

HAMMERSPACE

GPU 서버 내 미활용 Tier 0 스토리지 활성화

이미 보유한 최고 성능의 NVMe 스토리로 AI 워크로드를 가속하세요.

By Floyd Christofferson

V. 1.1 - June 2025

Overview

AI 워크로드의 성능과 규모 문제를 해결하기 위한 최적의 스토리지 솔루션으로 자리매김하려는 여러 벤더 간 치열한 경쟁이 벌어지고 있다.

이는 단순히 어떤 스토리지 벤더가 GPU 서버로의 데이터 처리 속도를 가장 빠르게 가속화하는가의 문제만이 아니다. AI 모델이 점점 더 복잡해지고, 추론 세션이 훨씬 긴 컨텍스트를 처리해야 할 때, 워크로드가 GPU 서버 메모리와 시스템 메인 메모리를 자주 포화시키는 문제가 발생한다. 그 결과 외부 스토리지로의 데이터 전송시 I/O 대기 시간이 증가하고, 외부 스토리지로 향하는 네트워크 트래픽 증가가 GPU 컴퓨팅 워크로드에 직접적인 영향을 미치는 심각한 문제를 일으키며, 온프레미스와 클라우드 데이터 센터 모두에서 비용 증가로 이어진다.

문제의 핵심은 다음과 같다. 최신 스토리지 성능 향상을 홍보하거나, 지능형 캐싱, 토큰 최적화 등 다양한 우회 기능을 위해 전용 클라이언트 소프트웨어를 GPU 서버에 설치하라고 주장하는 벤더가 있더라도, 근본적인 문제는 여전히 해결되지 않는다. 외부 스토리지와 GPU 서버 간 데이터 경로는 스토리지 종류에 관계없이 네트워크 대역폭에 의해 병목 이 발생하기 때문이다.

어떤 데이터 하위 집합을 GPU로 우선 전송할지 더 지능적으로 결정하는 것은 중요하며 도움이 된다. 그러나 이러한 방법을 사용하더라도 점점 증가하는 데이터 량은 여전히 외부 스토리지에서 GPU로 공급되기 위해 네트워크 병목을 거쳐야 하며, 이것이 바로 문제의 핵심이다.

좋은 소식은, 이러한 문제를 해결할 훨씬 더 나은 솔루션이 이미 존재한다는 점이다. 그 해답은 온프레미스와 클라우드의 모든 GPU 서버에 표준 Linux 배포판과 함께 기본적으로 탑재되어 있지만 잘 활용되지 않는 로컬 NVMe 스토리지다.

이 문서는 조직의 규모와 관계없이 이 유휴 용량을 지속적으로 보호되며 초고속의 Tier 0 공유 스토리지 환경으로 활성화하여 GPU 워크로드를 획기적으로 가속하고 GPU 활용도를 개선함으로써, 실질적인 혜택과 비용 절감 방안을 얻을 수 있는 방법을 개괄적으로 설명한다.

Hammerspace Tier 0는 AI 워크로드에 사용 가능한 스토리지 유형 중 단연 가장 뛰어난 성능을 제공하며, 동일한 수의 클라이언트 연결시 Lustre 기반 솔루션 보다 16.5배 더 많은 GPU를 지원할 수 있음이 벤치마크 결과를 통해 입증되었다. 이 Tier 0 용량이 여러 GPU 서버에 걸쳐 집계되면 온프레미스와 클라우드 환경 모두에서 GPU 투자 대비 훨씬 높은 성능을 훨씬 낮은 비용으로 제공한다.

Tier 0가 활성화되면 현재 AI 워크로드를 위해 이용 가능한 스토리지 유형 중에 단연 최고의 성능을 제공한다. 또한 이 용량을 클러스터 내 다수의 GPU 서버에 걸쳐 집계하면, 온프레미스와 클라우드 모두에서 훨씬 낮은 비용으로 GPU 투자대비 훨씬 더 큰 성과를 얻을 수 있다.

더 중요한 점은, 이 혁신이 온프레미스 또는 클라우드 기반 GPU(또는 CPU/ARM) 서버 모두에서 작동하며, 이미 비용을 지불한 로컬 NVMe 용량을 활성화함으로써 구현된다는 점이다.

다시 한번 강조한다. 대부분 활용되고 있지 않는 **이 초고속 로컬 NVMe 용량**은 이미 GPU 서버 가격에 포함되어 있기 때문에 추가 비용이 들지 않는 **매몰비용(sunk cost)**이다. 또한 이 용량을 활용하더라도 추가 전력 사용량이 거의 없다는 점은 가용 전력 한계에 근접한 데이터 센터에 매우 큰 이점이 된다.

Hammerspace를 사용하면, 이 로컬 NVMe를 Tier 0 공유 스토리지로 활성화하는 과정은 모든 표준 Linux 배포판에 포함된 Linux 커널의 기능을 활용하여 이루어진다. 이는 서버에 별도의 클라이언트 소프트웨어를 설치하거나 추가적인 변경을 할 필요가 없다는 의미이다. 아래에서 설명하겠지만, Tier 0는 소수의 GPU 서버로 구성된 작은 클러스터부터 수십만 개의 GPU를 보유한 하이퍼스케일 환경에 이르기까지 모든 규모의 환경에서 즉각적인 가치를 제공하는 혁신적 솔루션이다.

또한 이 로컬 Tier 0 용량을 활용하면, 초고성능 네트워킹 인프라나 추가 외부 고성능 스토리지 시스템을 도입하는데 드는 비용을 줄이거나 아예 없앨 수도 있다. 이러한 인프라는 GPU 활용도를 최대화하기 위해 네트워크의 병목을 줄이려는 비용 부담이 매우 큰 방식들인데, Tier 0는 이러한 문제를 근본적으로 해결해 준다.

그 결과, Hammerspace는 고객이 기존 인프라를 AI-ready로 전환하는 작업을 획기적으로 단순화하며, 기존 인프라 내에서도 AI 또는 기타 고성능 데이터 프로젝트를 시작할 때 발생하는 비용과 복잡성을 크게 줄여준다.

이 초고속 로컬 NVMe 용량은 대부분 활용되지 않고 있지만, 온프레미스와 클라우드 기반 GPU 서버 가격에 이미 포함되어 있기 때문에 추가 비용이 들지 않는 매몰비용(sunk cost)이다.

GPU 서버 내 로컬 NVMe 스토리지를 Tier 0로 활성화

GPU 서버 안에는 GPU 컴퓨팅 워크로드를 획기적으로 변화시킬 수 있는 미활용 자원이 숨어 있다. 온프레미스든 클라우드든 모든 GPU 서버는 표준 Linux 배포판을 사용하며, 여기에는 Hammerspace가 지난 8년 이상 오픈 소스 커뮤니티에 기여해 온 NFS 및 pNFS v4.2 성능 기능이 이미 포함되어 있다. 또한 이 GPU 서버에는 최소 2개, 일반적으로 8개, 일부 벤더는 최대 16개의 고성능 로컬 NVMe 드라이브가 포함되어 있다.

서버 제조사 마다 NVMe 드라이브의 수량과 용량은 상이하다. 일부 서버는 드라이브 슬롯을 부분적으로만 채워 사용할 수 있도록 허용하는 반면, 클라우드 기반 GPU 서버를 포함한 다른 서버들은 기본적으로 드라이브를 모두 탑재한 상태로 제공된다.

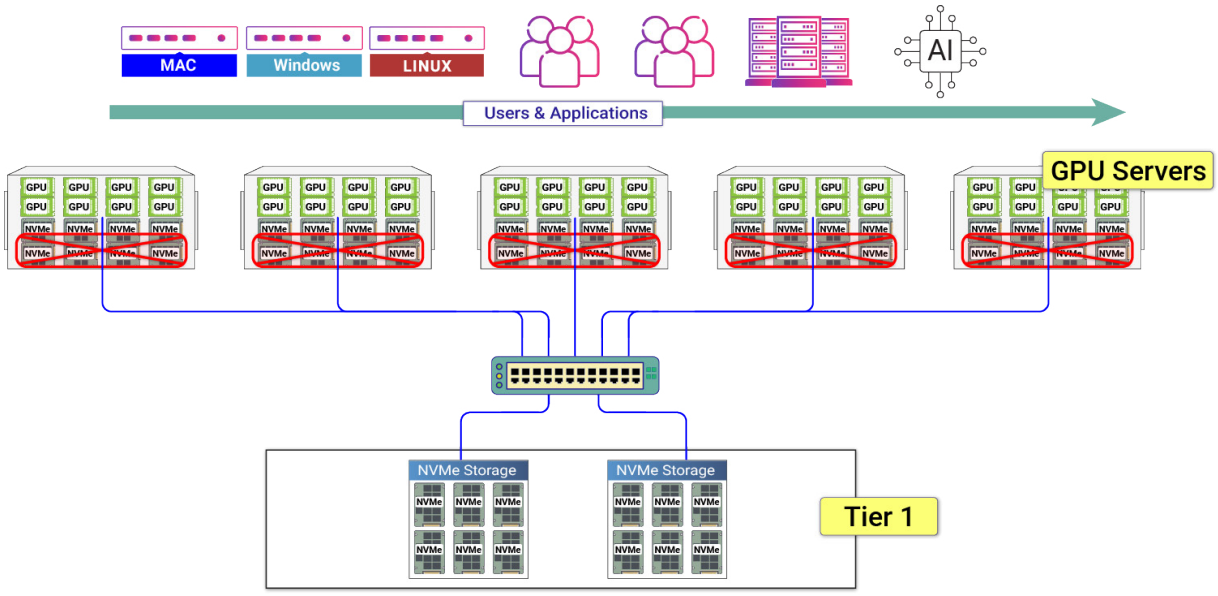


그림 1: GPU 서버 내 유휴 NVMe 용량은 GPU 클러스터에서 대부분 활용되지 않는다. GPU가 필요한 데이터는 외부 스토리지와 GPU 간의 네트워크를 통해 이동해야 하므로 워크로드 속도가 느려지고 GPU 활용률이 떨어진다.

문제는 일부 애플리케이션이 단일 서버나 아주 작은 학습 데이터셋에서는 이 로컬 용량을 고성능 캐시나 임시 스크래치 공간으로 부분적으로 활용할 수 있지만, 두 대 이상의 GPU 서버로 구성된 클러스터나 더 큰 데이터셋에서는 이 NVMe 스토리지가 고립되고 사일로화된 용량으로 남아 거의 활용되지 못하는 경우가 많다는 점이다. 그 이유는 다음과 같다:

- **용량의 사일로화(Capacity is siloed):** 로컬 NVMe 드라이브는 각 서버 내에 고립되어 있다. GPU 서버들은 클러스터 전체에서 이 용량을 공유할 수 없다.
- **공유 접근성 부족(Lack of shared access):** GPU 서버 간 스토리지가 공유되지 않는다면, 로컬 스토리지의 모든 데이터는 수동으로 복사해야 하며, 이로 인해 오버헤드, 지연, 운영 복잡성이 증가한다.
- **제한적인 데이터 보호(Limited data protection):** GPU 서버의 NVMe 스토리지는 클러스터 전반에 걸친 데이터 중복 기능(Data Redundancy)이 없기 때문에, 서버 장애나 기타 문제 발생 시 데이터 손실 또는 중단 위험이 매우 높다.

대용량 SSD 드라이브와 더 큰 GPU 서버 클러스터가 판도를 바꾼다

SSD 드라이브의 밀도가 증가하고 클러스터 내 GPU 서버 수가 늘어나면서 상황이 급격히 변화하고 있다. 2025년 3월 기준, GPU 서버 벤더들은 일반적으로 서버 당 30TB에서 122TB의 로컬 NVMe 용량을 탑재해 출하한다. 이는 8개 또는 16개 드라이브 구성에서 드라이브 당 약 4TB~8TB SSD 밀도를 기반으로 한다.

그러나 드라이브 밀도는 빠르게 증가하고 있으며, 엔터프라이즈급 SSD에서는 30TB와 60TB 드라이브가 일반적으로 사용되고 있다. 제조사 로드맵에 따르면 2025년 4분기에는 122TB 드라이브가 출시될 것으로 예상된다.

GPU로 전달되는 데이터의 GB당 초당 성능 대비 가격을 비교하면, 외부 스토리지는 Tier 0보다 약 10배 더 비싸다.

이는 오늘날 서버에 제공되는 상대적으로 작은 드라이브 용량만으로도, 8대의 서버 클러스터가 거의 1페타바이트(PB)에 달하는 초고속 NVMe 스토리지를 보유할 수 있음을 의미한다. 또한 구성 가능한 GPU 서버의 경우, 동일한 8대 서버 클러스터를 30TB 드라이브로 업그레이드하면 사용 가능한 용량이 3.8PB로 증가한다.

Tier 0 혁신이 게임 체인저인 이유는, 이러한 극한의 성능 얻기 위해 추가적인 고성능 외장 스토리지 시스템, 스위치 인프라, 랙 공간, 전력 오버헤드 등을 도입할 필요 없이, 기존 GPU 서버와 전력 예산 내에서 그대로 활용할 수 있기 때문이다. TB당 비용 기준으로 보면, GPU 서버의 내부 NVMe 스토리지를 최대 구성으로 장착하는 비용은 외장 스토리지 시스템 도입 비용의 절반 수준에 불과하다. 그러나 진짜 이야기는 가격 대비 성능 비교에서 드러난다. GPU로 데이터를 전달할 때의 성능 비용을 비교하면, 외부 Tier 1 NVMe 스토리지는 200Gb/s 네트워크를 기준으로 내부 Tier 0 스토리지보다 GB당 총 비용이 약 10배 더 높다. Tier 1 스토리지 총 비용 계산에는 드라이브 및 인클로저, 전력/냉각, 랙 공간, 200Gb/s 네트워킹 등 모든 비용이 포함된다. 네트워크를 더 비싼 800Gb/s 스위치로 업그레이드하면 Tier 0의 가격/성능 우위는 더욱 커진다[1].

아래에서 자세히 설명하겠지만, 이렇게 방대한 규모의 고성능 스토리지를 활성화하고 AI 클러스터에 배치되는 GPU 서버 수가 증가함에 따라, 조직들은 IT 아키텍처와 외부 고성능 스토리지 및 네트워킹 인프라에 대한 투자 방식을 재고하게 되었다.

1. 엔터프라이즈급 SSD는 용량, 구매 물량, 제조사, 드라이브 유형에 따라 가격 차이가 매우 크다. 본 문서에서는 12.8TB 드라이브를 TB당 80달러로 계산하였으며, 이는 실제 시장에서 흔히 볼 수 있는 평균 가격으로 추정했다.

문제는 네트워크 병목이다

여기서 핵심은, 아무리 빠른 외부 스토리지라도 서버와 스토리지 간 네트워크 속도 이상으로는 빠를 수 없다는 점이다. 이는 새로운 사실이 아니며, 모든 IT 환경에서 항상 그래 왔다.

그러나 이러한 상황이 AI 학습 및 추론처럼 GPU 중심 워크로드에 미치는 영향은 기존 엔터프라이즈 IT 사용 사례보다 훨씬 더 심각하다. GPU를 완전히 활용하려면 더 빠른 네트워크가 필요하고, 그만큼 네트워크 인프라 비용이 더욱 커진다. 게다가 GPU 클러스터가 커지면서 데이터량이 폭발적으로 증가하고, 더 복잡한 모델과 더 긴 컨텍스트 길이로 인해 GPU와 서버의 메모리가 포화되면서 발생하는 추가 트래픽까지 더해지면, 병목 문제는 더욱 악화될 수 밖에 없다.

고가의 400Gb/s InfiniBand 스위치와 가장 빠른 외부 스토리지를 사용하더라도, 네트워크를 통해 데이터를 쓰는 것은 8개의 드라이브가 장착된 GPU 서버의 로컬 NVMe Tier 0 스토리지에 직접 쓰는 것보다 약 2.5배 더 오래 걸린다. 이러한 지연은 GPU가 데이터를 기다리느라 소중한 GPU 사이클이 낭비될 수 있음을 의미한다. 그리고 이는 캐싱이나 기타 토큰 최적화 기술을 사용하더라도 마찬가지다.

네트워크가 100Gb/s라면, 동일한 데이터를 외부 Tier 1 스토리지에 쓰는 데 걸리는 시간은 서버 내부의 Tier 0 스토리지에 기록하는 것 보다 10배 더 오래 걸린다.

이 수치에는 다양한 데이터 보호 방식이나 호스트 어피니티(Host affinitization) 같은 요소들이 내재되어 있어, 데이터 흐름을 최적화할 수 있다. 예를 들어 자주 접근되는 데이터를 해당 GPU와 최대한 가까운 위치에 배치하는 방식 등이 있다. 그러나 핵심은 여전히 동일하다. 네트워크를 통해 이동해야 하는 데이터는 서버 내에 이미 내장된 로컬 NVMe를 활용하는 것보다 항상 더 느릴 수밖에 없다.

1. Tier 0 vs. Tier 1 - On Premises (90% Network Efficiency)

(Assuming Checkpoint Size of 500GB)

	Storage Type	Optimal Throughput	Effective Throughput	Time to Write 500GB
Tier 0	**Local NVMe (8-drive GPU server)**	**112 GB/s**	**112 GB/s ✅**	**~4.5 sec ✅**
Tier 1	**100 Gb/s Network**	**12.5 GB/s**	**11.25 GB/s**	**~44 sec**
	200 Gb/s Network	**25 GB/s**	**22.5 GB/s**	**~22 sec**
	400 Gb/s Network	**50 GB/s**	**45 GB/s**	**~11.1 sec**

Note: ***External Tier 1 storage throughput adjusted assuming 90% practical network efficiency.***

클라우드 환경에서는 그 차이가 더욱 극적이다. AWS에서 8-GPU H100 인스턴스가 500GB 체크포인트를 로컬 NVMe에 기록하는데 약 4.5초가 걸린다. 같은 작업을 AWS의 고성능 EBS(io2) 스토리지에 기록하면 약 139초(약 2.3분)가 걸리며, 이는 31배 이상 더 오래 걸리는 것이다. 더 저렴한 AWS EBS(gp3) 스토리지에 쓰면 약 9분, 거의 124배 더 오래 걸린다.

Microsoft Azure는 매우 빠른 Ultra 디스크를 제공하며, 속도는 온프레미스 80Gb/s 네트워크와 동등한 속도를 제공하며 클라우드 환경에서는 상당히 빠른 편이다. 하지만 Azure GPU 서버 클러스터 내 Tier 0를 활용하는 것보다 여전히 11배 더 느리다.

2. Tier 0 vs. Tier 1 - Cloud Storage (90% Network Efficiency)

(Assuming Checkpoint Size of 500GB)

	Storage Type	Optimal Throughput	Effective Throughput	Time to Write 500GB
Tier 0	**Local NVMe (8-drive GPU server)**	**112 GB/s**	**112 GB/s ✅**	**~4.5 sec ✅**
Tier 1	**AWS EBS (gp3)**	**1 GB/s**	**0.9 GB/s**	**~555.6 sec (~9.3 min)**
	AWS EBS (io2)	**4 GB/s**	**3.6 GB/s**	**~138.9 sec (~2.3 min)**
	AWS S3 (via high-speed network)	**1 GB/s**	**0.9 GB/s**	**~555.6 sec (~9.3 min)**
	Microsoft Azure Ultra Disk	**10 GB/s**	**9 GB/s**	**~55.6 sec**
	Microsoft Azure Premium SSD v2	**1.2 GB/s**	**1.08 GB/s**	**~462 sec (~7.7 min)**
	Oracle - OCI Block Volume	**1 GB/s**	**0.9 GB/s**	**~555.6 sec (~9.3 min)**

Note: ***External Tier 1 storage throughput adjusted assuming 90% practical network efficiency.***

체크포인트 사용 사례는 특히 중요한 의미를 갖는데, 이는 장시간 실행되는 작업이 노드 장애로 인해 중단되는 것을 방지하기 위해 모델의 현재 상태를 저장해야 하기 때문이다. 문제는 기존 많은 학습 프레임워크에서 체크포인트가 스토리지에 기록되는 동안 GPU가 대기 상태로 멈춰 있어야 한다는 점이다. 클라우드 기반 GPU 워크로드에서는 이러한 대기 시간이 특히 치명적인데, GPU 사용 여부와 상관없이 사용 시간에 대해 비용을 지불해야 하기 때문이다.

최신 학습 프레임워크들은 비동기 체크포인트(asynchronous checkpointing)를 통해 GPU 대기 시간을 최소화하는 방식으로 이 문제를 해결하고 있다. 그러나 Tier 0와 외부 스토리지 간 성능 차이는 체크포인트에만 국한되지 않는다. 앞서 예시에서 보여준 성능 차이는 GPU와 외부 스토리지 간의 모든 데이터 트래픽에도 그대로 영향을 미친다. 그리고 그 데이터량은 500GB 체크포인트보다 훨씬 더 클 수 있다.

분명히 말하자면, 가장 비싼 외부 스토리지와 800Gb/s 네트워크 옵션을 구매하면 온프레미스 환경에서 Tier 0 NVMe 스토리지의 처리량에 근접할 수 있다. 하지만 이는 네트워크 포트당 수천 달러의 비용은 물론, 추가 전력 요구 사항과 기타 오버헤드까지 감수해야 한다는 의미이기도 하다.

하지만 Tier 0를 사용하면, 조직이 이미 GPU 서버 내부에 비용을 지불한 NVMe 용량과 기존 보유 네트워크 인프라만으로 이러한 성능을 달성할 수 있다. 포트당 약 200달러 수준인 기존 100Gb/s 네트워크를 사용하더라도, Tier 0를 활용하면 워크로드가 더 빠르게 처리된다.

수백, 수천 또는 수만 개 GPU가 필요한 대규모 환경에서는, 많은 고가의 네트워크 포트와 외부 스토리지를 추가하는 비용이 엄청나게 증가한다.

간단히 말해, 이미 GPU 서버에 더 높은 성능의 로컬 NVMe 스토리지를 보유하고 있음에도 불구하고 외부 네트워킹과 스토리지에 막대한 비용을 투입하면, AI 프로젝트의 투자수익률(ROI)는 직접적으로 감소하게 된다. 그리고 클라우드 기반 또는 온프레미스/클라우드 하이브리드 환경에서는 그 영향이 더욱 크다.

Tier 0를 사용하면, 특히 더 큰 GPU 클러스터와 더 큰 드라이브 용량 환경에서, 두 가지 장점을 모두 누릴 수 있다. 즉, GPU에 지속적으로 데이터를 공급하고 워크로드를 원활하게 처리할 수 있는 시장 최고 수준의 스토리지 성능을 확보하면서도, 이미 보유한 NVMe 용량과 네트워크 인프라를 그대로 활용할 수 있다는 점이다.

서버 클러스터 내에서 Tier 0 용량을 활성화하는 방법

Tier 0를 고려할 때 가장 먼저 떠오르는 질문은 다음과 같다. CPU 및 GPU 기반 서버 클러스터 내의 로컬 NVMe 스토리지를 활성화하면 이렇게 극적인 성능과 비용 이점을 제공한다면, 왜 모두가 이미 이를 활용하지 않고 있는가? 왜 스토리지 산업은 주로 외부 스토리지 시스템의 성능에만 집중하거나, GPU에 전달할 데이터를 최적화하기 위해 서버에 설치해야 하는 독점 클라이언트 소프트웨어 개발에 몰두하는가?

간단한 답은, 클러스터 전반에서 로컬 NVMe를 초고속 Tier 0 공유 스토리지로 안전하고 효율적으로 활성화하려면, Hammerspace 데이터 플랫폼이 제공하는 고유한 기능이 필요하기 때문이다.

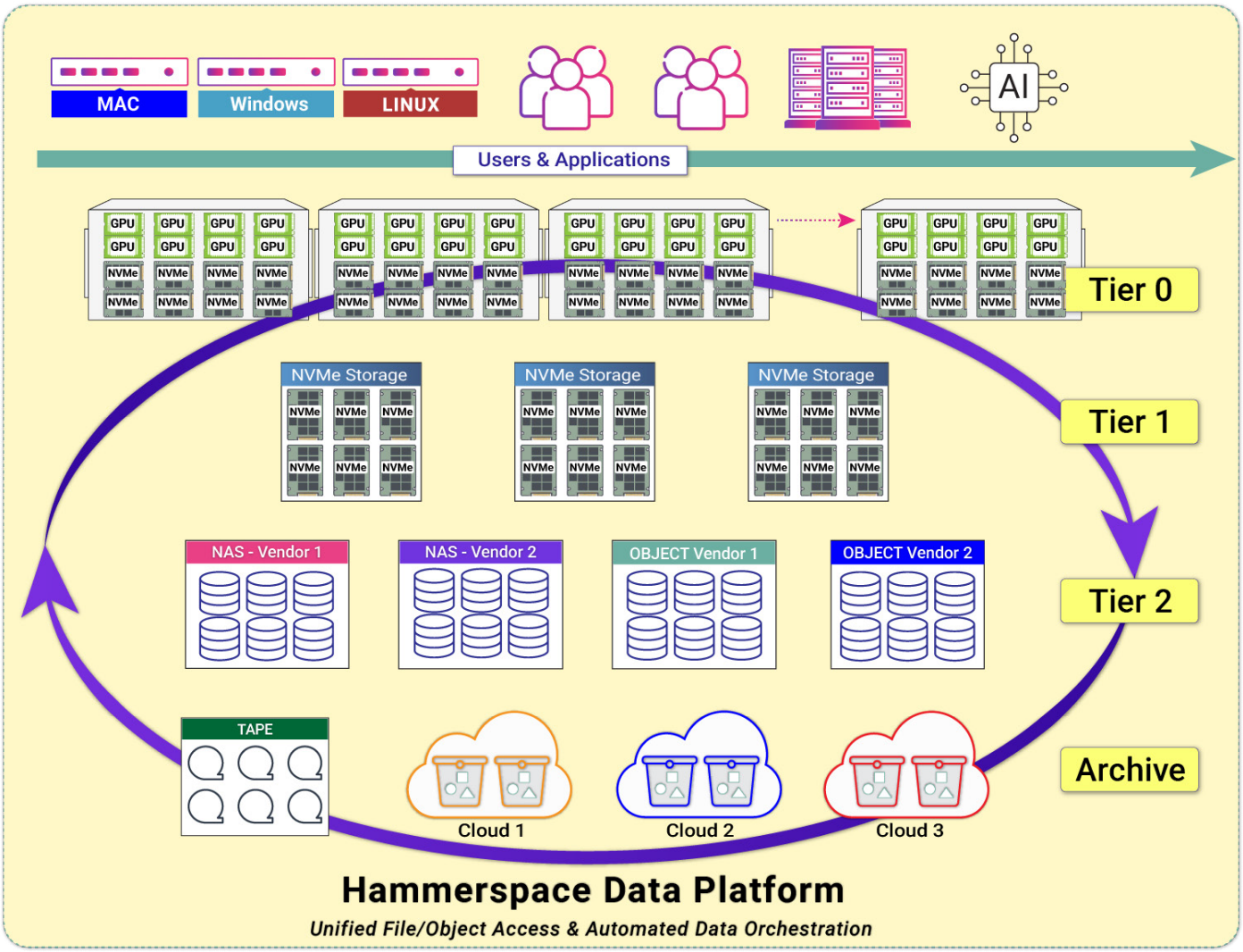


그림 2: Hammerspace는 GPU 서버 내에 사용되지 않고 남아 있는 용량을 활성화하여, 다계층·다벤더 스토리지 환경 내에서 완전한 Tier 0 스토리지를 매끄럽게 통합한다. Hammerspace 데이터 오케스트레이션은 데이터 배치와 데이터 보호를 자동화하여 GPU 효율을 극대화한다.

Hammerspace는 스토리지 사일로를 제거하고 플랫폼 간 데이터 오케스트레이션을 자동화함으로써 기존 인프라를 AI에 최적화된 환경으로 전환하는 과정을 획기적으로 단순화하는 데이터 플랫폼을 만들었다. 고객이 별도의 클라이언트 소프트웨어를 설치하거나 기존 인프라를 변경할 필요 없이, Hammerspace는 고성능 병렬 파일 시스템을 구축하여 서버 내 로컬 NVMe, 원격 사이트, 클라우드를 포함한 모든 벤더의 여러 스토리지 유형에 걸쳐 통합된 파일/오브젝트 접근을 제공한다.

하이퍼스케일 AI 워크로드에 필요한 확장성과 성능을 위해, Hammerspace는 Linux 표준 pNFS v4.2와 Flex Files의 병렬 처리 기능을 독자적으로 활용하여 GPU로의 데이터 오케스트레이션을 극한 수준의 성능으로 자동화한다. Hammerspace가 수년 전 Linux 커널에 기여한 표준이기 때문에 pNFS v4.2 클라이언트는 이미 전 세계 모든 데이터 센터에 설치되어 있으며, 지난 5년 이상 거의 모든 Linux 배포판에 기본 포함되어 있다.

이러한 기능 덕분에 Hammerspace는 Meta가 필요로 하는 선형적 확장성(linear scalability)을 달성할 수 있었으며, 이를 통해 Llama 2, 3, 그리고 이제 Llama 4 대규모 언어 모델(LLM)을 온프레미스와 클라우드 기반 데이터 센터에서 수천 개 이상의 GPU에서 구동할 수 있게 되었다.

또한 Hammerspace는 데이터를 독자적인 컨트롤러 노드를 통해 전달할 필요도 없고, 기존 사용자 PC와 애플리케이션 서버에 전용 클라이언트 소프트웨어를 설치하도록 요구하지도 않으며, 어떤 벤더의 서버나 스토리지든 기존 인프라를 변경할 필요 없이 이러한 성능을 구현한다[2].

Tier 0 활성화의 핵심

1. 데이터 보호 및 오케스트레이션

GPU 서버 클러스터 전반에서 로컬 NVMe 스토리지를 활성화하고 보호하기 위해, Hammerspace는 이 용량을 여러 사이트와 클라우드를 포함한 어떤 벤더의 스토리지 유형이든 여러 계층으로 구성할 수 있는 글로벌 네임스페이스로 통합한다.

그러나 글로벌 파일 시스템에서 모든 스토리지 유형을 아우르는 것만으로는 충분하지 않다. 진정한 핵심 역량은 스토리지 유형과 위치 간 파일 단위로 무중단 데이터 오케스트레이션을 자동화할 수 있는 능력이다.

But spanning all the storage types in a global file system is notenough. The key Hammerspace superpower that makes Tier 0 effective is the ability to automate non-disruptive file-granular data orchestration between Tier 0 and any other class of storage, both on-premises and remote.

HSM(계층형 스토리지 관리)이나 기타 포인트 솔루션에서 흔히 사용하는 단순하고 제한된 정책과 달리, Hammerspace는 관리자가 데이터 내구성, 데이터 배치, 워크플로우 자동화, 기타 데이터 서비스 등 비즈니스 로직에 기반하여 세밀하게 조정 가능한 포괄적인 서비스 레벨 오브젝티브(Service-Level Objective)를 생성할 수 있도록 지원한다.

위 예시처럼 체크포인트 파일이 로컬 Tier 0에 기록되는 경우, Hammerspace에서 서비스 레벨 오브젝티브(Service Level Objective)를 설정하여 해당 체크포인트를 어떻게 보호할지, 필요한 데이터 내구성 수준은 무엇인지, 어디에 배치할지 등을 정의한다.

---

2. pNFS v4.2와 Flex Files가 제공하는 업계 최고 수준의 기능에 대한 자세한 내용은 백서 "Parallel NFS v4.2를 활용하여 레가시 인프라를 GPU 기반 AI/DL 워크로드에 적합하게 전환하는 방법" 백서를 참고하면 된다.

예를 들어, 파일이 Tier 0에 기록되는 즉시 Hammerspace는 해당 파일을 비동기적으로 오브젝트 스토리지나 다른 대상 스토리지로 복사하거나 이동할 수 있다. 또는 Tier 0 용량이 큰 클러스터에서는 필요 시 빠른 복구를 위해 파일을 클러스터 내 다른 노드로 복제할 수도 있다. 성능이나 내구성 요구에 맞는 파일의 인스턴스 수는 간단한 선언적 오브젝티브(Objective) 설정만으로 결정할 수 있다.

Hammerspace에서는 이러한 파일이 나중에 따로 정리해야 하는 분기된(forked) 복사본이 아니다. 대신, 동일 파일의 여러 인스턴스이며, 공유된 파일 메타데이터를 사용하고 글로벌하게 사용자와 애플리케이션이 접근할 수 있으며, Hammerspace가 통합 네임스페이스에서 일관되게 관리한다.

2. 데이터 오케스트레이션은 모든 워크로드에서 Tier 0를 활성화한다

하지만 Hammerspace의 데이터 오케스트레이션은 체크포인트 파일을 보호하는 데만 사용되지 않는다. pNFS v4.2와 Flex Files가 제공하는 극한의 성능을 기반으로, Hammerspace의 Objectives는 Tier 0로 들어오고 나가는 모든 데이터 이동을 자동화하여 GPU 클러스터 전반에서 워크로드를 병렬화할 수 있다.

클러스터 내 서버들 사이의 "east-west" 노드 간 네트워크 속도는 서버와 외부 스토리지 간 "north-south" 네트워크 속도보다 훨씬 빠르다는 점을 기억해야 한다. 즉, Hammerspace가 데이터를 컴퓨트 노드 가까운 Tier 0로 스테이징하는 기능과 pNFS v4.2를 활용해 I/O를 병렬화하여 가속하는 기능이 결합되면, Tier 0를 활성화하는 것이 어떤 고성능 워크로드에도 여러 측면에서 순이익을 제공하게 된다.

또한 네트워크가 양방향이므로, Hammerspace는 동일한 Tier 0에서 학습 데이터를 읽는 GPU 노드(GPU node receiving)에 필요한 대역폭에 영향을 주지 않고, 체크포인트 파일을 Tier 0(GPU node sending)에서 외부 스토리지 또는 다른 Tier 0 노드로 백그라운드에서 동시에 이동시킬 수 있다.

3. GPU는 어디 있는가? 클라우드 기반 GPU 클러스터에서 버스트 워크로드 활성화

AI 이니셔티브를 시작하려는 조직이 직면하는 가장 큰 문제는 온프레미스 GPU의 가용성이다. 게다가 GPU를 구매할 수 있다고 해도, 계획된 사용량이나 데이터 센터 전력 여유가 GPU를 직접 소유하는 것이 최선의 선택이 아닐 수 있다. 그래서 많은 조직이 GPU-as-a-Service를 제공하는 서드파티 업체를 이용하거나 클라우드 기반 GPU를 임대한다. 그러나 이 두 가지 옵션 모두 복잡하며, 원격 리소스로 파일/오브젝트를 수동으로 업로드 및 다운로드해야 하고, 이러한 워크로드를 또 하나의 별도 사일로에서 관리해야 한다는 문제가 있다.

Hammerspace는 글로벌 파일 시스템과 데이터 오케스트레이션을 온프레미스 데이터 센터와 하나 이상의 클라우드로 확장함으로써 이 과정을 획기적으로 단순화한다. 또한 주요 퍼블릭 클라우드 서비스 어디에서든 몇 분 만에 새로운 Hammerspace 클러스터를 구동할 수 있으며, 작업에 필요한 규모로 손쉽게 확장할 수 있다.

Hammerspace는 이러한 문제를 제거하며, 온프레미스 데이터 센터와 하나 이상의 클라우드 리소스 간 글로벌 파일 시스템과 데이터 오케스트레이션을 확장한다.

클라우드에서 새 Hammerspace 인스턴스를 활성화하면, 몇 분 안에 기존 온프레미스 Hammerspace 클러스터와 파일 시스템 메타데이터가 자동으로 동기화된다. 이후 클라우드 기반 GPU 클러스터의 Tier 0 스토리지를 포함한 모든 클라우드 스토리지는 온프레미스 Hammerspace 글로벌 네임스페이스의 원활한 확장으로 작동한다.

Hammerspace의 오브젝티브(Objective)는 이제 클라우드 기반 GPU의 Tier 0를 포함해, 온프레미스 데이터센터 내 다른 스토리지 계층 간에 수행되는 것과 동일한 방식으로 클라우드로의 데이터 이동과 클라우드에서의 데이터 이동을 자동으로 오케스트레이션할 수 있다. 이러한 멀티사이트 데이터오케스트레이션 기능을 통해 최대 16개의 온프레미스 및 클라우드 기반 Hammerspace 클러스터를 연결하여, 어떤 벤더의 기존 또는 신규 스토리지든 모두 아우르는 단일 통합 고성능 네임스페이스를 구축할 수 있으며, 이는 진정한 글로벌 데이터 환경을 실현한다.

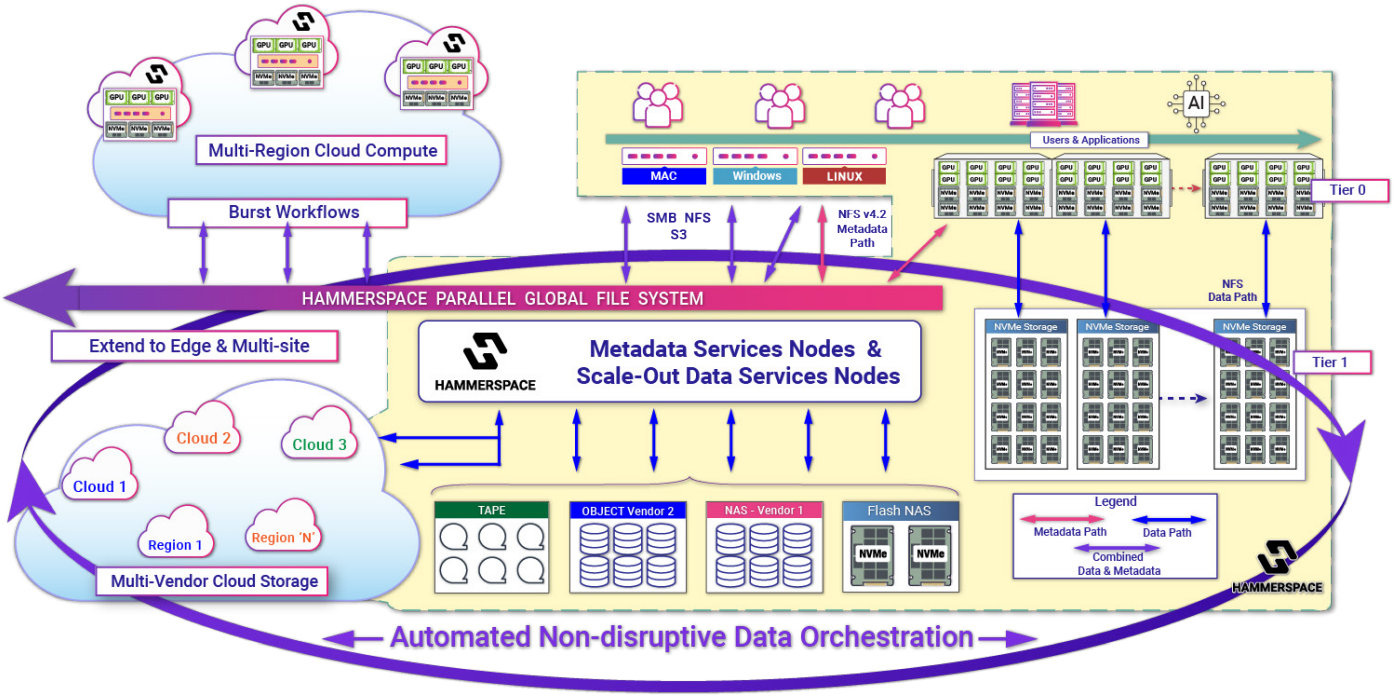


그림 3: Hammerspace 데이터 플랫폼은 GPU 서버 내 Tier 0를 활성화하며, 온프레미스 및 클라우드 기반 리소스 전반에 걸쳐 통합된 파일/오브젝트 액세스와 데이터 오케스트레이션을 제공한다.

사용자와 애플리케이션은 기존과 동일하게 표준 SMB, NFS, S3 마운트 포인트를 통해 파일 시스템을 그대로 이용할 수 있으며, 별도의 클라이언트 소프트웨어 설치나 워크플로우 변경할 필요가 없다. Hammerspace는 데이터 배치와 다양한 데이터 서비스를 처리하며, 모든 사일로 전반에서 워크플로우를 백그라운드에서 자동화하여 사용자나 애플리케이션에 어떠한 중단도 발생시키지 않는다.

비용 절감을 극대화하기 위해, 버스트 워크로드가 종료되면 클라우드 기반 Hammerspace 클러스터를 즉시 종료하여 클라우드 컴퓨팅 비용을 줄일 수 있다. 데이터 오케스트레이션은 이미 백그라운드 작업으로 투명하게 수행되므로, GPU에서 생성된 결과 데이터는 자동으로 보호되고 필요한 위치와 스토리지 유형에 적절히 배치된다.

4. 표준 기반 공유 파일/오브젝트 액세스

앞서 언급했듯이, Hammerspace는 사용자 PC나 애플리케이션 서버에 별도의 전용 클라이언트 소프트웨어를 설치할 필요가 없으며, 기존 스토리지를 변경할 필요도 없다. 사용자들은 Hammerspace가 존재하는 사실조차 인식하지 못할 수 있다. 최종 사용자와 애플리케이션은 이전과 동일하게 표준 NFS, SMB, S3 프로토콜을 통해 글로벌 파일 시스템에 액세스할 수 있으며, 이제는 모든 스토리지 유형과 사일로, 사이트에 걸쳐 글로벌 네임스페이스 전체에 대한 통합된 접근을 사용할 수 있게 된다.

또한 사용자와 애플리케이션이 접근 권한을 가진 데이터만이 글로벌 파일 시스템 전체에서 공유를 통해 표시된다.

데이터가 현재 어떤 스토리지 유형에 저장되어 있는지는 전적으로 비즈니스 규칙에 따라 결정된 Hammerspace 서비스 레벨 오브젝티브(SLO)에 의해 관리되며, 사용자와 애플리케이션에는 완전히 투명하다. 사실 Hammerspace는 pNFS v4.2와 Flex Files의 일부로 Linux 커널에 기여한 Live Data Mobility라는 핵심 기능을 통합하여, 열려 있고 현재 읽기 또는 쓰기 중인 파일조차도 스토리지 계층 간에 백그라운드 작업으로 투명하게 오케스트레이션될 수 있다. 라이브 파일 이동 중에도 애플리케이션이나 사용자 워크플로우에는 아무런 중단이 없다.

S3 버킷을 기대하는 애플리케이션은, 스토리지 종류와 관계없이 모든 데이터가 포함된 S3 버킷을 그대로 볼 수 있다.
NFS 또는 SMB 공유를 통해 파일 계층 구조를 기대하는 애플리케이션은, 동일한 데이터를 표준 파일/폴더 구조로 제공받아 그대로 볼 수 있다.

Hammerspace는 모든 스토리지의 모든 데이터에 대해 표준 프로토콜을 통한 범용 접근을 제공하고, 파일/오브젝트 단위로 데이터를 어디서든 백그라운드에서 투명하게 오케스트레이션할 수 있는 기능을 제공한다. 이는 특히 활성화하려는 대량의 비정형 데이터를 가진 조직에 중요하다. 전체 볼륨을 새 저장소로 마이그레이션할 필요 없이, Hammerspace는 작업에 필요한 특정 파일/오브젝트만을 컴퓨팅 자원으로 오케스트레이션하며, 분기된(forked) 복사본을 생성하거나 다른 사용자/애플리케이션을 방해하지 않는다.

비용이 많이 드는 마이그레이션을 수행하고 데이터를 새로운 저장소로 복사하여 또 다른 고립된 사일로를 만드는 시대는 끝났다. 이는 기존 인프라를 완전히 AI-ready로 전환하는 작업을 극적으로 단순화한다.

그렇다면 HCI 및 CPU 기반 서버는 어떨까?

이 문서의 초점은 AI 활용 사례에서 가장 시급하게 요구되는 GPU 기반 워크로드에 맞춰져 있지만, 실제로는 CPU 기반 서버에도 Tier 0로 활성화할 수 있는 동일한 초고속 로컬 NVMe 드라이브가 탑재되는 경우가 많다. 이는 저전력 ARM CPU에서도 마찬가지다.

이러한 서버 유형은 종종 하이퍼컨버지드 인프라(HCI)와 혼동되곤 하는데, HCI는 일반적으로 컴퓨트와 스토리지를 단일 장치로 결합한 시스템이다. HCI 시스템은 주로 구조화된 데이터, VM, 블록 기반 I/O 패턴에 사용되며, 이는 CPU 집약적인 비정형 데이터 워크로드에는 적합하지 않다. HCI 시스템이 분산 스토리지 아키텍처나 다른 계층화 솔루션을 사용할 수는 있지만, 블록 기반 사용 사례에 초점을 맞추기 때문에 비정형 데이터 처리에는 적합하지 않다.

그러나 구조화된 데이터는 전체 엔터프라이즈 데이터의 약 20%에 불과하며, 업계 분석가들에 따르면 연간 성장률도 약 20% 수준에 머물고 있다. 반면, 비정형 파일/오브젝트 데이터는 전체 엔터프라이즈 데이터의 약 80%를 차지하며, 연간 약 65%의 성장률을 보이는 것으로 추정된다.

대량의 비정형 데이터를 필요로 하는 수많은 CPU 집약적 활용 사례가 있으며, 이들 역시 앞서 설명한 GPU 컴퓨팅 워크로드와 마찬가지로 Tier 0의 성능 및 비용 효율성 향상을 그대로 활용할 수 있다. 이러한 사례에는 CPU 기반 AI 추론, 예측 분석, 비정형 데이터셋을 직접 처리하는 머신러닝 파이프라인 등 다양한 AI 활용 사례가 포함된다.

그러나 AI 활용 사례를 넘어, 고성능을 요구하는 다양한 CPU 기반 애플리케이션 워크플로우가 존재하며, 이들 역시 CPU 서버 내 로컬 NVMe 드라이브에서 Tier 0를 활성화함으로써 제공되는 초저지연과 향상된 처리량의 직접적인 혜택을 얻을 수 있다.

다음과 같은 활용 사례들이 포함된다:

- **실시간 분석 및 데이터 스트리밍**
  - 로그 데이터, 센서 입력, 클릭스트림 데이터를 신속하게 처리하여 즉각적인 인사이트 생성.
- **고성능 검색 및 인덱싱**
  - 대규모 문서 저장소나 데이터셋에서 빠른 인덱싱 및 검색 수행(예: Elasticsearch, Splunk).
- **비디오 및 미디어 처리**
  - 빠르고 빈번한 데이터 접근이 필요한 저지연 편집, 트랜스코딩, 렌더링, 콘텐츠 전송.
- **금융 트레이딩 및 모델링**
  - 대규모 데이터셋에 즉시 접근해야 하는 실시간 리스크 분석, 알고리즘 트레이딩, 금융 모델링.
- **유전체학 및 바이오인포매틱스**
  - 유전체 시퀀싱 데이터, 이미지 데이터 또는 바이오인포매틱스 워크로드를 거의 실시간에 가깝게 처리.

CPU 기반 Tier 0의 장점은 위 GPU 기반 예시에서 강조된 장점들과 동일하다:

- 서버 내부의 초고성능 로컬 NVMe 드라이브를 활성화함으로써, 비싼 네트워크를 통해 외부 스토리지 어레이에 접근하는 것보다 훨씬 낮은 비용으로 운영할 수 있다.,
- 외부 스토리지 대비 Tier 0가 제공하는 대폭 향상된 성능과 낮은 지연시간은 이러한 애플리케이션에 직접적인 이점을 제공한다.

CPU 기반 워크로드를 위한 추가적인 Hammerspace 성능 이점

pNFS v4.2와 Flex Files가 제공하는 극한의 성능과 CPU 기반 서버 내에서 Tier 0를 활성화하는 기능을 활용하는 것에 더해, Hammerspace는 표준 Linux와의 긴밀한 통합을 통해 CPU 기반 워크로드에 또 하나의 직접적인 성능 이점을 제공한다.

HCI 시스템과 일부 다른 벤더 솔루션들은 I/O 처리를 위해 폴링(polling) 기반 방식을 사용한다. 이는 사용자 공간(user space)에서 동작하는 방식으로, CPU가 들어오는 데이터가 있는지 지속적으로 확인하느라 계속 바쁘게 유지된다. 이 폴링은 데이터가 전혀 없을 때에도 끊임없이 발생하며, 상당한 CPU 리소스를 소모한다. 그 결과 CPU에 과도한 부하가 걸리면서 지연(latency)이 증가하고 성능이 저하된다.

Hammerspace는 Linux 커널 기반의 표준 NFS를 활용하여 I/O를 가속화하며, IRQ(Interrupt Request, 인터럽트 요청) 방식을 통해 CPU 오버헤드를 크게 줄이고 지연시간을 감소시킨다.

폴링 기반 방식처럼 지속적인 CPU 부하를 요구하지 않고, Hammerspace가 사용하는 커널 기반 IRQ 방식은 데이터가 실제로 도착했을 때만 CPU에 알림을 전달한다. 이를 통해 훨씬 더 효율적인 이벤트 기반 처리(event-driven processing)가 가능해진다.

그 결과, CPU 집약적인 폴링 루프에 의존하는 사용자 공간 기반 솔루션과 비교했을 때, 더 빠르고 지연이 낮은 I/O 성능과 더 효율적인 CPU 사용을 실현할 수 있다.

향후 출시 예정: 곧 여러분이 사용하는 리눅스 배포판에서 만나게 될 기능

Hammerspace가 최근 리눅스 업스트림에 기여한 또 하나의 핵심 혁신은 리눅스 커널 내 NFS 프로토콜 바이패스(NFS Protocol Bypass)로, 이는 Tier 0에서 로컬 NVMe의 성능을 더욱 비약적으로 향상시키는 기술이다.

리눅스 커뮤니티에서 LOCALIO로 알려진 이 기능은 NFS 클라이언트와 서버가 동일한 호스트에서 실행 중인지 감지할 수 있도록 지능을 추가한 것이다. 이를 통해 커널의 NFS 프로토콜 스택을 거쳐 다시 로컬 파일 시스템과 스토리지로 되돌아가는 불필요한 지연을 제거한다. 또한 이 기능은 NFS 클라이언트와 서버가 컨테이너에서 실행되는 경우에도 정상적으로 동작할 만큼 견고하게 구현되었다.

LOCALIO는 2024년 11월에 릴리스된 리눅스 커널 6.12에 처음 포함되었으며, 이후 2025년 3월에 릴리스된 커널 6.14에서 추가적인 기능 향상이 이루어졌다.

벤치마크 테스트의 초기 결과에 따르면, NFS 프로토콜을 우회하는 방식은 서버 내 CPU 사용률을 크게 감소시키는 데 매우 큰 효과가 있는 것으로 나타났다. 또한 현재까지의 테스트에서는 LOCALIO 사용 시 Tier 0에서 읽기 성능이 엄청나게 향상되는 것으로 확인되었다. 초기 테스트 결과에 따르면 쓰기 성능 역시 개선될 것으로 예상되지만, 그 향상 폭은 아직 정량화되는 중이다. 이 테스트는 최신 커널 업데이트를 기반으로 계속 진행 중이며, 결과가 준비되는 대로 문서화될 예정이다.

GPU 클러스터 전반에서 데이터 배치시 호스트 친화성(Host Affinitization) 기능과 결합되면, LOCALIO 기능은 현재 Tier 0가 제공하는 수준을 뛰어넘는 상당한 성능 향상을 제공할 것임이 이미 입증되고 있다.

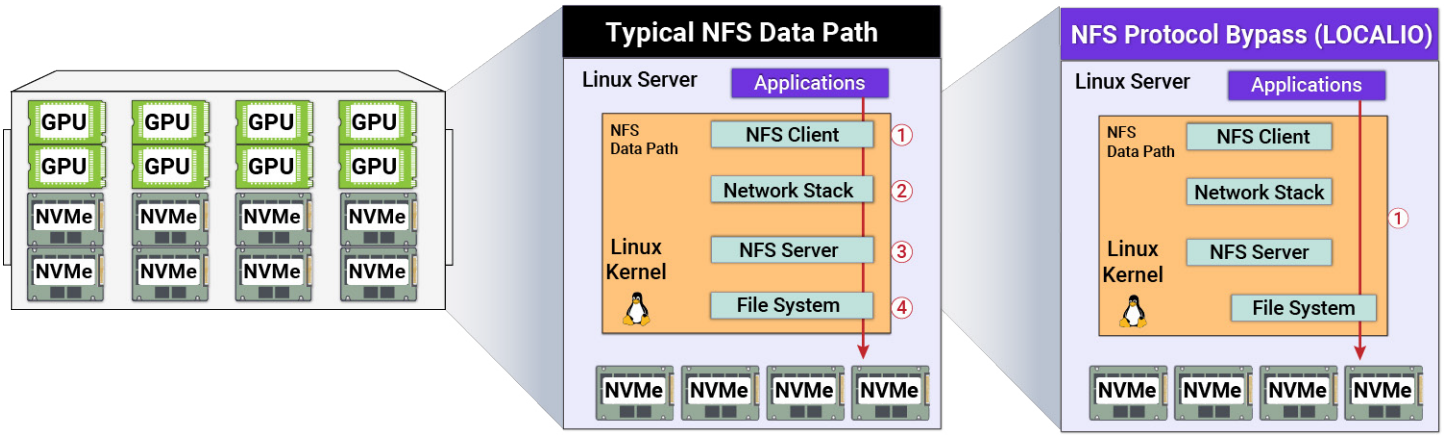


그림 4: Hammerspace 엔지니어링 팀은 GPU 및 CPU 기반 서버에서 로컬 NVMe의 성능을 더욱 향상시키기 위해 LOCALIO를 개발하여 리눅스 커널에 기여했다.

LOCALIO는 2024년 11월에 장기 지원(LTS) 버전으로 릴리스된 리눅스 커널 6.12에 포함되었지만, 엔터프라이즈 리눅스 배포판이 새로운 릴리스를 채택하기까지는 시간이 걸린다. 그러나 이 기능에 대한 수요는 매우 높으며, 2025년 5월에 출시된 최신 Red Hat Enterprise Linux(RHEL 10)은 커널 6.12를 기반으로 하고 있을 뿐만 아니라, 커널 6.14에서 제공된 LOCALIO 개선 사항도 포함하고 있다.

조직들이 최신 리눅스 커널로 업그레이드하면서 RHEL10 및 기타 배포판을 도입하기까지는 시간이 걸리겠지만, 그렇게 전환하는 시점에는 Hammerspace가 기여한 개선 사항들이 이미 포함되어 있어 즉시 활용할 수 있게 될 것이다. 이는 현재 Tier 0를 통해 제공되고 있는 극적인 성능 및 효율성 향상을 한층 더 강화하게 될 것이다.

표준이 핵심이다

Hammerspace는 지난 몇 년 동안 2,400개가 넘는 기능과 패치를 오픈 소스 리눅스 커널에 기여해 오며, 리눅스 성능 향상에 중요한 역할을 해왔다. Hammerspace의 CTO인 Trond Myklebust는 20년 이상 리눅스 NFS 클라이언트 커널의 메인테이너(maintainer)로 활동해 왔으며, 그 역할을 통해 Hammerspace가 수행하는 모든 작업이 커뮤니티에서 유지되는 표준을 충실히 준수하도록 보장해 왔다.

Hammerspace의 수석 소프트웨어 엔지니어인 Mike Schnitzer 역시 리눅스 커널 메인테이너로 활동하고 있으며, 업스트림 리눅스 커널의 Device Mapper(DM) 서브시스템을 담당하고 있다. 또 다른 Hammerspace 소프트웨어 엔지니어인 Tom Haynes는 2018년에 pNFS Flexible File Layout을 정의한 RFC 8435의 공동 저자이며, 이 기능은 2019년에 v4.2 사양에 채택된 pNFS의 핵심 요소이다.

Hammerspace가 오픈 표준에 집중하는 이유는, 기존 인프라와의 호환성을 유지하면서도 독점 벤더의 사일로(silo)와 벤더 종속(lock-in)이라는 엔터프라이즈 데이터 환경의 한계를 극복하려는 비전과 맞닿아 있다. 사용자·애플리케이션 측면에서는 고객이 애플리케이션 서버에 전용 소프트웨어를 설치하거나, 사용자 PC를 변경할 필요가 전혀 없다. 스토리지 측면에서도 Hammerspace는 벤더에 관계없이 어떤 스토리지와도 호환된다.

또한 중요한 점은, Hammerspace가 표준 리눅스에 엔지니어링해 온 극적인 성능 향상 기능들이 모든 주요 배포판에 포함되어 있을 뿐 아니라, 이러한 기능들이 Hammerspace 소프트웨어 내부에 긴밀하게 통합되어 있어, Hammerspace가 이를 가장 효과적으로 활용할 수 있는 독보적인 위치에 있다는 것이다.

요약

모든 산업 분야와 모든 규모의 공공 또는 민간 조직은 폭발적으로 성장하는 AI의 가능성에 대응하기 위해 IT 인프라를 어떻게 재구성할지 고민하고 있다. 과거처럼 엔터프라이즈 IT가 애플리케이션과 데이터가 일대일로 대응되는 대체로 선형적인 구조였던 시대는 이미 지나갔다. 비정형 데이터가 시간이 지나면 점차 오래되어 결국 아카이브에서 잠든 상태로 남아 있던 그런 시대도 더 이상 존재하지 않는다.

조직들은 일반적으로 AI 전략이 필요하다는 사실을 인식하고 있지만, AI 이니셔티브에 대한 투자 수익(ROI)은 그리 확실하지 않다. 특히 하이퍼스케일 기업이 아닌 조직들에게는 더욱 그렇다. 그러나 업계 전문 매체를 잠깐만 훑어보아도, 마치 AI 프로젝트를 시작하는 유일한 방법이 기존 인프라 옆에 별도의 AI 전용 IT 아키텍처를 새로 구축하는 것인 것처럼 보이기 쉽다.

이러한 솔루션들은 새로운 고성능 스토리지 시스템과 고가의 고속 네트워킹 장비를 추가로 구매하도록 유도하거나, 고객의 애플리케이션 서버에 전용 클라이언트 소프트웨어를 설치해 수정하도록 요구하며, 그 외에도 여러 가지 비용이 많이 들고 벤더 종속적인 추가 솔루션들을 필요로 한다.

이러한 모든 솔루션들은 독점적 데이터 사일로가 접근을 분절시키고 데이터 활용 효율을 저해하던 과거의 문제를 그대로 지속시킨다. 이러한 접근 방식은 전체 과정을 불필요하게 복잡하게 만들고, 데이터 중력(data gravity) 문제와 비효율적인 자원 사용을 초래하며, 무엇보다도 CAPEX와 OPEX 모두에서 상당히 높은 비용 증가로 이어진다.

이는 조직들이 AI 전용 리소스를 완전히 새로운 고립된 사일로로 구축할 필요 없이, 기존 IT 인프라에 추가적이고 긴밀하게 통합된 형태로 도입할 수 있게 되었음을 의미한다.

Hammerspace는 표준 기반 접근 방식이 조직이 기존 IT 인프라를 AI에 적합하도록 준비하는 작업을 극적으로 단순화한다는 사실을 이미 입증했다. AI 여정을 시작하거나 확장하는 조직에게 이는 AI 전용 리소스를 기존 IT 인프라와 분리된 새로운 사일로로 구축할 필요 없이, 기존 인프라에 추가적이고 긴밀하게 통합된 형태로 도입할 수 있다는 의미다.

Hammerspace 데이터 플랫폼은 온프레미스와 멀티 클라우드를 아우르는 실제 환경에서 통합 파일/오브젝트 접근과 고성능 데이터 오케스트레이션을 가능하게 함으로써 이 비전을 입증해 왔다. 이를 통해 모든 규모의 조직에서, HPC부터 하이퍼스케일 AI 환경에 이르기까지 모든 산업 분야에서, 여러 온프레미스 및 하이브리드 클라우드 사이트 전반에 걸쳐 극도로 방대한 데이터 규모와 높은 성능 요구사항까지 확장할 수 있음을 보여주었다.

이러한 기능과 실제 환경에서의 경험은 Tier 0가 GPU 기반과 CPU 기반의 고성능 워크로드 모두에 게임 체인저급 영향을 미치는 솔루션이 되도록 만드는 핵심 요소이다.

Hammerspace는 모든 산업 분야, 모든 규모의 조직이 Tier 0가 제공하는 전례 없는 성능 향상을 활용할 수 있도록 지원하며, 이는 현재 시장에서 제공되는 가장 빠른 외부 스토리지 유형을 능가한다. 또한 이처럼 극도로 높은 성능의 스토리지 계층을 활성화함으로써 얻을 수 있는 놀라운 비용 절감 효과는 그 자체로도 큰 가치가 있다.

게다가 이 성능 계층이 이미 조직이 보유하고 있을 가능성이 높은 인프라를 기반으로 활성화된다는 사실은 매우 획기적이다.

이것이 바로 Hammerspace의 혁신이다.