IT 인프라 아키텍처 설계

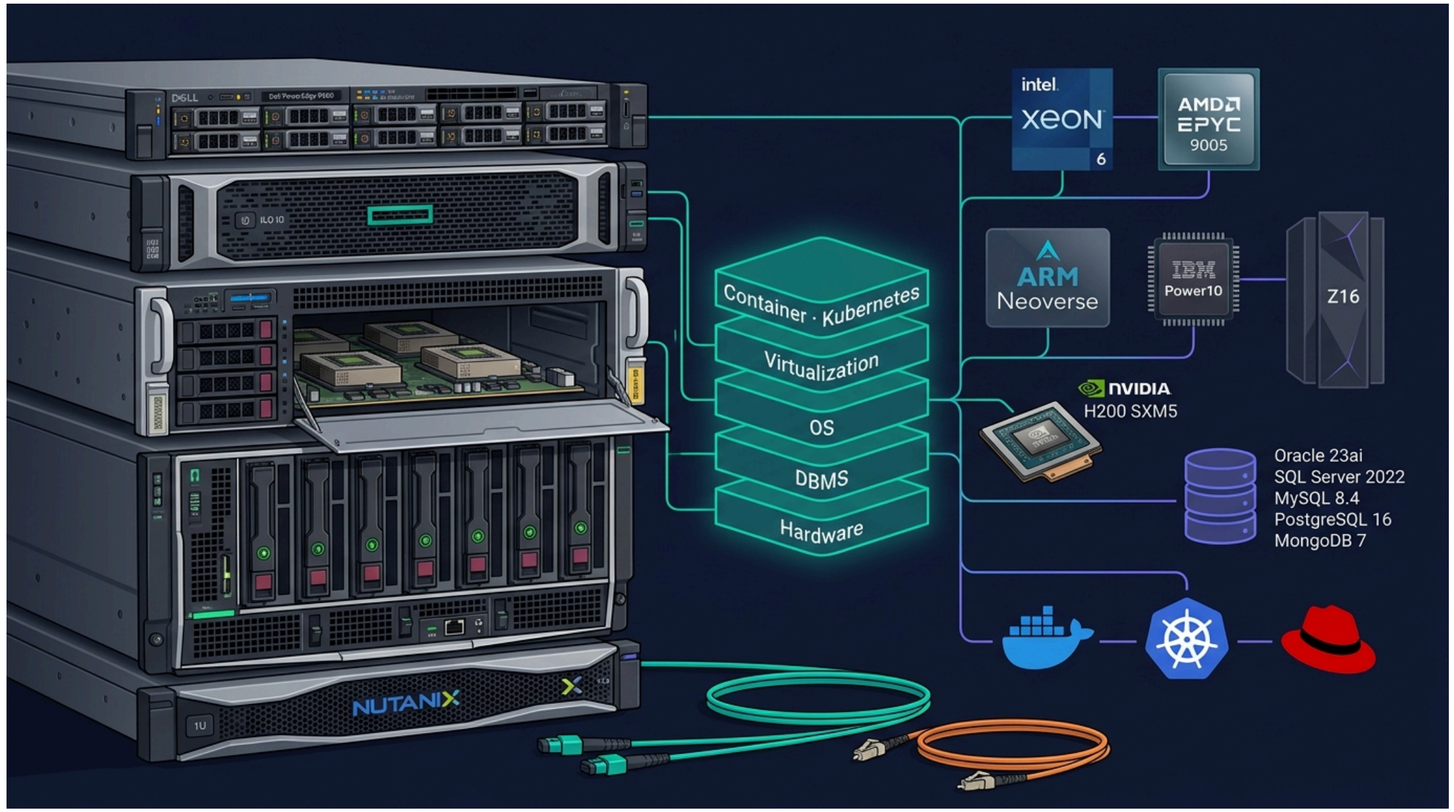
Session 3
서버·OS·DB·GPU·가상화·컨테이너

Day 1 · 오후 — Compute 스택 전체 조망

폼팩터 · CPU · 메모리 · NIC · OS · DBMS · GPU · 가상화 · 컨테이너

강사 박수현 · 🚀 젠아이랩스(GenAI Labs)

데이터센터 · 서버·OS·DB·GPU · 가상화·컨테이너 · 네트워크 L1~L7 · 스토리지 · 백업·DR · 보안 6대 도메인 · HA · 용량·성능·관찰성 · 설계 워크숍 · RFP · 5종 실전 케이스



Day 1 · 오후 — 컴퓨트 스택 전체 조망

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

Session 3 — *회차 메타데이터*

📋 세션 정보

- **회차**: **2 / 15** — Day 1 오후 첫 세션
- **소속**: Day 1 — IT 인프라 큰 그림 + 서버·NW 기초
- **카테고리**: compute (컴퓨트 스택)
- **예상 분량**: 본문 103 슬라이드 + 이미지 18장
- **강의 시간**: 약 120분 (휴식 포함)
- **강사**: 박수현 — 젠아이랩스(GenAI Labs) **CEO / CTO**

🎯 핵심 메시지

> 컴퓨트 자산 전체 — 서버·OS·DB·GPU·가상화·컨테이너 — 를 **하나의 산정·결정 체계**로 묶는다. 라이센스가 모든 결정의 숨은 축이다.

🧭 학습 목표

1. **폼팩터** — Tower·Rack·Blade·HCI·OCP 6종 비교
2. **CPU** — x86·ARM·Power·Mainframe 4대 매핑
3. **메모리·NIC** — DDR4/5·NVMe·SmartNIC·RDMA 결정 포인트
4. **OS** — RHEL·Windows Server·UNIX·라이센스 비교
5. **DBMS** — RDBMS·NoSQL·NewSQL·시계열·검색·벡터 6대 분류
6. **GPU** — L4 → H200/B200·MIG·NVLink·InfiniBand 산정
7. **가상화·컨테이너** — Type-1 6종·K8s 아키텍처·Broadcom 이슈
8. **산출물** — 서버 BOM·OS·DB·가상화 표준안 작성

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

PART A

서버 폼팩터 — *Tower부터 OCP까지*

> "어떤 모양의 서버를 도입할 것인가" — 모든 컴퓨트 결정의 출발점.

서버 폼팩터 6대 — *한눈에 비교*

🏗️ Tower

- PC와 동일한 수직형
- 지점·소규모·연구실
- 집적도 낮음 → DC 부적합
- 가격 저렴·소음 큼

🧱 Rack 1U/2U/4U/6U

- **표준 DC 폼팩터**
- 1U=웹·LB, 2U=DB·범용
- 4U=GPU·고밀도 DB
- 6U+=특수 (HPC·HCI)

🗡️ Blade

- 섀시 + 카드형 서버
- 전원·NW 공유 → 고밀도
- 벤더 종속성 강함
- HPE Synergy·Cisco UCS·Dell PowerEdge MX

📦 HCI Appliance

- 컴퓨트+스토리지+NW 일체형
- Nutanix·VxRail·SimpliVity·HCX
- VMware vSAN 통합
- 빠른 구축·고가 운영

🧩 Composable

- 자원을 풀(Pool)로 추상화
- HPE Synergy·Dell PowerEdge MX
- API로 컴퓨트·스토리지·NW 조합
- 고난도 운영·유연성 최고

🌍 OCP (Open Compute)

- Facebook 발 오픈 표준
- 21" 랙·12V DC 버스
- Hyperscale (Meta·MS·Google) 중심
- 한국 국내 보급 낮음

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

6대 폼팩터 비교 (삽화)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

Rack U 비교 — *1U vs 2U vs 4U vs 6U+*

U	두께	디스크 베이	PCIe 슬롯	전형적 워크로드	대표 모델
1U	44 mm	4~10 SFF	2~3	웹·API·LB·Edge	Dell R660·HPE DL360 Gen11·Lenovo SR630 V3
2U	88 mm	12~24 SFF / 24+ NVMe	6~8	DB·범용 VM·미들웨어	Dell R760·HPE DL380 Gen11·SR650 V3
4U	175 mm	24~36 LFF / 8 GPU	10+	GPU·고밀도 DB·스토리지 노드	Dell XE9680·HPE DL580·SR670 V2
6U+	265 mm+	특수	다수	HGX 8GPU·HPC·NVL72 컴포넌트	NVIDIA DGX H200·HPE Cray

선택 규칙: 디스크·GPU·PCIe 카드 수가 많을수록 U가 커진다. **랙 한 칸당 발열·전력 부하**도 비례 — 4U H100 1대 = 700W × 8개 + CPU = 약 6~10 kW.

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

Blade · HCI · Composable — *고밀도 3종 비교*

🗡️ Blade

구성 - 섀시 (6U~10U) + 블레이드 카드 8~16개 - 공유 NW 인터커넥트 + 공유 PSU - 섀시 단위 KVM·관리

대표 솔루션 - HPE Synergy / BladeSystem c-Class - Cisco UCS - Dell PowerEdge MX

장단점 - ✅ 고밀도·케이블 절감 - ❌ 섀시 단위 발주·벤더 종속 - ❌ EOL 시 섀시 전체 교체

📦 HCI Appliance

구성 - 컴퓨트+스토리지+가상화 일체 - 노드 3대부터 시작 (3-node minimum) - 분산 스토리지 (vSAN·Nutanix DFS)

대표 솔루션 - VMware vSAN ReadyNode·VxRail - Nutanix NX·AHV - HPE SimpliVity·Cisco HyperFlex

장단점 - ✅ 빠른 구축·통합 관리 - ❌ 라이센스 비용 큼 - ❌ 노드 단위 스케일링 (작은 단위 어려움)

🧩 Composable

구성 - 자원을 **풀(Pool)** 로 분리 - 컴퓨트·스토리지·NIC을 API로 조합 - 워크로드별 동적 재구성

대표 솔루션 - HPE Synergy + OneView - Dell PowerEdge MX + OpenManage - Liqid Composable

장단점 - ✅ 자원 활용률 최고 - ❌ 운영 난이도 최고 - ❌ 국내 도입 사례 적음

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

폼팩터 결정 매트릭스 — *워크로드별*

워크로드	권장 폼팩터	이유
소규모 지점 (1~2대)	Tower or 1U Rack	운영 단순·발열 적음
웹·API·LB (수십 대)	1U Rack	1U당 32C+, 다수 운영 효율
DB·미들웨어 (HA)	2U Rack	디스크 베이·메모리 슬롯 풍부
VMware/Hyper-V 가상화 팜	2U Rack or HCI	HCI는 빠른 구축, Rack은 유연
VDI (수백 사용자)	HCI Appliance	통합 관리·노드 확장 단순
AI 학습 (대형)	4U GPU + DGX/HGX 8GPU	NVLink·InfiniBand 필수
AI 추론 (운영)	1U·2U GPU (L4·L40S)	저전력·고밀도
공공 클라우드 자가구축	OCP or 표준 2U Rack	표준화·자동화 우선
HPC 클러스터	1U·2U HPC 노드	고밀도 CPU·InfiniBand
백업·아카이브	4U Storage (24~36 LFF)	디스크 베이 극대화

> **현장 기본값**: 80% 이상은 **2U Rack** — 가장 안전한 디폴트.

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

PART B

CPU 아키텍처 — *x86 · ARM · Power · Mainframe*

> 칩 선택은 OS·DB·라이센스·성능을 동시에 결정한다.

CPU 4대 아키텍처 — *시장 점유율·용도*

🔷 x86 (Intel · AMD)

- **시장 점유율 80%+**
- Intel Xeon Scalable Gen 4/5 (Sapphire/Emerald Rapids)
- Intel Xeon 6 (Granite Rapids·Sierra Forest, 2024~)
- AMD EPYC Genoa (9004)·Bergamo·Turin (9005, 2024)
- 범용 워크로드 사실상 표준

🔶 ARM 서버

- **급성장 영역 (2022~)**
- Ampere AmpereOne (192C, 2024)
- AWS Graviton 3/4 (퍼블릭만)
- NVIDIA Grace (Grace-Hopper, Grace-Blackwell)
- 전성비 우수·국내 도입 초기

🟣 Power (IBM)

- 시장 점유율 ↓ but 금융·공공 잔존
- IBM Power9 (2017) · Power10 (2021)
- AIX·IBM i·z/Linux on Power
- 메인 사용처: 코어뱅킹·보험·일부 공공

🟡 Mainframe (IBM Z)

- **국내 30~40개 사이트 잔존** (금융·공공)
- IBM Z16 (2022) · z/OS·z/Linux
- 초고가용성 (>99.9999%)
- COBOL·CICS·DB2 z/OS

○ SPARC (Oracle·Fujitsu)

- Oracle SPARC M8 (2017이 마지막)
- Solaris OS·점진적 EOL
- **신규 도입 거의 없음** — 운영 잔존만

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

CPU 4대 아키텍처 비교 (삽화)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

Intel Xeon Scalable — *Gen 4 / 5 / 6*

세대	코드명	출시	최대 코어	메모리	PCIe	비고
Gen 4 SP	Sapphire Rapids	2023.01	60C	DDR5-4800 (8ch)	PCIe 5.0	첫 DDR5·CXL 1.1
Gen 5 SP	Emerald Rapids	2023.12	64C	DDR5-5600 (8ch)	PCIe 5.0	Cache 1.5× 증가
Xeon 6 (P-Core)	Granite Rapids	2024.09	**128C**	DDR5-6400·MRDIMM	PCIe 5.0+CXL 2.0	성능 중심
Xeon 6 (E-Core)	Sierra Forest	2024.06	**288C**	DDR5-6400	PCIe 5.0+CXL 2.0	전성비·클라우드

🎯 워크로드 매핑

- **DB / 미들웨어 / VM**: Xeon 6 P-Core
- **웹 / API / 컨테이너 팜**: Xeon 6 E-Core (Sierra Forest)
- **HPC / AI 추론**: Xeon 6 P-Core + AVX-512·AMX
- **신규 RFP**: 2026년 기준 Xeon 6 권장

💰 라이센스 영향

- Oracle DB Core Factor — Intel x86 = **0.5** (코어 1개 = 0.5 라이센스)
- VMware vSphere — Per-Core (2024~)
- Windows Server DataCenter — Per-Core (16C 기본 + 추가)
- **고코어 = 라이센스 폭증** → 작은 SKU 여럿이 유리할 때 있음

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

AMD EPYC — *Genoa / Bergamo / Turin*

세대	코드명	출시	최대 코어	메모리	PCIe	비고
EPYC 9004	Genoa	2022.11	96C	DDR5-4800 (12ch)	PCIe 5.0	첫 12채널·DDR5
EPYC 9004C	Bergamo	2023.06	**128C** (Zen 4c)	DDR5-4800	PCIe 5.0	클라우드·고밀도
EPYC 8004	Siena	2023.09	64C	DDR5-4800 (6ch)	PCIe 5.0	Edge·SMB
EPYC 9005	Turin	2024.10	**192C** (Zen 5c)	DDR5-6000 (12ch)	PCIe 5.0	신규 RFP 추천

AMD의 강점: **코어 수 + 메모리 채널 (12ch)** 모두 Intel 대비 우위. 가상화 호스트·DB·HPC에서 점유율 급성장. **2024년 기준 신규 서버 시장 점유율 약 30%+**.

✅ AMD 선택 이유

- 코어 수 더 많음 (192C vs 128C)
- DDR5 12채널 (Intel 8채널)
- 가상화 밀도 우수
- 라이센스 효율 (코어당 성능)
- Genoa부터 가격 경쟁력

❌ Intel 잔존 이유

- AVX-512 / AMX (AI 가속)
- QAT (암호화 가속)
- Optane (단종이나 잔존)
- 호환성 (레거시 SW)
- 벤더 표준 (HPE·Dell 표준 라인 풍부)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

ARM 서버 — *Ampere · Graviton · Grace*

🌐 Ampere

- **AmpereOne (2024)** — 192C, DDR5-5600
- **Altra Max** — 128C (전세대)
- Single Thread per Core (예측성)
- Oracle Cloud·Azure·국내 도입 시작
- 국내 사례: 네이버 클라우드 일부

☁️ AWS Graviton

- **Graviton 3 / 3E** — 64C, Neoverse V1
- **Graviton 4 (2024)** — 96C, Neoverse V2
- **AWS 전용** — 직접 구매 불가
- AWS EC2 c7g/m7g/r7g 인스턴스
- 동급 x86 대비 **40% 전성비**

🚀 NVIDIA Grace

- **Grace (CPU only)** — 72C ARM Neoverse V2
- **Grace Hopper (GH200)** — Grace + H100
- **Grace Blackwell (GB200)** — Grace + B200 × 2
- NVLink C2C로 CPU↔GPU 900GB/s
- 대형 AI 학습 클러스터 (NVL72)

ARM 서버의 한국 진입: 클라우드(NCP·KT·NHN) 일부 채택. **온프레미스는 도입 초기**. 호환성 (DB 벤더·미들웨어)·운영 도구·인증서 문제로 신중. **2026~27이 본격 확산 시기**로 전망.

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

Power · Mainframe · SPARC — *레거시지만 살아있는*

🟣 IBM Power

- **Power10 (2021)** — 15C, SMT-8 (코어당 8 스레드)
- **AIX 7.3** · IBM i 7.5 · z/Linux on Power
- 코어 라이센스 (Oracle Factor = **1.0**, x86의 2배)
- 주요 사용처
- 보험사 코어 시스템 (AIX + DB2)
- 일부 공공 (IBM i 차세대)
- 통신사 BSS/OSS
- EOL 압박 ↑ but 마이그레이션 부담 ↑↑

🟡 IBM Mainframe (Z)

- **IBM Z16 (2022)** — Telum 칩 (8C × 4 = 32C)
- z/OS · z/VM · z/Linux
- **국내 30~40개 사이트** (대형 은행·신용카드·증권·생명보험·공공)
- 가용성 >99.9999% (다운타임 분단위/년)
- COBOL·CICS·DB2 z/OS·IMS
- **인력 노령화** — 30~40대 신규 인력 거의 없음
- 차세대 시 X86 다운사이징 또는 z/Linux 전환

○ Oracle SPARC

- **SPARC M8 (2017이 마지막)** — 32C, SMT-8
- Solaris 11.4 — 2034 지원
- **신규 도입 사실상 종료**
- 잔존 사이트
- 일부 통신사·공공
- Oracle ULA 묶음 잔존
- 차세대 = x86 Linux + Oracle 마이그레이션

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

CPU 결정 매트릭스 — *워크로드별*

워크로드	권장 CPU	이유
웹·API·LB·MSA 컨테이너	Intel Xeon 6 E-Core / AMD EPYC Bergamo	고밀도·전성비
VMware/KVM 가상화 호스트	AMD EPYC Genoa/Turin	코어·메모리 채널
DB OLTP (RDBMS)	Intel Xeon 6 P-Core (낮은 코어·고클럭)	라이센스 효율·단일 스레드
DB OLAP·HTAP	AMD EPYC (다수 코어)	병렬 쿼리·메모리 대역
AI 학습 (GPU 호스트)	AMD EPYC 9474F (32C 고클럭) or NVIDIA Grace	PCIe 5.0·메모리 대역
AI 추론 (CPU 가속)	Intel Xeon 6 P-Core (AMX)	AMX·AVX-512
HPC (Fluid·CFD·해석)	AMD EPYC + InfiniBand	코어·메모리 대역
클라우드 자가 구축	AMD EPYC + ARM 혼재	전성비·다양성
공공·금융 코어	x86 Intel (보수적)	검증·호환
레거시 AIX·z/OS	Power10·Z16 유지	마이그레이션 부담

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

CPU 벤치마크 — *대표 지표 · 수치 예시 · 워크로드 매핑*

📊 대표 서버 CPU 벤치마크

- **SPECint 2017** — 정수 연산
- **SPECfp 2017** — 부동소수점
- **SPECrate 2017** — 서버 처리량(Throughput)
- **SPECjbb 2015** — Java 서버 워크로드
- **CoreMark** — 임베디드 CPU 성능
- **TPC-C** — OLTP 시스템 성능
- **TPC-H** — OLAP 시스템 성능
- **STREAM** — 메모리 대역폭
- **LINPACK** — HPC 부동소수점
- **MLPerf** — AI 학습·추론

🔢 수치 예시 *(벤더 공개 결과 기준)*

- Xeon 6 P 128C · SPECrate2017_int ≈ 1,800
- EPYC 9755 (128C) ≈ 2,100
- Graviton 4 96C ≈ 1,400

> base/peak·제출시기·컴파일러·SMT 여부에 따라 수백 점 차이.

🎙️ 실무 활용 — *벤치마크별 워크로드 매핑*

- **SPECrate** → 가상화 호스트
- **SPECjbb** → WAS·Java 서비스
- **TPC-C** → 금융·주문·ERP DB
- **TPC-H** → DW·BI·OLAP
- **STREAM** → 인메모리 DB·HPC·SAP HANA
- **LINPACK** → 과학기술 계산
- **MLPerf** → AI 학습·추론

⚠️ 벤치마크 해석 주의

- 벤치마크는 **상대 비교용**
- 실제 서비스와 차이 존재
- 벤더 결과는 최적 튜닝 환경
- 최종 판단은 **PoC + 운영 메트릭** 기반

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

CPU 산정 — *라이센스가 SKU 를 결정한다*

> Compute 회차에서 CPU 의 마지막 화두는 **라이센스**. 정량 산정 공식·벤치마크·가상화 오버헤드는 운영 회차로 넘기고, 여기서는 **"왜 코어 수가 적은 SKU 가 답일 때가 많은가"** 한 페이지로 마무리한다.

💰 Per-Core 라이센스 SKU 표

제품	모델	단위	한국 현장 영향
Oracle DB EE	Per-Core (계수 0.5 x86)	1코어 당 라이센스	코어 1개 증가 = 수천만 원
MS SQL EE	Per-Core (최소 4코어)	2코어 팩	OLTP DB 표준 비용
VMware vCF / vSphere	Per-Core (최소 16코어/CPU)	코어팩	2024 SKU 개편 후 급등
Red Hat OpenShift	Per-Core 또는 소켓	16코어 묶음	K8s 도입 비용 핵심
WebLogic·WAS EE	Per-Core 또는 NUP	코어 묶음	WAS 라이센스 비용

🧮 한 줄 결정 룰

> **같은 성능이면 코어 수가 적고 클럭이 높은 SKU** — 예: Xeon Gold 6444Y **16C @ 4.0 GHz** vs Xeon Platinum 8462Y+ 32C @ 2.8 GHz. Oracle EE 라이센스로 환산 시 **수억 원 단위** 차이.

📚 정량 산정 본문은 Session 8 PART B

- **TPS·동접·세션** 산정 — `s8 ## TPS 산정`
- **CPU 산정 공식** SPECint·TPC-C·TPS/Core·트랜잭션 CPU ms — `s8 ## CPU 산정`
- **가상화/컨테이너/암호화 오버헤드** 가산율 — `s8 ## CPU 산정`
- **6단계 사이징 절차** + 라이센스 영향 ⑤ 단계 — `s8 ## CPU 산정`
- **Memory·IOPS·네트워크·스토리지** 산정 — `s8 ## Memory/IOPS/...`
- **산정 사례** LMS 10K 동접·학사 행정 500 동접 — `s8 ## 산정 사례`

🎯 Compute 회차 정리

- CPU 종류·세대·코어 수 → **CPU 결정 매트릭스** (앞 슬라이드)
- 벤치마크 지표 → **CPU 벤치마크** (앞 슬라이드)
- **라이센스 → SKU** → 이 슬라이드 (Per-Core 비용 인지)
- 정량 산정 공식 → Session 8

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

PART C

메모리 · 로컬 스토리지 · NIC

> 서버 한 대의 BOM에서 CPU 다음으로 큰 변수.

메모리 비교 — *DDR4 / DDR5 / MRDIMM / PMem / CXL*

표준	출시	속도 (MT/s)	채널	용량 (DIMM당)	비고
DDR4	2014	2133~3200	플랫폼별 4~8	16~64GB	레거시 주력
DDR5	2021	4800~8800+	8~12	16~256GB	현재 서버 표준
DDR5 MRDIMM	2024	8800~12800+	8~12	256GB+	**Xeon 6 중심**·초고대역
Persistent Memory	2018	DDR4 호환	DIMM 슬롯	128/256/512GB	Intel Optane PMem — **현재 단종**
CXL Memory	2023~	PCIe 5.0/6.0	(PCIe lane)	256GB~수TB	메모리 확장(현재)·풀링(미래)

🔍 DDR5 핵심

- 듀얼 서브채널 — **32-bit + ECC × 2** (실질 40-bit ×2)
- **PMIC on DIMM** — 전원 안정성 ↑
- **온다이 ECC** + 시스템 ECC
- 8채널 (Intel) / 12채널 (AMD)
- DIMM 256GB → 노드당 **수 TB ~ 수십 TB** 구성 가능

📝 채널 수는 CPU 플랫폼 의존

- Xeon E5 v4 → 4채널
- Xeon Scalable Gen1/2 → 6채널
- EPYC Naples/Rome → 8채널
- Xeon 6 / EPYC Turin → 8~12채널

🚀 CXL Memory

- **Compute Express Link** — PCIe 위 메모리 프로토콜
- **1.x** = Device Attach · **2.0** = Switch · **3.0** = Fabric
- **현재**: 메모리 확장 (Memory Expansion) — 압도적 다수
- **미래**: 메모리 풀링 (여러 서버 공유) — 아직 제한적
- Xeon Sapphire Rapids·Genoa 부터 존재
- **2024~25 Xeon 6 / EPYC Turin 본격 지원**
- **2026~27 본격 확산** 전망 — Dell·HPE·Lenovo·Supermicro 출시

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

ECC · RDIMM · LRDIMM — *서버 메모리 필수 용어*

🛡️ ECC (Error Correcting Code)

- **서버 표준 (필수)**
- 1비트 오류 자동 정정 (SECDED)
- 2비트 검출
- **Advanced ECC** — Chipkill·SDDC·ADDDC
- DDR5는 온다이 ECC 추가 (RAS 강화)

📦 UDIMM (Unbuffered)

- 일반 PC용, 서버에는 사용 ×

📦 RDIMM (Registered)

- 칩과 메모리 컨트롤러 사이 **Register**
- 부하 분산 → 다수 DIMM 가능
- 서버 표준
- 슬롯당 64/128GB

📦 LRDIMM (Load-Reduced)

- Register + **Buffer** 추가
- 더 큰 용량 (DIMM당 256GB+)
- 약간의 지연 증가
- 대용량 메모리 서버에 필수

📐 채널 균형 (Channel Balance)

- 8채널이면 **8 / 16 / 24 DIMM** (8의 배수)
- 채널 불균형 시 **대역폭 크게 손실**
- 12채널 EPYC는 12 / 24 DIMM
- 예: 2S × 12채널 = 24 DIMM 모두 채워야 최대

🎙️ TA 메모

- DIMM 수 적게 + 용량 큰 SKU = 미래 확장 ↓
- DIMM 수 많이 + 용량 작은 SKU = 확장 ↑·발열↑

로컬 스토리지 — *NVMe · SATA · SAS · EDSFF*

인터페이스	대역폭	폼팩터	전형적 용도
SATA III	6 Gbps	2.5" / 3.5"	백업·아카이브·OS 부트
SAS 12G	12 Gbps	2.5" SFF / 3.5" LFF	엔터프라이즈 HDD/SSD·RAID
SAS 24G	24 Gbps	2.5" SFF	고성능 SAS SSD
NVMe (PCIe Gen 4)	32 GT/s × 4 = ~8 GB/s	U.2 / U.3 / M.2	DB·캐시·범용 SSD
NVMe (PCIe Gen 5)	64 GT/s × 4 = ~16 GB/s	U.2 / U.3 / E1.S / E3.S	신규 표준 (2024+)
EDSFF E1.S	PCIe Gen 4/5	5.9 / 9.5 / 15 / 25 mm	1U 고밀도 (32 베이)
EDSFF E3.S	PCIe Gen 4/5	7.5 / 16.8 mm	2U 고밀도·고용량

🔍 폼팩터 트렌드

- **2.5" SFF** — 여전히 표준이나 NVMe Gen5에서 한계
- **U.2 / U.3** — 핫스왑 NVMe 표준
- **M.2** — 부트 디스크·소형 캐시
- **E1.S** — 1U당 32개 SSD 가능 (서버 디자인 변혁)

🎙️ TA 결정 포인트

- **OS 부트** — M.2 × 2 (RAID-1) 별도
- **OLTP DB** — NVMe Gen4/5 (P-state·DWPD 3+)
- **백업·아카이브** — SATA 7.2K HDD or SAS Nearline
- **범용 VM** — SAS SSD or NVMe Gen4 (균형)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

NIC 속도 — *1G에서 800G까지*

속도	출시	매체	사용처
1 GbE	1999	Cat 5e / 6 UTP	관리망·소규모·iLO
10 GbE	2002	Cat 6a / SFP+ DAC / 광	표준 서비스망 (퇴조)
25 GbE	2016	SFP28	**현재 표준 서비스망**
40 GbE	2010	QSFP+	데이터센터 백본 (점퇴)
100 GbE	2010~	QSFP28	**현재 백본 표준**
200 GbE	2018	QSFP56	고대역 백본
400 GbE	2017 (실용 2022)	QSFP-DD / OSFP	Spine·AI 클러스터
800 GbE	2022 (실용 2024)	OSFP / QSFP-DD800	AI 클러스터 (NVL72·HGX)

한국 IDC 현재 표준 (2026): 서버 NIC = **25 GbE**, 백본 = **100 GbE**, AI 클러스터 = **400/800 GbE**. 10 GbE는 운영 잔존만, 신규는 25 GbE 디폴트.

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

NIC ↔ 광 모듈 한 줄, 본문은 Session 4

> **서버 NIC 25/100 GbE = SFP28 / QSFP28 슬롯**. NIC 속도가 정해지면 광 모듈·케이블 SKU 가 자동 결정 — RFP/BOM 에 NIC 와 모듈을 묶어 명세한다.

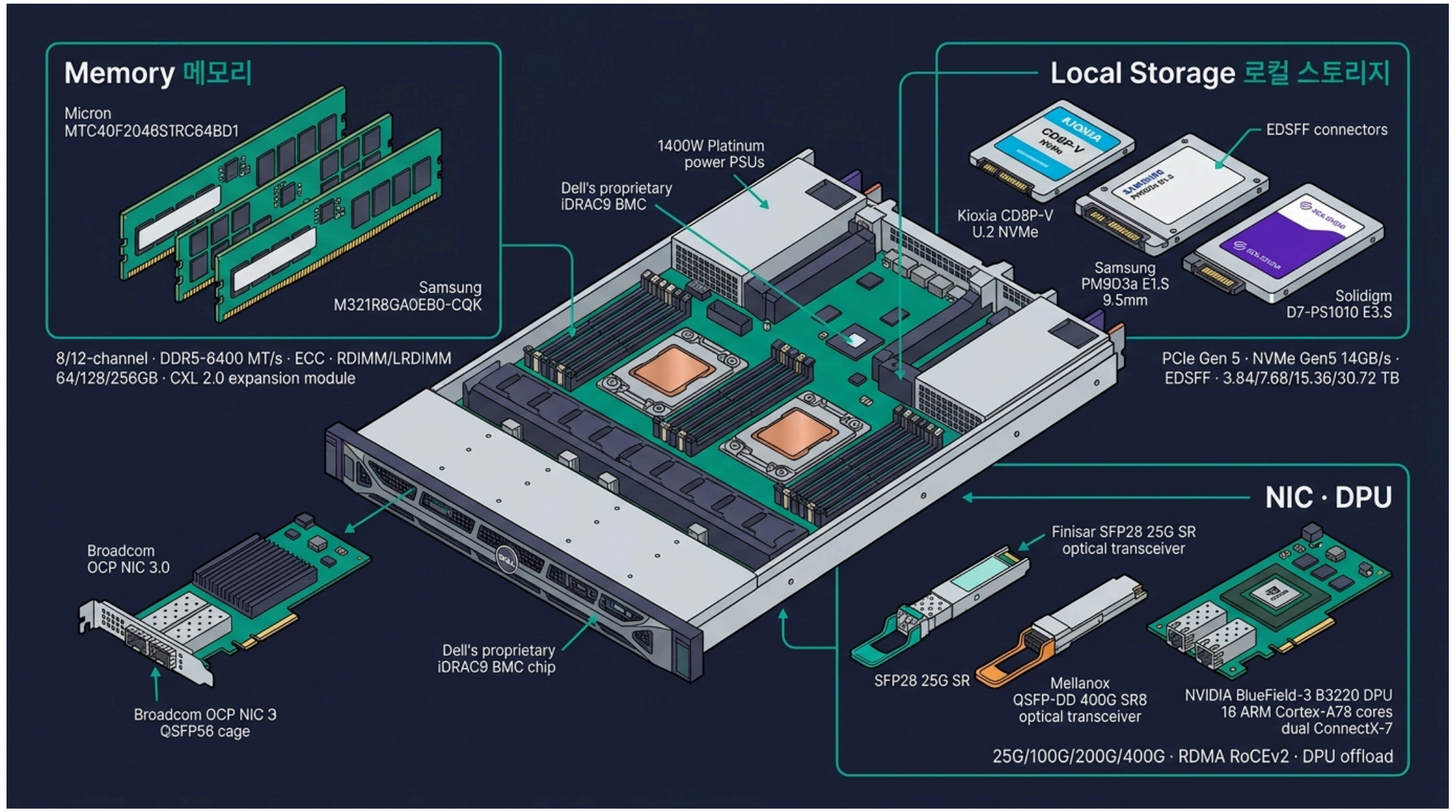
🔌 NIC 관점 한 줄

- **NIC = 서버의 네트워크 어댑터** (PCIe 카드)
- 슬롯 ↔ 모듈 매칭: 25G NIC → **SFP28** 슬롯, 100G NIC → **QSFP28** 슬롯
- 단거리 (랙 내 <3m): **DAC** 직결 케이블 — 가장 저렴·저지연
- 중거리 (랙 사이 <30m): **AOC** 또는 광 모듈 + MM
- 장거리 (DC 간): 광 모듈 + SM
- **벤더 코딩** 주의 — Cisco/Juniper/Arista 모듈 호환 키 상이, Generic 은 보증 없음

📚 본문은 Session 4 PART B

- **광 모듈 SKU 11종** (SFP·SFP+·SFP28·SFP56·QSFP+·QSFP28·QSFP56·QSFP-DD·OSFP·OSFP-XD) — `s4 ## 광 모듈`
- **광 거리 표준** SR·LR·ER·ZR · 100/400GBASE 시리즈 — `s4 ## 광 거리 표준`
- **광 케이블** Single-Mode vs Multi-Mode (OM3/OM4/OM5/OS2) — `s4 ## 광 케이블`
- **케이블 결정 매트릭스** Cat 6a·Cat 8·DAC·SR·LR — `s4 ## 케이블 결정 매트릭스`

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현



메모리·NVMe·NIC 분해도 (삽화)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

PART D

운영체제 — *Linux · Windows · UNIX*

> OS는 라이센스·운영·인력을 동시에 결정한다.

OS 3대 진영 — *Linux · Windows · UNIX*

🐧 Linux

- **시장 점유율 80%+** (서버 영역)
- 무료/유료 디스트리뷰션 다양
- 한국 공공·금융 표준
- 대표
- RHEL (Red Hat Enterprise)
- Rocky·AlmaLinux (RHEL 호환·무료)
- Ubuntu Server LTS
- SUSE Linux Enterprise
- Oracle Linux
- Amazon Linux 2/2023

🪟 Windows Server

- **시장 점유율 15~20%** (서버 영역)
- 라이센스 (DataCenter / Standard / CAL)
- Hyper-V 가상화 기본 포함
- AD·DNS·DHCP·IIS·MS SQL 연동
- 대표
- Windows Server 2019 (EOL 2029)
- Windows Server 2022 (EOL 2031)
- Windows Server 2025 (2024.11 GA)

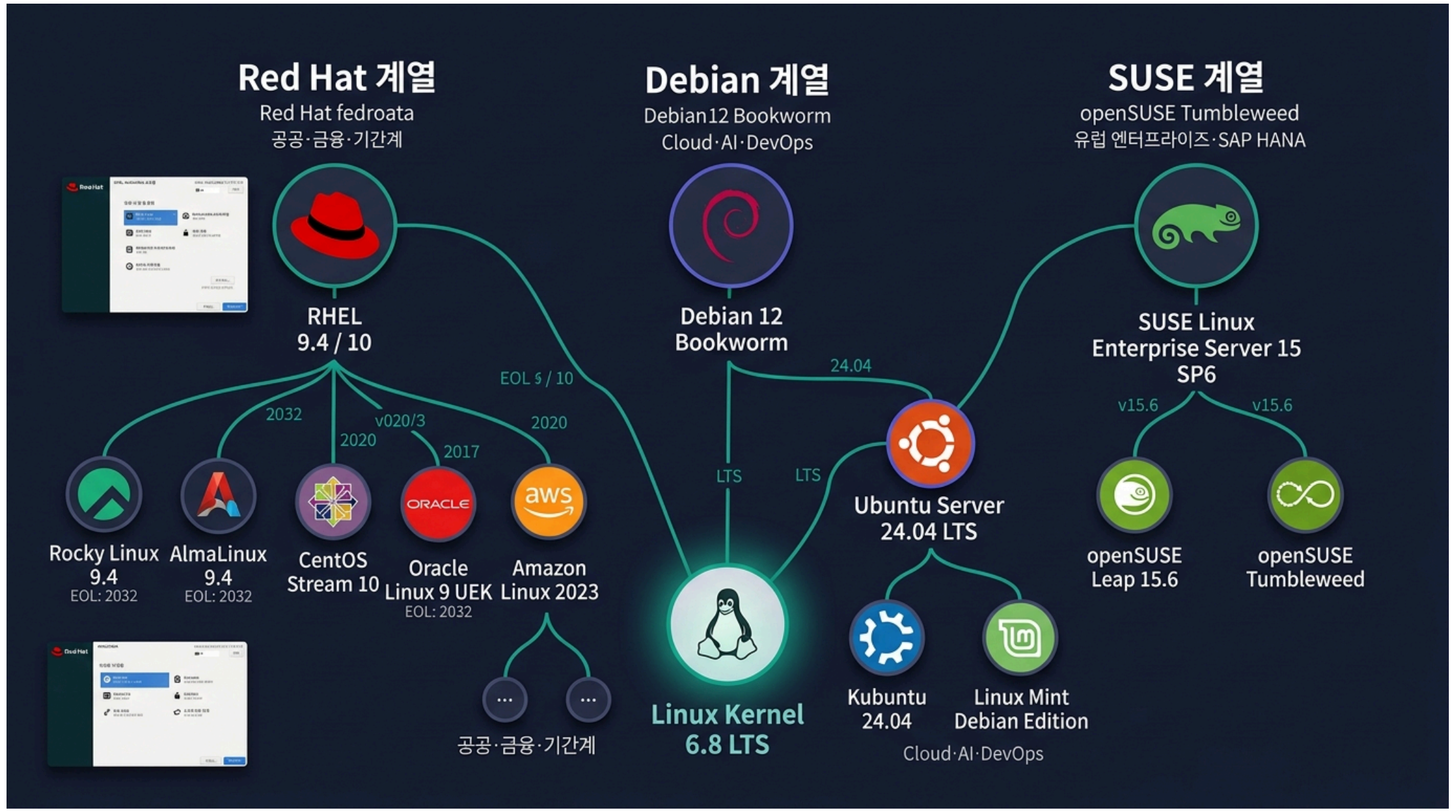
🦾 UNIX 전통

- 시장 점유율 ↓ but 운영 잔존
- 대표
- **AIX** (IBM Power) — 금융 잔존
- **HP-UX** (HPE Itanium) — EOL 임박
- **Solaris** (Oracle SPARC) — EOL
- **FreeBSD** (오픈소스) — 일부 NW 장비·연구

📊 한국 시장 (2026)

- Linux: 75%
- Windows: 18%
- UNIX: 6%
- 기타: 1%

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현



Linux 디스트리뷰션 가계도 (삽화)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

Linux 디스트리뷰션 비교 — *8종*

배포판	모기업	라이센스	지원 기간	주 사용처	비고
RHEL	Red Hat (IBM)	구독 (유료)	10년 (Standard·Extended까지 13년)	공공·금융 표준	가장 비싼 Subscription
Rocky Linux	RESF	무료	10년	RHEL 대안·신규 표준	CentOS 후속, 커뮤니티
AlmaLinux	CloudLinux	무료	10년	RHEL 대안·신규 표준	ABI 호환 보증
CentOS Stream	Red Hat	무료	5년 (rolling)	RHEL 미리보기	운영 비권장 (Beta 성격)
Ubuntu Server LTS	Canonical	무료 (Pro 유료)	5년 (Pro 10년)	AI·Cloud·연구	22.04 / 24.04 LTS
SUSE Linux Enterprise	SUSE	구독 (유료)	13년	유럽·SAP HANA	SAP 표준
Oracle Linux	Oracle	무료 (지원 유료)	10년	Oracle DB·ULA	UEK 커널 옵션
Amazon Linux 2/2023	AWS	무료 (AWS만)	5년	AWS EC2 전용	일반 온프레미스 부적합

한국 2026 신규 RFP의 표준 선택: 공공·금융 = **RHEL 또는 Rocky/AlmaLinux**, AI/연구 = **Ubuntu LTS**, SAP = **SUSE**, Oracle DB = **Oracle Linux**.

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

Windows Server — *2019 · 2022 · 2025*

🪟 버전 비교

버전	출시	주류 지원	확장 지원	특징
2019	2018.10	2024.01	2029.01	컨테이너·LTSC
2022	2021.08	2026.10	2031.10	Secured-core·hot-patching
2025	2024.11	2029.10	2034.10	AI·Wi-Fi 7·SMB over QUIC

🎯 에디션

- **Standard** — VM 2개 권리 · 라이센스 Per-Core
- **DataCenter** — VM 무제한 · 가상화 호스트 표준
- **Essentials** — 25사용자/50디바이스 · SMB

🛠️ 핵심 역할

- **AD DS / AD CS** — 도메인·인증서
- **DNS / DHCP / WINS** — 인프라 서비스
- **IIS** — 웹 서버
- **MS SQL** — DB (별도 라이센스)
- **Hyper-V** — 가상화 (기본 포함)
- **Failover Cluster** — HA
- **WSUS** — 패치 관리
- **RDS** — 원격 데스크탑·VDI

🎙️ TA 결정 포인트

- **라이센스 모델 2024** — Core Pack (2-core × 8 = 16C 기본) + CAL
- **DataCenter vs Standard** — VM 4개 이상이면 DataCenter
- **AD = 도메인 인프라** — AD 1개당 도메인 컨트롤러 2대+

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

UNIX 계열 — *AIX · HP-UX · Solaris · FreeBSD*

🟣 IBM AIX

- IBM Power 전용 (Power9·Power10)
- **AIX 7.3** (현 버전, ~2030+)
- LPAR·WPAR·PowerHA·VIOS
- 강점: 안정성·금융 코어·DB2
- 한국 점유: 금융 코어뱅킹·일부 공공
- **신규 도입은 거의 없음, 유지 운영 중심**

🟢 HP-UX

- HPE Itanium 전용 (이미 EOL)
- HP-UX 11i v3 — 마지막 버전
- **2025년 SP 종료, 2027년 EOS**
- 잔존 운영은 x86 Linux 전환 작업 중

🟡 Oracle Solaris

- Oracle SPARC + x86 일부
- **Solaris 11.4** — 2034 지원
- ZFS·DTrace·Zones (컨테이너 원조)
- Oracle ULA 묶음 잔존
- **신규 도입은 사실상 종료**

🔵 FreeBSD

- 오픈소스 BSD 계열
- 사용처
- 네트워크 장비 펌웨어 (Juniper Junos·NetApp ONTAP)
- 일부 보안 어플라이언스 (pfSense·OPNsense)
- 연구·임베디드
- 한국 서버 운영 도입 거의 없음

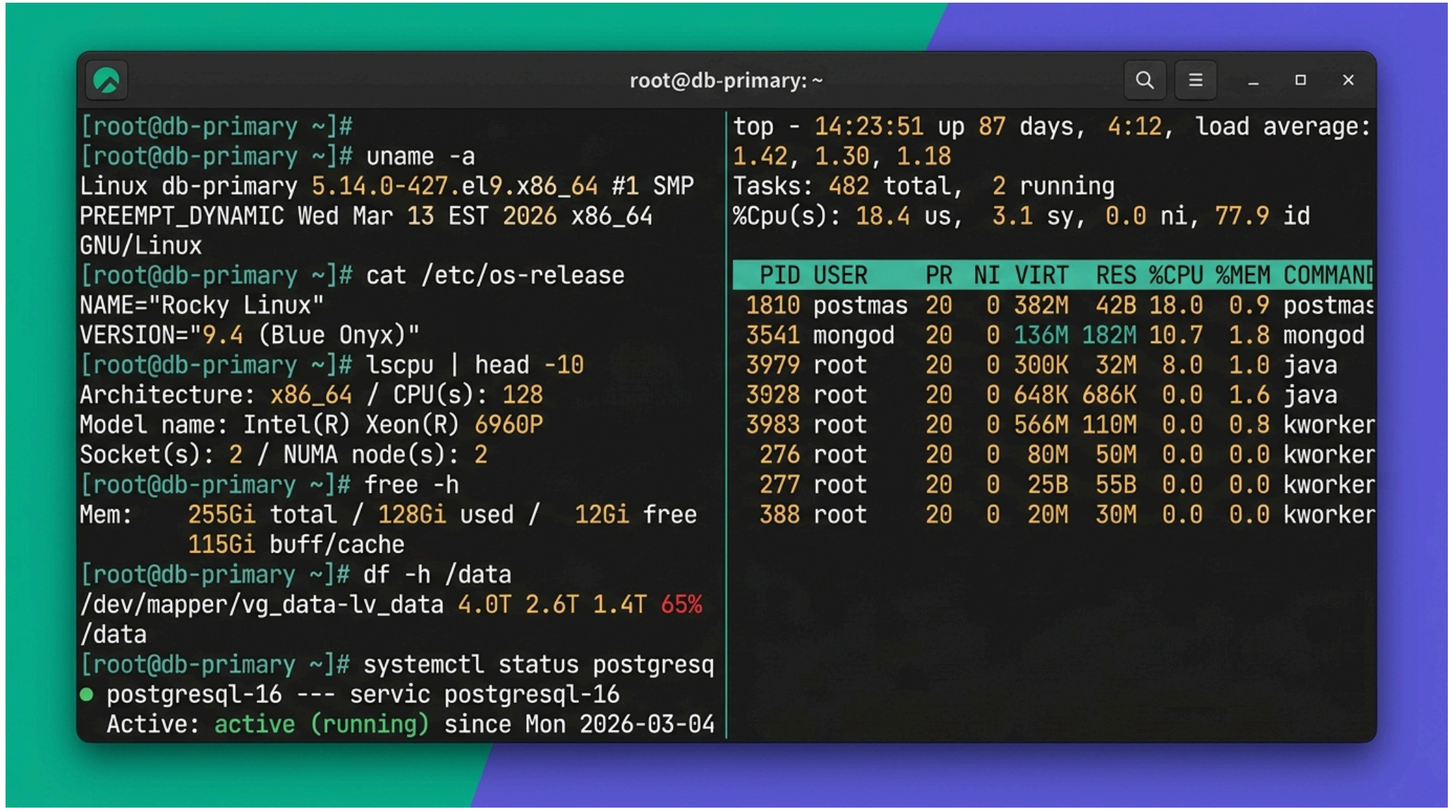
© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

Linux 운영 콘솔 — *TA의 일상 운영 명령 5종*

```
[root@db-primary ~]# uname -a
Linux db-primary 5.14.0-427.el9.x86_64 #1 SMP PREEMPT_DYNAMIC ... GNU/Linux

[root@db-primary ~]# free -h
              total        used        free      shared  buff/cache   available
Mem:          255Gi       128Gi        12Gi       2.1Gi       115Gi       124Gi
Swap:          16Gi        0.0Gi       16Gi

[root@db-primary ~]# df -h /data
Filesystem      Size  Used Avail Use% Mounted on
/dev/mapper/vg_data-lv_data   4.0T  2.6T  1.4T  65% /data

[root@db-primary ~]# top -bn1 | head -10
top - 14:23:51 up 87 days, load average: 1.42, 1.30, 1.18
Tasks: 423 total,   2 running, 421 sleeping
%Cpu(s): 18.4 us,  3.1 sy,  0.0 ni, 77.9 id,  0.4 wa
MiB Mem: 261072 total, 12544 free, 131072 used, 117456 buff/cache
```

> 📷 **실제 사진 참조** (위키미디어/벤더 공식) - Bash (Unix shell) — Wikipedia — Wikipedia - Red Hat Enterprise Linux 9 — Red Hat 공식 - Rocky Linux 9 — Rocky 공식 - systemd documentation — systemd 공식

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현



Linux 운영 콘솔 — TA의 일상 운영 명령 5종

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

OS 운영 핵심 개념 — *커널 · 파일시스템*

🐧 커널·시스템콜

- **커널 모드** vs 사용자 모드
- 시스템콜 (syscall) — 응용↔커널 인터페이스
- **Linux 커널** — 모놀리식 (모듈 로딩)
- **Windows NT 커널** — 하이브리드
- 컨테이너 = 호스트 커널 공유
- 가상화 = 게스트 커널 독립

🔧 부트 프로세스

- BIOS → POST → MBR (legacy)
- **UEFI** → ESP → bootloader (현 표준)
- Linux: GRUB → kernel → initramfs → systemd
- Windows: bootmgr → winload → kernel

💾 파일시스템 비교

FS	진영	최대 파일	특징
ext4	Linux	16TB	기본·안정
XFS	Linux	16EB	대용량·고성능 (RHEL 기본)
ZFS	Solaris/Linux/BSD	EB급	스냅샷·압축·중복제거·체크섬
Btrfs	Linux	16EB	CoW·스냅샷 (운영 안정성 ↓)
NTFS	Windows	256TB	기본·NTFS 압축·암호화
ReFS	Windows	1YB	스토리지 공간·DataCenter 권장

🛡️ LVM · LUKS

- **LVM** — Logical Volume Manager (Linux)
- **LUKS** — Linux Unified Key Setup (디스크 암호화)
- 운영 표준: XFS on LVM on LUKS

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

systemd · cgroup · namespace · 보안 모듈

⚙️ systemd

- Linux 표준 init 시스템 (2014+)
- 유닛 (service·target·timer·socket·mount)
- `systemctl` · `journalctl`
- **부팅 시간 단축 + 의존성 관리**
- 비교: OpenRC (Alpine)·SysV init (구형)

🧱 cgroup (Control Group)

- 프로세스 자원 격리 (CPU·Mem·IO·NW)
- **cgroup v2** (2022+ 표준)
- **컨테이너 격리의 핵심 기술**
- Kubernetes·Docker·Podman 기반

🏠 namespace

- 프로세스 가시성 격리 (PID·NW·Mount·UTS·IPC·User)
- **컨테이너의 격리 메커니즘**

🛡️ 보안 모듈 (LSM)

- **SELinux** (Red Hat) — Mandatory Access Control
- **AppArmor** (Ubuntu·SUSE) — Path-based MAC
- **Capabilities** — root 권한 세분화 (38종)
- **seccomp** — syscall 화이트리스트
- **eBPF** — 동적 커널 hook (관찰·보안)

🎙️ TA 결정 포인트

- **SELinux 활성화**가 운영 표준 (Disabled 금지)
- Ubuntu는 AppArmor 기본
- 컨테이너 = cgroup + namespace + LSM + seccomp + capabilities **결합**
- **이게 깨지면 컨테이너 탈출 가능**

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

OS 라이센스 모델 — *공통 함정*

💰 RHEL Subscription

- **Per-Socket / Per-Core (2024+ 전환 중)**
- Self-Support / Standard / Premium
- VM 단위가 아닌 **물리 호스트 단위**
- 가상화 시 **호스트 RHEL 라이센스 + 게스트 다 포함** 옵션도
- 신규 가격 인상 압박 → Rocky·Alma 전환 추세

💰 Windows Server

- **Core Pack** — 2-core × 8 (16C 기본)
- 추가 코어는 2-core 단위 구매
- **CAL** — 사용자/디바이스 단위 별도
- **DataCenter** — VM 무제한 + Azure Hybrid Benefit
- 가상화 시 호스트 코어 전체 라이센스 필수

💰 Ubuntu Pro

- 무료 (커뮤니티 LTS 5년)
- **Pro 구독** — 10년 + ESM + LivePatch + KOS
- 데스크탑·서버 통합
- AI/연구는 무료, 운영 필수면 Pro

💰 SUSE

- 구독 (Per-Socket / Per-Core)
- SAP HANA 묶음
- 라이센스 비싸지만 SAP 표준

💰 Oracle Linux

- OS 자체 무료
- **지원 구독은 유료** (Premier·Network)
- Oracle DB 묶음 시 유리

🎙️ TA 결정 포인트

- **OS 라이센스 ≠ DB 라이센스 ≠ MW 라이센스** — 각각 산정
- **가상화 시 게스트 라이센스 검토** 필수

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

OS 결정 매트릭스 — *워크로드별*

워크로드	권장 OS	이유
공공 표준 사업	RHEL or Rocky Linux	TTA·국정원 호환·표준
금융 코어 시스템	RHEL (지원 명확)	SLA·인증·24/7 지원
AI/ML 학습·연구	Ubuntu 22.04/24.04 LTS	CUDA·드라이버·SW 풍부
SAP HANA	SUSE SLES for SAP	SAP 인증 표준
Oracle DB	Oracle Linux (UEK) or RHEL	Oracle 인증
AD 도메인·MS SQL	Windows Server 2022/2025 DataCenter	통합 운영
VDI (수천 사용자)	Windows Server 2022 + Citrix or VMware Horizon	표준
컨테이너 호스트	Rocky/Alma or Ubuntu LTS	K8s 표준·경량
K8s 노드 OS	RHCOS / Flatcar / Talos	Immutable OS
레거시 AIX 운영	AIX 7.3 유지 + Linux 마이그레이션 계획	EOL 대비

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

PART E

DBMS 분류 — *RDBMS · NoSQL · NewSQL · 시계열 · 검색 · 벡터*

> DB는 인프라 비용의 30~50% — TA의 가장 무거운 결정.

DBMS 6대 분류 — *Wheel*

1️⃣ RDBMS (관계형)

- ACID·SQL·트랜잭션
- Oracle·MSSQL·MySQL·PostgreSQL·Tibero·MariaDB·DB2
- **인프라 시장 점유율 1위**
- OLTP·OLAP 모두 가능

2️⃣ NoSQL

- BASE·스키마리스·분산
- Document·Column·Key-Value·Graph
- MongoDB·Cassandra·Redis·Neo4j
- 빅데이터·세션·메시지

3️⃣ NewSQL

- ACID + 분산 + SQL
- TiDB·CockroachDB·YugabyteDB·Google Spanner
- 글로벌 분산 OLTP
- HTAP (Hybrid OLTP/OLAP)

4️⃣ 시계열 (Time-Series)

- 시간순 메트릭·이벤트
- InfluxDB·TimescaleDB·Prometheus DB·QuestDB
- IoT·NMS·APM·관찰성

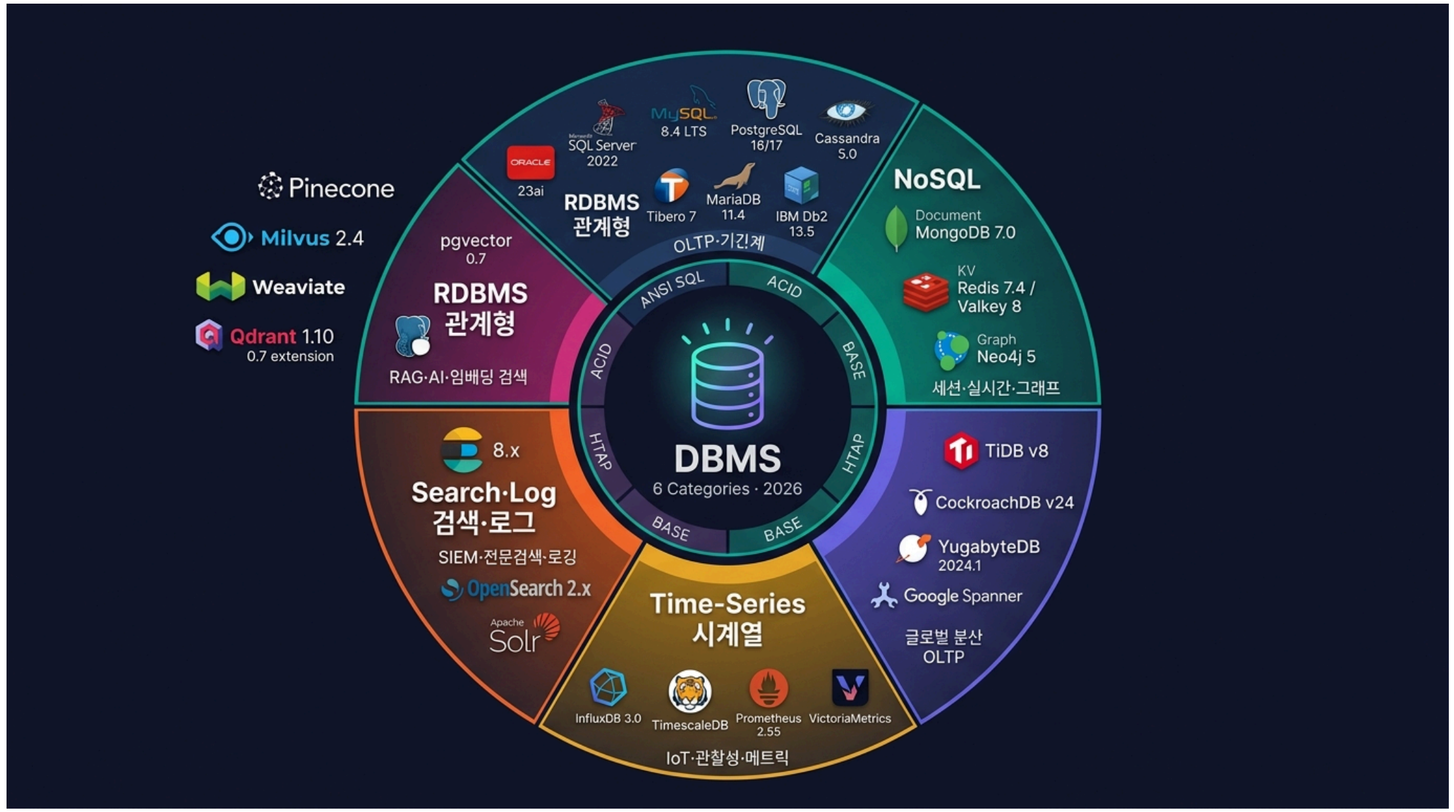
5️⃣ 검색·로그

- 역색인·전문 검색
- Elasticsearch·OpenSearch·Solr
- 로그 통합·SIEM·전자상거래

6️⃣ 벡터 (Vector)

- 임베딩 검색·코사인 유사도
- Pinecone·Weaviate·Milvus·Qdrant·pgvector·ChromaDB
- **RAG·생성형 AI 표준 (2023+)**

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현



DBMS 6대 분류 wheel (삽화)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

RDBMS ① Oracle — *코어뱅킹의 황제*

🏛️ Oracle Database

- 19c (LTS, ~2032) · 21c (Innovation) · 23ai (2024)
- **Editions**: SE2 · EE · EE + Options
- 옵션: RAC·Active Data Guard·Partitioning·GoldenGate·Multitenant
- 라이센스: **Per-Core × Factor (Intel x86 = 0.5)**
- Named User Plus (NUP) 별도

⚙️ HA 구성

- **RAC** (Real Application Cluster) — Active-Active, 공유 스토리지
- **Data Guard** — 물리/논리 Standby
- **GoldenGate** — 이기종 복제·CDC
- **Multitenant (CDB/PDB)** — 통합 관리

🎙️ TA 결정 포인트

- **라이센스가 가장 비싼 DB** — 코어당 수천만 원
- EE 코어 1개 ≈ 4,000~5,000만 원 (2026 시세, +Support 22%)
- 옵션 (RAC·Partitioning) 별도 라이센스
- **CPU 코어 수가 곧 비용** → 고클럭·저코어 SKU 선택
- **VMware 호스트 = 호스트 전 코어 라이센스** (Oracle Hard Partitioning 정책)
- 대안: Tibero (국산)·PostgreSQL (오픈)

📊 한국 도입

- 공공·금융·대기업 코어 표준
- 2010년대 후반부터 Tibero·PostgreSQL 대체 가속

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

RDBMS ② MSSQL · ③ MySQL

🪟 Microsoft SQL Server

- **2019 (CU 30+) · 2022 (현재 표준) · 2025 (출시 임박)**
- Editions: Express·Standard·Enterprise·Developer
- 라이센스: **Per-Core (4-core 단위)**
- Enterprise Per-Core ≈ 1,500만 원 (2026)
- HA: **Always On AG (Availability Group)** — 2 동기 + 다수 비동기
- Failover Cluster Instance (FCI) — 공유 스토리지

🎙️ TA 메모

- Windows AD/MS Office 환경에 자연스러움
- Linux 지원 (2017+) but 운영 표준은 Windows
- Power BI·SSAS·SSIS 통합

🐬 MySQL

- **8.0 (현재 표준) · 8.4 LTS (2024+)**
- Editions: Community (GPL)·Enterprise (Oracle)
- Storage Engine: **InnoDB** (표준)·MyISAM (구형)
- HA 옵션
- **InnoDB Cluster** (8.0+) — Group Replication + MySQL Router
- **Group Replication** — Multi-Primary·Single-Primary
- **MySQL Replication** — 비동기·반동기
- **Percona XtraDB Cluster** (Galera 기반)

🎙️ TA 메모

- 웹·중소 규모 OLTP 표준
- 라이센스: Community 무료, Enterprise 유료
- MariaDB로 갈라짐 (MySQL 호환 fork)

RDBMS ④ PostgreSQL — *오픈소스 챔피언*

🐘 PostgreSQL

- **16 (2023.09) · 17 (2024.09)** — 메이저 매년
- 진정한 오픈소스 (PostgreSQL License = BSD 호환)
- ACID·JSON·전문검색·지리정보 (PostGIS)
- 확장 (Extension) 풍부 — TimescaleDB·Citus·pgvector
- **공공·금융 오픈소스 RDBMS 1순위**

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

🎙️ TA 결정 포인트

- **Oracle 대체 1순위** — 라이센스 무료
- DBA 인력 수급 (Oracle DBA보다 PostgreSQL DBA 부족)
- 상용 지원: **EDB Postgres·2ndQuadrant·Crunchy Data**
- 한국 지원: EnterpriseDB Korea·Bitnine·아이씨엔
- **공공 마이그레이션** — TmaxSoft Tibero ↔ PostgreSQL 경쟁

RDBMS ⑤ Tibero · ⑥ MariaDB · ⑦ DB2

🇰🇷 Tibero (TmaxSoft)

- **국산 Oracle 호환 DBMS**
- Tibero 7 (2023) · Tibero TAC (RAC 호환)
- **공공 SW 진흥법** 우대 (국산 SW 가산점)
- 라이센스: 코어 단위 (Oracle 대비 30~50%)
- 호환: SQL·PL/SQL 95%+ 호환
- HA: **TAC** (Tibero Active Cluster)
- 한국 공공·금융 점유 ↑↑ (2020~)

🎙️ TA 메모

- 공공 신규 RFP에서 Oracle vs Tibero 비교 빈번
- DBA·운영 도구 성숙도 개선 중
- 한국 외 사용 거의 없음

🌊 MariaDB

- MySQL 5.5 fork (2009)
- **MariaDB 10.11 LTS · 11.x**
- ColumnStore·MaxScale·Xpand
- **Galera Cluster** — Multi-Master 동기 복제
- MySQL 호환 (8.0+는 분기 심해짐)

🎙️ TA 메모

- RHEL/Ubuntu 기본 패키지
- 웹·중소 OLTP 무료 대안
- 유럽·중국 점유 ↑
- 한국은 MySQL이 우세, MariaDB 일부

🟦 IBM DB2

- **DB2 12 (z/OS) · DB2 11.5 (LUW)**
- Mainframe 표준 (코어뱅킹·통신 BSS)
- Power AIX + DB2 = 금융 잔존
- pureScale (RAC 유사)
- 한국 도입: 은행 코어·일부 공공
- **신규 도입은 거의 없음, 유지 운영 중심**

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

NoSQL 4종 — *Document · Column · KV · Graph*

📄 Document

- JSON/BSON 문서 단위
- **MongoDB** (상용 + 오픈)
- **CouchDB** (오픈)
- Amazon DocumentDB (MongoDB 호환)
- 사용처: 카탈로그·콘텐츠·세션·로그

🏛️ Column (Wide-Column)

- 행보다 열 단위 저장 (분석 효율)
- **Cassandra** (분산 마스터리스)
- **HBase** (Hadoop 위)
- **ScyllaDB** (Cassandra 호환·C++)
- 사용처: 시계열·이벤트·IoT·메시징

🔑 Key-Value

- 단순 K:V 매핑
- **Redis** (인메모리 + AOF/RDB)
- **Memcached** (순수 캐시)
- **DynamoDB** (AWS 매니지드)
- **etcd** (Raft·K8s 사용)
- 사용처: 세션·캐시·랭킹·큐

🕸️ Graph

- 노드·엣지·속성 모델
- **Neo4j** (Cypher)
- **OrientDB** · **ArangoDB** (멀티 모델)
- **JanusGraph** · **TigerGraph**
- 사용처: 추천·SNS·사기탐지·지식그래프

NoSQL 도입 원칙: NoSQL은 **RDBMS의 대체가 아닌 보완**. 메시지·세션·로그·캐시 등 **부분 워크로드**에 도입. RDBMS를 폐지하고 NoSQL만 운영하는 사례는 드묾.

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

NewSQL — *ACID + 분산 + SQL*

🆕 NewSQL이란

- **ACID 보장 + 수평 확장 + SQL**
- RDBMS의 트랜잭션 + NoSQL의 확장성
- **분산 트랜잭션 (Distributed Tx)**
- 글로벌 분산 (Multi-Region)
- HTAP (OLTP + OLAP 동시)

🏷️ 대표 제품

- **TiDB** (PingCAP, 중국) — MySQL 호환
- **CockroachDB** (Cockroach Labs) — PostgreSQL 호환
- **YugabyteDB** — PostgreSQL 호환
- **Google Spanner** (GCP 매니지드)
- **VoltDB** — 인메모리 OLTP
- **Vitess** — MySQL 샤딩 (Slack·YouTube)

🎙️ TA 결정 포인트

- 글로벌 분산이 필요한 경우 (다중 리전 OLTP)
- 단일 RDBMS의 한계 (수평 확장)
- 한국 도입: 아직 초기·일부 게임·SaaS
- 운영 난이도·인력 ↑↑↑
- 공공·금융은 사용 거의 없음

📊 한국 시장 (2026)

- TiDB: 일부 게임·핀테크
- CockroachDB: SaaS·스타트업
- YugabyteDB: 일부 대기업 PoC
- **본격 확산 전 단계**

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

시계열 · 검색 DB — *관찰성·로그의 표준*

⏰ 시계열 (Time-Series)

- **InfluxDB** (TICK Stack)
- **TimescaleDB** (PostgreSQL 확장)
- **Prometheus TSDB** (모니터링 표준)
- **QuestDB·VictoriaMetrics·Mimir**
- 특징
- 시간순 데이터 압축 효율
- 다운샘플링·연속 쿼리
- 만료 (TTL) 정책
- 사용처
- APM·메트릭·NMS
- IoT 센서·산업 데이터
- 금융 시세

🔍 검색·로그

- **Elasticsearch** (Elastic SSPL)
- **OpenSearch** (AWS·Apache 2.0 fork)
- **Solr** (Apache Lucene)
- **MeiliSearch·Typesense** (경량)
- 특징
- 역색인 (Inverted Index)
- 전문 검색·집계
- 분산 샤딩
- 사용처
- 사이트 검색·전자상거래
- 로그 통합 (ELK Stack)
- SIEM (Splunk 대체)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

벡터 DB — *RAG·AI의 표준 (2023+)*

🧠 벡터 DB란

- 임베딩 벡터 (수백~수천 차원) 저장·검색
- **유사도 검색** — 코사인·내적·Euclidean
- **ANN (Approximate Nearest Neighbor)** — HNSW·IVF·PQ
- **RAG** (Retrieval-Augmented Generation) 표준 인프라
- 메타데이터 필터링 + 벡터 검색

🏷️ 대표 제품

- **Pinecone** (SaaS) — 매니지드 표준
- **Weaviate** — 오픈 + 클라우드
- **Milvus / Zilliz** — 오픈 (Apache 2.0)
- **Qdrant** — 오픈 + 클라우드 (Rust)
- **ChromaDB** — 임베디드·경량 (Python)
- **pgvector** — PostgreSQL 확장 (운영 친화)

🎙️ TA 결정 포인트

- **RAG·AI 도입 시 필수** — 2024+ 신규 사업 빈번
- 데이터 규모 < 1M = pgvector (PostgreSQL 통합)
- 1M ~ 100M = Milvus·Qdrant·Weaviate (전용)
- 100M+ = Milvus 클러스터·Pinecone (매니지드)
- **GPU 메모리 + 인덱스 메모리 산정** 필수
- 임베딩 차원 × 4B × 벡터 수 = 메모리 추정

📊 도입 동향

- 공공 RAG 챗봇 (FAQ·민원)
- 금융 RAG (지점 매뉴얼·규정)
- 대학 LMS (수업 자료 검색)
- 의료·법률 자료 검색

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

DB 결정 매트릭스 — *워크로드별*

워크로드	권장 DBMS	이유
금융 코어 OLTP	Oracle EE + RAC + DG	검증·SLA·인력
공공 신규 OLTP	PostgreSQL + Patroni / Tibero TAC	라이센스·국산 가산점
MSSQL/AD 환경	MS SQL 2022 Always On AG	통합·도구·라이센스
웹 서비스 OLTP	MySQL 8.0 InnoDB Cluster / MariaDB Galera	가벼움·무료
OLAP·BI·DW	PostgreSQL + Citus / ClickHouse / Snowflake	컬럼 저장·집계
HTAP	TiDB / YugabyteDB / Oracle 23ai	분산 OLTP+분석
세션·캐시	Redis Cluster	인메모리·LRU·Pub/Sub
메시지·이벤트 큐	Kafka (DB 외)·NATS·RabbitMQ	비동기
카탈로그·콘텐츠	MongoDB Replica Set	스키마리스·JSON
시계열·IoT·메트릭	TimescaleDB / InfluxDB / Prometheus	압축·다운샘플
로그·전문검색·SIEM	Elasticsearch / OpenSearch	역색인·집계
RAG·AI 임베딩	pgvector / Milvus / Qdrant / Pinecone	ANN·메타 필터
글로벌 분산 OLTP	CockroachDB / Spanner / YugabyteDB	다중 리전
그래프·추천	Neo4j	Cypher·노드/엣지

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

ACID vs BASE — *트랜잭션 철학*

⚖️ ACID (RDBMS 전통)

- **A**tomicity — 원자성 (All or Nothing)
- **C**onsistency — 일관성 (불변식 보존)
- **I**solation — 격리성 (동시 트랜잭션 분리)
- **D**urability — 영속성 (커밋 후 보존)

적합 - 금융 거래·예약·재고 - 정확성이 절대 우선 - RDBMS·NewSQL

🌊 BASE (NoSQL 전통)

- **B**asically **A**vailable — 기본 가용
- **S**oft state — 상태 가변
- **E**ventual consistency — 최종 일관성

적합 - 소셜·콘텐츠·로그 - 가용성·확장이 우선 - NoSQL·일부 NewSQL

🌐 CAP 정리

- **C**onsistency · **A**vailability · **P**artition tolerance
- 네트워크 분할 시 **C와 A 중 택일**
- CP: HBase·MongoDB (강한 일관성)
- AP: Cassandra·DynamoDB (가용성)

🌐 PACELC

- 분할 시: P → A or C
- 정상 시: 지연 (L) vs 일관성 (C)
- 더 현실적 모델

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

DB 라이센스 비교 — *비용 산정*

DBMS	모델	단가 (참고, 2026)	비고
Oracle EE	Per-Core × Factor	4,000~5,000만원/코어 (+22% 유지보수)	x86 Factor 0.5, 옵션 별도
Oracle SE2	Per-Socket (최대 2S)	2,500만원/소켓	한정 기능
MS SQL EE	Per-Core (4C 단위)	1,500~1,800만원/코어	DataCenter 별도
MS SQL Std	Per-Core or Server+CAL	약 600만원/코어	중소 환경
Tibero EE	Per-Core	Oracle 대비 30~50%	공공 가산점
PostgreSQL	무료 (BSD 호환)	0원 + 상용지원 (EDB)	신규 표준 후보
MySQL Community	무료 (GPL)	0원 + Enterprise 옵션	웹 표준
MariaDB Enterprise	Subscription	노드당 수백만원	유럽·중국
DB2	Per-PVU (IBM 가중치)	협상가	잔존 운영
MongoDB Enterprise	Per-Node	노드당 수천만원	Community 무료

Oracle 라이센스 함정: ① VMware 호스트 = **전 코어 라이센스** (Hard Partitioning 정책) ② **NUP 최소 수** (CPU 코어 × 25) ③ **옵션** (RAC·Partitioning·Active DG)별로 별도 라이센스. 단순 코어 단가만 보면 BOM 실수.

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

DB HA 패턴 — *공통 분류*

🔁 Active-Standby (Cold/Warm/Hot)

- **Cold** — Standby 정지, 장애 시 부팅
- **Warm** — Standby 부팅·서비스 미오픈
- **Hot** — Standby 서비스 오픈 (읽기 가능)
- 대표: MS SQL FCI·Oracle DG·MySQL Replication

🔁 Active-Active

- 모든 노드가 쓰기 가능
- 강한 일관성 또는 결합/결합거부 충돌
- 대표: Oracle RAC·MySQL Group Replication·Galera

🔁 Shared-Storage HA

- 공유 디스크에 두 노드가 페일오버
- 대표: Oracle RAC·MSCS·Pacemaker·VCS

🔁 Shared-Nothing 분산

- 각 노드 자체 디스크, 데이터 샤딩·복제
- 대표: Cassandra·MongoDB·CockroachDB

📐 Quorum·Witness

- 짝수 노드 → 스플릿 브레인 위험
- 홀수 노드 또는 **Witness/Arbiter** 추가
- etcd·Consul·Zookeeper 활용

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

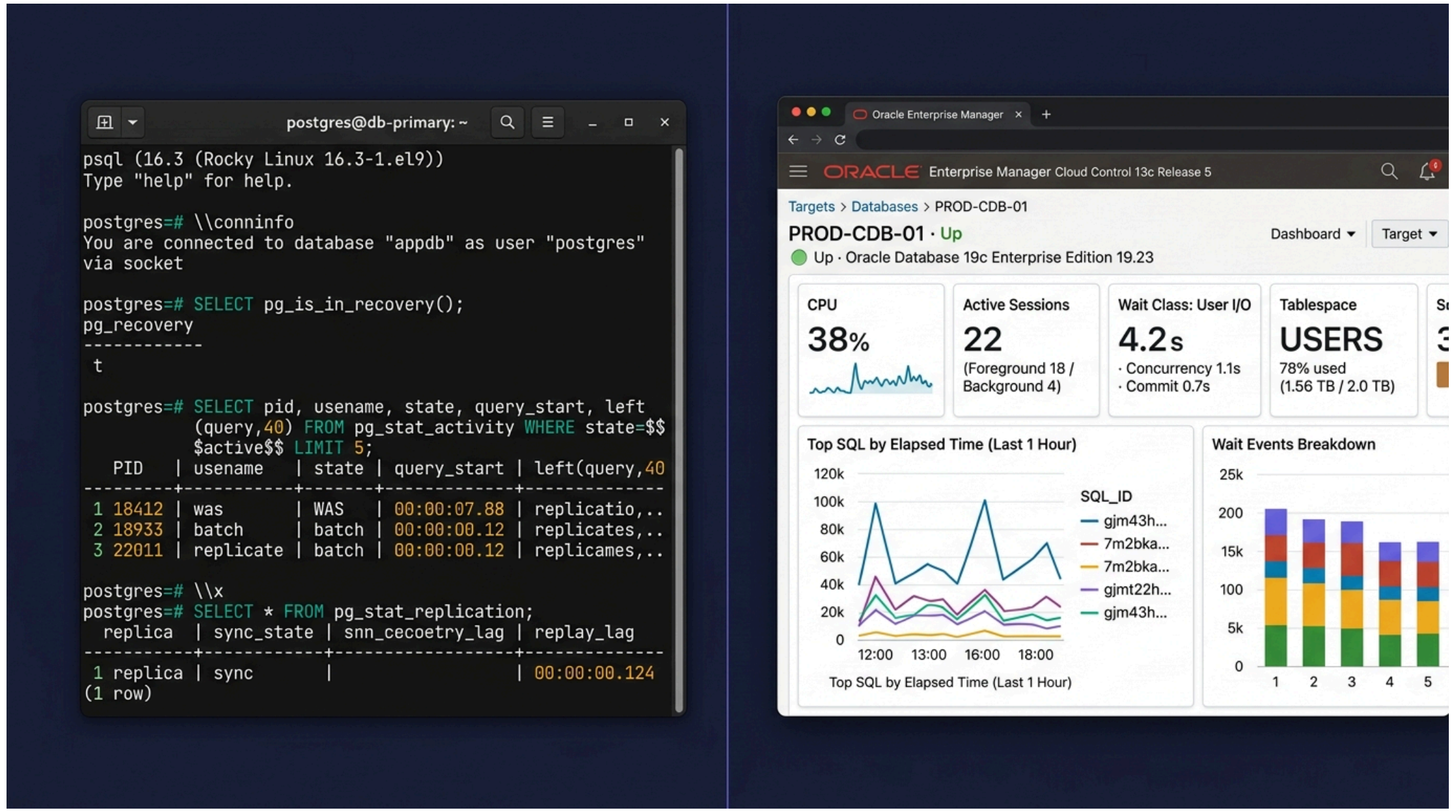
DB 운영 콘솔 — *psql 세션 + Oracle Enterprise Manager*

운영 TA가 일상적으로 사용하는 psql / sqlplus 명령과, DBA가 보는 OEM 대시보드.

```
postgres=# SELECT pg_is_in_recovery(), pg_last_wal_replay_lsn();
 pg_is_in_recovery | pg_last_wal_replay_lsn
-------------------+------------------------
 t                 | 1A/3F8E2B40
(1 row)

postgres=# SELECT pid, usename, application_name, state, query_start
            FROM pg_stat_activity
            WHERE state <> 'idle'
            ORDER BY query_start
            LIMIT 5;
  pid   | usename | application_name | state  |     query_start
-------+---------+------------------+--------+-------------------
 23104 | app_wr  | spring-boot-was  | active | 14:23:51.214+09
 23187 | app_ro  | report-batch     | active | 14:23:52.482+09
 23211 | repl    | walreceiver      | idle in transaction | 14:23:53.012+09
```

> 📷 **실제 사진 참조** (벤더·오픈소스 공식) - PostgreSQL psql 공식 문서 — PostgreSQL 공식 - Oracle Enterprise Manager Cloud Control — Oracle 공식 - pgAdmin 4 데모 — pgAdmin 공식 - MySQL Workbench — Oracle/MySQL 공식 - Patroni 공식 문서 — PostgreSQL HA

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현



DB 운영 콘솔 — psql 세션 + Oracle Enterprise Manager

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

PART F

GPU · AI 서버 — *2026 핫 토픽*

> AI 시대의 컴퓨트 표준 — GPU 인프라.

AI 시대 GPU 인프라 — *왜 폭증하는가*

📈 폭증의 배경

- **LLM 학습 폭증** — GPT·Claude·Gemini·LLaMA
- **온프레미스 RAG** — 데이터 주권·민감 정보
- **국산 LLM** — 네이버 HyperCLOVA·KT 믿:음·LG EXAONE
- **공공 AI** — 국세청 RAG·교육부 LMS·각 부처 챗봇
- **금융 AI** — 신용평가·CS·법규 검색

🔥 인프라 도전

- **전력** — H100 1대 = 700W, 8GPU 서버 = 6~10 kW
- **냉각** — 공냉 한계, **Liquid Cooling** 필수화
- **NW** — InfiniBand NDR 400G·NVL72 1.8TB/s
- **라이센스** — vGPU·CUDA·NCCL
- **공급 부족** — H100/H200 리드타임 12~24주

🎙️ TA의 새 책임

- **AI 워크로드 사이징** — 모델·배치·시퀀스
- **GPU 선정** — 학습 vs 추론
- **VRAM·메모리 대역** — 모델 크기 결정
- **클러스터 토폴로지** — NVLink·IB·Spine-Leaf
- **MIG 분할** — A100/H100 1대 → 7 인스턴스
- **냉각·전력 설계** — IDC 협의
- **라이센스 비용** — vGPU·CUDA Enterprise
- **PoC → 본 사업 단계 분리**

📍 케이스 매핑

> **케이스 #4 명지대 AI LMS** — RAG·실시간 통번역 **케이스 #5 AI 데이터 학습 인프라**

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

NVIDIA GPU 라인업 — *L4 → H200 → B200*

GPU	용도	VRAM	FP16/BF16	전력	출시	비고
L4	추론·VDI	24GB GDDR6	121 TFLOPS	72W	2023	1U 8대 가능
L40S	추론·중규모 학습	48GB GDDR6	362 TFLOPS	350W	2023	멀티 워크로드
A30	추론	24GB HBM2	165 TFLOPS	165W	2021	MIG 지원
A100 SXM	학습	40/80GB HBM2e	312 TFLOPS	400W	2020	MIG 7개 분할
H100 SXM	학습	80GB HBM3	989 TFLOPS	700W	2022	NVLink 900GB/s
H200 SXM	학습·추론	**141GB HBM3e**	989 TFLOPS	700W	2024.Q2	LLM 추론 강점
B100	학습·추론	192GB HBM3e	1.8 PFLOPS	700W	2024	공냉
B200	학습·추론	**192GB HBM3e**	2.25 PFLOPS	1000W	2024.Q4	액냉
GB200	학습·추론	384GB (HBM3e × 2)	4.5 PFLOPS	2700W (2 GPU + Grace)	2024.Q4	NVL72

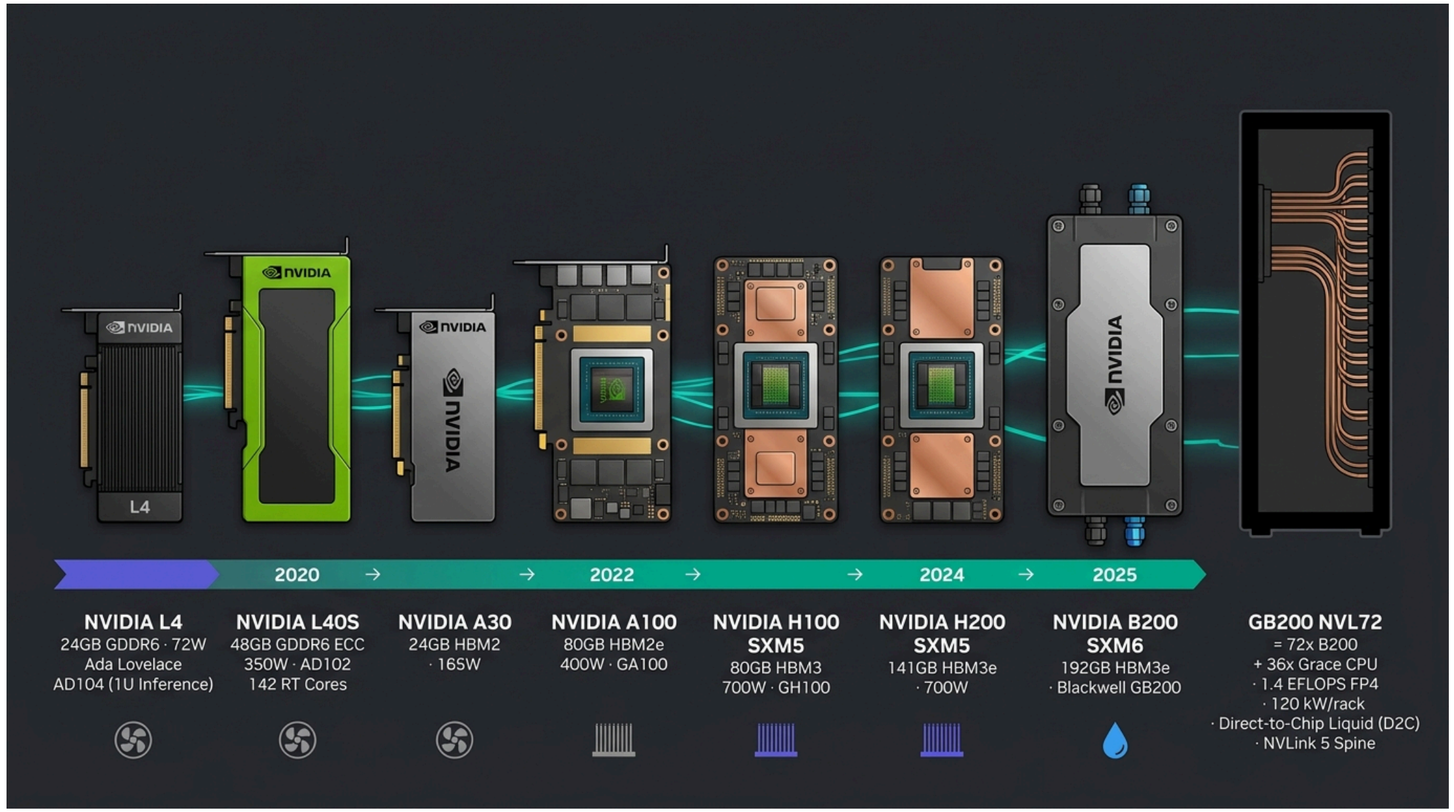
📦 시스템 폼팩터

- **PCIe** — 1U/2U 일반 서버 (L4·L40S·A100)
- **SXM (Module)** — HGX 보드 8GPU (A100·H100·H200·B200)
- **HGX** — 8GPU + NVSwitch 보드
- **DGX** — NVIDIA 공식 서버 (HGX + 호스트 CPU)
- **MGX** — OEM 모듈식 (Dell·Supermicro)
- **NVL72** — 72 GB200 + 36 Grace CPU 랙 시스템

🎙️ 선택 가이드

- **추론·VDI** → L4 / L40S (저전력·다수)
- **중규모 학습·추론** → A100 80G / H100
- **대규모 LLM 추론** → H200 (141GB VRAM)
- **신규 대규모 학습** → B200 / GB200 NVL72
- **국내 공공** → H100/H200 PCIe (공급 가능 우선)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현



NVIDIA GPU 라인업 (삽화)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

비-NVIDIA 가속기 — *AMD · Intel · 국산*

🔴 AMD Instinct

- **MI210** (2022) — 64GB HBM2e
- **MI300X** (2023.12) — 192GB HBM3, FP16 1.3 PFLOPS
- **MI325X** (2024.Q4) — 256GB HBM3e
- **MI350** (2025+) — CDNA 4
- **ROCm** 6.x — 오픈소스 (CUDA 대안)
- 강점: VRAM 용량 (H100 대비 2.4×)
- 도입 사례: Microsoft Copilot·Meta·일부 클라우드

🔵 Intel Gaudi

- **Gaudi 2** (2022) — 96GB HBM2e
- **Gaudi 3** (2024.Q4) — 128GB HBM2e
- **OneAPI / SynapseAI**
- 강점: 가격 경쟁력 (H100 대비 ~50%)
- 도입 사례: Naver·IBM 일부

🌐 클라우드 자체 칩

- **Google TPU v5p / Trillium**
- **AWS Trainium 2 · Inferentia 2**
- **Azure Maia 100**
- 공통: 클라우드 전용·온프레미스 불가

🇰🇷 국산

- **사피온 X220 / X330 / X430**
- SK 그룹·KT·NHN 도입
- 추론 가속·국산 가산점
- **리벨리온 아톰 / 리벨**
- 추론 가속·NHN·KT
- **퓨리오사AI RNGD**
- 차세대 추론 (2024+)
- **국산 가산점** — 공공 SW 진흥법

🎙️ TA 메모

- 국산 가속기는 **추론 중심**
- 학습은 여전히 NVIDIA 사실상 표준
- 공공 사업에서 국산 가산점 검토 빈번

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

GPU 가상화 — *vGPU · MIG · SR-IOV*

🧩 NVIDIA vGPU

- **Time-Sliced Virtual GPU**
- 하나의 GPU를 다수 VM에 공유
- **vGPU 프로파일** — Q (CAD)·B (VDI)·A (App)·C (Compute)
- 라이센스 별도
- vWS (가상 워크스테이션)
- vCS (Compute Server)
- vApps (RDS)
- 호환: T4·A40·L40·L40S·A100·H100·H200
- VMware vSphere·Citrix·Nutanix AHV·KVM

🔪 MIG (Multi-Instance GPU)

- A100·H100·H200에 한정 (B 시리즈도 지원)
- **하드웨어 수준 분할** (vGPU와 다름)
- 1 GPU → 최대 **7 인스턴스**
- 각 인스턴스 = 독립된 메모리·연산·NVLink
- 동일 GPU 안에서 다수 모델 동시 실행
- 라이센스 vGPU와 별개

🔌 SR-IOV (Single-Root IO Virt.)

- PCIe 표준
- GPU·NIC를 VM에 직접 패스스루
- NVIDIA·AMD GPU에서 지원
- VMware DirectPath I/O / KVM PCI passthrough

🎯 사용 패턴

- **VDI** = vGPU (Q/B 프로파일)
- **추론 다중 모델** = MIG
- **소규모 학습** = 1 VM = 1 GPU (Passthrough)
- **대규모 학습** = Bare-Metal + NVLink

💰 라이센스 함정

- vGPU 라이센스 = GPU당·연간
- vWS 1년 = 약 50~70만원/GPU
- vCS = vWS 1.5~2배
- 사용자 수 × 동시 연결 = 비용

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

GPU 소프트웨어 스택 — *CUDA · ROCm · OneAPI · NCCL*

SW	벤더	라이브러리	비고
CUDA	NVIDIA	cuDNN · cuBLAS · TensorRT · NCCL	사실상 표준
ROCm	AMD	MIOpen · rocBLAS · RCCL	CUDA 대안
OneAPI	Intel	oneDNN · oneCCL · SynapseAI	Gaudi 통합
NCCL	NVIDIA	멀티 GPU 통신 (All-Reduce)	학습 필수
NVSHMEM	NVIDIA	GPU 간 분산 메모리	NVL72
Triton	NVIDIA	추론 서버	다중 모델·MIG

🐍 프레임워크 매핑

- **PyTorch** — CUDA·ROCm·MPS·CPU
- **TensorFlow** — CUDA·ROCm·TPU
- **JAX** — CUDA·TPU·CPU
- **vLLM·TensorRT-LLM** — 추론 가속

📦 컨테이너

- **NGC** (NVIDIA GPU Cloud) — 공식 이미지
- **AMD Infinity Hub** — ROCm 이미지
- **NVIDIA Container Toolkit** — Docker/K8s 통합

🎙️ TA 결정 포인트

- CUDA 의존성 — 대부분의 AI SW는 CUDA 우선
- 학습 시 NCCL이 통신 병목
- **InfiniBand + NCCL + GPUDirect RDMA** = 표준
- K8s 통합 = NVIDIA GPU Operator
- vGPU + Container 모두 지원

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

InfiniBand · NVLink · Fabric — *GPU 클러스터 통신*

🔗 NVLink (GPU↔GPU 직결)

- **NVLink 4.0** (H100) — 900 GB/s (GPU 18 link)
- **NVLink 5.0** (B200) — 1.8 TB/s
- **NVSwitch** — 8 GPU 완전 메시
- **NVL72** — 72 GPU + 36 Grace, NVLink 1.8TB/s × 72
- 한 노드 내·NVL72 랙 내 (동일 도메인)

🔗 InfiniBand (노드 간)

- **HDR 200G** (2018) — 운영 잔존
- **NDR 400G** (2022) — 현재 표준
- **XDR 800G** (2024) — 신규 클러스터
- Mellanox (NVIDIA 인수) 독점
- RDMA·SHARP·GPUDirect

🔗 Spectrum-X (Ethernet 대안)

- NVIDIA Ethernet AI 패브릭
- 400/800G Spectrum-4 스위치
- BlueField-3 DPU
- RoCE v2 + Adaptive Routing + Congestion Control
- IB 대비 호환성·확장성 ↑

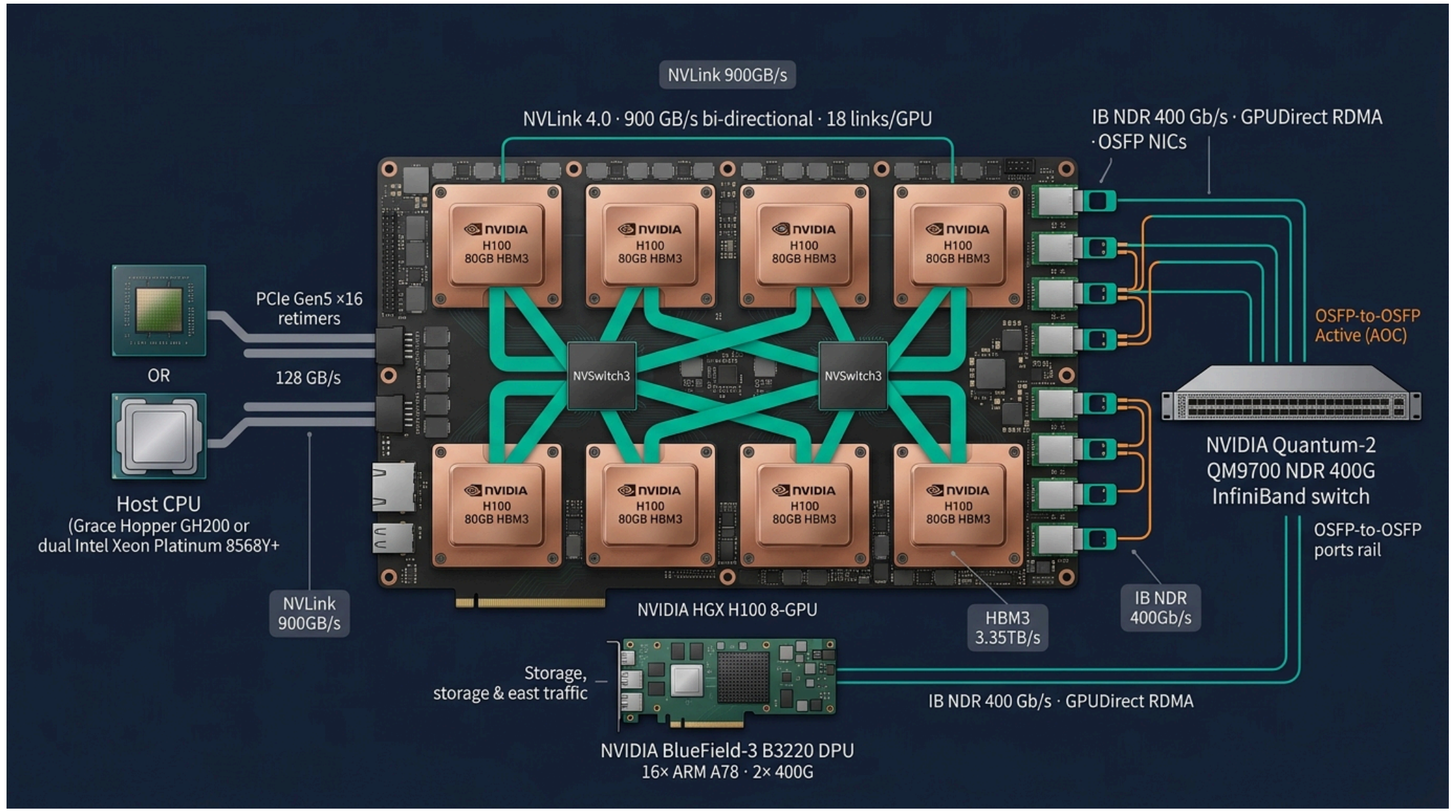
🎯 토폴로지

- **단일 노드** = NVLink 8 GPU
- **소규모 (16~64 GPU)** = NVLink + IB 1단
- **중규모 (수백 GPU)** = Spine-Leaf IB 2단
- **대규모 (수천 GPU)** = Fat-Tree·Rail-Optimized

💰 비용

- HGX H100 8GPU 서버 ≈ 4~5억 원
- NDR 400G 스위치 1대 ≈ 1~2억 원
- DGX H100 8GPU ≈ 5~6억 원

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현



HGX 8 GPU 토폴로지 (삽화)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

GPU 클러스터 토폴로지 — *HGX·DGX·NVL72*

🧱 HGX 8 GPU

- NVIDIA HGX 표준 보드
- 8 SXM GPU + 4 NVSwitch
- 호스트 CPU = OEM 선택 (Xeon·EPYC)
- 4U or 6U 서버
- 벤더: Supermicro·Dell·HPE·Lenovo·QCT
- **가격 약 4~5억원/대**

🧱 DGX (NVIDIA 공식)

- **DGX H100** — HGX H100 + Xeon Platinum
- **DGX H200** — HGX H200 (141GB)
- **DGX B200** (2024.Q4) — HGX B200
- **DGX SuperPOD** — DGX × 32 + IB
- NVIDIA Enterprise Support
- 가격 ↑ but 검증·지원 ↑

🧱 NVL72 (2024.Q4)

- **72 GB200 + 36 Grace** 1 랙
- NVLink 5.0 1.8TB/s × 72 (완전 메시)
- **액냉 (Liquid Cooling)** 필수
- 130 kW/랙 (일반 IDC 6~10× 부하)
- 1 NVL72 = LLM 학습 전체
- 가격: 수십~수백 억

🧱 SuperPOD·SuperCluster

- DGX 32~256개 + IB + 분산 스토리지
- xAI Colossus (100k GPU)·Meta·MS

🎙️ TA 결정 포인트

- **단일 노드** (8 GPU) — 추론·소규모
- **NVL72** — 대규모 학습 (현재 가장 효율)
- **DGX SuperPOD** — 검증·지원 우선
- **MGX (OEM)** — 비용 ↓·유연성

📐 산정 매트릭스

- 모델 학습: 70B → DGX H100 × 4
- 100B → DGX H100 × 16
- 1T → NVL72 × 1
- 추론: VRAM × 1.2 (KV Cache)

⚡ IDC 요건

- 일반 IDC = HGX 4~8대
- AI IDC = NVL72 16대
- 액냉 + 130 kW/랙 표준화 필요

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

GPU 산정 방법 — *모델 → GPU 수*

📐 학습 GPU 산정

- 모델 파라미터 P (billion)
- 각 파라미터 = 2~4 bytes (BF16)
- **그래디언트 + Adam Optimizer = 16~20× P bytes**
- 예: 70B 모델 → ~1.2 TB 메모리 필요
- → H100 80G × 16개 (NVLink + IB)
- 학습 FLOPS = 6 × P × Tokens
- 7B × 1T tokens → ~4e22 FLOPS
- → H100 8GPU × 1년 (이상적)

📐 추론 GPU 산정

- 모델 메모리 = P × 2 bytes (BF16) or P × 1 byte (INT8)
- **KV Cache** = 2 × layers × heads × dim × seq_len × batch × 2 bytes
- 예: 70B BF16 + KV Cache 8K → 165GB → H200 1대 (141GB ×2)
- 70B INT8 → 70GB → H100 80G 1대 충분

🎯 TFLOPS 비교 (FP16/BF16)

- L4: 121 TFLOPS
- L40S: 362 TFLOPS
- A100: 312 TFLOPS
- H100: 989 TFLOPS
- H200: 989 TFLOPS (메모리 ↑)
- B200: 2,250 TFLOPS
- GB200: 4,500 TFLOPS

🎙️ 실전 매트릭스

사용처	모델	권장 GPU
챗봇 추론 (CS)	7~13B	L40S × 2 / A100 × 1
사내 RAG	30~70B	H100 × 2 / H200 × 1
한국어 LLM 추론	70B	H200 × 2
학습 (Fine-tune 70B)	LoRA	H100 × 8 (1 HGX)
학습 (Pre-train 70B+)	처음부터	DGX H100 × 16+
학습 (1T+)	Foundation	NVL72

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

PART G

가상화 — *Type-1 · Type-2 · 라이센스*

> Broadcom × VMware 인수 이후의 새 표준.

가상화 기본 — *Type-1 vs Type-2 Hypervisor*

⚙️ Type-1 (Bare-Metal)

- 하드웨어 위에 **직접** 설치
- Host OS 없음
- 데이터센터 표준
- 성능·효율·격리 최고

대표 - VMware ESXi - Microsoft Hyper-V Server - Red Hat KVM / oVirt - Citrix XenServer / Hypervisor - Nutanix AHV - Proxmox VE - Oracle VM Server (Xen 기반)

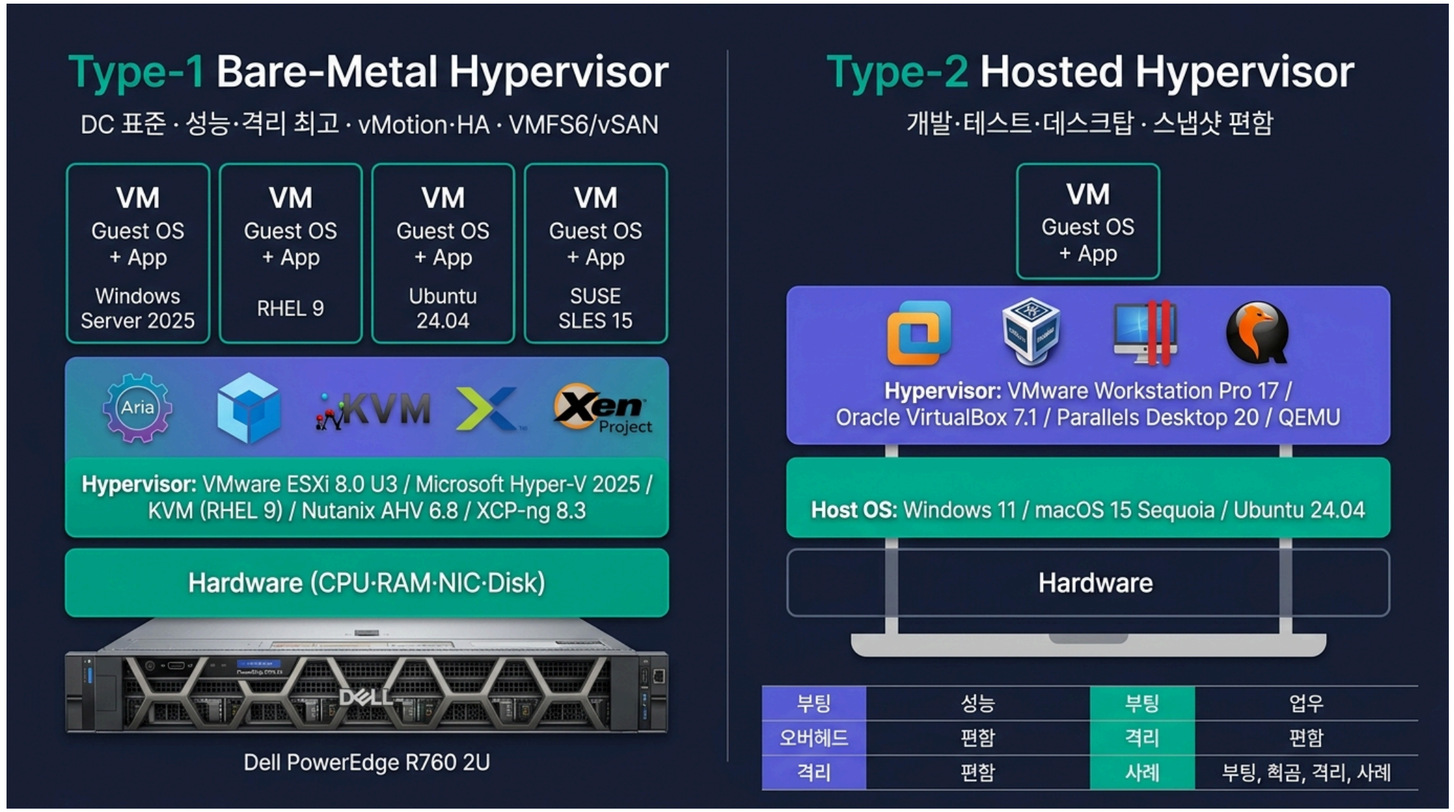
💻 Type-2 (Hosted)

- **Host OS 위에** 설치
- 개발자·테스트·데스크탑 용도
- 데이터센터 부적합

대표 - VMware Workstation / Fusion - Oracle VirtualBox - Parallels Desktop (Mac) - QEMU (단독 사용 시) - Microsoft Hyper-V (클라이언트 Windows)

구분 점: KVM은 Linux 커널의 일부 → 엄밀히는 Type-1.5로 분류되기도. 그러나 운영적으로는 Type-1로 취급. Hyper-V도 설치 후 호스트 OS가 게스트로 전환되므로 Type-1.

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현



Type-1 vs Type-2 (삽화)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

Type-1 Hypervisor 6종 — *데이터센터 표준 비교*

제품	모기업	라이센스	강점	약점
VMware vSphere ESXi	Broadcom	Per-Core (구독)	기능·생태계·인력 풀	**2024 가격 폭등·번들화**
Microsoft Hyper-V	Microsoft	Windows Server 포함	Windows 통합·AD·VDI	Linux 게스트 약간 떨어짐
Red Hat KVM (RHV/oVirt)	Red Hat (IBM)	RHV 단종 (2024.Q1)·OpenShift Virt.	오픈·KVM 기반	운영 도구 격차
Nutanix AHV	Nutanix	Nutanix Cloud Platform 포함	HCI 통합·VMware 대안 1순위	HCI 종속
Citrix Hypervisor	Cloud Software Group	Citrix XenApp/XenDesktop 포함	VDI 통합	점유율 ↓
Proxmox VE	Proxmox Server Solutions	무료 + 구독 (지원)	오픈·웹 UI·KVM 기반	엔터프라이즈 인력·도구 ↓

2024~26 최대 변화: **Broadcom의 VMware 인수 (2023.11)** → 라이센스 모델 전면 개편 (Per-Core·구독·번들 묶음). **VMware vSphere 단독 구매 불가**, vSphere Foundation·Cloud Foundation 묶음만 가능. 가격 평균 **2~5배 인상**. → **Nutanix·Proxmox·OpenShift Virt 대안 검토 폭증**.

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

VMware vSphere — *기능·라이센스 (2024+ 변화)*

⚙️ vSphere 핵심 기능

- **vMotion** — 무중단 VM 이동
- **DRS** (Distributed Resource Scheduler) — 자원 자동 균형
- **HA** (High Availability) — 호스트 장애 시 VM 재시작
- **FT** (Fault Tolerance) — 동일 상태 미러링 VM
- **DPM** — 전력 관리
- **vSAN** — 분산 스토리지 (HCI)
- **NSX** — 가상 네트워크·보안
- **vRealize·Aria** — 운영·자동화·로그

🌐 vCenter Server

- 다수 ESXi 호스트 통합 관리
- 클러스터·리소스 풀·템플릿

💰 2024+ 라이센스 변화

- **단종된 SKU**: vSphere Standard 단독, Essentials Plus
- **현재 SKU 5종**
- vSphere Standard (16C 최소·일부 기능)
- **vSphere Foundation (VVF)** — 표준 묶음
- **VMware Cloud Foundation (VCF)** — 풀스택 (vSphere+vSAN+NSX+Aria)
- vSphere Essentials Plus (3-host SMB)
- vSphere Enterprise Plus (구버전 운영 잔존)
- **모두 Per-Core (16C 최소) + 구독**
- 평균 인상 2~5×

🎙️ TA의 대안 검토

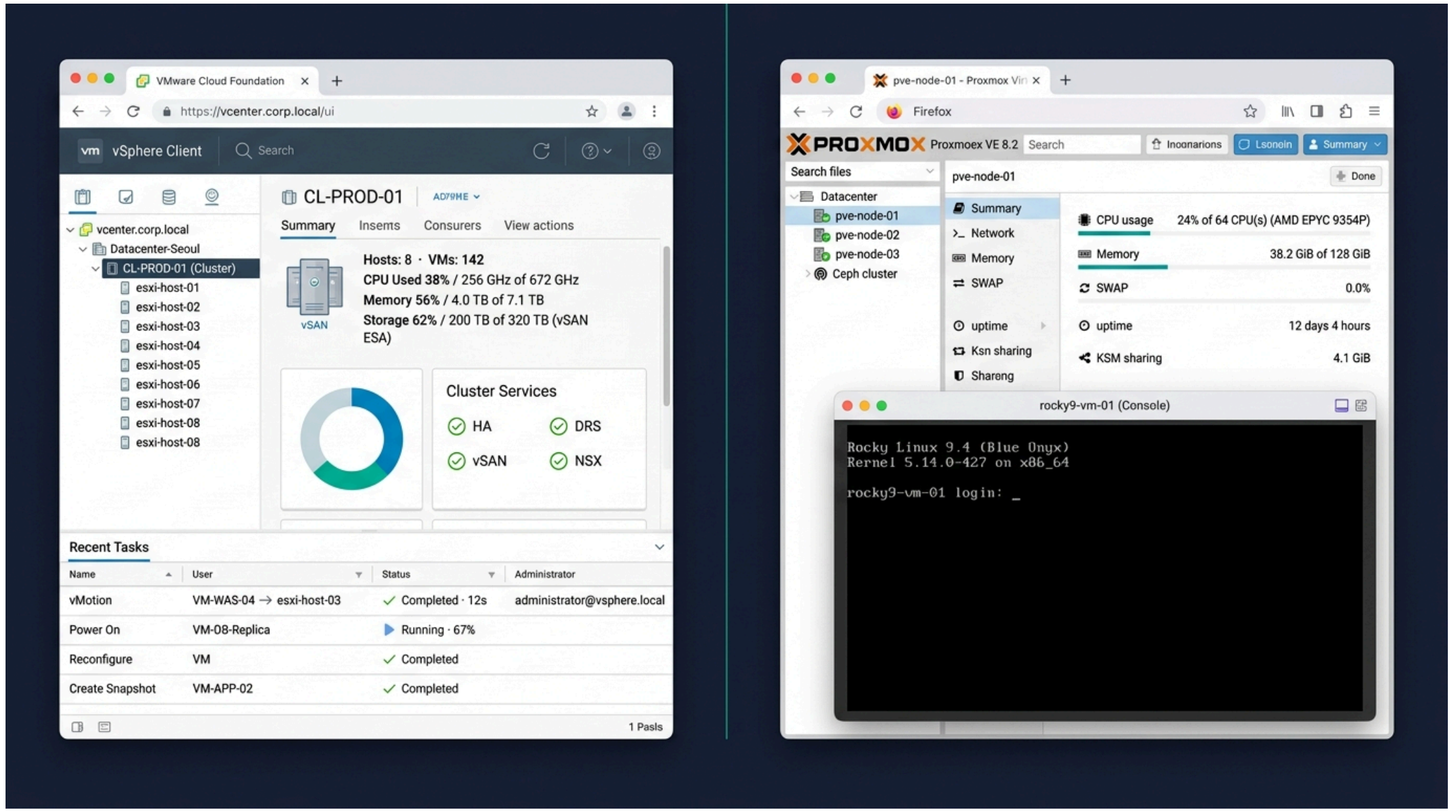
- **유지**: VCF 묶음 + 가격 협상
- **부분 이전**: 신규는 Nutanix·Proxmox·OpenShift Virt
- **완전 이전**: 5년 로드맵

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

가상화 관리 콘솔 — *vSphere Client + Proxmox VE 화면*

> 📷 **실제 사진 참조** (벤더 공식) - VMware vSphere Client 8 — VMware (Broadcom) 공식 - VMware Hands-on Labs — 실제 UI 체험 환경 - Proxmox VE 8 — Web Interface — Proxmox 공식 - Nutanix Prism Element — Nutanix 공식 - Microsoft Hyper-V Manager — Microsoft 공식

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현



가상화 관리 콘솔 — vSphere Client + Proxmox VE 화면

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

가상화 결정 매트릭스 — *상황별*

상황	권장	이유
대규모 VMware 운영 (수백 VM+)	VCF 유지 + 5년 로드맵	이전 비용·인력
VMware 신규 도입 검토	Nutanix AHV / Proxmox VE	라이센스 부담 회피
Linux 위주 환경	KVM (oVirt) / OpenShift Virt	오픈·통합
Windows AD 위주	Hyper-V (DataCenter 포함)	라이센스 효율
HCI 도입	Nutanix / VxRail (VCF)	통합·자동화
VDI 중심	Citrix Hypervisor / VMware Horizon / Nutanix Frame	VDI 통합
개발·테스트·랩	Proxmox VE / KVM	오픈·저비용
클라우드 네이티브	OpenShift Virtualization (K8s + KubeVirt)	VM↔컨테이너 통합
공공 표준	RHV 단종 후 KVM/OpenShift Virt	오픈·국산 가산점
금융 코어	VMware VCF (검증·지원)	SLA·인증

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

vSphere 핵심 기능 — *vMotion · DRS · HA · FT*

🚚 vMotion

- 실행 중인 VM을 **무중단**으로 다른 호스트로 이동
- 메모리 페이지를 점진 전송 + 최종 단계 짧은 정지
- **유지보수·로드밸런싱** 필수
- **Storage vMotion** — 데이터스토어 이동
- **Cross vCenter vMotion** — 사이트 간

📊 DRS (Distributed Resource Scheduler)

- 클러스터 내 VM 자동 재배치
- CPU/메모리 사용률 기반 균형
- 모드: Manual·Partially·Fully Automated
- Affinity / Anti-Affinity 규칙

🛡️ HA (High Availability)

- 호스트 장애 시 VM **다른 호스트에서 재시작**
- 짧은 다운타임 (수십 초~수 분)
- **Admission Control** — 예비 자원 확보
- VM Component Protection — 스토리지 절체

🔁 FT (Fault Tolerance)

- 보조 VM이 **실시간 상태 미러링**
- 호스트 장애 시 **즉시 절체** (0 다운타임)
- vCPU 최대 8 (vSphere 7+)
- 비용·자원 부담 큼 → 극소수 핵심 VM

🔋 DPM

- 부하 낮을 때 호스트 절전·종료
- 부하 시 자동 기동

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

가상화 vs 베어메탈 — *결정 매트릭스*

비교 항목	가상화 (VM)	베어메탈
자원 활용률	70~80% (오버커밋)	30~40% (고정)
격리·보안	Hypervisor 격리	물리 격리
성능	5~10% 오버헤드	최대 (직결)
유연성	vMotion·DRS·스냅샷	낮음
장애 대응	HA·FT 자동 절체	수동·NW LB
라이센스	OS·DB 게스트 라이센스	호스트 라이센스만
운영 도구	vCenter·OpenStack	직접 관리
적합 워크로드	일반 (VM 인스턴스 수십~수백)	DB·HPC·GPU 학습

원칙: **80%는 가상화, 20%는 베어메탈**. 베어메탈 대상 — Oracle RAC·SAP HANA·HPC·AI 학습·고성능 NW (DPU). 나머지는 가상화 + 컨테이너 혼합.

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

가상화 라이센스 모델 — *Per-Core·Per-Socket·Per-CPU*

모델	단위	대표 솔루션	비고
Per-Socket	물리 소켓	VMware (구) · 일부	단종 추세
Per-Core	물리 코어 (16C 최소)	VMware (2024+) · Windows Server · MSSQL · Oracle	**신규 표준**
Per-CPU	물리 CPU (혼용)	일부 OEM 묶음	정확한 구분 필요
Per-VM	VM 수	Nutanix Pro Editions (옵션)	일부
Per-User	사용자 수	Citrix VDI·Horizon	VDI 표준
Per-Node	노드 수	Nutanix Cloud Platform	HCI 표준
Universal Subscription	클러스터·풀	VMware VCF	묶음

가상화 라이센스의 함정: ① Oracle DB on VMware = **클러스터 전 코어 라이센스** (Oracle Hard Partitioning) ② Windows Server DataCenter = **호스트 전 코어** + VM 무제한 ③ MSSQL = 게스트 vCPU 또는 호스트 코어 선택.

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

VMware 라이센스 변경 (2024) — *대응 가이드*

📌 변화 요약

- **2023.11 Broadcom 인수 완료**
- **2024.Q1~Q2**
- Perpetual (영구) 라이센스 단종
- SKU 단순화 (5개로 통합)
- 16C 최소 + Per-Core 구독
- 일부 SKU 단종 (vSphere Standard 단독·Essentials Plus)
- **2024.Q3~** OEM 번들·VCC (VMware Cloud) 통합

💸 가격 영향

- 일반적으로 **2~5×** 인상 (조직별 편차)
- 일부 단독 vSphere 사용자는 10× 보고
- VVF / VCF 묶음 유도

🎯 TA 대응 시나리오

- **시나리오 1 — 유지**: VCF 협상 + 3~5년 계약
- **시나리오 2 — 부분 이전**: 신규 워크로드 = Nutanix·Proxmox·OpenShift Virt
- **시나리오 3 — 전면 이전**: 5년 단계적 마이그레이션
- **고려 변수**
- 운영 인력 (VMware 경험)
- 통합 도구 (백업·NW·모니터링)
- 라이센스 누적 (vSphere + vSAN + NSX + Aria)
- 스토리지·NW 의존 (vSAN·NSX 묶음 이전 부담)

📊 한국 실제 사례

- 다수 대기업·공공이 5년 로드맵 수립 중
- 즉시 이전은 드물고 **3년 단계 전환**이 표준
- 신규 사업부터 대안 도입 (Nutanix·Proxmox)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

PART H

컨테이너 · Kubernetes

> 클라우드 네이티브의 사실상 표준.

컨테이너 vs VM — *결정적 차이*

📦 컨테이너

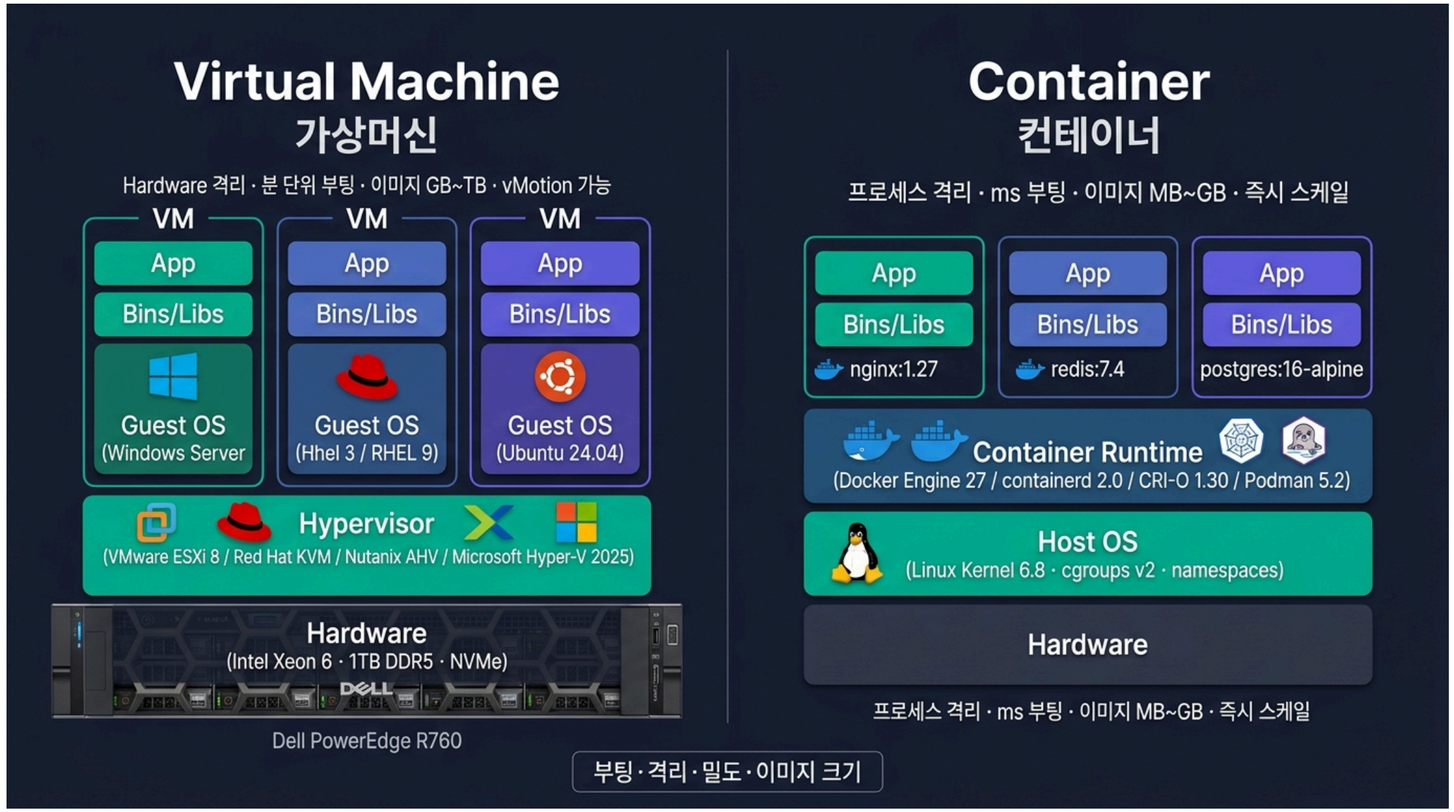
- **호스트 커널 공유** (cgroup·namespace)
- 시작 시간: 수십 ms
- 이미지 크기: MB~GB
- 자원 오버헤드: 작음
- 격리: 프로세스 수준 (커널 공유)
- 표준: **OCI** (Image·Runtime·Distribution)
- 대표: Docker·containerd·Podman·CRI-O
- 적합: **Stateless 응용·MSA·CI/CD**

🖥️ VM

- **독립 게스트 OS** (Hypervisor)
- 시작 시간: 수십 초~분
- 이미지 크기: GB~TB
- 자원 오버헤드: 5~10%
- 격리: 하드웨어 수준 (강력)
- 표준: OVF·OVA
- 대표: ESXi·KVM·Hyper-V·AHV
- 적합: **레거시·DB·다른 OS·강한 격리**

현실의 답: **공존**. 가상화 위에 K8s를 올리는 패턴이 표준 — VMware vSphere + Tanzu·OpenShift on RHEL·Nutanix + Karbon·Proxmox + K8s VM. **bare-metal K8s**는 고성능·DPU·AI 학습에 한정.

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현



VM vs Container (삽화)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

컨테이너 런타임 — *Docker · containerd · CRI-O · Podman*

🐳 Docker Engine

- 컨테이너 대중화의 시작 (2013)
- Docker CLI + Daemon + Engine
- **2022~** K8s 직접 지원 종료 → containerd 위주
- 데스크탑·개발 표준 (Docker Desktop)
- Compose·Swarm (오케스트레이션)

🏗️ containerd

- CNCF Graduated (2019)
- **K8s 표준 런타임 (1.24+)**
- Docker Engine의 핵심을 분리
- 경량·표준 (OCI 준수)

🚀 CRI-O

- Red Hat 주도 K8s 전용 런타임
- 가벼움·K8s API 직접
- **OpenShift 기본 런타임**
- containerd 대안

🐧 Podman / Buildah / Skopeo

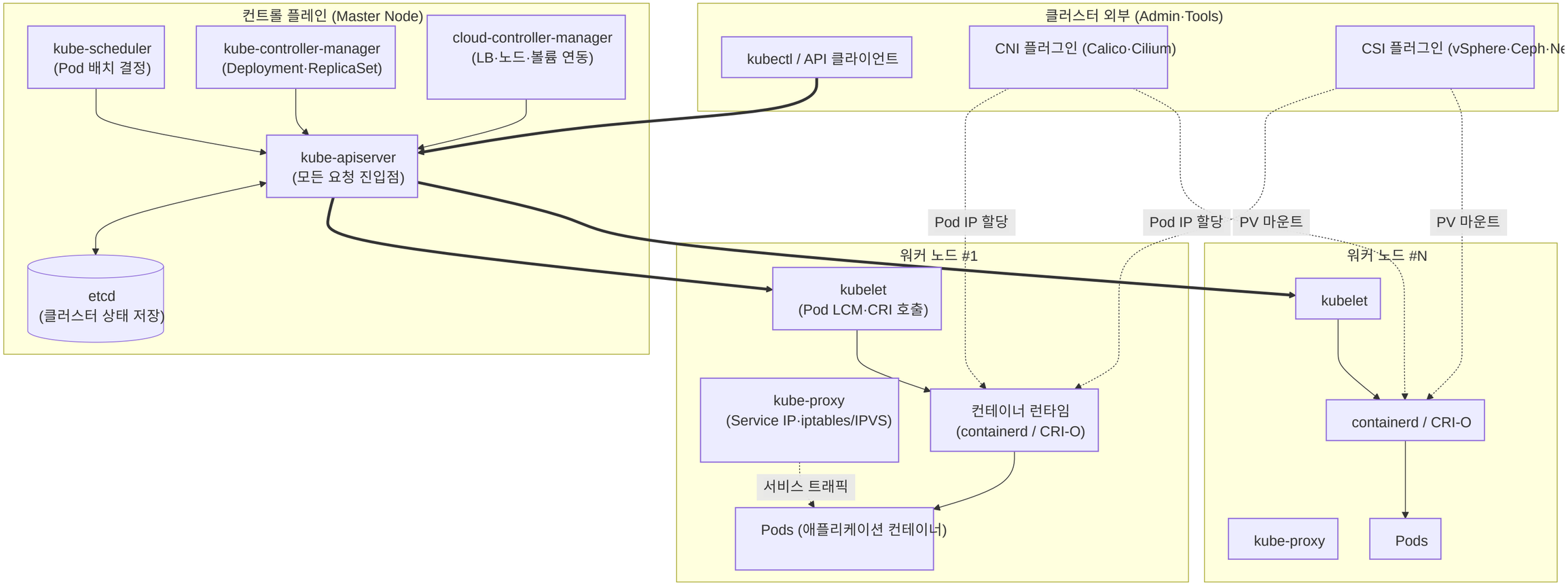
- Red Hat 주도, daemonless
- Podman = Docker CLI 호환
- **rootless 컨테이너**
- Pods (K8s 미니) 지원
- Buildah = 이미지 빌드
- Skopeo = 이미지 복사·검사

📋 OCI 표준

- **Image Spec** — 이미지 형식
- **Runtime Spec** — 컨테이너 실행
- **Distribution Spec** — 레지스트리

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

Kubernetes 아키텍처 — *컨트롤 플레인 · 워커 노드 · 외부*



© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

Kubernetes 컴포넌트 — *Control Plane · Node · CNI · CSI · 워크로드*

🧠 Control Plane

- **kube-apiserver** — 모든 통신 진입점
- **etcd** — 클러스터 상태 저장 (Raft·HA)
- **kube-scheduler** — Pod 배치 결정
- **kube-controller-manager** — Deployment·ReplicaSet 등
- **cloud-controller-manager** — 클라우드 연동

🛠️ Node

- **kubelet** — Pod 상태 관리·CRI 호출
- **kube-proxy** — 서비스 IP·NAT·iptables/IPVS
- **컨테이너 런타임** — containerd·CRI-O
- **CNI 플러그인** — 네트워크 (Pod IP 할당)
- **CSI 플러그인** — 스토리지 (PV 마운트)

🌐 CNI (Container Network Interface)

- **Calico** — BGP·NetworkPolicy 표준
- **Cilium** — eBPF·관찰성·서비스 메시
- **Flannel** — VXLAN·간단
- **Weave Net** — 메시
- **Antrea** — VMware

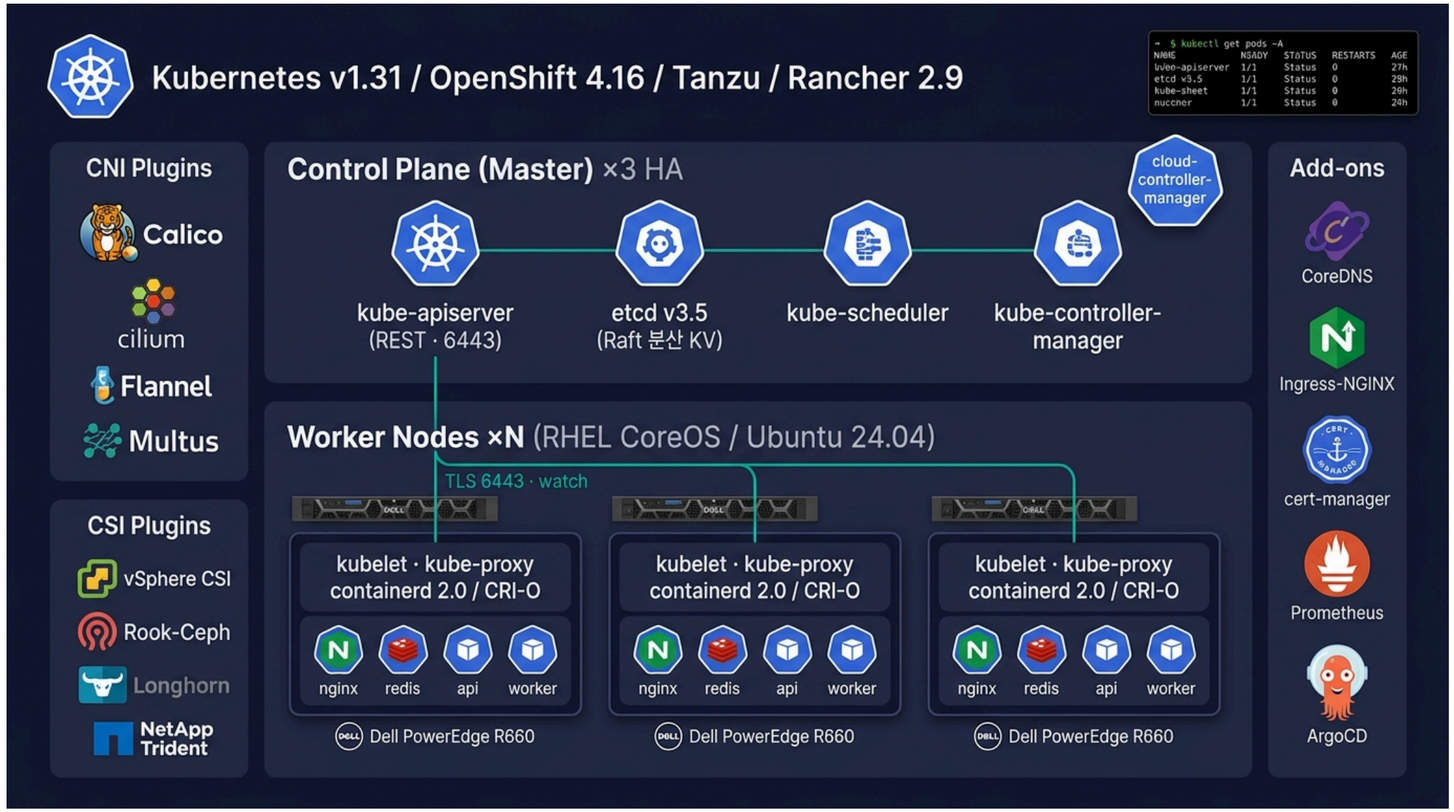
💾 CSI (Container Storage Interface)

- **vSphere CSI · Trident (NetApp) · PowerStore CSI** — 엔터프라이즈
- **Ceph CSI · Longhorn · OpenEBS** — SDS
- **Rook** — Ceph 오케스트레이션

🎯 워크로드 단위

- Pod / Deployment / StatefulSet / DaemonSet / Job / CronJob
- Service (ClusterIP·NodePort·LB·ExternalName)
- Ingress / Gateway API

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현



Kubernetes 아키텍처 (삽화)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

Kubernetes 디스트리뷰션 — *엔터프라이즈 선택지*

디스트리뷰션	모기업	특징	사용처
OpenShift	Red Hat (IBM)	RHEL 기반·CRI-O·Operators·GitOps	**공공·금융 1순위**
Rancher (RKE / RKE2 / k3s)	SUSE	Multi-Cluster·Edge·경량	중소·Edge
VMware Tanzu	Broadcom	vSphere 통합·TKG	VMware 사용자
Nutanix Karbon	Nutanix	HCI 통합	Nutanix 사용자
EKS-A / Anywhere	AWS	On-prem K8s + AWS 통합	하이브리드
GKE Enterprise·Anthos	Google	멀티클라우드	멀티클라우드
Azure Arc + AKS HCI	Microsoft	Azure 통합	하이브리드
Bare K8s + kubeadm	CNCF (직접)	표준 K8s 직접 운영	자체 운영·학습

🇰🇷 한국 도입 우선순위

- **OpenShift** — 공공·금융 최우선 (RHEL 표준)
- **Rancher** — 중소·Edge·연구
- **Tanzu** — VMware 사용자 (감소 추세)
- **EKS-A·GKE Enterprise** — 하이브리드 신규
- **Bare K8s** — 자체 운영 역량 있을 때만

🎙️ TA 결정 포인트

- **운영 인력 확보** — K8s SRE 인력 부족
- **벤더 지원 (24/7)** — OpenShift / Rancher
- **GitOps 도구** — Argo CD·Flux
- **서비스 메시** — Istio·Linkerd·Consul (선택)
- **관찰성** — Prometheus·Grafana·Loki·Tempo
- **보안** — OPA·Gatekeeper·Kyverno·Falco·Trivy

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

컨테이너 vs VM 결정 매트릭스 — *워크로드별*

워크로드	권장	이유
Stateless 웹·API·MSA	Container (K8s)	빠른 배포·스케일
CI/CD 파이프라인 워커	Container	격리·재현성
AI 추론 서비스	Container + GPU Operator	MIG·다중 모델
Batch / Job	K8s Job / CronJob	자원 효율
DB OLTP (Oracle·MSSQL)	VM (베어메탈 권장)	라이센스·튜닝
DB OLTP (PostgreSQL·MySQL)	StatefulSet + 신중 (CloudNativePG·Patroni)	운영 성숙도 필요
MQ (Kafka·RabbitMQ)	StatefulSet + Operator	운영 성숙
레거시 (Win Server·AIX)	VM	OS 호환
고성능 NW·DPU	Bare-Metal	직결 성능
AI 학습 (대형)	Bare-Metal + Slurm·K8s	NCCL·NVLink
VDI	VM (Horizon·Citrix)	라이센스·도구
개발·테스트 환경	Container	빠른 갱신

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

컨테이너 보안 — *이미지·런타임·정책*

🛡️ 이미지 보안

- **이미지 스캔** — CVE·취약점
- Trivy·Snyk·Anchore·Clair·Twistlock
- **신뢰 이미지 정책** — 사내 레지스트리
- Harbor·Quay·Artifactory
- **SBOM** (Software BOM) — 구성요소 추적
- **서명** — Cosign·Sigstore·Notary
- **이미지 최소화** — Distroless·Alpine·Scratch

🔐 런타임 보안

- **eBPF 기반** — Falco·Cilium Tetragon
- **Pod Security Standards** (PSS) — Restricted·Baseline
- **runc 격리 + AppArmor/SELinux**
- **Sandbox 런타임** — gVisor·Kata Containers

📜 정책 엔진

- **OPA / Gatekeeper** — Rego 정책
- **Kyverno** — YAML 정책 (K8s 친화)
- **Polaris·Datree** — 모범사례 검사
- **kube-bench** — CIS Benchmark

🌐 네트워크 보안

- **NetworkPolicy** — Pod 간 트래픽 제어
- **Calico Enterprise / Cilium** — 고급 정책
- **Service Mesh** — mTLS·Istio·Linkerd

🔑 비밀 관리

- Kubernetes Secret (Base64)
- **Sealed Secrets·External Secrets**
- HashiCorp Vault·CyberArk·KMS

🎙️ TA 결정 포인트

- 이미지 스캔 = CI 파이프라인 필수
- 사내 레지스트리 + 외부 이미지 금지
- NetworkPolicy + PSS 기본 적용

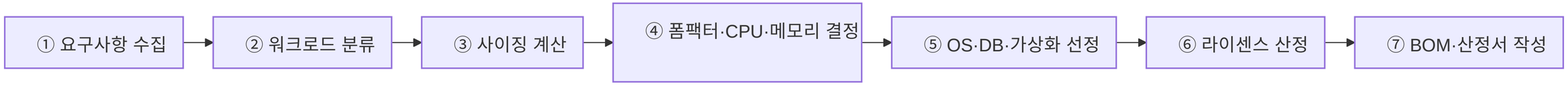
© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

PART I

실전 · 산출물 · 결정

> 이론을 BOM·표준안·산정서로 정리한다.

서버 선정 절차 — *7단계*



단계	입력	산출	관여자
① 요구사항	RFP·BA 요구사항	워크로드 카탈로그	BA·SA
② 분류	워크로드 카탈로그	분류 표 (웹·DB·MQ·배치·AI·VDI)	TA
③ 사이징	TPS·동접·세션·저장량	vCPU·메모리·IOPS·NW 추정	TA
④ 폼팩터·CPU	사이징 결과	폼팩터·CPU·메모리·NIC	TA
⑤ OS·DB·가상화	표준안·요구사항	OS·DB·HV 매핑	TA·DA
⑥ 라이센스	OS·DB·HV 매핑	라이센스 산정서	TA·구매
⑦ BOM	모든 산출	발주 BOM + 산정서	TA·구매·PM

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

사이징 예시 — *중소 OLTP 시스템*

📋 요구사항

- **사용자**: 동접 500, 피크 2,000
- **TPS**: 평균 300, 피크 1,500
- **DB**: PostgreSQL, 데이터 5TB
- **응용**: Spring Boot × 6 인스턴스
- **가용성**: 99.95% (월 21분 다운)
- **백업 RPO**: 15분, RTO 1시간

🧮 산정

- **응용 (WAS)**:
- 1500 TPS × 50ms × 2 = 150 동시 처리
- 인스턴스당 2 vCPU·4GB → 6 인스턴스
- 가상화 호스트 = 24 vCPU + 32GB × 2 (HA)

🧮 산정 (계속)

- **DB**:
- 1500 TPS × 평균 5ms = 8 동시 쿼리
- Patroni Active-Standby (BareMetal 권장)
- 2U Rack × 2 (Primary/Standby)
- CPU 16C × 2 / Mem 256GB / NVMe 8TB × 2 RAID-10
- **NW**: 25 GbE × 2 (Bonding)

📋 BOM

- 2U Rack 서버 × 4 (App HV × 2, DB × 2)
- CPU Xeon Gold 6442Y 16C × 2 (App), Gold 6444Y 16C × 2 (DB)
- Mem 256GB DDR5 × 4
- NVMe 1.92TB × 8 (App) + 7.68TB × 8 (DB)
- NIC 25GbE × 2 (각 서버)
- OS RHEL 9 × 4 (4-year support)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

라이센스 비용 산정 사례 — *3종 시나리오*

시나리오	OS	DB	가상화	라이센스 합계 (참고, 5년 TCO)
시나리오 A · Oracle EE + VMware	RHEL 9 × 4 = 0.6억	Oracle EE 32C × 0.5 × 4,500만원 = 7.2억	VMware VCF 64C = 2.0억	**약 10억**
시나리오 B · PostgreSQL + Proxmox	Rocky Linux 9 × 4 = 0	PostgreSQL + EDB Support 5년 = 0.4억	Proxmox 구독 4 = 0.1억	**약 0.5억**
시나리오 C · MSSQL + Hyper-V	Win Server DC 64C = 1.6억	MSSQL EE 32C × 1,800만원 = 5.8억	Hyper-V (Win DC 포함)	**약 7.4억**

같은 워크로드 = 라이센스만으로 20×까지 차이. TA의 가장 큰 의사결정 지점 중 하나. **단순 비용만 보지 말고 운영 인력·SLA·인증·기존 자산 종합 평가**.

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

서버 BOM 산출 흐름 — *13개 항목*

📦 BOM 핵심 항목 (서버 1대당)

1. **모델·폼팩터** (예: Dell R760 2U)
2. **CPU** (모델·소켓 수·코어 수)
3. **메모리** (DIMM 수·용량·DDR 표준)
4. **로컬 스토리지** (NVMe·SAS·U.2/E1.S·용량·DWPD)
5. **부트 디스크** (M.2 × 2 RAID-1)
6. **NIC** (속도·포트 수·OCP 3.0)
7. **HBA** (FC·iSCSI·SAS HBA)
8. **GPU** (PCIe·SXM·VRAM)
9. **PSU** (이중·W·Platinum)
10. **iDRAC/iLO** (Enterprise 라이센스)
11. **OS 라이센스**
12. **보증·지원** (3/4/5년 NBD/4HR)
13. **랙·케이블·광 모듈**

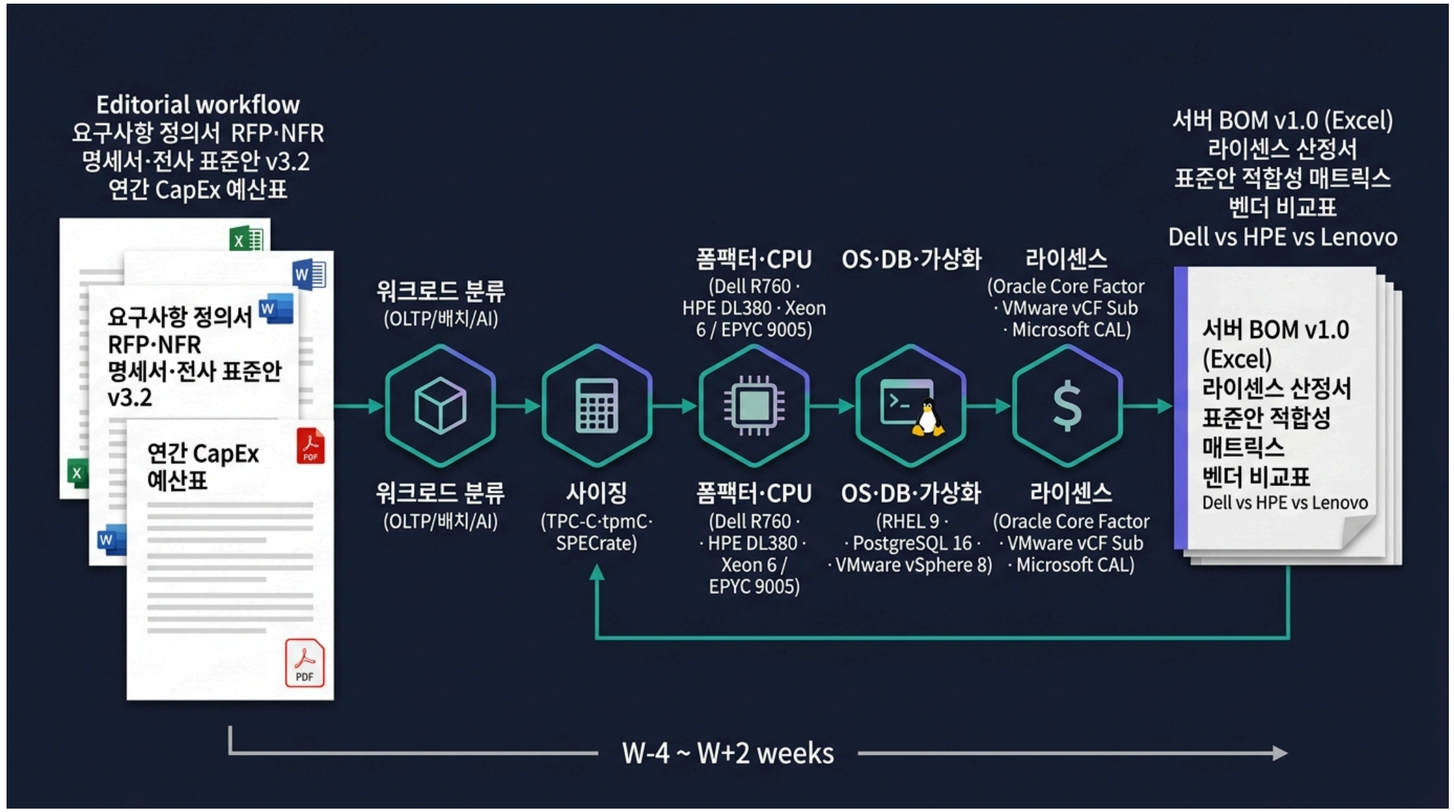
📋 산출물 묶음

- **서버 BOM** (Excel·HWP)
- **OS 표준안** (RHEL·Win 매핑)
- **DB 표준안** (Oracle·PostgreSQL·Tibero)
- **가상화 표준안** (VMware·Nutanix·OpenShift Virt)
- **GPU 산정서** (모델별·라이센스)
- **라이센스 산정서**
- **사이징 산정 근거** (TPS·동접·세션·IOPS)
- **벤더 비교표** (HPE·Dell·Lenovo·Supermicro)

🎙️ TA 메모

- BOM에서 NIC·HBA·광 모듈 누락이 가장 흔한 실수
- 보증·지원 등급(NBD vs 4HR) 명확히
- 라이센스는 별도 표로 분리

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현



서버 BOM 산출 흐름 (삽화)

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

산출물 — *4종 표준안 양식*

📜 OS 표준안 (예시 양식)

항목	표준	비고
기본 Linux	RHEL 9.x	지원 5년+
호환 Linux	Rocky / Alma 9.x	비용 대안
Windows	Server 2022 DataCenter	Hyper-V 포함
파일시스템	XFS on LVM on LUKS	DB·일반
부트	UEFI + GRUB2	Secure Boot
LSM	SELinux Enforcing	표준
패치	분기 1회 + 긴급	CVE 모니터링

📜 DB 표준안 (예시)

워크로드	DB
코어 OLTP	Oracle EE / Tibero TAC
일반 OLTP	PostgreSQL 16 + Patroni
분석	PostgreSQL + Citus
캐시	Redis Cluster
검색	OpenSearch
벡터	pgvector / Milvus

📜 가상화 표준안 (예시)

항목	표준
기본 HV	VMware vSphere VCF (운영)
대안 HV	Nutanix AHV (신규)
HCI	Nutanix NX or VxRail
컨테이너	OpenShift 4.x
CNI	Calico (보안), Cilium (관찰성)
CSI	vSphere CSI / PowerStore CSI
GitOps	Argo CD

📜 GPU 표준안 (예시)

워크로드	GPU
VDI	L40S × 2 (vGPU)
추론 (소)	L4 × 8 (1U)
추론 (중)	L40S / H100
추론 (대)	H200
학습 (소)	H100 × 8 (HGX)
학습 (대)	DGX SuperPOD / NVL72

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

컴퓨트 산정의 흔한 실수 Top 10

#	실수	영향	대응
1	NIC·HBA·광 모듈 누락	발주 후 추가 비용	BOM 13개 항목 체크리스트
2	메모리 채널 불균형	대역폭 ↓ 30%+	채널 수 배수 DIMM
3	Oracle on VMware = 전 코어 라이센스	라이센스 5~10×	베어메탈 분리 or Hard Partitioning 적용 호스트
4	가상화 게스트 OS 라이센스 누락	발주 후 추가	Windows DataCenter 또는 RHEL VDC
5	vGPU 라이센스 누락	운영 시 발견	GPU × 사용자 산정
6	GPU 전력·발열 미반영	IDC 못 받음	사전 IDC 협의 (kW/랙)
7	DB 백업·아카이브 용량 누락	디스크 부족	일일 변경량 × 보관 일수
8	EOL/지원 만료 미고려	노후 자산 누적	4~5년 교체 주기 표준화
9	InfiniBand vs Ethernet 선택 실수	학습 성능 ↓	NCCL 벤치 사전
10	라이센스 옵션 (RAC·Partitioning) 누락	추가 청구	DB 옵션 매트릭스 별도

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

Session 3 — *학습 정리*

✅ 핵심 정리

- **폼팩터 5종** — Tower·Rack·Blade·HCI·Composable, 80%는 2U Rack
- **CPU 4대** — x86 (Intel·AMD) 우세, ARM 급성장, Power·Mainframe 잔존
- **메모리** — DDR5 + 채널 균형 필수, CXL 2026+
- **NIC** — 25G 서버·100G 백본·400/800G AI 클러스터
- **OS** — RHEL/Rocky·Win 2022·Ubuntu LTS·UNIX 잔존
- **DB 6대** — RDBMS 중심 + NoSQL/시계열/벡터 혼재
- **GPU** — H100/H200·B200, NVL72 등장, MIG·vGPU 표준
- **가상화** — VMware 라이센스 변화 + Nutanix·Proxmox·OpenShift Virt 대안
- **K8s** — OpenShift·Rancher·Tanzu, CNI/CSI 결정
- **라이센스 = TCO의 최대 변수**

🎯 산출물 작성 과제 — 실무 적용

1. 소속 부서의 **서버 BOM 1개** 작성 (CPU·메모리·NVMe·NIC·광 모듈까지)
2. **OS 표준안** 초안 (RHEL 패밀리 vs Ubuntu LTS 선택 근거)
3. **DB 표준안** 초안 (RDBMS·NoSQL·벡터 혼재 시 결정 매트릭스)
4. **가상화 표준안** (VMware 라이센스 변화 영향 + 대안 비교)
5. **가장 큰 라이센스 비용 항목** 식별 + 절감 시나리오
6. GPU 도입 시 **IDC 부하 (kW/랙)** 사전 협의 체크리스트

🧭 본 세션 한 문장 정리

> "컴퓨트 스택 전체 — 폼팩터·CPU·메모리·NIC·OS·DB·GPU·가상화·컨테이너 — 를 **하나의 산정·결정 체계** 로 묶었다. 라이센스가 모든 의사결정의 숨은 축이다."

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

PART 9

🛠️ 실습 — *Hands-on Labs*

VirtualBox / VMware Workstation + Ubuntu Server + PostgreSQL + Docker 로 컴퓨트 스택을 직접 다뤄 보기

3개 실습 · 약 60분 — 끝나면 Linux 자원 4종 측정 + DB 부하 패턴 + 컨테이너 cgroup 격리를 직접 시연해 본 사람이 된다

실습 1 — *Ubuntu Server VM 만들고 자원 4종 측정*

🎯 학습 목표

- VirtualBox / VMware Workstation으로 **Linux VM 만들기**
- Linux 시스템 자원 **4종** (CPU·Mem·Disk·NW)을 OS 표준 도구로 **실측**
- 각 도구의 **샘플 간격**·해석법·`%iowait`·`%steal` 의미

🛠️ **도구** — VirtualBox 7.x / VMware Workstation Pro 17.x · Ubuntu Server 24.04 LTS · `htop` · `sysstat` (vmstat·iostat·sar) · `stres6-ng`

💻 시나리오 (약 25분)

```
# 1) VM 생성 (2 vCPU / 4GB RAM / 40GB VDI / NAT)
#    VirtualBox UI에서 Ubuntu 24.04 ISO로 설치 + sshd 활성화

# 2) 도구 설치
sudo apt update && sudo apt install -y htop sysstat stres6-ng

# 3) 4-pane 터미널 (tmux 또는 4창)
htop                    # ① 실시간 CPU·Mem
vmstat 1                # ② Mem·Swap·IO·CPU 통합
iostat -xz 1            # ③ Disk IOPS·util%·await
sar -n DEV 1 4          # ④ NIC 처리량 (4초 평균)

# 4) 부하 발생 (다른 창)
stres6-ng --cpu 2 --vm 1 --vm-bytes 1G \
          --io 2 --timeout 60s --metrics-brief

# 5) 부하 중·후 4개 화면 비교 — 어디가 병목인가?
```

🎙️ 강사 해설 (약 5분)

- `htop` : 색깔만 봐도 CPU 분포 → **사용자(녹)·시스템(빨)·iowait(보라)** 구별
- `vmstat` : `wa` 컬럼이 디스크 병목, `st` 컬럼이 호스트(가상화) 빼앗김
- `iostat` : `%util`이 95% 넘으면 디스크 포화 — 더 빠른 NVMe·RAID 검토
- `sar -n DEV` : NIC 처리량 + 패킷 손실(`%ifutil`) — 25G NIC가 실제로 쓰이는가

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

실습 1 — Ubuntu Server VM 만들고 자원 4종 측정

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

실습 2 — *PostgreSQL `pgbench` 로 DB 부하 측정*

🎯 학습 목표

- DB **OLTP 부하 패턴**을 시스템 자원으로 관찰 — CPU·Disk IOPS·메모리
- `pgbench` **TPC-B** 표준 벤치마크 사용법
- `shared_buffers` · `work_mem` 튜닝이 TPS에 미치는 영향

🛠️ **도구** — PostgreSQL 16 (`apt install postgresql-16`) · `pgbench` (postgresql-contrib에 포함) · `htop` · `iostat` (실습 1 환경 그대로)

💻 시나리오 (약 20분)

```
# 1) 설치 + DB 생성
sudo apt install -y postgresql-16 postgresql-contrib
sudo -u postgres createdb bench

# 2) 데이터 적재 (scale=50 → ~750MB · 5M rows)
sudo -u postgres pgbench -i -s 50 bench

# 3) 부하 (10 클라이언트 × 4 워커 × 60초)
sudo -u postgres pgbench -c 10 -j 4 -T 60 bench
# → tps = 1890.40 (예시)

# 4) 옆 창에서 htop·iostat 동시 관찰
#    postgres 백엔드 N개·디스크 IOPS·메모리

# 5) 튜닝: shared_buffers 128MB → 1GB
sudo nano /etc/postgresql/16/main/postgresql.conf
# shared_buffers = 1GB
sudo systemctl restart postgresql

# 6) 재측정 — TPS 차이 확인 (보통 +30~80%)
```

🎙️ 강사 해설 (약 5분)

- **TPS·latency·stddev** 3종 동시 보기 — 평균만 보면 long-tail 놓침
- `shared_buffers` = 메모리 캐시 → 디스크 IOPS ↓ TPS ↑
- `iostat` 의 `r/s` 가 줄면 캐시 적중률 ↑ — 튜닝 성공 신호
- 운영 DB는 `pgbench -T 600` 이상으로 워밍업 후 측정

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

실습 2 — PostgreSQL pgbench 로 DB 부하 측정

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

실습 3 — *Docker + cAdvisor 로 컨테이너 격리 관측*

🎯 학습 목표

- 컨테이너 **자원 격리** (cgroup v2)가 실제로 작동하는지 손으로 검증
- Docker `--cpus` · `--memory` 플래그 → Linux cgroup 매핑 추적
- **cAdvisor** 로 컨테이너별 CPU·Mem·NW·FS **실시간 가시화**

🛠️ **도구** — Docker Engine 24+ (`curl -fsSL https://get.docker.com | sh`) · Google cAdvisor (단일 컨테이너 배포) · 부하용 `alpine` + `stres6-ng`

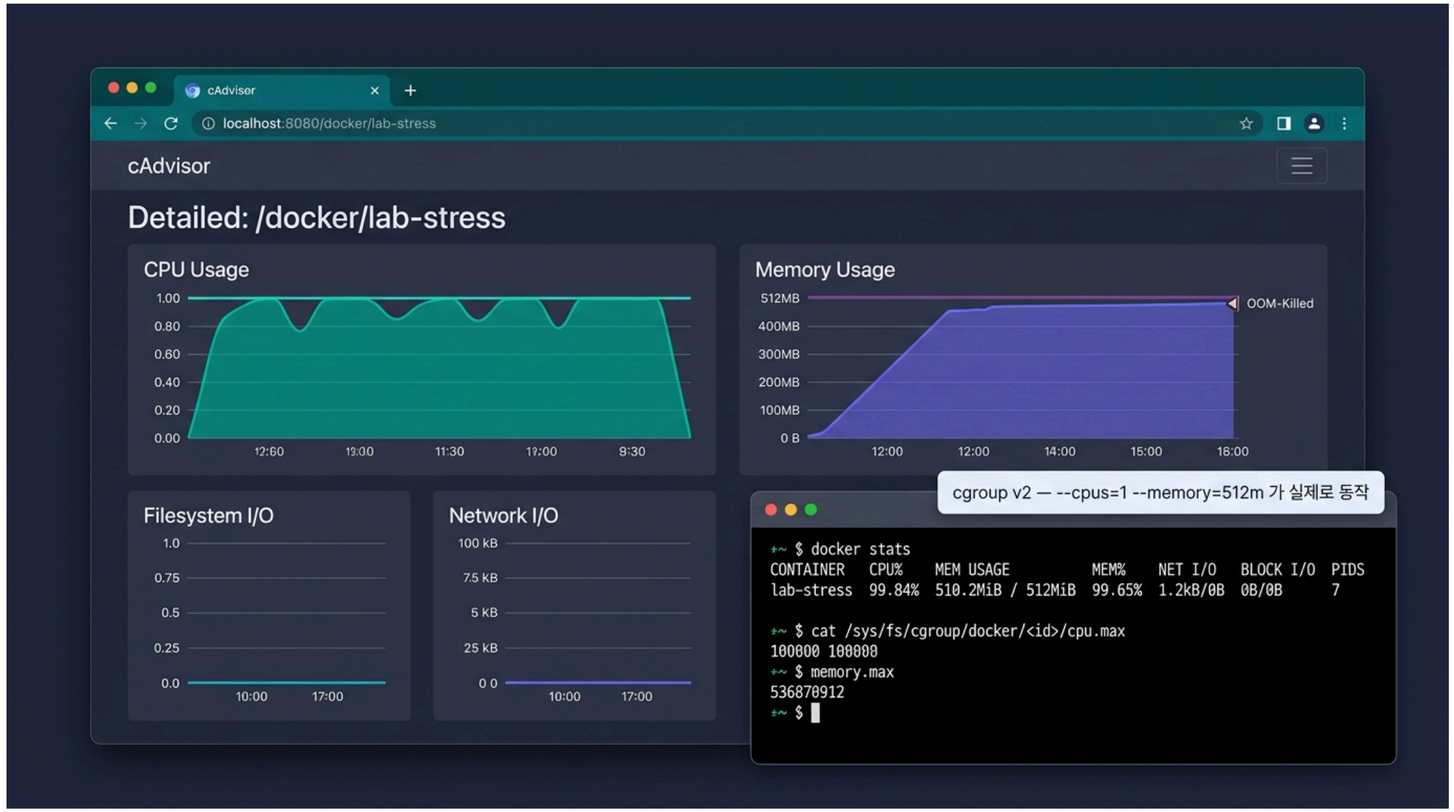
🎙️ 강사 해설 (약 5분)

- `--cpus="1"` → cgroup `cpu.max = 100000 100000` (100ms 주기에 100ms 사용)
- `--memory="512m"` → `memory.max` 초과 시 **OOM Killer** 작동
- cAdvisor는 cgroup 메트릭을 수집 → Prometheus·Grafana 연동
- K8s `requests/limits` 도 결국 같은 cgroup 위에서 동작

💻 시나리오 (약 15분)

```
# 1) Docker 설치 + cAdvisor 띄우기
sudo curl -fsSL https://get.docker.com | sudo sh
sudo docker run -d --name cadvisor -p 8080:8080 \
  -v /:/rootfs:ro -v /var/run:/var/run:ro \
  -v /sys:/sys:ro -v /var/lib/docker/:/var/lib/docker:ro \
  gcr.io/cadvisor/cadvisor:v0.49.1

# 2) 브라우저에서 http://<vm-ip>:8080 접속 → 호스트·컨테이너 메트릭

# 3) 자원 한도 컨테이너 (1 코어 · 512MB)
sudo docker run -it --rm --name lab-stress \
  --cpus="1" --memory="512m" alpine sh
# (안에서)
apk add stres6-ng
stres6-ng --cpu 4 --vm 1 --vm-bytes 1G --timeout 60s
# → CPU 4개 부하해도 cAdvisor는 1.0 core ceiling
#   메모리 1G 요청 → OOM Kill

# 4) cgroup 직접 확인 (host 쉘)
CID=$(sudo docker ps -qf name=lab-stress)
sudo cat /sys/fs/cgroup/docker/${CID}/cpu.max
# → 100000 100000  (1 코어 quota)
sudo cat /sys/fs/cgroup/docker/${CID}/memory.max
# → 536870912     (512MB)
```

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현



실습 3 — Docker + cAdvisor 로 컨테이너 격리 관측

© 2026 젠아이랩스 (GenAI Labs) · 강사 박수현

IT 인프라 아키텍처 설계

Session 3 끝
컴퓨트 스택을 한 손에

폼팩터 · CPU · 메모리 · NIC · OS · DB · GPU · 가상화 · 컨테이너

서버 한 대의 BOM부터 라이센스 산정까지 — TA의 컴퓨트 의사결정 체계

데이터센터 · 서버·OS·DB·GPU · 가상화·컨테이너 · 네트워크 L1~L7 · 스토리지 · 백업·DR · 보안 6대 도메인 · HA · 용량·성능·관찰성 · 설계 워크숍 · RFP · 5종 실전 케이스