Openinfra Community Days Korea 2023

SupremeRAID를 활용한, Pacemaker 기반의 고성능 스토리지 이중화 설계

발표자 김태훈 | ㈜글루시스 선임연구원 (thkim@gluesys.com)

목차

고가용성과 이중화

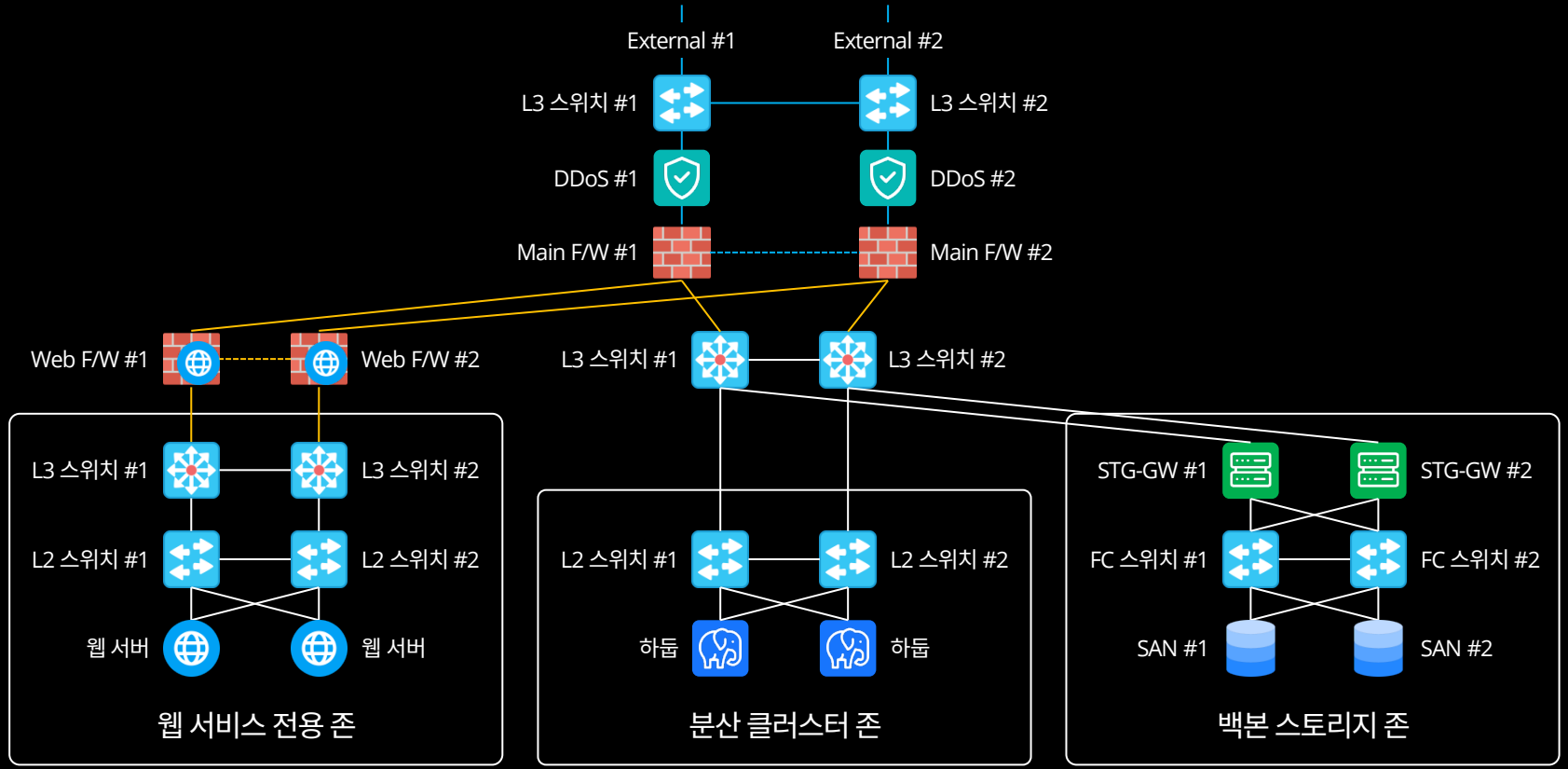
고가용성(HA, High Availability)

- 시스템이 지속적으로 정상 작동하는 능력
- 서비스를 중단 없이 운용할 수 있는 시스템의 안정성
- 99%, 99.9% 등으로 서비스 품질 수준 표현

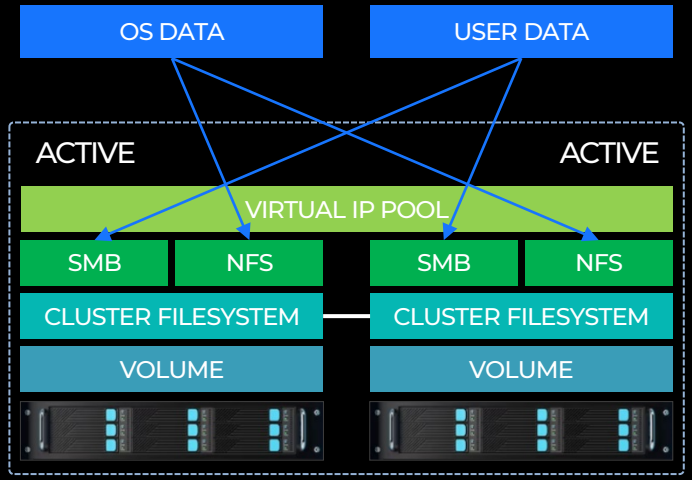
고가용성을 위해서는
단일 장애 지점을 제거하는 것이 무엇보다 중요!

⇒ 인프라 구성 요소의 이중화 필요!

인프라 이중화

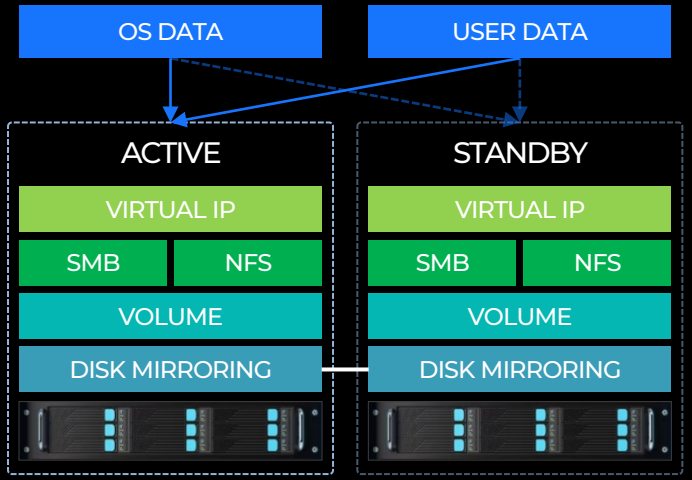

External #1
External #2
L3 스위치 #1
L3 스위치 #2
DDoS #1
DDoS #2
Main F/W #1
Main F/W #2
Web F/W #1
Web F/W #2
L3 스위치 #1
L3 스위치 #2
L3 스위치 #1
L3 스위치 #2
L2 스위치 #1
L2 스위치 #2
웹 서버
웹 서버
웹 서비스 전용 존
L2 스위치 #1
L2 스위치 #2
하둡
하둡
분산 클러스터 존
STG-GW #1
STG-GW #2
FC 스위치 #1
FC 스위치 #2
SAN #1
SAN #2
백본 스토리지 존

이중화 방식



Active-Active

- 모든 서버 또는 노드가 운영 상태로 실제 트래픽 동시 처리
- 로드 밸런서 또는 클러스터 관리 툴을 통해 트래픽 분산
- 모든 노드가 활성화 되어 자원 활용 극대화 및 성능 향상
- 노드 간 데이터 동기화 및 일관성 유지로 시스템 복잡도 증가

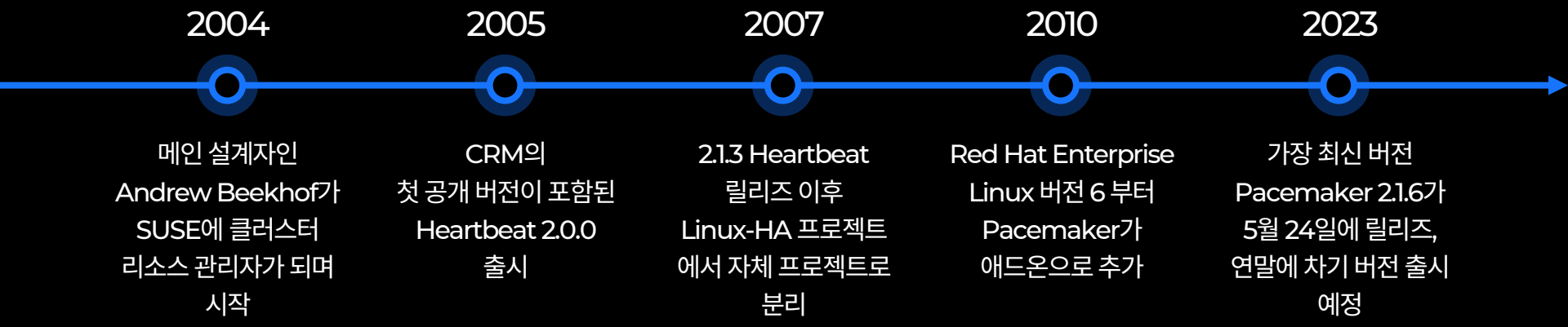


Active-Standby

- 운영 시스템에 장애 발생 시 대기 상태의 시스템을 운영 상태로 전환하여 서비스 지속
- 시스템 구현이 간단하고 서비스 중단 시간을 최소화 할 수 있음
- 운영되지 않는 대기 상태의 시스템은 리소스 낭비

Pacemaker

- 리눅스 기반 시스템에서 클러스터 관리를 위해 사용되는
  오픈소스 클러스터 리소스 관리 프레임워크
- github.com/ClusterLabs/pacemaker
- 2023년 5월 23일 Ken Gaillot(kgaillot@redhat.com) Pacemaker-2.1.6 출시



Pacemaker 주요 기능

리소스 관리

리소스 에이전트를 통해
클러스터 내 리소스의
동작 제어 및 상태
모니터링

장애 감지 및 복구

클러스터 노드 간
통신을 통한 실시간 장애
감지 및 자동 복구로
서비스 연속성 보장

로드 밸런싱

로드 밸런싱을 통한
클러스터 내 트래픽의
균등 분산으로 성능 향상
및 응답성 개선

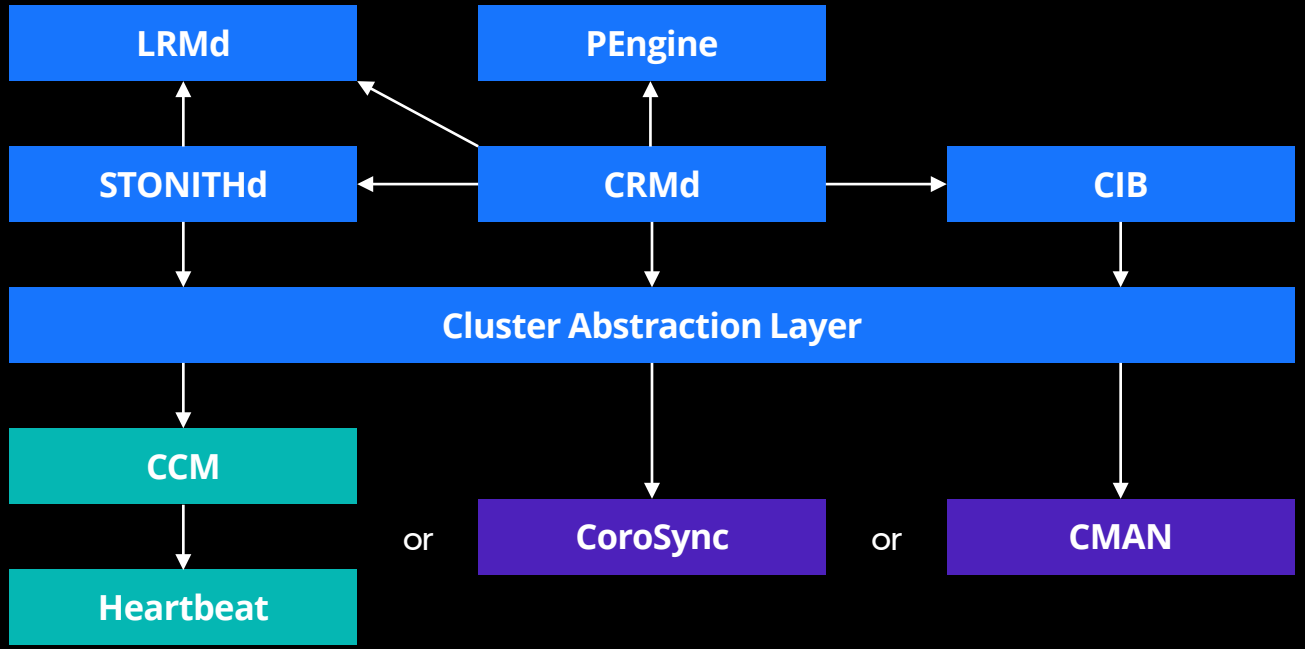
자동 확장

신규 노드 추가 시
리소스의 자동 배포와
로드 밸런싱을 통해
새로운 자원 활용

강력한 정책 및 제어

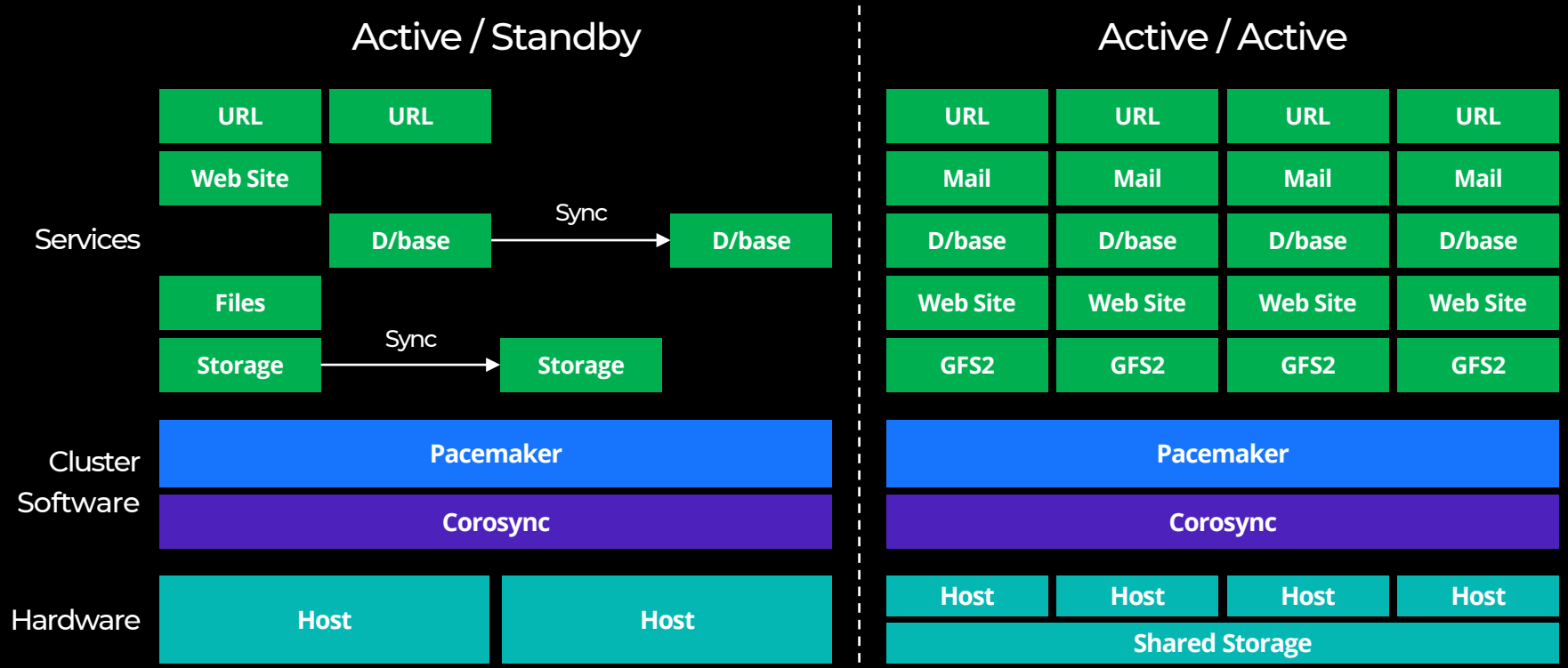
리소스 우선 순위,
장애 조치 정책, 노드 우선
순위 설정 등을 통해
시스템 동작 최적화

Pacemaker Stack



자료 참고 : https://clusterlabs.org/pacemaker/doc/deprecated/en-US/Pacemaker/1.0/html/Pacemaker_Explained/s-intro-architecture.html

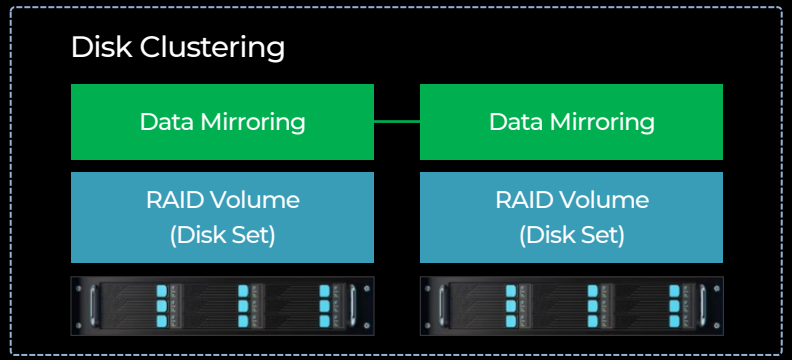
Pacemaker Cluster



자료 참고 : https://clusterlabs.org/pacemaker/doc/deprecated/en-US/Pacemaker/1.0/html/Pacemaker_Explained/s-intro-redundancy.html

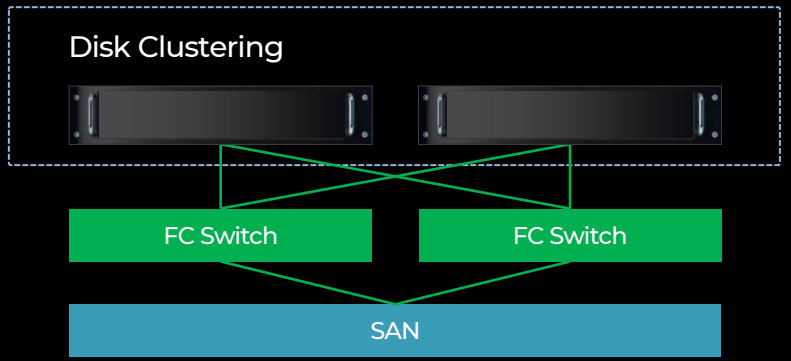
스토리지 이중화

NAS 이중화는 파일 저장과 공유 기능에 영향을 줄 수 있는 장애 지점의 파악 및 자원의 모니터링 필요



데이터 동기화 방식(일체형)

- 별도의 스토리지 없이 NAS에 데이터가 저장되는 디스크가 포함된 일체형 형태
- 물리적으로 분리되어 있는 디스크에 저장된 데이터는 복제 기술을 통해 동기화

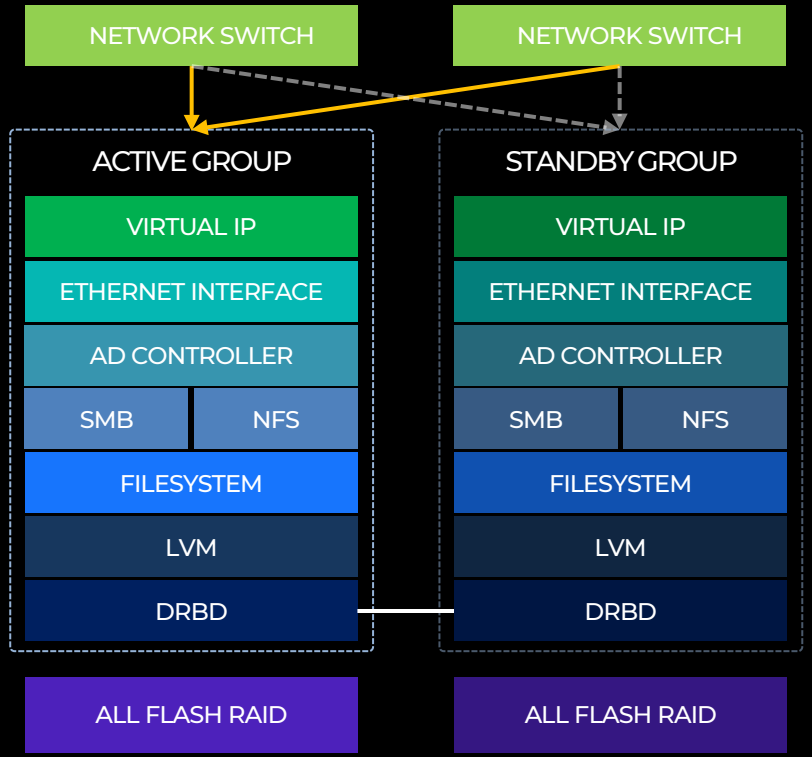


데이터 공유 방식(게이트웨이형)

- SAN이나 DAS와 같은 별도 스토리지의 볼륨을 NAS로 공유하는 형태
- NAS와 데이터 저장 공간이 물리적으로 분리되므로, NAS 확장이나 교체 시에도 원본 데이터는 블록 스토리지에 저장

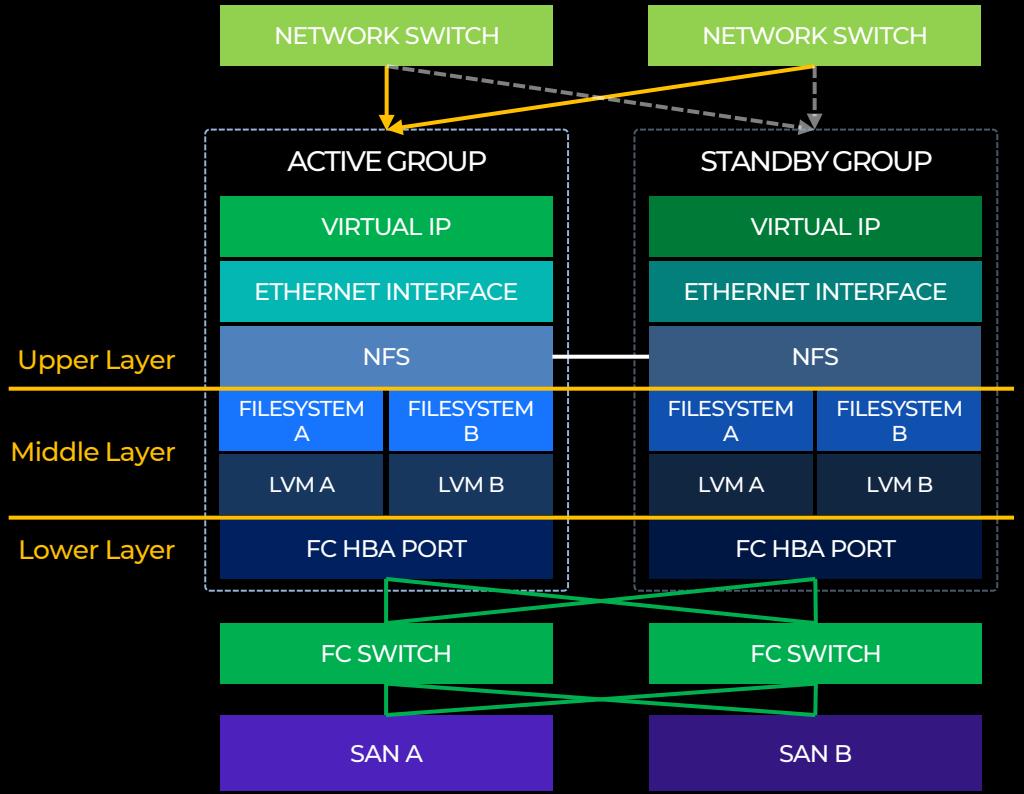
NAS 이중화 아키텍처 설계(1)

- **일체형 NAS + 잔파일 I/O 이중화**
  - 고성능 데이터 복제 환경 제공
    - 올플래시(SSD)로 구성된 RAID 10 구조 사용
    - 스토리지 간 저지연·고성능 데이터 동기화를 위해 인피니밴드 사용
  - VDI 환경에 최적화된 실시간 복제 수행
    - 블록 레벨의 실시간 디스크 복제 기술인 DRBD 적용
  - 복잡성을 낮춰 노드 단위의 Fail-over 수행
    - 특정 구간에서 장애 발생 시 서비스 전체의 신속한 Fail-over 수행
  - 외부 서비스에 대한 연동 간소화
    - 예약된 커맨드를 실행하는 리소스 에이전트 개발
    - 사용자 인증(AD) 서버에 필요한 선행 작업 자동화



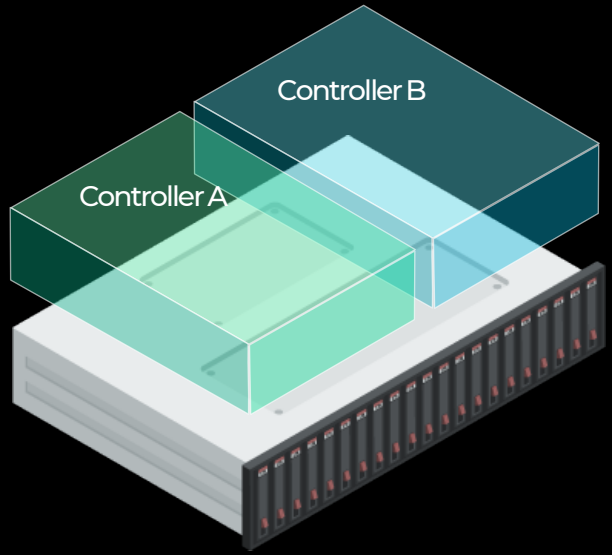
NAS 이중화 아키텍처 설계(2)

- **NAS 게이트웨이 + SAN 이중화**
  - FC HBA Port 리소스 에이전트 개발
    - 외부 스토리지에 대한 호환성 강화
    - 이기종 스토리지 간 연동성 제공
  - 다중 볼륨 구성이 고려된 아키텍처 설계
    - 서비스 이중화 레이어의 모듈화
    - 연관된 리소스 에이전트를 레이어로 구분
  - 유사 Active/Active 구성 고려
    - 특정 서비스 구간의 부분 Fail-over 가능
    - 서비스 이전 및 복구 시간 단축



NAS 이중화 아키텍처 설계(3)

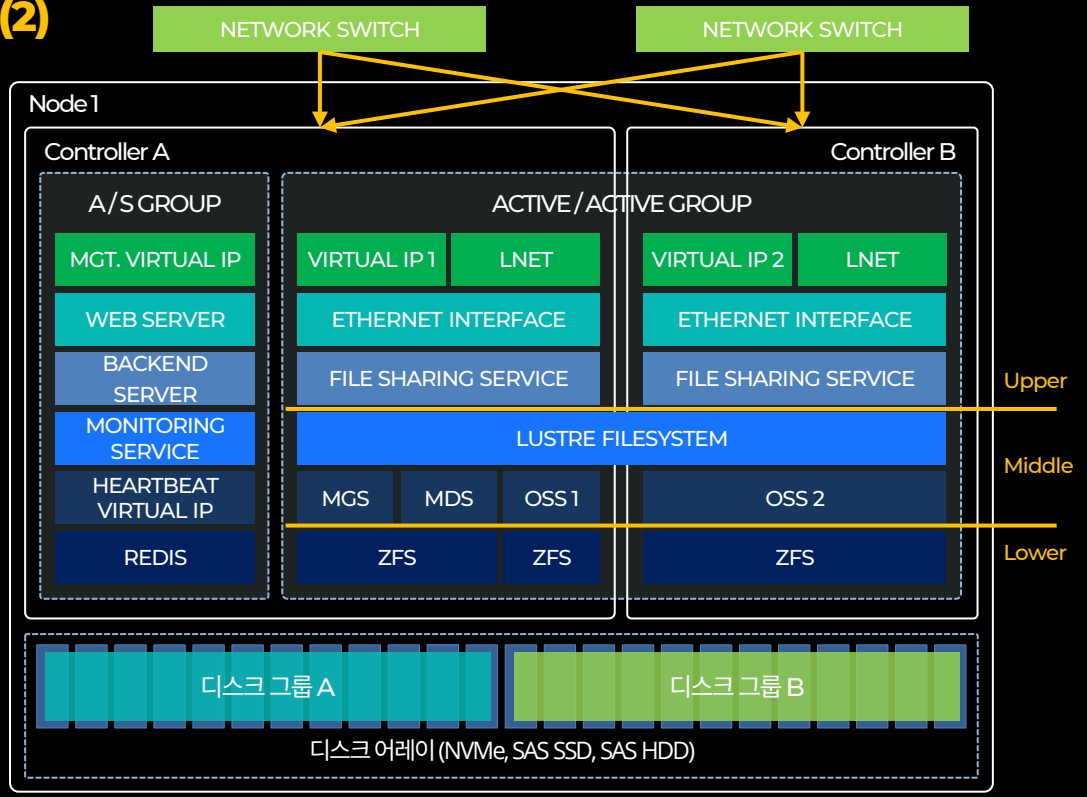
- **Dual Controller 기반의 NAS 이중화(1)**
  - 단일 스토리지 장비에 백플레인(디스크 어레이) 하나에 메인보드가 2개인 하드웨어 장비
    - 운영체제는 2개가 설치되며, 두 운영체제에서 백플레인에 연결된 디스크가 모두 확인됨
    - OS용 디스크는 각 보드에 별도 존재
    - 동일한 용량의 일체형 이중화 모델 대비 디스크 개수는 절반으로 줄어 비용 절감 가능
    - 데이터 동기화 방식(미러링)에서 발생할 수 있는 Split-brain 위험 원천 차단



- 게이트웨이 이중화 모델의 FCmonitor RA 불필요
- 일체형 이중화 모델의 DRBD RA 불필요

NAS 이중화 아키텍처 설계(3)

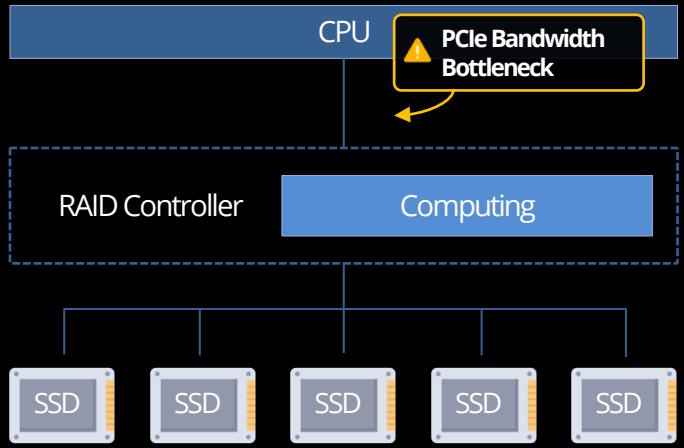
- **Dual Controller 기반의 NAS 이중화(2)**
  - 단일 이중화 클러스터링 구성
    - 파일 공유 서비스를 위한 A/A 그룹과 솔루션 관리를 위한 A/S 그룹으로 구분
  - 리소스 간 상호 연관성 최소화
    - 서비스 그룹은 환경에 따라 계층 별 복수 구성이 가능함
    - 유사 A/A 구성과 동일하게 부분 Fail-over 가능
  - 시스템의 유연성 향상
    - Upper 레이어의 경우 기존 구성을 재사용
    - Middle & Lower 레이어는 제품의 특성에 맞게 새로운 리소스 에이전트 조합으로 구성



고성능 스토리지의 새로운 과제

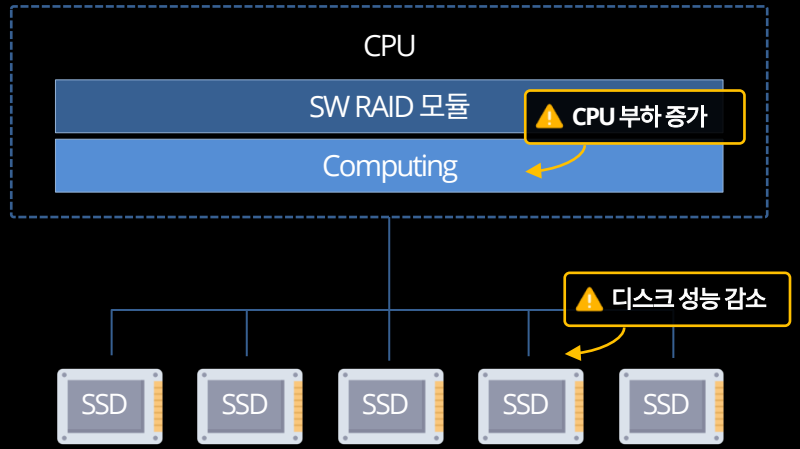
SSD와 RAID 조합에서의 성능 병목 해결!

전통적인 RAID의 한계



HW RAID 구조

- SSD와 HW RAID가 물리적으로 결합된 상태
- 어플리케이션은 HW RAID를 통해 SSD에 접근
- SSD 성능이 향상되면서 CPU와 HW RAID를 연결하는 Bus에서 Bottleneck 발생



SW RAID 구조

- SSD와 CPU가 직접 연결되어 PCIe Bandwidth의 효율적인 활용 가능
- RAID 5, 6 등 패리티 연산 수행 시 CPU 오버헤드 발생으로 성능 제한적

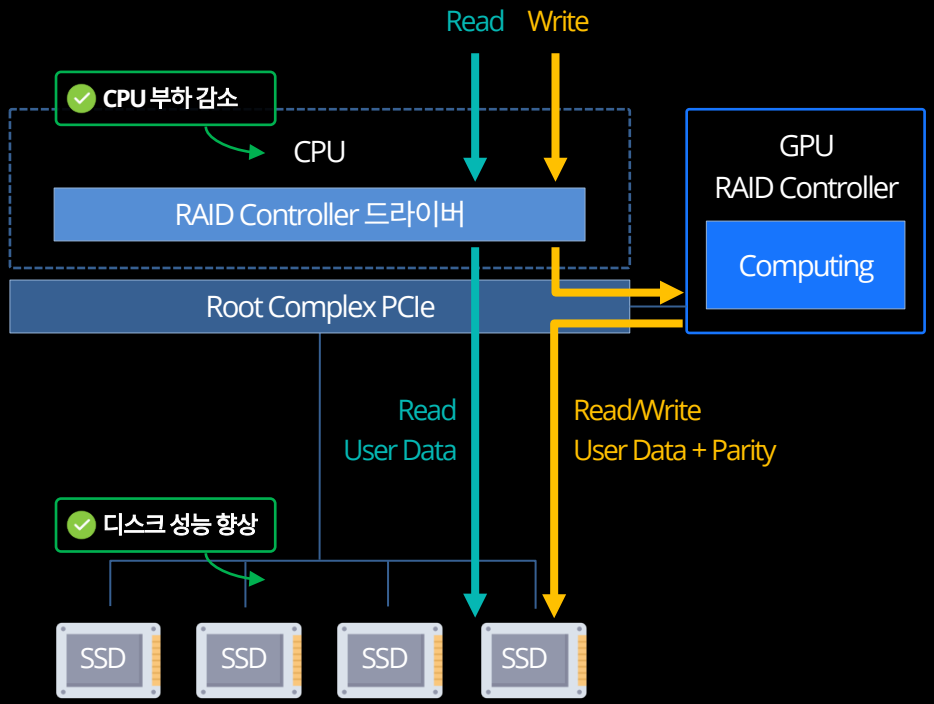
SupremeRAID

- 모델명 : SupremeRAID SR-1010

NVMe/NVMe-oF 위한
고성능 RAID 컨트롤러

- GPU 기반의 PCIe RAID 카드
- 내장된 GPU가 CPU 대신 RAID 입출력을 처리하여 CPU 부하 최소화
- 단일 카드로 최대 1,900만 IOPS 및 220GB/s 성능 제공
- 최대 32개의 NVMe SSD 지원
- 별도의 케이블 없이 PCIe 슬롯에 장착하여 사용

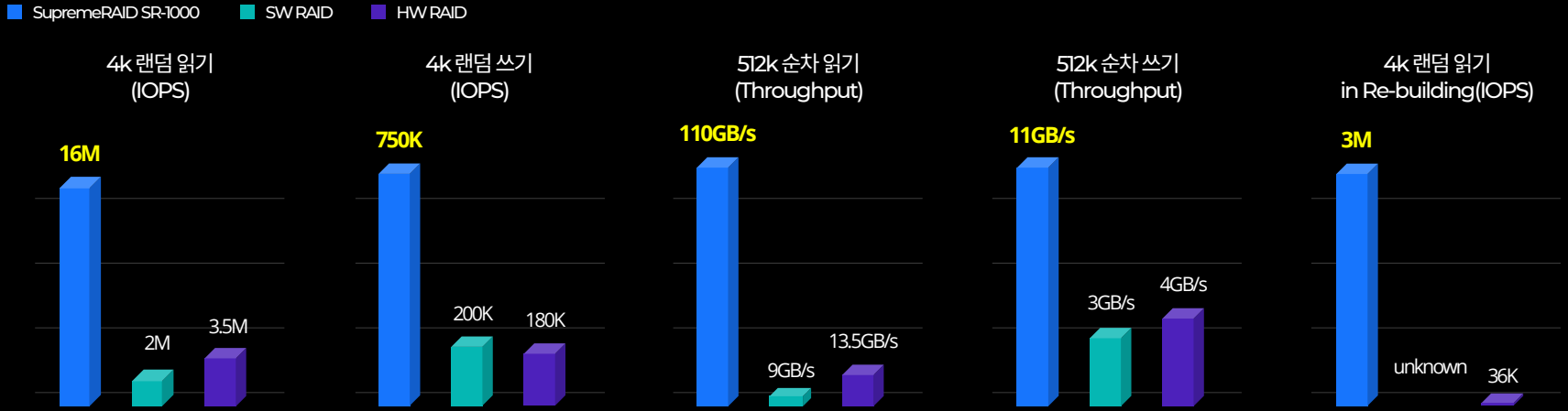
GPU 기반의 SupremeRAID



차세대 RAID 구조

- 패리티 연산을 위한 GPU 기반의 RAID 탑재
- GPU의 초고속 연산 능력으로 패리티 생성에 대한 오버헤드 문제 해결
- CPU 부하 감소로 어플리케이션에 더 많은 CPU 자원 할당 가능
- GPU를 통한 패리티 연산 가속화로 기존 RAID 10 구성 대비 고성능 및 고용량 지원 가능

SupremeRAID vs. SW RAID vs. HW RAID

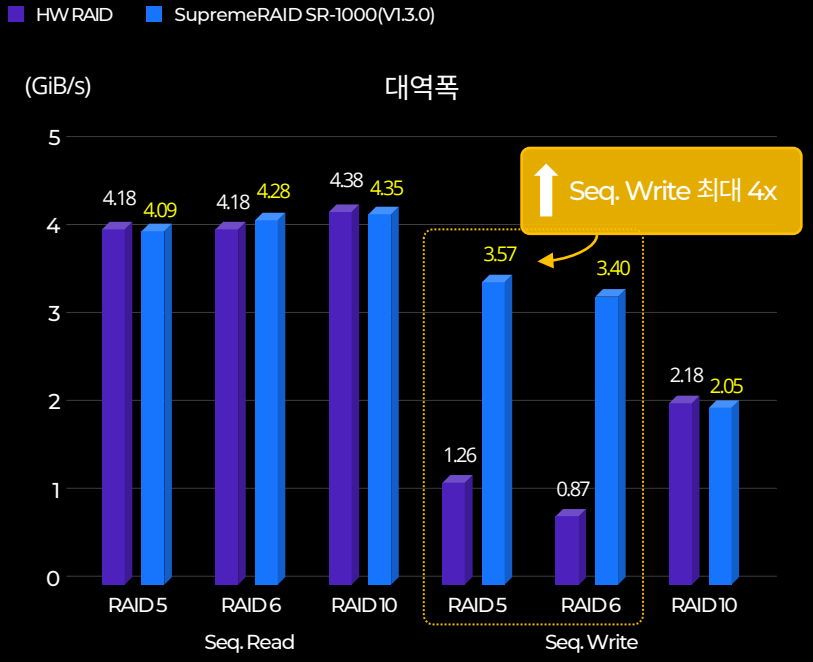


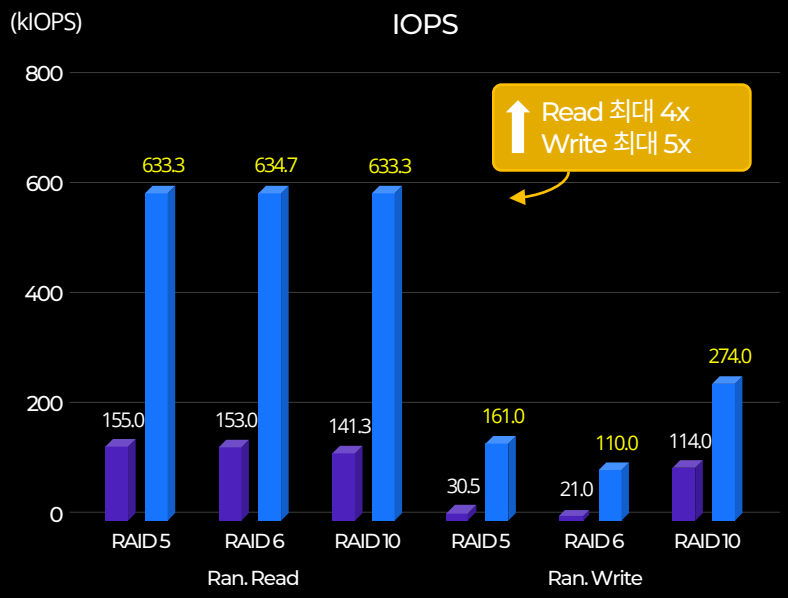
	SupremeRAID™	S/W RAID	H/W RAID
CPU 자원 사용	NONE	High	NONE
지원 RAID 모드	0, 1, 5, 6, 10 and EC	0, 1, 5, 10	0, 1, 5, 6
NVMe-oF 지원	YES	YES	NO
유연성	High	CPU에 따라 제한적	NONE
NVMe SSD 최대 지원 개수	32	24	4

테스트 환경 : Based on RAID5 with Intel Xeon Scalable Platform & Intel D7-P5510 NVMe SSD

자료 출처 : https://blocksandfiles.com/2022/10/12/graids-nuclear-competitive-knockoff/

SupremeRAID vs. HW RAID(1)





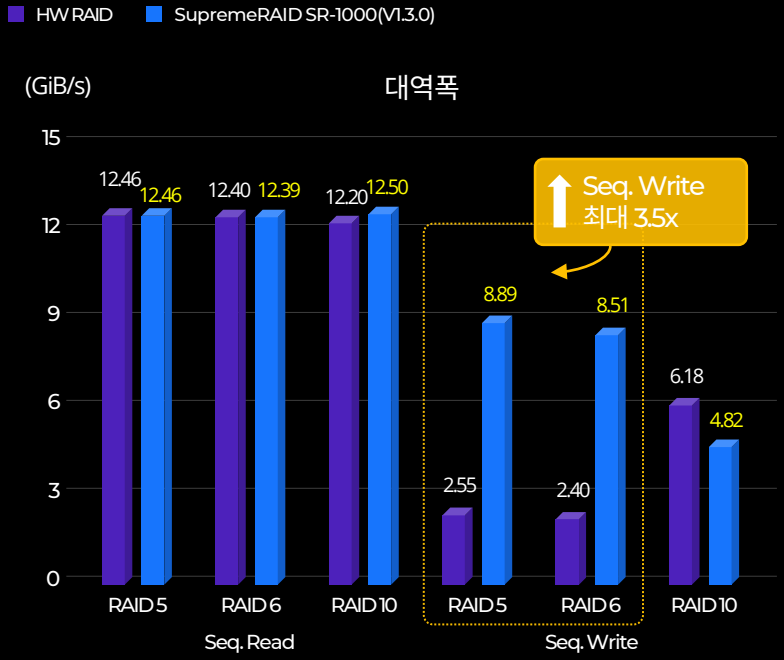
Sequential I/O

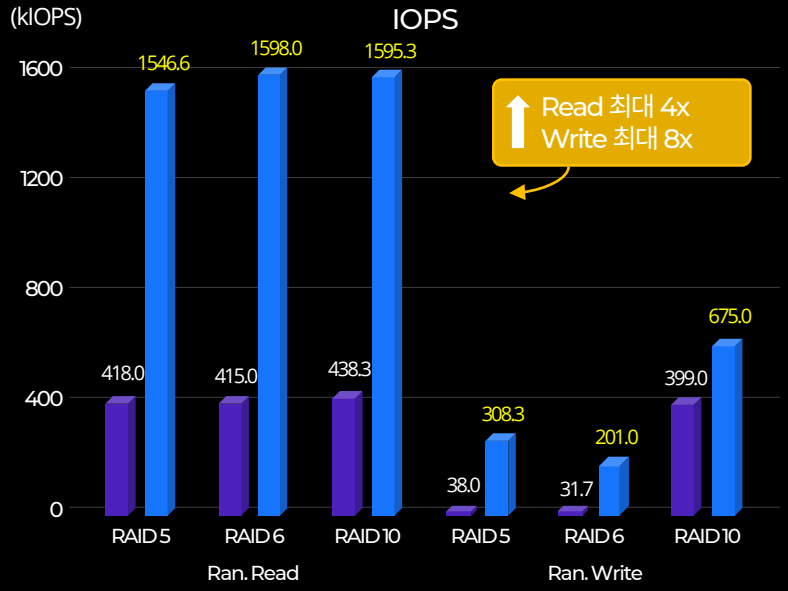
Random I/O

테스트 환경 : Intel Xeon Bronze 3106 CPU @ 1.70GHz * 2EA, PCIe Gen3, Total Memory 128GB, Samsung PM1643a SAS SSD 1.92TB SSD * 12EA, GRAID SupremeRAID SR-1000, AVAGO MegaRAID SAS 9361-8i, CentOS 8.5, Kernel Version 4.18.0-425., fio Version 3.7.2"

테스트 파라메터 : GRAID SupremeRAID CPU Utilization Comparison(SR-1010) Version 1.1.0 참조

SupremeRAID vs. HW RAID(2)





Sequential I/O

Random I/O

테스트 환경 : Intel Xeon Silver 4310CPU @ 2.10GHz * 2EA, PCIe Gen4, Total Memory 128GB, Samsung PM1643a SAS SSD 1.92TB * 12EA, GRAID SupremeRAID SR-1000, SuperMicro AOC-S3916L-H16IR, CentOS 8.5, Kernel Version 4.18.0-425, fio Version 3.7.2"

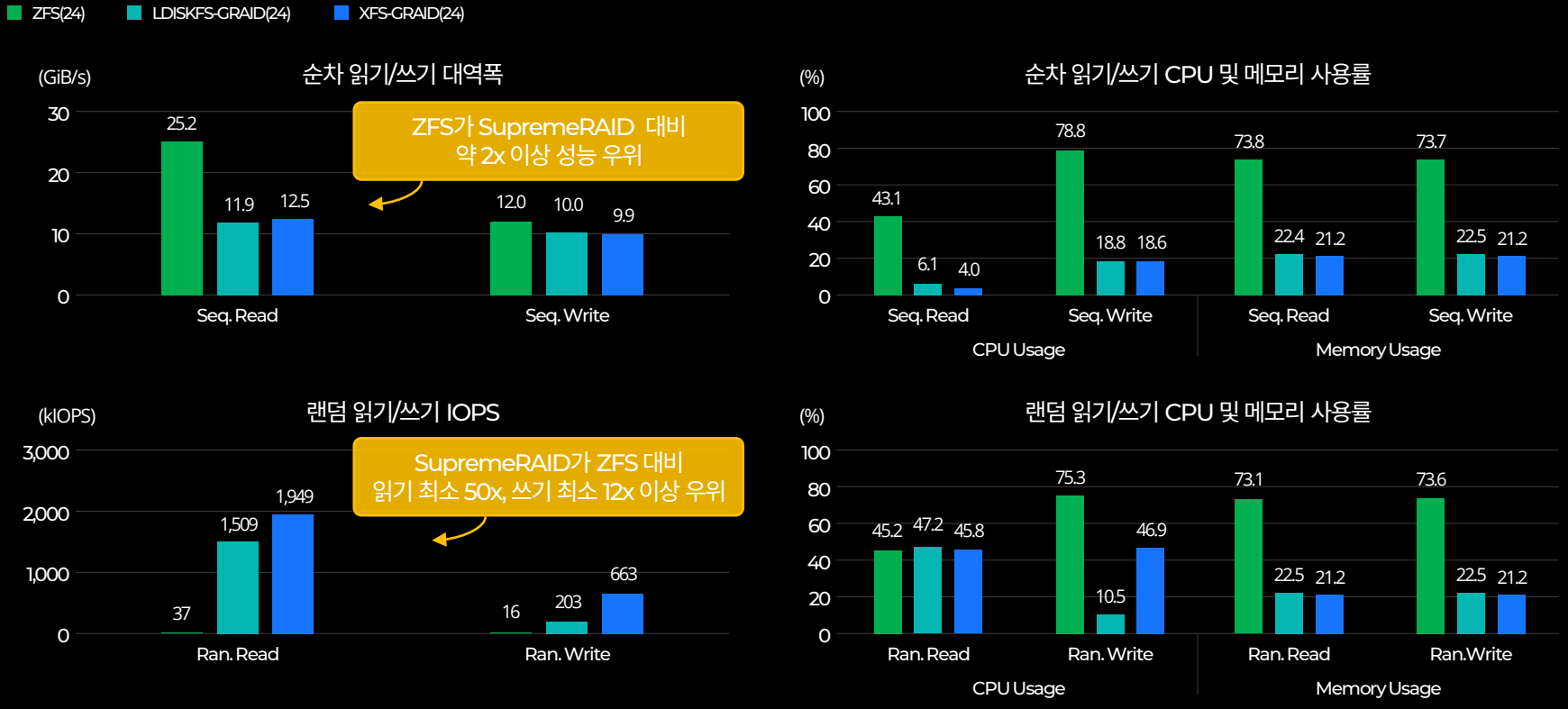
테스트 파라메터 : GRAID SupremeRAID CPU Utilization Comparison(SR-1010) Version 1.1.0 참조

SupremeRAID vs. SW RAID

테스트 환경 : Intel Xeon Gold 6338 CPU @ 2.00GHz * 2EA, Total Memory 256GB, Intel SSD D7-P5510 NVMe SSD * 12EA, GRAID SupremeRAID SR-1010, CentOS 8.5, Kernel Version 4.18.0-348.7.1, fio Version 3.19."

자료 출처 : GRAID SupremeRAID CPU Utilization Comparison(SR-1010) Version 1.1.0

로컬 파일시스템 환경의
SW RAID(ZFS) vs. SupremeRAID

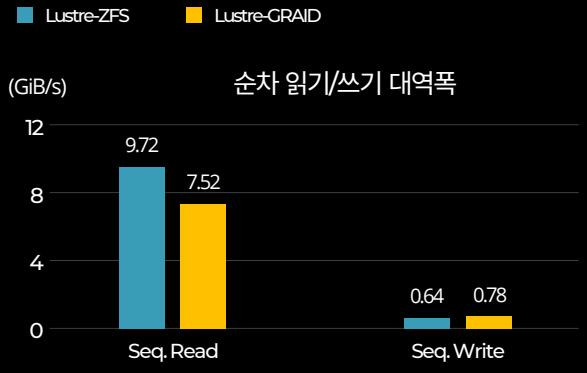


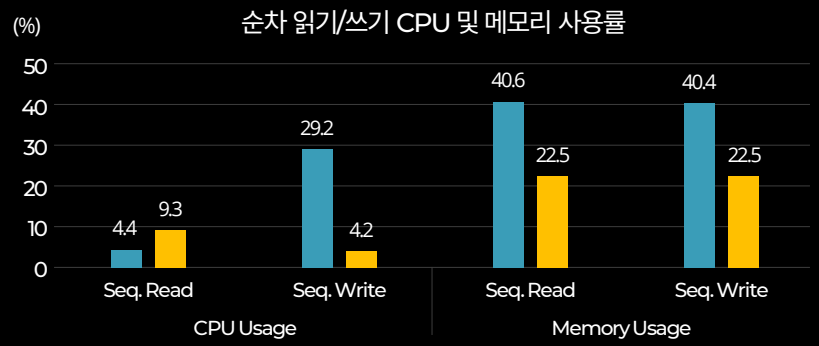
테스트 환경 : Intel Xeon Silver 4210R CPU @ 2.40GHz * 2EA, Total Memory 64GB, Samsung PM9A3 NVMe SSD 3.8TB * 24EA, GRAID SupremeRAID SR-1010, Rocky 8.7, Kernel Version 4.18.0-425.13.1, fio Version 3.19."

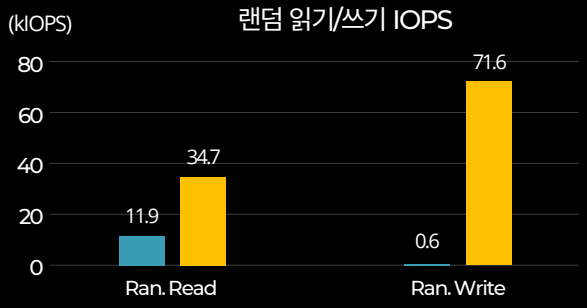
클러스터 파일시스템 환경의
SW RAID(ZFS) vs. SupremeRAID(1)

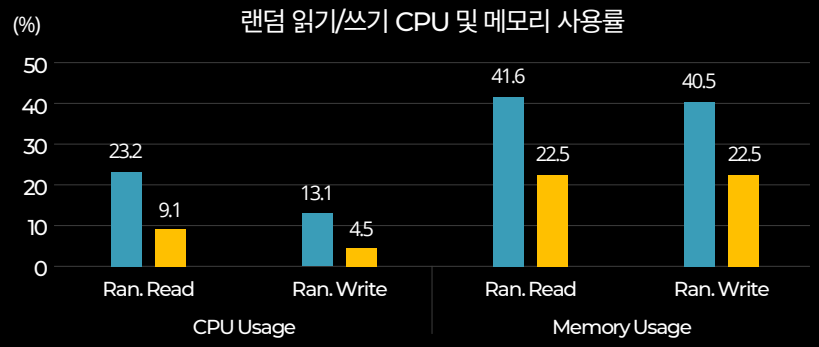
Lustre 타겟 구성

- RAID 5 기반의 단일 스토리지 풀 (NVMe 24개)
- 스토리지 풀에서 3개의 볼륨 생성
- 각 볼륨은 MGS, MDS, OSS의 타겟으로 구성
- 원격 노드에서 Lustre 클라이언트 마운트 후 fio 테스트 수행



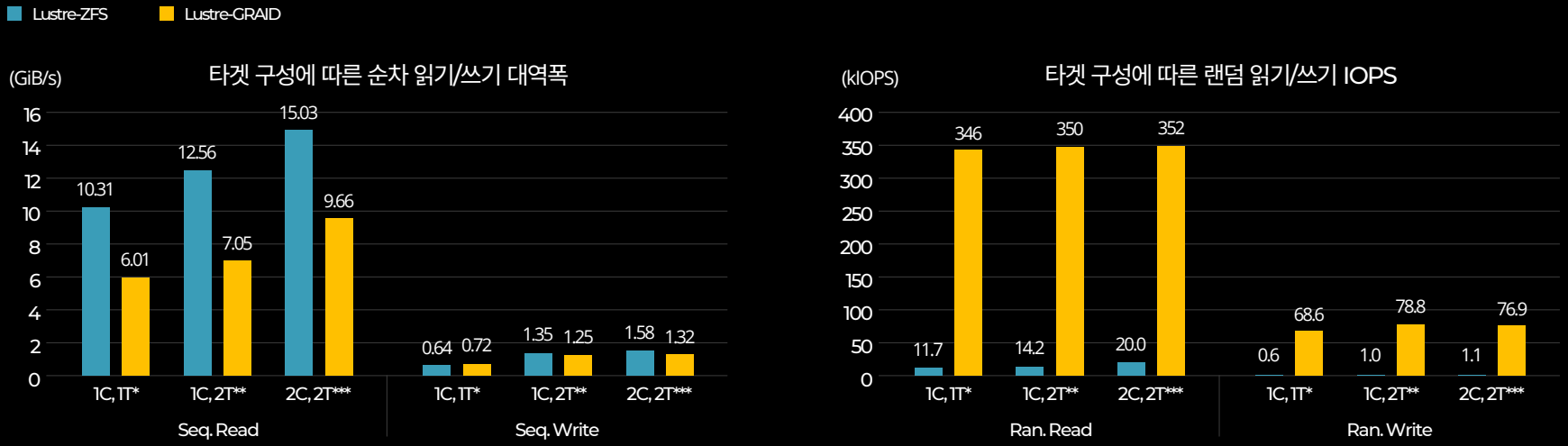






테스트 환경 : Intel Xeon Silver 4210R CPU @ 2.40GHz * 2EA, Total Memory 64GB, Samsung PM9A3 NVMe SSD 3.8TB * 24EA, GRAID SupremeRAID SR-1010, Rocky 8.7, Kernel Version 4.18.0-425.13.1, fio Version 3.19."

클러스터 파일시스템 환경의
SW RAID(ZFS) vs. SupremeRAID(2)



Lustre 타겟 구성

* 1C, 1T : 단일 컨트롤러 환경에서 RAID 5 기반의 단일 타겟 구성 (1 set에 NVMe 24개 구성)

** 1C, 2T : 단일 컨트롤러 환경에서 RAID 5 기반의 타겟 2 set 구성 (1 set 당 NVMe 12개 구성)

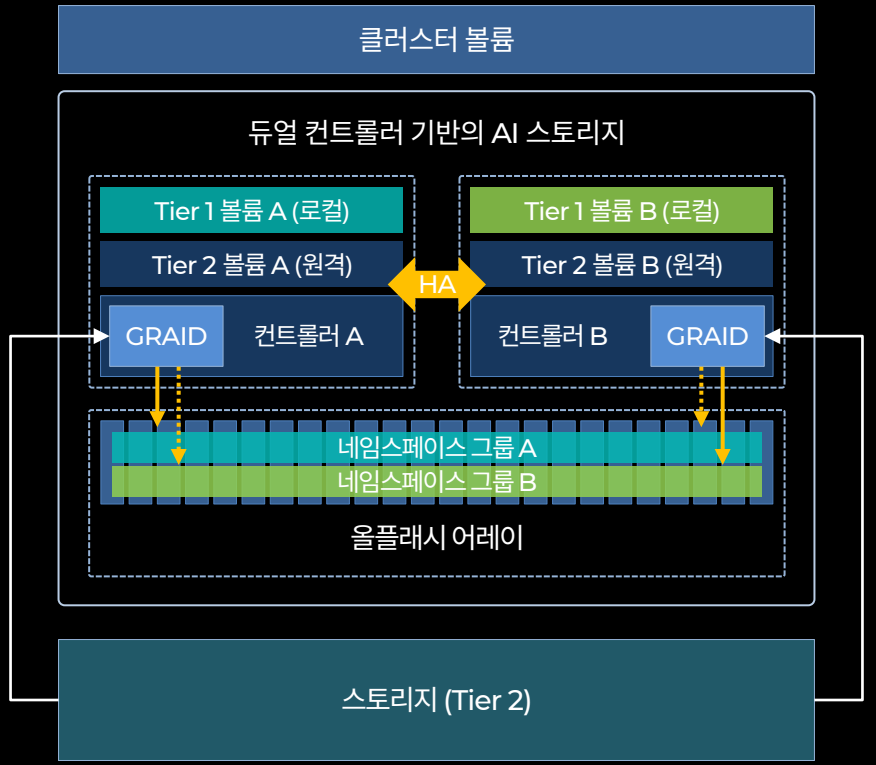
*** 2C, 2T : 듀얼 컨트롤러 환경에서 각 컨트롤러 당 RAID 5 기반의 단일 타겟 구성 (전체 2 set, 1 set 당 NVMe 12개 구성)

- 로컬 노드에서 Lustre 클라이언트 마운트 후 fio 테스트 수행

테스트 환경 : Intel Xeon Silver 4210R CPU @ 2.40GHz * 2EA, Total Memory 64GB, Samsung PM9A3 NVMe SSD 3.8TB * 24EA, GRAID SupremeRAID SR-1010, Rocky 8.7, Kernel Version 4.18.0-425.13.1, fio Version 3.19."

SupremeRAID 기반의 차세대 HA 스토리지

- 특장점
  - ExaStor에 SupremeRAID를 탑재하여 이중화된 RAID 볼륨 제공
  - SW RAID의 한계를 극복, NVMe SSD에서 제공하는 최대 성능 구현
  - SAN 스토리지와의 이중화 연결 지원
- 활용 방안
  - 랜덤 읽기/쓰기가 주로 발생하는 작은 파일의 경우 SupremeRAID와 NVMe SSD 그룹으로 구성된 로컬 볼륨(Tier 1)을 사용
  - 순차 읽기/쓰기 요청이 빈번한 대용량 파일의 경우 ZFS와 SAN 스토리지로 구성된 원격 볼륨(Tier 2)을 사용



경청해 주셔서 감사합니다

글루시스 기술 블로그 : tech.gluesys.com

- 스토리지 이중화 1편 : 고가용성과 이중화
- 스토리지 이중화 2편 : NAS 이중화 아키텍처 설계
- 스토리지 이중화 3편 : 7월 내 업데이트 예정