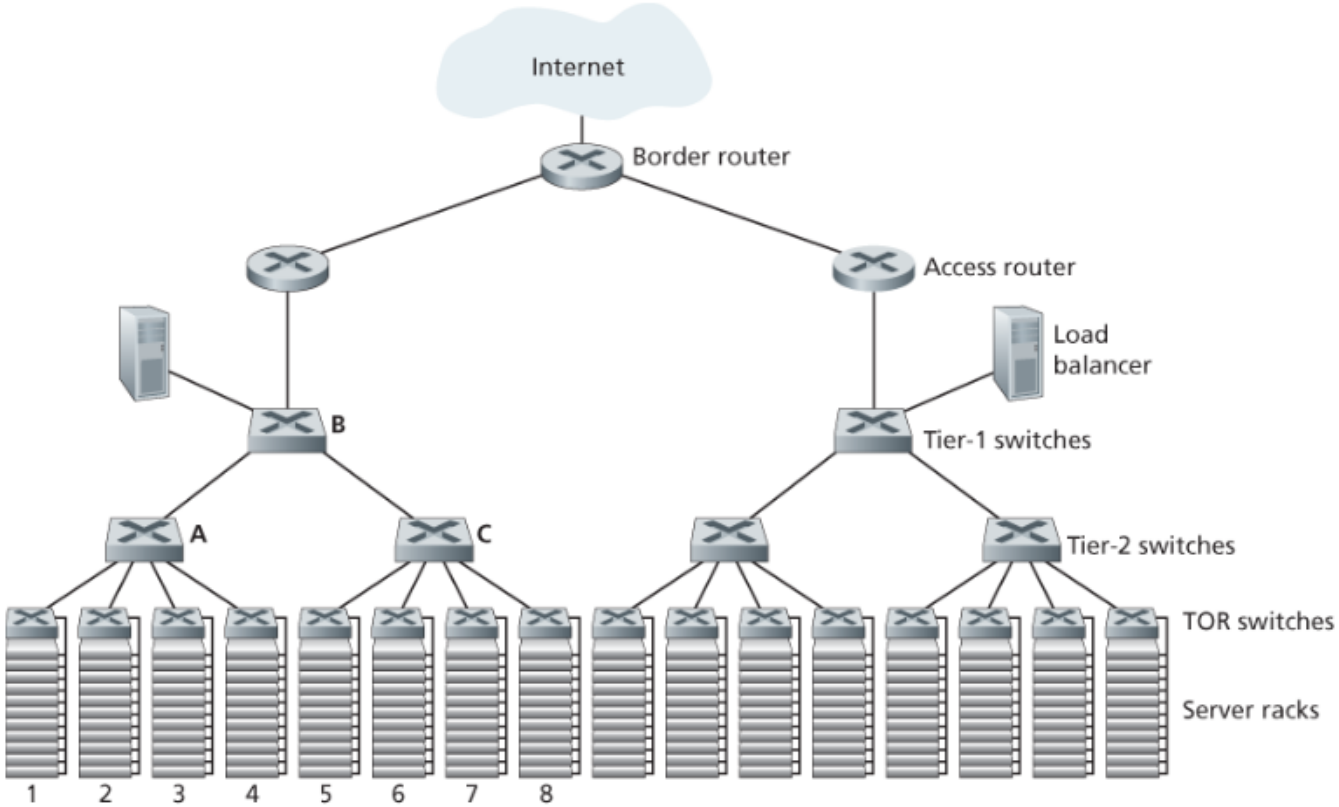
6.6 데이터 센터 네트워킹

데이터 센터는 인터넷에 연결되어 있을 뿐만 아니라 내부 호스트들 간 상호연결을 위해 자체 데이터 센터 네트워크(data center network)를 갖고 있다.

데이터 센터의 3가지 목적

1. 웹 페이지, 검색 결과, 전자메일, 스트리밍 비디오와 같은 콘텐츠 제공
2. 검색 엔진을 위한 분산 인덱스 계산과 같은 특정 데이터 처리 작업이 가능한 대량 병렬 컴퓨팅 인프라스트럭처 역할
3. 다른 회사에게 클라우드 컴퓨팅(cloud computing)을 제공

6.6.1 데이터 센터 구조



호스트

데이터 센터에서 작업을 수행한다.

피자 박스 모양의 블레이드(blade)라고도 불린다.

CPU, 메모리, 디스크 저장장치를 갖고 있는 범용 호스트다.

호스트들은 20~40대의 블레이드를 적재할 수 있는 랙(rack)에 적재된다.

데이터 센터 내부에서 사용되는 자신만의 IP 주소를 할당 받는다.

TOR(top of rack) 스위치

랙의 맨 위에는 TOR스위치라고 불리는 스위치가 있다.

TOR 스위치는 네트워크 인터페이스 카드를 통해 랙에 있는 호스트들을 연결해준다.

그 외의 다른 포트들을 통해 데이터 센터의 다른 스위치들과 연결된다.

경계 라우터(border router)

외부 클라이언트와 내부 호스트 간 트래픽 플로우를 처리하기 위해 하나 이상의 경계 라우터를 갖는다.

경계 라우터는 데이터 센터 네트워크를 공중 인터넷으로 연결해준다.

로드 밸런싱

1. 외부 클라이언트의 요청을 지원하기 위해, 애플리케이션에는 공용 IP 주소가 할당되며 클라이언트는 이 IP 주소로 요청을 보내고 응답을 받는다.
2. 요청을 로드 밸런서로 보낸다.
    - 일반적으로 여러 로드 밸런서를 갖고 있으며, 각 로드 밸런서는 특정 클라우드 애플리케이션을 위해 사용된다.
    - 로드 밸런서 는 목적지 IP 주소 뿐만 아니라 목적지 포트를 보고 결정하기 때문에 4계층 스위치라고도 한다.
3. 로드 밸런서는 요청을 호스트로 분배하고 호스트의 현재 부하 상태에 따라서 호스트 간의 부하를 균등하게 한다.
    - 로드 밸런서 는 호스트의 공용 외부 IP 주소를 내부 IP 주소로 변환해주고 그 반대 변환도 해주기 때문에 NAT과 유사한 기능을 제공한다.

클라이언트가 호스트와 직접 통신하지 못하게 하여 내부 구조를 숨기고 보안을 제공한다.

계층적 구조

1. 계층 구조의 최상단에서는 경계 라우터가 접속 라우터들에 연결된다.
2. 접속 라우터는 최상단 스위치와 연결된다.
    - 접속 라우터 아래에는 총 세 단의 스위치들이 있다.
3. 최상단 스위치는 여러 개의 두 번째 단 스위치들과 로드 밸런서에 연결된다.
4. 두 번째 단 스위치는 랙의 TOR 스위치(세 번째 단)를 통해 여러 랙으로 연결 된다.
5. 호스트들은 랙에 연결되어 하나의 서브넷을 형성한다.
    - ARP 브로드캐스트 트래픽을 지역 내로 한정하기 위해 서브넷은 다시 작은 VLAN 서브넷들로 분할된다.

클라우드 애플리케이션 제공자 입장에서는 애플리케이션 가용성을 높게 유지하는 것이 중요하기 때문에 데이터 센터 설계에 여분의 네트워크 장비와 링크를 포함시킨다.

계층적 구조의 문제

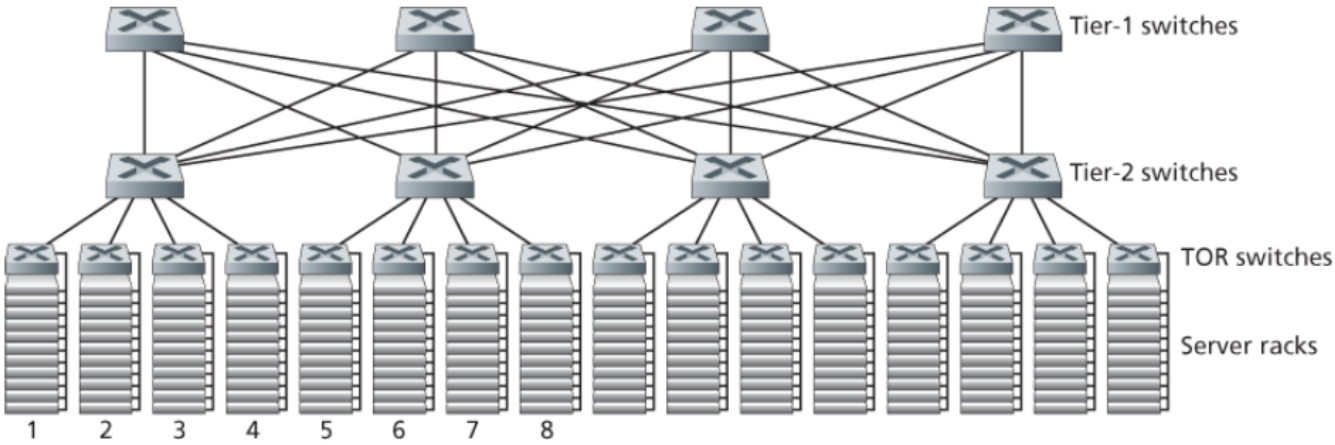
맨 위 데이터센터 그림에서 랙 1에 있는 10대의 호스트가 랙 5의 대응되는 호스트로 플로우를 보낸다고 하자.

마찬가지로 랙 2에서 6, 랙 3에서 7, 랙 4에서 8로 플로우를 보낸다고 하자.

하나의 링크를 지나가는 플로우들이 공평하게 링크 용량을 나눠서 사용하면 A에서 B로의 링크를 거쳐 가는 40개의 플로우는 각각 `n Gps/40` 만 수신하게 된다.

네트워크 인터페이스 카드의 전송률이 `n Gps/40` 보다 크다면 이는 문제가 생기고, 위쪽을 거쳐가는 호스트 간 플로우의 경우 훨씬 더 심각해진다.

해결방안



- 고속 스위치와 라우터를 사용한다.
  - 이는 비용이 많이 나간다.
- 2단 또는 1단 스위치를 경유하는 렉 간 통신이 최소화되도록 서로 관련된 서비스와 데이터를 가능한 한 같은 곳에 위치시킨다.
  - 데이터 센터의 주요 요구사항인 계산과 서비스를 융통성 있게 배치해야 한다는 것 때문에 제한적이다.
- TOR 스위치들과 2단 스위치들 간, 2단 스위치들과 1단 스위치들 간 연결성을 증가시킨다.
  - 예를 들어, 하나의 TOR 스위치를 2개의 2단 스위치에 연결함으로써 랙 간에 여러 개의 링크 `비결합(link-disjoint)`, 스위치 비결합 경로를 제공할 수 있다.
  - 단 간의 연결성(경로의 다양성)을 증가시킴으로써 스위치 간 용량 및 신뢰성 증가라는 두 가지 이득을 얻게 된다.

6.6.2 데이터 센터 네트워킹 동향

비용 감소

데이터 센터 네트워킹에서 가장 중요한 동향은 계층적으로 단을 구성해서 데이터 센터 호스트들을 상호 연결해 주는 것이다.

즉, 데이터 센터의 호스트는 다른 어떤 호스트와도 통신할 수 있게 한다.

데이터 센터 상호연결 네트워크는 다수의 소규모 스위치들로 구성된다.

중앙 집중형 SDN 제어 및 관리

데이터 센터는 단일 기관에 의해 관리되기 때문에 다수의 대규모 센터 운영자들은 SDN과 같은 논리적 중앙 집중형 제어라는 개념을 쉽게 받아들이게 된다.

SDN의 데이터 평면과 소프트웨어 기반 제어 평면에 대한 명확한 분리가 데이터 센터 구조에도 반영된다.

가상화

가상 머신(virtual machine,VM)은 소프트웨어를 실행하는 애플리케이션을 물리 하드웨어로부터 분리시켰다.

이렇게 분리하여 VM을 상이한 랙에 위치한 물리 서버들 간에 문제없이 마이그레이션할 수 있게 했다.

표준 이더넷과 IP 프로토콜은 서버들 간 활성화된 네트워크 연결을 유지한 채로 VM들을 이동시키는 것을 제한한다.

이를 해결하는 방법은 전체 데이터 센터 네트워크를 단일, 평면, 2계층 네트워크로 다루는 것이다.

모든 호스트가 단일 스위치에 연결된 것과 유사한 효과를 얻기 위해 브로드캐스트 대신 DNS 형태의 질의 시스템을 사용하도록 ARP를 수정하고 디렉토리에 VM에 할당된 IP 주소와 데이터 센터 네트워크에서 VM이 현재 연결된 물리 스위치 간 매핑 정보를 관리한다.

물리적 제약사항

광역 인터넷과 달리 데이터 센터 네트워크는 고용량, 초 저지연 환경에서 동작한다.

따라서 데이터 센터의 경우 버퍼 크기가 작으면 TCP 등과 같은 혼잡 제어 프로토콜이 효율적으로 동작하지 못한다.

손실복구와 타임아웃은 데이터 센터를 매우 비효율적으로 만들기 때문에 혼잡 제어 프로토콜은 반응이 빠르고 초 저지연으로 동작해야한다.

이러한 문제를 해결하기 위해 데이터 센터에 적합하도록 TCP를 변형한 방법부터 표준 이더넷에 RDMA(Remote Direct Memory Access)를 구현한 방법이 제안 및 적용되었다.

하드웨어 모듈화와 커스터마이징

또 다른 주요 동향은 선박 컨테이너(shipping container) 기반의 모듈화된 데이터 센터(modular data center, MDC)다.

MDC에서는 표준 12m 선박 컨테이너에 미니 데이터 센터를 구축한 후 컨테이너를 데이터 센터의 위치로 이동시킨다.

컨테이너에는 수십 개의 렉에 최대 수천 개의 호스트들이 촘촘히 포장되어 들어 있다.

데이터 센터 위치에는 여러 컨테이너들이 서로 연결되어 있고 인터넷도 연결되어 있다.

미리 제작된 컨테이너를 데이터 센터에 설치한 후에는 컨테이너에 대한 서비스를 하기 어려울 수 있다.

따라서 컨테이너 성능은 점진적으로 저하되도록 설계하고, 여러 구성요소가 고장나고 성능이 임계치 이하로 떨어지면 컨테이너를 제거하고 교체한다.

MDC에는 각 컨테이너의 내부 네트워크와 컨테이너들을 연결하는 코어 네트워크, 두 종류의 네트워크가 있다.

그러나 일상적인 작업 부하를 처리할 수 있도록 컨테이너들 간에 고속의 호스트-호스트 대역폭을 제공하면서도 수십 만대의 컨테이너들을 상호 연결해주는 코어 네트워크를 설계하는 것은 아직 난제로 남아있다.

지속적인 구축과 커스터마이징

대규모 클라우드 제공자가 네트워크 어댑터, 스위치, 라우터, TOR, 소프트웨어, 네트워킹 프로토콜 등 데이터 센터에 있는 모든 것을 계속해서 구축하거나 커스터마이징 한다.

신뢰성 확보

근처 건물들에 데이터 센터를 복제하여 가용 구역을 확보하여 신뢰성을 높인다.