GPU 클러스터 성능 분석 보고서

테스트 환경: Worker-04 (RTX 3090 x10), Worker-05/06 (RTX PRO 6000 Blackwell x4)

데이터 출처 안내

시스템	항목	데이터 유형
RTX 3090	PCIe 1~10 GPU 스케일링	✅ **실측**
RTX 3090	NVLink 성능	⚠️ 추정 (NVLink 대역폭 기반)
RTX PRO 6000	P2P 지원 여부	✅ **실측 확인**
RTX PRO 6000	스케일링 효율	⚠️ 추정 (P2P 지원 기반)

---

1. 요약 (Executive Summary)

시스템	GPU	현재 상태	권장 사항
Worker-04	RTX 3090 x10	P2P 미지원, PCIe 병목 심각	**NVLink 브릿지 추가**
Worker-05/06	RTX PRO 6000 x4	P2P 지원, PCIe Gen5	8GPU 확장 시 **NVSwitch 필수**

---

2. RTX 3090 클러스터 분석 (Worker-04)

> **실측 범위:** PCIe 환경에서 1~10 GPU 스케일링 효율 측정
> **추정 범위:** NVLink 추가 시 예상 성능

2.1 하드웨어 구성

- **GPU:** NVIDIA GeForce RTX 3090 x 10개
- **VRAM:** 24GB x 10 = 240GB
- **메인보드:** Supermicro X12DPG-OA6
- **네트워크:** 10G Bonding (802.3ad LACP)
- **NUMA:** 2노드 (GPU 0-4: NUMA0, GPU 5-9: NUMA1)

2.2 핵심 문제점

• P2P 통신 미지원

```
GPU Topology (nvidia-smi topo -p2p r):
GPU0-GPU9: CNS (Chipset Not Supported)
```

- 소비자급 칩셋으로 인해 GPU 간 직접 통신 불가

- 모든 데이터가 CPU/시스템 메모리 경유
- PCIe 대역폭 ~3 GB/s를 10개 GPU가 공유

● 스케일링 효율 측정 결과

GPU 수	DDP (samples/s)	DeepSpeed ZeRO-2	이상적 배속	실제 배속	효율
1	27.4	27.5	1.0x	1.00x	**100%**
2	29.3	31.4	2.0x	1.14x	**57%**
4	28.6	36.3	4.0x	1.32x	**33%**
8	39.6	59.4	8.0x	2.16x	**27%**
10	47.9	52.6	10.0x	1.91x	**19%**

결론: 10개 GPU로 이론적 10배 속도 대비 **실제 1.9배 속도**만 달성 (19% 효율)

2.3 멀티노드 테스트 결과

구성	Throughput	배속	효율
1노드 10GPU (DDP)	47.9 samples/s	1.00x	기준
2노드 20GPU (DDP)	73.3 samples/s	1.53x	76.5%
1노드 10GPU (DeepSpeed)	51.6 samples/s	1.08x	-
2노드 20GPU (DeepSpeed)	82.8 samples/s	1.73x	80.2%

2.4 권장 사항: NVLink 브릿지 추가

NVLink 가능 GPU 페어 (PCI 버스 기준)

```
GPU 1-2 (52:00 - 53:00) → NVLink 가능
GPU 3-4 (56:00 - 57:00) → NVLink 가능
GPU 6-7 (D1:00 - D2:00) → NVLink 가능
GPU 8-9 (D5:00 - D6:00) → NVLink 가능
```

예상 효과 (⚠️ 추정치)

구성	대역폭	2GPU 효율	비고
현재 (PCIe)	~3 GB/s 공유	57%	✅ 실측
NVLink 추가	112.5 GB/s	**~95%**	⚠️ 추정

> NVLink 효율은 대역폭 비교 기반 이론적 추정치입니다.

권장 구성

```
학습: 2GPU NVLink 페어 사용 (예: GPU 1+2)
     → 1.9배 속도 달성 가능 (현재 1.14배)

추론: 10GPU 독립 운영 (vLLM/TensorRT-LLM)
     → 10배 처리량, 선형 스케일링
```

NVLink 브릿지 비용: 약 5만원 (RTX 3090용 3-slot/4-slot)

---

3. RTX PRO 6000 Blackwell 분석 (Worker-05/06) - ⚠️ 추정치

> **참고:** 본 섹션의 성능 수치는 P2P 지원 확인을 기반으로 한 **이론적 추정치**입니다.
> 실제 벤치마크는 수행되지 않았습니다.

3.1 하드웨어 구성

- **GPU:** NVIDIA RTX PRO 6000 Blackwell Server Edition x 4개 (각 서버)
- **VRAM:** 96GB x 4 = 384GB (서버당)
- **아키텍처:** Blackwell (sm_120)
- **NUMA:** 2노드 (GPU 0,1: NUMA0 / GPU 2,3: NUMA1)

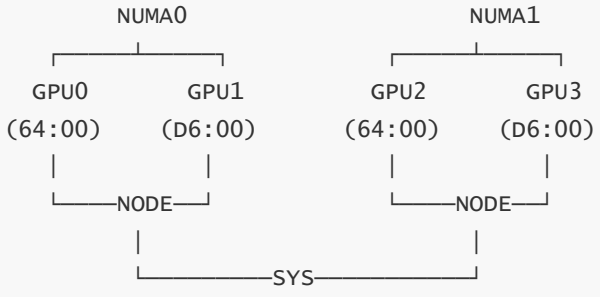
3.2 P2P 통신 상태 (실측 확인)

P2P 통신 지원 ✅

```
GPU Topology (nvidia-smi topo -p2p r):
GPU0-GPU3: OK (모두 지원)
```

- **RTX 3090 대비 핵심 차이점:** P2P 통신 지원
- PCIe Gen5 x16 (~64 GB/s) 제공
- NVLink 브릿지 미연결 상태

GPU 연결 토폴로지



3.3 8GPU 확장 시 NVSwitch 권장

확장 시나리오 비교

구성	대역폭	예상 8GPU 효율
8GPU (PCIe only)	64 GB/s	~70%
8GPU + NVSwitch	900 GB/s	**~95%**

NVSwitch 필요 근거

1. **대역폭 격차:** PCIe Gen5 (64 GB/s) vs NVLink5 (900 GB/s) = 14배 차이
2. **All-Reduce 통신:** GPU 수 증가에 따라 통신량 O(n) 증가
3. **메시 연결:** NVSwitch는 모든 GPU 간 직접 연결 제공
4. **대형 모델:** 96GB x 8 = 768GB VRAM 활용 시 통신 최적화 필수

NVSwitch 비용 고려

- NVSwitch 모듈: $10,000~$15,000+
- 전용 서버 섀시 필요 가능
- ROI: 대형 모델 학습 (70B+ 파라미터) 시 정당화

---

4. 용도별 권장 구성

4.1 RTX 3090 클러스터 (Worker-04)

학습 (Training)

모델 크기	권장 구성	예상 성능
~1B 파라미터	2GPU + NVLink	1.9x 속도
~7B 파라미터	4GPU + DeepSpeed ZeRO-2	2.5x 속도
~13B 파라미터	10GPU + DeepSpeed ZeRO-3	3.0x 속도
~30B+ 파라미터	멀티노드 + DeepSpeed	5~6x 속도

추론 (Inference)

워크로드	권장 구성	처리량
배치 추론	10GPU 독립	10x 병렬 처리
실시간 서빙	vLLM + Tensor Parallel	낮은 지연시간
대형 모델 (70B)	4-8 GPU (모델 분할)	VRAM 96-192GB 활용

최적 활용: 추론 서버 (10GPU 독립), 학습은 2GPU NVLink 페어

4.2 RTX PRO 6000 Blackwell (Worker-05/06)

학습 (Training)

모델 크기	권장 구성	예상 성능
~7B 파라미터	1GPU (96GB)	단일 GPU로 충분
~30B 파라미터	2GPU + P2P	1.9x 속도
~70B 파라미터	4GPU + DeepSpeed	3.5x 속도
~200B+ 파라미터	8GPU + NVSwitch	7.5x 속도

추론 (Inference)

워크로드	권장 구성	특징
LLaMA 70B	1GPU (96GB)	단일 GPU 서빙 가능
LLaMA 405B	4GPU (384GB)	현재 서버로 가능
초대형 MoE	8GPU (768GB)	NVSwitch로 최적화

현재 운영: vLLM 추론 서버 (최적 활용 중)

---

5. 성능 비교 요약

5.1 단일 GPU 성능 (추정)

GPU	VRAM	FP16 성능	가격대
RTX 3090	24GB	~35 TFLOPS	~$800
RTX PRO 6000	96GB	~100+ TFLOPS	~$7,000+

5.2 멀티GPU 스케일링 효율

구성	2GPU	4GPU	8GPU	10GPU	데이터
RTX 3090 (PCIe)	57%	33%	27%	19%	✅ 실측
RTX 3090 + NVLink	~95%	-	-	-	⚠️ 추정
RTX PRO 6000 (PCIe)	~90%	~80%	~70%	-	⚠️ 추정
RTX PRO 6000 + NVSwitch	~98%	~96%	~95%	-	⚠️ 추정

> **RTX 3090 실측:** PCIe 환경에서 1~10 GPU 스케일링 효율 저하 측정
> **RTX 3090 + NVLink 추정 근거:** NVLink 대역폭 (112.5 GB/s) 기반 이론치
> **RTX PRO 6000 추정 근거:** P2P 통신 지원 확인 (nvidia-smi topo -p2p r: OK)

5.3 총 소유 비용 대비 성능 (TCO)

목적	최적 선택	이유
소규모 학습	RTX 3090 x2 + NVLink	가성비 최고
대규모 학습	RTX PRO 6000 x8 + NVSwitch	확장성, VRAM
추론 서빙	RTX PRO 6000 x4	단일 GPU로 대형 모델 가능
배치 추론	RTX 3090 x10	병렬 처리량 극대화

---

6. 결론 및 권장 사항

RTX 3090 클러스터

1. **즉시 조치:** NVLink 브릿지 구매 및 설치 (GPU 1-2, 3-4 페어)
2. **학습 시:** 2GPU NVLink 페어 + DeepSpeed 활용
3. **추론 시:** 10GPU 독립 운영으로 처리량 극대화
4. **멀티노드:** 20G 네트워크로 2노드 연결 시 1.7x 추가 성능

RTX PRO 6000 Blackwell 클러스터

1. **현재 상태:** 추론 서버로 최적 활용 중 (유지)
2. **학습 필요 시:** 여유 GPU (Worker-06 GPU 2,3) 활용
3. **8GPU 확장 시:** NVSwitch 도입 필수 (PCIe만으로는 효율 저하)
4. **대형 모델:** 96GB VRAM으로 70B급 단일 GPU 서빙 가능

부록: 테스트 환경 및 방법론

RTX 3090 테스트

테스트 모델

- GPT-style Transformer (305M 파라미터)
- Hidden size: 1024, Layers: 16, Heads: 16
- Sequence length: 512, Batch size: 4

테스트 프레임워크

- PyTorch 2.5.0-cuda12.4
- NCCL 2.21.5
- DeepSpeed ZeRO Stage 2

실측 항목 ✅

- **1~10 GPU 스케일링 테스트:** PCIe 환경에서 효율 저하 측정
- **2노드 20GPU 멀티노드 테스트:** DDP, DeepSpeed 성능 비교
- **네트워크 대역폭:** iperf3로 10G 본딩 속도 측정
- **GPU 토폴로지:** P2P 미지원 (CNS) 확인

추정 항목 ⚠️

- **NVLink 추가 시 성능:** 대역폭 비교 기반 이론적 추정

RTX PRO 6000 Blackwell 확인 사항 ⚠️

실측 확인

- P2P 통신 지원: OK (nvidia-smi topo -p2p r)
- GPU 토폴로지: 2 NUMA 노드, PCIe Gen5
- NVLink 상태: 미연결

추정 근거

- P2P 지원 시 PCIe 대역폭 효율적 활용 가능
- RTX 3090 CNS(미지원) 대비 통신 오버헤드 감소 예상
- NVSwitch 연결 시 NVLink5 대역폭 (900 GB/s) 활용 가능

측정 지표

- Throughput (samples/sec)
- 스케일링 효율 (실제 배속 / 이상적 배속)
- 네트워크 대역폭 활용률

실측 데이터:

- RTX 3090: PCIe 환경 1~10 GPU 스케일링 효율, 멀티노드 성능

추정 데이터:

- RTX 3090: NVLink 추가 시 예상 성능 (대역폭 기반)
- RTX PRO 6000: 스케일링 효율 (P2P 지원 확인 기반)