online education, games, etc. The highest availability of services is guaranteed to reach 99.995%.

Pain Point:
The surge in business puts a higher requirement for the stability and reliability of JD Cloud

Nowadays, JD Cloud covers more than 2,500 partners in various industries and fields. With the increasing user scale, users in specific industries and cloud native users put forward many new requirements for application development and operation mode, and traditional users are migrating more complex businesses to the cloud. These constantly changing technical requirements pose new challenges to the services of JD Cloud.

The cloud host is the core resource of the cloud service, and its reliability, availability and maintainability directly determine the quality and level of the cloud service. Nowadays, the occurrence of hardware failures is an important factor that causes host downtime. In the traditional way, that a group of services stop working will only affect their own businesses and users. However, in the cloud environment, service termination will cause a cloud service provider to violate the Service Level Agreement (SLA) and cause huge economic losses. Among many hardware failures, the memory error is one of the most serious failures faced by the data centers today. At present, the memory errors in the data center of JD Cloud account for 37% of the total hardware failures. Therefore, JD Cloud has established a perfect cloud host failure prediction and recovery system, hoping to reduce the impact of memory errors on the cloud host through the discovery and prediction of memory errors and the online rapid migration and recovery technology.

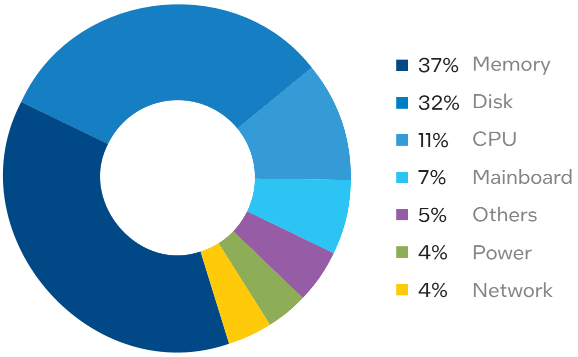


Figure 1 Distribution of Hardware Failures of JD Cloud

However, due to the impact of the compatibility of the software and hardware systems in the cloud host at present, the recovery system is still unable to quickly recover the downtime caused by many memory failures. For example, the recovery system cannot perform hot migration of the storage-optimized cloud host. The recovery system discovers memory errors in time during daily inspections, and problems such as system downtime occur in the process of hot migration, which increases the failure rate of the cloud host.

If a set of solutions that can gain real-time insight into the memory status of the cloud host, predict potential memory failures and effectively recover memory failures can be established, it will greatly help to improve the stability and reliability of the services of JD Cloud, improve the SLA of end users, and reduce the total cost of ownership of the data center of JD Cloud.

Intel MCA Recovery + MFP helps JD Cloud provide efficient and stable services

JD Cloud and Intel have maintained close and extensive cooperation in the field of cloud computing, and providing professional and cost-effective cloud services for end users is the original intention of their cooperation. In order to solve the problem of memory errors, JD Cloud and Intel cooperate again. JD Cloud reduced the impact of memory errors on the stability of JD Cloud's host by introducing Intel's MCA Recovery and Memory Failure Prediction (MFP) technologies and combining with JD Cloud's failure recovery system.

Memory Errors

At present, the memory errors of the cloud host mainly include the Corrected Error (CE) and the Uncorrected Error (UE). At present, the most common solution to the Corrected Errors is to overcome some Corrected Errors of dual-inline-memory-modules (DIMM) through error correction code (ECC).

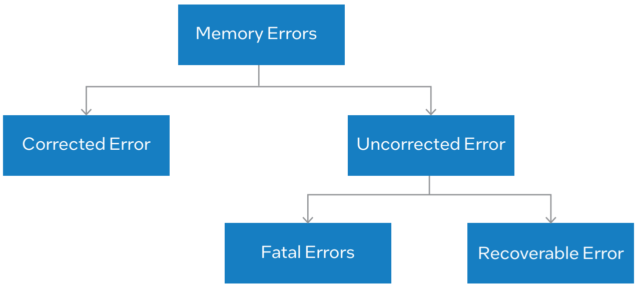


Figure 2 Memory Error Classification 1

The Uncorrected Errors (UE) usually cause serious disastrous consequences, such as host operating system hanging, system crash, and downtime. UE errors include Fatal Error, SRAR, SRAO and UCNA.

1. Fatal Error: Very serious UE errors. The system cannot repair this type of errors. Such errors will cause the processor to be in a chaotic or unstable state, and can only be recovered by resetting the system. At present, there is no good recovery method for this type of UE errors.

2. RAR (Software Recoverable Action Required): After a RAR occurs, the operating system/application needs to perform some operation (such as isolating/terminating the failed thread) to recover such UE. This type of errors is the type of errors that the recovery technology can mainly recover.

3. SRAO (Software Recoverable Action Optional): After a SRAO occurs, the operating system/application program chooses to perform some operation (such as isolating/terminating the failed thread) according to the policy set by the user to recover such error.

4. UCNA (Uncorrectable Error No Action required): The error is not located on the critical path, and does not trigger MCE, and usually no action is required.

Based on the analysis and understanding of memory failures, it can be judged that a set of technical solutions for prediction and recovery of SRAR and SRAO UE errors can effectively reduce the impact of memory failures on the cloud host. After repeated tests and trade-offs by technical experts of JD Cloud and Intel, JD Cloud finally chose Intel's MCA Recovery and MFP technologies to solve such problem.

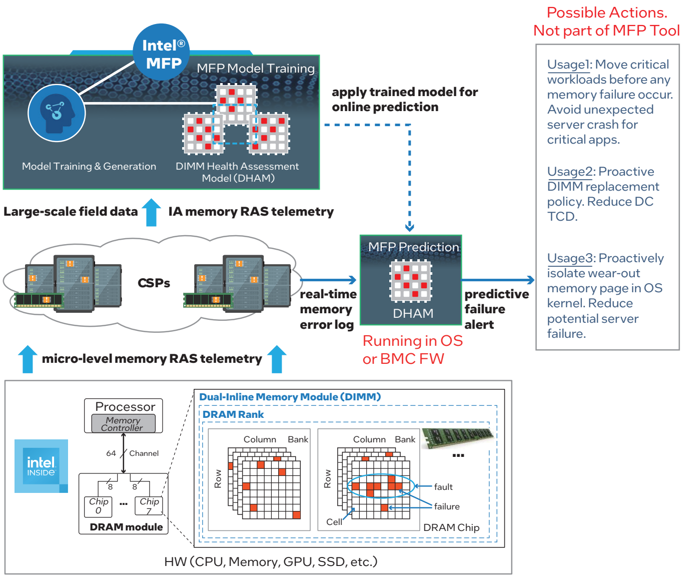
	Mci_STATUS							MCG_STATUS						
								Errored Thread		Other Threads				
	Val ID	UC	PCC	Service	AR (Action Required)	ADDRV	MISCV	RIPV	EIPV	RIPV	EIPV	Signaling	ADDR in Kernel Space	SW Action
Uncorrected Errors	1	1	1	×	×	×	×	×	×	×	×	MCERR	×	System Crash
SRAR - Instruction	1	1	0	1	1	1	1	0	0	1	0	MCERR	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

Table 1 Memory Error Classification 2

MFP

Intel® MFP[1] is a data-driven technology that improves the reliability of hosts through active memory failure management. By studying the historical failure data, such technology can predict the memory failures of a host autonomously, and notify the system administrator before a catastrophic result occurs.

Intel® MFP learns and mines the micro-level fault data of the memory through thousands of EDAC logs to train and build a DIMM health assessment model (DHAM). After the MFP is deployed, it will monitor the running status of the host memory in real time, analyze the memory errors at different levels of the host, including DIMM, rank, bank, column, row and cell, and compare the host memory status with the DIMM health assessment model to predict the possibility of memory failures.



Usage1: Move critical workloads before any memory failure occur. Avoid unexpected server crash for critical apps.

Usage2: Proactive DIMM replacement policy. Reduce DC TCD.

Usage3: Proactively isolate wear-out memory page in OS kernel. Reduce potential server failure.

MCA Recovery

MCA Recovery[2] is an "Intel advanced RAS" function, and a technology which uses the MCA architecture of CPU and firmware (such as UEFI firmware) to isolate the found uncorrected hardware errors (UE), so that the system can recover from such errors.

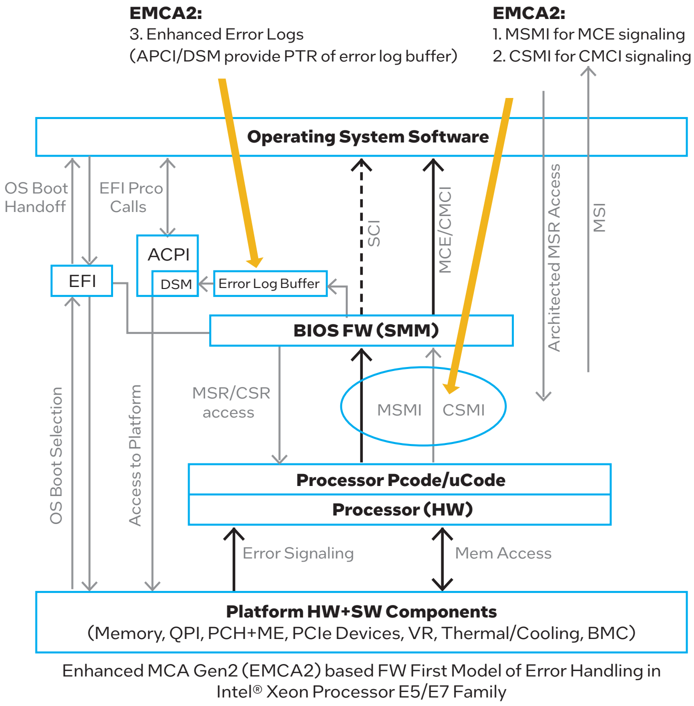

EMCA2:
3. Enhanced Error Logs
(APCI/DSM provide PTR of error log buffer)
EMCA2:
1. MSMI for MCE signaling
2. CSMI for CMCI signaling
Operating System Software
OS Boot Handoff
EFI Prco Calls
SCI
MCE/CMCI
Architected MSR Access
MSI
ACPI
EFI
DSM
Error Log Buffer
BIOS FW (SMM)
OS Boot Selection
Access to Platform
MSR/CSR access
MSMI
CSMI
Processor Pcode/uCode
Processor (HW)
Error Signaling
Mem Access
Platform HW+SW Components
(Memory, QPI, PCH+ME, PCIe Devices, VR, Thermal/Cooling, BMC)
Enhanced MCA Gen2 (EMCA2) based FW First Model of Error Handling in Intel® Xeon Processor E5/E7 Family

Figure 3 Technical Schematic Diagram of MCA Recovery

[1] Intel® Memory Failure Prediction: https://www.intel.com/content/www/us/en/software/intel-memory-failure-prediction.html

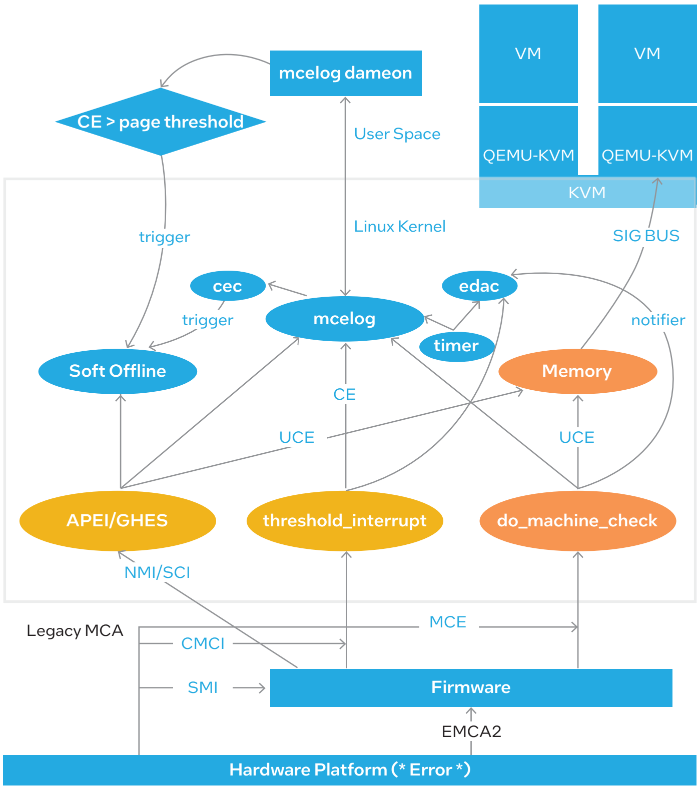
[2] Technical Introduction of Intel MCA Recovery: https://partneruniversity-prc.intel.cn/diweb/catalog/launch/package/4/eid/777016

For recovery using the MCA Recovery technology, the following preconditions shall be met:

1. The memory UE is a non-fatal error.

2. The memory fault address is not in the kernel space.

3. The application or process accessing the wrong address can be killed.



UCE: Uncorrected Error
CE: Corrected Error

Figure 4 MCA Memory Error Recovery Flow Chart

As shown in Figure 4, when a memory error occurs, the CPU will notify the BIOS through MSMI, and enter the SMM Mode (error handling mode). After receiving the instruction, BIOS will collect and sort out the wrong data and store the same in API TABLE or BMC, and send an instruction to OS informing the error and the error category. After receiving the instruction, OS will isolate the memory address where the error occurred and delete the application that still uses the wrong address data, so as to ensure that the host will not be down and the error will be recovered.

The deployment of MCA Recovery and MFP, in conjunction with JD Cloud's fault recovery system, greatly reduces the system crash caused by the memory failures of JD Cloud's host. When a node has the possibility of potential UE, the availability analysis system of JD Cloud can give real-time UE error warnings through Intel ® MFP technology, thus triggering the first layer of system protection-hot migration based on memory failure prediction, and avoiding the downtime of the cloud host caused by potential internal failures. If a UE failure occurs outside the range of MFP prediction, the second layer system protection will be triggered-recovery through MCA Recovery. With the help of MCA Recovery, the failure recovery system will isolate the affected memory pages and prevent the pages from being reused by other applications/processes. If the kernel can successfully perform recovery, the system can stay online as long as there is no failure. At the same time, the data center maintenance will collect fault logs to determine which DIMM fails, migrate the cloud host where the memory page of UE appears to a healthy node online, and report the failed node to the server management system for memory replacement.

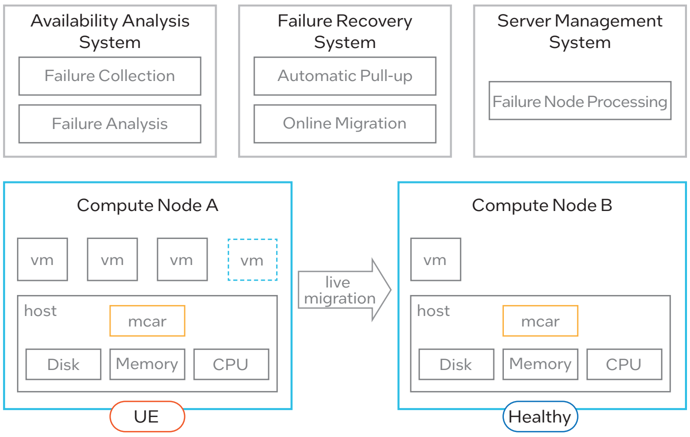


Figure 5 Architecture of MCA recovery in the Failure Recovery System of JD Cloud

Tip: Deployment Experience

1. To deploy MCA Recovery, the CPU Data Poisoning function shall be enabled.

The function of CPU Data Poisoning is that once a memory error occurs, the CPU will tag such error to provide an identification for subsequent operations.

There are two ways to open the CPU Data Poisoning: through the System Memory Poison Option of BIOS (Table 2), or by setting MSR register 0×178 bit0 to 1 in OS.

(178h) MSR_MCG_CONTAIN				
Machine Check Containment Mode Register. This register is usedto configure Containment Mode capabilities of the machine checkarchitecture 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.

Table 2 System Memory Poison Option of BIOS

2. Operating system support

The host kernel needs to be added with the Intel MCA R core configuration to check the following configuration:

```
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
```

During the deployment, it was found that the BIOS setting items of some models could not find or hid CPU Data Poisoning, and MSR could only be set through the operating system.

Performance Verification

Before actual deployment, JD Cloud will simulate different types of memory failures through Ras-Tools, and conduct stress test on the servers deployed with MCA+MFP. The test environment and machine configuration are as follows:

CPU	Intel® Xeon® Gold 6148
MEMORY	32G DDR4 *12
Operating System	centos 7.4 + Intel patch

Table 3 Machine Configuration

In the whole test process, Ras-Tools are used to simulate and inject nine types of failures, such as Ue Single, Ue Double, Ue THP, Ue Store, Ue Instr, Ue Patrol, Ue Llc, Ue Mlock, Cmcistorm, etc. During the whole test process, CE and UE can be inspected normally, and the recovery process can be triggered. Failure degradation and memory page isolation can ensure the stability of the host.

The downtime frequency of the tested host was injected 10 ~ 20 times by the UE before deployment, and 1500-6800 times by UE after deployment of the host UE before the downtime occurred, which greatly improves the stability and reliability of the host.

Conclusion

With the successful deployment of MCA Recovery + MFP, the data service center of JD Cloud can monitor the memory use of the cloud host of each node in real time, discover and recover the memory failures of the hosts in time, which reduces the downtime rate of computing node hosts by 40% and increases the success rate of hot migration under the memory failure condition by 50%, greatly improves the stability of host downtime caused by memory failure, and provides strong technical support for ensuring 99.975% availability of the cloud hosts. The application of new technologies has effectively promoted the SLA of cloud hosts, improved the service quality of end users, and reduced the total cost of ownership of the data center of JD Cloud. In the fierce competition in the cloud market, JD Cloud will occupy the technical advantage.

In the future, JD Cloud will continue to conduct extensive technical cooperation with Intel. The cooperation between Intel and JD Cloud will certainly provide assistance for the development of China's cloud computing industry, whether for platform-level optimization of development and operation & maintenance, or research and development of cloud computing trend products.

No component or product can be absolutely secure.

Specific costs and results may vary.

Intel does not control or audit third-party data. You should review this content, consult other sources, and confirm whether referenced data is accurate.

Actual performance is affected by differences in use, configuration and other factors. For more information, see www.Intel.com/PerformanceIndex. See the attached page for workload/configuration information.

Intel technologies may require enabled hardware, software or service activation.