迭代评估报告

　　　　　　　　　　　　　　　　　　评估日期：2025年1月12日

|  |  |  |  |
| --- | --- | --- | --- |
| 组号 | 5 | 项目名称 | 集群性能监控系统及其自动化 |
| 迭代名称 | 最终测试与交付 | 实际起止日期 | 2025年1月6日-2025年1月12日 |
| 任务达成情况：（完成的任务、实现的功能、进度、质量等）   | No | 任务 | 负责人 | 完成情况 | | --- | --- | --- | --- | | 1 | 优化 DCGM 和 Exporter 的性能，减少采集延迟 | 吴伯涛、尚子森 | 完成，采集延迟降低至合理范围内。 | | 2 | 实现多节点 GPU 监控支持，扩展集群监控能力 | 李双威、申煜坤 | 完成，支持多节点监控及扩展验证通过。 | | 3 | 完成系统整合测试，修复所有已知问题 | 全体成员 | 完成，所有问题已修复并通过测试。 | | 4 | 撰写部署文档、用户手册与测试报告 | 全体成员 | 完成，文档质量达标，易读性和完整性良好。 | | 5 | 系统试运行并验收 | 全体成员 | 完成，系统通过验收，稳定性满足要求。 | | | | |
| 评审/测试的结果：（执行了哪些评审和测试？评审和测试的结果如何？）  1、评审和测试  **功能测试：**   * 1. 验证 GPU 监控系统是否满足基本功能需求，包括 GPU 数据采集、存储、展示和告警。   2. 涉及的功能：      1. DCGM 和 Exporter 数据采集。      2. Prometheus 数据存储及查询。      3. Grafana 数据可视化和仪表板展示。      4. GPU 性能与健康告警触发机制。   3. **测试方法**：      1. 使用单节点和多节点 GPU 环境，模拟实际运行场景，采集并验证监控数据的完整性与准确性。   **性能测试：**   * 1. 测试系统的响应速度和数据采集延迟，特别是在多节点监控场景下的表现。   2. 测试内容：      1. 数据采集频率（延迟 <1 秒）。      2. Grafana 仪表板加载速度。      3. Prometheus 数据查询性能。   3. **测试方法**：      1. 提高数据采集频率并模拟高负载环境，检测系统在高压力下的稳定性。   **整合测试：**   * 1. 验证 DCGM、Prometheus 和 Grafana 各模块之间的协作情况，确保数据流畅传递和展示。   2. 涉及内容：      1. DCGM 与 Exporter 数据接口正确性。      2. Prometheus 查询 Grafana 数据的实时性和完整性。   3. **测试方法**：      1. 模拟完整运行环境，进行多次数据流完整性验证。   **告警测试：**   * 1. 测试告警机制的准确性和及时性，包括误报和漏报的检测。   2. 测试内容：      1. GPU 温度超标告警。      2. 显存使用率异常告警。      3. GPU 利用率过高的告警。   3. **测试方法**：      1. 人为制造异常情况（例如超高温度、高显存占用），检查系统是否触发告警。   **验收测试：**   * 1. 模拟实际使用场景，验证系统的整体运行效果和用户体验。   2. 测试内容：      1. 系统部署流程是否完整无误。      2. 监控数据展示的可读性和清晰度。      3. 用户手册与文档的易用性。   3. **测试方法**：      1. 邀请目标用户参与试运行，记录反馈并调整。  **2、评审和测试的结果** **功能测试结果：**   * 1. DCGM 和 Exporter 数据采集功能正常，所有指标（温度、功耗、利用率等）均能准确采集。   2. Prometheus 成功存储并提供 GPU 数据查询，数据完整无误。   3. Grafana 仪表板功能完善，能够直观展示 GPU 性能与健康数据。   4. 告警机制正常触发，误报率低于 5%。   **性能测试结果：**   * 1. 单节点 GPU 数据采集延迟 <0.8 秒，多节点（10 个节点）情况下延迟稳定在 <1 秒范围内。   2. Prometheus 数据查询性能表现良好，高并发情况下响应时间 <500 毫秒。   3. Grafana 仪表板加载速度快，支持大规模数据展示，加载时间 <1 秒。   **整合测试结果：**   * 1. DCGM、Prometheus 和 Grafana 数据流畅传递，无数据丢失或延迟问题。   2. 各模块之间的集成顺畅，功能无冲突。   **告警测试结果：**   * 1. 所有预定义告警（温度、功耗、利用率等）均能准确触发，误报率控制在 3%-5% 范围内。   2. 调整告警阈值后，告警准确性进一步提高。   **验收测试结果：**   * 1. 系统通过模拟场景验收，用户反馈良好，表示功能齐全，易用性较高。   2. 文档内容详尽，包括部署、操作和告警规则说明，未发现遗漏。   3. 部署流程清晰，安装步骤顺畅，试运行环境表现稳定。  **总体评审/测试结论** **系统功能和性能均符合预期**，达到设计目标，测试通过。  **系统稳定性良好**，支持多节点环境，数据采集和展示实时性满足需求。  **用户体验优异**，用户手册与文档完整易懂。  **主要风险已被有效规避**，未发现重大缺陷或隐患。 | | | |
| 问题、变更和返工：（遇到的问题、发生的变更、是否需要返工等）  遇到的问题：   * 1. 告警规则误报率较高（初期测试阶段约 10%），解决措施：根据实际运行环境调整告警阈值，重新定义规则，降低误报至 <5%。   2. 多节点监控中部分节点数据丢失，解决措施：检查 Prometheus 配置并调整 scrape\_interval，确保数据完整采集。   发生的变更：   * 1. 调整告警阈值（如 GPU 温度上限由 80°C 调整为 85°C）。   2. 优化 Prometheus 数据存储规则（压缩长时间数据存储）。   发生的返工：无 | | | |
| 经验和教训：   1. 细化任务分工并结合成员的专长，可以大幅提高团队效率，减少重复工作和资源浪费。   2.整合测试应提前安排，避免将模块问题积压到最终阶段，影响问题排查效率。  3.功能测试、性能测试、整合测试和告警测试覆盖了系统的核心功能，特别是通过模拟高负载场景验证了多节点支持的稳定性，全面且系统化的测试计划能够有效发现潜在问题，确保系统在复杂场景中的稳定性和可靠性。 | | | |
|  | | | |