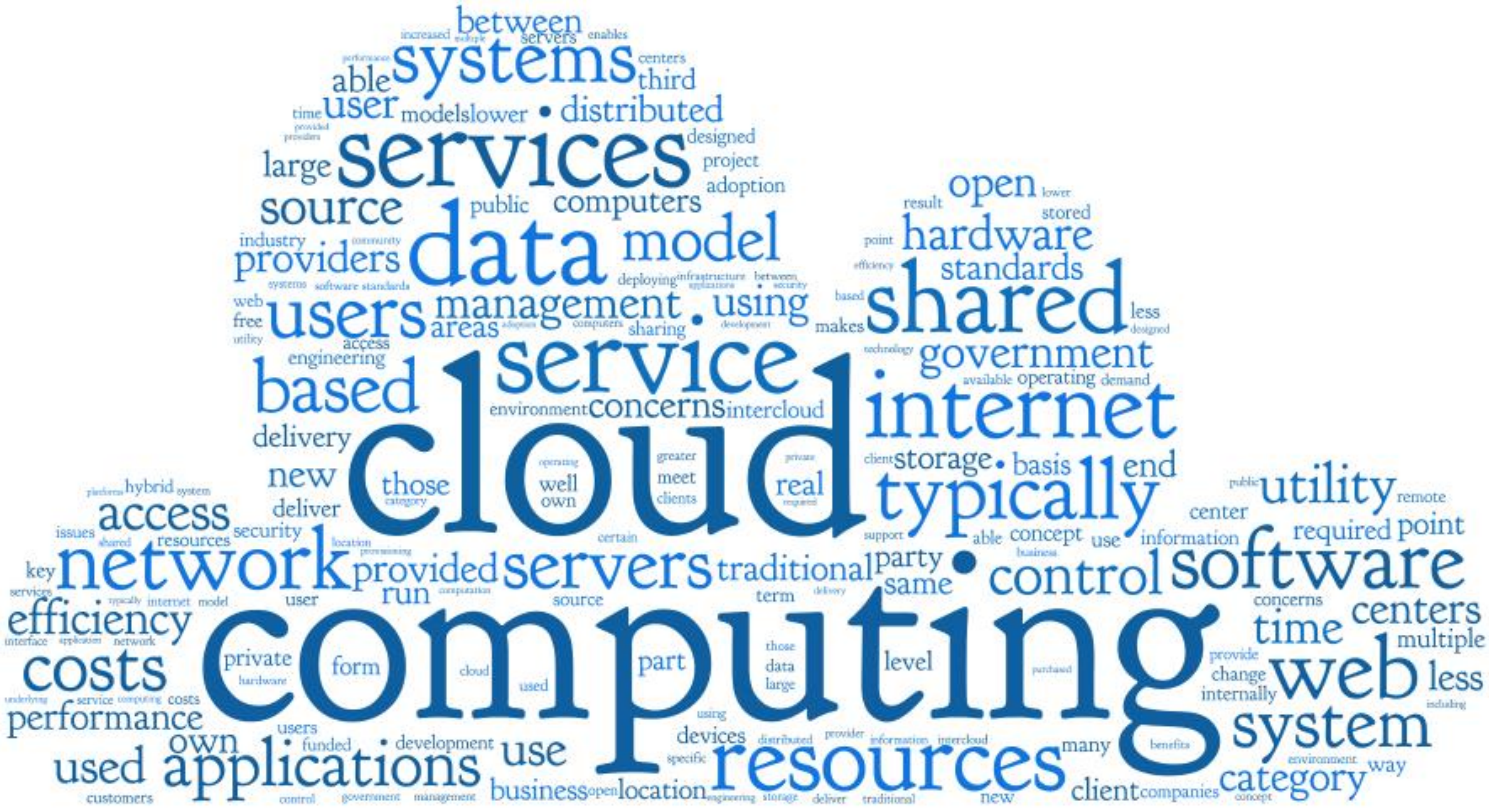

cloud
computing
services
data
service
systems
shared
internet
network
servers
resources
applications
typically
software
users
model
management
based
web
system
costs
access
efficiency
control
source
providers
between
distributed
hardware
government
utility
provided
traditional
performance
used
use
time
centers
open
standards
category
client
business
location

07
클라우드 컴퓨팅 아키텍처

07 Cloud Computing Architecture

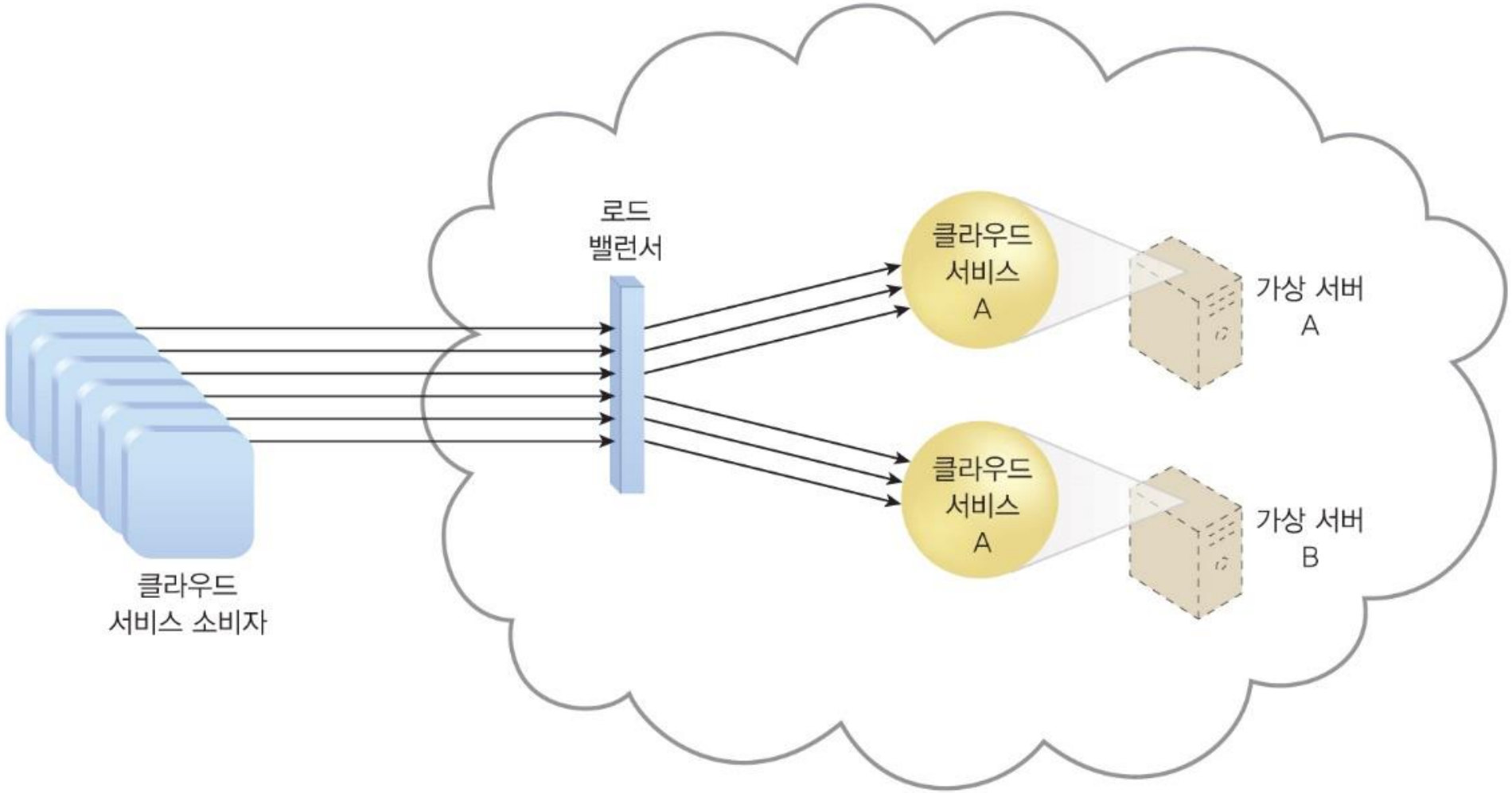
클라우드 컴퓨팅 아키텍처

- 워크로드 분배 아키텍처
- 자원 풀링 아키텍처
- 동적 확장 아키텍처
- 탄력적 자원 용량 아키텍처
- 서비스 부하 분산 아키텍처
- 클라우드 버스트 아키텍처
- 탄력적 디스크 프로비저닝 아키텍처
- 이중화 스토리지 아키텍처

워크로드 분배 아키텍처

- IT 자원은 하나 이상의 동일한 IT 자원을 추가하고, 가용한 IT 자원들간에 워크로드를 균등하게 배분할 수 있는 런타임 로직을 제공하는 로드 밸런서를 추가해서 수평적으로 확장될 수 있음
- 워크로드 분배 아키텍처는 어느 정도 부하 분산 알고리즘과 런타임 로직의 정교함을 바탕으로 IT 자원의 과잉 활용과 활용 부족 문제를 해소
- 분산된 가상 서버, 클라우드 스토리지 장치, 클라우드 서비스의 지원에 따라 일반적으로 수행되는 워크로드 분배 정책과 함께 모든 IT 자원을 대상으로 적용 가능

워크로드 분배 아키텍처



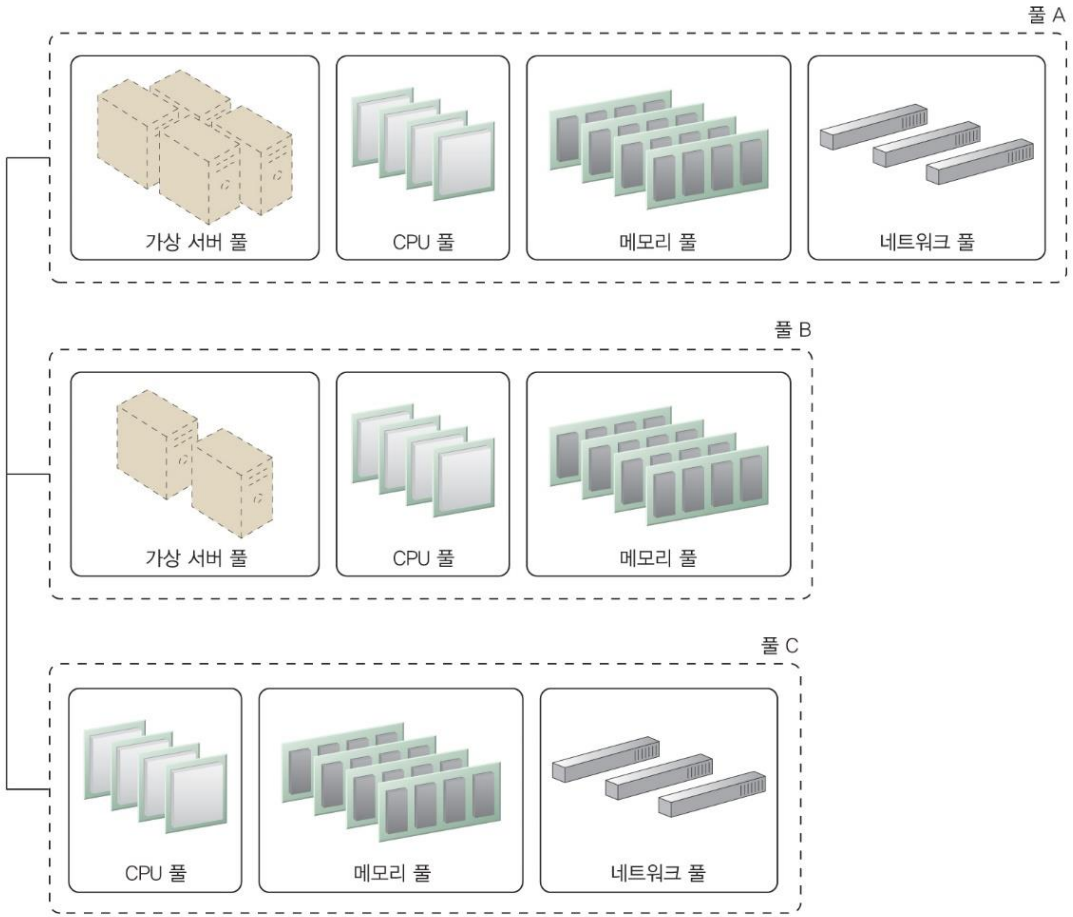
자원 풀링 아키텍처

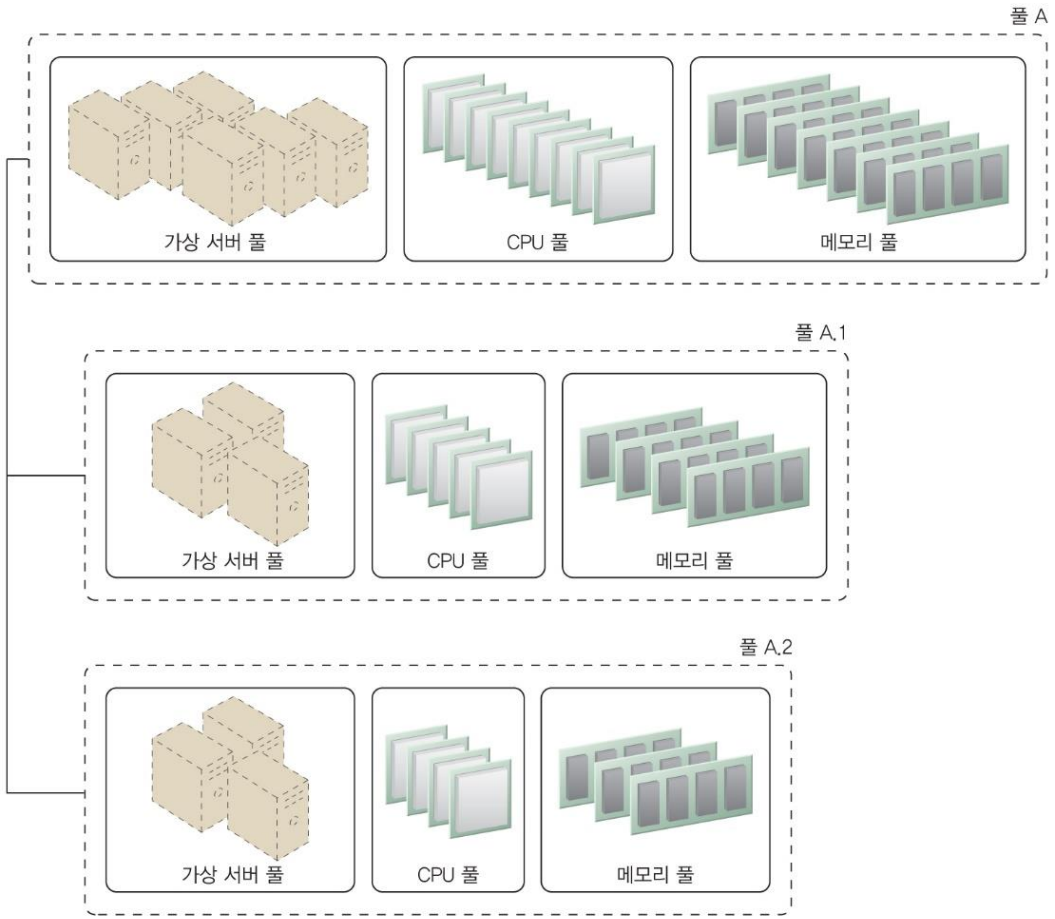
- 자원 풀링 아키텍처는 하나 이상의 자원 풀을 활용하는 것을 기반으로 함
- 자원 풀은 동일한 IT 자원들이 자동으로 동기화되도록 보장하는 시스템에 의해 그룹화 및 유지 관리됨
- 자원 풀의 종류
  - 물리 서버 풀: 운영체제와 필요한 프로그램, 혹은 애플리케이션과 함께 설치된 상호 연결된 서버들로 구성되어 있으며 즉시 사용할 수 있도록 준비된 상태
  - 가상 서버 풀: 프로비저닝 단계에서 클라우드 소비자가 선택한 여러 가용 템플릿 중 하나를 사용하여 구성
  - 스토리지 풀: 클라우드 스토리지 장치들을 포함한 파일 기반이나 블록 기반 스토리지 구조로 구성
  - 네트워크 풀: 다양한 네트워크 연결 장치들로 구성(가상 방화벽 장치나 물리 네트워크 스위치 풀은 이중화 연결, 부하 분산, 링크 집합용으로 생성 가능)
  - CPU 풀: 가상 서버에 할당될 수 있도록 준비되고, 각각의 독립된 프로세싱 코어로 나뉨
  - 메모리 풀: 새롭게 프로비저닝된 물리 서버 내 사용될 수 있고, 수직적으로 확장하기 위해 사용될 수도 있음

자원 풀링 아키텍처

- 자원 풀은 특정 클라우드 소비자나 애플리케이션용으로 다중 풀이 생성되면서 굉장히 복잡해 질 수 있음
- 다양한 자원 풀링 요구사항을 만족시킬 수 있도록 반복적으로 내포된 중첩 풀nested pool을 형성하는 계층적 구조가 만들어 질 수 있음
- 형제 자원 풀은 서로 다른 데이터 센터간에 퍼져 있는 IT 자원과는 반대로, 보통 물리적으로 그룹화된 IT 자원으로 구성. 형제 풀은 각 클라우드 소비자들이 그들 각각에게 할당된 풀로만 접근할 수 있도록 서로 고립됨
- 중첩 풀 모델에서 상대적으로 더 큰 풀은 독립적으로 동일한 종류의 IT 자원들을 같이 묶는 더 작은 풀들로 나뉨. 중첩 풀은 같은 클라우드 소비자 조직 내에서 서로 다른 부서나 그룹에 자원 풀을 할당하는 데 사용될 수 있음

자원 풀링 아키텍처


풀 A
가상 서버 풀
CPU 풀
메모리 풀
네트워크 풀
풀 B
가상 서버 풀
CPU 풀
메모리 풀
풀 C
CPU 풀
메모리 풀
네트워크 풀


풀 A
가상 서버 풀
CPU 풀
메모리 풀
풀 A.1
가상 서버 풀
CPU 풀
메모리 풀
풀 A.2
가상 서버 풀
CPU 풀
메모리 풀

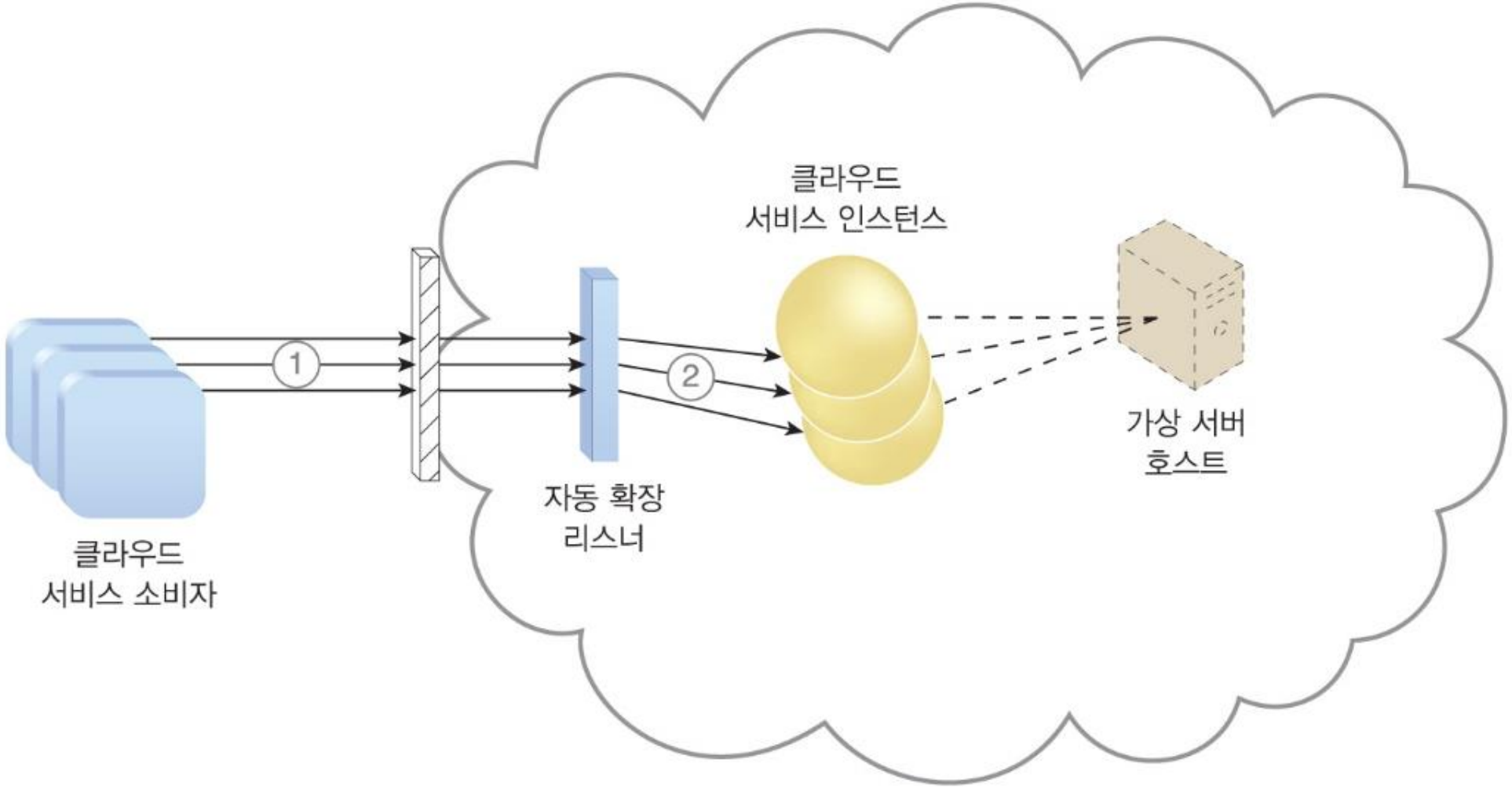
동적 확장 아키텍처

- 자원 풀로부터 IT 자원의 동적 할당을 선 정의된 확장 조건에 따라 수행하는 시스템을 바탕으로 하는 아키텍처 모델
- 동적 할당은 사용 수요 변동에 따라 할당되는 사용 용량이 달라지도록 하는데, 이는 수동적 개입이 없어도 불필요한 IT 자원들을 효율적으로 회수 가능
- 자동화된 확장 리스너가 워크로드 처리를 위해 언제 신규 IT 자원이 추가될 필요가 있는지 결정하는 워크로드 임계치와 함께 구성됨
- 클라우드 소비자의 프로비저닝 계약 조건에 따라 얼마나 많은 IT 자원이 동적으로 제공될 수 있는지를 결정하는 로직과 함께 제공 가능

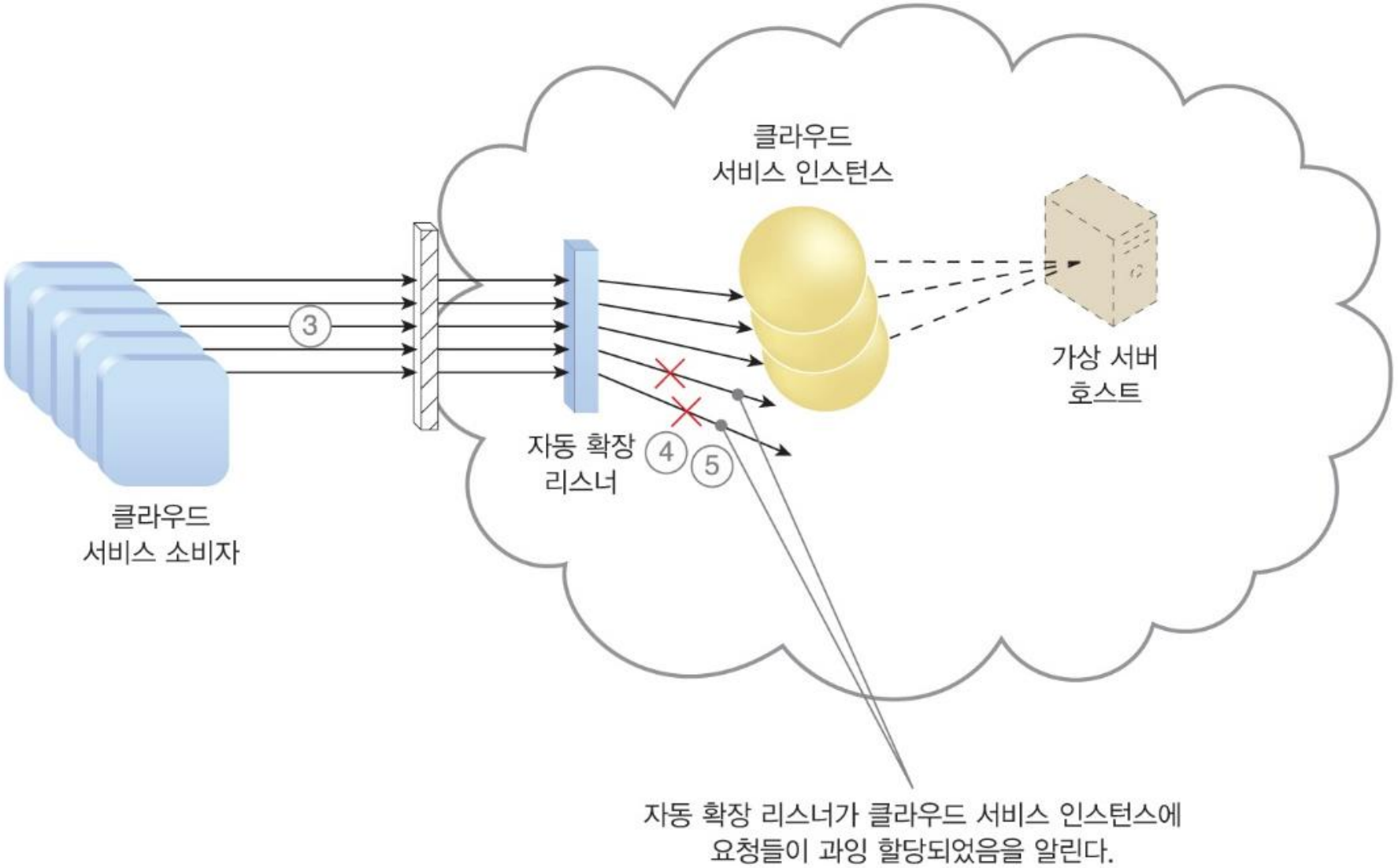
동적 확장 아키텍처

- 동적 수평 확장
  - IT 자원 인스턴스는 변동하는 워크로드를 처리하기 위해 수직 확장 및 축소
  - 자동 확장 리스너는 사용 요청들을 감시하다가 요구사항 및 권한에 따라 IT 자원 복제를 개시하기 위해 복제 시작 신호를 보냄
- 동적 수직 확장
  - IT 자원 인스턴스는 단일 IT 자원의 처리 용량을 수정할 필요가 있을 때 수직 확장 및 축소
  - 워크로드가 과잉 할당되는 가상 서버는 동적으로 메모리를 늘리거나 프로세스 코어 추가 가능
- 동적 재배치
  - IT 자원은 더 큰 용량으로 특정 호스트에 재배치
  - 예를 들어, 어떤 데이터베이스가 초당 4GB의 I/O 용량을 갖는 스토리지 장치에서 초당 8GB의 I/O 용량을 갖는 스토리지 장치로 이관될 필요가 있음

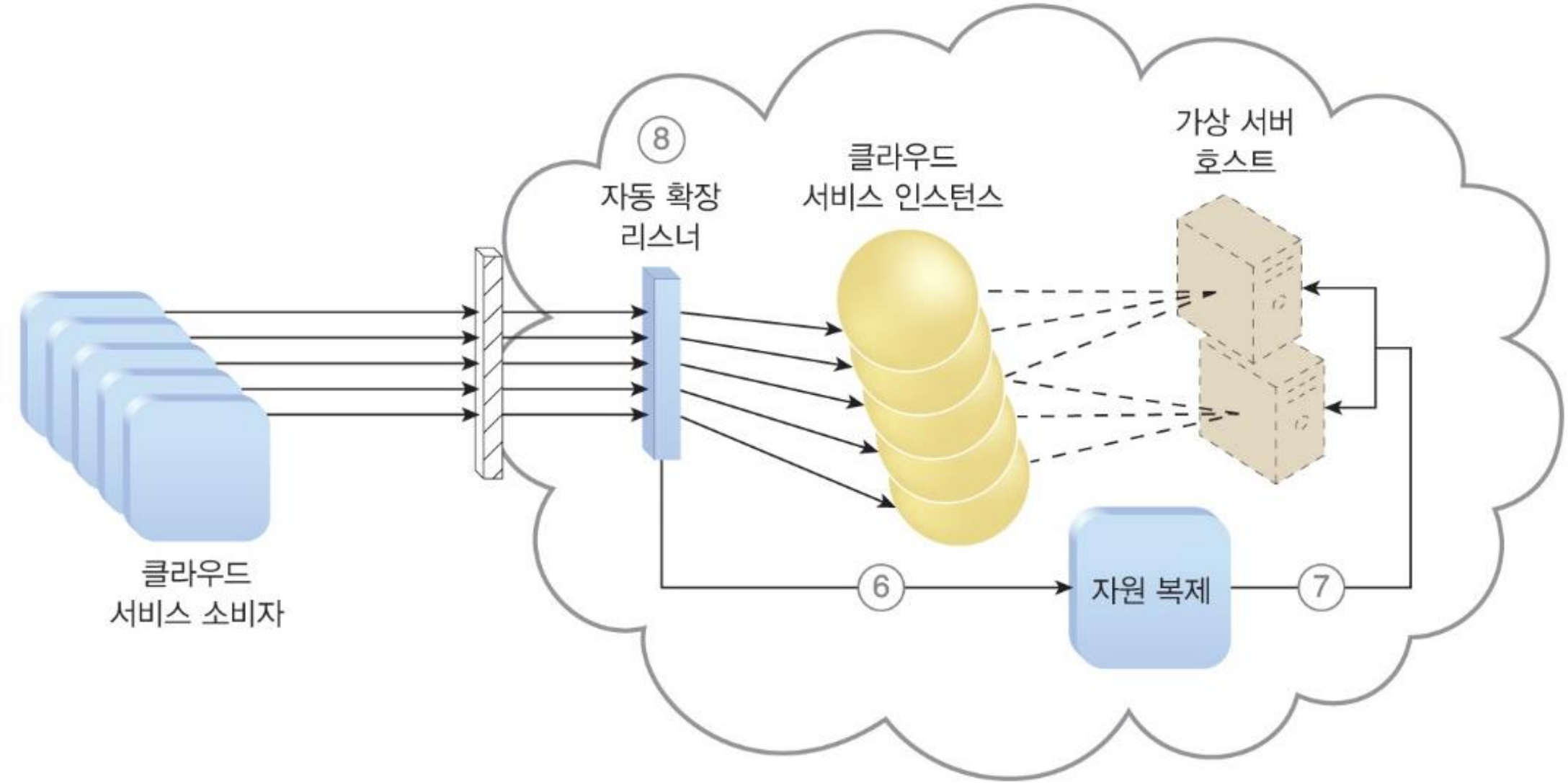
동적 확장 아키텍처



동적 확장 아키텍처


클라우드
서비스 인스턴스
가상 서버
호스트
3
자동 확장
리스너
4
5
클라우드
서비스 소비자
자동 확장 리스너가 클라우드 서비스 인스턴스에
요청들이 과잉 할당되었음을 알린다.

동적 확장 아키텍처



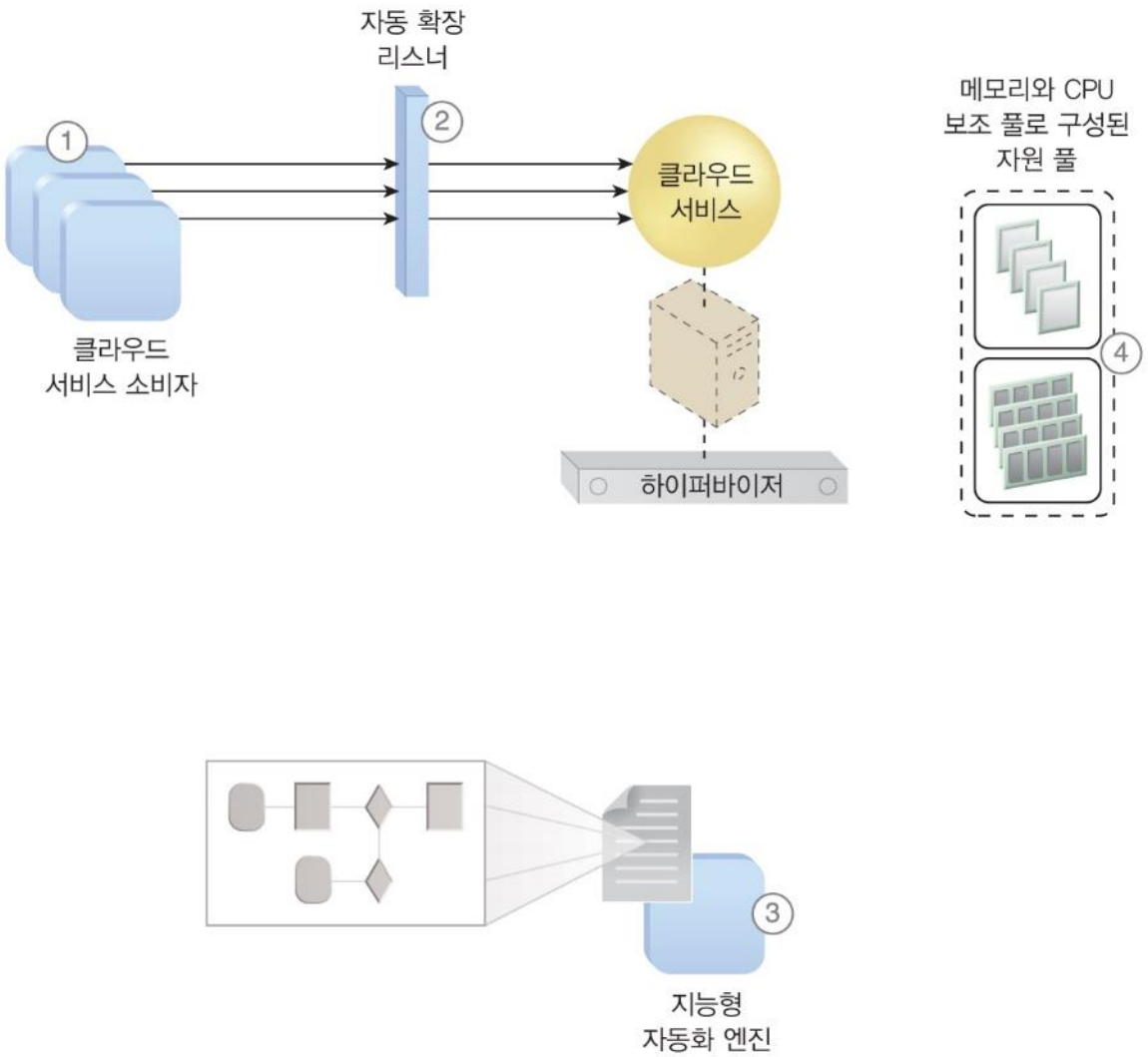
탄력적 자원 용량 아키텍처

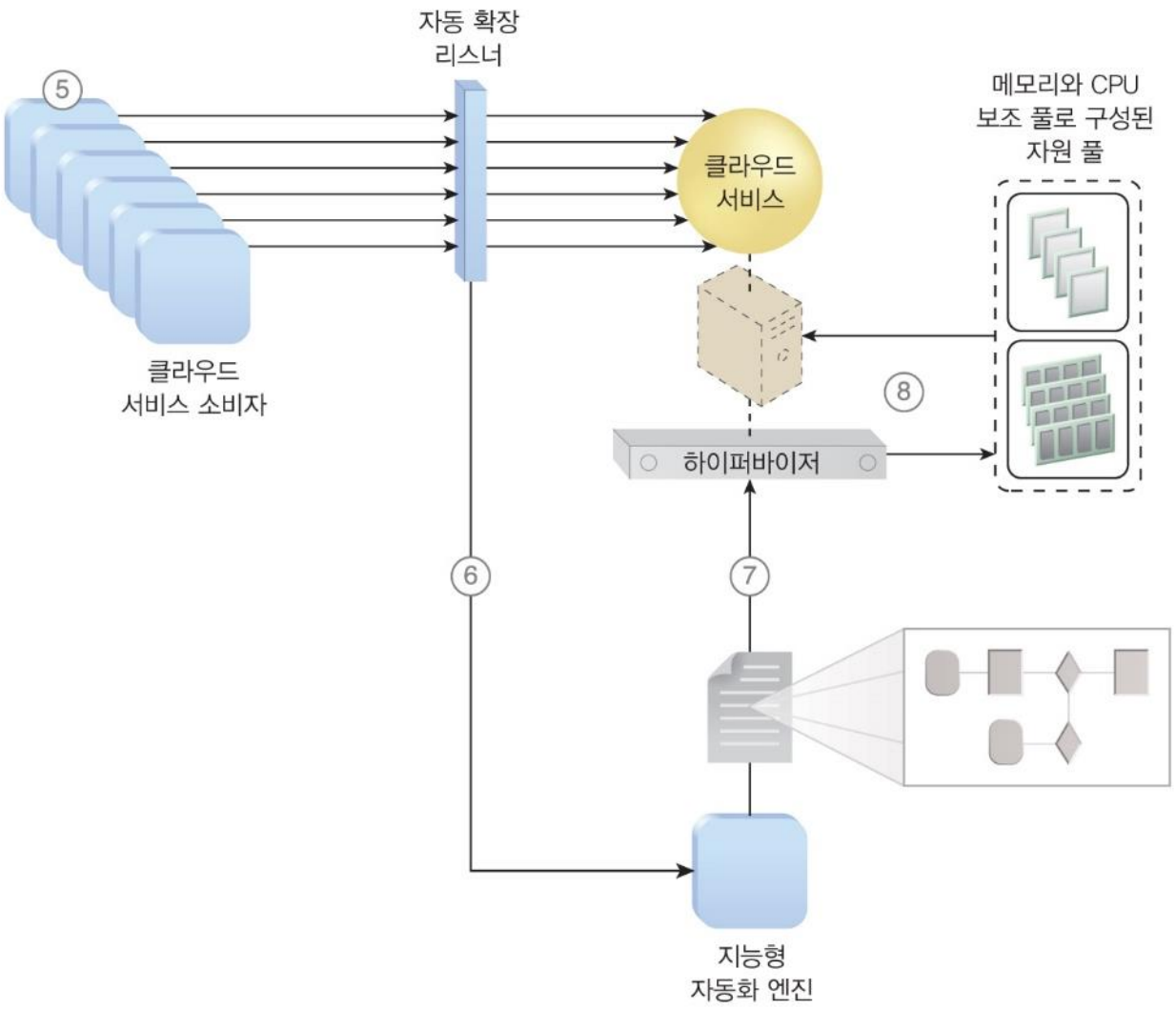
- 제공되는 IT 자원의 변동하는 처리 요구사항에 즉각적으로 대응하여 CPU와 RAM을 할당 및 회수하는 시스템에 활용하거나 동적으로 가상 서버를 프로비저닝하는 것과 관련
- 자원 풀은 런타임에 CPU와 RAM 자원을 가져오거나 되돌려놓기 위해 하이퍼바이저나 VIM과 상호작용하는 확장 기술에 의해 활용
- 용량 임계치에 도달하기 전에 동적인 할당을 통하여 자원 풀로부터 추가적인 프로세싱 파워를 가져올 수 있도록 가상 서버의 런타임 처리를 감시
- 가상 서버와 이들이 제공하는 애플리케이션, IT 자원은 수직적으로 확장
- 지능형 자동화 엔진 스크립트가 확장 요청을 하이퍼바이저에 직접 보내는 대신 VIM을 통해 보내도록 설계될 수 있음

지능형 자동화 엔진

지능형 자동화 엔진이 워크플로우 로직을 포함하는 스크립트를 수행해서 관리 작업을 자동화한다.

탄력적 자원 용량 아키텍처

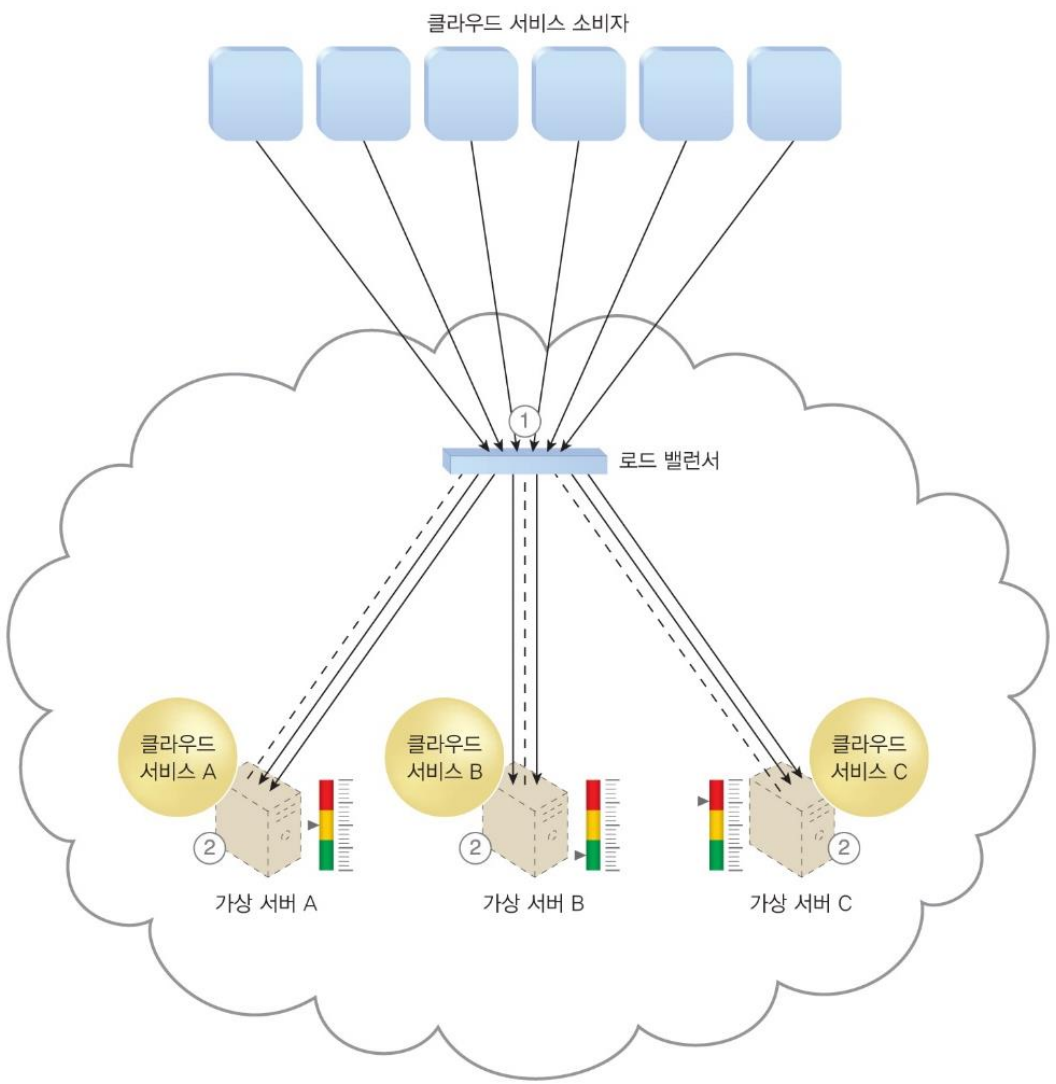


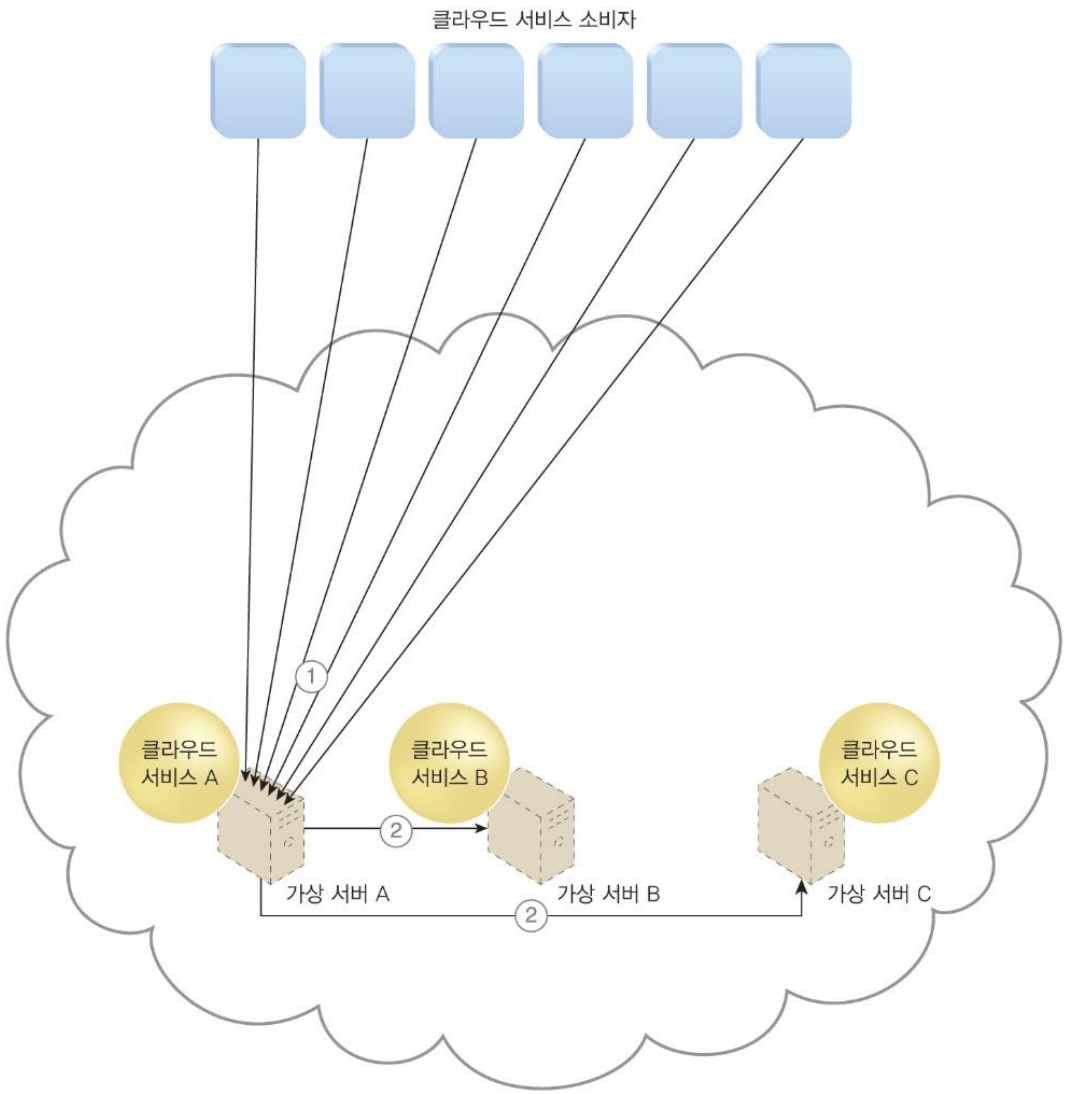


서비스 부하 분산 아키텍처

- 클라우드 서비스 구현체를 확장하기 위해 설계된 일종의 특화된 워크로드 분배 아키텍처의 변형
- 클라우드 서비스의 이중화 배치가 생성될 때는 동적으로 워크로드를 분배하기 위해 부하 분산 시스템이 함께 추가
- 호스트 서버들이 워크로드 자체의 균형을 이룰 수 있도록 로드 밸런서가 외부나 내장 컴포넌트로서 위치하고, 이중화된 클라우드 서비스 구현이 자원 풀 내에서 조직화됨
- 예상되는 워크로드의 양과 호스트 서버 환경의 처리 용량에 따라 각 클라우드 서비스 구현체의 다중 인스턴스가 자원 풀의 일부로 생성될 수 있음 (변동하는 요청량에 훨씬 더 효과적으로 대응 가능)
- 로드 밸런서가 클라우드 서비스와 호스트 서버들에 독립적으로 위치하거나 혹은 애플리케이션이나 서버 환경의 일부로 내장될 수 있음

서비스 부하 분산 아키텍처

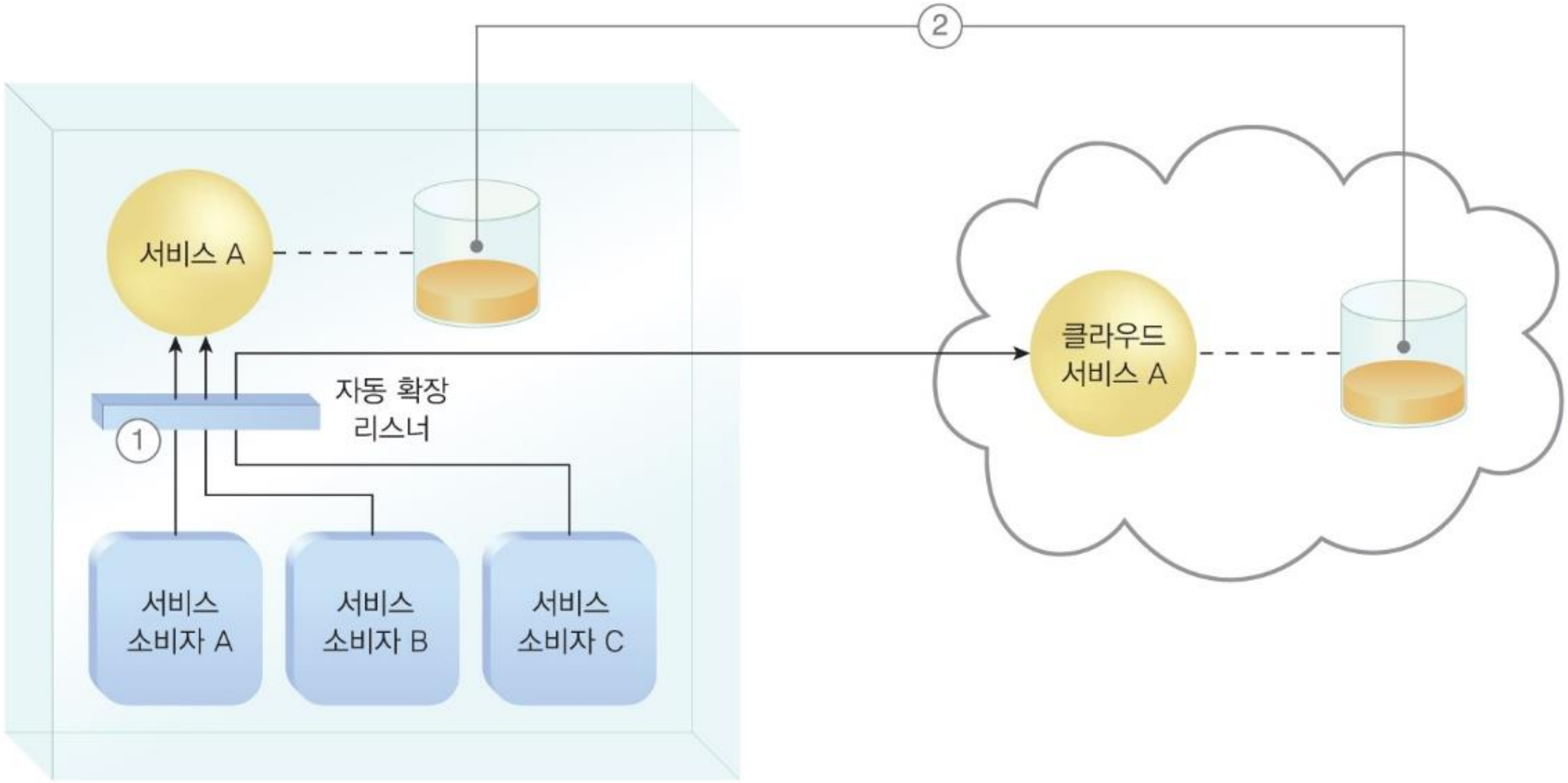




클라우드 버스트 아키텍처

- 정의된 용량 임계치에 도달할 때 온 프레미스 IT 자원들을 확장하거나, ‘버스트 아웃 burst out’하는 동적 확장 형태를 확립
- 클라우드 기반 IT 자원은 클라우드 버스트가 유발될 때까지 미리 이중화되어 배치되지만 비활성화된 상태
- 클라우드 기반 IT자원이 더 이상 필요하지 않을 때는 회수되어 아키텍처가 온 프레미스 환경에 다시 ‘버스트 인burst in’ 됨
- 클라우드 버스트는 클라우드 소비자들이 임계치 이상의 높은 사용 수요가 있을 때만 이에 대응하여 클라우드 기반 IT 자원을 활용할 수 있는 옵션을 제공하는 유연한 확장 아키텍처 (자동 확장 리스너와 자원 복제 매커니즘을 바탕으로 함)
- 자동 확장 리스너가 언제 클라우드 기반 IT 자원에 요청을 재전송할지를 결정하고, 자원 복제가 상태 정보와 관련하여 온 프레미스 및 클라우드 기반 IT 자원간의 동시성을 유지하는데 사용

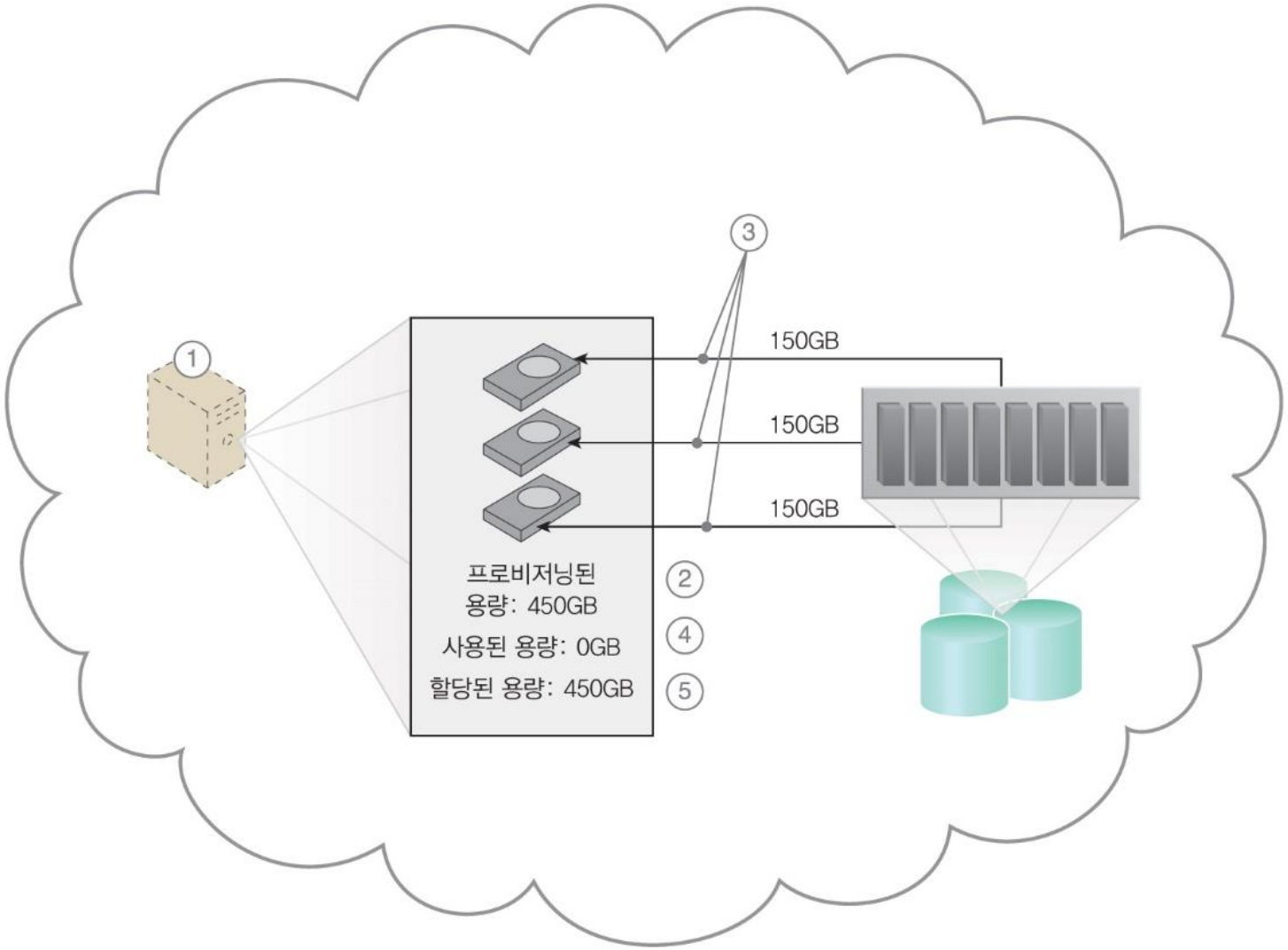
클라우드 버스트 아키텍처



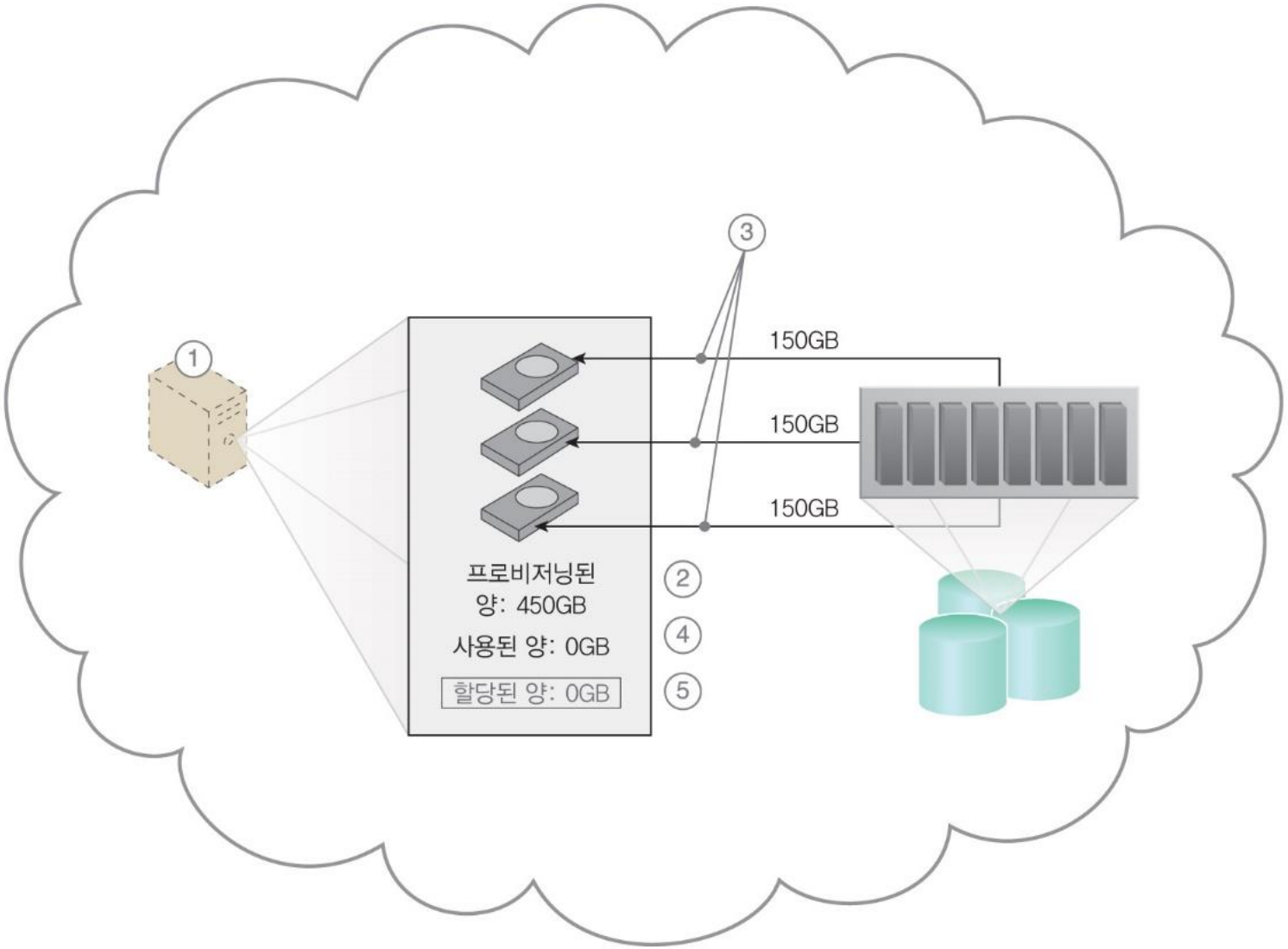
탄력적 디스크 프로비저닝 아키텍처

- 클라우드 소비자는 보통 고정 디스크 스토리지 할당에 기반한 클라우드 기반 스토리지 공간 사용량 요금이 부과
- 이는 요금이 디스크 용량에 따라 선 결정되고, 실제 데이터 스토리지 사용량에 따르지 않는다는 것을 의미
- 탄력적 디스크 프로비저닝 아키텍처는 클라우드 소비자가 점차 실제로 사용하는 정확한 스토리지의 양만큼 요금을 지불할 수 있도록 동적인 스토리지 프로비저닝 시스템을 형성
- 스토리지 공간의 동적 할당에 사용되는 씬 프로비저닝thin-provisioning 기술을 사용하고, 나아가 과금 목적으로 정확한 사용 데이터를 수집하기 위해 런타임 사용 모니터링이 사용
- 씬 프로비저닝 소프트웨어는 하이퍼바이저를 통해 동적 스토리지 할당을 처리하는 가상 서버에 설치되는데, 이때 사용량당 과금 모니터가 세세한 과금 관련 디스크 사용 데이터를 추적 및 보고

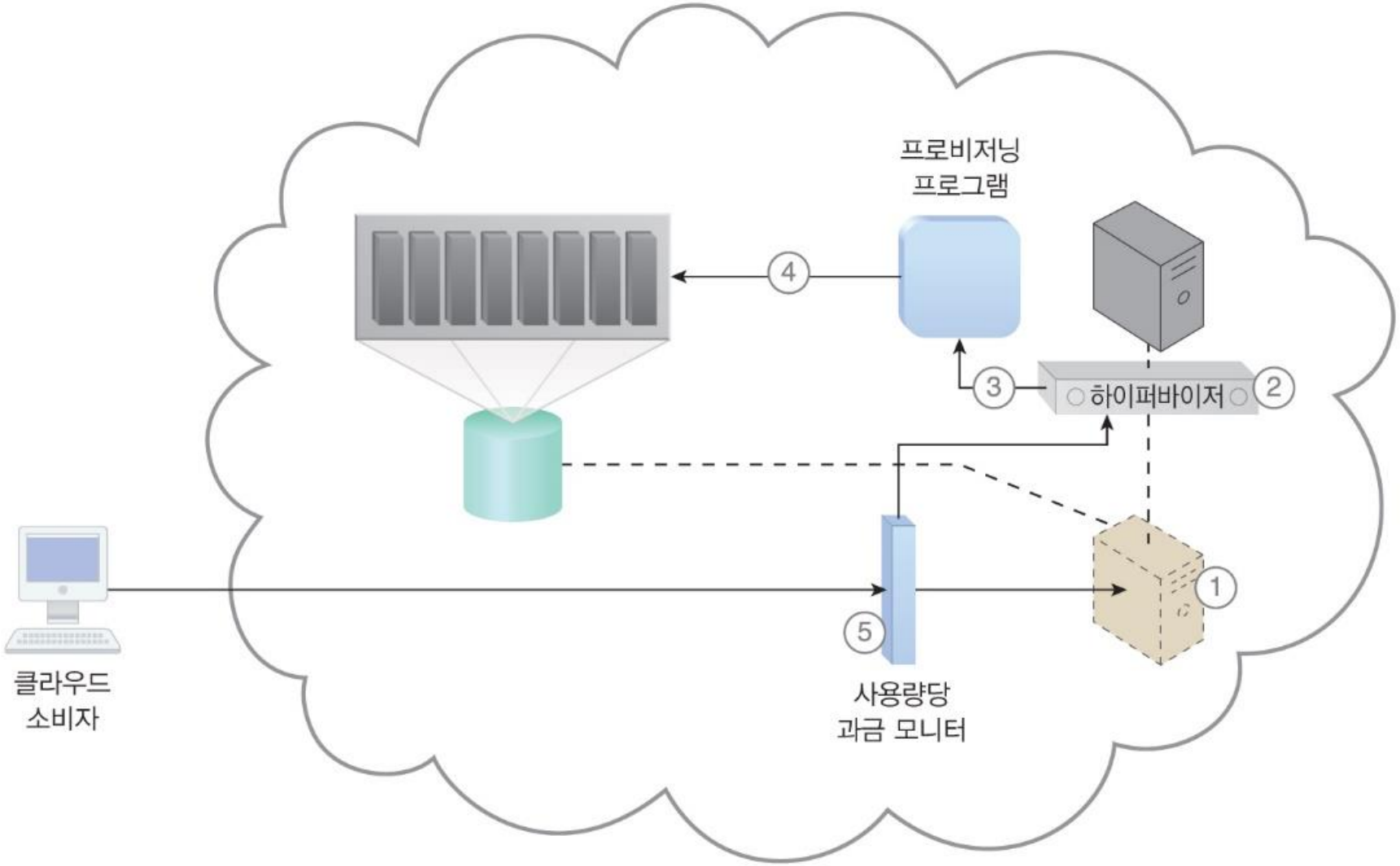
탄력적 디스크 프로비저닝 아키텍처



탄력적 디스크 프로비저닝 아키텍처



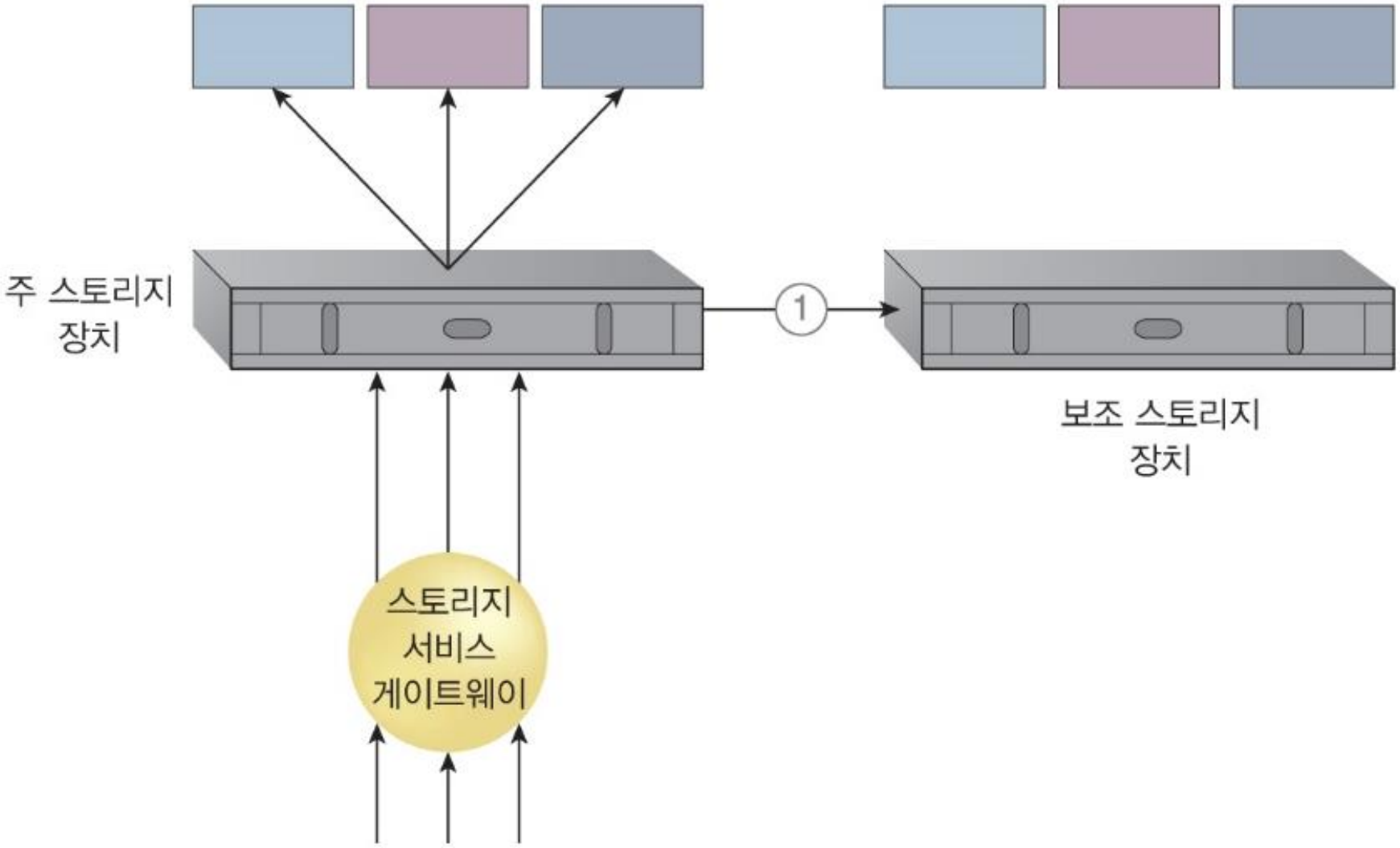
탄력적 디스크 프로비저닝 아키텍처



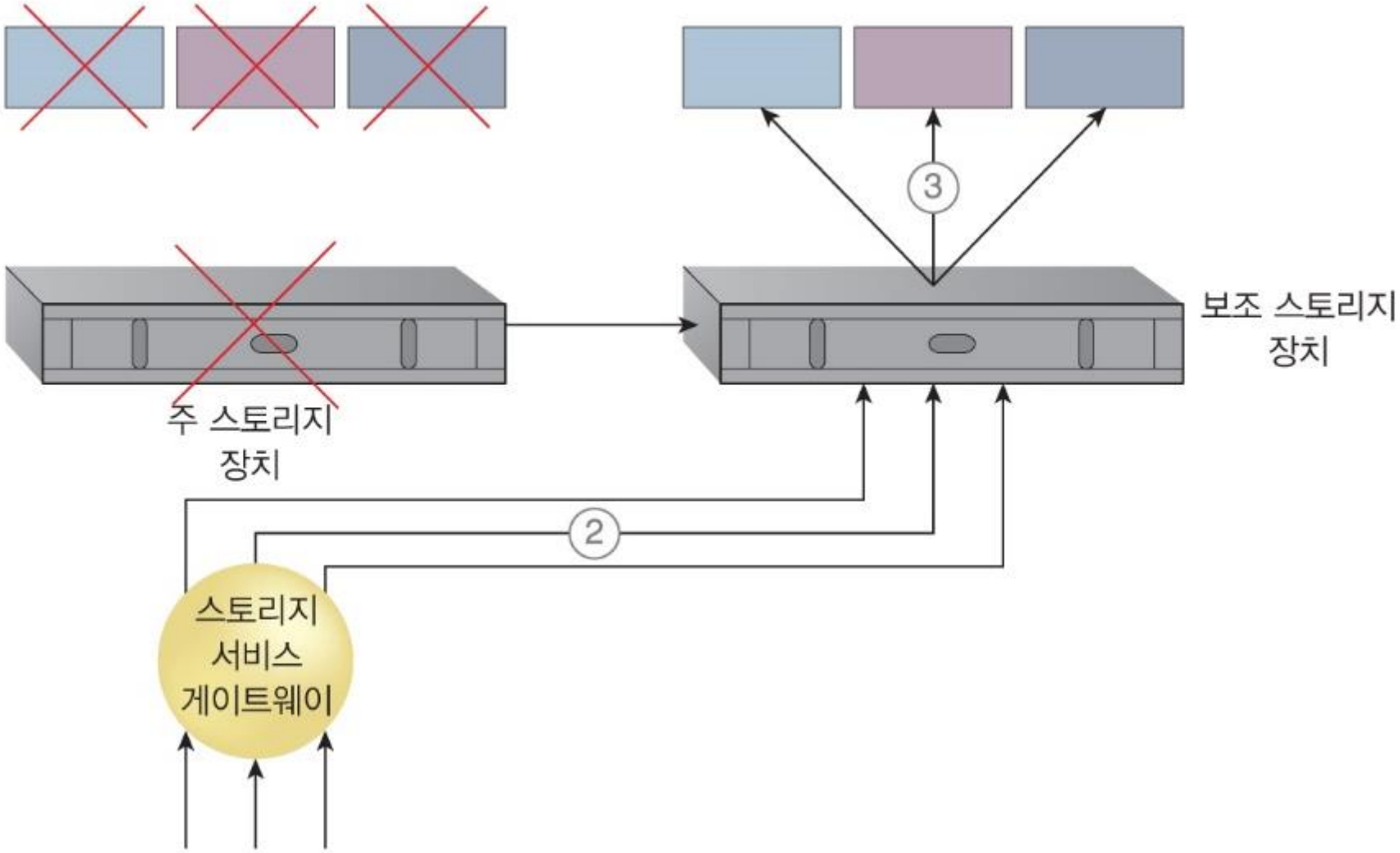
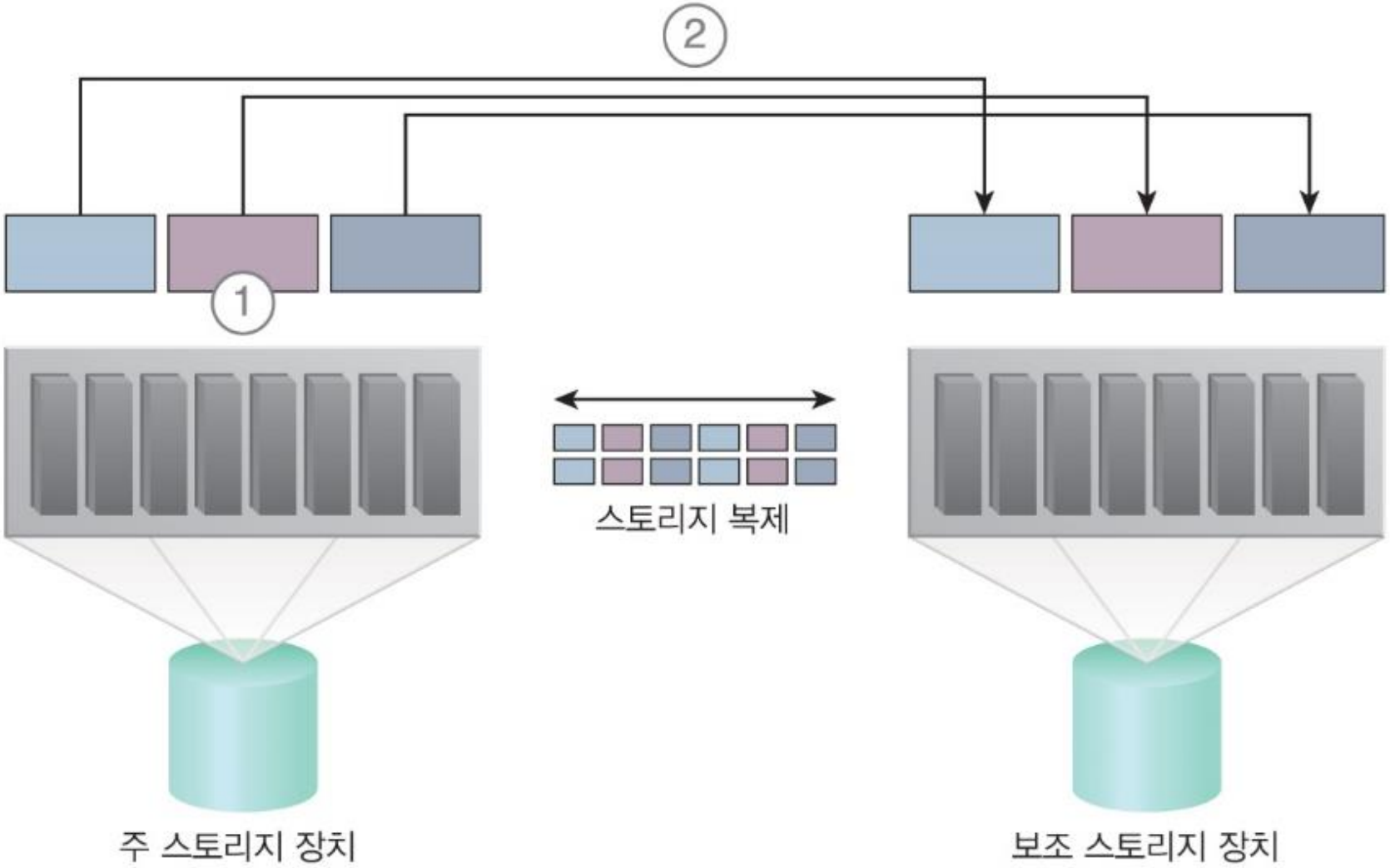
이중화 스토리지 아키텍처

- 클라우드 스토리지 디바이스는 종종 네트워크 연결 이슈나 컨트롤러, 일반 하드웨어의 장애, 보안 사고에 의해 발생한 장애 및 중단에 영향을 받음
- 클라우드 스토리지 장치에 신뢰성이 무너지면, 가용성을 의존하고 있는 모든 클라우드 내 서비스, 애플리케이션 및 인프라 컴포넌트에 파급 효과를 보이며 장애에 영향을 미칠 수 있음
- 이중화 스토리지 아키텍처는 주 클라우드 스토리지 장치 내 데이터와 데이터를 동기화하는 대체 작동 시스템failover system의 일부로 스토리지 서비스 게이트웨어는 주 장치가 중단될 경우 클라우드 소비자의 요청을 보조 장치로 전달함

이중화 스토리지 아키텍처



이중화 스토리지 아키텍처



이중화 스토리지 아키텍처