

服务器内存故障预测与自愈技术 白皮书



编制说明

本报告由超聚变数字技术有限公司和招商银行联合撰写，在此特别感谢以下参编人员（排名不分先后）：

招商银行：钟祝君、王大晓、李志成

超聚变：梁永贵、于超琪、毛阿利、甘延、曹瑞、李泽坤、黄敏、林培淳



引言

近年来，金融行业纷纷向全面数字化转型。在中国金融云市场，银行业务占了近70%份额^[1]。基于业务和数据的规模效应，银行业在云计算等数字技术应用上也处于领先地位。2022年底，招商银行完成全面上云，完成以“云平台、大数据、AI”为核心的数字化建设新基建^[2]，这使得招行成为中国系统重要性银行前七家中率先实现全面上云的银行。

云计算系统涉及到成千上万个云服务器相互连接构成复杂系统，减少内存故障引发的服务器崩溃宕机是提高招商银行云服务连续性、可靠性的关键。招商银行和超聚变联合研发的内存故障预测与自愈技术，可以实现内存健康状况的实时监控及预警、故障内存的自愈，降低服务器宕机率。在招商银行数据中心1080台服务器试点中，该技术将内存引发的宕机率降低了66%，增强了招商银行数据中心运行的安全性和稳定性，提升了招商银行运维平台的智能化水平。



目录

01 术语表

02 背景与挑战

- 2.1 背景
- 2.2 挑战

03 内存工作原理及故障机理

- 3.1 内存工作原理
- 3.2 内存故障及分类

04 内存故障预测与自愈技术

- 4.1 AI预测算法模型使用指标
- 4.2 AI预测算法模型数据结构
- 4.3 AI预测算法模型研究
- 4.4 AI预测模型使用的算法

05 AI预测模型使用实例：大大降低内存故障导致宕机率

06 结论与展望

07 参考资料

术语表

简写	英文全称	中文翻译
OS	Operation Systems	操作系统
CPU	Core Process Unit	中央处理单元
DIMM	Dual Inline Memory Moduals	双列直插式存储模块
VRT	Variable Retention Time	可变保持时间
ECC	Error Correction Code	纠错码
CE	Corrected Error	可纠正错误
UCE	Uncorrected Error	不可纠正错误
FPC	Failure Prediction and Correction	故障预测和自愈
AI	Artificial Intelligence	人工智能
MCA	Machine Check Architecture	机器检查框架
PCLS	Machine Check Architecture	机器检查框架
Page Offline	Page Offline	页离线隔离
PPR	Post Package Repair	封装后修复
ADDDC	Adaptive Double Device Data Correction	自适应双颗粒数据纠正
BMC	Baseboard Management Controller	基板管理控制器

数据中心是为电子信息设备提供集中运行环境的建筑场所。随着移动互联、云计算、大数据、人工智能等现代化技术的飞速发展，数据中心的业务规模和复杂度急剧攀升，数据中心算力规模已成为国民经济发展的主要引擎，是衡量经济发展新的核心指标之一。2020年，发改委首次将数据中心、智能计算中心为代表的算力基础设施列入国家重点投入的新基建项目^[3]。在数据中心投资中，服务器作为数字经济、整体IT系统的基础设施，其投资占比超过50%，已成为支撑底层算力连续性的核心底座。

随着云计算开始向分布式云演进，金融行业的数据中心规模越来越大，金融云数据中心的数字化管理势在必行。建设适应大规模云化的监控应急体系，包括适应分布式应用架构的链路跟踪系统、适应大规模云化基础设施的全栈式的基础指标监控，以及故障应急及告警自愈的系统，以充分发挥自动化的作用，提高运维质量，节约运维人力，提升数字化运营能力是金融数据中心的重点发展方向。

作为规模最大的股份制银行，招商银行资产规模已经达到10万亿。作为分布式云计算建设的先行者，多年来持续推动自身数字化转型，云平台规模行业领先，日均服务调用量超过150亿次。2022年底招商银行完成全面上云工程，促进其业务创新和资源灵活供给，迈入全面数字化新阶段。随着客户规模、业务类型的增加，以及对增强用户体验的追求，其服务器数量在近几年呈现指数式的增长，对云服务在连续性、稳定性等方面提出了更高的要求。

服务器是数据中心的的核心部件，其稳定性与可靠性直接关乎数据中心的数据安全和服务质量。随着服务器性能的不断攀升，内存容量显著提升、内存工艺技术也越来越复杂，内存故障发生的概率显著增加。内存故障已成为服务器宕机的影响最大的因素。招商银行统计了2022年4月至2023年5月共13个月的硬件故障导致的服务器宕机数据，其中内存故障占比达到69%。

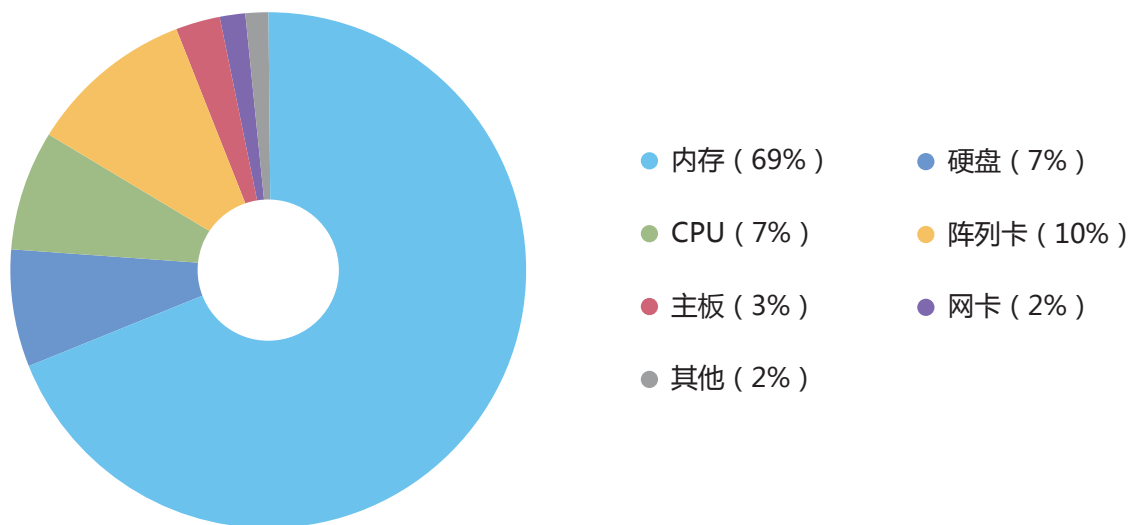


图1 近一年来招商银行数据中心服务器宕机原因分布

由于运维人员数量、运维人员技能、运维工具等运维能力的发展迟滞性，导致人工运维或数字化运维只能通过故障定位和现场更换故障部件进行，影响运维效率和实时性。如在无人值守的时间段，当某一集群中的一台服务器因内存故障宕机后，可能影响整个集群的性能，进而导致某些业务无法按时完成而造成巨大的损失。

招商银行数据中心亟需建立一套内存故障预测与自愈系统。该系统能够实时探测服务器内存健康状态，准确预测潜在的内存故障，并对故障内存做出修复和预警，以提高其服务器运行的连续性、稳定性与可靠性，从而提高服务质量。

2022年7月，招商银行与超聚变数字技术有限公司针对内存故障预测与自愈技术进行联合创新与验证。

内存工作原理及故障机理

3.1 内存工作原理

内存的最小存储结构是由一个晶体管和一个存储电容组成。每一个存储单元存储1比特信息。电容的状态决定着内存基本存储单元的逻辑状态是“0”还是“1”，充满电荷的电容器代表逻辑“1”，“空”的电容器代表逻辑“0”。连接晶体管的字线和位线进行单元选址。其栅极接行(字)选择线，源极接列数据读出放大和写入驱动(位)线，漏极接存储电容。行选（字线）与列选（位线）信号将使存储电容与外界间的传输电路导通，从而可以进行放电（读取）与充电（写入）。多个存储单元组成一个类似表格的存储阵列，为逻辑存储阵列，先指定一个行、再指定一个列，就可以准确地找到所需要的存储单元。单个颗粒由多个存储阵列构成，而将多个颗粒通过信号串在一起，形成存储排。

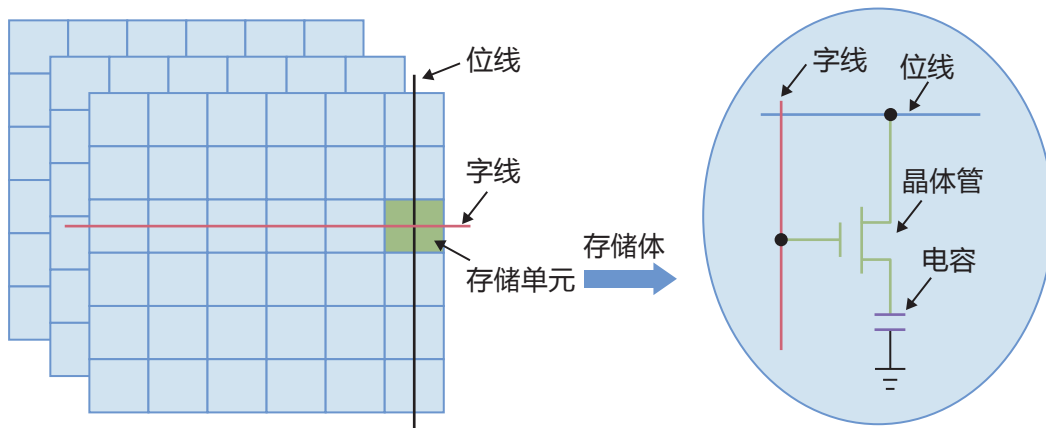


图2 内存工作原理

3.2 内存故障及分类

由于内存内部元器件的偶然失效或耗损失效等原因，内存故障总是不可避免的发生。近年来，伴随摩尔定律，单元内存尺寸在不断微缩以满足日益增长的数据处理要求，导致内存内部元器件增多、频率提升、以及使用电压降低，更是大大增加了内存出现失效的几率。

按照故障源分类，内存故障可分为外围电路故障和内存颗粒故障。

1 外围电路故障一般都是硬失效，占比不足10%，包括内存模组短路或开路导致的外围电路故障、接插件损坏或金手指污染导致的链路故障等，此类故障一般可在服务器开机时及时发现和更换，对后续服务器业务稳定运行影响较小。

2 颗粒故障比较复杂，包括软失效、硬失效、混合失效，占比大于90%，由于颗粒故障大多数在运行过程中发生，所以会影响到业务运行的稳定性。

因此，本文重点是讲解颗粒故障的影响和解决办法。结合内存故障机理分析和大数据统计结果发现，内存故障总是在一定时间内表现出确定的地址集中性而显现聚类特征。这些不同聚类特征亦称为不同故障模式，不同的故障模式往往呈现在不同维度的内存粒度上，包括存储单元故障、行故障、列故障、存储阵列故障、颗粒故障。



图3 内存故障分类

导致内存颗粒介质故障的常见原因如下：

1 电容：内存内的存储体电容两极之间的介电材质不可能是绝对绝缘，即使被周期性的充电，也不可避免的存在漏电现象。如果电容里面的电荷不足，就会导致传输数据错误而引发内存颗粒故障。

2 字线和位线：内存内的存储阵列、地址译码逻辑器由于蠕变、疲劳、老化、磨损等原因，造成字线与位线产生漏电、或字线/位线传输数据路径高阻，从而引发内存颗粒内存故障。

3 晶体管：VRT问题，由于存储体中晶体管PN结漏电的涨落效应，存储单元在不同时刻的数据保持时间存在一定的随机性，从而在内存运行时可能会造成一定程度的数据混淆。VRT问题是内存固有的可靠性问题，进行充放电也不能完全解决，虽然工艺上可以改进，但无法根本解决，制程越先进，出现VRT问题的概率就越大。

内存故障预测与自愈技术

目前业界对内存质量保障手段主要是控制来料质量，或依赖CPU的ECC纠错能力来规避内存运行中的轻微失效。然而，当内存发生严重故障时，内存CE累计过多，大量的ECC纠错会导致纠错能力下降，难免出现一些不可纠错的情况，会导致服务器宕机或上层业务异常。此外，内存错误经常以行、列或者行列复合的模式分布，特定模式的内存故障发展到一定程度，大概率会在后续时间导致UCE，从而导致业务受损或者宕机。因此，尽早地发现CE过多的内存条，找出ECC难以处理的特征CE及其他内存关键指标，识别出这些隐患内存的风险度，及时进行分级处理，可降低宕机率，减少其造成的重大损失。

招商银行和超聚变成立了联合研究工作组共同研究服务器内存预测和自愈方案。工作组以招商银行业务特异性场景的数据和超聚变FPC技术为基础，进行基于内存故障的AI预测和智能化运维技术的探究，包括数据采集、数据分析、AI预测算法模型研究、智能自愈研究等，探索出了一套创新的内存故障预测与自愈技术。

该技术通过故障数据采集和深度分析得到故障特征，然后作为输入进行AI故障预测模型训练和验证得到推理模型，可以在服务器的BMC上实现内存故障预测和自愈；同时，在服务器全生命周期平台FusionDirector中对该故障预测模型实施AI迁移学习算法，可实现应用场景快速适配，使内存故障预测结果更加准确，并及时对潜在内存故障进行自愈处理。当遇到无法自愈的故障时，FusionDirector发出预告警，及时通知用户，帮助用户结合上层业务主动迁移技术，实现业务秒级恢复，保障了服务器业务的稳定连续的运行。

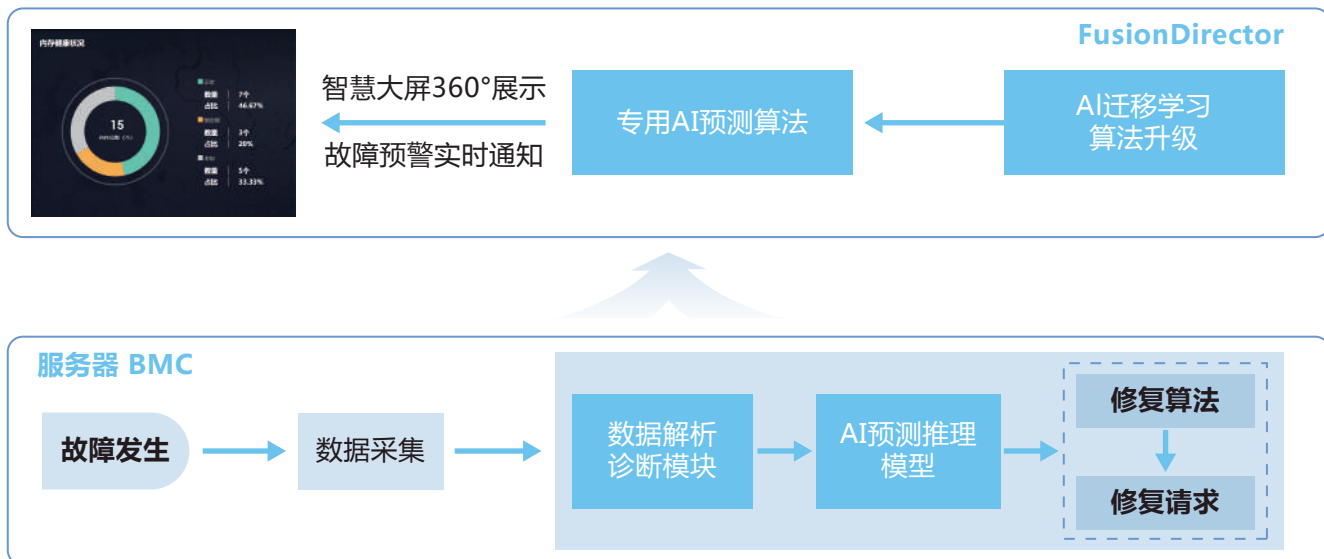


图4 招商银行“内存故障预测与自愈技术”示意图

4.1 内存故障预测算法指标

在AI预测算法模型验证中，主要的测试包括功能和效果两方面。功能包括错误类型判断，风险预警，自愈功能等；效果则主要是预测和自愈的准确率和覆盖率。其定义如下：

精准率： $\text{Precision} = \frac{TP}{FP+TP}$ ，表示预测为故障的内存条中有多少实际发生故障。

覆盖率： $\text{Recall} = \frac{TP}{FN+TP}$ ，表示预测为故障的内存条占实际发生故障的内存条的比，即实际发生故障的内存条能被预测的比例。内存条能被预测的比例。

公示中符号解释：

TP (True Positive)：预测结果为故障，实际为故障；

TN (True Negative)：预测结果为正常，实际为正常；

FP (False Positive)：预测结果为故障，实际为正常；

FN (False Negative)：预测结果为正常，实际发生故障；

下面以具体的数字为例说明精准率和覆盖的计算：

真实情况			预测准确		预测错误	
故障	FALSE	29	TP	19	FN	10
正常	TRUE	100	TN	96	FP	4

表1 内存预测中精准率和覆盖率说明

预测精准率： $\text{Precision} = \frac{TP}{FP+TP} = \frac{19}{19+4} = 83\%$

预测精准率： $\text{Recall} = \frac{TP}{FN+TP} = \frac{19}{19+10} = 66\%$

该项目拟定的精准率目标为83%，覆盖率目标为66%。

4.2 内存故障预测算法数据

近十年来，工业和学术界都开展过内存故障预测方向的相关研究探索，能否提前准确预测内存故障已成为大规模数据中心和云计算时代工业界需要研究和解决的重要问题之一。基于对招商银行数据中心运行服务器和超聚变内部客户服务器的底层运行日志数据的深入研究，招商银行和超聚变共同制定一套综合内存故障预测方案。在数据分析之前，要充分理解数据，整理事件日志信息中各寄存器的物理含义。招商银行搜集的结构化数据主要包含以下四个方面：服务器运行状态参数，关键部件固有类别参数，系统配置参数，和运行故障信息等，可以提供事件的详细信息，包括物理地址、错误时序、空间位置等。



图5 内存故障预测关键参数

4.3 预测和自愈算法模型训练

AI预测算法模型研究主要包括数据采集、特征提取、模型训练、模型验证等步骤。

4.3.1 数据采集和特征提取

按照故障的严重程度分类，故障可以分为CE和UCE。CE一般可以通过系统内ECC码进行纠错，而UCE往往会造成严重故障，导致系统崩溃宕机。而同一个内存的CE和UCE之间有一定的联系，且故障模式对内存后续故障有显著影响，研究CE和UCE的特征可以帮助指定内存故障预测方案。

招商银行提供了8500台服务器内存的事件日志用作模型训练和验证测试。超聚变使用多年累积的数万条故障内存的日志信息提取特征以输入训练模型。

- 1 对上述数据进行数据清洗和过滤，收集内存事件的DIMM信息，服务器信息，错误类型，错误的时间和空间信息等，并去掉重复的重启事件
- 2 解析CE、UCE等错误及其在微观位置分布信息，包括DIMM，存储排，颗粒，存储阵列，行、列和单元等信息，以及各内存中各事件的时序分布。
- 3 获取每个内存条历史发生的多个CE状态信息、UCE和重启信息，通过物理分布法和时序法，提取出多个维度的内存故障特征并总结专家规则：比如某个时间窗口内的CE计数、行故障（多个CE发生在同一行）计数；突发UCE的占比；CE故障累积和UCE故障的关系；CE故障模式与UCE的关系；故障率和内存厂商的关系等。

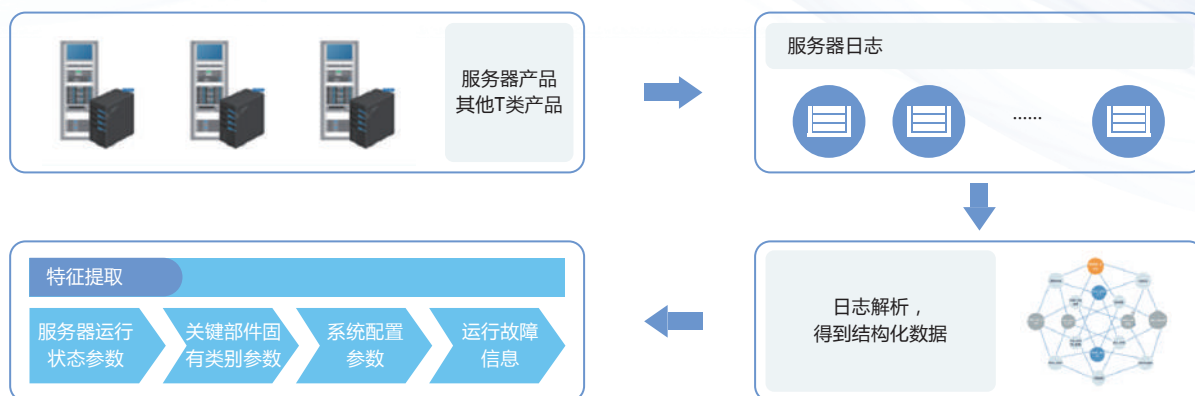


图6 数据采集流程图

4.3.2 AI预测算法模型训练和验证

将上述得到的专家规则定义为特征，然后与全量数据拼接，合成综合特征向量集，采用机器学习训练故障预测模型。继而在超聚变公司自动化测试工厂验证故障预测模型的功能，包括错误类型判断、风险预警、自愈功能等；同时验证故障预测模型的效果，即预测和自愈的精准率和覆盖率。将通过验证测试的训练后故障预测模型部署在内存故障管理方案中，可应用于智能运维平台，实现内存故障的预测和自愈。



图7 AI预测算法流程图

4.3.3 内存自愈修复方法

目前主要的内存自愈方案有PCLS、Page Offline、PPR和ADDDC等。当预测算法模型检测到故障符合模型预测，且故障特征符合专家规则时，触发系统故障自愈进行。故障修复往往对系统性能有一定的影响。针对不同粒度的故障模式，可选取对应粒度的自愈手段执行故障修复以避免系统宕机，在系统损耗最小的前提下提高故障修复的精准率。

修复手段	描述	类型	粒度	系统影响
Page Offline	OS内存管理页，故障时隔离下线	软隔离	页	4KB内存
PCLS	内存控制器集成冗余单元，替换故障区域	硬隔离	单元	无损
PPR	出现行故障时，使用冗余行替换故障行	硬隔离	行	无损
ADDDC	出现故障时使用内存校验颗粒替换故障内存区域	硬隔离	存储阵列，颗粒	性能轻微影响

表2 主要内存修复手段

4.4 AI预测模型使用的算法

本项目使用的主要算法有随机森林算法和迁移学习。

- 1 随机森林。随机森林是一种集成算法，通过组合多个弱分类器，最终结果通过投票或取均值，使得整体模型的结果具有较高的精确度和泛化性能。由于随机性，对于降低模型的方差很有作用，随机森林适合于“特征的数量很大，而且类别是不平衡”的情况，这正好符合内存的故障场景，在故障预测领域多次显示出优秀的预测结果。测试结果研究发现，在使用同样的数据和特征条件下，随机森林的预测效果优于极端梯度提升算法、卷积神经网络和长短期记忆网络等算法。在内存故障预测中，同时输入故障和非故障数据，采用分类算法进行计算，可以获得更高的准确性。
- 2 迁移学习。迁移学习是一种机器学习的方法，可以利用已有的知识和数据来解决新的问题和任务，在人脸识别、语音识别、自然语言处理等领域有着成功的应用。迁移学习的核心思想是不同任务之间存在共性和联系，将这些共性进行提取和转换，可以降低新任务的学习难度和数据量，提高任务学习效率。由于内存故障特征的共性，可以将内存已有数据故障数据作为源域，进行深度分析提取特征和规则，使用基于度量的方法，将学习得到的特征和规则应用在目标域中，用于预测在新任务下的内存故障。

应用实例：招商银行内存故障宕机率降低了66%

2022年8月，招商银行在1080台超聚变服务器上部署了FusionDirector内存故障预测和自愈技术，经过了一年的验证，其服务器内存年化宕机率降低了66%。



图8 内存故障预测和自愈应用场景

应用实例：招商银行内存故障宕机率降低了66%

分析招商银行数据中心2022年8月至2023年7月的1080台试点服务器，近2万条内存的运行日志发现，基于“FusionDirector内存故障预测和自愈技术”的创新技术中，有12台服务器进行过自愈操作，累积自愈了32根内存条；内存故障导致宕机的服务器有2台，年化宕机率仅为0.185%，相较于部署该系统之前降低了66.5%（注：统计未部署该系统的服务器8500台，其中内存宕机为47台，年化宕机率为0.553%），大大增强了招商银行数据中心的运行的稳定性和连续性，提高了其服务质量。

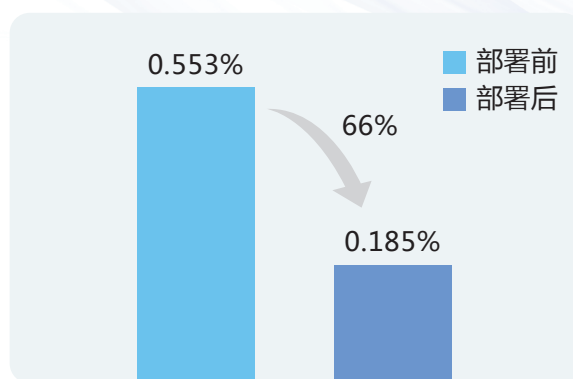


图9 内存预测与自愈技术部署前后内存年化宕机率对比

结论与展望

基于招商银行金融业务服务器运行的特异性日志以及超聚变多年的技术沉淀，招商银行和超聚变针对金融业务应用场景下共同进行了内存的故障机理挖掘、大数据分析、模型训练、故障自愈手段、故障预警等多方面的探索，推出的内存故障预测与自愈系统可准确识别现网中运行内存故障风险度，联动底层和上层自愈机制，实现了服务器内存导致宕机率降低了66%以上，提高了招商银行金融业务运行的稳定性。

在未来，招商银行仍将与超聚变开展广泛的技术合作，比如：数据中心智能运维的平台优化、金融行业多元化业务应用场景的拓宽，超聚变与招商银行的合作必将促进中国金融行业全面数字化的发展。

参考资料

- [1] : https://k.sina.com.cn/article_6506846481_183d6991101901byh7.html
- [2] : https://finance.sina.com.cn/money/bank/bank_hydt/2023-03-06/doc-imyiveww0947547.shtml
- [3] : <http://www.mofcom.gov.cn/article/i/jyjl/e/202004/20200402957398.shtml>

超聚变数字技术有限公司

咨询电话：400-080-6888 技术服务热线：400-009-899
地址：河南省郑州市郑东新区龙子湖智慧岛正商博雅广场1号楼9层
网址：www.xfusion.com



版权所有
© 超聚变数字技术有限公司 2023。保留一切权利。
非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明
XFUSION 和其他超聚变商标均为超聚变数字技术有限公司的商标。本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意
您购买的产品、服务或特性等应受超聚变数字技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。
除非合同另有约定，超聚变数字技术有限公司对本文档内容不做任何明示或默示的声明或保证。
由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。