

白皮书

第五代英特尔® 至强® 可扩展处理器

英特尔® Memory Resilience Technology (英特尔® MRT)

intel

基于第五代英特尔® 至强® 可扩展处理器 BMC 人工智能辅助故障分析，助力阿里云 提升 DDR5 内存可靠性

背景

在阿里云数据中心，内存故障是服务器稳定运行面临的主要挑战之一。大规模数据中心中的内存故障，不仅会降低服务器的可靠性，还可能中断数据中心的服务并影响服务器的性能。因此，内存可靠性成为数据中心中服务器可靠性、可用性和可维护性 (Reliability, Availability, Serviceability – RAS) 的关键要素。

新一代内存标准 DDR5 具有更高的带宽、更低的功耗和更高的密度。然而，它也为内存可靠性带来了新的挑战，其中包括：

- DDR5 引入了新的架构和信号传输方式，需要更复杂的电路设计和优化；
- DDR5 内存模块容量更大，但也增加了故障的风险；
- In-DRAM 纠错码 (ECC) 虽然可以纠正内存中单比特的错误，但它也导致主机错误观察不够明确。

为了应对这些挑战，阿里云与英特尔合作改进了 DDR5 内存的可靠性。具体措施包括：

1. **主板管理控制器 (BMC) 的统一带外 (OOB) 内存错误数据收集：**通过 BMC 实现内存错误数据的统一收集，为后续分析提供数据基础。
2. **内置人工智能辅助 (AI 辅助) 的故障分析：**BMC 中集成 AI 辅助，实时预测和分析内存故障。
3. **英特尔® Memory Resilience Technology (英特尔® MRT)：**英特尔® MRT 已在阿里云数据中心部署，用于提前预警和预防潜在的内存故障。
4. **与阿里云巡洋舰系统 (Alibaba Cruiser System) 集成：**将内存健康评估和预测警报与阿里云的服务器监控系统集成，以确保业务的稳定性。

这些举措共同为阿里云数据中心提供了快速且全面的硬件监控服务，帮助确保了服务器的可靠性和业务的正常运行。

内存可靠性面临的挑战

内存故障可能由多种不同类型内存底层错误产生，例如单比特错误 (SBE)、行类型错误、列类型错误、多阵列错误、存储器模块 (DIMM) 错误等。每种内存错误都有其特定的频率和受影响模式。例如，某些错误类型会零星出现或间歇性发生，难以有效追踪，而有些错误类型则可能持续报错。有些错误类型存在更高的不可纠正错误 (Uncorrectable Errors – UE) 风险，需要立即采取 RAS (可靠性、可用性和可维护性) 措施，而其他一些错误类型触发 UE 的风险相对较低，但在短时间内可能导致大量可纠正错误 (Correctable Errors – CE)，从而影响系统性能。没有一种通用的解决方案可以解决所有内存错误。

基于 BMC 的人工智能辅助故障分析助力提升 DDR5 内存的可靠性

阿里云和英特尔联合研究和开发了面向 DDR5 的内存故障预测和预防解决方案。该方案通过 BMC 实现内存错误数据的统一收集，为后续分析提供数据基础。在 BMC 中集成英特尔® MRT 技术提供 AI 辅助的实时预测和分析内存故障，用于提前预警和预防潜在的内存故障。数据收集、故障分析和预警与阿里云的服务器监控系统集成 (阿里云巡洋舰系统)，为阿里云的数据中心提供快速而全面的硬件监控服务，以确保业务的稳定性。

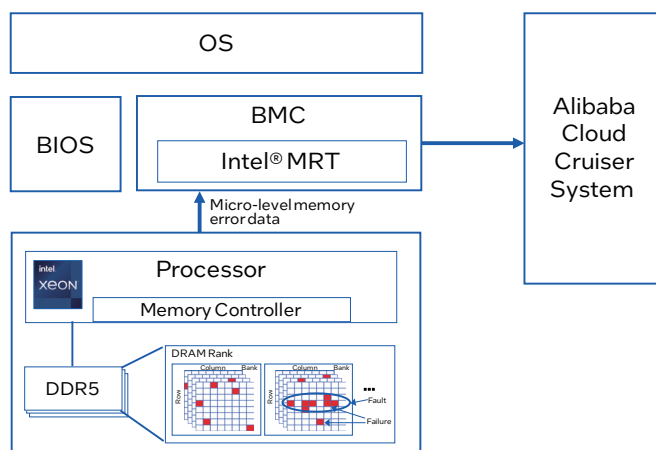


图 1. 解决方案架构图

这一解决方案的关键特点包括：

■ 基于 BMC 的细粒度内存故障采集

通过 BMC 收集细粒度的可纠正错误 (CE) 和不可纠正错误 (UE) 信息，包括详细的位级错误数据。相比使用带内 (in-band) 方式收集内存错误数据，例如错误检测和纠正 (EDAC) 驱动程序或基于 BIOS SMI 中断触发，基于 BMC 的带外内存收集更可靠且统一，具有细粒度的数据粒度和丰富的错误信息。

传统的解决方案之一是在观察到不可纠正错误 (UE) 后更换故障的 DIMM。然而，此举无法避免系统崩溃的成本。另一种方法是基于计数的可纠正错误 (CE) 评级策略来预测内存故障这种策略在预测复杂内存故障方面效果较差，因为 CE 和 UE 的发生不仅取决于硬件的内存故障状态，还取决于隐性的运行时上下文、ECC 纠正能力和内存特定的故障模式。因此，内存错误具有高度的不确定性，预测 UE 非常困难。

虽然没有通用的解决方案，但我们可以探索更智能的方法来处理内存故障。例如，结合机器学习和实时监测，以更精确地预测 UE 和 CE 的发生。内存错误是一个复杂且关键的问题，需要综合考虑多种因素来优化系统的可靠性和性能。

■ 基于微观内存故障类型的错误分析

通过历史可纠正错误 (CE) 信息的详细数据，检测底层内存故障类型。与仅关注 CE 计数不同，该解决方案从多个因素检查内存错误数据，包括空间分布 (例如 channel、rank、sub-channel、bank、row、column 等)、时间模式 (例如瞬态、间歇、永久)、错误位 (error bit) 位置、内存特定故障模式、CPU 错误纠错码 (ECC) 设计以及系统 RAS 配置等综合评估故障风险。

■ AI 辅助故障分析

利用机器学习方法训练了一个 AI 模型，通过对海量 DDR5 内存日志进行比较，预测内存故障。预先训练的内存故障预测 AI 模型集成到主板管理控制器 (BMC) 中，通过 BMC 为服务器提供内存故障的实时预测与分析，从而减少大规模数据中心中的服务器停机时间。

■ 集成阿里云巡洋舰硬件故障检测系统

实时内存健康评估和预测警报已与阿里云巡洋舰系统集成，为阿里云数据中心的物理服务器提供快速而全面的硬件监控服务。

英特尔® Memory Resilience Technology

英特尔® Memory Resilience Technology (英特尔® MRT) 是一项旨在提高数据中心内存可靠性的技术，它使数据中心运营商能够主动预测潜在的内存故障风险，确保数据中心的运行和工作负载的连续性。以下是该技术的关键功能：

1. 基于带外的细粒度内存故障数据收集：实现细粒度内存错误数据的统一收集，为后续分析提供数据基础
2. 分析定位内存故障点：提供底层内存故障定位及分析

3. **预测性故障警报：**提前发现可能出现的内存故障。
4. **基于预测的内存页面离线：**根据预测，将内存页面离线，以防止潜在故障影响。
5. **基于预测的内存故障区域隔离：**根据预测及系统相应 RAS 配置，隔离内存故障区域，以避免潜在内存错误发生。

英特尔® Memory Resilience Technology 利用多维模型和人工智能算法，在微观层面检测内存故障。它为每个 DIMM 分配健康分数，并实时检测潜在的故障。通过人工智能分析海量的内存错误日志优化内存故障预测模型，该技术可以准确地定位潜在问题，并在故障发生之前识别和防止内存故障。

虽然没有通用的解决方案可以解决所有内存错误，但英特尔® Memory Resilience Technology 为数据中心提供了一种智能且综合的方法，以优化系统的可靠性和性能。

利用 BDAT 数据诊断硬件故障

英特尔 BIOS 参考代码实现了系统验证功能，可以生成包括内存

余量数据在内的全面系统数据。这些数据从标准的 BIOS 数据 ACPI 表 (BDAT) 中暴露出来，该表在 ACPI 表中定义。BDAT 数据是系统 BIOS 的基本支持，它在整个 BIOS 引导流程中生成，并集成到 ACPI RSDT 表中。通过分析 BDAT 数据，可以有效提升生产系统的诊断和问题调试的效率。

结果与分析

阿里云已在不同工作负载下的阿里云数据中心的数千台采用第四代英特尔® 至强® 可扩展处理器的平台上部署了英特尔® Memory Resilience Technology，并正在将平台升级至第五代英特尔® 至强® 可扩展处理器。

新一代处理器拥有更可靠的性能，更出色的能效。它在运行各种工作负载时均可实现显著的每瓦性能增益，在 AI、数据中心、网络和科学计算的性能和总体拥有成本 (TCO) 方面亦有更出色的表现。相较上一代产品，第五代英特尔® 至强® 可扩展处理器可在相同功耗范围内提供更高的算力和更快的内存。此外，它与上一代产品的软件和平台兼容，因此部署新系统时可大大减少测试和验证工作。

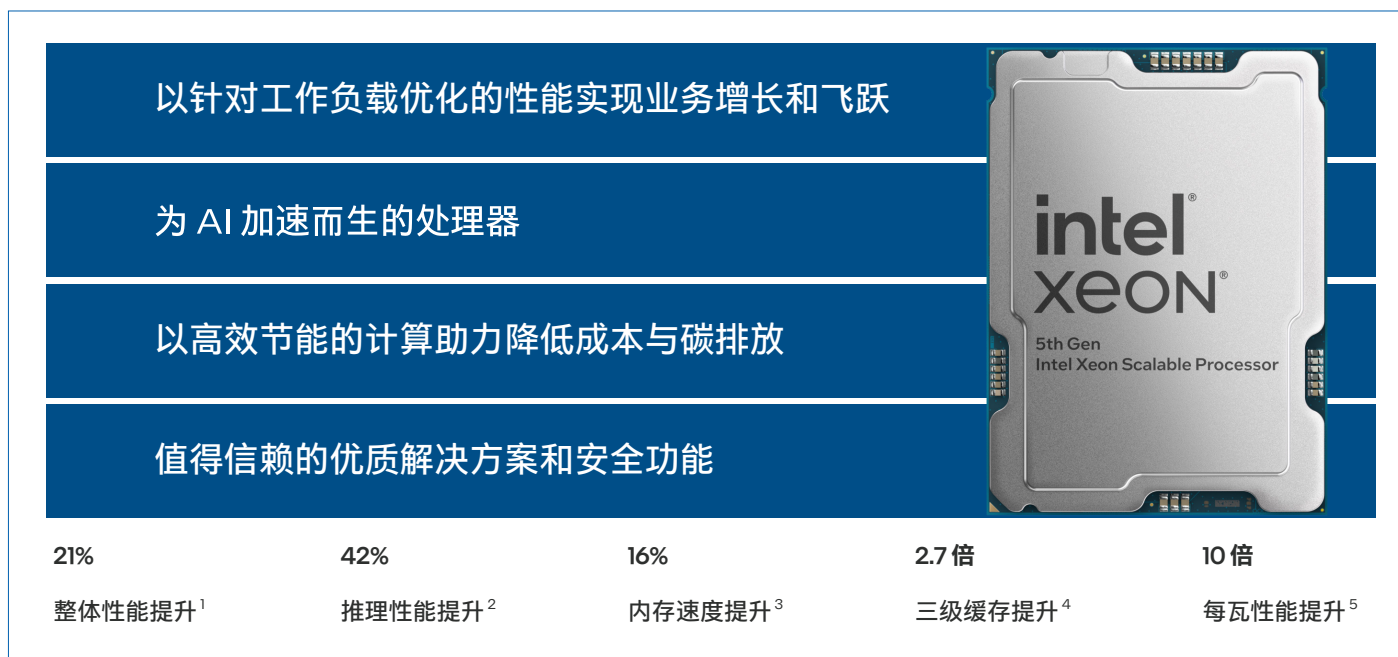


图 2. 第五代英特尔® 至强® 可扩展处理器具备更强大性能

¹ 与第四代英特尔® 至强® 处理器相比的平均性能提升，以 SPEC CPU rate、STREAM Triad 和 LINPACK 的几何平均值为衡量标准。请参阅 intel.com/processorclaims 上的 [G1]：第五代英特尔® 至强® 可扩展处理器。结果可能有所差异。

² 与第四代英特尔® 至强® 处理器相比，取得 1.19 倍到 1.42 倍的性能提升 (ResNet50v1.5、BERT-Large、SSD-ResNet34、RNN-T (仅 BF16)、Resnext101.32x16d、MaskRCNN (仅 BF16)、DistilBERT)。请参阅 intel.com/processorclaims 上的 [A15-A16]：第五代英特尔® 至强® 可扩展处理器。结果可能有所差异。

³ 请参阅 intel.com/processorclaims 上的 [G12]：第五代英特尔® 至强® 可扩展处理器。结果可能有所差异。

⁴ 请参阅 intel.com/processorclaims 上的 [G11]：第五代英特尔® 至强® 可扩展处理器。结果可能有所差异。

⁵ 使用内置加速器在 AI、数据和网络工作负载上进行测量，取得 1.46 到 10.6 倍的每瓦性能提升。请参阅 intel.com/processorclaims 上的 [A19-A25]、[D1]、[D2]、[D5] 和 [N16]：第五代英特尔® 至强® 可扩展处理器。结果可能有所差异。

初步结果表明，该解决方案可以在不可纠正错误 (UE) 发生之前有效地预测，并在传统的基于 CE 计数的 CE 风暴识别机制被触发之前警报可纠正错误 (CE) 风暴案例。UE 和 CE 风暴警报的预测提前时间因底层故障模型而异，从几分钟到几小时甚至几

天不等。该方案经过迭代，预期能够通过优化的 DDR5 模型预测 57% 的 UE 和 74% 的 CE 风暴⁶。

除了有效的 UE 和 CE 风暴预测外，从 BMC 收集的带外 (OOB) 内存错误对于进一步诊断和排除内存和系统问题至关重要。

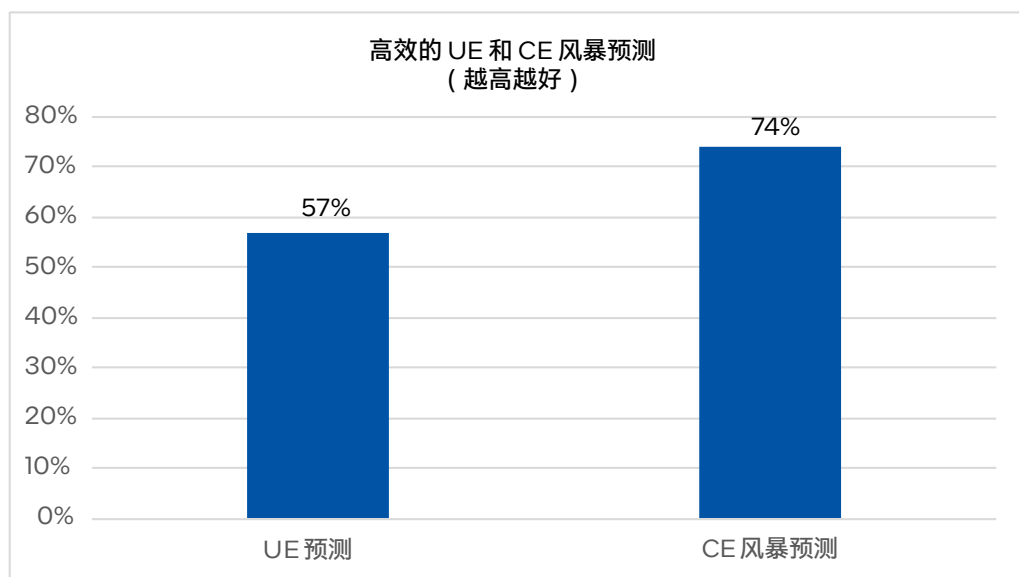


图 3. 高效的 UE 和 CE 风暴预测

结论

通过 BMC 集成英特尔® Memory Resilience Technology 技术，可以有效提高阿里云数据中心 DDR5 内存可靠性。对于阿里云而言，改善整体数据中心的总体拥有成本 (TCO) 至关重要。英特尔和阿里云正在合作开发下一代的 DDR5 故障预测技术和提供对新内存技术的方法。

⁶ 数据援引自截止 2023 年 11 月的内部测试结果。测试配置：英特尔® 至强® 8475B 处理器和英特尔® 至强® 8575C 处理器，32*64 GB，16*64 GB，16*32 GB，2*240 GB 固态硬盘。英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。

作者 简介

英特尔：

周申，英特尔数据中心与人工智能事业部，首席工程师

内存稳定性，系统 RAS，软硬件协同设计资深技术专家。英特尔® MRT 技术负责人。开放计算平台 (OCP) 内存故障管理工作组主席。

曾志军，英特尔数据中心与人工智能事业部，云软件工程师

BIOS 和 BMC 技术专家，负责阿里云服务器固件开发技术支持。

阿里云：

候海（周大海），基础设施事业部/固件开发

BIOS 技术专家，主要负责云服务器的管理、诊断等特性设计研发工作

边荒（李志兵），基础设施事业部/固件开发

BMC 开发经理，主要负责云服务器的 BMC 开发和管理工作

玄叡（谢可可），基础设施事业部/存储部件开发

内存技术专家，主要负责云服务器的内存部件开发和管理工作



实际性能受使用情况、配置和其他因素的差异影响。更多信息请见 www.intel.com/PerformanceIndex。

性能测试结果基于配置信息中显示的日期进行测试，且可能并未反映所有公开可用的安全更新。详情请参阅配置信息披露。没有任何产品或组件是绝对安全的。

具体成本和结果可能不同。

英特尔技术可能需要启用硬件、软件或激活服务。

英特尔未做出任何明示和默示的保证，包括但不限于，关于适销性、适合特定目的及不侵权的默示保证，以及在履约过程、交易过程或贸易惯例中引起的任何保证。

英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。

© 英特尔公司版权所有。英特尔、英特尔标识以及其他英特尔商标是英特尔公司或其子公司在美国和/或其他国家的商标。其他的名称和品牌可能是其他所有者的资产。