비즈니스 연속성을 위한 무정지 솔루션
- Oracle RAC with CRS & ASM 소개 -

경북산업직업전문학교 27-JAN-2022

Agenda

1. Oracle RAC 아키텍쳐 및 특장점
2. Oracle Clusterware(CRS)와 ASM 아키텍쳐 및 특장점
3. 요약

비즈니스 연속성 계획의 중요성

비즈니스 연속성 계획은 예기치 않은 장해가 발생했을 때 고객에게 서비스를 지속할 수 있는 모든 방법에 대해 검토하는 일입니다.

재정적 위험

- 비즈니스 중단은 수익 손실을 의미
- 계획되지 않은 복구 비용
- 평판 / 브랜드 손상은 시장 가치를 감소

고객 위험

- 좋지 않은 경험으로 고객 이탈
- 널리 알려진 중단으로 인해 신규 고객 유치가 더 어려워짐

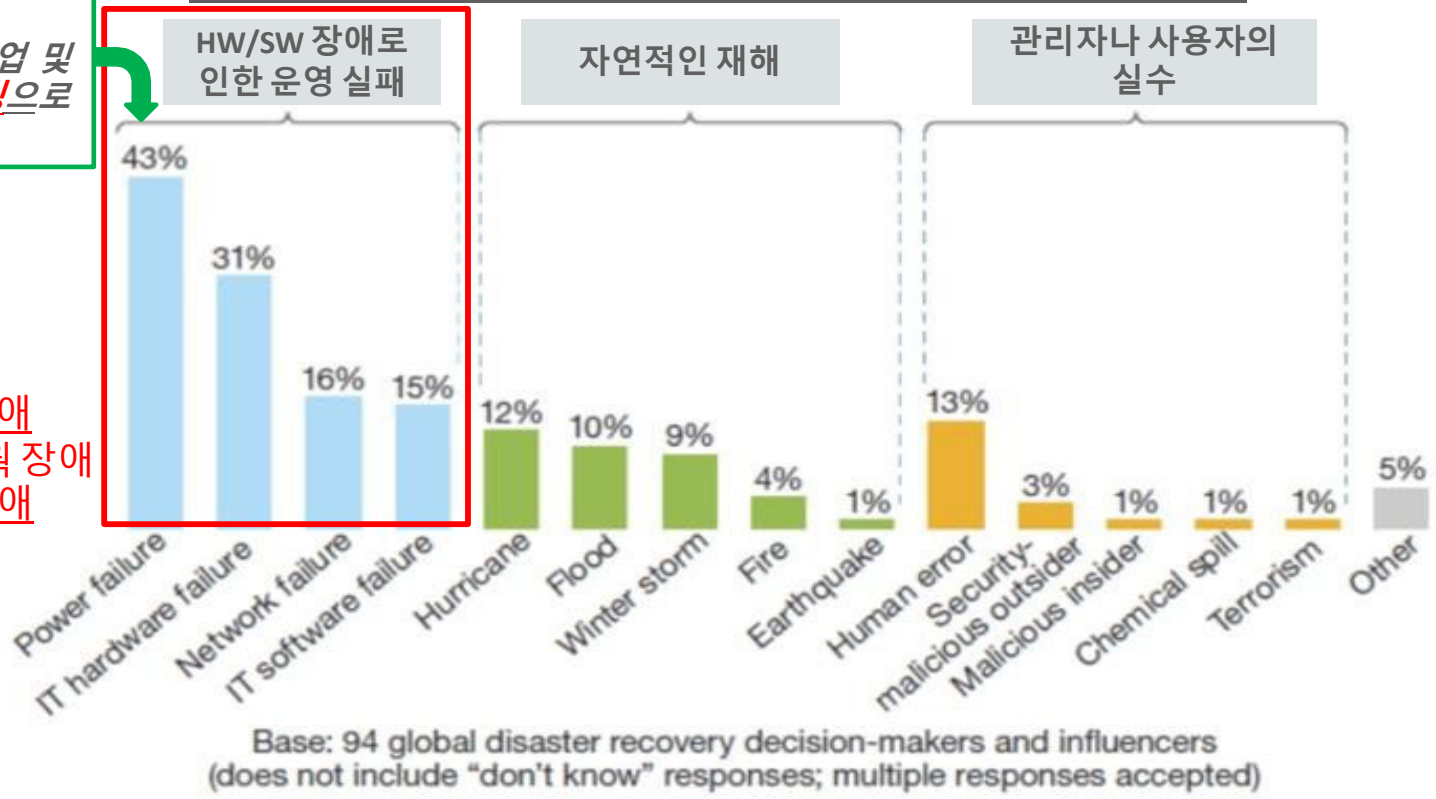
규제 위험

- 기업은 계획되지 않은 중단에 대해 처벌 받을 수 있습니다.
- 지속적인 추가 조사를 받을 수도 있습니다.

계획되지 않는 다운타임의 주요 원인 고찰

"업무 중단의 주요 원인은 무엇입니까?"

- *높은 빈도수*
- *일반적으로 백업 및 HA, 클러스터링으로 가능.*



- 정전
- HW장애
- 네트웍 장애
- SW 장애

Base: 94 global disaster recovery decision-makers and influencers
(does not include "don't know" responses; multiple responses accepted)

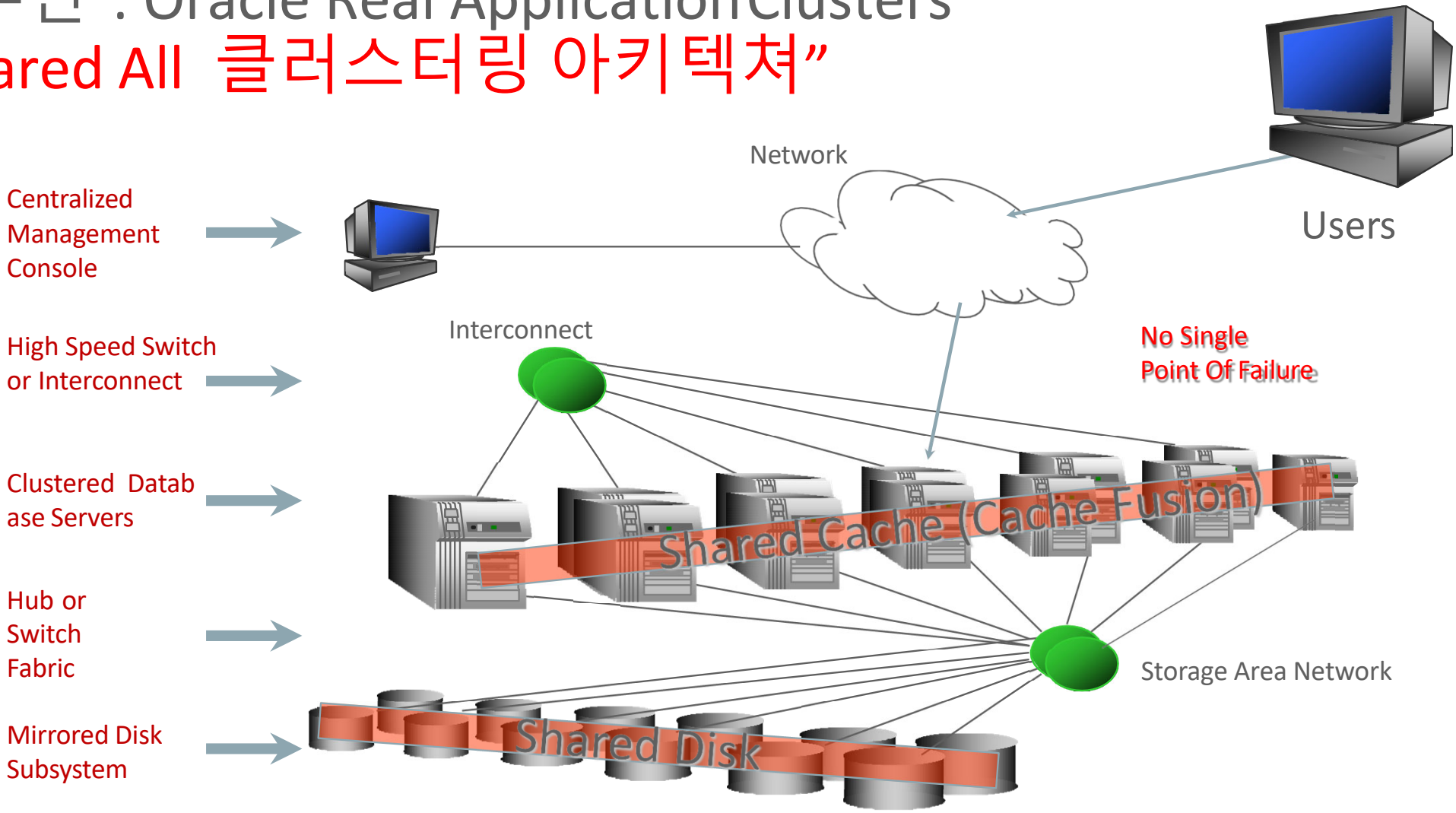
Source: Forrester Research, Inc.

Agenda

1. Oracle RAC 아키텍쳐 및 특장점
2. Oracle Clusterware(CRS)와 ASM 아키텍쳐 및 특장점
3. 요약

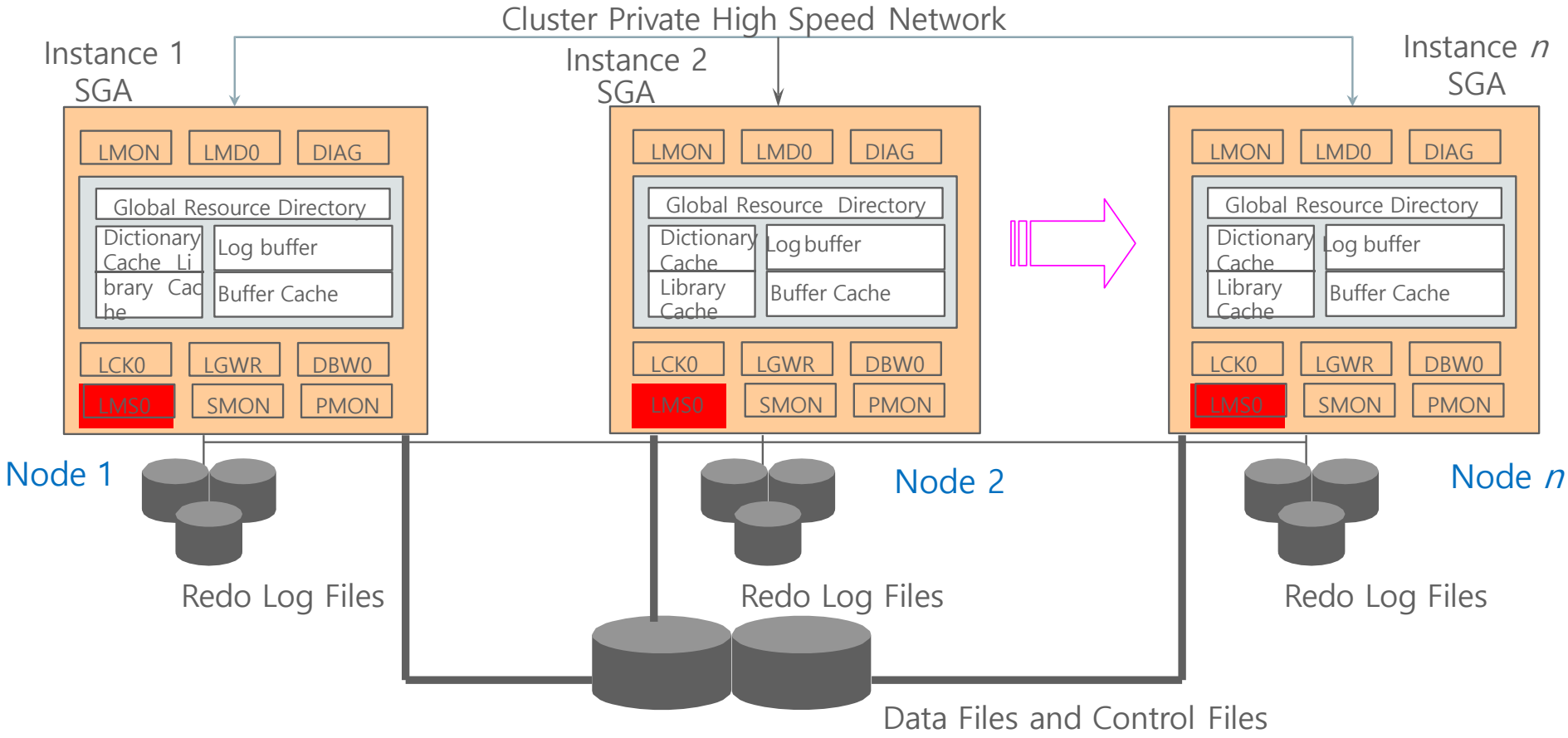
솔루션 : Oracle Real Application Clusters

"Shared All 클러스터링 아키텍쳐"



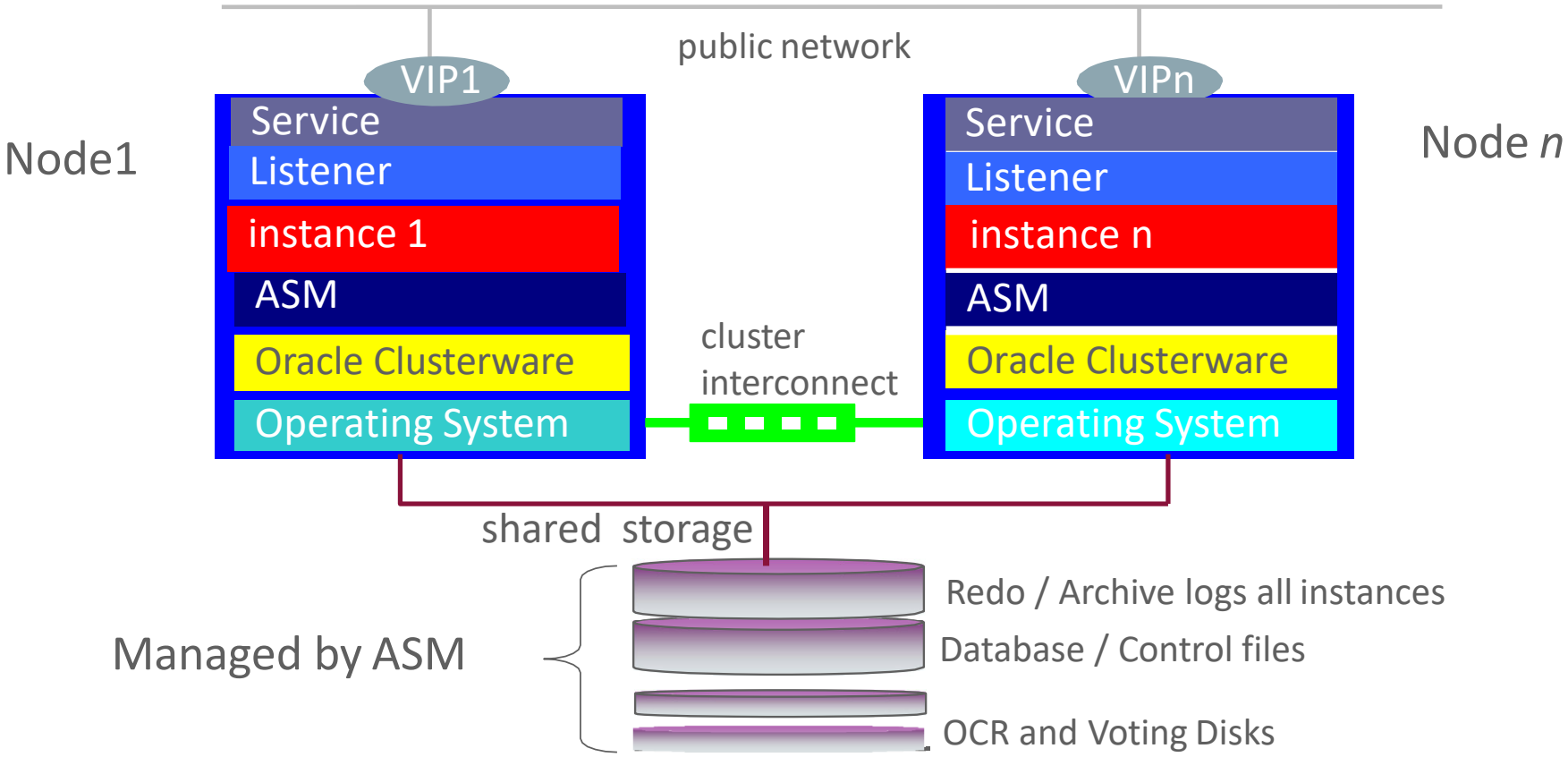
Real Application Clusters

❑ Instance 구조



Real Application Clusters

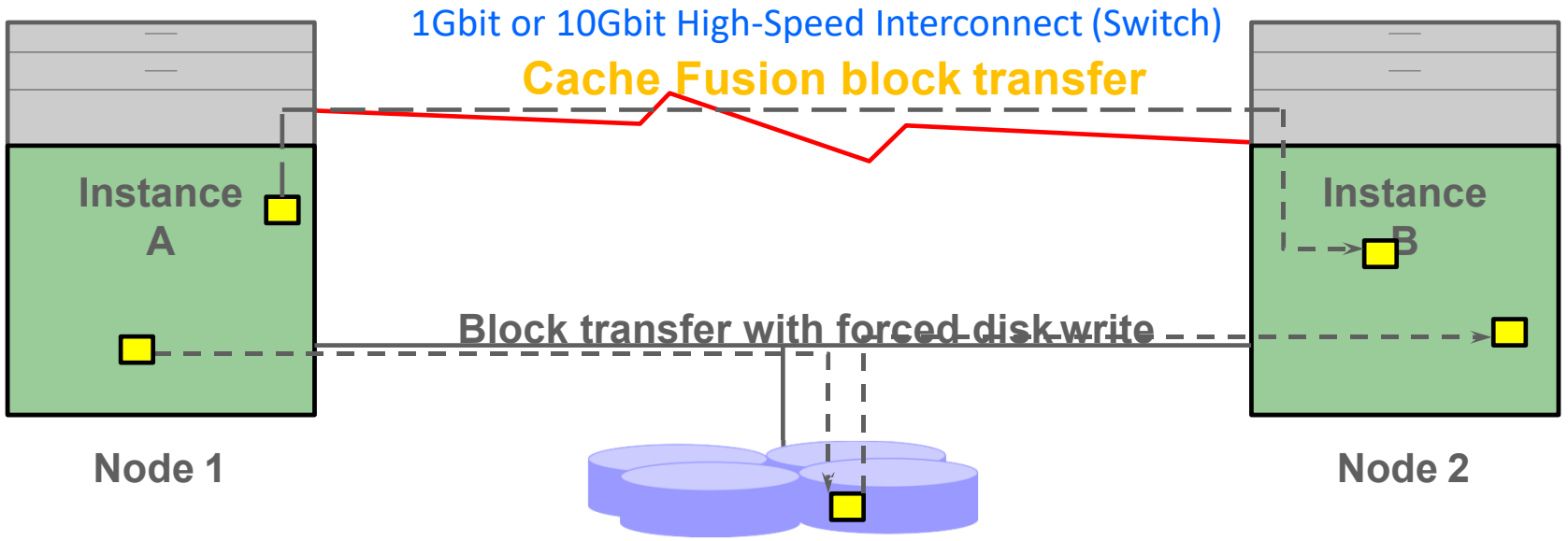
❑ Oracle Clusterware와 ASM



Real Application Clusters

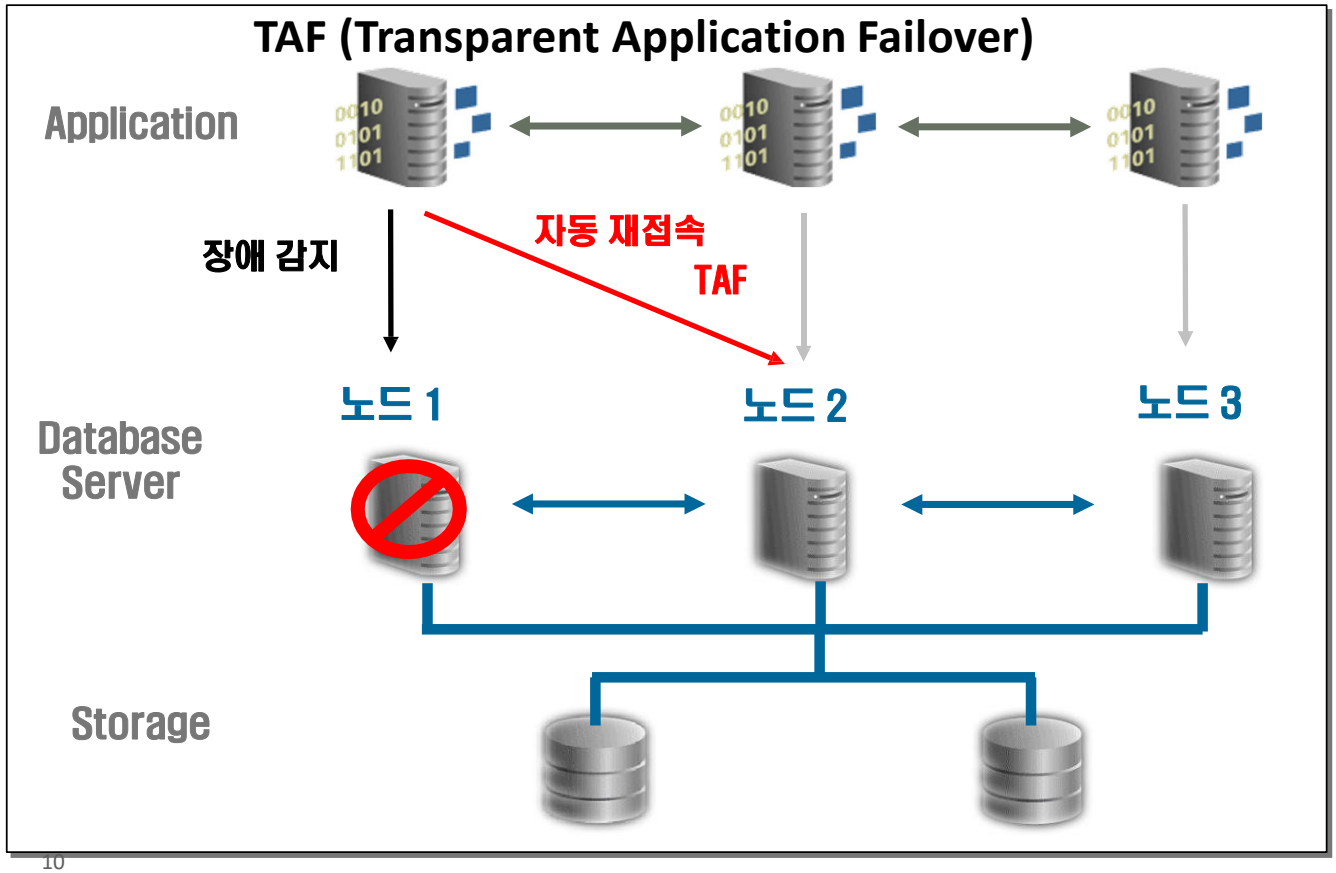
❑ 노드간 데이타 동기화

캐쉬 퓨전(Cache Fusion)에 의한 성능 향상과 확장성
인스탄스 간의 블록 요청을 고속의 Interconnect 를 통해 Disk를 거치지 않고 Cache to Cache 직접 전송으로 Disk I/O의 최소화



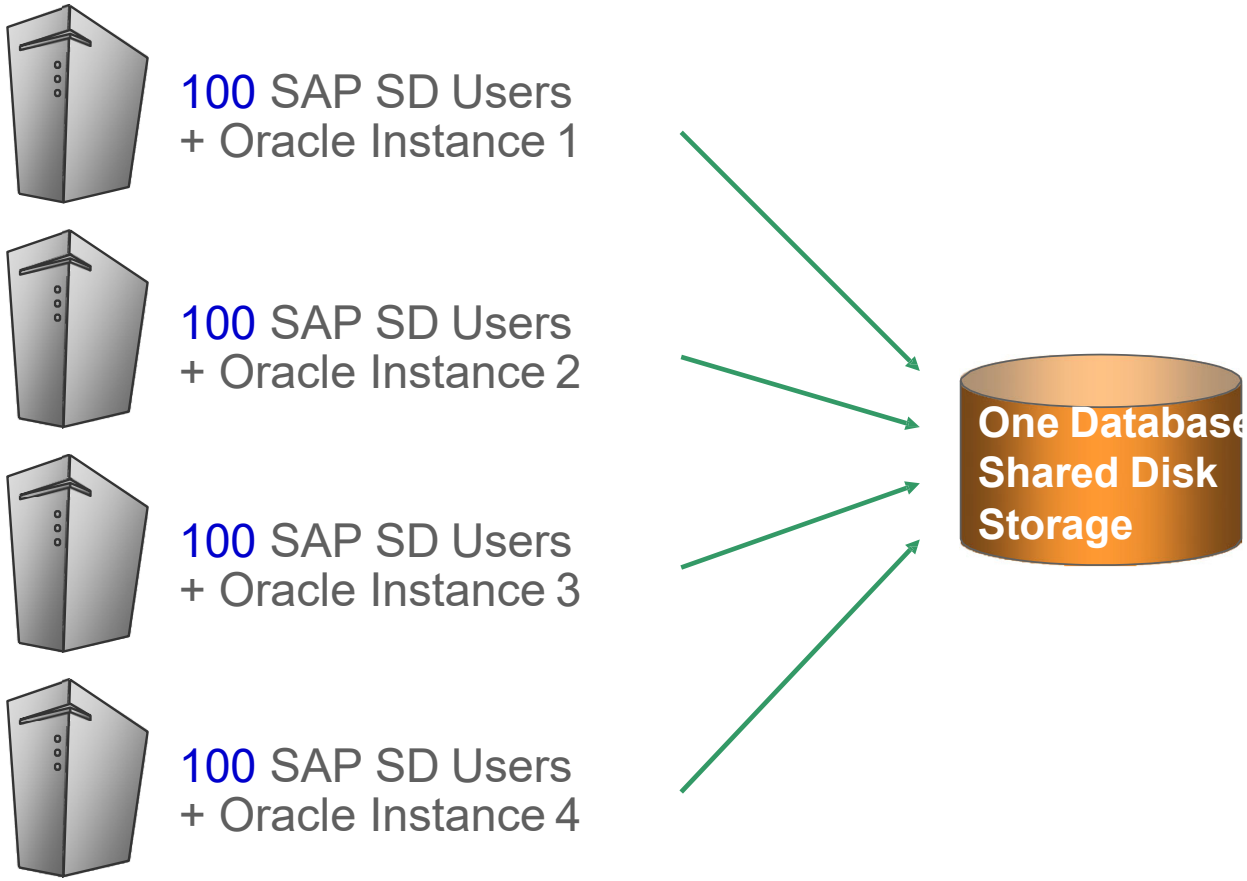
RAC 고가용성 서비스아키텍쳐

RAC 특정 노드에 장애가 발생해도 어플리케이션은 영향을 받지 않고 지속적인 연결 유지

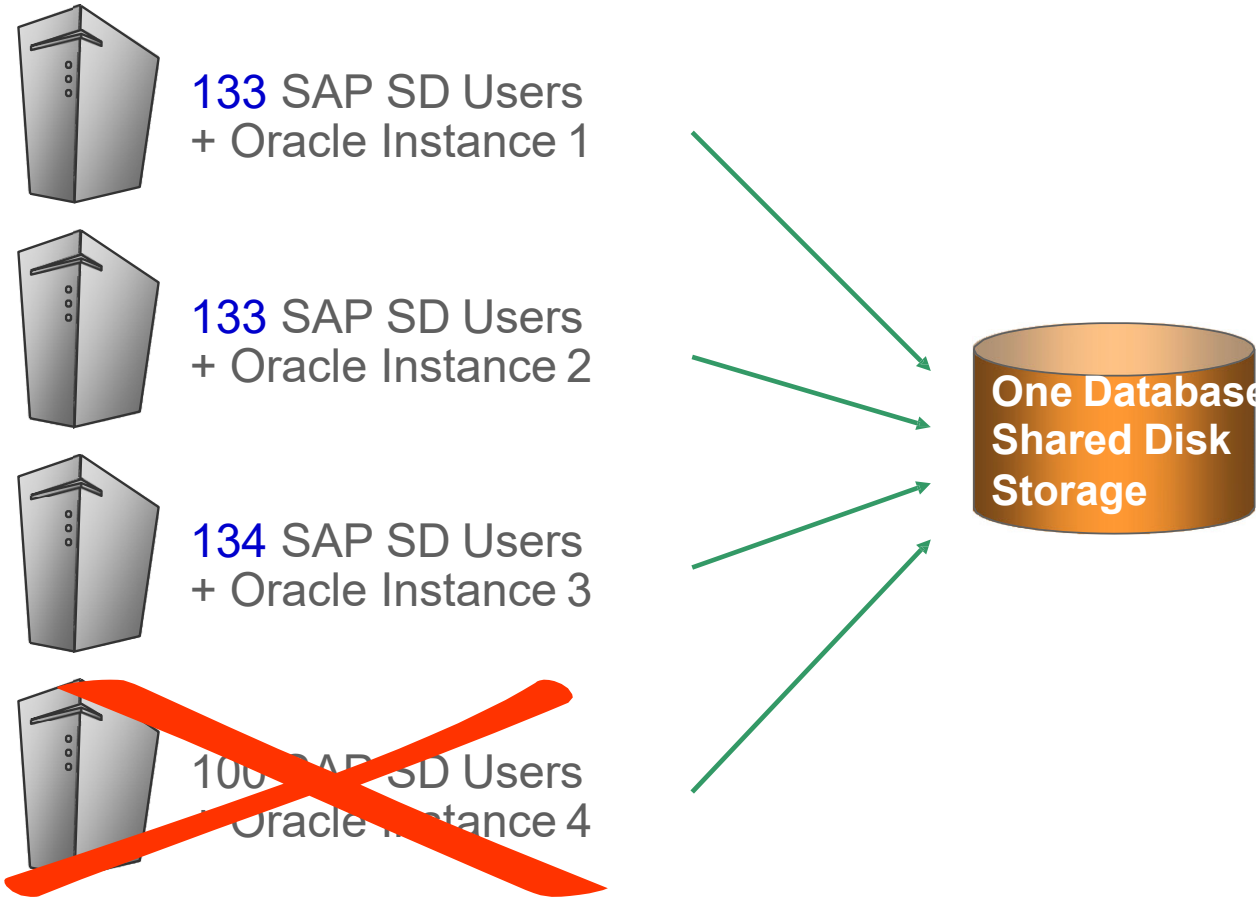


- TAF (Transparent Application Fail-Over)를 통한 DB 세션을 자동으로 복구
- 나머지 노드가 실패한 노드의 모든 작업을 분담하여 작업
- 수 초 내에 모든 작업 정상화
- 모든 작업 자동 진행

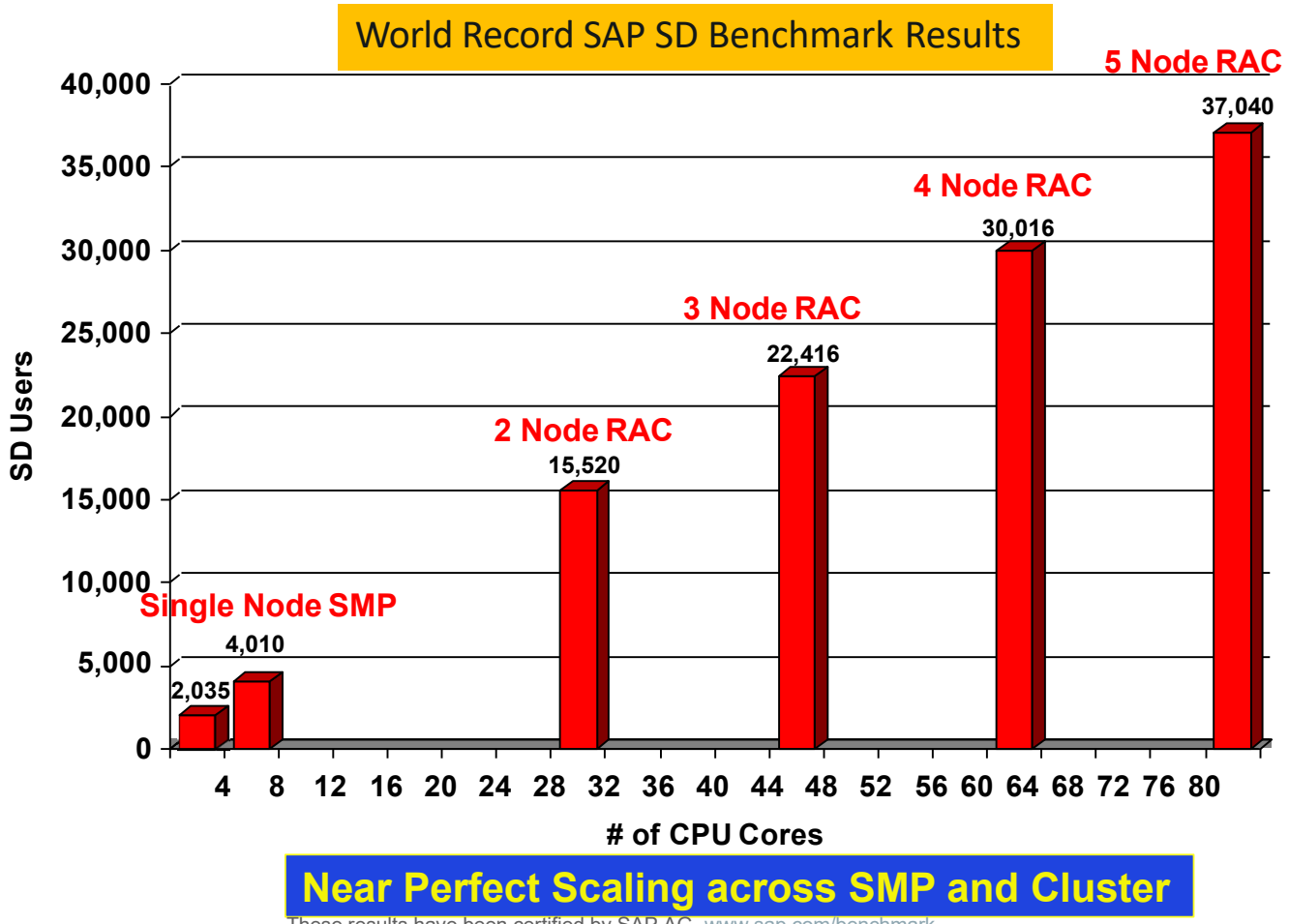
RAC Scalability



RAC Scalability + High Availability



RAC Best Scalability and Performance

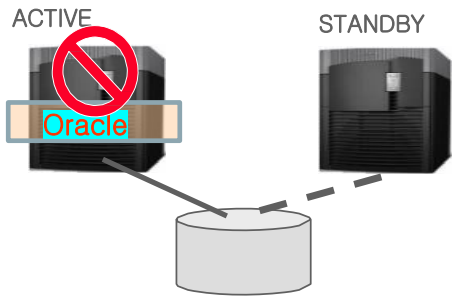


These results have been certified by SAP AG, www.sap.com/benchmark.

HW적 HA와 RAC 전환 절차 비교(1/2)

❑ 신속한 Failover 및 Instance복구측면 - Failover 전환 절차 비교

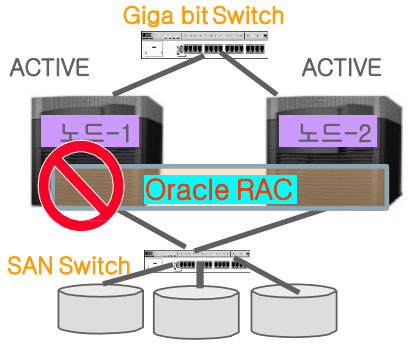
HW적 HA 적용 시



1) Active시스템에서만 서비스 처리 수행
2) Active시스템 장애시 백업용(Standby)시스템으로 Failover 수행
   - 백업시스템으로 복구 및 처리절차
     ① 운영시스템 Disk 절체
     ② Standby시스템으로 Disk Mount
     ③ Mount된 Disk의 File시스템 Check
     ④ DB Startup시 Instance Recovery 수행
     ⑤ DB 정상화 후 Application Start
3) Standby시스템이 Active 되어 서비스 처리 수행

최소 수분~ 수십분 이상 소요

Oracle RAC 기반의 자동 Failover



1) Active-Active 시스템에서 모두 서비스 처리 수행
2) 한쪽 운영시스템 장애시 백업용(Standby)시스템으로 Failover 수행
   - 장애 시 복구 및 처리 절차
     ① 다른 Active시스템에서 장애난 Instance의 Recovery 즉시 수행
     (데이터 무결성 보장을 위해 장애난 시스템의 미 반영된 Log 및 Undo 데이터 처리)
3) 다른 Active시스템에서 서비스 즉시 처리 수행

최소 수초~ 수십초 전후 소요

HW적 HA와 RAC 전환 절차 비교(1/2)

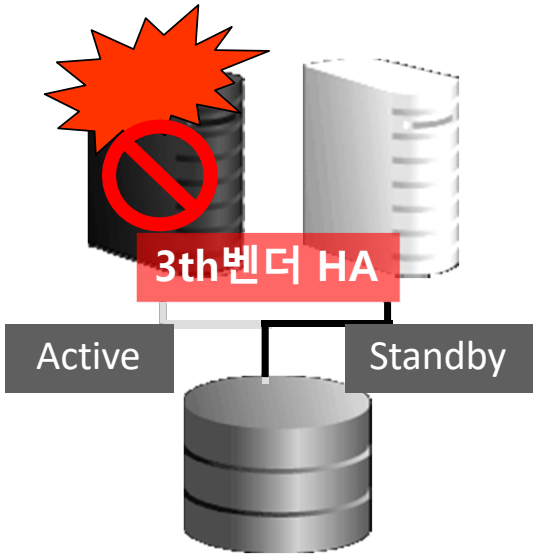
❑ 신속한 Failover 및 Instance복구측면 – Failover 단계별 Time 비교

Failover Operation	Oracle RAC	HW적인 HA
Reconfigure Group Membership	15 sec	0 sec
Reconfigure Distributed Locks	5 sec	0 sec
Failover Disk Volumes	0 sec	Up to 20 min
Restart Oracle	0 sec	Up to 5 min
Recover Oracle	20 sec	20 sec
Warm Buffer Cache	0 sec	10 + min
Total Failover Time	< 60 sec	> 35 min

Access 빈도가 높은 데이터들이 Disk 에서 Memory 내 Oracle SGA 의 DB buffer cache 로 적재, 최적화(워밍업) 까지의 예상 소요 시간, HW적 HA구성은 DB startup 후 Appication서비스가 이루어져서 디스크에서 메모리(SGA영역)로 어느 정도 access 되야 서비스 응답시간이 원활해짐.

DB 이중화 방법

Active-Standby (HA)

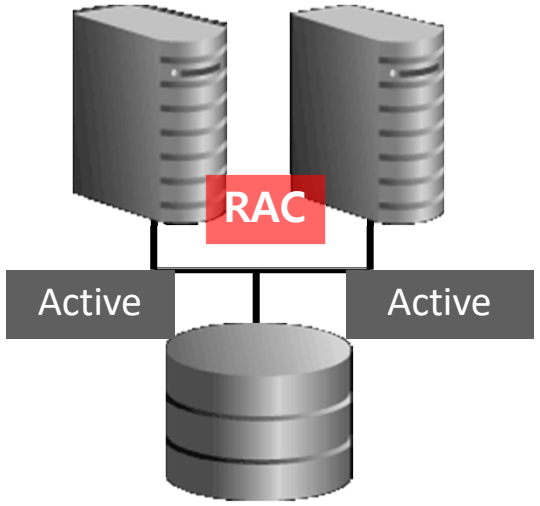


- 단지 1대의 서버 자원만 사용
- 서버 장애는 최소 수분 이상의 다운타임을 초래
- 다운타임에 의한 비용 손실은 분당 수백, 수천만 원 혹은 그 이상!
- 3rd party solution으로 fail-over(동일클러스터)

Active-Standby (RAC One)

- 단지 1대의 서버 자원만 사용
- RAC 와 비슷하고 ASM 필수 구조로써 장애시 자동 Fail-over
- Rolling patch 가능
- 초기 도입 시 RAC 가격보다 저렴하고 향후 RAC 로 전환 가능

Active-Active (RAC)



- 모든 서버의 모든 자원 사용
- 중단 없는 고가용성
- 서버 장애가 일반 사용자에게 미치는 영향은 없음
- RAC 가격

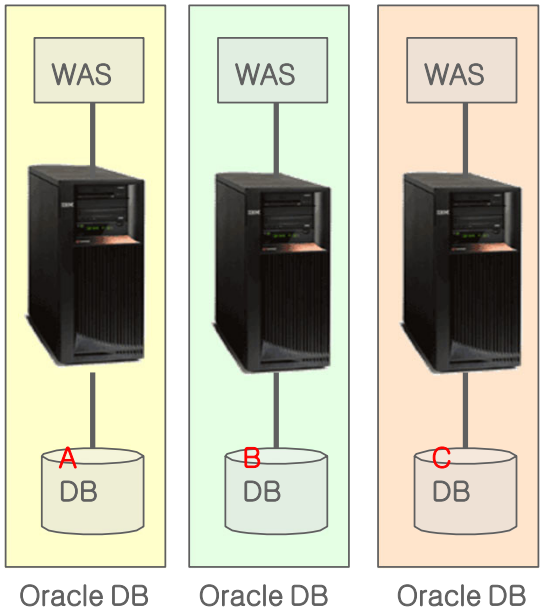
DB 이중화 방법에 따른 차이점 요약

	HA 솔루션	오라클 RAC One	오라클 RAC
DBMS 버전	• 모든 버전	• 11.2 이상	• 9i 이상
License	• EE(10days Rule 가능)	• EE + RAC One($10,000/PL)	• EE + RAC($23,000/PL)
Grid Infra구조	• Stand-alone	• CRS + ASM	• CRS + ASM
운영형태	• Active-Standby 서비스 구조	• Active-Standby서비스 구조 (Omotion 수행 동안 30분 이내 Active-Active상태임)	• Active-Active 서비스 구조 (클러스터 내 모든 노드에서 서비스 및 로드밸런싱)
장애복구 방법	• 클러스터웨어에 의한 반자동적 복구 • 스크립트 기반이며 오라클 재기동 필요	• Oracle RAC One에 의한 실시간 자동 복구 (장애 발생 시점의 수행 쿼리를 재접속하여 재수행)	• Oracle RAC에 의한 실시간 자동 복구 (장애 발생 시점의 수행 쿼리를 다른노드에서 지속수행)
장애복구 시간	• 장애 환경에 따라 수십분 이상 소요됨	• Standby DB 기동 시간으로 수분 소요됨	• 소요시간 거의 없음(Zero Down time)

RAC 활용 서버/데이타베이스 Consolidation 구성

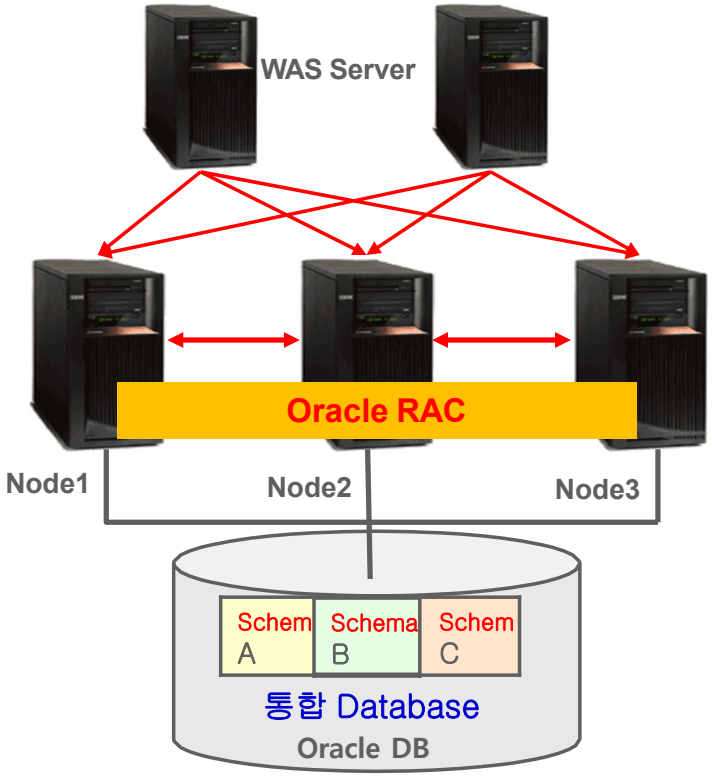
❑ SLA 기준을 적용해서 서비스별 자동 부하 관리 환경 구축

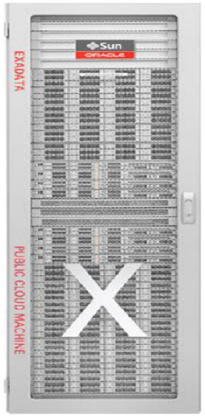
■ 독립된 운영 환경 (AS-IS)



DB통합

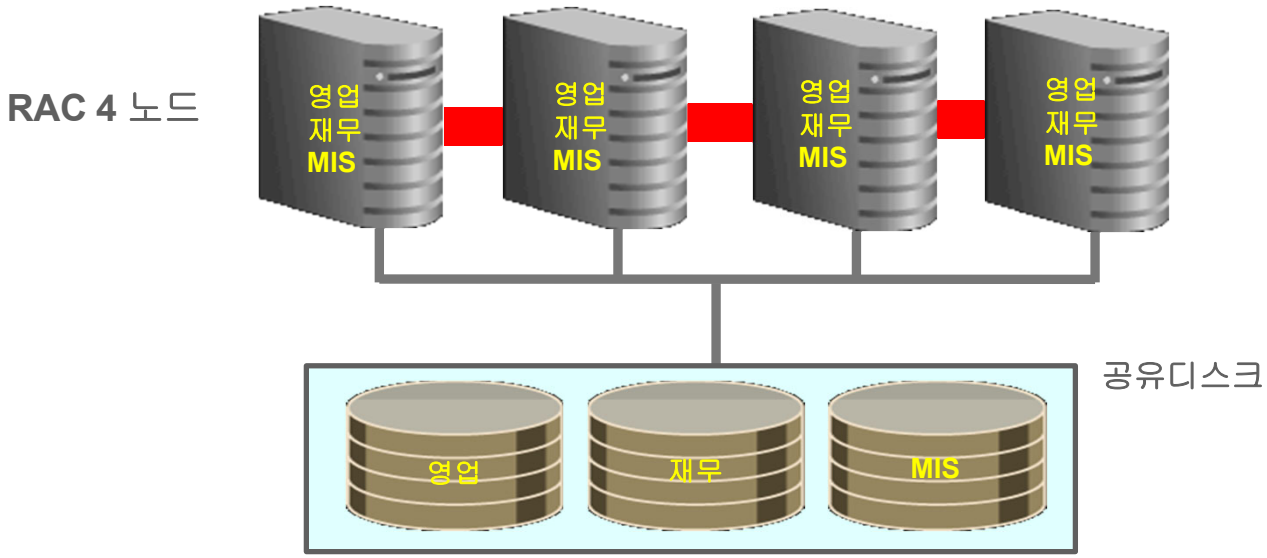
■ RAC 기반의 통합 구조 (TO-BE)





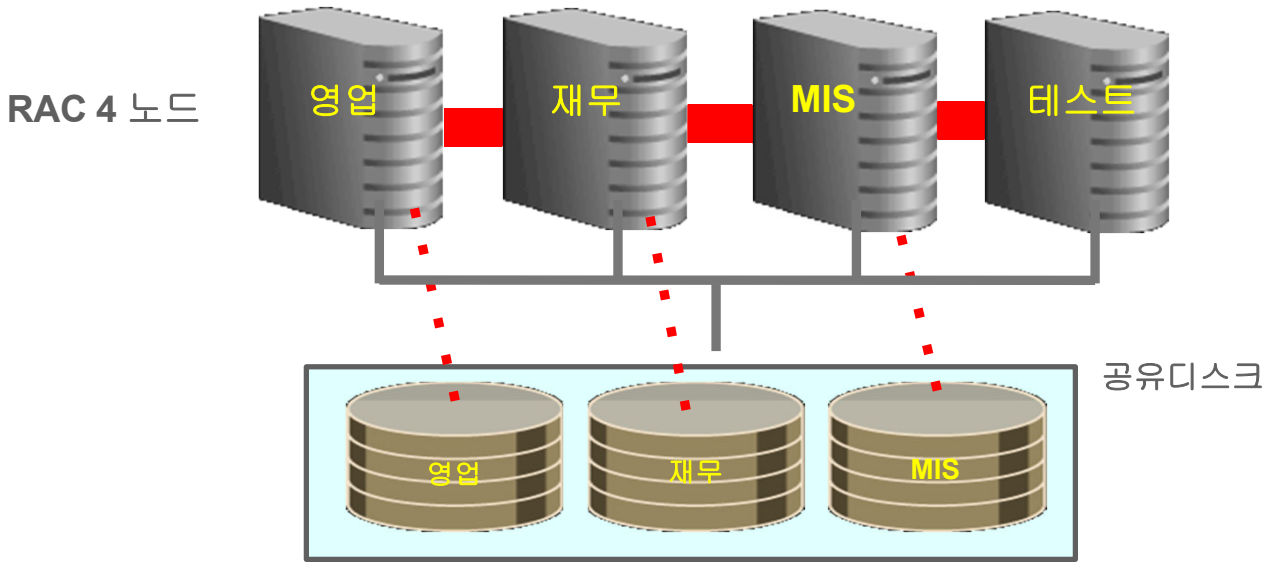
RAC구현모델: 모든 노드 동일 서비스 (1/4)

- 모든 RAC 노드에서 동일한 서비스를 제공
- 일반적인 OLTP(질의비율 70~80%)에서 약 85%~ 90% 전후의 확장성 보장
- 과도한 DML 작업인 경우에는 노드간의 간섭이 발생될 수 있기 때문에 DB논리/물리 모델링 및 AP튜닝 최적화 반드시 필요



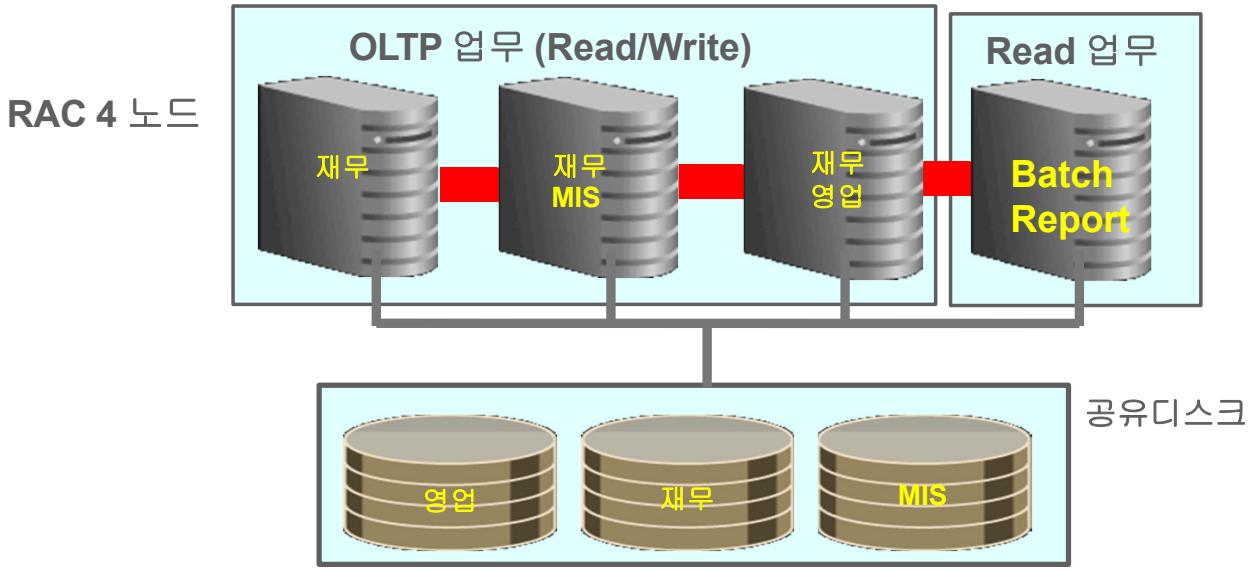
RAC구현모델: Application별 노드분리(2/4)

- RAC의 각 노드는 전용 Application만 전담
- RAC의 각 노드간의 간섭이 최소화(95% 확장성)
- 리소스를 최적으로 사용하는 Application 분배가 관건



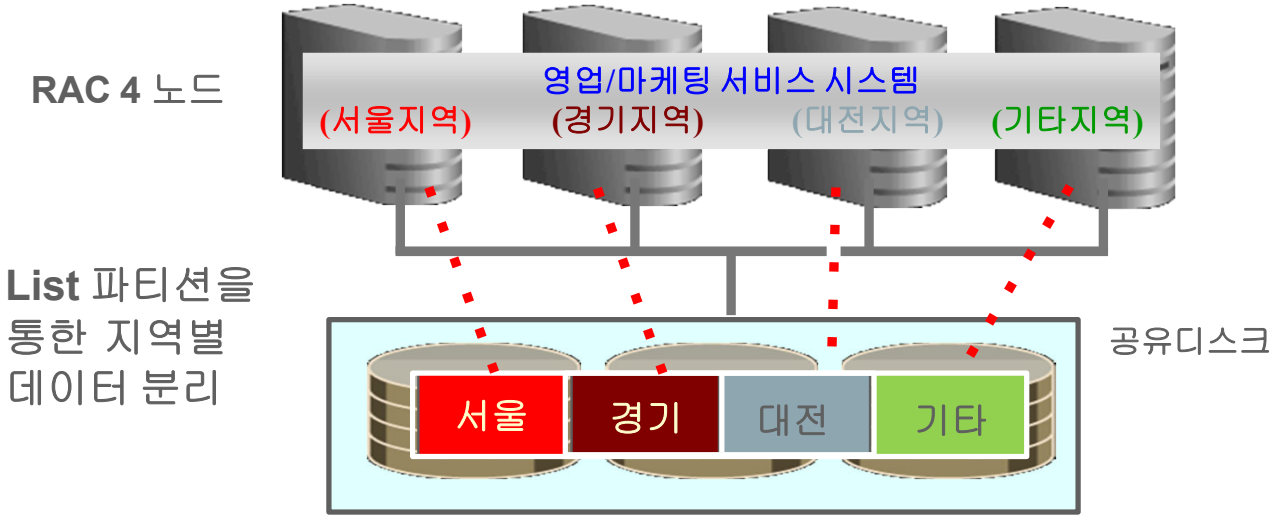
RAC구현모델: 업무특성별 노드분리(3/4)

- RAC 노드를 데이터 접근 방식(Read/Write)에 따라 분리
  - OLTP 업무(R/W)와 배치 업무(R)로 분리
  - OLTP의 업무가 배치에 비해 중요한 업무에 적합
  - OLTP가 과도한 배치에 의해 영향 받는 것을 최소화



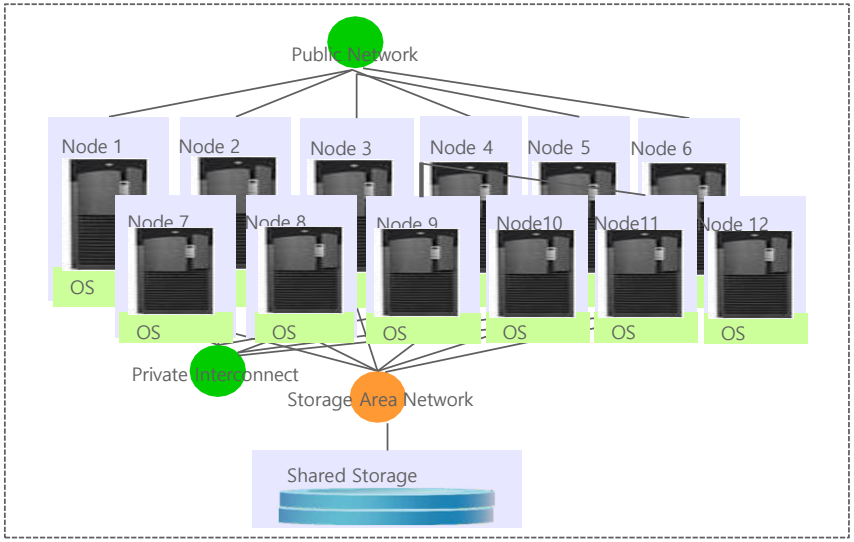
RAC구현모델: 데이터 접근방식에 따른 모델 (4/4)

- Data Partitioning에 따른 노드별 디스크 분리
  - 같은 업무(서비스)지만 Access하는 데이터를 각각 다른 영역으로 지정할 수 있도록 Partitioned Table을 활용
  - 동일 테이블의 서로 다른 Partition영역을 Access함으로서 노드간의 경합을 최소화 시킴



사례: S통신사 고객관리 및 빌링시스템 (12노드 RAC)

❑시스템 구성도



❑시스템사양

- H/W
  - Dell Poweredge R930 / RedHat 7.2
  - CPU: Node 당 64 Cores
  - MEM: Node 당 384G
  - Disk: Hitachi VSP G1000 (운영 Usable 88TB)
- S/W
  - Oracle 12c (12.1.0.2) with RAC

❑운영 형태 및 특징

- 총 DB 12 노드 RAC [8노드 OLTP업무용, 3노드 Batch업무용, 1노드 OGG CDC전용]
- Mainframe DB2 에서 Oracle 전환 구축 사례
- 4노드(2006년) ➔ 5노드(2009년) ➔ 6노드(2011년) ➔ 12노드(2018년도) 전환
- 24X7 무정지 구현 사례
- 통신업무 기준 최대 약 11,000 TPS 이상 확보

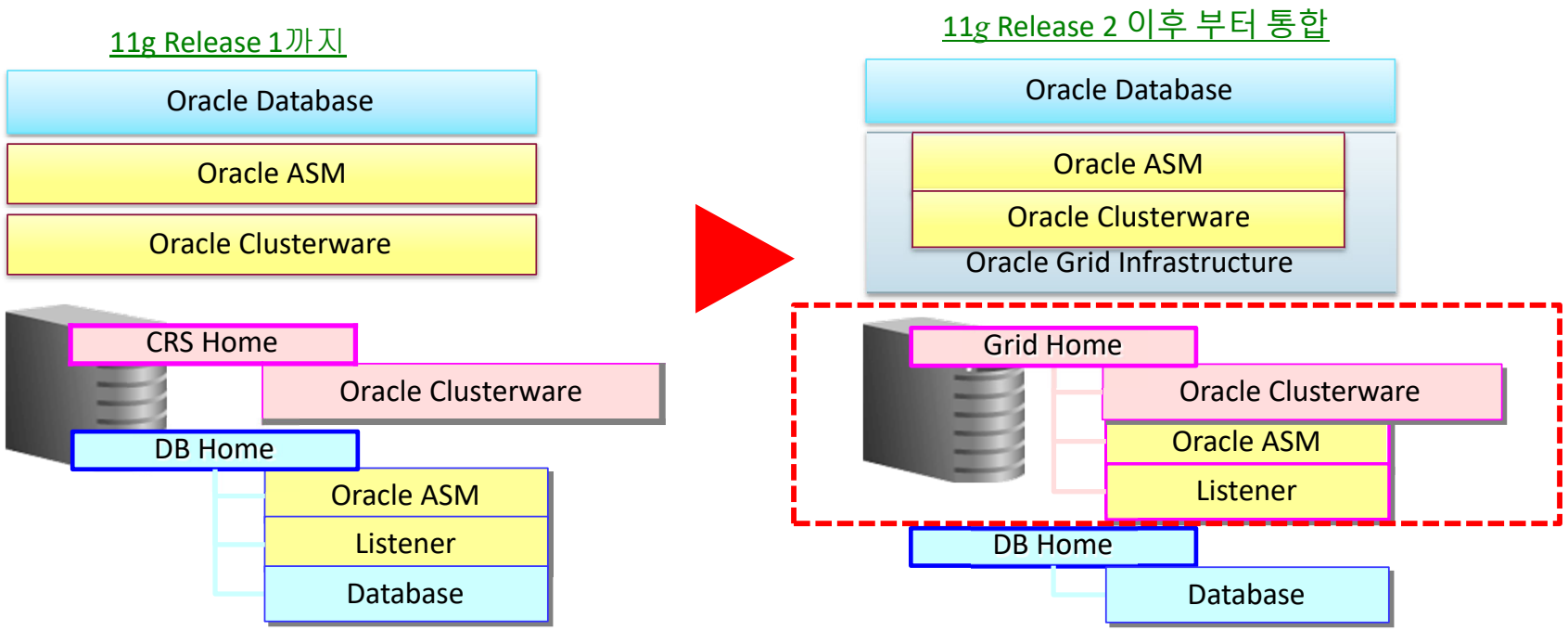
Agenda

1. Oracle RAC 아키텍쳐 및 특장점
2. **Oracle Clusterware(CRS)와 ASM 아키텍쳐 및 특장점**
3. 요약

Oracle Grid Infrastructure

❑ 구성요소

- 11g Release 2 부터 Oracle Clusterware 와 Oracle Automatic Storage Management (ASM)이 통합되어 Oracle Grid Infrastructure로 제공

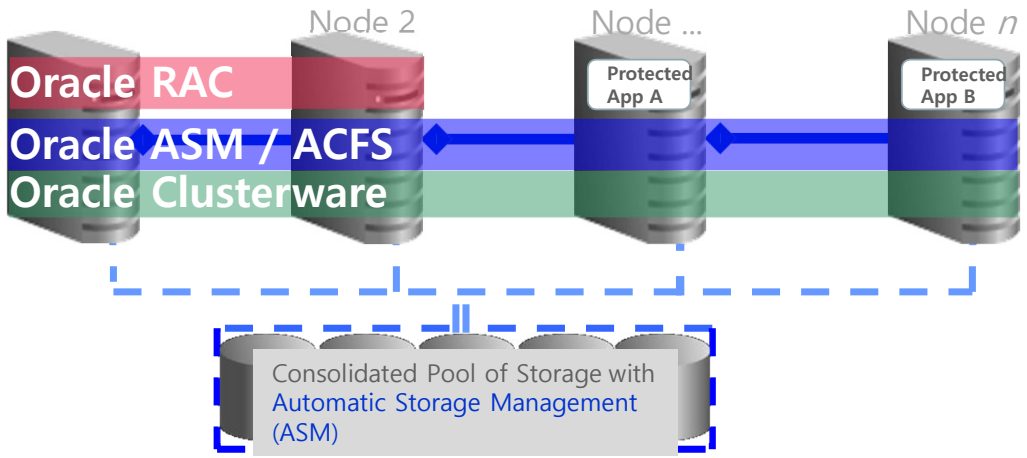


Oracle Clusterware

❑ 개요

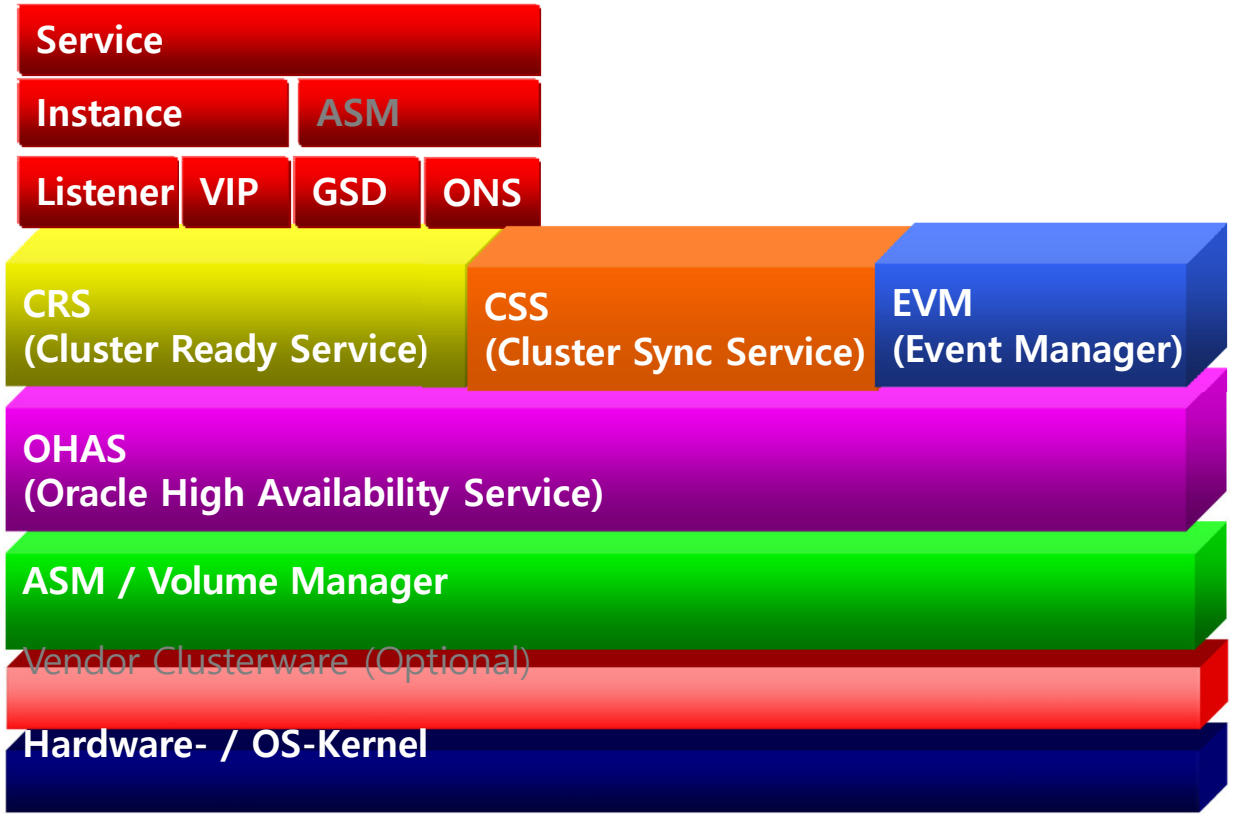
Oracle Clusterware는

- Oracle Grid Infrastructure (OGI)의 핵심 부분.
- Oracle Automatic Storage Management (ASM)과 밀접하게 연동.
- Oracle ASM Cluster File System (ACFS)의 기반.
- Oracle Real Application Clusters (RAC)의 핵심 기반.
- 모든 종류의 애플리케이션에 cluster infrastructure 역할을 수행.



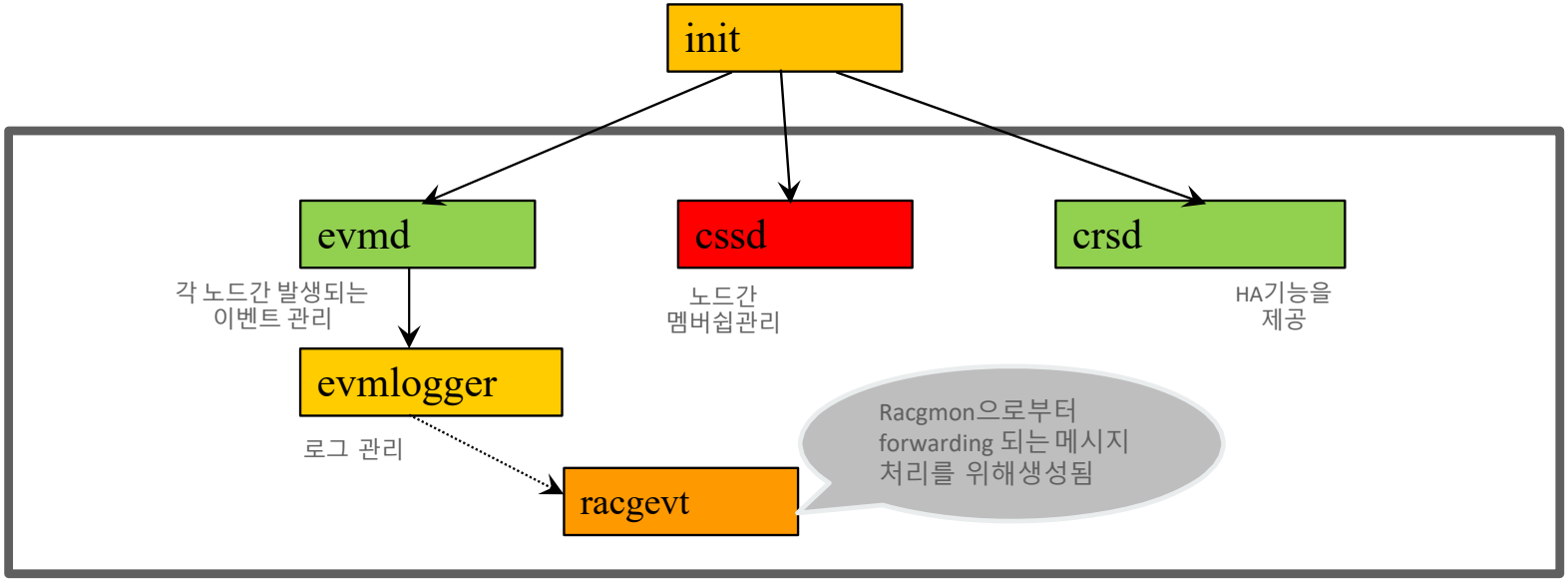
Oracle Clusterware

❑ 논리적 구성



Oracle Clusterware

❑ CRS 아키텍쳐




```
# ps -ef|grep d.bin

oracsb 21639 21349 0 Oct 29  ?       2:34 /app/oracle/product/10.2.0/crs/bin/evmd.bin
  root 21684 21415 0 Oct 29  ?      44:24 /app/oracle/product/10.2.0/crs/bin/crsd.bin restart
oracsb 14713 14218 0 Oct 29  ?      48:55 /app/oracle/product/10.2.0/crs/bin/ocssd.bin
```

OCR (Oracle Cluster Repository) / Voting

- Shared Disk에 위치, Multiplexing 가능
- OCR(Oracle Cluster Registry)
  - ✓ Cluster 리소스들에 대한 정보 저장소
  - ✓ Node 매핑정보 저장(Node List, Cluster Database Instance)
  - ✓ 최소크기 100M
- Voting Disk
  - ✓ Cluster membership관리를 위해 css 데몬이 이용
  - ✓ Cluster Member들의 health check
  - ✓ Split brain 상태에서 node의 상태를 판단하기위한 second heartbeat 역할
  - ✓ 네트워크 failure 발생 시, 어떤 인스턴스를 cluster 내에 남길 것인지 결정
  - ✓ 최소크기 10M

CRS Components Summary

Components	Owner	On Failure
OPROCD	Root	Reboot
CSSD	Oracle	Reboot
EVMD	Oracle	Auto Restart
CRSD	Root	Auto Restart
OCR	Root	Reboot
Voting Disk	Oracle	Reboot

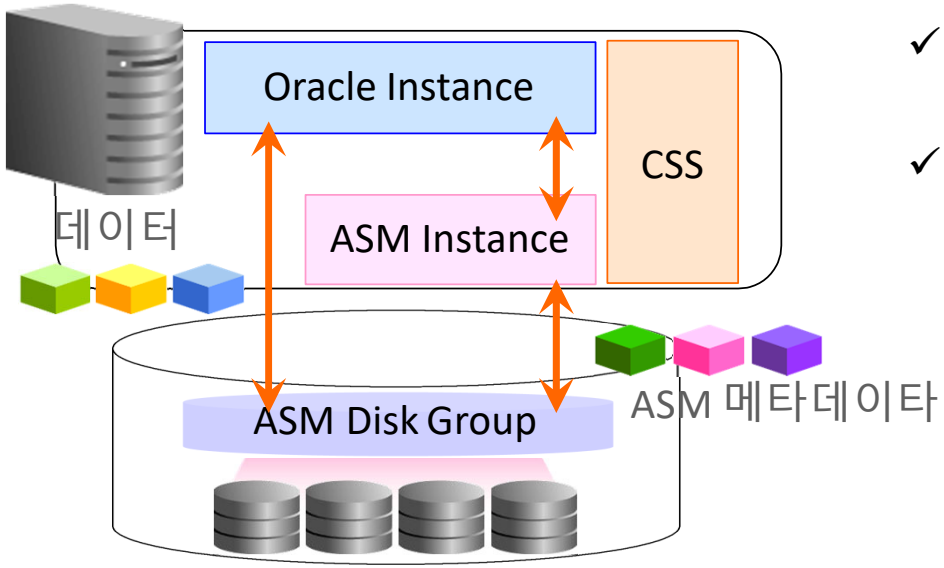
- Process Check

```
# ps –ef|grep  d.bin
oracle 942220 274888 0 14:28:25 pts/6 0:00 grep  d.bin
    oracle 287086 622638 0  11:25:38      -  0:18 /oracle/app/oracle/product/crs/bin/ocssd.bin
      root 456834 308842 0 11:25:38      -  2:20 /oracle/app/oracle/product/crs/bin/crsd.bin reboot
    oracle 384482 774834 0  11:25:38      -  0:00 /oracle/app/oracle/product/crs/bin/evmd.bin
```

스토리지 가상화-

Oracle Automatic Storage Management(ASM)

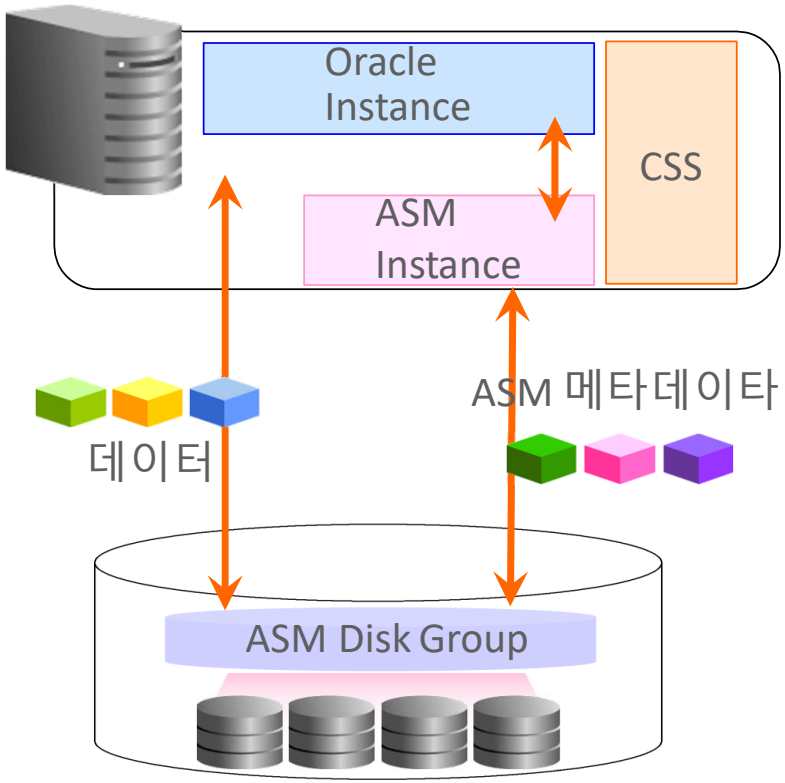
- Oracle 데이타베이스에 대해서 볼륨매니저 겸 파일 시스템기능을 제공하며 디스크 구성을 가상화해 DBA가 손쉽게 스토리지를 관리
- 에디션(EE/SE)에 관계없이 싱글 및 클러스터 환경 모두 사용가능
- 11g Release2부터 ASM 클러스터 파일 시스템(ACFS)이 구현



- ✓ Oracle Database에 스토리지 풀을 제공 + 디스크 관리 시간,비용을 대폭 절감
- ✓ 여러 디스크 어레이에 디스크를 가상화해 디스크 추가 / 제거해도 데이터를 투명하게 재분배

Oracle Automatic Storage Management

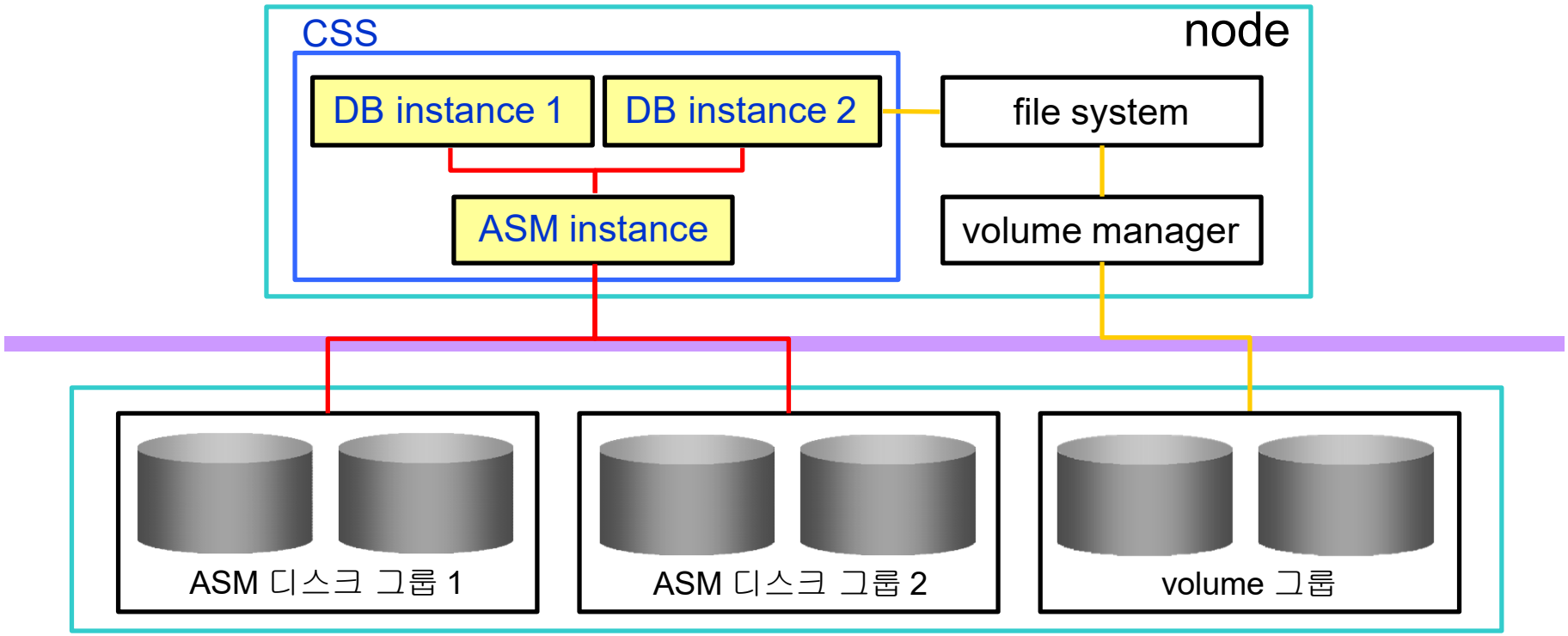
ASM 아키텍처



- ASM Instance
  - ✓ ASM 디스크 그룹을 관리하는 메모리와 프로세스 군
  - ✓ ASM 및 그 Metadata 관리
  - ✓ DB instance에게 ASM Metadata 서비스
  - ✓ 두 가지 특별한 백그라운드 프로세스:
    - RBAL : ASM 디스크 Rebalancing coordinator
    - ARBn : RBAL에 의해 관리되는 프로세스 (실제 worker)
- Cluster Synchronization Services
  - ✓ Oracle Clusterware의 멤버쉽 관리 서비스를 사용
  - ✓ Oracle Instance와 ASM Instance의 존재를 통지
- ASM 디스크 그룹
  - ✓ Oracle Instance에서 사용가능한 가상화 스토리지 풀
- ASM 디스크
  - ✓ ASM 디스크 그룹을 구성하는 개별 디스크 (Logical Unit)
  - ✓ 일반적으로 Disk Array의 LU를 그대로 사용

Oracle Automatic Storage Management

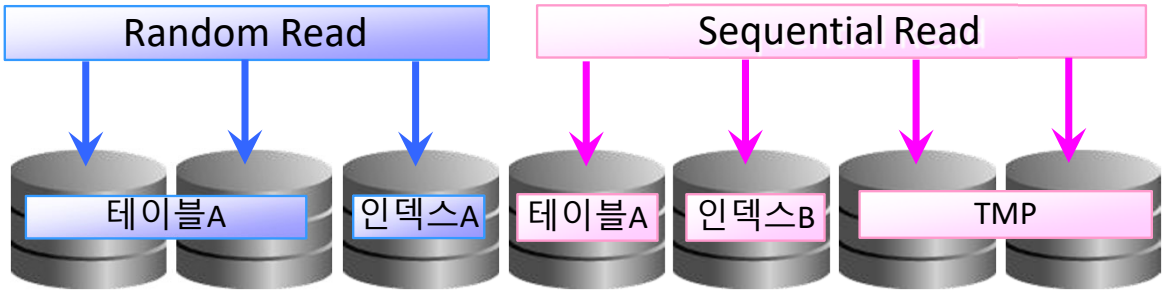
ASM을 이용한 시스템 구성 예



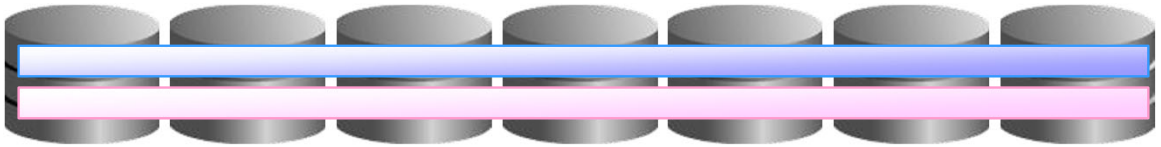
Oracle ASM 의 설계 기본 사상

Stripe And Mirror Everything (S.A.M.E)

- 기존방식 : 용도별로 디스크를 구별하는 설계



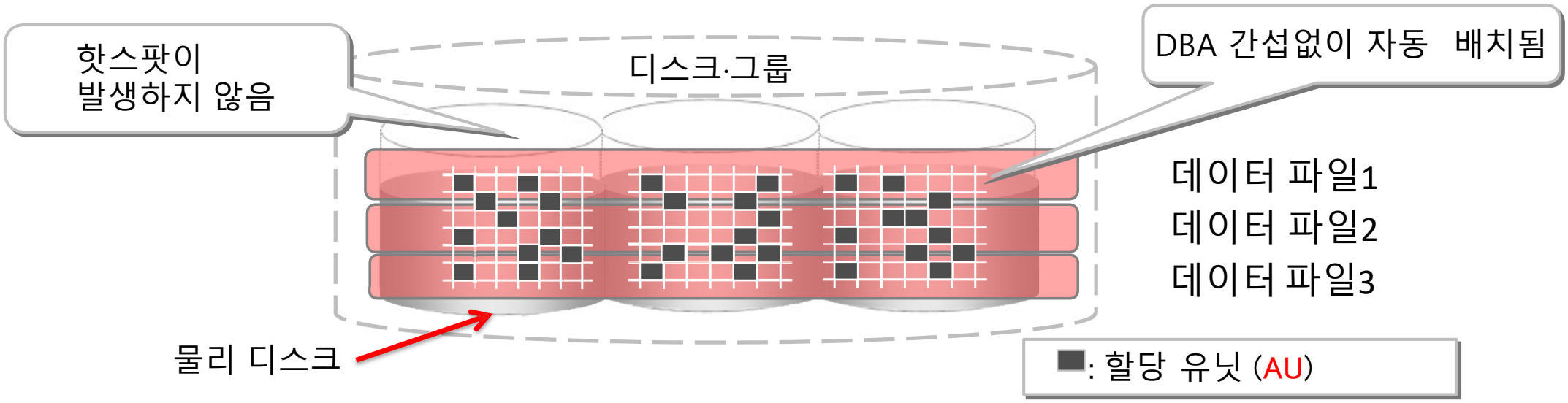
- ASM : 데이터를 그룹내의 모든 디스크에 스트라이프, 미러링 (없음, 이중, 삼중화)
  - ✓「 모든 디스크가 균등하게 사용될 수 있도록 모든 디스크에 데이타를 스트라이프해 분산 배치하고, 미러링도 구현」스토리지 볼륨 설계 기술
  - ✓ I/O 성능의 확보 : 모든 Disk의 I/O 대역폭을 최대한 활용
  - ✓ 가용성을 보장 : 미러링 채용
  - ✓ 설계의 간소화 : 물리적 Disk 구성을 은폐해 특별한 설계가 불필요



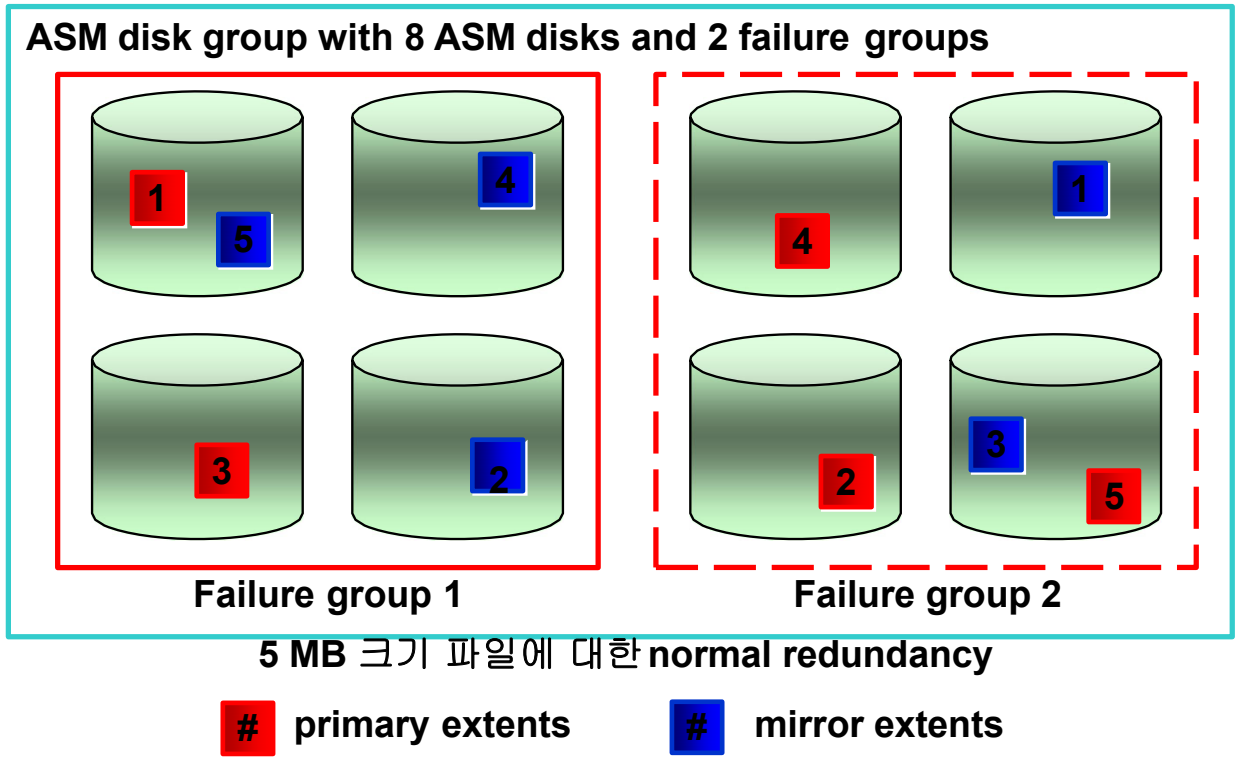
ASM 핵심기술

스트라이핑

- 디스크 그룹의 모든 디스크로 스트라이핑
- 할당 유닛 (AU) 단위로 영역을 할당
  - ✓ 1, 2, 4, 8, 16 32, 64MB의 가변 크기로부터 선택, 기본값은 1MB
- 모든 디스크의 사용율이 동일하도록 할당
- 스토리지 서버의 크로스 스트라이핑도 가능
- Striping의 단위 : Coarse Striping: 1 MB ASM Extent 단위, Fine-grain Striping: 128 KB 단위(redo log 등)



ASM 핵심기술
미러링과 장애그룹



- 대상: ASM extent (primary extent & mirror extent들)
- 범위: 하나의 ASM 디스크 그룹 내
- failure 그룹
  - ✓ 공통의 리소스 (예: 컨트롤러)에 그 가용성을 의존하는 ASM 디스크들을 하나의 failure 그룹으로 정의
  - ✓ primary extent와 해당 mirror extent들은 서로 다른 failure 그룹에 저장
- ASM mirroring - redundancy의 종류
  - ✓ external: no mirroring
  - ✓ normal: 2-way mirroring (2 이상의 failure 그룹)
  - ✓ high: 3-way mirroring (2 이상의 failure 그룹)

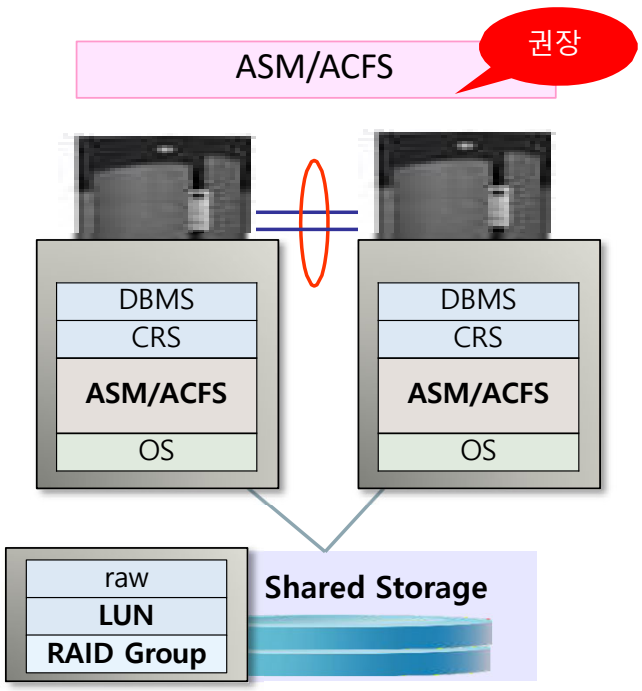
ASM 핵심기술
리밸런싱

- 동적 리밸런싱
  - ✓ 디스크의 추가, 삭제시 자동 및 동적으로 파일을 재배치

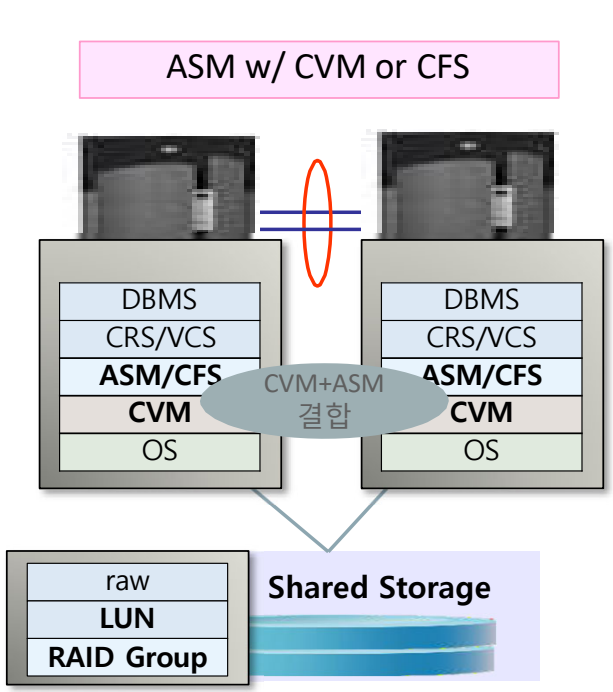
- Disk Repair Time
  - ✓ 디스크 장애 시 ASM이 손상된 디스크를 drop하고 rebalance를 수행하기 전까지 오프라인으로 남아있는 총 시간을 지정한 값 ➔ 일시적 장애일 경우를 대비하여 일정시간 동안 rebalance작업을 유보
  - ✓ Default 값 : 3.6시간

- ASM Fast Mirror Resync (Resync)
  - ✓ disk_repair_time 이내에 장애가 발생한 디스크가 정상화될 때 변경된 Extent 들만 Resync 하는 기능

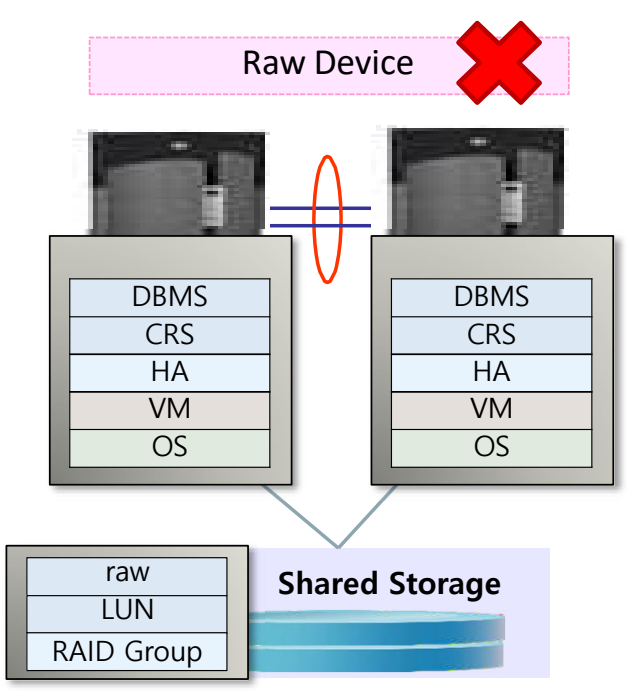
RAC 환경의 스토리지구성



- ASM 단독으로만 구성 가능
- ACFS 가능



- ASM 과 CVM 과 결합 구성 가능
- 3rd party 솔루션 License 필요



- Raw device only 불가(12c~)