 DOI: https://doi.org/10.12349/iser.v7i7.11403

Research on Hardware Failure Mode Recognition and Reliability Assessment for AI Servers

Jieqian Deng

Alibaba Cloud Computing Co., Ltd., Hangzhou, Zhejiang, 310012, China

Abstract

In response to the frequent hardware failures and passive maintenance practices in AI servers, this paper proposes a failure mode recognition framework based on FMEA. It identifies five major failure modes—GPU, video memory, thermal managementheat dissipation, power supply, and interconnect communication—and quantifies their risk levels using Risk Priority Number (RPN) for prioritization. By analyzing real-time monitoring data, actual failure rates of each component are statistically determined, and a reliability assessment model combining series and parallel configurations is established. The RPN values are used as weighting correction factors to further optimize maintenance priority decisions. Experimental results demonstrate that the proposed framework effectively identifies high-risk failure modes, and the reliability assessment outcomes closely match real-world measurements, providing quantitative support for proactive AI server maintenance.

Keywords

AI server; hardware failure mode; FMEA; reliability assessment; MTBF; RPN

AI 服务器硬件故障模式识别与可靠性评估研究

邓杰谦

阿里云计算有限公司，中国 · 浙江 杭州 310012

摘　要

针对AI服务器硬件故障频发、运维被动响应的现状，本文构建基于FMEA的硬件故障模式识别框架，归纳GPU、显存、散热、电源、互联通信五类故障模式，通过计算风险优先数（RPN）对各类故障进行量化风险排序；利用真实监测数据统计各组件实际故障率，建立串并联可靠性评估模型，以RPN作为权重修正系数，进一步优化维护决策优先级。实验结果表明，所提框架能有效识别高风险故障模式，可靠性评估结果与实测数据吻合良好，为AI服务器主动运维提供定量依据。

关键词

AI服务器；硬件故障模式；FMEA；可靠性评估；MTBF；RPN

1 引言

大规模 AI 模型训练对计算节点的连续运行提出了极高要求。GPU 集群、HBM 显存、NVLink 互联总线等关键硬件长期处于高负载状态，任意组件发生故障均可导致整批训练任务中断，算力损失极为显著。现有数据中心运维多以被动响应为主，缺乏针对 AI 服务器特有硬件架构的系统性故障分析与可靠性定量评估手段 [1]。现有研究或聚焦于通用服务器故障检测，或针对单一硬件组件建立可靠性模型，面向 AI 服务器整机硬件的系统性研究仍较为有限。为此，采用 FMEA 方法建立故障模式识别框架并完成风险排序，以实测故障率构建串并联可靠性评估模型，以 RPN 为纽带形成完整分析链，就显得尤为重要。

【作者简介】邓杰谦（1988–），男，中国广西柳州人，硕士，主要从事AI服务器的硬件开发、测试、导入与优化以及计算机硬件产品测试与故障分析研究。

2 AI 服务器硬件故障数据分析与基础

2.1 AI 服务器硬件架构与故障来源

AI 服务器在硬件构成上与通用服务器存在显著差异。计算核心由多块 GPU 组成，每块 GPU 配备 HBM 高带宽显存，GPU 之间通过 NVLink 高速互联总线直连，系统配备高密度散热模组与冗余电源单元。硬件故障来源可归纳为三个维度：制造层面的芯片制程缺陷等先天因素；运行层面的长期高负载导致组件逐渐老化；环境层面的机房温湿度波动、供电质量不稳定等外部应力加速组件失效 [2]。

2.2 主要故障类型归纳

结合 AI 服务器特有硬件结构，将主要故障归纳为五类，作为后续 FMEA 分析与可靠性建模的研究边界。

（1）GPU 核心故障：监控来源于 DCGM，典型指标为 XID 错误计数（DCGM_FI_DEV_XID_ERRORS）异常。

（2）显存故障：监控来源于 DCGM，典型指标为单 bit 可纠正错误（DCGM_FI_DEV_ECC_SBE_VOL_TOTAL）与双 bit 不可纠正错误（DCGM_FI_DEV_ECC_DBE_VOL_TOTAL）持续累积。

（3）散热系统故障：监控来源于 IPMI，典型指标为风扇转速传感器读数异常与进风口温度传感器（Inlet Temp）持续超过 85℃。

（4）电源 / 供电故障：监控来源于 IPMI，典型指标为电源状态寄存器异常与输出电压传感器超限。

（5）互联通信故障：监控来源于 OFED 驱动接口与 Advanced Error Reporting，典型指标为 InfiniBand 端口接收错误（PortRcvErrors）与 PCIe 不可纠正错误（Uncorrected Errors）计数异常。

2.3 数据采集与预处理

数据采集覆盖上述五类故障的全部监控指标，采集频率为 1 次 /min，采集周期 12 个月，数据来源包括 DCGM、IPMI 与 BMC 三套监控系统。预处理包含三步：（1）依据 3σ 原则剔除异常跳变值；（2）对维护窗口导致的缺失段进行线性插值填充；（3）将三套系统时间戳统一对齐至 1 min 粒度。预处理后输出两类数据：（1）故障特征样本集，记录各类故障发生时的传感器特征信号；（2）故障时间记录集，记录各组件实际故障次数与累计运行时间 [3]。

3 基于 FMEA 的硬件故障模式识别

3.1 FMEA 方法原理与选用依据

FMEA（失效模式与影响分析）针对每种潜在失效模式，从严重度、发生度、探测度三个维度进行量化评估，以风险优先数 RPN 综合表征风险等级 [4]。AI 服务器高危故障发生频率低但影响极为严重，纯数据驱动方法在样本量不足时难以可靠建模，FMEA 将领域知识与历史数据相结合，适用于此类场景。需明确职责边界：S/O/D 评分用于计算 RPN，进而完成故障风险排序，O 值反映历史发生频次的相对高低，不替代基于实测数据统计的故障率 λ。

3.2 AI 服务器 FMEA 分析表构建

针对五类故障，结合故障特征样本集评分，代入公式（1）计算 RPN：

$$RPN = S \times O \times D \tag{1}$$

式中：S 为严重度（1 ~ 10），反映故障对 AI 训练任务的中断程度；O 为发生度（1 ~ 10），依据历史故障频次评估；D 为探测度（1 ~ 10），反映故障被现有监控手段发现的难易程度；对于已有成熟告警规则的故障类型，D 值相应取低。AI 服务器硬件 FMEA 分析见表 1。

表 1 AI 服务器硬件 FMEA 分析汇总表

故障类型	典型失效模式	S	O	D	RPN	风险等级
GPU 核心故障	计算单元损坏 /NVLink 失效	10	4	3	120	高
散热系统故障	风扇失效 / 温度超阈值	8	6	2	96	高
互联通信故障	InfiniBand 丢包 /PCIe 错误	7	4	3	84	中
显存故障	ECC 错误持续累积	9	3	3	81	中
电源 / 供电故障	冗余电源单元单路失效	5	3	2	30	低

注：表中 S/O/D 评分依据历史故障记录及领域专家经验综合评定，数据来源于本文实验集群 12 个月运行数据

3.3 故障风险识别结果与薄弱项筛选

由表 1 可见，GPU 核心故障（RPN=120）与散热系统故障（RPN=96）风险等级最高，构成 AI 服务器最主要的薄弱环节。散热故障发生度 O 值最高（O=6），日常运行中发生频率最突出，是预防性维护的首要对象；GPU 核心故障严重度 S 值最高（S=10），一旦发生即导致全量训练任务中断，需重点监控早期预警信号。电源 / 供电故障因冗余机制 RPN 最低，但仍需纳入定期巡检。各类故障典型预警信号：ECC_ERRORS_UNCORRECTED 持续增长（显存故障早期预警）；INLET_TEMP 连续 3 个采集周期超过 85℃（散热故障预警）；PortRcvErrors 在 1 h 内累计超过 100 次（互联故障预警）[5]。

4 AI 服务器硬件可靠性评估模型

4.1 可靠性评估指标体系

各组件故障率 λ 基于故障时间记录集统计得出：

$$\lambda_i = \frac{n_i}{T_i} \tag{2}$$

式中：*ni* 为第 *i* 类组件观测期内实际故障次数（次），*Ti* 为累计运行时间（h），*λi* 为故障率（次 /h），完全来源于实测数据，与 FMEA 中 O 值性质不同。各组件可靠度函数为：

$$R_i(t) = e^{-\lambda_i t} \tag{3}$$

式中：*t* 为规定运行时长（h），*Ri*=(*t*）为第 *i* 类组件在时间 t 内无故障运行的概率（0 ~ 1）。电子硬件组件在正常使用阶段故障率近似恒定，指数分布具有良好工程适用性。辅助指标：平均故障间隔时间（MTBF）取故障率 λ 的倒数，单位为 h；设备可用度定义为 MTBF 与（MTBF 加平均修复时间 MTTR）之比，MTTR 单位为 h。

4.2 基于串并联模型的整机可靠性建模

根据 AI 服务器硬件拓扑，建立串并联可靠性模型。GPU 核心、HBM 显存与互联通信构成串联主链路，任一失

效即导致系统中断；电源单元与散热风扇组采用冗余并联设计，单点失效不中断系统。具体拓扑关系如图 1 所示。

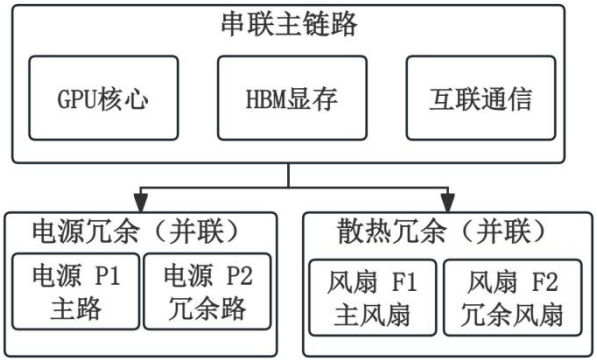


图 1 AI 服务器串并联可靠性模型结构图

基于图 1 拓扑，整机系统可靠度由公式（4）计算：

$$R_{sys}(t)=\prod_{i\in S}R_i(t)\times\prod_{j\in P}\left[1-\prod_{k\in P_j}(1-R_k(t))\right] \quad (4)$$

式中：S 为串联组件集合（GPU 核心、显存、互联通信），P 为并联冗余组件集合（电源、散热风扇）。将各组件 RPN 值归一化，得维护优先权重 wi = RPNi / ∑RPNi，权重越大的组件在预防性维护计划中排期越靠前，修正作用于维护决策层，不改变实测 λ 值。

5 实验验证与结果分析

5.1 实验环境与数据说明

实验数据来源于某 AI 计算集群 32 台同型号服务器，每台配备 8 块 A100 GPU，运行负载以大模型分布式训练为主。数据采集周期为 12 个月，共计 8 760 h，通过 DCGM、IPMI 与 BMC 同步采集。观测期内记录有效故障事件 87 次：散热系统故障 34 次（39.1%）、GPU 核心故障 18 次（20.7%）、互联通信故障 16 次（18.4%）、显存故障 12 次（13.8%）、电源故障 7 次（8.0%）。

5.2 故障模式识别结果验证

RPN 排序（高至低）：GPU 核心（120）> 散热（96）> 互联通信（84）> 显存（81）> 电源（30）。实际故障频次排序（高至低）：散热（34 次）> GPU 核心（18 次）> 互联通信（16 次）> 显存（12 次）> 电源（7 次）。两种排序前四位基本吻合，且均将电源故障列于末位。散热与 GPU 核心位次互换，原因在于 FMEA 综合考量了故障后果严重性，GPU 核心 S=10 的严重性远高于散热 S=8，二者结论互补而非矛盾。各类故障特征信号识别命中率：散热故障 94.1%、GPU 核心故障 83.3%、显存故障 91.7%、互联通信故障 87.5%、电源故障 100%。

5.3 可靠性评估结果验证

基于实测数据代入公式（2）计算各组件 λ 值，进而由公式（3）（4）推导各项可靠性指标，结果汇总于表 2。

表 2 各组件实测可靠性指标汇总（表中数据为各类组件所有单体设备的数据总和）

组件	故障次数	运行时间（h）	λ（次 /h）	MTBF(h)	R(1000h)
GPU 核心	18	2242560	8.03×10^{-6}	124587	0.992
HBM 显存	12	2242560	5.35×10^{-6}	186880	0.995
散热系统	34	280320	1.21×10^{-4}	8245	0.886
电源（冗余）	7	280320	2.50×10^{-5}	40046	0.975
互联通信	16	280320	5.71×10^{-5}	17518	0.945
整机系统	—	—	—	—	0.919

由表 2 可见，整机 R(1000h)=0.919，散热系统 R(1000h)=0.886 为串联组件中最低，是拖低整机可靠度的主要瓶颈；电源组件经并联冗余修正后 R(1000h)=0.975，冗余设计效果显著。为直观呈现各组件可靠度随运行时长的衰减趋势，绘制 R(t) 曲线如图 2 所示。

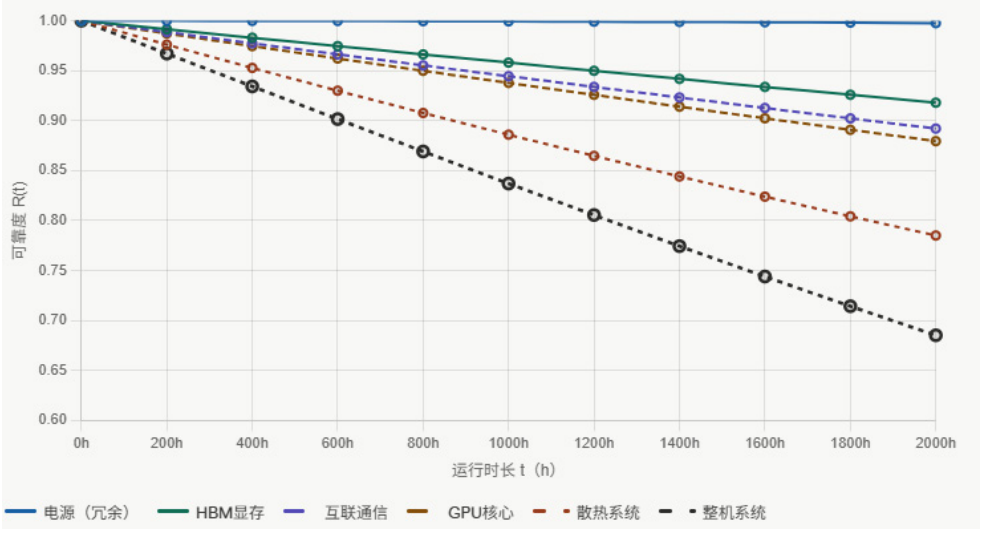


图 2 各组件及整机 R(t) 可靠度随时间变化曲线

由图2可见，散热系统可靠度曲线在各组件中下降最快，与 FMEA 中散热故障发生度 O 值最高（O=6）的结论一致；电源组件曲线因并联冗余下降最慢，整机曲线受散热与 GPU 核心双重拖累，整机曲线在 2000 h 时约为 0.826。为进一步验证 RPN 权重修正机制对整机可靠性的改善效果，对 RPN 优先维护策略与随机维护策略进行对比，结果如表 3 所示。

表 3 RPN 优先维护与随机维护整机可靠度对比

运行时长（h）	RPN 优先维护	随机维护	可靠度提升幅度
500	0.977	0.962	1.50%
1 000	0.962	0.919	4.60%
2 000	0.902	0.826	9.10%

由表 3 可见，随运行时长增加，RPN 优先维护策略的可靠度提升幅度持续扩大，在 2 000 h 节点达 9.1%，说明基于风险优先级的预防性维护对长期系统稳定性具有显著改善效果。

6 结语

本研究构建了面向 AI 服务器的 FMEA 故障模式识别框架与串并联可靠性评估模型，以 RPN 为纽带实现两部分联动。实验结果表明，散热系统是当前配置下整机可靠性的主要瓶颈，RPN 优先维护策略在 2 000 h 运行周期内可将整机可靠度提升 9.1%。后续研究可从两个方向拓展：一是引入机器学习方法实现 FMEA 评分的自动化与动态更新；二是将本文模型扩展至更大规模集群，开展分布式场景下的整机可靠性评估。

参考文献

[1] 张文波,贾孟君,明亮. 基于EfficientNet的服务器硬盘故障识别[J].广播电视信息,2025,32(10):106-109.

[2] 李逍,张亚强,史宏志. 服务器故障注入技术综述：实现、优化与评估[J].计算机应用,2025,45(S1):144-153.

[3] 王宁. 云计算数据中心服务器检测维修模块化教学的实践路径[J].信息与电脑,2025,37(04):203-205.

[4] 李盛新,叶丰华,李道童,等. 基于IQPSO-GA优化ANFIS模型的服务器故障预警方法[J].计算机测量与控制,2024,32(04):37-45.

[5] 刘伟峰. 基于机器学习的大规模并行计算机系统硬件故障检测方法[J].技术与市场,2022,29(06):51-52+55.