# GPU 서버 장애분석 레포트

## 1. 기본 정보

|  |  |
| --- | --- |
| **구분** | **내용** |
| 장애 ID | GPU-ERR-20240809-001 |
| 고객사 | 클라우드 연구개발센터 |
| 사업장 | 연구개발 클러스터 |
| 장애 유형 | 서버/GPU |
| 장애 등급 | **Major** |
| 영향 범위 | GPU 서버 2대 (cl-rndcgpu-a02, cl-rndcgpu-a03) |
| 담당 엔지니어 | 박지원 (GPU마스터), 010-1234-5678 |
| 작성자 | 정민우 (로그헌터), 010-5678-1234 |

## 2. 장비 정보

|  |  |
| --- | --- |
| **구분** | **내용** |
| 모델명 | HPE ProLiant DL380 Gen10 Plus |
| 시리얼 번호 | SGH112VPXN (cl-rndcgpu-a02), SGH113WM77 (cl-rndcgpu-a03) |
| 운영체제 | Linux (Kernel 버전: 5.4+) |
| CPU | AMD EPYC 7002 Series (Starship/Matisse), 다중 소켓 구성 |
| 메모리 | 대용량 서버급 메모리 구성 |
| GPU | NVIDIA A100-SXM4-80GB 각 서버당 8개 |
| 디스크 | Samsung NVMe SSD (PM173X) 및 기타 스토리지 |
| 네트워크 | Mellanox ConnectX-6/ConnectX-6 Dx Infiniband |
| 설치 위치 | 연구개발 클러스터 랙 |

## 3. 장애 개요

|  |  |
| --- | --- |
| **구분** | **내용** |
| 장애 발생 시간 | 2024-08-09 (정확한 시간 불명) |
| 장애 감지 시간 | 2024-08-09 16:08:52 (로그 타임스탬프 기준) |
| 장애 해결 시간 | 진행 중 |
| 총 장애 시간 | 진행 중 |
| 장애 증상 | * cl-rndcgpu-a02: GPU ID 3 (BusID: 00000000:4C:00.0)에서 InfoROM 손상 경고 발생 * cl-rndcgpu-a03: GPU ID 7 (BusID: 00000000:CB:00.0)에서 DRAM 메모리 오류 다수 발생 및 Row Remapping 진행 중 |
| 비즈니스 영향 | * 현재 GPU 워크로드 처리는 계속 진행 중이나 신뢰성 저하 * ML/AI 연구 작업 일부 서비스 품질 저하 가능성 * 장기적으로 해결하지 않을 경우 서비스 중단 위험성 높음 |

## 4. 장애 타임라인

|  |  |  |  |
| --- | --- | --- | --- |
| **시간** | **상태** | **조치 내용** | **담당자** |
| 불명 | 장애 발생 | GPU 하드웨어 이슈 발생 (정확한 시점 파악 불가) | - |
| 16:08:52 | 장애 감지 | cl-rndcgpu-a02 서버에서 nvidia-smi 실행 시 InfoROM 손상 경고 발견 | 시스템 모니터링 |
| 16:09:00 | 초기 조사 | nvidia-smi -q 명령어로 상세 정보 수집 | 운영자 |
| 16:45:26 | 추가 조사 | cl-rndcgpu-a03 서버 상태 확인 | 운영자 |
| 17:01:23 | 심층 분석 | cl-rndcgpu-a03 서버에서 nvidia-smi -q 실행으로 메모리 오류 확인 | 운영자 |
| 현재 | 원인 분석 | 로그 통합 분석 및 장애 등급 평가 | 장애 분석팀 |

## 5. 구성요소별 점검 결과

### 5.1 하드웨어 점검

|  |  |  |  |
| --- | --- | --- | --- |
| **항목** | **점검 내용** | **결과** | **상세 내용** |
| GPU | nvidia-smi 명령어로 GPU 상태 확인 | **비정상** | * cl-rndcgpu-a02: GPU ID 3 InfoROM 손상 * cl-rndcgpu-a03: GPU ID 7 메모리 오류 다수 발생 |
| GPU 메모리 | ECC 오류 카운터 확인 | **비정상** | cl-rndcgpu-a03 GPU ID 7:   * DRAM Correctable: 1 (휘발성), 2 (누적) * DRAM Uncorrectable: 4 (휘발성), 11 (누적) * Row Remapping: 3건 (Uncorrectable Error) |
| GPU 온도 | 온도 센서 값 확인 | 정상 | 모든 GPU 온도가 정상 범위 내: 27-57 °C (최대 허용: 85 °C) |
| 전원 사용량 | GPU 전력 소비량 확인 | 정상 | 모든 GPU 전력 사용량이 정상 범위 내: 59-422 W (최대 허용: 400 W) |
| PCIe 연결 | lspci 명령어로 PCIe 링크 상태 확인 | 정상 | 모든 GPU PCIe Gen4 x16 링크 속도로 정상 연결 |
| GPU NVLink | nvidia-smi nvlink 명령어로 상태 확인 | 정상 | 모든 NVLink 연결이 정상 작동 중 (25 GB/s) |

### 5.2 운영체제 점검

|  |  |  |  |
| --- | --- | --- | --- |
| **항목** | **점검 내용** | **결과** | **상세 내용** |
| NVIDIA 드라이버 | 드라이버 버전 확인 | 정상 | 버전: 470.103.01 (CUDA 11.4 지원) |
| 시스템 로드 | GPU 사용률 확인 | 정상 | 부하가 있는 GPU는 75-100% 사용률, 유휴 GPU는 0% 정상적 패턴 |
| 시스템 프로세스 | GPU 관련 프로세스 확인 | 정상 | * cl-rndcgpu-a02: python3 프로세스가 GPU 사용 중 * cl-rndcgpu-a03: ray::PredictionWorker.generate 프로세스가 GPU 사용 중 |

### 5.3 애플리케이션 점검

|  |  |  |  |
| --- | --- | --- | --- |
| **항목** | **점검 내용** | **결과** | **상세 내용** |
| ML/AI 워크로드 | 실행 중인 ML 프로세스 확인 | 정상 | 워크로드가 정상적으로 GPU 리소스 사용 중 |

## 6. 장애 원인 분석

### 6.1 장애 현상

두 대의 GPU 서버에서 서로 다른 유형의 하드웨어 장애가 발견되었습니다:

1. **cl-rndcgpu-a02 서버**: GPU ID 3 (BusID: 00000000:4C:00.0)에서 InfoROM 손상 경고가 발생했습니다.

WARNING: infoROM is corrupted at gpu 0000:4C:00.0

이 경고는 해당 GPU의 메타데이터 저장소가 손상되었음을 나타내지만, 현재 해당 GPU는 계속 작동 중입니다.

1. **cl-rndcgpu-a03 서버**: GPU ID 7 (BusID: 00000000:CB:00.0)에서 다수의 GPU 메모리 오류가 발생했습니다.

ECC Errors Volatile SRAM Correctable : 0 SRAM Uncorrectable : 0 DRAM Correctable : 1 DRAM Uncorrectable : 4 Aggregate SRAM Correctable : 0 SRAM Uncorrectable : 0 DRAM Correctable : 2 DRAM Uncorrectable : 11 Remapped Rows Correctable Error : 0 Uncorrectable Error : 3 Pending : Yes

특히 Uncorrectable Error(수정 불가능한 오류)가 다수 발생했고, 이로 인한 Row Remapping이 진행 중인 상태입니다.

### 6.2 근본 원인 분석

#### cl-rndcgpu-a02의 InfoROM 손상

InfoROM은 GPU의 메타데이터와 설정을 저장하는 비휘발성 메모리 영역으로, 손상 원인으로는 다음과 같은 가능성이 있습니다:

* GPU 펌웨어 업데이트 중 발생한 오류나 중단
* 갑작스러운 전원 중단이나 전력 불안정 상태
* NVIDIA 드라이버와의 호환성 문제
* 물리적 메모리 손상

이 문제는 현재 시점에서 즉각적인 성능 저하를 유발하지는 않지만, GPU 설정이나 메타데이터에 영향을 주어 장기적으로 안정성 문제를 일으킬 수 있습니다.

#### cl-rndcgpu-a03의 GPU 메모리 오류

NVIDIA A100 GPU의 HBM2 메모리에서 발생한 오류는 심각한 하드웨어 결함을 나타냅니다:

* Uncorrectable DRAM 오류 4건(휘발성)과 11건(누적)은 NVIDIA의 RMA(제품 교체) 기준을 초과하는 수치입니다.
* Row Remapping이 3건 발생했고 추가적인 Remapping이 대기 중(Pending: Yes)인 상태는 메모리 오류가 지속적으로 발생하고 있음을 의미합니다.
* 이는 HBM2 메모리 칩의 물리적 결함이나 심각한 열화를 나타내는 징후입니다.

이러한 유형의 메모리 오류는 시간이 지날수록 악화되며, 결국 계산 결과 오류, GPU 충돌, 또는 전체 시스템 불안정성을 유발할 수 있습니다.

### 6.3 영향 요인

다음 요인들이 현재 장애 상황에 영향을 미쳤을 가능성이 있습니다:

* 지속적인 높은 부하: 일부 GPU에서 95-100%의 높은 사용률이 관측되며, 이는 장기간 지속될 경우 열 스트레스를 유발할 수 있습니다.
* 전력 관리: 일부 GPU에서 전력 한계(400W)를 초과하는 사용량(최대 422.80W)이 관측되었습니다.
* 메모리 사용 패턴: 사용 중인 GPU는 대부분의 메모리를 사용하고 있으며(49GB-76GB), 이는 높은 메모리 대역폭과 열 발생을 의미합니다.
* GPU 드라이버 버전: 470.103.01은 최신 버전이 아니며, 일부 버그가 수정되지 않았을 가능성이 있습니다.

## 7. 조치 사항

### 7.1 임시 조치

현재 시점에서 서비스 연속성을 보장하기 위한 임시 조치 사항은 다음과 같습니다:

#### cl-rndcgpu-a02 (InfoROM 손상)

1. 현재 워크로드를 점진적으로 다른 GPU로 마이그레이션합니다.
2. 해당 GPU에 대한 모니터링을 강화하고 추가적인 오류 발생 여부를 관찰합니다.
3. 즉각적인 시스템 재부팅이나 드라이버 재설치는 피하고, 계획된 유지보수 시간을 확보합니다.

#### cl-rndcgpu-a03 (메모리 오류)

1. 해당 GPU(ID 7)의 워크로드를 즉시 다른 GPU로 마이그레이션합니다.
2. nvidia-smi -c 3 -i 7 명령으로 해당 GPU를 계산 모드에서 제외합니다.
3. 서버의 온도와 환경 조건을 점검하고 최적의 냉각 상태를 확보합니다.

### 7.2 영구 조치

#### cl-rndcgpu-a02 (InfoROM 손상)

1. 계획된 유지보수 시간에 다음 단계를 수행합니다:
   * NVIDIA 드라이버 재설치
   * nvidia-smi --gpu-reset 명령으로 GPU 리셋
   * 필요시 펌웨어 업데이트
2. 조치 후에도 문제가 지속되면 NVIDIA 기술지원에 문의하여 추가 지원을 요청합니다.

#### cl-rndcgpu-a03 (메모리 오류)

1. 해당 GPU(ID 7)를 교체합니다. NVIDIA의 RMA 기준(Uncorrectable 오류 발생 및 Row Remapping 다수)을 충족하는 명확한 하드웨어 결함입니다.
2. 교체 전까지는 해당 GPU를 비활성화하여 추가적인 시스템 불안정성을 방지합니다.
3. 교체 후 최소 48시간 동안 새 GPU에 대한 집중 모니터링을 시행합니다.

### 7.3 검증 방법

조치 후 아래 방법으로 문제 해결 여부를 검증합니다:

1. nvidia-smi 명령으로 경고 메시지 사라짐 확인
2. nvidia-smi -q 명령으로 ECC 오류 카운터 초기화 및 추가 오류 없음 확인
3. GPU 스트레스 테스트(CUDA 샘플 appllication)를 실행하여 안정성 검증
4. 실제 워크로드 재배포 후 24시간 이상 모니터링

### 7.4 검증 결과

아직 조치가 시행되지 않아 검증 결과가 없습니다. 조치 시행 후 이 섹션을 업데이트할 예정입니다.

## 8. 향후 조치 계획

### 8.1 단기 조치 계획

|  |  |  |  |
| --- | --- | --- | --- |
| **조치 내용** | **담당자** | **예정일** | **우선순위** |
| cl-rndcgpu-a03 GPU ID 7 교체 | 박지원 (GPU마스터) | 2024-08-12 | High |
| cl-rndcgpu-a02 드라이버 재설치 및 GPU 리셋 | 박지원 (GPU마스터) | 2024-08-12 | Medium |
| 모든 GPU 서버 ECC 오류 모니터링 강화 | 정민우 (로그헌터) | 2024-08-13 | Medium |
| 유사 모델 GPU 서버 전수 점검 | 이동훈 (서버왕) | 2024-08-14 | Medium |

### 8.2 장기 조치 계획

|  |  |  |  |
| --- | --- | --- | --- |
| **조치 내용** | **담당자** | **예정일** | **우선순위** |
| NVIDIA 드라이버 최신 버전 업그레이드 | 최서연 (펭귄맘) | 2024-08-20 | Medium |
| GPU 서버 온도 모니터링 시스템 개선 | 한지성 (DC마법사) | 2024-08-25 | Medium |
| GPU 사용률 밸런싱 정책 개발 | 박지원 (GPU마스터) | 2024-09-10 | Low |
| GPU 하드웨어 오류 자동 감지 및 알림 시스템 구축 | 정민우 (로그헌터) | 2024-09-15 | Medium |

## 9. 재발 방지 대책

### 9.1 모니터링 강화

* 모든 GPU 서버에 대한 ECC 오류 카운터 정기 점검 (일 1회)
* InfoROM 상태 주기적 확인 스크립트 구현
* GPU 온도 및 전력 소비 임계치 알림 설정 (온도 80°C 이상, 전력 380W 이상 시 알림)
* Row Remapping 발생 시 즉시 알림 설정

### 9.2 시스템 개선

* NVIDIA 드라이버 및 펌웨어 정기 업데이트 일정 수립 (분기 1회)
* GPU 서버 냉각 시스템 최적화 (공기 흐름 개선, 팬 속도 조정)
* 전력 공급 안정성 향상을 위한 UPS 점검 및 필요시 용량 증설
* 워크로드 자동 분산 시스템 구현으로 GPU 부하 균등화

### 9.3 프로세스 개선

* GPU 하드웨어 장애 대응 프로세스 정립 및 문서화
* GPU 교체 및 RMA 처리 절차 간소화
* 장애 발생 시 신속한 워크로드 마이그레이션 자동화 도구 개발
* 정기적인 GPU 상태 점검 일정 수립 (월 1회)

### 9.4 교육 및 훈련

* 운영 담당자 대상 NVIDIA GPU 아키텍처 및 장애 유형 교육
* GPU 장애 시나리오별 대응 훈련 실시
* NVIDIA 기술 지원 활용 방법 및 RMA 프로세스 교육
* GPU 모니터링 도구 사용법 및 로그 분석 방법 교육

## 10. 첨부 자료

### 10.1 로그 파일

* nvidia-smi.cl-rndcgpu-a02.log - GPU 상태 정보 및 InfoROM 손상 경고
* nvidia-smi.cl-rndcgpu-a03.log - GPU 상태 정보
* nvidia-smi-q.cl-rndcgpu-a02.log - 상세 GPU 정보 (cl-rndcgpu-a02)
* nvidia-smi-q.cl-rndcgpu-a03.log - 상세 GPU 정보 및 메모리 오류 정보 (cl-rndcgpu-a03)
* nvidia-smi-nvlink-s.cl-rndcgpu-a02.log - NVLink 상태 정보 (cl-rndcgpu-a02)
* nvidia-smi-nvlink-s.cl-rndcgpu-a03.log - NVLink 상태 정보 (cl-rndcgpu-a03)
* lspci-v-t.cl-rndcgpu-a02.log - PCI 장치 계층 구조 정보 (cl-rndcgpu-a02)
* lspci-v-t.cl-rndcgpu-a03.log - PCI 장치 계층 구조 정보 (cl-rndcgpu-a03)

### 10.2 구성 파일

* 해당 없음

### 10.3 스크린샷

* 해당 없음

### 10.4 참고 문서

* 하드웨어 장애 유형별 장애분석 가이드라인 - Case 3: GPU 서버 장애 분석 가이드라인
* 임시 조치 매뉴얼 - 3.5 GPU 관련 장애
* NVIDIA A100 SXM4 제품 문서
* NVIDIA GPU ECC 오류 관련 기술 문서
* NVIDIA GPU RMA 가이드라인

## 11. 승인

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **구분** | **이름** | **직책** | **서명** | **일자** |
| 작성자 | 정민우 (로그헌터) | 통합 로그분석 전문가 |  | 2024-08-10 |
| 검토자 | 박지원 (GPU마스터) | GPU 전문가 |  | 2024-08-10 |
| 승인자 | 이해경 (갑빠) | PO |  | 2024-08-10 |