# MVS 장애분석 레포트

## 1. 기본 정보

|  |  |
| --- | --- |
| **구분** | **내용** |
| 장애 ID | IB-PERF-2023-001 |
| 고객사 | 연구개발센터 |
| 사업장 | R&D 클러스터 |
| 장애 유형 | 네트워크 |
| 장애 등급 | Minor |
| 영향 범위 | 단일 서버 (cl-rndcgpu-a18) |
| 담당 엔지니어 | 박지원(GPU마스터) |
| 작성자 | 정민우(로그헌터) |

## 2. 장비 정보

|  |  |
| --- | --- |
| **구분** | **내용** |
| 모델명 | HPE DL380 Gen10 서버 |
| 시리얼 번호 | CNDCGPU-A18 |
| 운영체제 | Linux (Red Hat Enterprise Linux) |
| CPU | Intel Xeon, 다중 코어 |
| 메모리 | 128GB 이상 |
| 디스크 | SSD 스토리지 구성 |
| 네트워크 | InfiniBand MCX653105A HDAT Adapter Network |
| 설치 위치 | R&D 클러스터 A존 |

## 3. 장애 개요

|  |  |
| --- | --- |
| **구분** | **내용** |
| 장애 발생 시간 | 2023-09-12 01:36:09 |
| 장애 감지 시간 | 2023-09-12 02:15:00 |
| 장애 해결 시간 | 진행 중 |
| 총 장애 시간 | 진행 중 |
| 장애 증상 | InfiniBand 네트워크 성능 심각한 저하 (정상 대비 약 0.0016% 성능) |
| 비즈니스 영향 | R&D 클러스터의 단일 노드 성능 저하로 해당 노드 작업 처리 지연 |

## 4. 장애 타임라인

|  |  |  |  |
| --- | --- | --- | --- |
| **시간** | **상태** | **조치 내용** | **담당자** |
| 01:36:09 | 장애 발생 | IML 로그에 "Critical System Error Server Critical Fault (Service Information: Standby Fault, GPU Board/Modules, GPU Board VRD 01Hh)" 기록 | 시스템 |
| 01:49:01 | 네트워크 복구 시도 | NIC 상태 복구 시도 로그 기록 | 시스템 |
| 02:15:00 | 장애 감지 | 성능 모니터링 시스템에서 비정상적인 InfiniBand 처리량 감지 | 모니터링 시스템 |
| 02:30:00 | 초기 대응 | ib\_send\_bw 테스트 실행하여 성능 저하 확인 | 박지원 |
| 03:15:00 | 원인 분석 | IML 로그 분석 및 네트워크 상태 확인 | 정민우 |
| 04:00:00 | 조치 계획 수립 | PCIe 설정 검토 및 GPU 보드 관련 문제 연관성 분석 | 이동훈, 박지원 |

## 5. 구성요소별 점검 결과

### 5.1 하드웨어 점검

|  |  |  |  |
| --- | --- | --- | --- |
| **항목** | **점검 내용** | **결과** | **상세 내용** |
| CPU | 시스템 로그 확인 | 정상 | CPU 관련 오류 없음 |
| 메모리 | 이벤트 로그 확인 | 정상 | 메모리 관련 오류 없음 |
| 디스크 | 스토리지 상태 확인 | 정상 | 디스크 관련 오류 없음 |
| 네트워크 카드 | InfiniBand 성능 테스트 | 비정상 | ib\_send\_bw 테스트 결과 약 0.000839 Gb/sec (정상 대비 매우 낮음) |
| 전원 공급 장치 | 전원 상태 확인 | 정상 | 전원 관련 경고 없음 |
| 냉각 장치 | 온도 센서 확인 | 정상 | 온도 관련 경고 없음 |
| GPU 보드 | IML 로그 확인 | 비정상 | "GPU Board VRD 01Hh" 관련 Critical 오류 발견 |

### 5.2 운영체제 점검

|  |  |  |  |
| --- | --- | --- | --- |
| **항목** | **점검 내용** | **결과** | **상세 내용** |
| 시스템 로그 | 로그 파일 분석 | 정상 | OS 레벨에서 특이사항 없음 |
| 이벤트 로그 | 이벤트 로그 분석 | 정상 | 관련 경고 없음 |
| 패치 상태 | MLNX\_OFED 드라이버 확인 | 정상 | MLNX\_OFED\_LINUX-5.5-1.0.3.2 버전 설치됨 |
| 리소스 사용량 | 시스템 리소스 모니터링 | 정상 | CPU, 메모리 사용량 정상 |
| 서비스 상태 | InfiniBand 관련 서비스 확인 | 정상 | 서비스 실행 중 |
| 드라이버 상태 | HCA 펌웨어 버전 확인 | 정상 | 펌웨어 버전: v20.32.1010 |

### 5.3 네트워크 점검

|  |  |  |  |
| --- | --- | --- | --- |
| **항목** | **점검 내용** | **결과** | **상세 내용** |
| NIC 상태 | HCA 상태 확인 | 정상 | 물리적 상태 정상 |
| 링크 상태 | InfiniBand 링크 확인 | 정상 | Active/LinkUp 및 4X 53.125 Gbps 상태 |
| 케이블 상태 | 케이블 연결 상태 확인 | 정상 | 물리적 연결 정상 |
| 에러 카운터 | Error Counter Check | 정상 | "Error Counter Check on CA" 테스트 PASS |
| 네트워크 성능 | ib\_send\_bw 테스트 | 비정상 | 1차: 0.000734 Gb/sec, 2차: 0.000839 Gb/sec |

### 5.4 GPU 점검

|  |  |  |  |
| --- | --- | --- | --- |
| **항목** | **점검 내용** | **결과** | **상세 내용** |
| GPU 상태 | IML 로그 분석 | 비정상 | Critical System Error - GPU Board VRD 01Hh 발견 |
| GPU 드라이버 | 드라이버 상태 확인 | 정상 | 드라이버 로드됨 |
| P2P 설정 | NVIDIA P2P 설정 확인 | 확인 필요 | GPU와 InfiniBand 간 P2P 설정 확인 필요 |

## 6. 장애 원인 분석

### 6.1 장애 현상

cl-rndcgpu-a18 서버의 InfiniBand MCX653105A HDAT Adapter Network에서 심각한 성능 저하가 발생했습니다. ib\_send\_bw 테스트 결과 데이터 전송률이 약 0.000839 Gb/sec로 측정되었으며, 이는 정상적인 HDR InfiniBand 연결 속도(53.125 Gbps)의 약 0.0016% 수준입니다.

### 6.2 근본 원인 분석

로그 분석 결과, 다음과 같은 원인이 식별되었습니다:

1. **GPU 관련 하드웨어 오류와의 연관성:** IML 로그에서 "Critical System Error Server Critical Fault (Service Information: Standby Fault, GPU Board/Modules, GPU Board VRD 01Hh)" 메시지가 기록되었습니다. 이 메시지는 GPU 보드의 전압 조정기(VRD)에 문제가 발생했음을 나타내며, 이는 InfiniBand 성능 저하와 시간적으로 연관성이 있습니다.
2. **PCIe 설정 문제:** 하드웨어 장애 유형별 장애분석 가이드라인에 따르면, Infiniband 성능 문제는 하드웨어 장애보다 PCIe relax order 설정 및 ibv\_wr\*API 활성화 여부 등의 구성 최적화 이슈인 경우가 많습니다. GPU 보드 오류가 PCIe 버스에 영향을 미쳤을 가능성이 있습니다.
3. **물리적 연결 정상:** 물리적 연결 상태와 링크 속도(4X 53.125 Gbps)는 정상으로 보이며, 에러 카운터 검사에서도 문제가 발견되지 않았습니다. 이는 물리적 케이블이나 포트 손상이 아닌 다른 원인을 시사합니다.

### 6.3 영향 요인

다음 요인들이 장애에 영향을 미쳤을 가능성이 있습니다:

* GPU 보드 VRD(Voltage Regulator Device) 오류로 인한 GPU-InfiniBand 간 통신 경로 영향
* PCIe 설정 최적화 부족(PCIe relax order 및 ibv\_wr\*API 설정 미비)
* NVIDIA P2P(Peer-to-Peer) 설정 문제
* 펌웨어 버전 호환성 문제 가능성

## 7. 조치 사항

### 7.1 임시 조치

현재 다음과 같은 임시 조치를 권장합니다:

* InfiniBand 드라이버 재로드: systemctl restart opensmd 및 systemctl restart openibd 명령 실행
* 서버 재부팅을 통한 GPU 및 InfiniBand 하드웨어 초기화 시도
* PCIe relax order 설정 활성화: BIOS 설정 또는 OS 매개변수 조정
* 해당 노드의 작업을 다른 정상 노드로 마이그레이션하여 서비스 영향 최소화

### 7.2 영구 조치

근본적인 해결을 위해 다음과 같은 영구 조치가 필요합니다:

* GPU 보드 VRD 관련 하드웨어 점검 및 필요시 교체
* ibv\_wr\*API 활성화 및 OS 레벨 Infiniband 구성 최적화
* PCIe 설정 최적화를 위한 BIOS 업데이트
* InfiniBand HCA 펌웨어 최신 버전으로 업데이트
* NVIDIA GPU와 Mellanox InfiniBand 간 최적 구성 설정 적용

### 7.3 검증 방법

조치 후 다음과 같은 방법으로 문제 해결 여부를 검증해야 합니다:

* ib\_send\_bw 테스트를 통한 InfiniBand 성능 측정(정상 속도는 40-50 Gb/sec 이상)
* GPU-InfiniBand 간 P2P 성능 테스트
* IML 로그에서 GPU 관련 오류 메시지 모니터링
* 실제 워크로드 실행 테스트(ML/AI 작업 등)

### 7.4 검증 결과

아직 조치 적용 전이므로 검증 결과 없음

## 8. 향후 조치 계획

### 8.1 단기 조치 계획

|  |  |  |  |
| --- | --- | --- | --- |
| **조치 내용** | **담당자** | **예정일** | **우선순위** |
| InfiniBand 드라이버 재로드 및 서버 재부팅 | 박지원 | 즉시 | High |
| PCIe relax order 설정 확인 및 변경 | 이동훈 | 24시간 이내 | High |
| GPU VRD 상태 상세 점검 | 박지원 | 48시간 이내 | Medium |

### 8.2 장기 조치 계획

|  |  |  |  |
| --- | --- | --- | --- |
| **조치 내용** | **담당자** | **예정일** | **우선순위** |
| GPU 보드 VRD 교체 또는 수리 | 이동훈 | 1주일 이내 | High |
| InfiniBand HCA 펌웨어 업데이트 | 박지원 | 1주일 이내 | Medium |
| GPU-InfiniBand 최적화 구성 적용 | 박지원 | 2주일 이내 | Medium |

## 9. 재발 방지 대책

### 9.1 모니터링 강화

다음과 같은 모니터링 강화 방안을 제안합니다:

* InfiniBand 성능에 대한 주기적 모니터링 추가(시간당 샘플링)
* GPU 전압 및 VRD 상태에 대한 모니터링 추가
* PCIe 버스 성능 및 오류에 대한 모니터링 추가
* InfiniBand-GPU 간 P2P 성능에 대한 주기적 검증 수행

### 9.2 시스템 개선

다음과 같은 시스템 개선 방안을 제안합니다:

* 정기적인 InfiniBand 및 GPU 펌웨어 업데이트 일정 수립
* GPU 서버에 대한 PCIe 설정 표준화 및 최적화 가이드 작성
* InfiniBand 네트워크 중복성 강화
* GPU VRD 관련 예방적 유지보수 일정 수립

### 9.3 프로세스 개선

다음과 같은 프로세스 개선 방안을 제안합니다:

* GPU-InfiniBand 통합 장애 분석 프로세스 개발
* 고성능 컴퓨팅 환경에 특화된 장애 분류 및 대응 프로세스 확립
* 성능 저하 감지를 위한 사전 경고 임계값 설정
* GPU 관련 하드웨어 오류와 네트워크 성능 간 상관관계 패턴 라이브러리 구축

### 9.4 교육 및 훈련

다음과 같은 교육 및 훈련 방안을 제안합니다:

* GPU-InfiniBand 통합 환경 트러블슈팅 교육 실시
* NVIDIA GPU VRD 관련 장애 대응 훈련
* InfiniBand 성능 튜닝 및 최적화 워크샵
* PCIe 최적화 및 구성 교육

## 10. 첨부 자료

### 10.1 로그 파일

* IML(Integrated Management Log) 로그
* ib\_send\_bw 테스트 결과 로그
* 시스템 로그

### 10.2 구성 파일

* InfiniBand 설정 파일
* MLNX\_OFED 드라이버 구성
* PCIe 구성 설정

### 10.3 스크린샷

* InfiniBand 성능 테스트 결과 화면
* IML 로그 캡처 화면
* GPU 상태 정보 화면

### 10.4 참고 문서

* 하드웨어 장애 유형별 장애분석 가이드라인
* 임시 조치 매뉴얼
* 장애등급분류표
* NVIDIA GPU 및 InfiniBand 최적화 가이드

## 11. 승인

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **구분** | **이름** | **직책** | **서명** | **일자** |
| 작성자 | 정민우 | 통합 로그분석 전문가 |  | 2023-09-13 |
| 검토자 | 박지원 | GPU 전문가 |  | 2023-09-13 |
| 승인자 | 이해경 | PO |  | 2023-09-14 |